Permutation Testing and Semiparametric Regression: Efficient Computation, Tests of Matrix Structure, and ℓ_1 Smoothing Penalties

by

Brian D. Segal

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2017

Doctoral Committee:

Professor Thomas Braun, Co-Chair
Professor Michael R. Elliott, Co-Chair
Professor Richard Gonzalez
Assistant Professor Hui Jiang

Brian D. Segal

bdsegal@umich.edu

ORCID iD: 0000-0002-2568-2541

Acknowledgements

I am grateful to everyone who has helped me to this point, particularly my advisors: Mike Elliott, Tom Braun, and Hui Jiang. Mike, Tom, and Hui have given me the perfect balance of freedom and guidance in my research, and have been extremely supportive and encouraging. I am extraordinarily thankful for the many stimulating and exciting conversations we have had.

I am also thankful to everyone else I have worked with at the University of Michigan, particularly Rich Gonzalez, Sarah Burgard, and Karl Jepsen. I have learned a tremendous amount from them. I also greatly enjoyed working with the CenHRS group at the Institute for Social Research, and am thankful to Trivellore Raghunathan for supervising me during my GSRA appointment and to Maggie Levenstein and Matthew Shapiro for allowing me to continue using the HRS office space after my GSRA appointment was over. If I have been productive, a large part of it is thanks to that office. I would also like to thank Maggie Hicken for collaborating with me to design a study of racism-related vigilance, which motivated the fourth chapter of this dissertation.

Finally, I would like to thank my family and friends, both in Ann Arbor and elsewhere. In particular, I would like to thank my parents for their never-ending love and wisdom; Nina Callens, who encouraged me at just the right moment to follow my own research interests, and whose own work in psychology inspired the third chapter of this dissertation; Efrén Cruz Cortés for keeping dance in my life, and much more; and Kimberly Huebner, who made the last part the best part.

Thank you.

Contents

List of Figures

List of Tables

Abstract

Part I: Permutation Testing

Chapters 1 and 2: Fast Approximation of Small p-values in Permutation Tests by Partitioning the Permutations

Researchers in genetics and other life sciences commonly use permutation tests to evaluate differences between groups. Permutation tests have desirable properties, including exactness if data are exchangeable, and are applicable even when the distribution of the test statistic is analytically intractable. However, permutation tests can be computationally intensive.

We propose both an asymptotic approximation and a resampling algorithm for quickly estimating small permutation p-values (e.g. $< 10^{-6}$) for the difference and ratio of means in two-sample tests. Our methods are based on the distribution of test statistics within and across partitions of the permutations, which we define. We present our methods and demonstrate their use through simulations and an application to cancer genomic data. Through simulations, we find that our resampling algorithm is more computationally efficient than another leading alternative, particularly for extremely small p-values (e.g. $< 10^{-30}$). Through application to cancer genomic data, we find that our methods can successfully identify up- and down-regulated genes. While we focus on the difference and ratio of means, we speculate that our approaches may work in other settings

Chapter 3: Tests of Matrix Structure for Construct Validation

Psychologists and other behavioral scientists are frequently interested in whether a questionnaire reliably measures a latent construct. Attempts to address this issue are referred to as construct validation. We describe nonparametric hypothesis testing procedures to assess matrix structures, which can be used for construct validation. These methods are based on a quadratic assignment framework, and can be used either by themselves or to check the robustness of other methods. We investigate the performance of these matrix structure tests

through simulations, and demonstrate their use by analyzing a big five personality traits questionnaire administered as part of the Health and Retirement Study. We also derive the rate of convergence for our overall test to better understand its behavior.

Part II: Semiparametric regression

Chapter 4: P-Splines with an ℓ_1 Penalty for Repeated Measures

P-splines are penalized B-splines, in which finite order differences in coefficients are typically penalized with an ℓ_2 norm. P-splines can be used for semiparametric regression and can include random effects to account for within-subject variability. In addition to ℓ_2 penalties, ℓ_1-type penalties have been used in nonparametric and semiparametric regression to achieve greater flexibility, such as in locally adaptive regression splines, ℓ_1 trend filtering, and the fused lasso additive model. However, there has been less focus on using ℓ_1 penalties in P-splines, particularly for estimating conditional means.

We demonstrate the potential benefits of using an ℓ_1 penalty in P-splines, with an emphasis on fitting non-smooth functions. We propose an estimation procedure using the alternating direction method of multipliers and cross validation, and provide degrees of freedom and approximate confidence bands based on a ridge approximation to the ℓ_1 penalized fit. We also demonstrate potential uses through simulations and an application to electrodermal activity data collected as part of a stress study.

Part I

Permutation Testing

Chapter 1

Fast Approximation of Small p-values in Permutation Tests by Partitioning the Permutations

1.1 Introduction and Motivation

Many researchers in the life sciences use permutation tests, for example, to test for differential gene expression (Doerge and Churchill, 1996, Morley et al., 2004, Stranger et al., 2005, 2007, Raj et al., 2014), and to analyze brain images (Nichols and Holmes, 2001, Bartra et al., 2013, Simpson et al., 2013). These tests are useful when the sample size is too small for large sample theory to apply, or when the distribution of the test statistic is analytically intractable. Permutation tests are also exact, meaning that they control the type I error rate exactly for finite sample size (Lehmann and Romano, 2005). However, permutation tests can be computationally intensive, especially when estimating small p-values for many tests. In this chapter, we present computationally efficient methods for approximating small permutation p-values (e.g. $< 10^{-6}$) for the difference and ratio of means in two-sample tests, though we speculate that our methods will also work for other smooth function of the means.

We denote the two groups of sample data as $\boldsymbol{x} = (x_1, \ldots, x_{n_x})'$ and $\boldsymbol{y} = (y_1, \ldots, y_{n_y})'$, with respective sample sizes n_x and n_y. We denote the full data as $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$, with total sample size $N = n_x + n_y$. Writing $\boldsymbol{z} = (z_1, \ldots, z_N)'$, we have that $z_i = x_i, i = 1, \ldots, n_x$, and $z_{n_x+j} = y_j, j = 1, \ldots, n_y$. In our setting, z_i are scalar values for all $i = 1, \ldots, N$. We use π to denote a permutation of the indices of $\boldsymbol{z}$, i.e. $\pi : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}$ is a bijection, and we denote the permuted dataset corresponding to π as $\boldsymbol{z}^* = (z_1^*, \ldots, z_N^*)'$, where $z^*_{\pi(i)} = z_i, i = 1, \ldots, N$. We use the term *correspondence* throughout this chapter, so for clarity, we define our use of the term in Definition 1.1.

Definition 1.1 (Correspondence)**.** Let $\boldsymbol{z} = (z_1, \ldots, z_N)'$ be the N-dimensional vector of observed data, and let $\pi : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}$ be a bijection (permutation) of the indices of $\boldsymbol{z}$. We say that the N-dimensional vector $\boldsymbol{z}^* = (z_1^*, \ldots, z_N^*)'$ *corresponds* to permutation π if $z_{\pi(i)}^* = z_i$ for all $i = 1, \ldots, N$.

It will also be useful to write the permuted dataset as $\boldsymbol{z}^* = (\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$, where $\boldsymbol{x}^* = (z_1^*, \ldots, z_{n_x}^*)'$ and $\boldsymbol{y}^* = (z_{n_x+1}^*, \ldots, z_N^*)'$ are the permuted group samples.

Let T be a test statistic, such that larger values are more extreme, and let $t = T(\boldsymbol{x}, \boldsymbol{y})$ be the observed test statistic. Similar to Lehmann and Romano (2005, p. 636), we denote the permutation p-value as $\hat{p} = \Pr(T \geq t|\boldsymbol{z}) = |\Psi|^{-1} \sum_{\pi \in \Psi} I[T(\boldsymbol{x}^*, \boldsymbol{y}^*) \geq t]$, where Ψ is the set of all permutations of the indices of $\boldsymbol{z}$, $|\Psi| = N!$ is the number of elements in Ψ, I is an indicator function, and for each π, $(\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$ is the corresponding permuted dataset. The randomization hypothesis (Lehmann and Romano, 2005, Definition 15.2.1) asserts that under the null hypothesis, the distribution of T is invariant under permutations $\pi \in \Psi$. This allows, for example, for the null hypothesis $H_0 : z_i \overset{\text{iid}}{\sim} P, i = 1, \ldots, N$, or more generally, for exchangeability, $H_0 : P(Z_1 = z_1, \ldots Z_N = z_n) = P(Z_1 = z_1^*, \ldots, Z_N = z_N^*)$ for all permuted datasets $\boldsymbol{z}^*$.

The set Ψ is typically too large to evaluate fully, so Monte Carlo methods are usually used to approximate $\hat{p}$. When resampling with replacement, also known as simple Monte Carlo resampling, the Monte Carlo estimate of $\hat{p}$ is $\tilde{p} = (B+1)^{-1} \left(\sum_{b=1}^{B} I\left[T_b \geq t\right] + 1 \right)$, where B is the number of resamples, and $T_b = T(\boldsymbol{x}^*, \boldsymbol{y}^*)$ for $(\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$ corresponding to the b^{th} randomly sampled permutation π_b. We refer to the above estimate as the adjusted $\tilde{p}$, because it adjusts the estimate to ensure it stays within its nominal level (Lehmann and Romano, 2005). However, for simplicity and to be consistent with other computationally efficient methods, particularly that of Yu et al. (2011), we use the unadjusted $\tilde{p}$, in which we remove the '+1' from the numerator and denominator.

While there may be many reasons for obtaining accurate small p-values, perhaps they are most often obtained in multiple testing settings, which are common in genetics. For example, in the analysis we present in Section 1.6, we analyze 15,386 genes for differential expression. With a Bonferroni correction and a type I error rate of $\alpha = 0.05$, to control the family-wise error rate (FWER), we would need to estimate p-values $< 0.05/15,386 \approx 3.25 \times 10^{-6}$. While one might want to use a different correction to control the FWER, false discovery rate (FDR), or other criteria, we would still need to calculate small p-values before implementing typical step-up or step-down procedures (for example, Holm (1979) to control FWER, or Benjamini and Hochberg (1995) to control FDR). These p-values, in combination with content area expertise and other statistical quantities, such as effect size, can be useful for prioritizing genes for further laboratory and statistical analysis.

As noted by Kimmel and Shamir (2006) and Yu et al. (2011), with simple Monte Carlo resampling, to estimate p-values on the order of $\hat{p} = 10^{-6}$ with a precision of $\sigma_{\hat{p}} = \hat{p}/10$, we need on the order of $B = 10^8$ resamples when using simple Monte Carlo resampling. For example, to separately estimate 5,000 p-values that are each on the order of 10^{-6}, we would need a total of $5,000 \times 10^8 = 5 \times 10^{11}$ resamples.

Several researchers have developed methods for reducing the computational burden of permutation tests, including Robinson (1982), Mehta and Patel (1983), Booth and Butler (1990), Kimmel and Shamir (2006), Conneely and Boehnke (2007), Li et al. (2008), Han et al. (2009), Knijnenburg et al. (2009), Pahl and Schäfer (2010), Zhang and Liu (2011), Jiang and Salzman (2012), and Zhou and Wright (2015). For comparisons with our method, we focus on the stochastic approximation Monte Carlo (SAMC) algorithm developed by Liang et al. (2007) and tailored to p-value estimation by Yu et al. (2011). Of the available methods, we found that SAMC was the most appropriate comparison, because: 1) we could directly apply it to the test static in our motivating application (see Section 1.6), 2) it is intended for very small p-values, and 3) it does not require derivations, so is more likely to be used in practice.

In this article, we propose alternative methods for quickly approximating small permutation p-values for the difference and ratio of the means in two-sample tests. Our approaches partition the permutations such that $\tilde{p}$ has a predictable trend across the partitions. Taking advantage of this trend, we develop both a closed form asymptotic approximation to the permutation p-value, as well as a computationally efficient resampling algorithm.

We find through simulations that our resampling algorithm is more computationally efficient than the SAMC algorithm, which in turn is 100 to 500,000 times more computationally efficient than simple Monte Carlo resampling (Yu et al., 2011). However, SAMC is a more general algorithm and can be used for a greater variety of statistics. The increase in efficiency is most notable for our algorithm when estimating extremely small p-values (e.g. $< 10^{-30}$). Our asymptotic approximation tends to be less accurate than our resampling algorithm but does not require resampling.

Before presenting our methods, we briefly explain the underlying properties that make them possible. The two basic components underlying our methods are 1) the partitions, which we define, and the distribution of permutations across these partitions, and 2) the limiting behavior of test statistics within each partition, and the trend in p-values across the partitions. We address the first component in Section 1.2 and the second in Section 1.3.

In Section 1.4, we introduce methods for estimating permutation p-values that take advantage of the properties discussed in Sections 1.2 and 1.3. In Section 1.5, we investigate the behavior of these methods through simulations and compare against the SAMC algorithm (additional simulations and comparisons against other methods are in Chapter 2). Then in Section 1.6, we use our proposed methods to analyze cancer genomic data. In Section 1.9, we end with a discussion of limitations and possible extensions. As noted under Supplementary material, we have implemented our methods in the R package `fastPerm`.

1.2 Partitioning the Permutations

1.2.1 Defining the Partitions

Let the smaller of the two sample sizes be $n_{\min} = \min(n_x, n_y)$. We define the distance between permutation π and the observed ordering of the indices $(1, 2, 3, \ldots, N)$ as the number of observations that are exchanged between $\boldsymbol{x}$ and $\boldsymbol{y}$ under the action of π. To be precise, let $\omega(\pi)$ be the set of indices that π places in one of the first n_x positions, i.e. $\omega(\pi) = \{i \in \{1, \ldots, N\} : \pi(i) \leq n_x\}$. Then we define the distance, denoted as $d(\pi)$, between permutation π and the observed ordering, as

$$d\,(\pi) = n_x - |\omega(\pi) \cap \{1, 2, \ldots, n_x\}|. \tag{1.1}$$

We define partition m, denoted as $\Pi(m)$, as the set of all permutations a distance of m away from the observed ordering, i.e. $\Pi(m) = \{\pi : d\,(\pi) = m\}$, $m = 0, 1, \ldots, n_{\min}$. As described below, our proposed methods focus on the permutation distributions of test statistics when resampling is restricted to permutations from a single partition.

To see why this definition of distance is useful, and to foreshadow our method, suppose that $\mu_x \neq \mu_y$, and note that as observations are exchanged between $\boldsymbol{x}$ and $\boldsymbol{y}$, the empirical distributions of the permuted samples $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ tend to become more similar. Consequently, test statistics that measure changes in the mean tend to become less extreme. For example, suppose that $n = n_x = n_y$ with n even, and let $\boldsymbol{z}^* = (\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$ be a permuted dataset corresponding to a permutation $\pi \in \Pi(n/2)$. Then half of the observations in $\boldsymbol{x}^*$ are from $\boldsymbol{x}$ and half are from $\boldsymbol{y}$, and the same is true for $\boldsymbol{y}^*$. Consequently, we would expect $\bar{x}^* \approx \bar{y}^*$, where $\bar{x}^*$ and $\bar{y}^*$ are the means of the permuted samples.

To make this explicit, and again assuming that $n = n_x = n_y$, let $\boldsymbol{\delta}_x^\pi = (\delta_{x,1}^\pi, \ldots, \delta_{x,n}^\pi)'$ and $\boldsymbol{\delta}_y^\pi = (\delta_{y,1}^\pi, \ldots, \delta_{y,n}^\pi)'$ be $n \times 1$ indicator vectors designating which observations are exchanged between $\boldsymbol{x}$ and $\boldsymbol{y}$ under the action of permutation π:

$$\delta_{x,i}^\pi = \begin{cases} 1 \text{ if } \pi(i) > n \\ 0 \text{ if } \pi(i) \leq n \end{cases}, i = 1, \ldots, n, \qquad \delta_{y,j}^\pi = \begin{cases} 1 \text{ if } \pi(n+j) \leq n \\ 0 \text{ if } \pi(n+j) > n \end{cases}, j = 1, \ldots, n.$$

Under the action of permutation π, $\bar{x}^* = n^{-1}\left[(\mathbf{1} - \boldsymbol{\delta}_x^\pi)'\boldsymbol{x} + \left(\boldsymbol{\delta}_y^\pi\right)'\boldsymbol{y}\right]$, where $\mathbf{1}$ is an $n \times 1$ vector of ones. Assuming uniform distribution of the permutations π, $\mathbb{E}\left[\boldsymbol{\delta}_x^\pi | \pi \in \Pi(m)\right] = (m/n)\mathbf{1}$, an $n{\times}1$ vector with all elements equal to m/n. Consequently, $\mathbb{E}[\bar{x}^* | \pi \in \Pi(m), \boldsymbol{x}, \boldsymbol{y}] = \bar{x} + (m/n)(\bar{y} - \bar{x})$ and $\mathbb{E}[\bar{y}^* | \pi \in \Pi(m), \boldsymbol{x}, \boldsymbol{y}] = \bar{y} + (m/n)(\bar{x} - \bar{y})$.

Then, for example, with the test statistic $T = \bar{x} - \bar{y}$, we have that $\mathbb{E}[T(\boldsymbol{x}^*, \boldsymbol{y}^*) | \pi \in \Pi(m), \boldsymbol{x}, \boldsymbol{y}] = (\bar{x} - \bar{y})(1 - 2m/n)$, where $\boldsymbol{x}^*, \boldsymbol{y}^*$ are the permuted samples corresponding to a permutation $\pi \in \Pi(m)$, $m = 0, \ldots, n$. This shows that the expected value of T is zero when for both $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ half of the observations are from $\boldsymbol{x}$ and half are from $\boldsymbol{y}$, i.e. in the $m = n/2$ partition. Similarly, the magnitude of T is $|\bar{x} - \bar{y}|$ when either none or all of the observations are exchanged between $\boldsymbol{x}$ and $\boldsymbol{y}$ (partitions $m = 0$ and $m = n$, respectively). This example demonstrates that test statistics tend to be less extreme when the permuted group samples, $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$, each contain a mixture of elements from the observed group samples, $\boldsymbol{x}$ and $\boldsymbol{y}$. Similar results hold for unbalanced sample sizes.

1.2.2 Distribution of the Partitions

Uniform sampling of the permutations π leads to a non-uniform distribution of the partitions $\Pi(m)$. The probability of drawing a permutation from partition m under uniform sampling, which we denote as $f(m), m = 1, \ldots, n_{\min}$, is given by

$$\begin{aligned} f\,(m) &\propto |\Pi(m)| && (\pi \sim \text{Uniform}) \\ &= \binom{n_x}{m}\binom{n_y}{m}, \end{aligned}$$

where the last line follows directly from the definition of $\Pi(m)$. The normalizing constant is $\sum_{j=0}^{n_{\min}} \binom{n_x}{j}\binom{n_y}{j} = \binom{N}{n_{\min}}$, so

$$f\,(m) = \binom{N}{n_{\min}}^{-1}\binom{n_x}{m}\binom{n_y}{m}. \tag{1.2}$$

As described in Section 1.4, in our proposed methods, we use f to weight the partition-specific p-values in order to obtain an overall p-value.

We note that in practice, directly using (1.2) to calculate $f(m)$ is not possible for large n_x and n_y, because the binomial coefficients become too large to represent on most computers. However, by noting the relationship between the gamma function and factorials, we can compute (1.2) for large sample sizes with the equivalent form:

$$\begin{aligned} f\,(m) = \exp\{&\log\Gamma(n_x + 1) - \log\Gamma(n_x - m + 1) \\ &+ \log\Gamma(n_y + 1) - \log\Gamma(n_y - m + 1) - 2\log\Gamma(m + 1) \\ &- \log\Gamma(N + 1) + \log\Gamma(N - n_{\max} + 1) + \log\Gamma(n_{\max} + 1)\}, \end{aligned}$$

where $\log \Gamma$ is the log gamma function.

1.3 Trend in p-values Across the Partitions

In this section, we describe the trend in p-values across the partitions both with asymptotic and simulated results. The results described in this section are given in greater detail in Section 1.8 and are the basis for our proposed methods.

Let T be a two-sided test statistic that is a function of the means, such that larger values are more extreme. In particular, we study $T = |\bar{x} - \bar{y}|$ and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$. T is a random variable, and we could calculate its value for all permutations of the data to get its permutation distribution.

We use two notations for the arguments to T: $T(\boldsymbol{x}, \boldsymbol{y})$ and $T(m)$. $T(\boldsymbol{x}, \boldsymbol{y})$ denotes the test statistic computed with data $\boldsymbol{x}, \boldsymbol{y}$, e.g. $T(\boldsymbol{x}, \boldsymbol{y}) = |\bar{x} - \bar{y}|$, and $T(m)$ denotes the test statistic computed with some permuted dataset $\boldsymbol{z}^*$, where $\boldsymbol{z}^*$ corresponds to a permutation $\pi \in \Pi(m)$. This notation facilitates further analysis in Section 1.8. We note that $\Pr\left(T(m) > t|\boldsymbol{z}\right) = \Pr\left(T(\boldsymbol{x}^*, \boldsymbol{y}^*) > t|\boldsymbol{z}, \pi \in \Pi(m)\right)$, i.e., $T(m) = T(\boldsymbol{x}^*, \boldsymbol{y}^*)$ restricted to permutations in partition m. To be concrete, we could in principle compute the partition-specific permutation p-value, $\Pr(T(m) > t|\boldsymbol{z})$, as $\hat{p}(m) = |\Pi(m)|^{-1} \sum_{\pi \in \Pi(m)} I[T(\boldsymbol{x}^*, \boldsymbol{y}^*) \geq t]$, where for each $\pi \in \Pi(m)$, $(\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$ is the corresponding permuted dataset.

While we are primarily interested in two-sided statistics T in this chapter, it helps to first note results for their one-sided counterparts, which we denote by R. In particular, $R = \bar{x} - \bar{y}$ and $R = \bar{x}/\bar{y}$. Similar to before, let $R(m) = R(\boldsymbol{x}^*, \boldsymbol{y}^*)$ restricted to permutations in partition m. As shown in Corollary 1.2 of Section 1.8, under certain regularity conditions and sufficiently large sample sizes, $R(m) \sim N(\nu(m), \sigma^2(m))$, where $\nu(m)$ and $\sigma^2(m)$ are functions of the partition m as well as the sample means and variances of $\boldsymbol{x}$ and $\boldsymbol{y}$. The regularity conditions are standard assumptions for finite sample central limit theorems and the delta method, requiring that the tails of the distributions of the data are not too large and that the derivative of R exists at the means.

As described in Corollary 1.3 of Section 1.8, a direct consequence of the limiting normality of $R(m)$ is that for n_x and n_y sufficiently large,

$$\Pr\left(T(m) \geq t|\boldsymbol{z}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right] - \Phi\left[\xi^{\text{conj}}\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right], \quad (1.3)$$

where Φ is the standard normal cumulative density function (CDF), $m_{\max} = \arg\max_m f(m)$, and ξ and ξ^{conj} are functions of the partition m and data $\boldsymbol{z}$, whose forms depend on the statistic T. The functions ξ and ξ^{conj} are identical in form but reverse the role of the means of the permuted samples $\bar{x}^*$ and $\bar{y}^*$. This accounts for the two-sided form of T. Equation 1.3 is the basis for our asymptotic approximation, which is described in Section 1.4.1.

The proof of (1.3) involves the fact that $\Pr\left(T(m) \geq t|\boldsymbol{z}\right)$, as a function of m, is approximately symmetric about $m_{\max}$. This symmetry is exact when $n_x = n_y$ and less accurate as the group sample sizes become imbalanced. Consequently, the accuracy of the approximation in (1.3) is best for equal group sample sizes and worsens as the group sample sizes become more imbalanced.

The result in (1.3) and the form for ξ and ξ^{conj} shown in Section 1.8 for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ give the smooth pattern shown in Figure 1.1 for $n_x = n_y = 100, \mu_x = \sigma_x^2 = 4$, and $\mu_y = \sigma_y^2 = 2$. In the case where $n_x \neq n_y$, the center of the trend shifts but is otherwise similar.

The smooth trend shown in Figure 1.1 is primarily an observation, though it holds with striking similarity for both $T = |\bar{x} - \bar{y}|$ and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ for a wide range of group sample sizes and parameter values. This observation is the basis for our resampling algorithm described in Section 1.4.2.

Figure 1.2 shows simulated results with $B = 10^3$ resamples within each partition for data coming from the following distributions with $n_x = n_y = 100$: Poisson with rates $\lambda_x = 4$ and $\lambda_y = 2$; exponential with rates $\lambda_x = 2$ and $\lambda_y = 1$; log normal with means $\mu_x = 2$ and $\mu_y = 1$ and variances $\sigma_x^2 = \sigma_y^2 = 1$, where μ and σ^2 are the means and variances of the log; and negative binomial with size $r_x = r_y = 3$ and probability of success $p = r/(r + \mu)$, where the means are $\mu_x = 4$ and $\mu_y = 2$. For visual comparison between theoretical and simulated results, Figure 1.1b shows the theoretical values cut off at 10^{-3}.

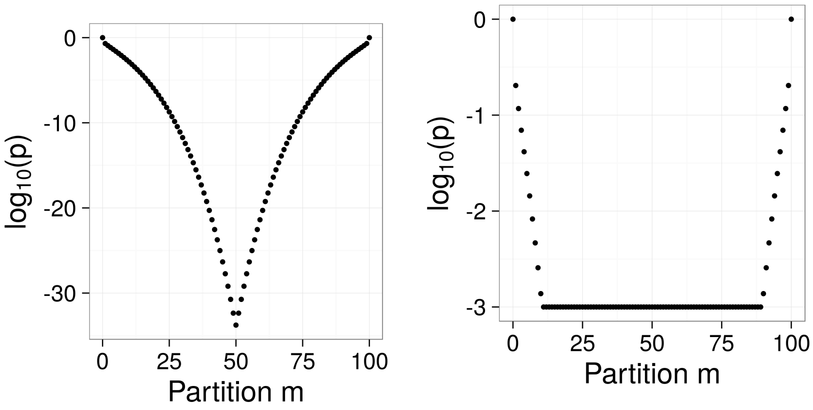
Note that the p-value for the $m = 0$ partition is always 1, as the only permutation in that partition is the observed test statistic. The same holds for partition $m = n_{\min}$ when $n_x = n_y$.

1.4 Proposed Methods

In this section, we propose two methods for approximating small permutation p-values: 1) a closed-form asymptotic approximation, and 2) a computationally efficient resampling algorithm. First, we note that we can express the permutation p-value as

$$\Pr(T \geq t|\boldsymbol{z}) = \sum_{m=0}^{n_{\min}} \Pr\left(T(m) \geq t\,|\boldsymbol{z}\right) f\left(m\right). \tag{1.4}$$

Both the asymptotic and resampling-based approaches involve approximations for the $\Pr\left(T(m) \geq t\,|\boldsymbol{z}\right)$ terms in (1.4). The asymptotic approach uses (1.3) to approximate these terms, whereas the resampling algorithm uses the trend across the partitions to predict the terms.



(a) Theoretical trend

(b) Theoretical trend cut off at 10^{-3}

Figure 1.1: Theoretical trend in p-values with $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ for $n_x = n_y = 100, \mu_x = \sigma_x^2 = 4$, and $\mu_y = \sigma_y^2 = 2$.

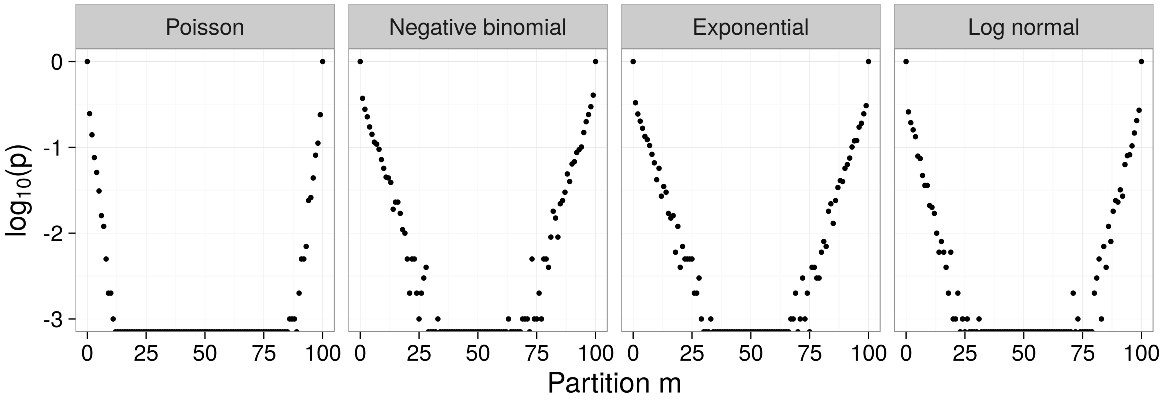


Figure 1.2: Simulated trend in p-values with $B = 10^3$ resamples within each partition and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$

If multiplicity corrections are needed, researchers can apply step-up or step-down procedures to the p-values produced by our method (e.g. Holm (1979) to control FWER, or Benjamini and Hochberg (1995) to control FDR).

1.4.1 Asymptotic Approximation

Our asymptotic approximation to the permutation p-value is given by $\hat{p}_{\text{asym}} = \sum_{m=0}^{n_{\min}} h(m)f(m)$, where $f(m)$ is given by (1.2) and

$$
\begin{aligned}
h(0) &= 1 \\
h(m) &= 2 - \Phi\left[\xi(\min\{m, 2m_{\max} - m\})\right] - \Phi\left[\xi^{\text{conj}}(\min\{m, 2m_{\max} - m\})\right], \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad m \in [1, n_{\min} - 1] \\
h(n_{\min}) &= \begin{cases} 1 & \text{if } n_x = n_y \\ 2 - \Phi\left[\xi(\min\{m, 2m_{\max} - m\})\right] - \Phi\left[\xi^{\text{conj}}(\min\{m, 2m_{\max} - m\})\right] & \text{otherwise} \end{cases}
\end{aligned}
$$

To see why $h(0) = 1$ always and $h(n_{\min}) = 1$ when $n_x = n_y$, note that the p-value is always 1 in the $m = 0$ partition, because this partition only contains the observed permutation. The same is true for the $n_{\min}$ partition when $n_x = n_y$, as T is a two-sided statistic.

Regarding notation, we use a hat in $\hat{p}_{\text{asym}}$ as opposed to a tilde to emphasize that we are not using Monte Carlo methods.

1.4.2 Resampling Algorithm

As noted in Section 1.3, we could in principle estimate each $\Pr(T(m) \geq t|\boldsymbol{z})$ term in (1.4) with Monte Carlo methods, but this would be more computationally intensive than directly estimating $\Pr(T \geq t|\boldsymbol{z})$ without conditioning on the partition. This is because for small p-values, $\Pr(T(m) \geq t|\boldsymbol{z})$ terms for m near $m_{\max}$ (the middle partition when $n_x = n_y$) are very small, so we would need to use an extremely large number of resamples to estimate these values (e.g. see Figure 1.1a).

However, by taking advantage of the trend in p-values across the partitions, we can avoid directly calculating $\Pr(T(m) \geq t|\boldsymbol{z})$ for m near $m_{\max}$. Instead, we use simple Monte Carlo resampling to estimate $\Pr(T(m) \geq t|\boldsymbol{z})$ sequentially for $m = 1, 2, \ldots, m_{\text{stop}}$, where m_{stop} is the stopping partition, which, as described below, is determined dynamically. We then use a Poisson model to predict the $\Pr(T(m) \geq t|\boldsymbol{z})$ terms for the remaining partitions (as well as for partitions $m = 1, \ldots, m_{\text{stop}}$) under the assumption that the log of the partition-specific p-values is linear in m.

We then take a weighted sum across the predicted partition-specific p-values, as in (1.4), to obtain an overall p-value. We denote the resulting p-value as $\tilde{p}_{\text{pred}}$, where the tilde emphasizes the use of Monte Carlo methods and the subscript emphasizes that the estimate is based on predicted counts within each partition.

As described in Algorithm 1, we set the number of Monte Carlo resamples within partitions at B_{pred} (e.g. we use $B_{\text{pred}} = 10^3$) and estimate $\Pr(T(m) > t|\boldsymbol{z})$ for $m = 1, \ldots, m_{\text{stop}}$, where m_{stop} is the first partition in which none of the resampled statistics are larger than the observed statistic. We stop at partition m_{stop} because the exponential decrease in p-values across the partitions, shown in Figure 1.1a, makes it nearly certain that we would not obtain a p-value greater than zero in partitions larger than m_{stop} using only $B_{\text{pred}} = 10^3$ resamples. In other words, it would be a waste of resources to continue sampling from additional partitions. Furthermore, since the trend is symmetric about $m_{\max}$, we can estimate the p-values in partitions $m = m_{\max} + 1, \ldots, n_{\min}$ using the p-values in partitions $m = 1, \ldots, m_{\max}$.

Regarding the Poisson model, this is a natural choice for count data (the number of resampled statistics larger than the observed statistic within each partition) and also enforces a log-linear trend. Furthermore, we found that Poisson regression worked best in the simulations. In addition to our current approach of using a slope and intercept term in the Poisson model, we experimented with using higher order polynomials and B-splines and selecting the optimal order or degrees of freedom based on AIC. However, we found that this approach was too sensitive to noise in the data and sometimes gave highly erroneous results (e.g. p-values > 1).

In Algorithm 1, we represent vector indices by square brackets $[\cdot]$ and begin the index at zero because our partitions begin at $m = 0$. We use the vector $\boldsymbol{c}$ to store the count of permuted test statistics in each partition that are as large or larger than the observed test statistic as obtained with simple Monte Carlo resampling and use $\boldsymbol{c}_{\text{pred}}$ to store predicted counts based on a fitted model. We use B_{pred} to denote that number of resamples within each partition.

Algorithm 1 $\tilde{p}_{\text{pred}}$

```
set $m \leftarrow 1$ and $\boldsymbol{c}[0] \leftarrow B_{\text{pred}}$
while ($m \leq m_{\max}$ and $\boldsymbol{c}[m-1] > 0$) do
    for $b = 1, \ldots, B_{\text{pred}}$, sample $\pi_b \in \Pi(m)$ uniformly and calculate $T_b(m) = T(\boldsymbol{x}^*, \boldsymbol{y}^*)$ for $\boldsymbol{x}^*, \boldsymbol{y}^*$ corresponding to $\pi_b$
    set $\boldsymbol{c}[m] \leftarrow \sum_b I[T_b(m) \geq t]$ and update $m \leftarrow m + 1$
end while
set $m_{\text{stop}} \leftarrow m - 1$ and $m_{\text{reg}} \leftarrow \max_m \{m \in \{1 \ldots, m_{\max}\} : \boldsymbol{c}[m] > 0\}$
regress $\boldsymbol{c}[0 : m_{\text{reg}}]$ on $(0, \ldots, m_{\text{reg}})$ using a Poisson model with slope and intercept terms
predict $\boldsymbol{c}_{\text{pred}}$ for $m = 1, \ldots, n_{\min}$ with fitted model, $s.t.$ $\boldsymbol{c}_{\text{pred}}$ is symmetric about $m_{\max}$
set $\boldsymbol{c}_{\text{pred}}[0] \leftarrow B_{\text{pred}}$, and if $n_x = n_y$, then set $\boldsymbol{c}_{\text{pred}}[n_x] \leftarrow B_{\text{pred}}$
return $\tilde{p}_{\text{pred}} \equiv (1/B_{\text{pred}}) \sum_{m=0}^{n_{\min}} \boldsymbol{c}_{\text{pred}}[m] f(m)$
```

Our proposed algorithm runs in $O(B_{\text{pred}} m_{\text{stop}})$ time. As described in Section 1.7, we provide functions for estimating m_{stop}, and thus run-time, prior to running the algorithm.

1.5 Simulations

To investigate the behavior of our proposed methods, we conducted simulations with the statistics $T = |\bar{x} - \bar{y}|$ and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$. Given the extremely small p-values in our simulations, it was not feasible to compute the true permutation p-values for comparison. Instead, we used asymptotically equivalent p-values and large sample sizes.

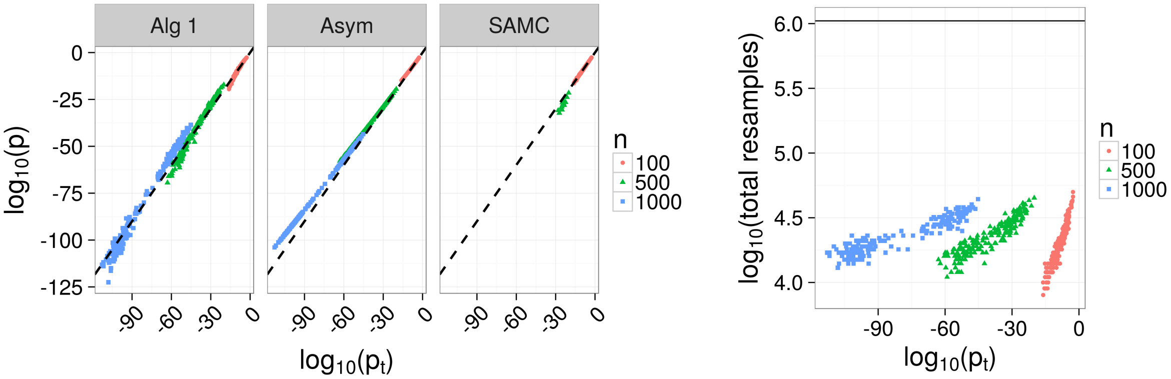
In Section 2.3, we show results from additional simulations for 1) small sample sizes, and 2) data generated under the null hypothesis, in which case we approximated the true permutation p-value with simple Monte Carlo resampling, and 3) data generated as Gamma random variables. In Section 2.4, we also show simulations with the moment-corrected correlation (MCC) method of Zhou and Wright (2015) using the statistic $T = |\bar{x} - \bar{y}|$, and compare our method with saddle point approximations (Robinson, 1982) by analyzing two small datasets ($n_x = n_y = 8$ and $n_x = 7, n_y = 10$), also using the statistic $T = |\bar{x} - \bar{y}|$. In Section 2.5, we show simulation results using our method with a studentized statistic to test null hypotheses regarding a single parameter as opposed to the full distribution, as described by Chung and Romano (2013). The results in Sections 2.3 and 2.4 show that the accuracy of our method is comparable to alternative methods, and the results in Section 2.5 show that by using a studentized statistic, our method can be extended to null hypotheses specifying equality in the means ($H_0 : \mu_x = \mu_y$), as opposed to equality in the entire distributions ($H_0 : P_x = P_y$).

1.5.1 Difference in Means

In this section, we consider the test statistic $T = |\bar{x} - \bar{y}|$ with normally distributed data of equal variance. Since the t-test is asymptotically equivalent to the permutation test in this setting (Lehmann and Romano, 2005, p. 642-643), we used the t-test as a baseline for comparison. We simulated data with both equal and unequal sample sizes ($n_x = n_y$ and $n_x \neq n_y$). In both cases, we generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(\mu_x, 1)$ and $Y_j \overset{\text{iid}}{\sim} N(\mu_y, 1)$ for various parameter values. For each combination of parameter values, we generated 100 datasets.

For equal sample sizes, we set $n = n_x = n_y = 100$, 500, or 1,000. For unequal sample sizes, we set $n_y = 500$, and $n_x = 50$, 200, or 350. In both cases we set $\mu_y = 0$ and $\mu_x = 0.75$ or 1. For each dataset, we applied our methods and did a t-test with the `t.test` function in R (R Core Team, 2017) (two-sided with equal variance). For our resampling algorithm, we used $B_{\text{pred}} = 10^3$ resamples in each partition.

For comparison, we also ran the SAMC algorithm using the R package `EXPERT` written



(a) p-values

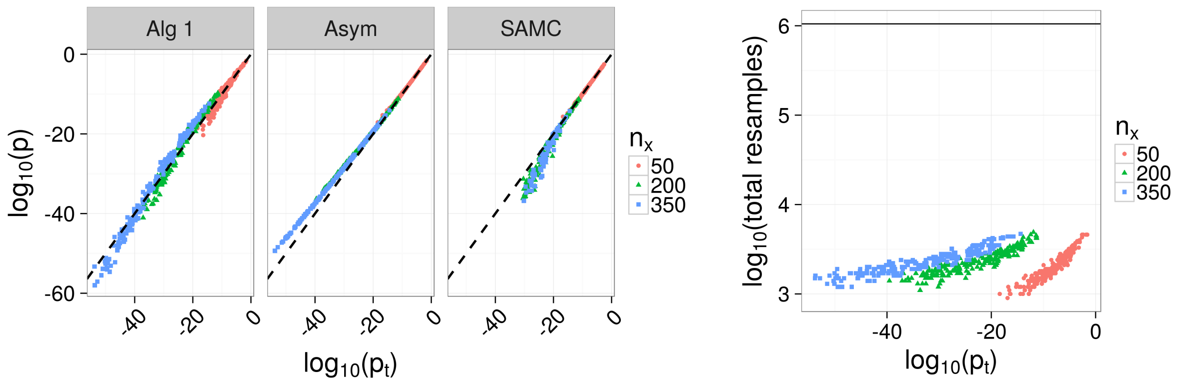
(b) Number of resamples in Alg 1

Figure 1.3: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with equal sample sizes of $n = n_x = n_y =$ 100, 500, 1,000. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 1.3b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 385 tests (points missing).

by Yu et al. (2011). We set the number of iterations (also resamples) in the initial round at 5×10^4 and the number of iterations in the final round at 10^6. Following the advice of Yu et al. (2011), we set the gain factor sequence to begin decreasing after the $1,000^{th}$ iteration, the proportion of data to be updated at each iteration at 0.05, and the number of regions at 101 for the initial run and 301 for the final run.

Results are shown in Figures 1.3 and 1.4. In the Figures, p_t denotes the p-value from a two-sided t-test with equal variance, and p denotes the p-value from either our methods or SAMC. The dashed line has a slope of 1 and intercept of 0, and indicates agreement between methods. The SAMC algorithm did not produce values for smaller p-values due to numerical problems, so these points are missing from Figures 1.3 and 1.4 (385 missing points in Figure 1.3, and 179 missing points in Figure 1.4). In order to estimate these points with the `EXPERT` implementation of the SAMC algorithm, we would need to increase the number of iterations.

As Figures 1.3 and 1.4 show, our resampling algorithm and asymptotic approximation are able to estimate extremely small p-values, which the SAMC algorithm is not able to estimate even though we set it to use approximately two orders of magnitude more resamples than our resampling algorithm. While our asymptotic approximation has less variance than our resampling algorithm, the asymptotic approximation appears to have more bias. We note that the scales are not the same in Figures 1.3 and 1.4, but in both cases, the p-values are



(a) p-values

(b) Number of resamples in Alg 1

Figure 1.4: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with unequal sample sizes, where $n_y = 500$ and $n_x = 50, 200, 350$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 1.4b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 179 tests (points missing).

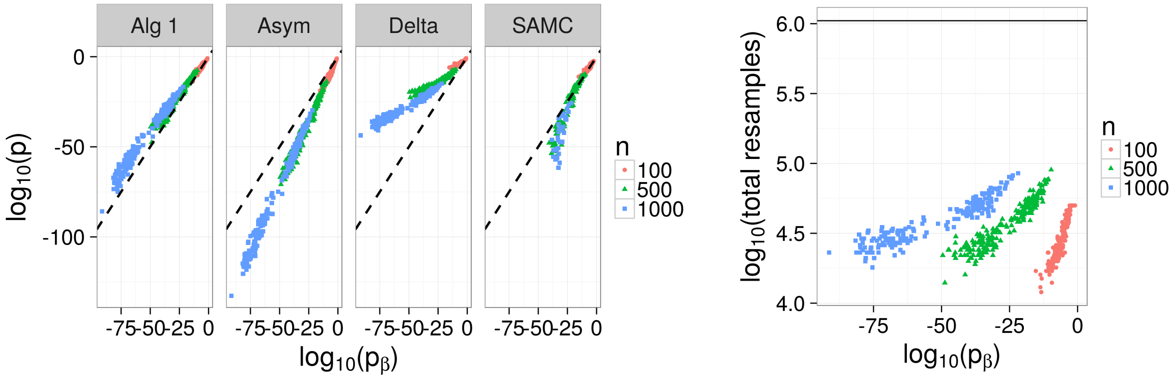
smaller than what would typically be estimated with resampling methods.

Figures 1.3b and 1.4b also demonstrate that our algorithm uses fewer permutations when estimating smaller p-values than when estimating larger p-values. This occurs because the trend in partition-specific p-values across the partitions tends to be steeper for smaller overall p-values, which leads to earlier stopping times.

1.5.2 Ratio of Means

In this section, we consider the test statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, both for $n_x = n_y$ and $n_x \neq n_y$. We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$, where $\text{Exp}(\lambda)$ is an exponential distribution with rate λ, i.e. $\mathbb{E}[X_i] = 1/\lambda_x$. We chose this setup because 1) having data with non-negative support ensures non-zero denominators in the ratio statistic, and 2) the resulting ratio statistic follows a beta prime distribution, also called a Pearson type VI distribution (Johnson et al., 1995, p. 248), which provides an approximate baseline for comparison (see Section 2.2).

For equal sample sizes, we set $n = n_x = n_y = 100$, 500, or 1,000. For unequal sample sizes, we set $n_y = 500$, and $n_x = 50$, 200, or 350. In both cases we set $\lambda_x = 1$ and $\lambda_y = 1.75$



(a) p-values

(b) Number of resamples in Alg 1

Figure 1.5: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with equal sample sizes of $n = n_x = n_y = 100$, 500, 1,000. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 1.5b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 246 tests (points missing).

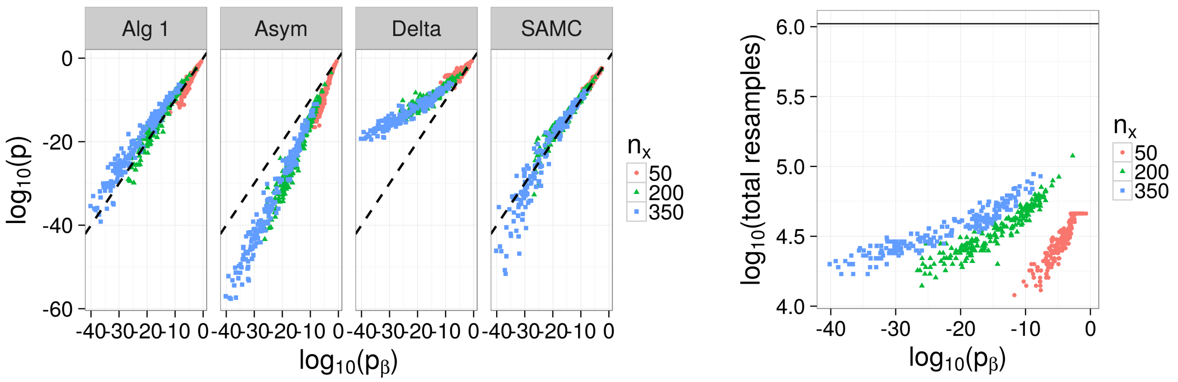
or 2.25. For all parameter combinations, we generated 100 datasets.

For each dataset, we applied our methods and computed the p-value from the beta prime distribution. For our resampling algorithm, we used $B_{\text{pred}} = 10^3$ resamples in each partition. We also computed p-values using the delta method (see Section 2.6) and ran the SAMC algorithm with the same specifications as described in Section 1.5.1.

Results are shown in Figures 1.5 and 1.6. In the Figures, p_β denotes the p-value from the beta prime distribution, and p denotes the p-value from either our methods, the delta method (see Section 2.6), or SAMC. The dashed line has a slope of 1 and intercept of 0, and indicates agreement between methods. As before, the SAMC algorithm did not produce values for smaller p-values, so these points are missing from Figures 1.5 and 1.6 (246 missing points in Figure 1.5, and 33 missing points in Figure 1.6).

As Figures 1.5 and 1.6 show, both our resampling algorithm and asymptotic approximation appear to have more bias in this setting than for the difference in means, though in this case, the asymptotic approximation is biased downward instead of upward. Our resampling algorithm tends to be biased upward.

As before, the SAMC algorithm had trouble estimating extremely small p-values with the number of iterations we allowed it. In the case of equal sample sizes, the SAMC algorithm began to have problems for p-values around 10^{-30}. In the case of unequal sample sizes, the



(a) p-values

(b) Number of resamples in Alg 1

Figure 1.6: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with unequal sample sizes, where $n_y = 500$ and $n_x = 50, 200, 350$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 1.6b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 33 tests (points missing).

SAMC algorithm appears to have performed similarly to our resampling algorithm, albeit with one to two orders of magnitude more resamples. Figures 1.5 and 1.6 also show that p-values from the delta method (see Section 2.6) are not reliable, even for large sample sizes.

Similar to Section 1.5.1, Figures 1.5b and 1.6b show that our resampling algorithm uses fewer resamples for smaller p-values. Also, as before, the scale of the p-values is not the same in Figures 1.5 and 1.6, but in both cases, they are smaller than what would typically be estimated with resampling methods.

1.6 Application to Cancer Genomic Data

To further demonstrate our methods, we analyzed RNA-seq data collected as part of The Cancer Genome Atlas (TCGA) (National Cancer Institute, 2015). In particular, we were interested in identifying genes that were differentially expressed in two different types of lung cancers: lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC).

We downloaded normalized gene expression data from the TCGA data portal. As described by TCGA, to produce the normalized gene expression data, tissue samples from patients with LUSC and LUAD were sequenced using the Illumina RNA Sequencing platform. The raw sequencing reads from all patient samples were processed and analyzed using the SeqWare Pipeline 0.7.0 and MapspliceRSEM workflow 0.7 developed by the University of North Carolina. Sequencing reads were aligned to the human reference genome using MapSplice (Wang et al., 2010), and gene level expression values were estimated using RSEM (Li and Dewey, 2011) with gene annotation file GAF 2.1. For each sample, RSEM gene expression estimates were normalized to set the upper quartile count at 1,000 for gene level estimates. For the analyses in this section, we used the normalized RSEM gene expression estimates.

For both LUAD and LUSC, TCGA contains normalized expression estimates for 20,531 genes (the same genes for both cancers). There were 548 subjects with LUAD observations, and 541 with LUSC observations. To ensure that our results would be biologically meaningful, we restricted our analysis to genes for which at least 50% of the subjects had expression levels above the 25^{th} percentile of all normalized gene expression levels (6.57). This reduced our analysis to 15,386 genes.

Let $P_{x,g}$ and $P_{y,g}$ be the underlying distributions that generated the normalized expression levels in LUAD and LUSC, respectively, for gene g. To test the two-sided hypothesis of $H_0 : P_{x,g} = P_{y,g}$ versus the alternative $H_1 : \mu_x/\mu_y \neq 1$, we used the fold-change statistic $T = \max(\bar{x}_g/\bar{y}_g, \bar{y}_g/\bar{x}_g)$. Here, μ_x and μ_y are the means of $P_{x,g}$ and $P_{y,g}$, respectively.

First, we conducted simple Monte Carlo permutation tests on all 15,386 genes with $B = 10^3$ resamples. This left us with 10,302 genes with p-values less than 10^{-3}, the minimum estimate possible with only $B = 10^3$ resamples. We then used our resampling algorithm to estimate p-values for the 10,302 genes that passed our preliminary screen.

Table 1.1 shows the results for the fifteen genes with the smallest p-values, as well as the deviance and AIC from the Poisson regression fit during the resampling algorithm. We report both the estimate from the initial, single run of our algorithm, as well as the $10^{th}, 50^{th}$, and 90^{th} quantiles from an additional 1,000 runs. Note that Table 1.1 reports the observed ratio of mean(LUAD)/mean(LUSC), not the max of the ratios that we used in the permutation test. Of the top 15 genes, none had elevated levels of LUAD. Point estimates for all genes are available as supplementary material.

Eleven of the these fifteen genes, shown in bold (*DSG3, KRT5, DSC3, CALM3, TP63, ATP1B3, KRT6B, TRIM29, PVRL1, FAT2*, and *KRT6C*), were also identified by Zhan et al. (2015) as being among the most effective genes for distinguishing between LUAD and LUSC. Like us, Zhan et al. (2015) used the TCGA dataset, though they based their analysis on the area under the curve from a Wilcoxon rank-sum test.

We emphasize that in presenting Table 1.1, we are not trying to promote the use of p-values as the sole source of information for making scientific decisions, such as ranking the

Table 1.1: Fifteen genes with the smallest p-values, and other output from our algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition. *Single run* is the value of $\log_{10}(\tilde{p}_{\text{pred}})$ from the initial run of our resampling algorithm. The quantiles are from 1,000 replicates. For the single run, m_{stop} is the partition at which our algorithm stopped, and deviance and AIC are from the Poisson regression fit during the algorithm. Genes shown in bold were identified by Zhan et al. (2015) as being among the most effective genes for distinguishing between LUAD and LUSC using the area under the curve from a Wilcoxon rank-sum test.

	$\log_{10}(\tilde{p}_{\text{pred}})$					
Gene name	Single run	Quantiles $(10^{th}, 50^{th}, 90^{th})$	$\frac{\text{mean(LUAD)}}{\text{mean(LUSC)}}$	m_{stop}	Deviance	AIC
DSG3	-212	(-217, -208, -200)	0.0100	5	40.1	68.1
KRT5	-210	(-223, -214, -205)	0.0107	4	12.5	38.2
DSC3	-197	(-212, -205, -197)	0.0175	6	41.5	72.1
CALML3	-195	(-198, -188, -179)	0.0138	6	57.8	90
TP63	-193	(-199, -192, -186)	0.0308	6	24.2	55.1
ATP1B3	-193	(-196, -188, -181)	0.225	5	28.6	57.7
S1PR5	-190	(-190, -181, -173)	0.0775	6	98.4	131
KRT6B	-185	(-189, -181, -173)	0.0173	5	45.4	76.1
TRIM29	-183	(-188, -181, -174)	0.0788	6	39.3	72
JAG1	-180	(-186, -179, -172)	0.170	5	60.7	92.2
PVRL1	-180	(-183, -177, -171)	0.110	6	8.33	39.2
CLCA2	-178	(-188, -180, -172)	0.0138	7	51.6	86.8
BNC1	-178	(-197, -188, -181)	0.0244	7	76.8	112
FAT2	-177	(-186, -179, -173)	0.0339	7	53.5	89
KRT6C	-177	(-188, -181, -174)	0.0183	6	84.8	119

importance of genes. Instead, we present Table 1.1 and make comparisons with the findings of Zhan et al. (2015) as a way of verifying the reasonableness of our results. Zhan et al. (2015) used different methods to analyze the TCGA data, so we do not expect our results to be exactly the same, but it is encouraging that our results appear to agree to some extent.

We also want to point out that our resampling algorithm can approximate extremely small p-values, but that in doing so, there is a large amount variability in the estimates. However, we think these estimates could still be used as an approximation of the order of magnitude, and note that they would be infeasible to estimate with existing Monte Carlo methods, including the SAMC algorithm.

1.7 Run Time and Sufficient Sample Size

In this section, we provide further details on the run-time of our resampling algorithm and guidance regarding the sample sizes necessary for our test to be reliable.

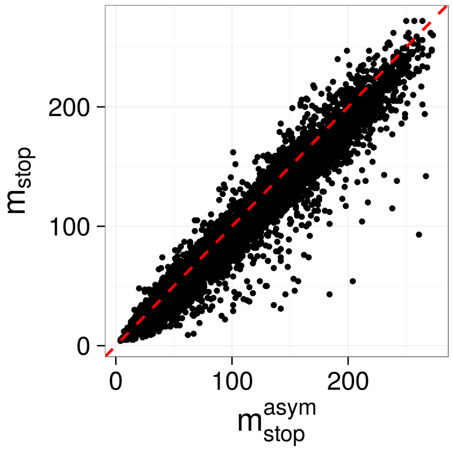


Figure 1.7: Comparison between $m^{\text{asym}}_{\text{stop}}$ and m_{stop} in the analysis of cancer genomic data. m_{stop} is the actual stopping partition, which our resampling algorithm determines dynamically. $m^{\text{asym}}_{\text{stop}}$ is our estimate of the stopping partition based on asymptotic approximations, and can be computed before running the algorithm. The dashed diagonal line has a slope of 1 and an intercept of 0, and indicates agreement.

Our resampling algorithm runs in $O(B_{\text{pred}} m_{\text{stop}})$ time. In our current implementation, we set B_{pred} a priori. Regarding m_{stop}, we obtain the following approximation for small p-values, in which we assume that $1 - \Phi(\xi(m)) \gg 1 - \Phi(\xi^{\text{conj}}(m))$. From Algorithm 1,

$$\begin{aligned} m^{\text{asym}}_{\text{stop}} &= \min_m \{m \in \{1, \ldots, m_{\max}\} : \boldsymbol{c}[m] < 1\} \\ &\approx \min_m \{m \in \{1, \ldots, m_{\max}\} : \Pr(T(m) \geq t | \boldsymbol{x}, \boldsymbol{y}) < 1/B_{\text{pred}}\} \quad \text{(for large } B_{\text{pred}}) \\ &\approx \min_m \{m \in \{1, \ldots, m_{\max}\} : 1 - \Phi(\xi(m)) < 1/B_{\text{pred}}\} \qquad (1.5) \\ &\approx \min_m \left\{m \in \{1, \ldots, m_{\max}\} : \Phi^{-1}(1 - 1/B_{\text{pred}}) < \xi(m)\right\} \quad \text{(for large } n_x, n_y) \qquad (1.6) \\ &\equiv m^{\text{asym}}_{\text{stop}}, \end{aligned}$$

where (1.5) follows from (1.12) and the assumption that $1 - \Phi(\xi(m)) \gg 1 - \Phi(\xi^{\text{conj}}(m))$.

In the R package `fastPerm`, we provide functions for computing $m^{\text{asym}}_{\text{stop}}$, which can help an analyst to approximate run-time before running the algorithm. We emphasize that $m^{\text{asym}}_{\text{stop}}$ is based on asymptotic approximations, and may not be the same as the actual stopping partition; $m^{\text{asym}}_{\text{stop}}$ is not used in Algorithm 1. As shown in Figure 1.7, the expected stopping distribution $m^{\text{asym}}_{\text{stop}}$ appears to be a reasonable estimate of the actual stopping partition m_{stop} in our analysis of cancer genomic data.

We can also use $m^{\text{asym}}_{\text{stop}}$ to provide guidance on sample size. Note that $m^{\text{asym}}_{\text{stop}}$ is the expected number of data points available to the Poisson regression in our resampling algorithm

Table 1.2: $\hat{n}$ for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, equal samples sizes $n_x = n_y = \hat{n}$, $B_{\text{pred}} = 1,000$, and $c = 4$.

$\mu_y = \sigma_y^2$	$\mu_x = \sigma_x^2$	$\hat{n}$	$\hat{p}_{\text{asym}}$
2	3	5	2.4×10^{-1}
	4	6	2.4×10^{-2}
	5	13	2.4×10^{-5}
	5.25	16	1.3×10^{-6}
	5.5	19	6.0×10^{-8}
	5.75	24	4.2×10^{-10}
	6	31	4.1×10^{-13}
	6.25	40	4.3×10^{-17}
	6.5	55	1.1×10^{-23}
	6.6	63	3.3×10^{-27}
	6.7	74	4.5×10^{-32}
	6.8	87	7.7×10^{-38}
	6.9	105	7.8×10^{-46}
	7	130	6.0×10^{-57}

for estimating the overall p-value. Large values of $m_{\text{stop}}^{\text{asym}}$ imply more reliable but slower estimates, and smaller values of $m_{\text{stop}}^{\text{asym}}$ imply less reliable but faster estimates. To ensure that the results of the sampling algorithm are reliable, we recommend that $m_{\text{stop}}^{\text{asym}} \geq c$ for some constant c. For example, we use $c = 4$. Then for equal sample sizes $n = n_x = n_y$, we set

$$\hat{n} = \min_n \{n \in \mathbb{N} : m_{\text{stop}}^{\text{asym}} \geq c\}.$$

While not explicit in the above notation, we note that $m_{\text{stop}}^{\text{asym}}$, and thus $\hat{n}$, is a function of $\sigma_x^2, \sigma_y^2, \mu_x, \mu_y$, and B_{pred}. Tables 1.2 and 1.3 show $\hat{n}$ and $\hat{p}_{\text{asym}} = \hat{p}_{\text{asym}}(\hat{n}, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y)$, the the p-value from our asymptotic approximation for the given set of parameter values and sample sizes. In Tables 1.2 and 1.3, we set $B_{\text{pred}} = 1,000$. As in Figure 1 in Section 1.3 and Figure 1.8 in Section 1.8, to obtain $\hat{p}_{\text{asym}}$, we substituted parameter values for sample quantities, e.g. μ_x for $\bar{x}$ and σ_x^2 for $(n_x - 1)^{-1} \sum_{i=1}^{n_x} (x_i - \bar{x})^2$. As can be seen in Tables 1.2 and 1.3, $\hat{n}$ and $\hat{p}_{\text{asym}}$ have an inverse relationship.

In general, we recommend that researchers check the output from `fastPerm` to ensure that $m_{\text{stop}} \geq 4$, and we note that the sample sizes required to achieve $m_{\text{stop}} \geq 4$ increase as the p-value decreases. Based on Tables 1.2 and 1.3, at least 15-20 observations in each group appears sufficient for p-values near 1×10^{-6}, and at least 70-90 observations in each group appears sufficient for p-values near 1×10^{-30}.

Table 1.3: $\hat{n}$ for $T = |\bar{x} - \bar{y}|$, $\sigma_x^2 = \sigma_y^2 = 1$, equal samples sizes $n_x = n_y = \hat{n}$, $B_{\text{pred}} = 1,000$, and $c = 4$.

μ_y	μ_x	$\hat{n}$	$\hat{p}_{\text{asym}}$
0	1.5	5	5.4×10^{-2}
	2	9	7.7×10^{-4}
	2.2	13	2.1×10^{-5}
	2.25	15	3.7×10^{-6}
	2.3	18	3.1×10^{-7}
	2.4	32	4.0×10^{-12}
	2.45	53	2.3×10^{-19}
	2.475	80	1.3×10^{-28}
	2.48	89	1.1×10^{-31}
	2.49	115	1.5×10^{-40}
	2.5	165	1.4×10^{-57}

1.8 Proofs

In this section, we find the limiting distribution of $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ and $T = |\bar{x} - \bar{y}|$ within each partition, and note the corresponding trend in p-values across the partitions. In the process, we prove the results discussed in Section 1.3. We structure this section around the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ to help to motivate our discussion, and then extend our results to the statistic $T = |\bar{x} - \bar{y}|$.

As before, we denote the total sample size as N, and we require that $N \geq 2$ to allow for at least one observation in each sample. Let $\{m^N\}_{N=2}^{\infty}$, $\{n_x^N\}_{N=2}^{\infty}$, and $\{n_y^N\}_{N=2}^{\infty}$ be sequences such that $m^N/N \to \tau$ and $n_x^N/N \to \lambda$ as $N \to \infty$, and for all N, $n_y^N = N - n_x^N$. We require that for all N, $0 < m^N \leq n_x^N \leq n_y^N < N$, and similarly, $0 < \tau \leq \lambda \leq 1 - \lambda < 1$. We denote the observed data as $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$, which are $n_x^N \times 1$ and $n_y^N \times 1$ vectors, respectively.

Let $\boldsymbol{\delta}_x^{m^N} = (\delta_{x,1}^{m^N}, \ldots, \delta_{x,n_x^N}^{m^N})'$ and $\boldsymbol{\delta}_y^{m^N} = (\delta_{y,1}^{m^N}, \ldots, \delta_{y,n_y^N}^{m^N})'$ be $n_x^N \times 1$ and $n_y^N \times 1$ indicator vectors, respectively, with 1's corresponding to indices of $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$ that are exchanged for a particular permutation π and zero elsewhere. To be specific, for a permutation $\pi \in \Pi(m^N)$, we define $\delta_{x,i}^{m^N}$ and $\delta_{y,j}^{m^N}$ as

$$\delta_{x,i}^{m^N} = \begin{cases} 1 \text{ if } \pi(i) > n_x^N \\ 0 \text{ if } \pi(i) \leq n_x^N \end{cases} \qquad i = 1, \ldots, n_x^N$$

$$\delta_{y,j}^{m^N} = \begin{cases} 1 \text{ if } \pi(n_x^N + j) \leq n_x^N \\ 0 \text{ if } \pi(n_x^N + j) > n_x^N \end{cases} \qquad j = 1, \ldots, n_y^N.$$

For completeness, we note that for fixed m and $i \neq j$, and dropping dependence on N,

$$\begin{aligned}
\mathbb{E}[\delta_{x,i}^m] &= m/n_x & \mathbb{E}[\delta_{y,i}^m] &= m/n_y \\
\mathrm{Var}(\delta_{x,i}^m) &= \frac{m}{n_x}\left(1 - \frac{m}{n_x}\right) & \mathrm{Var}(\delta_{y,i}^m) &= \frac{m}{n_y}\left(1 - \frac{m}{n_y}\right) \\
\mathrm{Cov}(\delta_{x,i}^m, \delta_{x,j}^m) &= \frac{-m(n_x - m)}{n_x^2(n_x - 1)} & \mathrm{Cov}(\delta_{y,i}^m, \delta_{y,j}^m) &= \frac{-m(n_y - m)}{n_y^2(n_y - 1)}
\end{aligned}$$

We denote the ratio of means as $R = \bar{x}/\bar{y}$. With the permutation test, for each permutation π in partition m^N, we calculate the statistic (ignoring for now the max function used earlier)

$$R(m^N) = \frac{\frac{1}{n_x^N}[(\mathbf{1} - \boldsymbol{\delta}_x^{m^N})'\boldsymbol{x}^N + \boldsymbol{\delta}_y^{m^N\prime}\boldsymbol{y}^N]}{\frac{1}{n_y^N}[\boldsymbol{\delta}_x^{m^N\prime}\boldsymbol{x}^N + (\mathbf{1} - \boldsymbol{\delta}_y^{m^N})'\boldsymbol{y}^N]}.$$

As for all permutation tests, $R(m^N)$ is conditional on the data. The random quantities are $(\boldsymbol{\delta}_x^{m^N}, \boldsymbol{\delta}_y^{m^N})$, which indexed by N, form a triangular array of identically distributed, dependent random variables. We can rewrite $R(m^N)$ as

$$\begin{aligned}
R(m^N) &= \frac{n_y^N}{n_x^N}\left(\frac{n_x^N\bar{x} + \left(\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N - \sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N\right)}{n_y^N\bar{y} - \left(\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N - \sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N\right)}\right) \\
&= g\left(\underbrace{\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N - \sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N}_{W(m^N)}\right). \qquad (1.7)
\end{aligned}$$

Writing $R(m^N)$ as a function of $W(m^N)$ will make it straightforward to generalize our results. We note that conditional on the observed data $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$, all terms in $R(m^N)$ are constant except for $W(m^N)$.

We can further split $W(m^N)$ into

$$W(m^N) = \underbrace{\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N}_{W_y(m^N)} - \underbrace{\sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N}_{W_x(m^N)} \qquad (1.8)$$

Following Theorem 2.8.2 in Lehmann (1999, p. 116), restated in Theorem 1.1 below, under certain conditions both $W_y(m^N)$ and $W_x(m^N)$ in (1.8) converge to normal random variables, in which case $W(m^N)$ also converges to a normal random variable.

We make a few observations before stating Theorem 1.1. The following statements focus on $W_y(m^N)$, but equivalent statements apply to $W_x(m^N)$. First, we note that conditional on $\boldsymbol{y}^N$, $W_y(m^N)$ is the sum of a random sample without replacement of m^N elements from a finite population $\boldsymbol{y}^N = (y_1^N, \ldots, y_{n_y^N}^N)'$. We consider a sequence of populations of increasing size, $\boldsymbol{y}^N, N = 2, 3, \ldots$, and random samples $\boldsymbol{v}^N = (v_1^N, \ldots, v_{m^N}^N)'$ from each $\boldsymbol{y}^N$. To be specific, for fixed $\boldsymbol{\delta}_y^{m^N}$, let $\mathcal{K} = \{j : \delta_{y,j}^{m^N} = 1\}$ be the set of indices corresponding to the selected elements of $\boldsymbol{y}^N$. Then writing $\mathcal{K} = \{k_1, \ldots, k_{m^N}\}$, we have $\boldsymbol{v}^N = (y_{k_1}^N, \ldots, y_{k_{m^N}}^N)'$.

Let $\bar{v}_{m^N} = (1/m^N) \sum_{k=1}^{m^N} v_k^N$, and $\bar{y}_{n_y^N} = (1/n_y^N) \sum_{j=1}^{n_y^N} y_j^N$. Then as shown by Lehmann (1999, p. 116-117),

$$\mathbb{E}[\bar{v}_{m^N} | \boldsymbol{y}^N] = \bar{y}_{n_y^N}$$

$$\text{Var}(\bar{v}_{m^N} | \boldsymbol{y}^N) = \frac{n_y^N - m^N}{m^N(n_y^N - 1)} \frac{1}{n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2.$$

We can now state Theorem 1.1.

Theorem 1.1 (Theorem 2.8.2, Lehmann (1999))**.**

$$\frac{\bar{v}_{m^N} - \mathbb{E}[\bar{v}_{m^N} | \boldsymbol{y}^N]}{\sqrt{\text{Var}(\bar{v}_{m^N} | \boldsymbol{y}^N)}} \to N(0,1)$$

provided that $m^N \to \infty$ and $n_y^N - m^N \to \infty$ as $N \to \infty$, and either of the following two conditions is satisfied:
i) m^N / n_y^N is bounded away from 0 and 1 as $N \to \infty$, and

$$\frac{\max(y_j^N - \bar{y}_{n_y^N})^2}{\sum_j (y_j^N - \bar{y}_{n_y^N})^2} \to 0$$

or
ii)

$$\frac{\max(y_j^N - \bar{y}_{n_y^N})^2}{\sum_j (y_j^N - \bar{y}_{n_y^N})^2 / n_y^N}$$

remains bounded as $N \to \infty$.

For a proof, please see Lehmann (1999) and references therein, particularly the corollary to Lemma 4.1 in Hájek (1961), and Example 4.1 and Section 5 in Hájek (1961). Our constraints on m^N, n_x^N, and n_y^N imply that $m^N \to \infty$ and $n_y^N - m^N \to \infty$ as $N \to \infty$. The other conditions in Theorem 1.1 require that the contribution of each deviance to the sum of deviances becomes negligible as the sample size becomes large. This excludes data coming from distributions with a non-finite variance, such as the Cauchy distribution.

Applying Theorem 1.1 to $W(m^N)$ we get Corollary 1.1.

Corollary 1.1. *Conditional on $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$, and assuming the conditions in Theorem 1.1 hold,*

$$\frac{W(m^N) - \mu(m^N)}{\sqrt{V(m^N)}} \to N(0,1),$$

where $\mu(m^N) = \mu_y(m^N) - \mu_x(m^N)$ and $V(m^N) = V_y(m^N) + V_x(m^N)$, with

$$\begin{aligned}\mu_y(m^N) &= \mathbb{E}[W_y(m^N)|\boldsymbol{y}^N] = m^N \bar{y}_{n_y^N} \\ \mu_x(m^N) &= \mathbb{E}[W_x(m^N)|\boldsymbol{x}^N] = m^N \bar{x}_{n_x^N}\end{aligned}$$

and

$$\begin{aligned}V_y(m^N) &= \operatorname{Var}(W_y(m^N)|\boldsymbol{y}^N) = m^N \frac{n_y^N - m^N}{\left(n_y^N - 1\right) n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2 \\ V_x(m^N) &= \operatorname{Var}(W_x(m^N)|\boldsymbol{x}^N) = m^N \frac{n_x^N - m^N}{\left(n_x^N - 1\right) n_x^N} \sum_{i=1}^{n_x^N} (x_i^N - \bar{x}_{n_x^N})^2.\end{aligned}$$

Before proving Corollary 1.1, we state Lemma 1.1.

Lemma 1.1. *For all m and N,* $\operatorname{Cov}\left(W_x(m^N), W_y(m^N)|\boldsymbol{x}, \boldsymbol{y}\right) = 0$.

Proof of Lemma 1.1. First note that for all m, N, i, and j, $\delta_{x,i}^{m^N} \perp \delta_{y,j}^{m^N}$. This is a direct consequence of the sampling procedure implied by the permutation, in which we condition on the number of elements to exchange (m), and then randomly select m elements of $\boldsymbol{x}$ and m elements of $\boldsymbol{y}$. Therefore, dropping dependence on N,

$$\begin{aligned}\mathbb{E}\left[W_x(m)W_y(m)|\boldsymbol{x}, \boldsymbol{y}\right] &= \mathbb{E}\left[\left(\sum_i \delta_{x,i}^m x_i\right)\left(\sum_j \delta_{y,j}^m y_j\right)|\boldsymbol{x}, \boldsymbol{y}\right] \\ &= \mathbb{E}\left[\sum_i \sum_j \delta_{x,i}^m x_i \delta_{y,j}^m y_j |\boldsymbol{x}, \boldsymbol{y}\right] \\ &= \sum_i \sum_j x_i y_j \mathbb{E}\left[\delta_{x,i}^m \delta_{y,j}^m\right] \\ &= \sum_i x_i \mathbb{E}\left[\delta_{x,i}^m\right] \sum_j y_j \mathbb{E}\left[\delta_{y,j}^m\right] \qquad (\delta_{x,i}^m \perp \delta_{y,j}^m) \\ &= \mathbb{E}\left[W_x(m)|\boldsymbol{x}\right] \mathbb{E}\left[W_y(m)|\boldsymbol{y}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{Cov}\left(W_x(m^N), W_y(m^N)|\boldsymbol{x}, \boldsymbol{y}\right) &= \mathbb{E}\left[W_x(m^N)W_y(m^N)|\boldsymbol{x}, \boldsymbol{y}\right] - \mathbb{E}\left[W_x(m^N)|\boldsymbol{x}\right]\mathbb{E}\left[W_y(m^N)|\boldsymbol{y}\right] \\ &= 0\end{aligned}$$

which proves the lemma. □

Now we prove Corollary 1.1.

Proof of Corollary 1.1. Working with the first term in (1.8), we have

$$W_y(m^N) = \sum_{j=1}^{n_y^N} \delta_{y,j}^{m^N} y_j^N = m^N \bar{v}_{m^N}$$

Therefore, as shown by Lehmann (1999, p. 116-117),

$$\mu_y(m^N) = \mathbb{E}[W_y(m^N)|\boldsymbol{y}^N] = m^N \bar{y}_{n_y^N}$$

and

$$V_y(m^N) = \text{Var}(W_y(m^N)|\boldsymbol{y}^N) = \left(m^N\right)^2 \frac{n_y^N - m^N}{m^N(n_y^N - 1)} \frac{1}{n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2.$$

$$= m^N \frac{n_y^N - m^N}{(n_y^N - 1)} \frac{1}{n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2.$$

Similarly, working with the second term in (1.8),

$$\mu_x(m^N) = \mathbb{E}[W_x(m^N)|\boldsymbol{x}^N] = m^N \bar{x}_{n_x^N}$$

$$V_x(m^N) = m^N \frac{n_x^N - m^N}{(n_x^N - 1)} \frac{1}{n_x^N} \sum_{i=1}^{n_x^N} (x_i^N - \bar{x}_{n_x^N})^2.$$

Applying Theorem 1.1, we have

$$\frac{W_y(m^N) - \mu_y(m^N)}{\sqrt{V_y(m^N)}} = \frac{\bar{v}_{m^N} - \mathbb{E}[\bar{v}_{m^N}|\boldsymbol{y}^N]}{\sqrt{\text{Var}(\bar{v}_{m^N}|\boldsymbol{y}^N)}} \to N(0,1).$$

Similarly, we have

$$\frac{W_x(m^N) - \mu_x(m^N)}{\sqrt{V_x(m^N)}} \to N(0,1).$$

By Lemma 1.1, we have

$$\text{Var}\left(W_y(m^N) - W_x(m^N) \middle| \, \boldsymbol{x}, \boldsymbol{y}\right) = V_y(m^N) + V_x(m^N).$$

Since uncorrelated normal random variables are independent, for N sufficiently large we also have $W_y(m^N) \perp W_x(m^N)$. Then since the sum of independent normal random variables is also normal, for N sufficiently large we have

$$W(m^N) = W_y(m^N) - W_x(m^N) \sim N\left(\mu_y(m^N) - \mu_x(m^N), V_y(m^N) + V_x(m^N)\right).$$

Equivalently, we have

$$\frac{W(m^N) - \mu(m^N)}{\sqrt{V(m^N)}} \to N(0,1)$$

which proves the corollary. □

In the rest of this section, we assume that N is sufficiently large for asymptotic normality to hold for any given partition m, so we drop N from the notation.

In Corollary 1.2 below, we apply the delta method to show that for sufficiently large N, the permutation distribution of the statistic $R(m)$ is normal within each partition.

Corollary 1.2. *Let $R = g(W)$, and suppose that $g'(\mu(m)) > 0$ exists. Also, suppose the conditions in Theorem 1.1 hold. Then conditional on the observed data $\boldsymbol{x}, \boldsymbol{y}$, and for N sufficiently large, $R(m) \sim N(\nu(m), \sigma^2(m))$, where the mean $\nu(m)$ and variance $\sigma^2(m)$ are functions of the partition m.*

Proof of Corollary 1.2. By Corollary 1.1, W is normal for N sufficiently large. Then by the delta method, $g(W)$ also converges to a normal distribution, which proves the corollary. □

The result in Corollary 1.2 for the one-sided statistic $R(m)$ leads directly to the following result for its two-sided counterpart $T(m)$, given in Corollary 1.3 below. However, we first define a new function g^{conj}, the conjugate of g.

Definition 1.2 (Conjugate g^{conj})**.** Let $g(W)$ be a function of W, in which the only other terms are the constants n_x, n_y, $\bar{x}$ and $\bar{y}$. The conjugate g^{conj} is formed by switching the place of n_x with n_y, and $\bar{x}$ with $\bar{y}$, and reversing the sign on each occurrence of W.

For example, for $R = \bar{x}/\bar{y}$, we have

$$g = \frac{n_y}{n_x}\left(\frac{n_x\bar{x} + W}{n_y\bar{y} - W}\right) \qquad g^{\text{conj}} = \frac{n_x}{n_y}\left(\frac{n_y\bar{y} - W}{n_x\bar{x} + W}\right)$$

and for $R = \bar{x} - \bar{y}$, as shown below, we have

$$g = \bar{x} - \bar{y} + \left(\frac{1}{n_x} + \frac{1}{n_y}\right)W \qquad g^{\text{conj}} = \bar{y} - \bar{x} - \left(\frac{1}{n_y} + \frac{1}{n_x}\right)W.$$

We note that $(g^{\text{conj}})^{\text{conj}} = g$.

Corollary 1.3. *Let $T(m) = \max\left(g(W(m)), g^{conj}(W(m))\right)$. Under the conditions of Theorem 1.1, and assuming $g'(\mu(m)) > 0$ and $(g^{conj})'(\mu(m)) > 0$ exist, then for N sufficiently large,*

$$\Pr\left(T(m) \geq t|\boldsymbol{x}, \boldsymbol{y}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{max} - m\right\}\right)\right] - \Phi\left[\xi^{conj}\left(\min\left\{m, 2m_{max} - m\right\}\right)\right], \tag{1.9}$$

where Φ is the standard normal CDF, $m_{max} = \arg\max f(m)$, and

$$\xi(m) = \frac{t - g\left(\mu(m)\right)}{g'\left(\mu(m)\right)\sqrt{V(m)}}, \qquad \xi^{conj}(m) = \frac{t - g^{conj}\left(\mu(m)\right)}{(g^{conj})'\left(\mu(m)\right)\sqrt{V(m)}}.$$

Proof of Corollary 1.3. For $m = 1, \ldots, m_{\max}$,

$$\begin{aligned}
\Pr(T(m) > t|\boldsymbol{x}, \boldsymbol{y}) &= \Pr\left(g(W(m)) > t\right) + \Pr\left(g^{\text{conj}}(W(m)) > t\right) \\
&= \Pr\left(Z > \frac{t - g\left(\mu(m)\right)}{g'\left(\mu(m)\right)\sqrt{V(m)}}\right) + \Pr\left(Z > \frac{t - g^{\text{conj}}\left(\mu(m)\right)}{(g^{\text{conj}})'\left(\mu(m)\right)\sqrt{V(m)}}\right) && (1.10) \\
&\approx 1 - \Phi\left(\xi(m)\right) + 1 - \Phi\left(\xi^{\text{conj}}(m)\right) && (1.11)
\end{aligned}$$

where Z is a standard normal random variable and $\mu(m)$ and $V(m)$ are given in Corollary 1.1. Line (1.10) follows from the delta method, and line (1.11) follows from Corollary 1.2 for N sufficiently large.

Furthermore, since the partition-specific p-values are approximately symmetric about $m_{\max}$ (the p-values are exactly symmetric for equal sample sizes, and the symmetry worsens as the sample sizes become more imbalanced), we can get the asymptotic p-value for any partition $m = 1, \ldots, \min(n_y, n_x)$ as

$$\Pr\left(T(m) \geq t|\boldsymbol{x}, \boldsymbol{y}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right] - \Phi\left[\xi^{\text{conj}}\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right]. \tag{1.12}$$

This proves the corollary. □

We also note that when $n_x = n_y$, the approximation in (1.9) is equally accurate for partitions both smaller and larger than $m_{\max}$. However, for unequal sample size, the approximation is less accurate for partitions larger than $m_{\max}$.

In summary, and to be explicit with all quantities, for the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, we have

$$\Pr\left(T(m) \geq t|\boldsymbol{x}, \boldsymbol{y}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right] - \Phi\left[\xi^{\text{conj}}\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right]$$

where Φ is the standard normal CDF, $m_{\max} = \arg\max_m f(m)$, $f(m) = \binom{N}{n_{\min}}^{-1}\binom{n_x}{m}\binom{n_y}{m}$,

$n_{\min} = \min(n_x, n_y)$, and [1]

$$\xi(m) = \frac{t - g\left(\mu(m)\right)}{g'\left(\mu(m)\right)\sqrt{V(m)}} \qquad \xi^{\text{conj}}(m) = \frac{t - g^{\text{conj}}\left(\mu(m)\right)}{(g^{\text{conj}})'\left(\mu(m)\right)\sqrt{V(m)}}$$

$$g(\mu(m)) = \frac{n_y}{n_x}\left(\frac{n_x\bar{x} + \mu(m)}{n_y\bar{y} - \mu(m)}\right) \qquad g^{\text{conj}}(\mu(m)) = \frac{n_x}{n_y}\left(\frac{n_y\bar{y} - \mu(m)}{n_x\bar{x} + \mu(m)}\right)$$

$$g'\left(\mu(m)\right) = \frac{n_y}{n_x}\left(\frac{n_y\bar{y} + n_x\bar{x}}{\left(n_y\bar{y} - \mu(m)\right)^2}\right) \qquad (g^{\text{conj}})'\left(\mu(m)\right) = -\frac{n_y}{n_x}\left(\frac{n_x\bar{x} + n_y\bar{y}}{\left(n_x\bar{x} + \mu(m)\right)^2}\right)$$

where

$$\mu(m) = m(\bar{y} - \bar{x})$$
$$V(m) = m\left[\frac{n_y - m}{n_y(n_y - 1)}\sum_{j=1}^{n_y}(y_j - \bar{y})^2 + \frac{n_x - m}{n_x(n_x - 1)}\sum_{i=1}^{n_x}(x_i - \bar{x})^2\right].$$

To get the expected trend shown Figure 1 of Section 3, we set $t = \bar{x}/\bar{y}$ (the observed test statistic), and substituted expected values for the sample quantities. For example, if we generated the elements of $\boldsymbol{x}$ as iid realizations of a random variable X, then we substituted $\mathbb{E}[X]$ for $\bar{x}$, and $\text{Var}(X)$ for $(n_x - 1)^{-1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$.

We note that we get similar results for $T = |\bar{x} - \bar{y}|$. In this case we can write $R(m) = \bar{x} - \bar{y}$ as

$$R(m) = \frac{1}{n_x}[(\mathbf{1} - \boldsymbol{\delta}_x)'\boldsymbol{x} + \boldsymbol{\delta}_y'\boldsymbol{y}] - \frac{1}{n_y}[\boldsymbol{\delta}_x{}'\boldsymbol{x} + (\mathbf{1} - \boldsymbol{\delta}_y)'\boldsymbol{y}]$$
$$= \bar{x} - \bar{y} + \left(\frac{1}{n_x} + \frac{1}{n_y}\right)W(m)$$

Therefore, (1.9) still holds, but with $g(\mu(m)) = \bar{x} - \bar{y} + \left(n_x^{-1} + n_y^{-1}\right)\mu(m)$ and $g'(\mu(m)) = \left(n_x^{-1} + n_y^{-1}\right)$, with the corresponding results for g^{conj} and $(g^{\text{conj}})'$. All other formula are the same as those given for the ratio of means. The resulting trend for $T = |\bar{x} - \bar{y}|$ is shown in Figure 1.8 with $n_x = n_y = 100$, $\mu_x = 4, \mu_y = 2$, and $\sigma_x^2 = \sigma_y^2 = 1$.

While this section shows that the nearly log linear trend holds for both $T = |\bar{x} - \bar{y}|$ and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, we speculate that the trend might be similar for other statistics that are smooth functions of the means. The results for $R = \bar{x}/\bar{y}$ and $R = \bar{x} - \bar{y}$ above suggest a general formulation of permutation statistics in terms of W, which might help with this effort. This general formulation is presented in Proposition 1.1, in which R could be any statistic of the sample means, and not necessarily the ratio or difference of means.

[1] Implementation note: In the fastPerm package, we use the same function to compute ξ and ξ^{conj}, reversing the order of the arguments related to x and y.

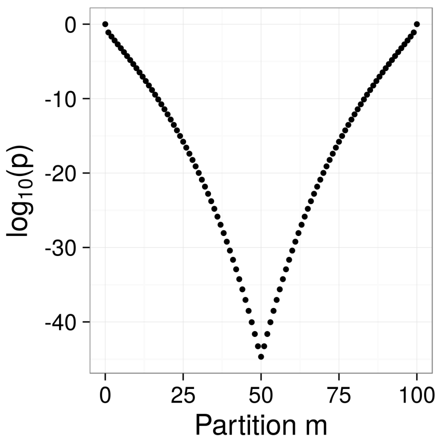


Figure 1.8: Theoretical trend in p-values across the partitions for $T = |\bar{x} - \bar{y}|$ with $n_x = n_y = 100$, $\mu_x = 4, \mu_y = 2$, and $\sigma_x^2 = \sigma_y^2 = 1$.

Proposition 1.1. *Let $R(m) = R(\bar{x}^*(m), \bar{y}^*(m)|\boldsymbol{x}, \boldsymbol{y})$ be any statistic of the permuted sample means conditional on observed data $\boldsymbol{x}, \boldsymbol{y}$, where $\bar{x}^*(m)$ and $\bar{y}^*(m)$ are the means of a permuted dataset $(\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$ corresponding to a permutation $\pi \in \Pi(m)$. Then we can always write $R(m) = g(W(m))$ for some function g that is conditional on the observed data $\boldsymbol{x}, \boldsymbol{y}$.*

Proof of Proposition 1.1. Noting that $\bar{x}^*(m) = \bar{x} + (1/n_x)W(m)$ and $\bar{y}^*(m) = \bar{y} - (1/n_y)W(m)$, we have

$$\begin{aligned} R\left(\bar{x}^*(m), \bar{y}^*(m)|\boldsymbol{x}, \boldsymbol{y}\right) &= R\left(\bar{x} + (1/n_x)W(m), \bar{y} - (1/n_y)W(m)|\boldsymbol{x}, \boldsymbol{y}\right) \\ &= g\left(W(m)\right) \end{aligned}$$

where the last line follows, because $\bar{x}$, $\bar{y}$, n_x, and n_y are constant conditional on $\boldsymbol{x}$ and $\boldsymbol{y}$, and can be absorbed into the functional form of R. This proves the proposition. □

Then for any one-sided statistic $R = g(W)$, in order for asymptotic normality to hold within each partition for the corresponding two-sided statistic T, we must check the conditions in Theorem 1.1 and Corollary 1.3. However, it remains to be shown what additional properties are required to ensure a log concave trend in p-values across the partitions, so we must currently check new statistics on a case-by-base basis.

1.9 Discussion

As we have demonstrated through simulations and an application to cancer genomic data, our methods can quickly approximate small permutation p-values (e.g. $< 10^{-6}$) for two-sample tests, where the test statistic is the difference or ratio of means. The computational efficiency of our resampling algorithm is particularly notable when estimating extremely small p-values (e.g. $< 10^{-30}$).

As is suggested in the example of Section 1.2, our methods can only detect changes in the mean. If $P_x \neq P_y$ but $\mu_x = \mu_y$, then the statistics $T = |\bar{x} - \bar{y}|$ and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ cannot detect differences. We also note that while our development focuses on the null hypothesis $P_x = P_y$, the simulations in Appendix E suggest that our methods extend to less restrictive null hypotheses, such as those considered by Janssen (1997) and Chung and Romano (2013).

As shown in Section 1.5 and Chapter 2, the accuracy of our resampling method is comparable to alternative methods, such as SAMC and MCC, though SAMC and MCC are applicable in situations where our methods are not. In particular, MCC can handle any statistic that can be expressed as, or is permutationally equivalent to, an inner product. In addition to these methods, researchers may want to consider the method of Fieller (1954) for obtaining confidence intervals for the ratio of means, and the approaches described by Cui and Churchill (2003) for using t-tests and ANOVA to analyze the mean log ratio.

While the reliability of our resampling algorithm will vary based on the empirical distribution of the data, in general, we recommend having at least 15-20 observations in each group for p-values near 1×10^{-6} and at least 70-90 observations in each group for p-values near 1×10^{-30} (see Section 1.7). As demonstrated in Section 1.6, there can be considerable variability in estimating extraordinarily small p-values (e.g. 1×10^{-200}). For these extraordinarily small p-values, we recommend that our method be used only to approximate the order of magnitude of the permutation p-value.

In choosing between our resampling algorithm and asymptotic approximation, we recommend using the resampling algorithm when possible for small p-values, as it appears to perform better in simulations. However, as demonstrated in Chapter 2, our asymptotic method may be preferable for large p-values, as it appears to be more conservative under the null. Both approaches work best for equal sample sizes, and we suggest caution when using with small and highly imbalanced samples.

Depending on a researcher's needs, our algorithm could be useful as a fast approximation of small p-values. This might be helpful, for example, in a screening study involving many genes, in which a researcher wants to quickly get a sense for which genes have p-values that are likely to be below a small threshold. It might also be helpful as a preliminary analysis to approximate the order of magnitude of a p-value, which could help a researcher to determine whether it would be feasible to follow-up with other Monte Carlo methods, such as SAMC, and if so, how many iterations they would need to use. For some situations, such as our analysis in Section 1.6, this could save considerable time and resources.

We want to emphasize that our methods are most useful for approximating small permutation p-values. For large p-values, our resampling algorithm is less computationally efficient than simple Monte Carlo resampling. In the context of genomics data, before using our methods, we recommend that researchers use simple Monte Carlo resampling with a small number of resamples (e.g. 10^3) to identify which genes have p-values below a certain threshold (e.g. 10^{-3}). However, this is not a requirement.

This chapter focuses on two-sample tests, and we plan to explore extensions to multiple samples in future work. As one way to handle multiple samples, we could conduct a union-intersection test (Casella and Berger, 2002, p. 380). For example, say we have k samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, and we wish to test the hypothesis $H_0 : \cap_{i \neq j} P_{x_i} = P_{x_j}$ versus the alternative $H_1 : \cup_{i \neq j} \mu_{x_i} \neq \mu_{x_j}$, where μ_i is the mean of P_{x_i}. Then we could use Algorithm 1 to compute p-values for all pairwise differences (or all pairwise ratios), and then take the minimum p-value. As another alternative, we could extend Algorithm 1 to use an omnibus statistic, similar to the ANOVA F-test, and use a multi-sample version of (1.2). For example, we might use $T = \sum_i n_i |\bar{x}_i - \bar{x}|/n$ where $\bar{x}_i$ and n_i are the mean and sample size, respectively, for group i, $\bar{x}$ is the overall mean, and $n = \sum_i n_i$. However, the extension of (1.2) to multiple samples is non-trivial. It is also unclear whether the p-values from the multi-sample case would follow the same trends across the partitions as in the two-sample case.

Returning to the two-sample case, while we have focused on the difference and ratio of the means, preliminary efforts to explain the nearly log-linear trend in p-values across the partitions suggests that the same pattern might hold for other smooth functions of the means. In future work, we plan to explore this further. We also plan to investigate potential diagnostics for assessing the reliability of the algorithm's output, possibly based on the AIC from the Poisson regression. Finally, we note that alternative Monte Carlo methods could be incorporated into our resampling algorithm. For example, the SAMC algorithm could be used in place of simple Monte Carlo resampling within each partition. This might further reduce run-time and increase accuracy.

1.10 R Package and Code

We have implemented our method in the R package `fastPerm` available at `https://github.com/bdsegal/fastPerm`. All code for the simulations and analyses in this chapter are available at `https://github.com/bdsegal/code-for-fastPerm-paper`.

Chapter 2

Fast Approximation of Small p-values in Permutation Tests by Partitioning the Permutations: Further Empirical Investigations

2.1 Introduction

In this chapter, we study the behavior of the methods we introduced in Chapter 1 under additional simulation scenarios and against additional alternative methods. In Section 2.2, we derive parametric p-values for ratios and differences of gamma random variables, which we use in Section 2.3 to evaluate the performance of our methods and alternative techniques. In Section 2.4, we compare against the moment corrected correlation (MCC) method of Zhou and Wright (2015), as well as a saddlepoint approximation to the permutation p-value. In Section 2.5, we evaluate the performance of our method under null hypotheses of a single parameter, as opposed to the entire distribution. In Section 2.6, we derive an asymptotic test for the ratio of the means via the delta method, and demonstrate its application to the cancer genomic data described in Section 1.6.

2.2 Parametric p-values for Ratios and Differences of Gamma Random Variables

The results in this section are used in our simulations of exponential and gamma random variables to obtain parametric approximations to the permutation p-value.

2.2.1 Ratio of Means

Let F be the beta prime CDF, also called a Pearson type VI distribution (Johnson et al., 1995, p. 248), and let f be the corresponding pdf. Following the form given by Becker and Klößner (2016), for $Z \sim F$,

$$f_Z(z;\alpha_1,\alpha_2,s,q) = \frac{\left(\frac{z-q}{s}\right)^{\alpha_1-1}\left(1+\frac{z-q}{s}\right)^{-\alpha_1-\alpha_2}}{sB(\alpha_1,\alpha_2)}$$

where B is the beta function. As we show in this section, if $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$, then $\bar{X}/\bar{Y}$ and $\bar{Y}/\bar{X}$ follow scaled beta prime distributions. This allows us to approximate the permutation p-value for the ratio statistic with the p-value from a beta prime. We note that the beta prime p-value is not conditional on the data, so is not the same as the permutation p-value, but simulation results suggest it is a reasonable approximation.

As in Section 1.5.2, let $x_i, i = 1,\dots,n_x$, and $y_j, j = 1,\dots,n_y$, be realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$. We consider the quantity $T = \max\left(\bar{X}/\bar{Y}, \bar{Y}/\bar{X}\right)$, and denote the observed statistic as $t = \max\left(\bar{x}/\bar{y}, \bar{y}/\bar{x}\right)$. Then under the null hypothesis that $\lambda_x = \lambda_y$, the p-value from the beta prime distribution is

$$\begin{aligned}
p_\beta &= \Pr(T \geq t) \\
&= \Pr\left(\max(\bar{X}/\bar{Y}, \bar{Y}/\bar{X}) \geq t\right) \\
&= \Pr\left(\left\{\bar{X}/\bar{Y} \geq t\right\} \cup \left\{\bar{Y}/\bar{X} \geq t\right\}\right) \\
&= \Pr\left(\bar{X}/\bar{Y} \geq t\right) + \Pr\left(\bar{Y}/\bar{X} \geq t\right) \qquad \text{(disjoint)} \qquad (2.1) \\
&= \Pr\left(\frac{n_y}{n_x}\frac{\sum_i X_i}{\sum_j Y_j} \geq t\right) + \Pr\left(\frac{n_x}{n_y}\frac{\sum_j Y_j}{\sum_i X_i} \geq t\right) \qquad (2.2) \\
&= 1 - F\left(t; \alpha_1 = n_x, \alpha_2 = n_y, s = n_y/n_x, q = 0\right) \qquad (2.3) \\
&+ 1 - F\left(t; \alpha_1 = n_y, \alpha_2 = n_x, s = n_x/n_y, q = 0\right).
\end{aligned}$$

The equality in (2.1) follows because $\bar{X}/\bar{Y} \geq t$ if and only if $\bar{Y}/\bar{X} < t$ (assuming $t \neq 1$, which occurs with probability one). Line 2.3 follows from well known properties, which we outline below.

Let $U_1 \sim \text{Gamma}(\alpha_1, \lambda_1)$ and $U_2 \sim \text{Gamma}(\alpha_2, \lambda_2)$, $U_1 \perp U_2$. Also, let $V_1 = h_1(U_1, U_2) = U_1/U_2$ and $V_2 = h_2(U_1, U_2) = U_2$ with respective inverse transformations $U_1 = h^{-1}(V_1, V_2) = V_1V_2$ and $U_2 = h^{-1}(V_1, V_2) = V_2$. Noting that the Jacobian of the transformation is

$$J = \begin{vmatrix} \partial u_1/\partial v_1 & \partial u_1/\partial v_2 \\ \partial u_2/\partial v_1 & \partial u_2/\partial v_2 \end{vmatrix} = \begin{vmatrix} v_2 & v_1 \\ 0 & 1 \end{vmatrix} = v_2,$$

we have

$$\begin{aligned} f_{V_1,V_2}(v_1, v_2) &= f_{U_1,U_2}\left(h_1^{-1}(v_1, v_2), h_2^{-1}(v_1, v_2)\right)|J| \\ &= \frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)}(v_1v_2)^{\alpha_1-1}e^{-\lambda_1v_1v_2}\frac{\lambda_2^{\alpha_2}}{\Gamma(\alpha_2)}v_2^{\alpha_2-1}e^{-\lambda_2v_2}v_2 \\ &= \frac{\lambda_1^{\alpha_1}\lambda_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v_1^{\alpha_1-1}v_2^{\alpha_1+\alpha_2-1}e^{-(\lambda_1v_1+\lambda_2)v_2}. \end{aligned}$$

Therefore,

$$\begin{aligned} f_{V_1}(v_1) &= \int_0^\infty f_{V_1,V_2}(v_1, v_2)dv_2 \\ &= \frac{\lambda_1^{\alpha_1}\lambda_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v_1^{\alpha_1-1}\int_0^\infty v_2^{\alpha_1+\alpha_2-1}e^{-(\lambda_1v_1+\lambda_2)v_2}dv_2 \\ &= \frac{\lambda_1^{\alpha_1}\lambda_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v_1^{\alpha_1-1}\frac{\Gamma(\alpha_1+\alpha_2)}{(\lambda_1v_1+\lambda_2)^{\alpha_1+\alpha_2}} \\ &= \frac{\left(\frac{v_1}{\lambda_2/\lambda_1}\right)^{\alpha_1-1}\left(1+\frac{v_1}{\lambda_2/\lambda_1}\right)^{-\alpha_1-\alpha_2}}{(\lambda_2/\lambda_1)B(\alpha_1, \alpha_2)}, \end{aligned}$$

which is a generalized beta prime distribution with shape parameters α_1 and α_2, location parameter $q = 0$, and scale parameter $s = \lambda_2/\lambda_1$. In the case where $\lambda_1 = \lambda_2$, this simplifies to the standard beta prime distribution with shape parameters α_1 and α_2. This shows that whenever $U_1 \sim \text{Gamma}(\alpha_1, \lambda)$, $U_2 \sim \text{Gamma}(\alpha_2, \lambda)$, and $U_1 \perp U_2$, we have $U_1/U_2 \sim F(\alpha_1, \alpha_2, 1, 0)$. We note that some sources report that for $U_1 \sim \text{Gamma}(\alpha_1, \lambda_1)$, $U_2 \sim \text{Gamma}(\alpha_2, \lambda_2)$, and $U_1 \perp U_2$, we have $U_1/U_2 \sim F(\alpha_1, \alpha_2, 1, 0)$ if $\lambda_1 = \lambda_2 = 1$ (e.g., Leemis and McQueston, 2008). However, as shown above, this also holds when $\lambda_1 = \lambda_2 \neq 1$.

Now let $Z = \left(\sum_{i=1}^{n_x} X_i\right) / \left(\sum_{j=1}^{n_y} Y_i\right)$. Since $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$, it follows that $\sum_{i=1}^{n_x} X_i \sim \text{Gamma}(n_x, \lambda_x)$ and $\sum_{j=1}^{n_y} Y_j \sim \text{Gamma}(n_y, \lambda_y)$. Then under the null of $\lambda_x = \lambda_y$, the results above give $Z \sim F(n_x, n_y, 1, 0)$ and $1/Z \sim F(n_y, n_x, 1, 0)$.

Now let $W = sZ$. Then by a change of variable, we have

$$f_W(w) = \frac{\left(\frac{w}{s}\right)^{n_x-1}\left(1+\frac{w}{s}\right)^{-n_x-n_y}}{sB(n_x, n_y)}$$

Applying this result to (2.2), we have

$$\frac{n_y}{n_x}\frac{\sum_{i=1}^{n_x} X_i}{\sum_{j=1}^{n_y} Y_j} \sim F(\cdot; n_x, n_y, n_y/n_x, 0)$$

and similarly,

$$\frac{n_x}{n_y}\frac{\sum_{j=1}^{n_y} Y_j}{\sum_{i=1}^{n_x} X_i} \sim F(\cdot; n_y, n_x, n_x/n_y, 0)$$

Then (2.3) follows directly from (2.2).

To compute the CDF values for the scaled beta prime, we used the `PearsonDS` package for R (Becker and Klößner, 2016).

Similarly, for gamma random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_x, \lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_y, \lambda_y)$, $\sum_{i=1}^{n_x} X_i \sim \text{Gamma}(n_x\alpha_x, \lambda_x)$ and $\sum_{j=1}^{n_y} Y_j \sim \text{Gamma}(n_y\alpha_y, \lambda_y)$. Then letting $Z = \left(\sum_{i=1}^{n_x} X_i\right) / \left(\sum_{j=1}^{n_y} Y_j\right)$, under the null of $H_0 : \lambda_x = \lambda_y, \alpha_x = \alpha_y = \alpha$, we have $Z \sim F(\cdot; n_x\alpha, n_y\alpha, 1, 0)$ and $1/Z \sim F(\cdot; n_y\alpha, n_x\alpha, 1, 0)$, so $(n_y/n_x)Z \sim F(\cdot; n_x\alpha, n_y\alpha, n_y/n_x, 0)$ and $(n_x/n_y)Z \sim F(\cdot; n_y\alpha, n_x\alpha, n_x/n_y, 0)$. Therefore,

$$\begin{aligned} p_\beta = \Pr(T \geq t) = 1 - F\left(t; n_x\alpha, n_y\alpha, n_y/n_x, 0\right) \\ + 1 - F\left(t; n_y\alpha, n_x\alpha, n_x/n_y, 0\right). \end{aligned}$$

In our simulations, we generated data under the alternative $H_1 : \lambda_x \neq \lambda_y, \alpha_x = \alpha_y = \alpha$ for various values of α. While we would ideally also simulate under the alternatives $H_1 : \lambda_x \neq \lambda_y, \alpha_x \neq \alpha_y$ and $H_1 : \lambda_x = \lambda_y, \alpha_x \neq \alpha_y$, in these scenarios it is not possible to compute p_β under $H_0 : \alpha_x = \alpha_y, \lambda_x = \lambda_y$, because α does not disappear in the beta prime density. Consequently, we would have to compute p_β under $H_0 : \alpha_x = \alpha_y = c, \lambda_x = \lambda_y$ for a specified constant c. This is more restrictive than the null hypothesis for the permutation test, and consequently, it would not be clear how to compute the parametric p-value to use as an approximation for the true permutation p-value.

2.2.2 Difference in Means

Let $M_X(t)$ be the moment generating function (MGF) for random variable X. Then for $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda), i = 1, \ldots, n$, $M_{\frac{1}{n}\sum_{i=1}^n X_i}(t) = M_{\sum_{i=1}^n X_i}(t/n) = \prod_{i=1}^n M_{X_i}(t/n) = \left(1 - \frac{1}{n\lambda}t\right)^{-n\alpha}$, which is the MGF for a Gamma distribution with shape parameter $n\alpha$ and rate parameter $n\lambda$. Therefore, $\bar{X} \sim \text{Gamma}(n\alpha, n\lambda)$.

Then for $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda), i = 1, \ldots, n_x$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda), j = 1, \ldots, n_y$, the distribution of $\bar{X} - \bar{Y}$, which we denote as G, is (Klar, 2015)

$$G(z) = \Pr(\bar{X} - \bar{Y} \leq z) = C \underbrace{\int_{\max\{0,-z\}}^{\infty} v^{n_y\alpha-1} e^{-n_y\lambda v} \gamma\left(n_x\alpha, n_x\lambda(v+z)\right) dv}_{A(z)}, \tag{2.4}$$

where $\gamma(a,b) = \int_0^b s^{a-1}e^{-s}ds$ is the lower incomplete gamma function, and $C = (n_y\lambda)^{n_y\alpha} / (\Gamma(n_x\alpha)\Gamma(n_y\alpha))$ is the normalizing constant. Klar (2015) also gives the density for $\bar{X} - \bar{Y}$, which was derived by Mathai (1993).

However, in our simulations we found that several scenarios led to numerical problems in computing (2.4) due to large gamma and incomplete gamma function values. These were not solved by computing $G(z) = \exp\{n_y\alpha \log(n_y\lambda) - \log\Gamma(n_x\alpha) - \log\Gamma(n_y\alpha) + \log(A(z))\}$ where $\log\Gamma$ is the log gamma function. As an alternative, we used a saddlepoint approximation for (2.4). As described below, the saddlepoint approximation is accurate and did not pose computational difficulties.

To compute the saddlepoint approximation, note that under $H_0 : \lambda_x = \lambda_y = \lambda, \alpha_x = \alpha_y = \alpha$, the MGF of $\bar{X} - \bar{Y}$ is

$$M_{\bar{X}-\bar{Y}}(t) = \left(1 - \frac{1}{n_x\lambda}t\right)^{-n_x\alpha} \left(1 + \frac{1}{n_y\lambda}t\right)^{-n_y\alpha} \qquad t \in (-n_y\lambda, n_x\lambda),$$

and the cumulant generating function is

$$K(t) = \log\left(M_{\bar{X}-\bar{Y}}(t)\right) = -n_x\alpha \log\left(1 - \frac{t}{n_x\lambda}\right) - n_y\alpha \log\left(1 + \frac{t}{n_y\lambda}\right).$$

After some algebra, we get the derivatives

$$K'(t) = \frac{\alpha(n_x+n_y)t}{(n_x\lambda - t)(n_y\lambda + t)}$$
$$K''(t) = \alpha(n_x+n_y)\frac{t^2 + n_x n_y \lambda^2}{\left[(n_x\lambda - t)(n_y\lambda + t)\right]^2}.$$

Let $\hat{t} = \hat{t}(z) \in (-n_y\lambda, n_x\lambda)$ be the solution to $K'(\hat{t}) = z$. Then as Butler (2007) describes, the saddlepoint approximation of the cumulative distribution for $z \neq \mathbb{E}[\bar{X} - \bar{Y}] = 0$ is (Lugannani and Rice, 1980)

$$\hat{G}(z) = \Phi(\hat{w}) + \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right), \tag{2.5}$$

where $\hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2\left[\hat{t}z - K(\hat{t})\right]}$, $\hat{u} = \hat{t}\sqrt{K''(\hat{t})}$, and Φ and ϕ are the standard normal distribution and density, respectively. The two-sided p-value is then $p_{\text{saddle}} = \Pr(T \geq t) = 1 - \hat{G}(t; n_x, n_y, \lambda, \alpha) + \hat{G}(-t; n_x, n_y, \lambda, \alpha)$.

Figure 2.1 compares the true distribution (2.4) and saddlepoint approximation (2.5) for $n_x = n_y = 100$, $\alpha = 1$, and $\lambda = 4$. Figure 2.1 shows agreement between the true distribution and saddlepoint approximation far into the tail. The trend is similar for other parameter values (not shown), and appears to be reliable up to quantile values of around 10^{-200}. We also note that through simulations, we found that both the true distribution and the saddlepoint approximation agreed with the empirical distribution for a variety of parameter values.

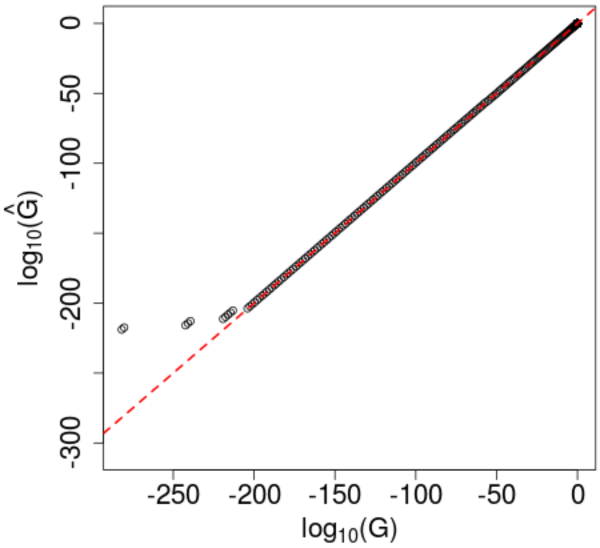


Figure 2.1: Comparison of true (G) and saddlepoint approximation ($\hat{G}$) distributions of the difference of gamma random variables. The diagonal dashed line has slope of 1 and an intercept of 0, and indicates agreement.

Both the true distribution (2.4) and saddlepoint approximation (2.5) are functions of α and λ. Neither parameter disappears under the null of $H_0 : \alpha_x = \alpha_y = \alpha, \lambda_x = \lambda_y = \lambda$, so we must set α and λ to fixed values to compute p-values. To do this in the simulations, we pooled the generated data, computed the maximum likelihood estimates (MLEs), and plugged the MLEs into (2.5). In the simulations, we found that allowing both α and λ to vary led to less reliable p-values from the saddlepoint approximation than allowing just one parameter to vary. To be consistent with our simulations for the ratio of gamma means, we fixed α and used the MLE estimate for λ in the simulations.

We note that this procedure for obtaining a parametric approximation to the permutation p-value involves three approximations: 1) approximating the permutation p-value (conditional on the data) with a parametric distribution (not conditional on the data), 2) approximating the parametric distribution with a saddlepoint approximation, and 3) approximating the general null $H_0 : \lambda_x = \lambda_y$ with the more restrictive null $H_0 : \lambda_x = \lambda_y = \hat{\lambda}$, where $\hat{\lambda}$ is the MLE from the pooled data.

To obtain the MLE estimates, let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ be the pooled data, $N = n_x + n_y$ be the total sample size, and $\bar{z} = N^{-1} \sum_{i=1}^{N} z_i$, $s^2 = (N-1)^{-1} \sum_i (z_i - \bar{z})^2$ be the sample mean and variance, respectively. Then assuming iid observations, the joint log likelihood is

$$\ell = N\alpha \log(\lambda) - N \log\left(\Gamma(\alpha)\right) + (\alpha - 1) \sum_i \log(z_i) - N\lambda\bar{z}.$$

Taking the derivative with respect to λ and setting to zero, we get $\lambda = \alpha/\bar{z}$. Then taking $\partial\ell/\partial\alpha$ and substituting in $\lambda = \alpha/\bar{z}$, we get

$$\begin{aligned} \ell'(\alpha) &= N \log\left(\frac{\alpha}{\bar{z}}\right) - N\Psi(\alpha) + \sum_i \log(x_i) \\ \ell''(\alpha) &= \frac{N}{\alpha} - N\Psi'(\alpha), \end{aligned}$$

where $\Psi(\alpha) = d\log(\Gamma(\alpha))/d\alpha$ is the digamma function, and $\Psi'(\alpha) = d\Psi(\alpha)/d\alpha$ is the trigamma function. We used Newton-Raphson until convergence of $\ell(\alpha)$ to get the MLE $\hat{\alpha}$, where each update is given by $\alpha^{k+1} = \alpha^k - \ell'\left(\alpha^k\right)/\ell''\left(\alpha^k\right)$, and then set $\hat{\lambda} = \hat{\alpha}\bar{z}$. To get initial values for α, we used the method of moments and set $\alpha^0 = \bar{z}^2/s^2$.

2.3 Additional Simulations

In this section, we present simulation results under additional scenarios.

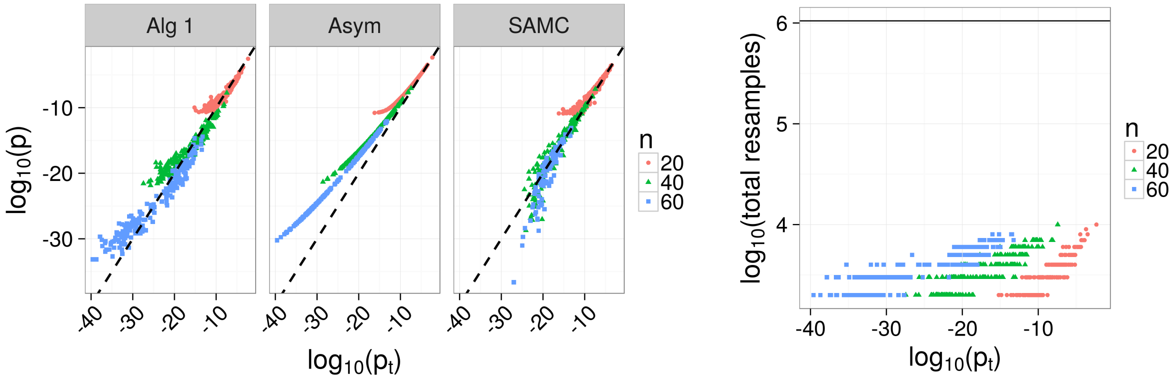
2.3.1 Difference in Means with Normal Data

In this subsection, we use the statistic $T = |\bar{x} - \bar{y}|$ with data generated as normal random variables.

Small Sample Sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(\mu_x, 1)$ and $Y_j \overset{\text{iid}}{\sim} N(\mu_y, 1)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes and for each each n or n_x, we set $\mu_x = 2$ or 3, and $\mu_y = 0$, and simulated 100 datasets for each combination of parameters. We used the p-value from a t-test with equal variance, denoted as p_t, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures 2.2 and 2.3, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The number of resamples used by our algorithm is shown in Figures 2.2b



(a) p-values

(b) Number of resamples in Alg 1

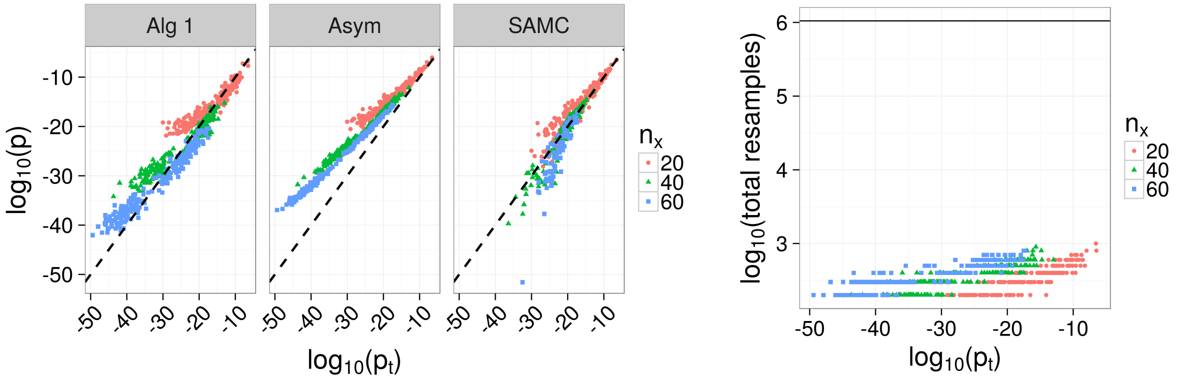
Figure 2.2: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data, $\mu_x = 2$ or 3, $\mu_y = 0$, and equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 2.2b shows the number of iterations used in the SAMC algorithm (set in advance and independent of p-value).

and 2.3b. We note that the bias shown in Figures 2.2a and 2.3a are similar to that obtained with moment-corrected correlation (MCC) (Zhou and Wright, 2015), shown in Figure 2.19 of Section 2.4.

Under the Null Hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(0,1)$ and $Y_j \overset{\text{iid}}{\sim} N(0,1)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes and for each each n or n_x, we simulated 1,000 datasets (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures 2.4 and 2.5, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* shows the p-value from a two-sided t-test with equal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. We compare p-values from the t-test against $\tilde{p}$, which shows close agreement. We do not show results from the SAMC algorithm, because the `EXPERT` package (Yu et al., 2011) does not provide



(a) p-values

(b) Number of resamples in Alg 1

Figure 2.3: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data, $\mu_x = 2$ or 3, $\mu_y = 0$, and unequal sample sizes, where $n_y = 100$ and $n_x = 20, 40, 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 2.3b shows the number of iterations used in the SAMC algorithm (set in advance and independent of p-value).

Table 2.1: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = |\bar{x} - \bar{y}|$ with normal data and equal sample sizes $n = n_x = n_y$. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *t-test* is a two-sided t-test with equal variance, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

signif level	n	MC	t-test	Alg 1	Asymptotic
0.01	20	0.010	0.010	0.015	0.010
	40	0.013	0.013	0.015	0.013
	60	0.010	0.010	0.011	0.010
0.05	20	0.048	0.050	0.064	0.050
	40	0.055	0.055	0.075	0.056
	60	0.049	0.050	0.061	0.050
0.1	20	0.098	0.098	0.14	0.11
	40	0.11	0.11	0.14	0.11
	60	0.10	0.10	0.12	0.10

results for p-values $> 10^{-3}$.

Tables 2.1 and 2.2 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample size simulations, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *t-test* is a two-sided t-test with equal variance, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

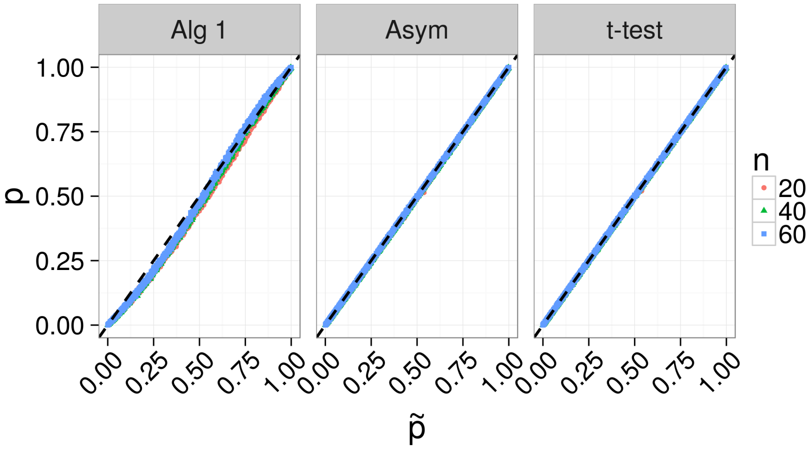


Figure 2.4: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data under the null $P_x = P_y$ with equal sample sizes of $n = n_x = n_y = 20,\ 40,\ 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* shows the p-value from a two-sided t-test with equal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

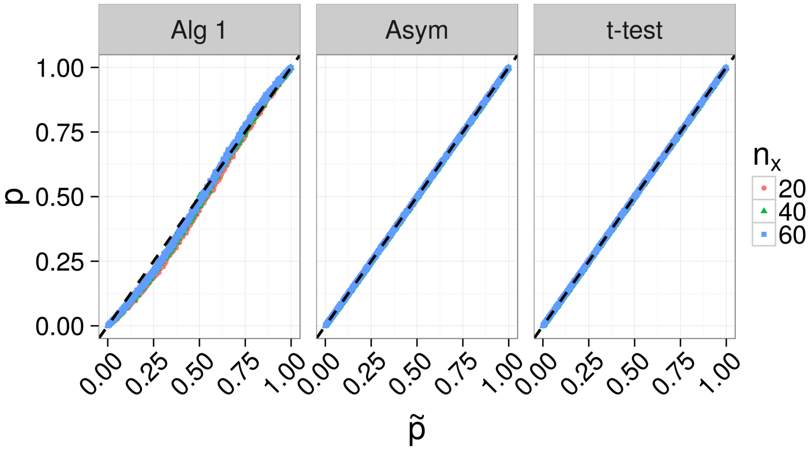


Figure 2.5: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data under the null $P_x = P_y$ with unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* shows the p-value from a two-sided t-test with equal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

Table 2.2: Type I error rates $\Pr(\text{p-value} \leq \text{signif level}|H_0)$ for $T = |\bar{x} - \bar{y}|$ with normal data and unequal sample sizes $n_x \neq n_y$ (n_x shown and $n_y = 100$). *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *t-test* is a two-sided t-test with equal variance, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

signif level	n_x	MC	t-test	Alg 1	Asymptotic
0.01	20	0.013	0.013	0.018	0.013
	40	0.016	0.016	0.018	0.016
	60	0.010	0.010	0.013	0.010
0.05	20	0.049	0.049	0.075	0.049
	40	0.047	0.047	0.066	0.047
	60	0.044	0.044	0.057	0.044
0.1	20	0.090	0.090	0.14	0.092
	40	0.10	0.10	0.14	0.11
	60	0.090	0.090	0.13	0.090

2.3.2 Ratio of Means with Exponential Data

In this subsection, we use the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with data generated as exponential random variables.

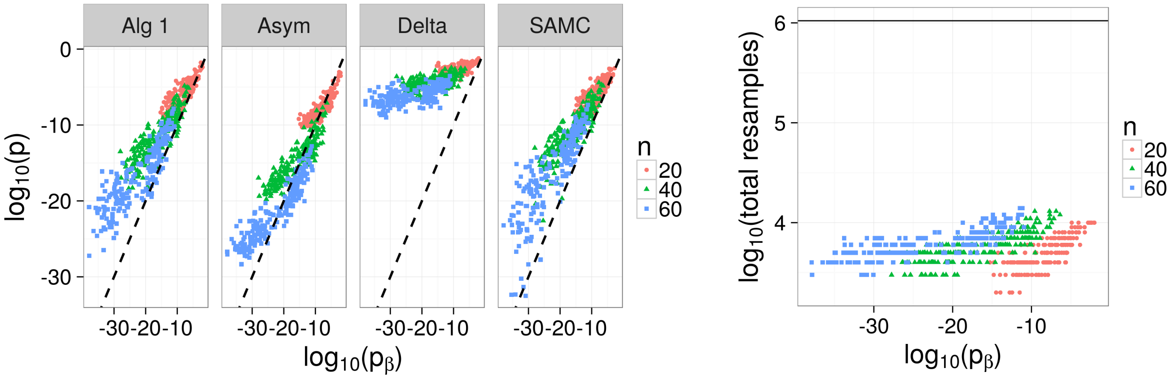
Small Sample Sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For each n or n_x, we set $\lambda_y = 5$ or 10, and $\lambda_x = 1$. For both equal and unequal sample sizes, we simulated 100 datasets for each combination of parameters. We used the p-value from the beta prime distribution, denoted as p_β (see Section 2.2), as an approximation to the true permutation p-value.

Results for equal and unequal sample size are shown in Figures 2.6 and 2.7, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The number of resamples used by our resampling algorithm is shown in Figures 2.6b and 2.7b.

Under the Null Hypothesis $P_x = P_y$

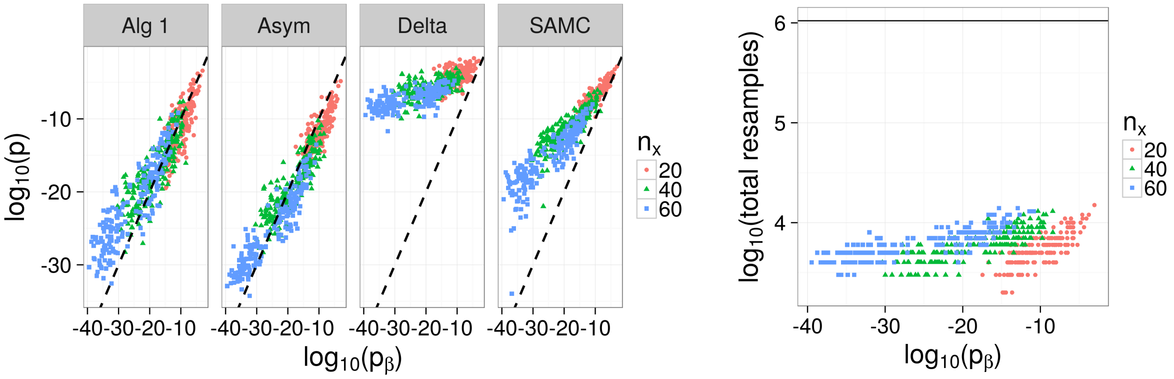
We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(1)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(1)$. For equal sample sizes, we set $n =$



(a) p-values

(b) Number of resamples in Alg 1

Figure 2.6: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data, $n = n_x = n_y = 20$, 40, 60, and rates $\lambda_y = 5, 10$ and $\lambda_x = 1$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 2.6b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 15 tests (points missing).



(a) p-values

(b) Number of resamples in Alg 1

Figure 2.7: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data, $n_x = 20$, 40, 60, $n_y = 100$, and rates $\lambda_y = 5, 10$ and $\lambda_x = 1$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 2.7b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value).

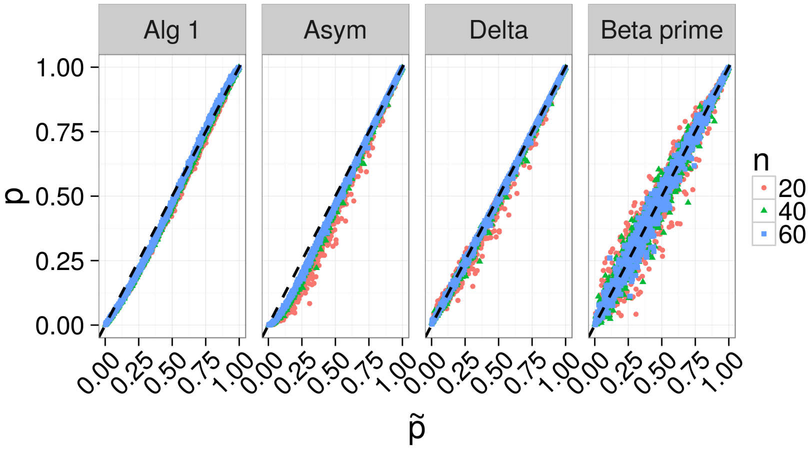


Figure 2.8: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data under the null of $P_x = P_y$ with equal sample sizes of $n = n_x = n_y =$ 20, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

$n_x = n_y = 20, 40, 60$. For unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, we simulated 1,000 datasets for each combination of parameters (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures 2.8 and 2.9, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. Given the large p-values, using 10^5 Monte Carlo resamples should be sufficient to obtain reliable estimates of the true permutation p-value. Therefore, this comparison demonstrates that the permutation p-value is not exactly the same as the p-value from the beta prime distribution. However, it appears reasonably close, so we use it as an approximation to the truth in other simulations in which the p-values are much smaller and simple Monte Carlo methods are not feasible.

We do not show results from the SAMC algorithm, because as noted above, the `EXPERT` package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Tables 2.3 and 2.4 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample size simulations, respectively. *MC* is the unadjusted p-value from

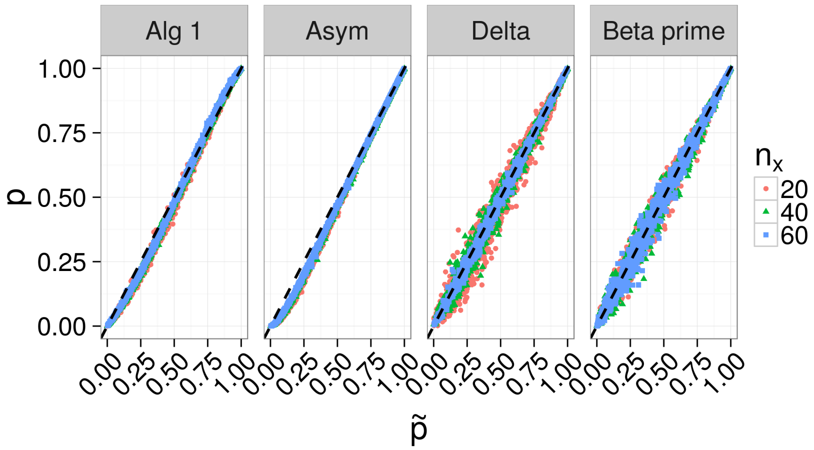


Figure 2.9: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data under the null of $P_x = P_y$ with unequal sample sizes of $n_x = 20$, 40, 60 and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

Table 2.3: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data and equal sample sizes $n = n_x = n_y$. *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asymptotic* is our asymptotic approximation, *Delta* is the delta method, and *Beta prime* is the the beta prime distribution.

signif level	n	MC	Alg 1	Asymptotic	Delta	Beta prime
0.01	20	0.010	0.016	0.066	0.003	0.009
	40	0.010	0.018	0.050	0.002	0.008
	60	0.013	0.013	0.031	0.006	0.015
0.05	20	0.064	0.084	0.14	0.045	0.058
	40	0.061	0.079	0.11	0.054	0.061
	60	0.051	0.063	0.091	0.050	0.047
0.10	20	0.11	0.15	0.21	0.12	0.11
	40	0.11	0.14	0.17	0.11	0.11
	60	0.093	0.11	0.14	0.095	0.092

simple Monte Carlo resampling and 10^5 resamples, *Beta prime* is the p-value from the beta prime distribution, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

Table 2.4: Type I error rates $\Pr(\text{p-value} \leq \text{signif level}|H_0)$ for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data and unequal sample sizes $n_x \neq n_y$ (n_x shown, and $n_y = 100$). *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asymptotic* is our asymptotic approximation, *Delta* is the delta method with, and *Beta prime* is the beta prime distribution.

signif level	n	MC	Alg 1	Asymptotic	Delta	Beta prime
0.01	20	0.011	0.016	0.054	0.008	0.012
	40	0.008	0.012	0.033	0.004	0.006
	60	0.012	0.016	0.035	0.007	0.014
0.05	20	0.061	0.082	0.127	0.065	0.056
	40	0.048	0.062	0.097	0.047	0.050
	60	0.047	0.065	0.083	0.044	0.051
0.10	20	0.12	0.16	0.19	0.14	0.12
	40	0.10	0.14	0.17	0.11	0.10
	60	0.091	0.12	0.14	0.093	0.088

2.3.3 Difference in Means with Gamma Data

In this subsection, we use the statistic $T = |\bar{x} - \bar{y}|$ with data generated as gamma random variables.

Small Sample Sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_y)$, where $\alpha = 0.5, 3, 5$, $\lambda_x = 1$, and λ is the rate parameter. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For $\alpha = 0.5$, we set $\lambda_y = 2.5, 3$ for all n or n_x. For $\alpha = 3$, we set $\lambda_y = 1.5, 1.75$ for all n or n_x. For $\alpha = 5$, we set $\lambda_y = 1.25, 1.5$ for all n or n_x. For both equal and unequal sample sizes, we simulated 100 datasets for each combination of parameters.

Results for equal and unequal sample size are shown in Figures 2.10 and 2.11, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* is a t-test with unequal variance, and *Saddle* is the saddlepoint approximation (see Section 2.2). SAMC results are not shown, as the EXPERT package does not provide p-values larger than 10^{-3}. We use the p-values from simple Monte Carlo resampling, denoted as $\tilde{p}$, with 10^5 resamples as a basis of comparison, and only show values for which $\tilde{p} > 10^{-3}$ to ensure that the $\tilde{p}$ are reliable (1,023 values shown in Figure 2.10, and 573 values shown in Figure 2.11).

We use a t-test with unequal variance because we anticipate that this is the test that

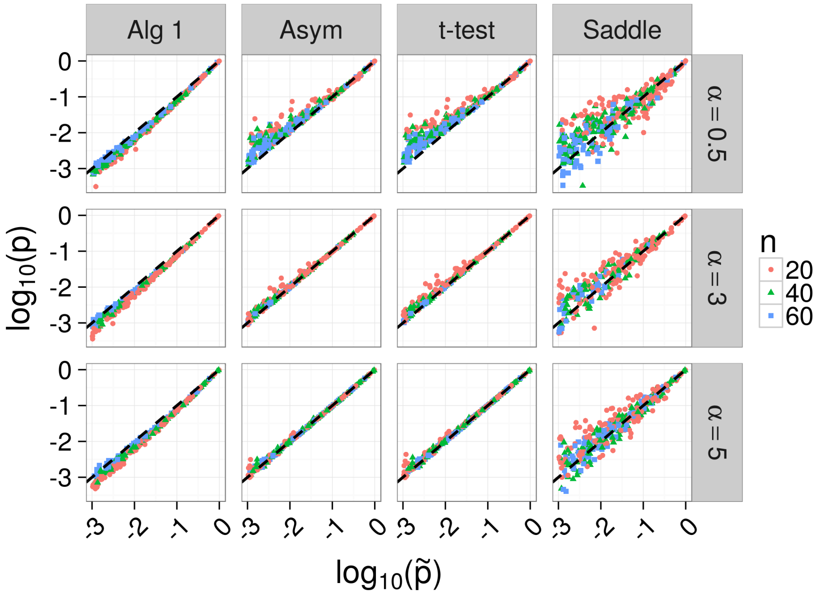


Figure 2.10: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data and equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* is a t-test with unequal variance, and *Saddle* is the saddlepoint approximation (see Section 2.2). $\tilde{p}$ is the p-values from simple Monte Carlo resampling with 10^5 resamples. SAMC results are not shown, as the EXPERT package does not produce p-values larger than 10^{-3}. Only simulations with $\tilde{p} > 10^{-3}$ shown (1,023 values shown). The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

would be used in practice, though we note that it tests a more general null hypothesis ($H_0 : \mu_x = \mu_y$) than the permutation test ($H_0 : P_x = P_y$). This puts our methods at a disadvantage.

Overall, Figures 2.10 and 2.11 suggest that our methods work well in this setting, though our resampling algorithm might be liberal for equal sample sizes and $\alpha = 0.5$. The t-test performs well in some scenarios, but tends to be too conservative, particularly for unequal sample sizes. Overall, the Saddlepoint approximation with fixed α and the MLE $\hat{\lambda}$ from the pooled data appears to have more variance than the other methods. Comparison with Figures 2.21 and 2.22 in Section 2.4 suggests that our resampling algorithm might be more reliable in this setting than moment corrected correlation (MCC) (Zhou and Wright, 2015) under the alternative and for unequal sample sizes.

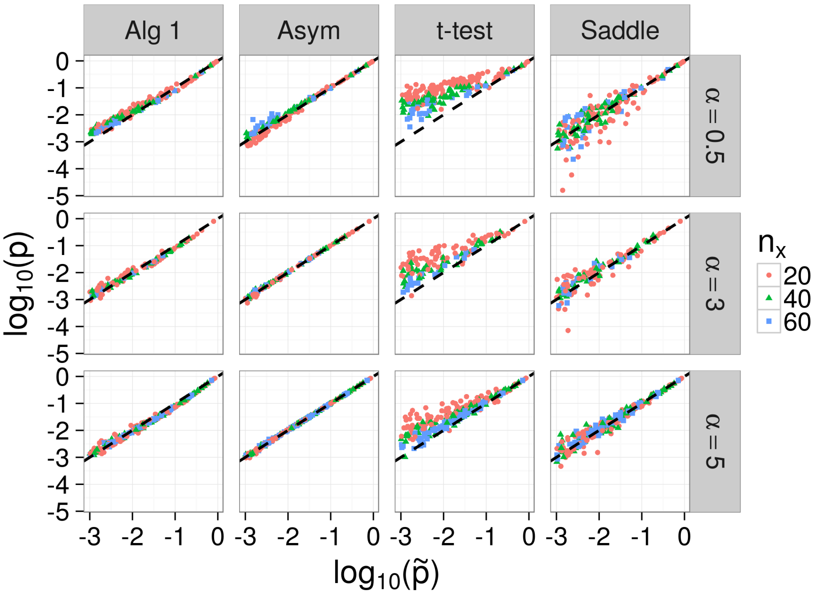


Figure 2.11: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data and unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. SAMC results are not shown, as the EXPERT package does not produce p-values larger than 10^{-3}. Only simulations with $\tilde{p} > 10^{-3}$ shown (573 values shown). The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

Under the Null Hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ for $\alpha = 0.5, 3, 5$ and $\lambda = 1, 5$, where λ is the rate parameter. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, and for each each n or n_x and combination of α and λ, we simulated 1,000 datasets (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures 2.12 and 2.13, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Saddle* is the saddlepoint approximation described in Section 2.2, *t-test* shows the p-value from a two-sided t-test with unequal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. We do not show results from the SAMC

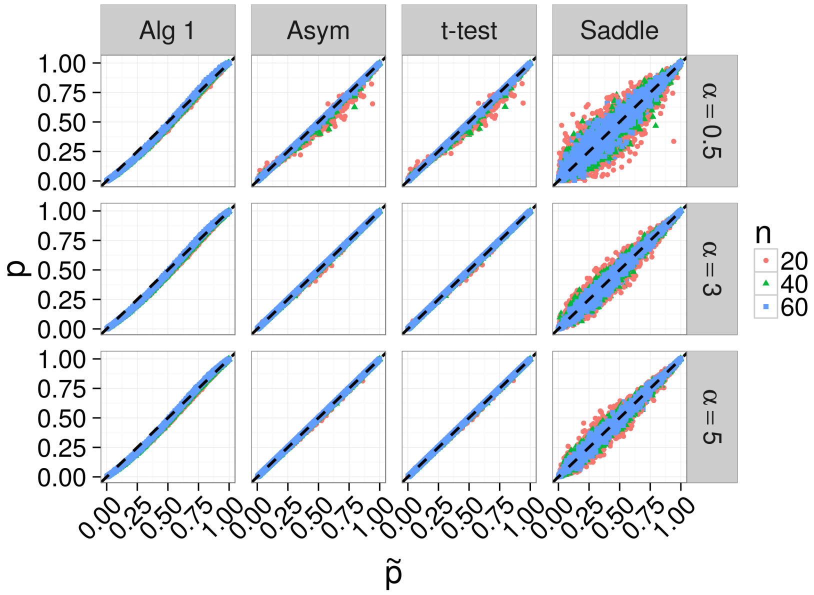


Figure 2.12: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data under the null $P_x = P_y$ with equal sample sizes of $n = n_x = n_y = 20, 40, 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Saddle* is the saddlepoint approximation described in Section 2.2, *t-test* shows the p-value from a two-sided t-test with unequal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

algorithm, because the EXPERT package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Figures 2.12 and 2.13 suggest that our methods work well in this setting, and have less variability than both the t-test and saddlepoint approximation (using fixed α fixed and the MLE $\hat{\lambda}$ from the pooled data).

Tables 2.5 and 2.6 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample size simulations, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling and 10^5 resamples, *Saddle* is the saddlepoint approximation described in Section 2.2, *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, and *t-test* shows the p-value from a two-sided t-test with equal variance.

Table 2.5: Type I error rates $\Pr(\text{p-value} \leq \text{signif level} | H_0)$ for $T = |\bar{x} - \bar{y}|$ with gamma data and equal sample sizes $n = n_x = n_y$. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *Saddle* is the saddlepoint approximation described in Section 2.2, *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, and *t-test* is a two-sided t-test with equal variance.

α	signif level	n_x	MC	Saddle	Alg 1	Asym	t-test
0.5	0.01	20	0.0110	0.0100	0.0165	0.0060	0.0045
		40	0.0125	0.0110	0.0150	0.0090	0.0085
		60	0.0115	0.0085	0.0140	0.0105	0.0105
	0.05	20	0.0495	0.0560	0.0665	0.0460	0.0410
		40	0.0515	0.0490	0.0660	0.0520	0.0485
		60	0.0455	0.0450	0.0595	0.0435	0.0425
	0.1	20	0.1000	0.1020	0.1280	0.1020	0.0945
		40	0.0995	0.0950	0.1260	0.1020	0.0975
		60	0.0980	0.0950	0.1230	0.0990	0.0965
3	0.01	20	0.0115	0.0070	0.0165	0.0095	0.0095
		40	0.0120	0.0115	0.0150	0.0120	0.0120
		60	0.0075	0.0075	0.0080	0.0070	0.0070
	0.05	20	0.0510	0.0465	0.0715	0.0515	0.0495
		40	0.0545	0.0575	0.0680	0.0560	0.0525
		60	0.0470	0.0475	0.0665	0.0480	0.0475
	0.1	20	0.0940	0.0990	0.1280	0.0980	0.0940
		40	0.0990	0.1000	0.1320	0.0990	0.0980
		60	0.0980	0.0985	0.1230	0.0980	0.0980
5	0.01	20	0.0115	0.0095	0.0175	0.0115	0.0115
		40	0.0090	0.0065	0.0130	0.0080	0.0080
		60	0.0045	0.0055	0.0085	0.0040	0.0040
	0.05	20	0.0525	0.0525	0.0675	0.0525	0.0505
		40	0.0525	0.0545	0.0715	0.0535	0.0520
		60	0.0460	0.0445	0.0580	0.0470	0.0470
	0.1	20	0.0965	0.0960	0.1220	0.0980	0.0955
		40	0.1070	0.1060	0.1370	0.1080	0.1080
		60	0.0925	0.0905	0.1300	0.0940	0.0915

Table 2.6: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = |\bar{x} - \bar{y}|$ with gamma data and unequal sample sizes $n_x \neq n_y$ (n_x shown and $n_y = 100$). α is the shape parameter in the gamma distribution, *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *Saddle* is the saddlepoint approximation described in Section 2.2, *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, and *t-test* is a two-sided t-test with equal variance.

α	signif level	n_x	MC	Saddle	Alg 1	Asym	t-test
0.5	0.01	20	0.0095	0.0095	0.0105	0.0085	0.0245
		40	0.0090	0.0060	0.0105	0.0070	0.0140
		60	0.0130	0.0160	0.0170	0.0105	0.0135
	0.05	20	0.0460	0.0465	0.0675	0.0440	0.0740
		40	0.0455	0.0470	0.0620	0.0445	0.0540
		60	0.0505	0.0500	0.0670	0.0495	0.0530
	0.1	20	0.0915	0.0930	0.1260	0.0845	0.1220
		40	0.0980	0.0945	0.1280	0.0960	0.1040
		60	0.1100	0.1080	0.1410	0.1100	0.1080
3	0.01	20	0.0085	0.0095	0.0155	0.0085	0.0135
		40	0.0135	0.0120	0.0185	0.0135	0.0140
		60	0.0070	0.0055	0.0090	0.0070	0.0070
	0.05	20	0.0440	0.0440	0.0665	0.0435	0.0480
		40	0.0480	0.0555	0.0695	0.0485	0.0530
		60	0.0470	0.0495	0.0635	0.0485	0.0460
	0.1	20	0.0875	0.0885	0.1260	0.0885	0.1000
		40	0.1050	0.1040	0.1350	0.1060	0.0975
		60	0.1040	0.1080	0.1370	0.1040	0.1040
5	0.01	20	0.0140	0.0110	0.0200	0.0140	0.0145
		40	0.0090	0.0100	0.0155	0.0090	0.0100
		60	0.0105	0.0090	0.0120	0.0110	0.0075
	0.05	20	0.0540	0.0535	0.0845	0.0540	0.0620
		40	0.0530	0.0525	0.0730	0.0525	0.0555
		60	0.0520	0.0510	0.0635	0.0520	0.0500
	0.1	20	0.1140	0.1160	0.1520	0.1140	0.1130
		40	0.0995	0.1000	0.1300	0.0995	0.1040
		60	0.1040	0.0985	0.1320	0.1050	0.1060

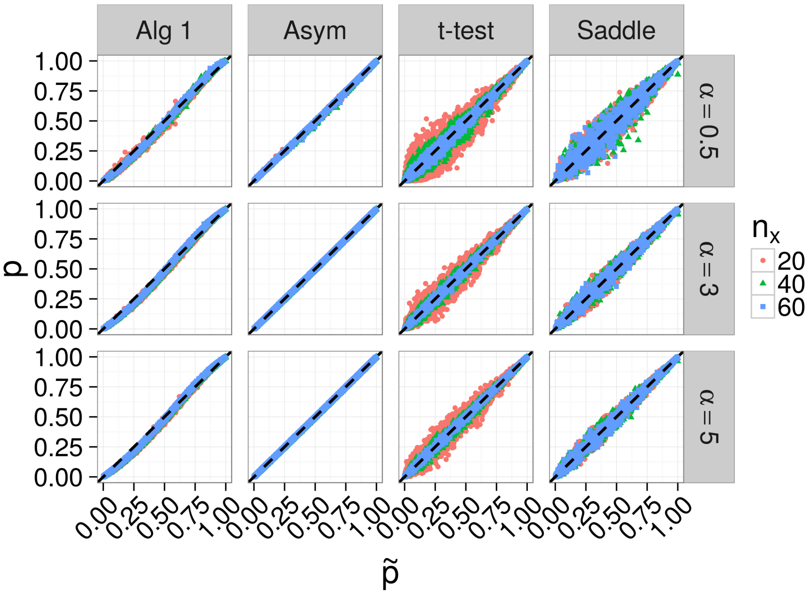


Figure 2.13: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data under the null $P_x = P_y$ with unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Saddle* is the saddlepoint approximation described in Section 2.2, *t-test* shows the p-value from a two-sided t-test with unequal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

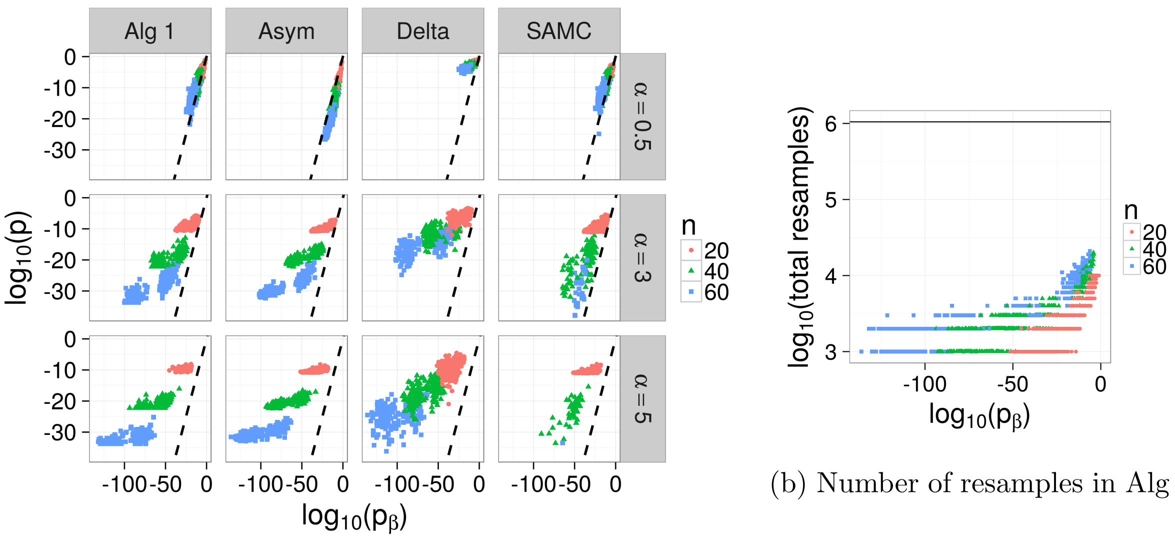
2.3.4 Ratio of Means with Gamma Data

In this subsection, we use the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with data generated as gamma random variables.

Small Sample Sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_y)$, where λ is the rate parameter, and $\alpha = 0.5, 3, 5$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For all simulations, we set $\lambda_x = 1$. For equal samples sizes, we set $\lambda_y = 7, 12.5$ for each n. For unequal sample sizes, we set $\lambda_y = 2.25, 2.75$ for all n_x for $\alpha = 0.5$, $\lambda_y = 2, 2.5$ for all n_x for $\alpha = 3$, and $\lambda_y = 1.75, 2.25$ for all n_x for $\alpha = 5$. We simulated 100 datasets for each combination of parameters.

Results for equal and unequal sample size are shown in Figures 2.14 and 2.15, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our



(a) p-values

(b) Number of resamples in Alg 1

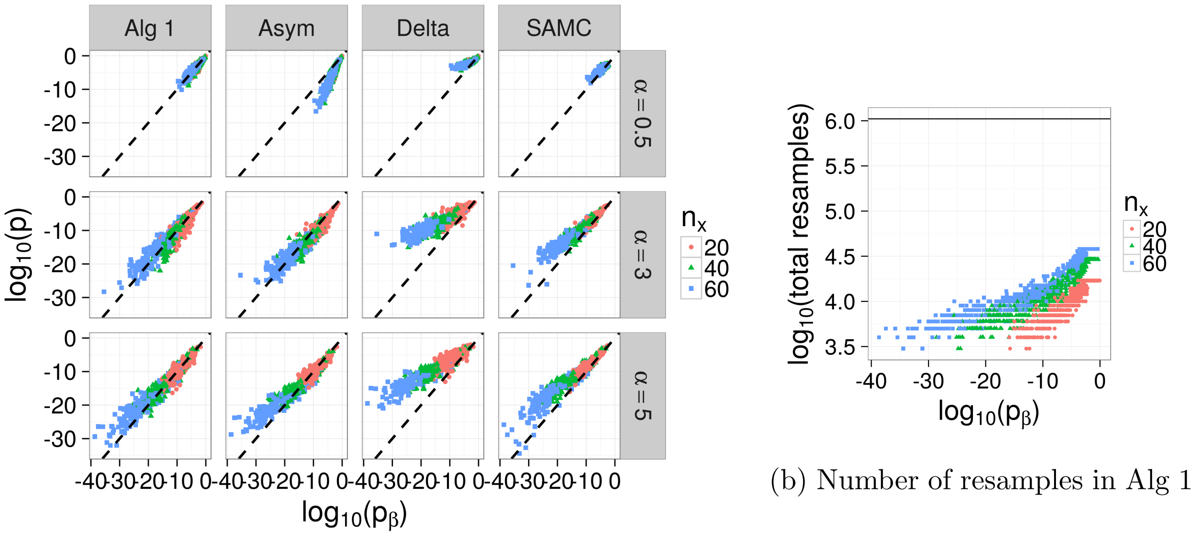
Figure 2.14: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and equal sample sizes of $n = n_x = n_y = 20,\ 40,\ 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 2.14b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 652 tests (points missing).

asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. Figures 2.14b and 2.15b show the number of resamples used by our resampling algorithm.

Under the Null Hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1)$ for $\alpha = 0.5, 3, 5$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$. For unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, we simulated 1,000 datasets for each combination of parameters (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures 2.16 and 2.17, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our



(a) p-values

(b) Number of resamples in Alg 1

Figure 2.15: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and unequal sample sizes of $n_x = 20$, 40, 60, $n_y = 100$, and rates $\lambda_y = 5, 10$, and $\lambda_x = 1$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in 2.15b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 304 tests (points missing).

asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. Given the large p-values, using 10^5 Monte Carlo resamples should be sufficient to obtain reliable estimates of the true permutation p-value. Therefore, this comparison demonstrates that the permutation p-value is not exactly the same as the p-value from the beta prime distribution. However, it appears reasonably close, so we use it as an approximation to the truth in other simulations in which the p-values are much smaller and simple Monte Carlo methods are not feasible.

We do not show results from the SAMC algorithm, because as noted above, the `EXPERT` package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Tables 2.7 and 2.8 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample sizes, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *Beta prime* is the p-value from the beta prime

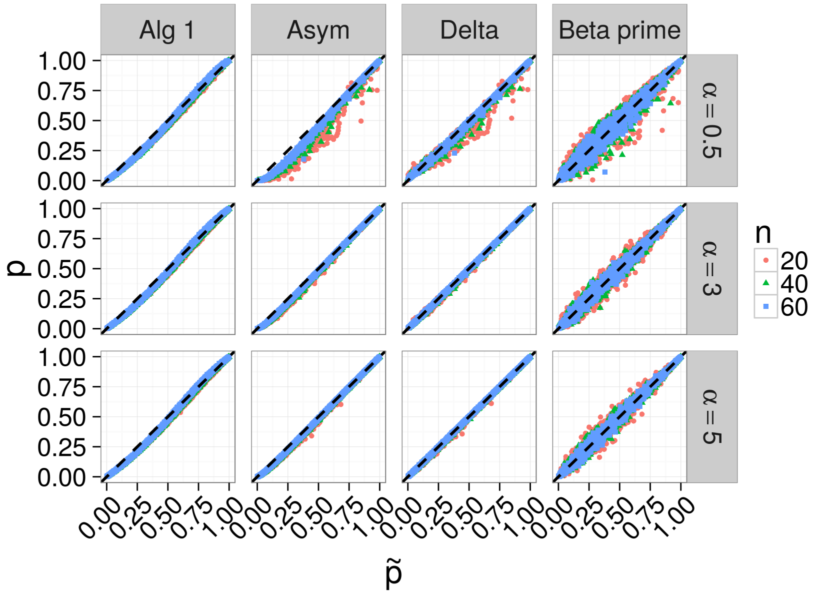


Figure 2.16: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data under the null of $P_x = P_y$ with equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

distribution, *Alg 1* is our resampling algorithm, and *Asym* is our asymptotic approximation.

2.4 Comparison with Additional Methods

2.4.1 Moment-Corrected Correlation

Moment-corrected correlation (MCC) (Zhou and Wright, 2015) is an analytical approximation to the permutation p-value, which is applicable in multiple testing situations in which the test statistic is permutationally equivalent to a single inner product. Where applicable, this approach is fast, as it does not involve resampling. However, if the test statistic of interest is not permutationally equivalent to an inner product, the MCC approach cannot be used.

The statistic $T = \bar{x} - \bar{y}$ fits into this setting, whereas, to the best of our knowledge, $T = \bar{x}/\bar{y}$ does not. To see this, let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ and $\boldsymbol{w} = (\underbrace{1/n_x, \ldots, 1/n_x}_{n_x}, \underbrace{-1/n_y, \ldots, -1/n_y}_{n_y})'$.

Table 2.7: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and equal sample sizes $n = n_x = n_y$. α is the shape parameter in the gamma distribution, *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asym* is our asymptotic approximation, *Delta* is the delta method, and *Beta prime* is the the beta prime distribution.

α	signif level	n	MC	Alg 1	Asym	Delta	Beta prime
0.5	0.01	20	0.013	0.018	0.093	0.002	0.015
		40	0.007	0.014	0.055	0.001	0.007
		60	0.007	0.010	0.047	0.002	0.011
	0.05	20	0.050	0.076	0.182	0.026	0.053
		40	0.050	0.072	0.135	0.037	0.055
		60	0.048	0.068	0.114	0.043	0.050
	0.1	20	0.110	0.136	0.243	0.106	0.108
		40	0.106	0.135	0.196	0.114	0.104
		60	0.096	0.127	0.178	0.101	0.097
3	0.01	20	0.007	0.012	0.027	0.003	0.006
		40	0.012	0.016	0.025	0.010	0.010
		60	0.012	0.015	0.025	0.012	0.008
	0.05	20	0.043	0.067	0.088	0.046	0.044
		40	0.053	0.062	0.073	0.052	0.051
		60	0.059	0.075	0.080	0.061	0.049
	0.1	20	0.095	0.126	0.143	0.103	0.090
		40	0.098	0.133	0.147	0.104	0.103
		60	0.095	0.115	0.116	0.097	0.093
5	0.01	20	0.009	0.015	0.023	0.009	0.009
		40	0.008	0.013	0.025	0.008	0.011
		60	0.012	0.012	0.019	0.012	0.013
	0.05	20	0.046	0.063	0.082	0.054	0.052
		40	0.048	0.063	0.066	0.050	0.043
		60	0.055	0.078	0.079	0.057	0.057
	0.1	20	0.093	0.130	0.139	0.106	0.099
		40	0.091	0.134	0.138	0.094	0.093
		60	0.115	0.138	0.136	0.116	0.112

Table 2.8: Type I error rates $\Pr(\text{p-value} \leq \text{signif level} | H_0)$ for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and unequal sample sizes $n_x \neq n_y$ (n_x shown and $n_y = 100$). α is the shape parameter in the gamma distribution, *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asym* is our asymptotic approximation, *Delta* is the delta method, and *Beta prime* is the beta prime distribution.

α	signif level	n_x	MC	Alg 1	Asym	Delta	Beta prime
0.5	0.01	20	0.011	0.015	0.065	0.006	0.011
		40	0.015	0.018	0.053	0.003	0.013
		60	0.008	0.011	0.042	0.003	0.012
	0.05	20	0.043	0.069	0.128	0.047	0.053
		40	0.057	0.072	0.133	0.048	0.056
		60	0.052	0.071	0.112	0.045	0.050
	0.1	20	0.098	0.121	0.179	0.109	0.091
		40	0.113	0.141	0.195	0.119	0.108
		60	0.106	0.126	0.172	0.109	0.098
3	0.01	20	0.011	0.016	0.023	0.012	0.011
		40	0.005	0.011	0.027	0.005	0.009
		60	0.011	0.013	0.017	0.011	0.011
	0.05	20	0.047	0.070	0.073	0.059	0.039
		40	0.058	0.065	0.069	0.057	0.054
		60	0.053	0.066	0.070	0.050	0.052
	0.1	20	0.088	0.128	0.135	0.104	0.087
		40	0.094	0.124	0.124	0.101	0.089
		60	0.094	0.119	0.117	0.097	0.097
5	0.01	20	0.010	0.014	0.022	0.007	0.009
		40	0.011	0.011	0.017	0.011	0.009
		60	0.015	0.020	0.025	0.015	0.018
	0.05	20	0.058	0.074	0.085	0.066	0.054
		40	0.046	0.057	0.059	0.048	0.052
		60	0.059	0.081	0.085	0.061	0.062
	0.1	20	0.110	0.145	0.143	0.121	0.114
		40	0.081	0.114	0.108	0.085	0.088
		60	0.113	0.145	0.138	0.118	0.115

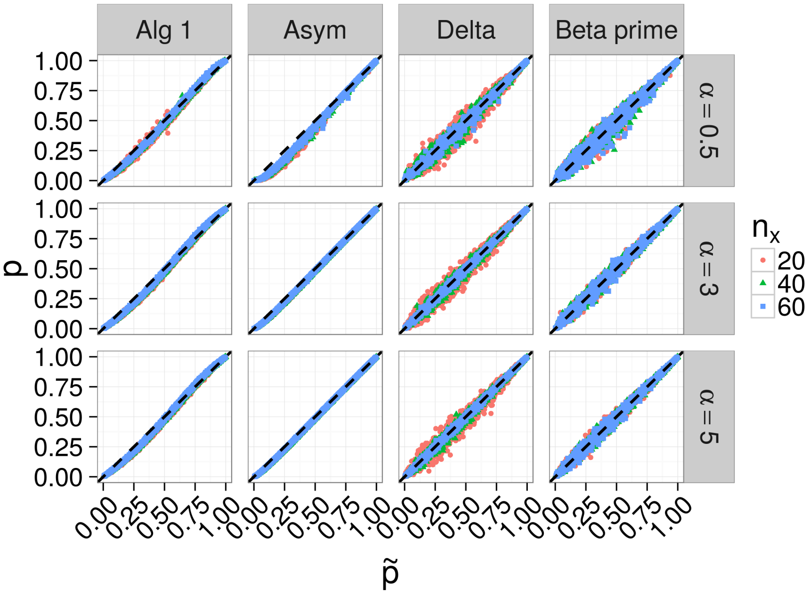


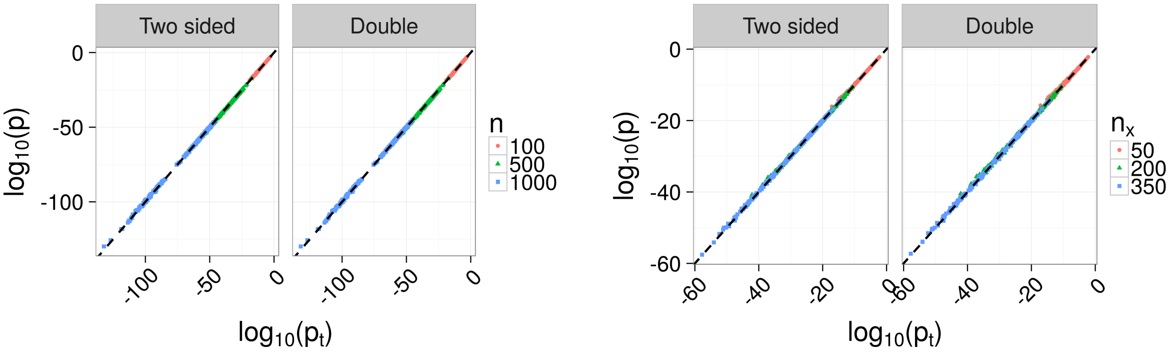
Figure 2.17: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data under the null of $P_x = P_y$ with unequal sample sizes of $n_x = 20$, 40, 60 and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

Then $\bar{x} - \bar{y} = \boldsymbol{z}'\boldsymbol{w}$. In contrast, $\bar{x}/\bar{y}$ cannot be written in this form, and we conjecture that it is not permutationally equivalent to any statistic that can be.

Figures 2.18 through 2.20 show simulation results for two-sided and doubled p-values, as described by Zhou and Wright (2015), using the `mcc` package (Zhou, 2014) under the same normal data settings as in Section 2.3.1. While MCC is more reliable for large sample sizes (Figure 2.18), MCC appears to suffer from the same bias as our methods for small sample sizes (Figure 2.19). Furthermore, we do not think that MCC can be used to obtain p-values for the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$.

Figures 2.21 and 2.22 show simulation results for two-sided and doubled p-values for small sample sizes and under the null, respectively, using the `mcc` package (Zhou, 2014) under the same gamma data settings as in Section 2.3.3. In Figure 2.21, we used $B = 10^5$ resamples to obtain the Monte Carlo estimate $\tilde{p}$ of the true permutation p-value, and only show results for $\tilde{p} > 10^{-3}$ to ensure reliable estimates (1,019 values shown in Figure 2.21a, and 705 values shown in Figure 2.21b).

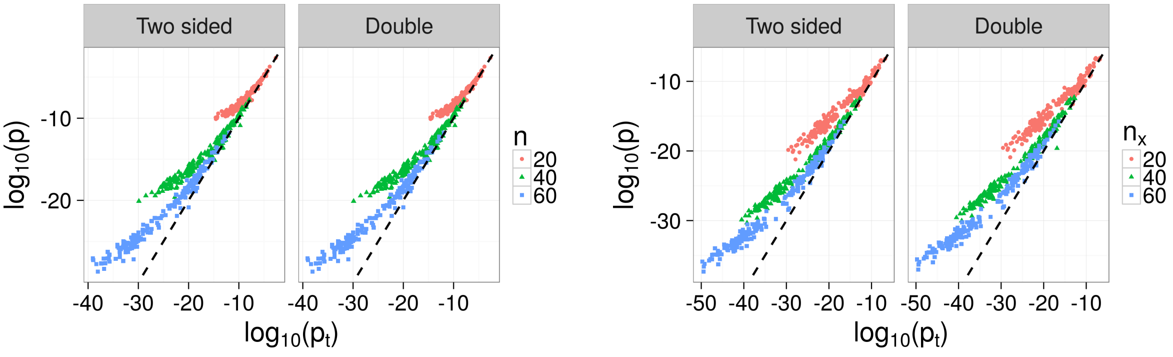
As seen in Figure 2.21, in many cases the MCC method substantially underestimated the permutation p-value for equal sample sizes $n_x = n_y$ and $\alpha = 0.5$. We did not observe this



(a) $n_x = n_y$

(b) $n_x \neq n_y$

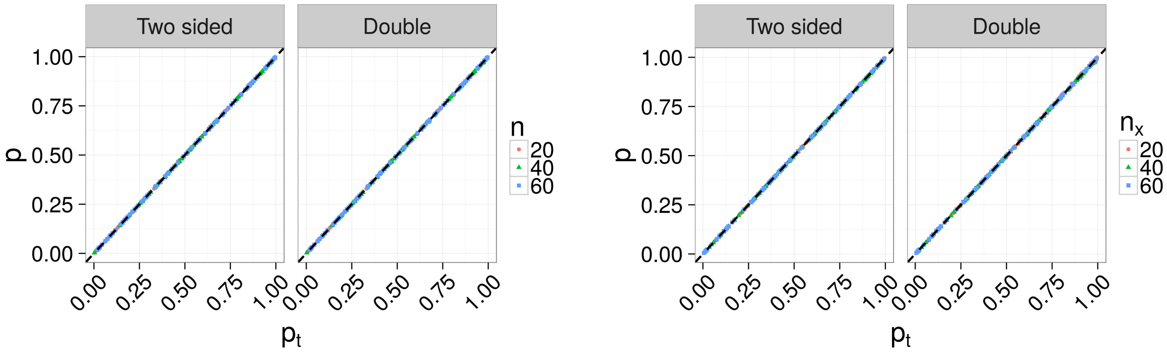
Figure 2.18: MCC with large sample sizes for $T = |\bar{x} - \bar{y}|$ with normal data and equal sample sizes of $n = n_x = n_y = 100, 500, 1,000$, and unequal sample sizes of $n_x = 50, 200, 350$ with $n_y = 500$. In both cases, data were simulated as normal random variables with $\mu_y = 0$, $\mu_x = 0.75, 1$ and $\sigma_x^2 = \sigma_x^2 = 1$. p_t is the p-value from a t-test with equal variance. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

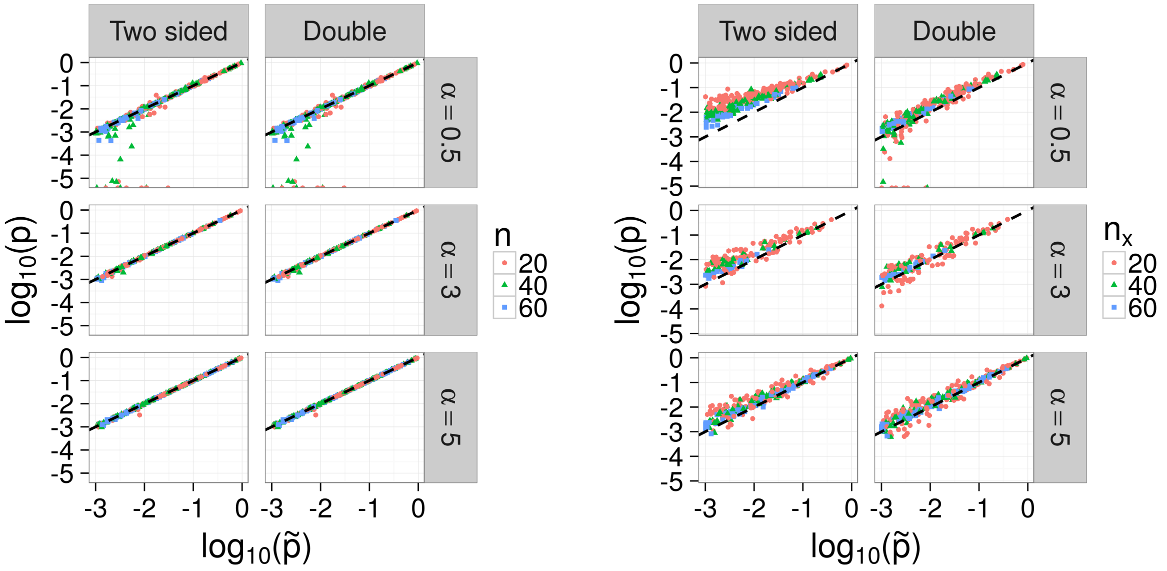
Figure 2.19: MCC with small sample size for $T = |\bar{x} - \bar{y}|$ with normal data and equal sample sizes of $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as normal random variables with $\mu_y = 0$, $\mu_x = 2, 3$ and $\sigma_x^2 = \sigma_x^2 = 1$. p_t is the p-value from a t-test with equal variance. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

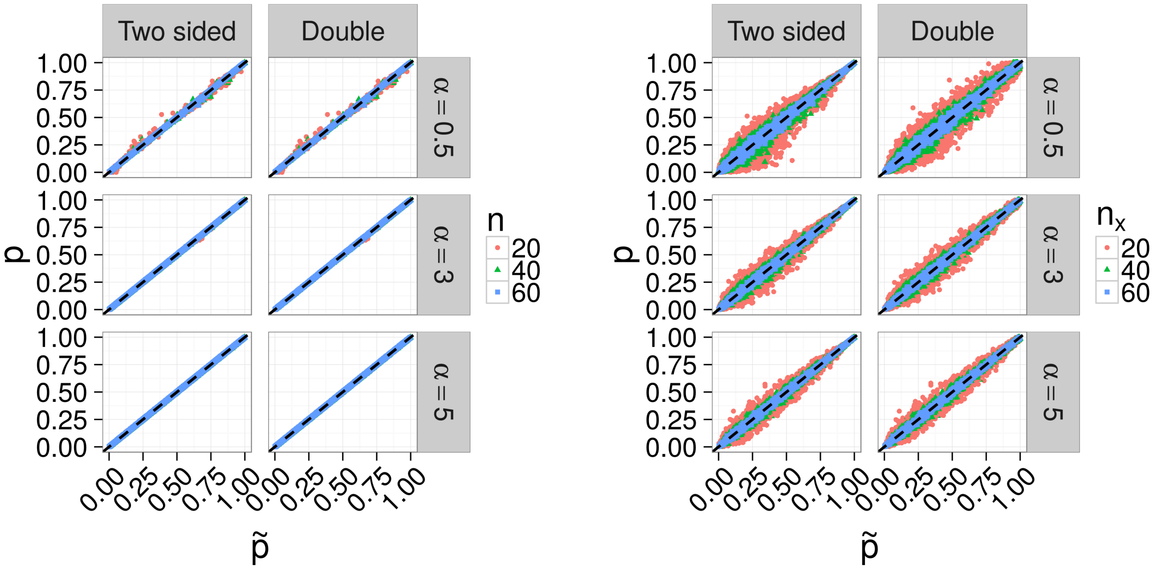
Figure 2.20: MCC under the null hypothesis for $T = |\bar{x} - \bar{y}|$ with normal data for equal sample sizes of $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as normal random variables with $\mu_y = \mu_x = 0$ and $\sigma_x^2 = \sigma_x^2 = 1$. p_t is the p-value from a t-test with equal variance. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

Figure 2.21: MCC with small sample size for $T = |\bar{x} - \bar{y}|$ with gamma data and equal sample size $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as gamma random variables, as described in Section 2.3.3. $\tilde{p}$ is the p-value from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

Figure 2.22: MCC under the null hypothesis for $T = |\bar{x} - \bar{y}|$ with gamma data for equal sample sizes of $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as gamma random variables, as described in Section 2.3.3. $\tilde{p}$ is the p-value from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement.

tendency with our resampling algorithm (see Figures 2.10 and 2.11).

2.4.2 Saddlepoint Approximations

Saddlepoint approximations can be used to estimate permutation p-values (Robinson, 1982). As shown in Table 2.9, estimates from our methods are comparable to those from saddlepoint approximations when using the statistic $T = |\bar{x} - \bar{y}|$. However, unlike saddlepoint approximations, our resampling algorithm requires no derivations.

2.5 Simulations Under Null Hypotheses for Single Parameters

Neuhaus (1993), Janssen (1997), Chung and Romano (2013), and others have extended permutation tests to be valid not only under the null $P_x = P_y$, but also under the more general null that $\theta(P_x) = \theta(P_y)$, where $\theta(P)$ is a single parameter. For example, for $X \sim$

Table 2.9: Comparison with Saddlepoint approximations for $T = |\bar{x}-\bar{y}|$. Datasets are from Robinson (1982, Table 2), who obtained them from Lehmann (1975). Dataset 1 pertains to hours of pain relief due to two different drugs ($n_x = n_y = 8$), and Dataset 2 pertains to the effect of an analgesia for two classes ($n_x = 7, n_y = 10$). The exact and saddlepoint p-values are from Robinson (1982). The the p-value from our resampling algorithm ($\tilde{p}_{\text{pred}}$) is the mean from 100 runs; the first and third quantiles were (0.080, 0.088) for dataset 1, and (0.011, 0.012) for dataset 2.

Method	Dataset 1	Dataset 2
Exact	0.102	0.012
First saddlepoint	0.089	0.010
Second saddlepoint	0.101	0.011
$\tilde{p}_{\text{pred}}$	0.083	0.012
$\hat{p}_{\text{asym}}$	0.092	0.013

$N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$, we might be interested in the alternative $H_1 : \mu_x \neq \mu_y$, even if $\sigma_x^2 \neq \sigma_y^2$.

As described by Chung and Romano (2013), in order to obtain a test procedure that is asymptotically valid in the above setting where $\sigma_x^2 \neq \sigma_y^2$, we need to replace $T = |\bar{x} - \bar{y}|$ with the studentized statistic

$$T = \frac{|\bar{x} - \bar{y}|}{\sqrt{s_x^2/n_x + s_y^2/n_y}} \tag{2.6}$$

where $s_x^2 = (n_x - 1)^{-1} \sum_i (x_i - \bar{x})^2$ and $s_y^2 = (n_y - 1)^{-1} \sum_j (y_j - \bar{y})^2$ are the sample variances. For each permutation, we compute the quantities $\bar{x}^*, \bar{y}^*, s_x^{*2}$, and s_y^{*2} with the permuted datasets. In this section, we conduct simulations using (2.6) when $P_x \neq P_y$ under the null $H_0 : \mu_x = \mu_y$ and alternative $H_1 : \mu_x \neq \mu_y$.

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(0, \sigma_x^2)$ and $Y_j \overset{\text{iid}}{\sim} N(0, \sigma_y^2)$, where $\sigma_x^2 = 9$ and $\sigma_y^2 = 1$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, we simulated 1,000 datasets for each combination of parameters. Figures 2.23 and 2.24 show the results with equal and unequal sample sizes, respectively.

As seen in Figures 2.23 and 2.24, the permutation test with the unstudentized statistic is relatively unaffected in our simulation under equal sample sizes, but is inaccurate for unequal sample sizes. This is as expected. By using a studentized statistic, our method is accurate even for unequal sample sizes. For comparison, Figures 2.23 and 2.24 also show the p-value from a t-test with unequal variance, as well as a Monte Carlo estimate using the unstudentized statistic $T = |\bar{x} - \bar{y}|$.

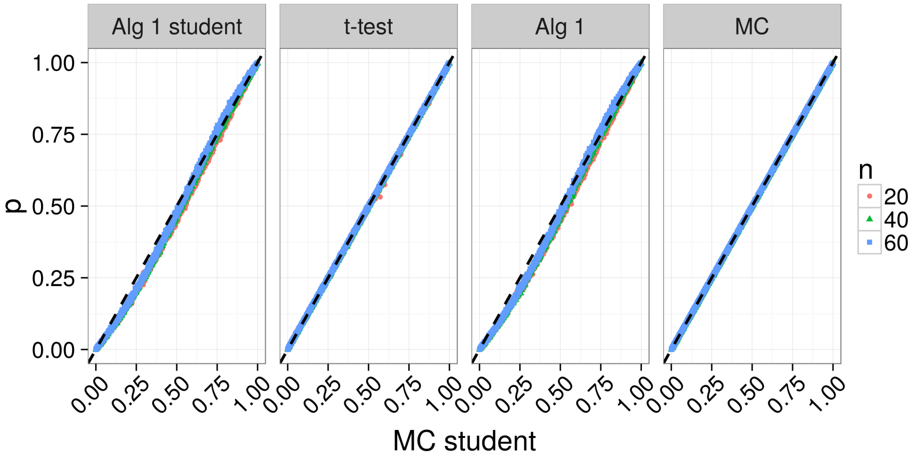


Figure 2.23: Simulation results under the null $\mu_x = \mu_y$ with normal data and unequal sample sizes of $n = n_x = n_y = 20, 40, 60$. *Alg 1 Student* and *Alg 1* are our resampling algorithm with the studentized (2.6) and unstudentized statistics, with $B_{\text{pred}} = 10^3$ resamples in each partition. *t-test* is the p-value from a two-sided t-test with unequal variance. *MC student* and *MC* are Monte Carlo estimates with the studentized (2.6) and unstudentized statistics, respectively, with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

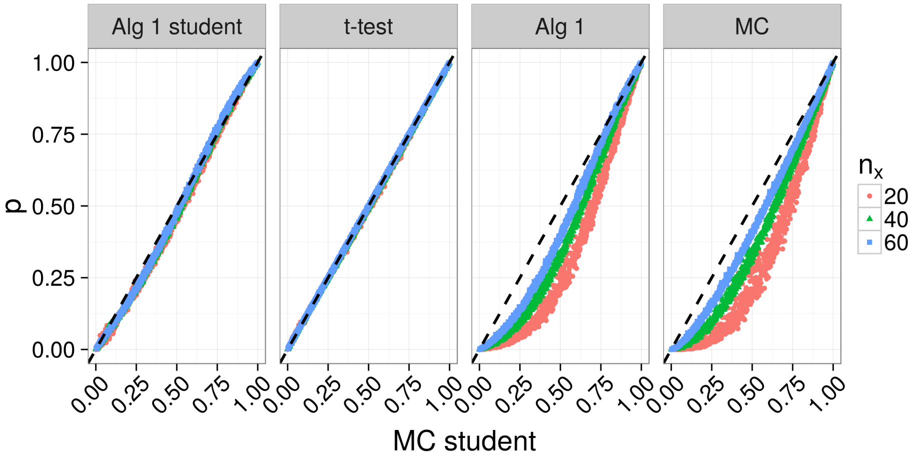


Figure 2.24: Simulation results under the null $\mu_x = \mu_y$ with normal data with unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1 Student* and *Alg 1* are our resampling algorithm with the studentized (2.6) and unstudentized statistics, and with $B_{\text{pred}} = 10^3$ resamples in each partition. *t-test* is the p-value from a two-sided t-test with unequal variance. *MC student* and *MC* are Monte Carlo estimates with the studentized (2.6) and unstudentized statistics, respectively, with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods.

2.6 Asymptotic Test of the Ratio of Means via the Delta Method and Application to Cancer Genomic Data

Let $\bar{x}$ and $\bar{y}$ be the sample means, and $s_x^2 = (n_x - 1)^{-1} \sum_i (x_i - \bar{x})^2$ and $s_y^2 = (n_y - 1)^{-1} \sum_i (y_i - \bar{y})^2$ be the sample estimates of variance. By the central limit theorem, for n_x, n_y sufficiently large, and assuming independence between samples,

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2/n_x & 0 \\ 0 & \sigma_y^2/n_y \end{bmatrix} \right).$$

Let $g(\bar{x}, \bar{y}) = (\bar{x}/\bar{y})$. Then $\nabla g = (1/\bar{y}, -\bar{x}/\bar{y}^2)'$, and by the delta method $\bar{x}/\bar{y} \to N(\theta, \tau_1^2)$, where $\theta = g(\mu_x, \mu_y) = \mu_x/\mu_y$ and

$$\tau_1^2 = \nabla g^T(\mu_x, \mu_y) \begin{bmatrix} \sigma_x^2/n_x & 0 \\ 0 & \sigma_y^2/n_y \end{bmatrix} \nabla g(\mu_x, \mu_y) = \frac{\sigma_x^2}{n_x} \frac{1}{\mu_y^2} + \frac{\sigma_y^2}{n_y} \frac{\mu_x^2}{\mu_y^4}.$$

Using unbiased estimates for the variance, we get

$$\hat{\tau_1}^2 = \frac{s_x^2}{n_x \bar{y}^2} + \frac{s_y^2 \bar{x}^2}{n_y \bar{y}^4}$$

where s_x^2 and s_y^2 are the sample variances for $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Similarly, we estimate the variance of $\bar{y}/\bar{x}$ as

$$\hat{\tau_2}^2 = \frac{s_y^2}{n_y \bar{x}^2} + \frac{s_x^2 \bar{y}^2}{n_x \bar{x}^4}.$$

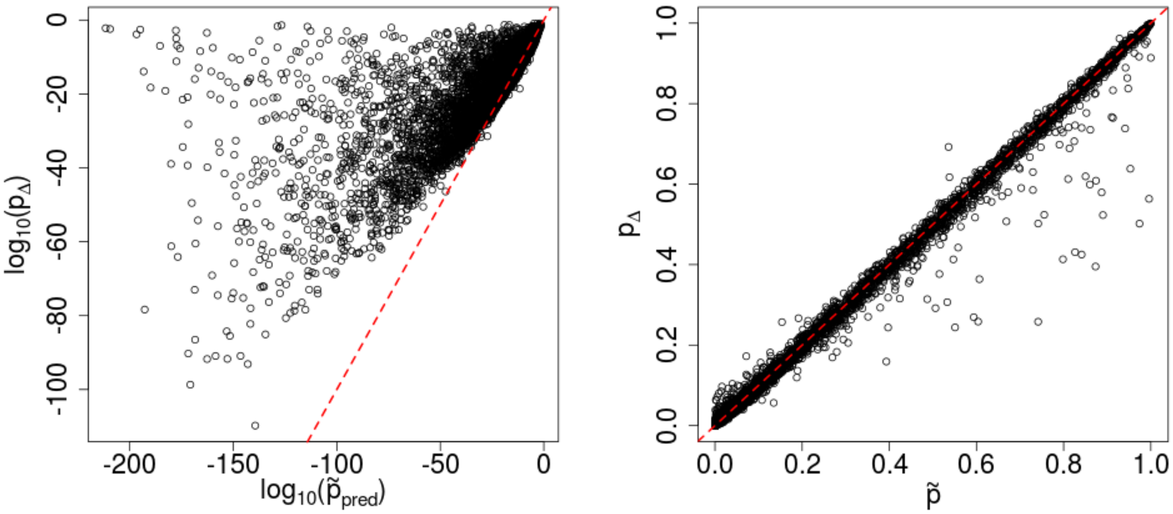
Therefore, to test the null $H_0 : \mu_x/\mu_y = 1$ versus the alternative $H_1 : \mu_x/\mu_y \neq 1$, the two-sided p-value using the delta method and unbiased estimates of variance is

$$p_\Delta = \begin{cases} \Pr(Z > \bar{x}/\bar{y}) + \Pr(U \leq \bar{y}/\bar{x}), & \bar{x}/\bar{y} \geq 1 \\ \Pr(U > \bar{y}/\bar{x}) + \Pr(Z \leq \bar{x}/\bar{y}), & \bar{x}/\bar{y} < 1 \end{cases},$$

where $Z \sim N(1, \hat{\tau_1}^2)$ and $U \sim N(1, \hat{\tau_2}^2)$. We use the Δ subscript in p_Δ to emphasize that the p-value is from the delta method. We note that p_Δ is potentially problematic, particularly if $\hat{\tau_1}^2$ or $\hat{\tau_2}^2$ are large, because the ratio is bounded below by zero, but the normal distribution is not.

We note that by allowing for unequal variance, we are testing a different null hypothesis than with the permutation test ($H_0 : P_x = P_y$). However, we expect that in practice, researchers would allow for unequal variance when using the delta method, which is why we use it as a basis for comparison. This comparison puts the permutation test at a disadvantage, but as shown in the simulations, the permutation test still performs better than the delta method.

Figure 2.25 compares estimates of the permutation p-values from our resampling algorithm ($\tilde{p}_{\text{pred}}$) to p_Δ for the cancer genomic data in Section 1.6. The dashed lines have an intercept of zero and slope of one, and indicate agreement. As seen in Figure 2.25, p_Δ tends to be an overestimate for small p-values, which is the same trend observed in the simulations. Out of the 100 genes with the smallest p_Δ, only three were identified by Zhan et al. (2015) as strongly distinguishing between LUAD and LUSC (*PVRL1*, *PERP*, and *ATP1B3*).



(a) Genes with $\tilde{p} \leq 1 \times 10^{-3}$ (10,302 genes)

(b) Genes with $\tilde{p} > 1 \times 10^{-3}$ (5,084 genes)

Figure 2.25: p-values for cancer genomic data: Comparison of results with the delta method (p_Δ) and our resampling algorithm ($\tilde{p}_{\text{pred}}$) with $B_{\text{pred}} = 10^3$ resamples within each partition, or with simple Monte Carlo ($\tilde{p}$) with a total of $B = 10^3$ resamples (see Section 6). The diagonal dashed lines have a slope of 1 and an intercept of 0, and indicate agreement between the methods.

Chapter 3

Tests of Matrix Structure for Construct Validation

3.1 Introduction

Psychologists and other behavioral scientists are frequently interested in whether a survey or questionnaire measures the concepts it purports to measure. Attempts to address this issue are referred to as construct validation. Since the construct cannot be directly observed, it is impossible to assess its validity directly. Instead, researchers divide construct validity into different aspects that can be addressed separately. These different aspects are called criterion-related validity, convergent validity, discriminant validity, and content validity. As Kline (2011) describes, criterion-related validity concerns the consistency of the test with external measures, convergent and discriminant validity refer to the magnitudes of correlations between test questions, and content validity is the degree to which the questions can be interpreted to represent the underlying scientific construct. By considering these different aspects of validity together, researchers can produce an overall body of evidence either in favor of or against validating a construct.

The statistical aspects of construct validation are covered by convergent and discriminant validity. Convergent validity occurs when the magnitudes of the correlations are high between items that are hypothesized to measure the same construct, and discriminant validity occurs when the magnitudes of the correlations are low between items hypothesized to measure different constructs (Kline, 2011). In this chapter, we describe tests for matrix structure that can be used to assess convergent and discriminant validity. These matrix structure tests can be used either by themselves or to check the robustness of other methods, such as confirmatory factor analysis (CFA).

In Section 3.2, we provide a motivating example. In Section 3.3, we describe methods for testing matrix structure based on the quadratic assignment framework of Hubert and Schultz (1976), and derive rates of convergence for the overall test. In Section 3.4, we discuss related methods, including linear models, pattern hypothesis tests of correlation coefficients, and CFA. In Section 3.5, we investigate the behavior of these methods through simulations, and in Section 3.6, we demonstrate these methods by analyzing the big five personality traits questionnaire conducted as part of the 2010 Health and Retirement Survey (HRS, 2016). In Section 3.8, we discuss the benefits and limitations of using tests of matrix structure for construct validation, as well as potential extensions. As noted in Section 3.9, we have implemented the methods described in this chapter in the `R` package `matrixTest`.

3.2 Motivating Example

As a motivating example, we analyze the big five personality traits questionnaire that was given as part of the 2010 Health and Retirement Study (HRS, 2016). HRS is a "longitudinal panel study that surveys a representative sample of approximately 20,000 Americans over the age of 50 every two years" (HRS, 2016). The big five personality traits questionnaire is given as part of the HRS Psychosocial and Lifestyle Questionnaire, which is administered to a rotating, random selection of 50% of the HRS respondents. The HRS data are publicly available at `http://hrsonline.isr.umich.edu`. The Psychosocial and Lifestyle Questionnaire is part of the core data release, in the file labeled `LB_R` (leave-behind, respondent).

In 2010, 7,215 respondents provided complete responses to the big five personality trait questionnaire, and an additional 1,050 subjects provided partial responses. The big five personality traits questionnaire contains 31 items, each of which was recorded on a four-point Likert scale. In what follows, we did a complete case analysis and did not incorporate sampling weights into the estimation of correlation coefficients, though this could be done in future analyses.

To assess convergent and divergent validity, we were interested in the magnitude of the correlations, but not the direction. Figure 3.1 shows the absolute values of Spearman's rank correlation matrix for the 31 items in the questionnaire, ordered by the hypothesized groups, which are outlined. From upper left to lower right, the outlined groups are: 1) neuroticism, 2) extroversion, 3) agreeableness, 4) openness to experience, and 5) conscientiousness. The questionnaire items are described in Appendix A.

From a visual inspection of Figure 3.1, the first block (neuroticism) appears to exhibit both convergent validity (high within block correlation) and divergent validity (low between-block validity). The second, third and fourth blocks (extroversion, agreeableness, and open-

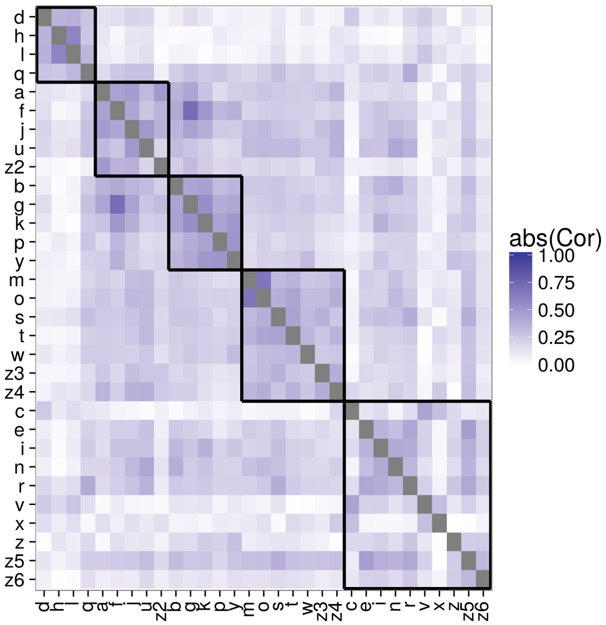


Figure 3.1: Absolute values of the Spearman rank correlation matrix for the HRS big five personality traits questionnaire ordered by hypothesized groups. From upper left to lower right, the groups are: 1) neuroticism, 2) extroversion, 3) agreeableness, 4) openness to experience, and 5) conscientiousness. Diagonal elements are all equal to 1, and are not included in the color gradient. Item labels (d, h, l, . . .) are taken from the HRS questionnaire. The items are described in Appendix A.

ness to experience) appear to exhibit convergent validity, though the relatively high correlations between these blocks makes it unclear whether they also exhibit divergent validity. The fifth block (conscientiousness) does not appear to exhibit either convergent or divergent validity. We next develop methods to formally test convergent and divergent validity using nonparametric tests of matrix structure.

3.3 Tests of Matrix Structure

Several authors have developed methods for testing matrix structure, including Bock and Bargmann (1966), Srivastava (1966), McDonald (1974) and Jöreskog (1978). The approach we describe has a similar goal to these methods, but differs in the way hypothesized matrix structures are assessed. Most notably, our approach sets up a traditional null hypothesis that researchers seek to reject, and does not use a goodness of fit (GOF) test or index to evaluate model fit.

3.3.1 Block Diagonal Structure

Let A be a $p \times p$ symmetric matrix. In our applications, A is typically the covariance or correlation matrix, or the absolute values of the covariance or correlation matrix. We are interested in whether A is approximately block diagonal:

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_K \\ & & & \end{pmatrix}$$

where blocks A_1 through A_K have respective dimensions $p_1 \times p_1, \ldots, p_K \times p_K$, and $\sum_{k=1}^{K} p_k \leq p$. When A is the covariance matrix, this is the structure implied by a CFA model in which each item loads onto no more than one latent variable. Throughout this chapter, we use the terms *group* and *block* interchangeably.

By approximately block diagonal, we mean that the elements in blocks $A_1, \ldots, A_K$ are larger in absolute value than elements in the non-blocks. If A were perfectly block-diagonal, all elements in blocks $A_1, \ldots, A_K$ would be non-zero, and all other elements would be zero. Figure 3.1 is an example where A is the elementwise absolute values of the correlation matrix, with $p = 31$ variables and a hypothesized $K = 5$ blocks of sizes $p_1 = 4$, $p_2 = 5$ and $p_3 = 5$, $p_4 = 7$, and $p_5 = 10$. If we exclude the fifth block from Figure 3.1, then $\sum_{k=1}^{4} p_k < p$, and the hypothesized block-diagonal structure would not extend all the way to the bottom right corner of the correlation matrix.

3.3.2 Hubert's Γ

Hubert's Γ (Hubert and Schultz, 1976) was originally proposed by Mantel (1967). Consequently, some authors, including Good (2000), refer to the statistic as Mantel's U. However, we follow most authors, including Jain and Dubes (1988), Halkidi et al. (2001) and Zaki and Jr. (2014) and refer to the statistic as Hubert's Γ, especially since our methods are based on the quadratic assignment framework of Hubert and Schultz (1976).

To define Hubert's Γ, let v_i be the label for the variable in row and column i of matrix A and let Δ be a $p \times p$ matrix with element δ_{ij} in row i and column j, where

$$\delta_{ij} = \begin{cases} 1 & \text{if variables } v_i \text{ and } v_j \text{ are hypothesized to belong to the same block} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we denote the element in row i and column j of A as a_{ij}. Let $N = p(p-1)/2$ be the number of upper triangular elements in A, where the upper triangular elements form the set $\{a_{ij} : i < j\}$. Let $\boldsymbol{a} = (a_{12}, a_{13}, a_{23}, a_{14}, a_{24}, \ldots, a_{N-1,N})^T$ be the $N \times 1$ vector of the upper triangular elements of A, and let $\boldsymbol{\delta} = (\delta_{12}, \delta_{13}, \delta_{23}, \delta_{14}, \delta_{24}, \ldots, \delta_{N-1,N})^T$ be the $N \times 1$ vector of the upper triangular elements of Δ. Since the A and Δ matrices are symmetric, we do not need to consider the lower triangular elements. Hubert's Γ is defined as the mean element-wise product between the upper triangular elements of A and Δ, given as $\Gamma = N^{-1}\sum_{i<j} a_{ij}\delta_{ij} = N^{-1}\boldsymbol{a}^T\boldsymbol{\delta}$.

We use the normalized Γ, which is more interpretable. Let $\bar{a} = N^{-1}\sum_{i<j} a_{ij}$ and $\hat{\sigma}_a^2 = (N-1)^{-1}\sum_{i<j}(a_{ij} - \bar{a})^2$ be the sample mean and variance of the elements in $\boldsymbol{a}$, let $\bar{\delta} = N^{-1}\sum_{i<j}\delta_{ij}$ and $\hat{\sigma}_\delta^2 = (N-1)^{-1}\sum_{i<j}(\delta_{ij} - \bar{\delta})^2$ be the sample mean and variance of the elements in $\boldsymbol{\delta}$, and let $\hat{\sigma}_{a\delta}^2 = (N-1)^{-1}\sum_{i<j}(a_{ij} - \bar{a})(\delta_{ij} - \bar{\delta})$ be the sample covariance between $\boldsymbol{a}$ and $\boldsymbol{\delta}$. Then the normalized Γ, which we denote as Γ_{norm}, is defined as the Pearson correlation between $\boldsymbol{a}$ and $\boldsymbol{\delta}$, given as

$$\Gamma_{\text{norm}} = \frac{\sum_{i<j}(a_{ij} - \bar{a})(\delta_{ij} - \bar{\delta})}{\sqrt{\sum_{i<j}(a_{ij} - \bar{a})^2 \sum_{i<j}(\delta_{ij} - \bar{\delta})^2}} = \frac{\hat{\sigma}_{a\delta}^2}{\hat{\sigma}_a \hat{\sigma}_\delta}. \tag{3.1}$$

Since Γ_{norm} is a correlation, $-1 \leq \Gamma_{\text{norm}} \leq 1$.

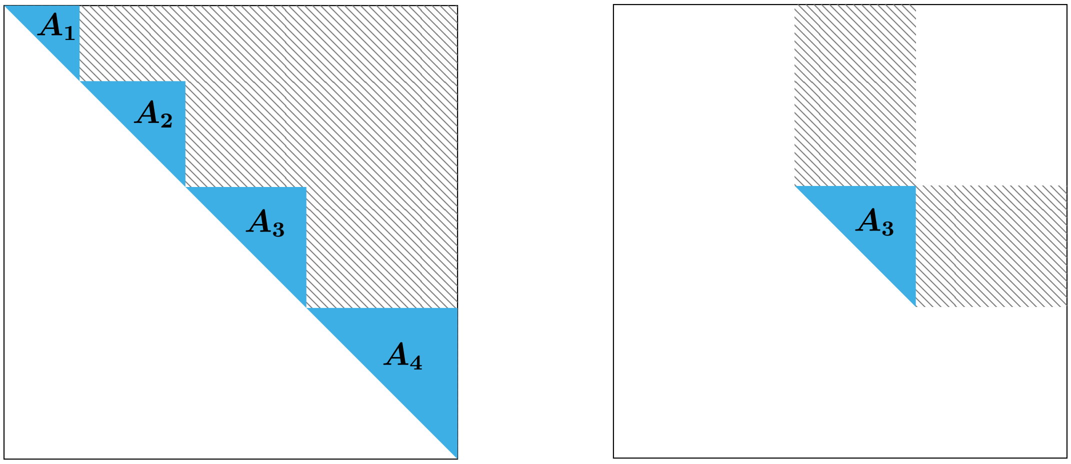
In general, the Δ matrix can be replaced by any conformable matrix in calculating Γ and Γ_{norm} depending on the hypothesis a researcher wants to test. As we show in Section 3.4.1, Γ_{norm} with Δ as defined above is related to the slope from a linear model that contrasts the within-block elements with the between-block elements.

Large positive values of Γ_{norm} (values near 1) indicate that overall, the clustering has a high degree of convergent and discriminant validity. If Γ_{norm} is near zero, then either the clusters have low levels of convergent validity, discriminant validity, or both. If Γ_{norm} is large and negative, then we have likely flipped blocks with non-blocks, and would have reason to revisit the exploratory analysis.

3.3.3 Permutation Test

The null hypothesis in our permutation test is that off-diagonal elements of A are exchangeable. Rejecting the null is evidence in favor of the hypothesized latent structure.

As described below and depicted in Figure 3.2, the overall test contrasts all the within block elements (solid) with all the between block elements (diagonal lines). The block-specific test contrasts the elements within a particular block with the elements between that block and the other blocks.



(a) Overall test

(b) Block-specific test for A_3

Figure 3.2: Elements contrasted by Γ_{norm} for a matrix A with $K = 4$ hypothesized blocks. The within block elements (solid) are contrasted with the between block elements (diagonal lines). Γ_{norm} does not include the diagonal elements in the contrast.

Overall Test

As before, let v_i be the label of the i^{th} column and row of A, and let $v = (v_1, \dots, v_p)$ be the ordered sequence of labels. For example, if A is the matrix of correlations among items on a questionnaire, then v_i would be the i^{th} item on the questionnaire. Also, let π be a permutation of the indices of v, and let $v^* = (v_1^*, \dots, v_p^*)$ be a permuted sequence of labels, where $v^*_{\pi(i)} = v_i, i = 1, \dots, p$. For example, in Section 3.2, the items in the Big Five questionnaire are labeled as $v = (v_1 = a, v_2 = b, \dots, v_{31} = z6)$, and under the hypothesized ordering shown along the rows and columns of Figure 3.1, $v^* = (v_1^* = d, v_2^* = h, v_3^* = l, \dots, v_{31}^* = z6)$.

In the permutation test, we keep the Δ matrix constant, permute the order of the labels in A, and recompute the test statistic Γ_{norm}. In keeping Δ constant, we are conditioning on the hypothesized number of blocks K and block sizes $p_k, k = 1, \dots, K$. This conditioning is an important constraint needed in the permutation test.

If we randomly sample B permutations $\pi_1, \dots \pi_B$ with replacement, then the Monte Carlo (MC) approximation to the permutation p-value is (Lehmann and Romano, 2005)

$$\tilde{p} = \frac{1}{B+1}\left[\sum_{b=1}^{B} \mathbb{1}\left(\left|\Gamma^b_{\text{norm}}\right| \geq \left|\Gamma^0_{\text{norm}}\right|\right) + 1\right],$$

where $\mathbb{1}$ is an indicator function, Γ^0_{norm} is the test statistic under the hypothesized clustering, and Γ^b_{norm} is the test statistic from the b^{th} randomly sampled permutation π_b. That is, $\tilde{p}$

represents the proportion of MC resamples with test statistics that exceed the observed test statistic under the hypothesized clustering.

Exchangeable off-diagonal elements implies a variety of matrix structures, including constant off-diagonal elements (referred to by Steiger (1980a) as equicorrelation in the case where A is the correlation matrix) and white noise. Under constant-off diagonal elements, A is of the form $A = a\mathbf{11'} + (1-a)I$ for some $a \in \mathbb{R}$ (for correlation matrices, $a \in [-1, 1]$), where $\mathbf{1}$ is a column vector of 1's and I is the identity matrix:

$$A = \begin{pmatrix} 1 & & & \text{\huge a} \\ & 1 & & \\ & & \ddots & \\ \text{\huge a} & & & 1 \end{pmatrix}.$$

More generally, under white noise we assume the off-diagonal elements $a_{ij} \sim P, i < j$ for some common distribution P. If A is a covariance or correlation matrix, then we have the additional constraint that A is positive semi-definite. If P has zero variance, we obtain constant off-diagonals.

Block-Specific Test

In addition to the overall test, we can test each block individually to see if the within-block elements are larger than the corresponding between-block elements. To this end, let $\Gamma_{\text{norm},k}$ be the same as above, except that the sum is restricted to (i, j) such that at least one of v_i, v_j is in block k. As before, we remove variance terms from the sum. To be precise, let $\mathcal{V}_k = \{v_i : a_{ii} \in A_k\}$ be the set of labels assigned to block k, and let $\mathcal{I}_k = \{(i, j) : v_i \in \mathcal{V}_k \text{ or } v_j \in \mathcal{V}_k, i < j\}$ be the set of ordered index pairs with at least one index in block k. Let $N_k = |\mathcal{V}_k|$ be the number of elements in $\mathcal{V}_k$, and let $\bar{a}_k = N_k^{-1} \sum_{(i,j)\in\mathcal{I}_k} a_{ij}$ and $\hat{\sigma}^2_{a,k} = (N_k - 1)^{-1} \sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)^2$ be the sample mean and variance of elements in the set $\{a_{ij} : (i, j) \in \mathcal{I}_k\}$, and $\bar{\delta}_k = N_k^{-1} \sum_{(i,j)\in\mathcal{I}_k} \delta_{ij}$ and $\hat{\sigma}^2_{\delta,k} = (N_k - 1)^{-1} \sum_{(i,j)\in\mathcal{I}_k} (\delta_{ij} - \bar{\delta}_k)^2$ be the sample mean and variance of elements in the set $\{\delta_{ij} : (i, j) \in \mathcal{I}_k\}$. Also, let $\hat{\sigma}^2_{a\delta,k} = (N_k - 1)^{-1} \sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)(\delta_{ij} - \bar{\delta}_k)$ be the sample covariance. Then we define

$$\Gamma_{\text{norm},k} = \frac{\sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)(\delta_{ij} - \bar{\delta}_k)}{\sqrt{\sum_{(i,j)\in\mathcal{I}_k} (a_{ij} - \bar{a}_k)^2 \sum_{(i,j)\in\mathcal{I}_k} (\delta_{ij} - \bar{\delta}_k)^2}} = \frac{\hat{\sigma}^2_{a\delta,k}}{\hat{\sigma}_{a,k}\hat{\sigma}_{\delta,k}}.$$

When testing multiple blocks, to control the family-wise error rate we follow Westfall and Young (1993) and for each permutation π_b set $\Gamma^{\max,b}_{\text{norm}} = \max_{k\in\{1,\ldots,K\}} |\Gamma^b_{\text{norm},k}|$, where $\Gamma^b_{\text{norm},k}$ is the computed statistic for block k under permutation π_b. We then compute the MC estimate of the permutation p-value for block k as

$$\tilde{p}_k = \frac{1}{B+1}\left[\sum_{b=1}^{B} \mathbb{1}\left(\Gamma_{\text{norm}}^{\max,b} \geq \left|\Gamma_{\text{norm},k}^{0}\right|\right) + 1\right].$$

3.3.4 Recommendations for Choosing Matrix A

In construct validation, the primary question concerns the magnitude of association, as opposed to the direction. Furthermore, in most questionnaires, the direction of correlation is arbitrary. For example, in the HRS big five personality questionnaire, some items are reverse coded to preserve positive correlations among items hypothesized to measure the same latent construct. Consequently, in some applications A could be set to the element-wise absolute correlations, as in the motivating example in Section 3.2. By using the absolute values of the correlations, we avoid potentially overlooking associations between items that are coded in such a way that their correlations are negative.

We use Spearman's rho so that our test is robust to non-normal data and non-linear associations. However, we speculate that other nonparametric correlation coefficients would also be reasonable, such as Kendall's tau and Goodman and Kruskal's gamma. Ultimately, we recommend that researchers use a matrix A that best measures the phenomenon of interest, which may differ across applications.

3.3.5 Convergence Rate

In data analyses, we use Monte Carlo methods to approximate the permutation p-value obtained with the estimated quantities $\boldsymbol{a}$. We denote the permutation p-value with the estimated quantities as $\hat{p}(\boldsymbol{a})$. However, we would ideally approximate the permutation p-value obtained with the true population values, which we denote as $\hat{p}(\boldsymbol{\rho})$, where $\boldsymbol{\rho}$ are the true population values. Assuming $\boldsymbol{a}$ is a consistent estimator of $\boldsymbol{\rho}$, $\boldsymbol{a} \to \boldsymbol{\rho}$ as $n \to \infty$. In this section, we address the rate at which the overall permutation p-value computed with the estimated values $\hat{p}(\boldsymbol{a})$ converges to the overall permutation p-value computed with the true values $\hat{p}(\boldsymbol{\rho})$. These results hold for the overall test.

As stated in Theorem 3.1, under fairly general conditions, the permutation p-value for the overall test has the same rate of convergence as the elements of $\boldsymbol{a}$.

Theorem 3.1. Let a_j be the sample estimates of ρ_j, $j = 1, \ldots, N$, and suppose that for all j, $|a_j - \rho_j| = O(g(n))$ with probability one for some strictly decreasing function g, such that $g(n) \to 0$ as $n \to \infty$. Also suppose that the permutation distribution $\hat{R}_N(t)$ has limiting distribution $R(t)$ such that the density of $R(t)$, denoted as $f(t)$, exists and $\sup_t f(t) < \infty$. Then for N sufficiently large, with probability one, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(g(n))$.

Furthermore, as described in Corollary 3.1, when $\boldsymbol{a}$ are Pearson's or Spearman's correlations, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(1/\sqrt{n})$. As described in Corollary 3.2, the same rate holds when using the absolute values of Pearson's or Spearman's correlations.

Corollary 3.1. Let $\boldsymbol{a}$ be Pearson's or Spearman's correlation coefficients estimated from n iid observations. Let $\tau_j^2 = \mathrm{Var}(a_j)$ and assume $\tau_j^2 < \infty$ for $j = 1, \ldots, N$. Also suppose that the permutation distribution $\hat{R}_N(t)$ has limiting distribution $R(t)$ such that the density of $R(t)$, denoted as $f(t)$, exists and $\sup_t f(t) < \infty$. Then for N sufficiently large, with probability one, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(1/\sqrt{n})$.

Corollary 3.2. Under the same conditions as Corollary 3.1, but with $\boldsymbol{a}$ and $\boldsymbol{\rho}$ replaced with absolute values of Pearson's or Spearman's correlations, we also have that with probability one $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(1/\sqrt{n})$.

For details and proofs, please see Section 3.7.

3.4 Comparison to Related Methods

3.4.1 Linear Model and t-test

To better understand and interpret Γ_{norm}, we note that because Γ_{norm} is a correlation, it is permutationally equivalent to the ordinary least squares coefficient from a simple linear regression model where the outcomes are the absolute values of the correlation coefficients $\boldsymbol{a}$ and the covariates are the indicators $\boldsymbol{\delta}$.

To see this, we write the linear model as

$$\mathbb{E}[\boldsymbol{a}] = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{\delta} \tag{3.2}$$

where $\mathbf{1}$ is an $N \times 1$ vector. The ordinary least squares estimate for (3.2) is $\hat{\beta}_1 = (\hat{\sigma}_a / \hat{\sigma}_\delta)\Gamma_{\text{norm}}$.

Let $\mathcal{W}_k = \{(i,j) : v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_k, i < j\}$ be the set of ordered index pairs for upper triangular elements such that both indices are in block k, let $N_{\text{in},k} = |\mathcal{W}_k|$ be the number of elements in $\mathcal{W}_k$, and $N_{\text{in}} = \sum_k N_{\text{in},k}$ be the total number of upper triangular within-block elements. Also, let $\mathcal{W}_{\text{out}} = \{(i,j) : (i,j) \notin \mathcal{W}_k, k = 1, \ldots, K, i < j\}$ be the set of ordered index pairs for upper triangular elements not in blocks, and $N_{\text{out}} = |\mathcal{W}_{\text{out}}|$ be the number of non-block elements. Then, because Δ is a matrix of zeros and ones, we have $\hat{\beta}_1 = \bar{a}_{\text{in}} - \bar{a}_{\text{out}}$, where $\bar{a}_{\text{in}} = N_{\text{in}}^{-1} \sum_k \sum_{(i,j) \in \mathcal{W}_k} a_{ij}$ and $\bar{a}_{\text{out}} = N_{\text{out}}^{-}1 \sum_{(i,j) \in \mathcal{W}_{\text{out}}} a_{ij}$ are the mean within-block and between-block elements, respectively. In the overall test, $\hat{\sigma}_a^2$ and $\hat{\sigma}_\delta^2$ are constant across permutations. Therefore, there is a one-to-one relationship between Γ_{norm} and $\hat{\beta}_1$, and they are permutationaly equivalent. In other words, $\hat{\beta}_1$ could be substituted for Γ_{norm} in the permutation test to obtain the same permutation p-value. When restricting to subsets of the matrix to evaluate $\Gamma_{\text{norm},k}$, $\hat{\sigma}_{a,k}^2$ is no longer constant across permutations, so $\Gamma_{\text{norm},k}$ and $\hat{\beta}_{1,k}$ are no longer permutationaly equivalent.

We also note that the t-statistic with unequal variance has potential advantages over the statistics Γ_{norm} and $\hat{\beta}_1$. In particular, the t-statistic with unequal variance controls the type I error rate in permutation tests under the null $H_0 : \bar{a}_{\text{in}} = \bar{a}_{\text{out}}$ versus $H_1 : \bar{a}_{\text{in}} \neq \bar{a}_{\text{out}}$ even if the variance of the within-block and between-block correlations are different (Chung and Romano, 2013). The t-statistic with unequal variance is given by

$$t = \frac{\bar{a}_{\text{in}} - \bar{a}_{\text{out}}}{\sqrt{\hat{\sigma}_{\text{in}}^2/N_{\text{in}} \ + \hat{\sigma}_{\text{out}}^2/N_{\text{out}}}} \tag{3.3}$$

where $\hat{\sigma}_{\text{in}}^2 = (N_{\text{in}} - 1)^{-1} \sum_k \sum_{(i,j)\in\mathcal{W}_k} (a_{ij} - \bar{a}_{\text{in}})^2$ and $\hat{\sigma}_{\text{out}}^2 = (N_{\text{out}} - 1)^{-1} \sum_{(i,j)\in\mathcal{W}_{\text{out}}} (a_{ij} - \bar{a}_{\text{out}})^2$ are the sample variances of the within-block and between-block upper triangular elements of A, respectively.

Due to the results of Chung and Romano (2013), it may be beneficial to use the studentized statistic t given by (3.3) in future work in place of Hubert's Γ, as it leads to permutation tests that are valid under a wider range of scenarios than those we examined in our simulations. However, in our simulations, the use of (3.3) in the permutation test gave nearly identical results to those obtained with Γ_{norm}.

3.4.2 Goodness of Fit (GOF) Tests

Several statistical methods used in construct validation rely on a goodness of fit (GOF) test, including CFA and pattern hypothesis tests (Steiger, 2007). Frequently, GOF tests are based on χ^2 statistics. In general terms, the null hypothesis in GOF tests is H_0: "the model fits" and the alternative is H_1: "the model does not fit." Under this framework, failure to reject the null is evidence in favor of the scientific theory. This is in contrast to the tests of matrix structure described in Section 3.3.3, for which rejection of the null is evidence in favor of the scientific theory.

Since GOF tests reverse the usual role of the null and alternative hypotheses, the interpretation of type I and II errors is also reversed. To guard against making false scientific claims, one needs to avoid accepting the null when the alternative is true – a type II error. Similarly, to increase the chances of finding evidence in favor of a scientific theory, one needs to avoid rejecting the null when the null is true. Given the analogy with statistical power,

Table 3.1: Comparison of interpretation of errors under traditional and GOF frameworks. Within each cell, the traditional interpretation is on the first line, and the GOF interpretation is on the second line in **bold**. H_0 is the null hypothesis and H_1 is the alternative hypothesis (H_0 rejected if p-value $< c$ for cutoff value c).

		Truth	
		H_0	H_1
Decision	H_0 (p-val $\geq c$)	Correct failure to reject H_0 **Type I power**	Type II error (missed opportunity) **GOF false alarm**
	H_1 (p-val $< c$)	Type I error (false alarm) **GOF missed opportunity**	Power **Correct rejection of H_0**

we refer to this as type I power. Since this is not a standard term, we define it in Definition 3.1.

Definition 3.1 (Type I power)**.** Type I power is the probability of failing to reject the null hypothesis when the null hypothesis is true: Pr(fail to reject $H_0|H_0$ true).

The reversal in GOF tests of the standard scientific interpretation of Type I and II errors may have several implications for the reliability of GOF tests in evaluating scientific hypotheses. In particular, failure to control type II errors in GOF tests could lead to higher than expected rates of false scientific claims, and low type I power would make it difficult to find evidence in favor of a scientific claim. Table 3.1 shows these differing interpretations, and proposes the terms "GOF false alarms" and "GOF missed opportunities" to describe the potential errors when conducting a GOF test. We are unaware of work aimed at controlling type II error rates in GOF tests, but several researchers have suggested ways to address low type I power. Contrary to standard statistical power, type I power decreases as sample size increases, making low Type I power a pervasive problem.

GOF Tests in Structural Equation Models (SEMs)

To address low type I power in structural equation models (SEMs), including CFA, researchers have developed alternative fit indices, most of which adjust the χ^2 GOF statistic based on the degrees of freedom, such as the comparative fit index (CFI) (Bentler, 1990) and Tucker-Lewis Index (TLI) (Tucker and Lewis, 1973). However, as shown in Section 3.5.1, the type I power of CLI and TLI decreases as sample size increases, though not as dramatically as for unadjusted χ^2 GOF statistics.

Many of the rules of thumb for interpreting fit indices have roots in the work of Hu and Bentler (1999). In particular, values above 0.95 are commonly considered to indicate acceptable fit for CFI and TLI (Hu and Bentler, 1999), though Hooper et al. (2008) notes that some researchers have suggested a cut-off value of 0.9 for CFI, and 0.8 for TLI. We show simulation results with all three cutoffs in Section 3.5.

As Barrett (2007) notes, some simulation studies, including Marsh et al. (2004), Beauducel and Wittmann (2005), Yuan (2005), and Fan and Sivo (2005), have cast doubt on the reliability of these rules of thumb for CFI and TLI. We note that the criticism of Barrett (2007) is controversial, and Steiger (2007) offers a rebuttal. Kline (2011) and Hu and Bentler (1999) offer discussions on fit statistics and indices for SEMs, and we refer the reader to these sources for details.

Pattern Hypothesis GOF Tests

As Steiger (1980b) describes, a pattern hypothesis is "any hypothesis that states that some of its elements are equal to each other and/or to specified numerical values." Using the same notation as before, let $\boldsymbol{a}$ be the $N \times 1$ vector of upper triangular elements of A. Pattern hypotheses are of the form (Steiger, 1980b)

$$H_0: \ \boldsymbol{a} = L\boldsymbol{\beta} + \boldsymbol{a}^*, \tag{3.4}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters to be estimated, $\boldsymbol{a}^*$ is $q \times 1$ vector of constants, and L is an $N \times q$ matrix of zeros and ones, with

$$L_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ element of } \boldsymbol{a} \text{ is hypothesized to equal } \beta_j \\ 0 & \text{otherwise,} \end{cases}$$

In the case where A is a covariance matrix, pattern hypothesis tests are related to the analysis of covariance structures (Bock and Bargmann, 1966).

If we set $q = 2$, then (3.4) would be a re-parameterization of (3.2). In this case, to recover (3.2) from (3.4), we would set $\boldsymbol{a}^*$ to zero and reparameterize L as $L = [\mathbf{1}, \boldsymbol{\delta}]$. This changes L from being a cell-means coding to a reference cell coding.

For the rest of this section, we assume A is the Pearson correlation matrix for underlying data $\boldsymbol{x}_l = (x_{l1}, \dots, x_{lp})^T, l = 1, \dots, n$, in which we have n observations of p variables. In particular, let $\bar{x}_i = n^{-1} \sum_{l=1}^n x_{li}$, $\hat{\sigma}_i^2 = (n-1)^{-1} \sum_{l=1}^n (x_{li} - \bar{x}_i)^2$ and $\hat{\sigma}_{ij}^2 = (n-1)^{-1} \sum_{l=1}^n (x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j)$. Then $a_{ij} = \hat{\sigma}_{ij}^2 / (\hat{\sigma}_i \hat{\sigma}_j)$. In this case, we set $\boldsymbol{r} = \boldsymbol{a}$ to use more familiar notation. If the underlying data are iid multivariate normal, then we can induce normality on the correlation coefficients by taking the Fisher r-to-z variance stabilizing transformation, denoted as $z(r)$, where (Fisher, 1921)

$$z(r) = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) = \operatorname{arctanh}(r).$$

The Fisher transformation improves the normal approximation to the distribution of the correlation coefficients, even if the underlying data are not normal, though the form of the $N \times N$ covariance matrix $\mathrm{Var}(\boldsymbol{z}(\boldsymbol{r}))$ may not be the same as for normal data (Hawkins, 1989).

Following (Steiger, 1980b), we test the null hypothesis (3.4) with the GOF χ^2 statistic

$$X_2 = (n-3)\left[\boldsymbol{z}(\boldsymbol{r}) - \boldsymbol{z}(\hat{\boldsymbol{r}}_{\mathrm{GLS}})\right]^T S_{LS}^{-1}\left[\boldsymbol{z}(\boldsymbol{r}) - \boldsymbol{z}(\hat{\boldsymbol{r}}_{\mathrm{GLS}})\right], \tag{3.5}$$

where $\hat{\boldsymbol{r}}_{\mathrm{GLS}} = L(L^T\hat{\Sigma}_{\mathrm{LS}}^{-1}L)^{-1}L^T\hat{\Sigma}_{\mathrm{LS}}^{-1}\boldsymbol{r}$, $\hat{\Sigma}_{\mathrm{LS}}$ is the covariance matrix with elements given by Steiger (1980b) with $\hat{\boldsymbol{r}}_{\mathrm{LS}} = L(L^TL)^{-1}L^T\boldsymbol{r}$ substituted for $\boldsymbol{r}$, and S_{LS} is the covariance matrix with elements also given by Steiger (1980b). Asymptotically, X_2 follows a χ^2 distribution with $N-2$ degrees of freedom (Steiger, 1980b). We note that the covariance formulas originated in the work of Pearson and Filon (1898) and are also given by Olkin and Finn (1990, 1995).

The permutation test with Γ_{norm} and the GOF χ^2 test with (3.5) are similar, but with important differences. In (3.4), and assuming $q = 2$, let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Then the permutation test is similar to obtaining a p-value for the null hypothesis $H_0 : \beta_1 = 0$, whereas (3.5) gives a p-value for the GOF null hypothesis H_0 : "the model fits." In addition, the permutation test is nonparametric and relies only on the exchangeability of off-diagonal elements, as opposed to the GOF test with (3.5), which relies on asymptotic approximations to obtain the reference distribution. The permutation test is also applicable for a variety of matrices A, whereas the asymptotic reference distribution for (3.5) is valid only for certain types of matrices.

3.5 Simulations

In this section, we simulated data under two scenarios: 1) block diagonal structure, and 2) constant off-diagonal values. For each scenario, we generated 1,000 datasets for each of three sample sizes (n = 10, 100, and 1,000). For all simulations, we used $K = 4$ blocks of sizes $p_1 = 5$, $p_2 = 7$, $p_3 = 9$, $p_4 = 11$, so that the total number of variables was $p = \sum_k p_k = 32$. In all figures, the block numbers begin in the upper left and end in the lower right, i.e., block $k = 1$ is in the top left corner, and block $k = 4$ is in the bottom right corner.

In the matrix structure testing framework, Sections 3.5.1 and 3.5.4 are under the alternative hypothesis (H_1 is true) and Sections 3.5.2 and 3.5.3 are under the null hypothesis (H_0 is true). In the GOF framework, the model is correctly specified in Section 3.5.1 (H_0 is true) and misspecified in Sections 3.5.2, 3.5.3, and 3.5.4 (H_1 is true).

We followed our recommendations in Section 3.3.4 and used absolute Spearman correlation coefficients when computing Γ_{norm}, though we acknowledge that other choices for the matrix A are possible. For the pattern hypothesis test, we used Pearson's correlation and Fisher's r-to-z transform to compute X_2, as described in Section 3.4.2. To obtain CFI and TLI, we fit CFA models with $K = 4$ latent factors and p_i items loading onto the the i^{th} factor, with p_i given above. In the CFA models, each item loaded onto exactly one factor. The tables in this section do not directly compare the results with the permutation test against those from CFI/CLI, because different types of errors are relevant for the two approaches, but we use the tables to compare the methods in Section 3.8.

3.5.1 Block Diagonal Structure

To simulate data under the scenario of a block diagonal correlation matrix, we began by generating the square root of the variance matrix $\Sigma^{1/2}$ such that variables within groups would be correlated with each other, and variables across groups would have minimal but non-zero correlations. In particular, we set $\Sigma_{ij}^{1/2} = \sum_k \mathbb{1}[v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_k] r_k + u_{ij}$, where $r_1 = 0.25$, $r_2 = 0.2$, $r_3 = 0.23$, $r_4 = 0.15$, and $u_{ij} \sim N(0, 0.01)$.

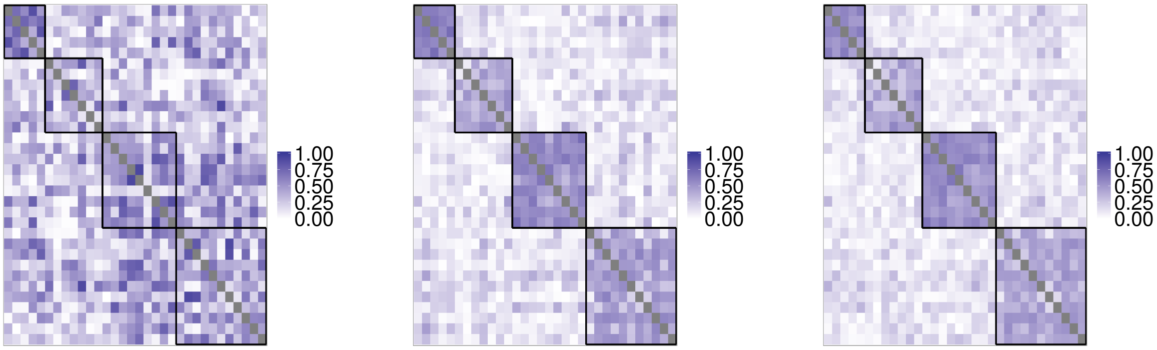
For each sample size of n = 10, 100, and 1,000, we simulated 1,000 $n \times p$ datasets, Y_t, $t = 1, \ldots, 1,000$, where

$$Y_t = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix},$$

and $\boldsymbol{y}_l = (y_{l1}, \ldots, y_{lp})^T$, $l = 1, \ldots, n$, were generated independently as $N(\mathbf{0}, \Sigma_t)$ random vectors with Σ_t generated as described above. We then created corresponding $n \times p$ datasets Z_t, t = 1,..., 1,000, of ordinal variables where for each dataset, $z_{li} = 1$ if $y_{li} < -2$, $z_{li} = 2$ if $-2 \leq y_{li} < -1$, $z_{li} = 3$ if $-1 \leq y_{li} < 0$, $z_{li} = 4$ if $0 \leq y_{li} < 1$, $z_{li} = 5$ if $1 \leq y_{li} < 2$, and $z_{li} = 6$ if $2 \leq y_{li}$.

For each dataset, we estimated Spearman's correlation matrix, which we denote as $C = C(Z)$, and conducted a permutation test with Hubert's Γ on $A = \text{abs}(C)$ where the absolute values are taken element-wise. We used $B = 10,000$ MC resamples for the permutation tests. We also computed X_2 with the Pearson correlation matrix of Z (treating the ordinal data as numeric), and fit a CFA model with the data Z (treating the data as ordinal) using the `lavaan` package (Rosseel, 2012) for R.

Figure 3.3 shows the estimated Spearman's absolute correlation matrices A from a single simulation for each sample size.



(a) $n = 10$

(b) $n = 100$

(c) $n = 1,000$

Figure 3.3: Block diagonal: estimated Spearman's correlation coefficients (absolute values) from a single simulation at each sample size.

Figure 3.4 shows the distribution of p-values from a permutation test with Γ_{norm} and $B = 10,000$ resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. As seen in Figure 3.4, the distribution of p-values from Γ_{norm} is heavily left-skewed, which is as expected under the alternative hypothesis. The p-values from the X_2 statistic quickly move from close to one to close to zero as the sample size increases, and the CFI values cluster around 0.8 to 0.9 for all sample sizes. However, as shown in Table 3.3, the distribution of CFI values shifts downward as sample size increases, though not as dramatically as for p-values from X_2.

Table 3.2 shows the power with Γ_{norm} and the permutation test under the alternative hypothesis of block diagonal structure for statistical significance levels of $\alpha = 0.01$ and 0.05. As seen in Table 3.2, the statistical power was 1 for all tests with sample sizes of 100 and 1,000.

Table 3.3 shows the percent of simulations with CFI and TLI above the cutoff value recommended by Hu and Bentler (1999) (0.95) as well as more liberal cutoff values noted by Hooper et al. (2008) (0.9 and 0.8). As can be seen in Table 3.3, the statistical power of TLI and CFI decreases as sample size increases, similar to the X_2 GOF test. Notably, the Type I power is at or near zero for both CFI and TLI for large sample sizes and cutoffs of 0.9 and 0.95.

3.5.2 Constant Off-Diagonal Correlation

For the scenario of constant off-diagonal correlation, we set $\Sigma_{t,ij} = 0.5$ if $i \neq j$ and 1 if $i = j$. We used $B = 1,000$ MC resamples for each test. The rest of the simulation is as described in Section 3.5.1.

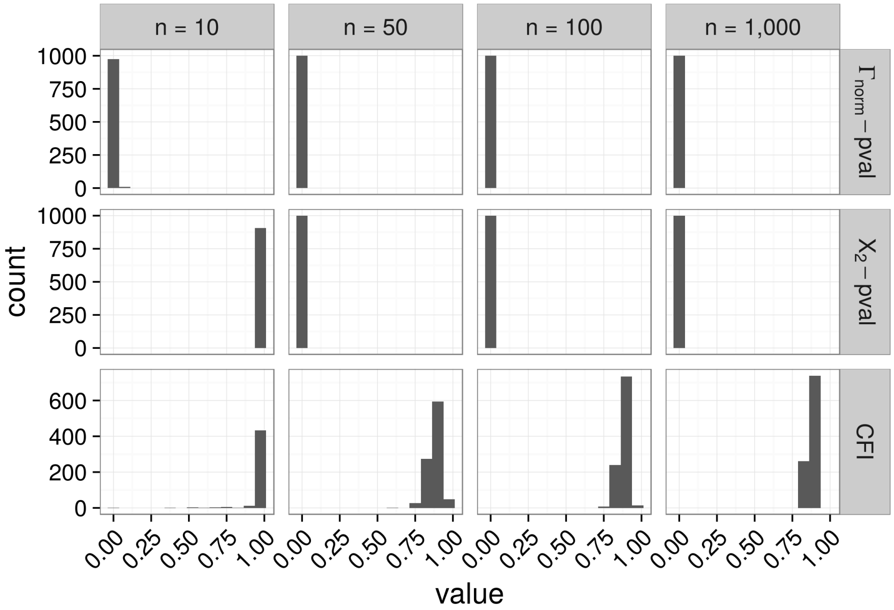


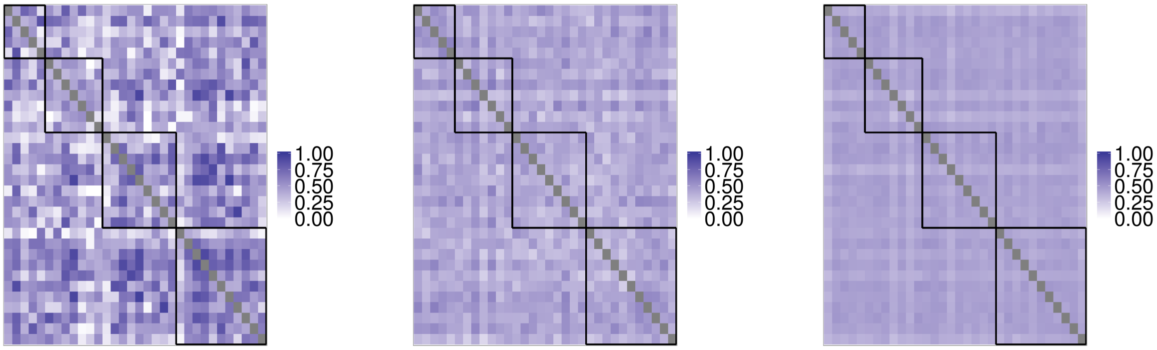
Figure 3.4: Overall test for block diagonal scenario: permutation p-values with Γ_{norm} and $B = 10,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. For each sample size we did 1,000 simulations. Results with TLI are similar to those for CFI and are not shown.

Table 3.2: Statistical power in block diagonal scenario using Γ_{norm} in a permutation test for significance levels of $\alpha = 0.01$ and 0.05. 1,000 simulations were run for each sample size.

$\alpha = 0.01$					
	Block				
n	Overall	$k = 1$	$k = 2$	$k = 3$	$k = 4$
10	0.97	0.30	0.31	0.71	0.36
50	1.0	0.93	0.96	1.0	0.98
100	1.0	0.98	0.99	1.0	1.0
1,000	1.0	1.0	1.0	1.0	1.0
$\alpha = 0.05$					
	Block				
n	Overall	$k = 1$	$k = 2$	$k = 3$	$k = 4$
10	0.98	0.44	0.49	0.81	0.52
50	1.0	0.97	0.99	1.0	1.0
100	1.0	1.0	1.0	1.0	1.0
1,000	1.0	1.0	1.0	1.0	1.0

Table 3.3: Type I power for block diagonal scenario: percent of simulation results above the cutoff value (CFI and TLI above the cutoff indicate good model fit)

		Cutoff		
Fit index	n	0.95	0.9	0.8
CFI	10	0.94	0.95	0.96
	50	0.016	0.33	0.96
	100	0.0020	0.32	0.99
	1,000	0.0	0.14	1.0
TLI	10	0.94	0.95	0.96
	50	0.0074	0.26	0.93
	100	0.0	0.22	0.97
	1,000	0.0	0.07	0.99



(a) $n = 10$

(b) $n = 100$

(c) $n = 1,000$

Figure 3.5: Constant off-diagonal: estimated Spearman's correlation coefficients (absolute values) from a single simulation at each sample size.

Figure 3.5 shows the estimated Spearman's absolute correlation matrices A from a single simulation at each sample size.

Figure 3.6 shows the distribution of p-values from a permutation test with Γ_{norm} and $B = 1,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. As seen in Figure 3.6, the distribution of p-values from Γ_{norm} is uniform, which is as expected under the null hypothesis. The p-values from the X_2 statistic move from close to one to close to zero as the sample size increases, though not as quickly as in the block diagonal scenario, and the CFI values cluster close to 1 for all sample sizes. In this scenario, the CFA model is misspecified, so large CFI values, indicating good model fit, represents a GOF false alarm (see Table 3.1).

Table 3.4 shows the type I error rates for Γ_{norm} and the permutation test for statistical significance levels of $\alpha = 0.01$ and 0.05. As seen in Table 3.4, the error rates are near their

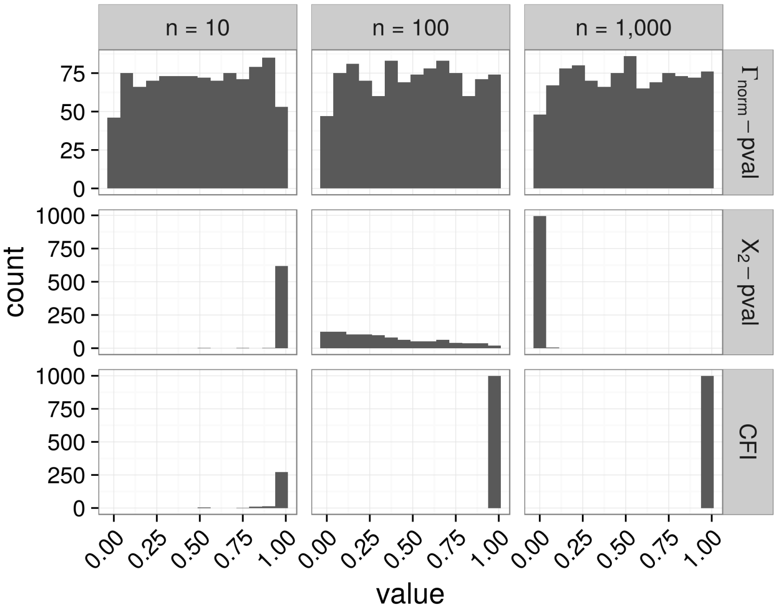


Figure 3.6: Overall test for constant off-diagonal scenario: permutation p-values with Γ_{norm} and $B = 1,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. For each sample size we did 1,000 simulations. Results with TLI are similar to those for CFI and are not shown.

Table 3.4: Type I error rates in constant off-diagonal scenario using Γ_{norm} in a permutation test for significance levels of $\alpha = 0.01$ and 0.05. 1,000 simulations were run for each sample size.

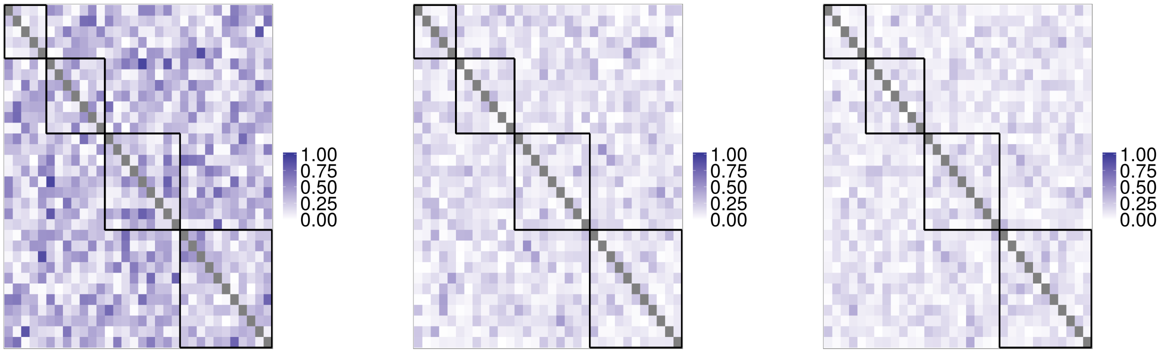
	n	Overall	Block-specific FWER
$\alpha = 0.01$			
	10	0.016	0.0092
	100	0.018	0.012
	1,000	0.011	0.011
$\alpha = 0.05$			
	10	0.062	0.060
	100	0.061	0.051
	1,000	0.057	0.047

nominal rates for all sample sizes.

Table 3.5 shows the percent of simulations with CFI and TLI above the cutoff value recommended by Hu and Bentler (1999) (0.95), as well as more liberal cutoff values noted by Hooper et al. (2008) (0.9 and 0.8). As seen in Table 3.5, The GOF false alarm rates are high for CFI and TLI in this simulation, and increase with sample size.

Table 3.5: GOF false alarm rate for constant off-diagonal scenario: percent of simulation results above the cutoff value (CFI and TLI above the cutoff indicate good model fit)

		Cutoff		
Fit index	n	0.95	0.9	0.8
CFI	10	0.89	0.92	0.98
	100	1.0	1.0	1.0
	1,000	1.0	1.0	1.0
TLI	10	0.88	0.92	0.97
	100	1.0	1.0	1.0
	1,000	1.0	1.0	1.0



(a) $n = 10$ (b) $n = 100$ (c) $n = 1,000$

Figure 3.7: White noise: estimated Spearman's correlation coefficients (absolute values) from a single simulation at each sample size.

3.5.3 White Noise

We generated Σ_t as purely white noise, where $\Sigma_{t,ij}^{1/2} \sim N(0,1)$ and $\Sigma_t = \left(\Sigma_t^{1/2}\right)^T \Sigma_t^{1/2}$. The rest of the simulation is as described in Section 3.5.1.

Figure 3.7 shows the estimated Spearman's absolute correlation matrices A from a single simulation at each sample size.

Figure 3.8 shows the distribution of p-values from a permutation test with Γ_{norm} and $B = 1,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. As seen in Figure 3.8, the distribution of p-values from Γ_{norm} is uniform, which is as expected under the null hypothesis. The p-values from the X_2 statistic move from close to one to close to zero as the sample size increases, though some simulates gave p-values close to 1 even for $n = 100$ and 1,000. The CFI values cluster close to 1 for all sample sizes. In this scenario, the CFA model is mispecified, so small CFI values for $n = 100$

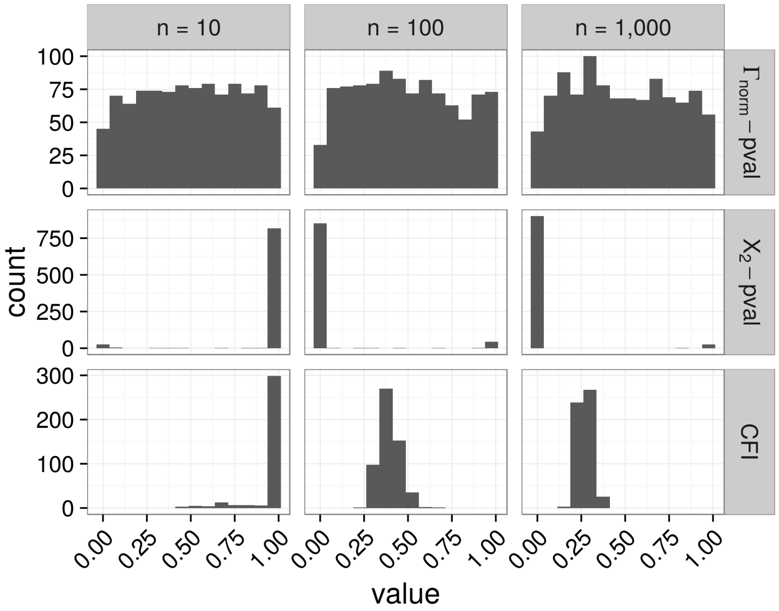


Figure 3.8: Overall test in white noise scenario: permutation p-values using Γ_{norm} and $B = 1,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. For each sample size we did 1,000 simulations. Results with TLI are similar to those for CFI and are not shown.

Table 3.6: Type I error rates for white noise scenario using Γ_{norm} in a permutation test for significance levels of $\alpha = 0.01$ and 0.05. 1,000 simulations were run for each sample size.

	n	Overall	Block-specific FWER
$\alpha = 0.01$			
	10	0.012	0.011
	100	0.010	0.0080
	1,000	0.012	0.011
$\alpha = 0.05$			
	10	0.054	0.055
	100	0.044	0.048
	1,000	0.058	0.054

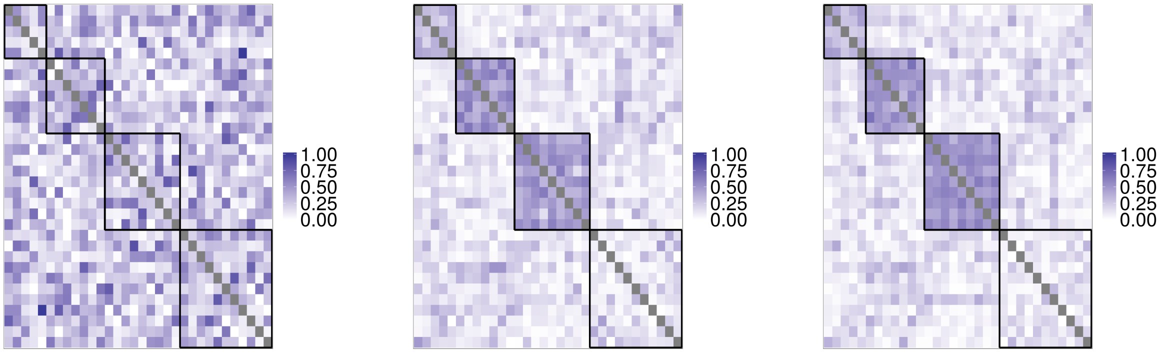
and 1,000 indicate a low GOF false alarm rate.

Table 3.6 shows the type I error rates for Γ_{norm} and the permutation test for statistical significance levels of $\alpha = 0.01$ and 0.05. As seen in Table 3.6, the error rates are near their nominal rates for all sample sizes.

Table 3.7 shows the percent of simulations with CFI and TLI above the cutoff value recommended by Hu and Bentler (1999) (0.95), as well as more liberal cutoff values noted by Hooper et al. (2008) (0.9 and 0.8). As seen in Table 3.7, The GOF false alarm is zero for

Table 3.7: GOF false alarm rate for white noise scenario: Percent of simulation results above the cutoff value (CFI and TLI above the cutoff indicate good model fit)

		Cutoff		
Fit index	n	0.95	0.9	0.8
CFI	10	0.87	0.88	0.91
	100	0.0	0.0	0.0
	1,000	0.0	0.0	0.0
TLI	10	0.87	0.88	0.90
	100	0.0	0.0	0.0
	1,000	0.0	0.0	0.0



(a) $n = 10$ (b) $n = 100$ (c) $n = 1,000$

Figure 3.9: Partial block diagonal: estimated Spearman's correlation coefficients (absolute values) from a single simulation at each sample size.

all sample sizes larger than $n = 10$.

3.5.4 Partial Block Diagonal Structure

For this scenario, we followed the simulation as described in Section 3.5.1, but set $r_4 = 0$, i.e., the last hypothesized block is not a true block.

Figure 3.9 shows the estimated Spearman's absolute correlation matrices A from a single simulation at each sample size.

Figure 3.10 shows the distribution of p-values from a permutation test with Γ_{norm} and $B = 10,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. As seen in Figure 3.10, the distribution of p-values from Γ_{norm} is left-skewed, which is as expected under the alternative hypothesis. The p-values from the X_2 statistic move from close to one to close to zero as the sample size increases, and the CFI

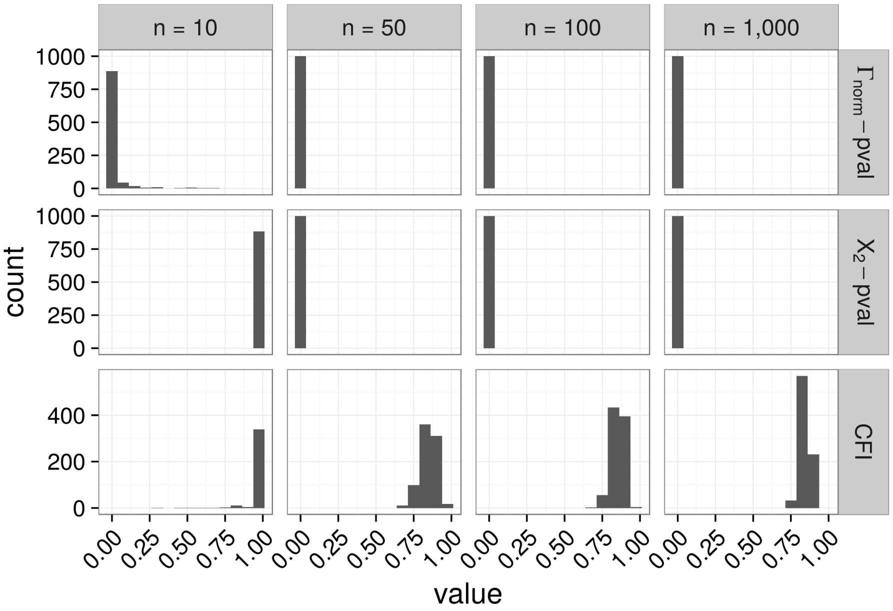


Figure 3.10: Overall test in partial block diagonal scenario: permutation p-values using Γ_{norm} and $B = 10,000$ MC resamples, p-values from the X_2 pattern hypothesis test, and CFI values from a CFA model. For each sample size we did 1,000 simulations. Results with TLI are similar to those for CFI and are not shown.

values cluster around 0.75 to 0.9 for all sample sizes.

Table 3.8 shows the power (overall and blocks 1, 2, 3) and type I error rate (block 4) using Γ_{norm} in a permutation test for statistical significance levels of $\alpha = 0.01$ and 0.05. As seen in Table 3.8, the statistical power is high for blocks 1, 2, and 3, and the type I error rate is low for block 5.

Table 3.9 shows the percent of simulations with CFI and TLI above the cutoff value recommended by Hu and Bentler (1999) (0.95), as well as more liberal cutoff values noted by Hooper et al. (2008) (0.9 and 0.8). As seen in Table 3.9, The GOF false alarm rate decreases as sample size increases. However, these results do not by themselves show that three of the four block are correctly modeled, and only the fourth is incorrectly modeled.

3.6 Application

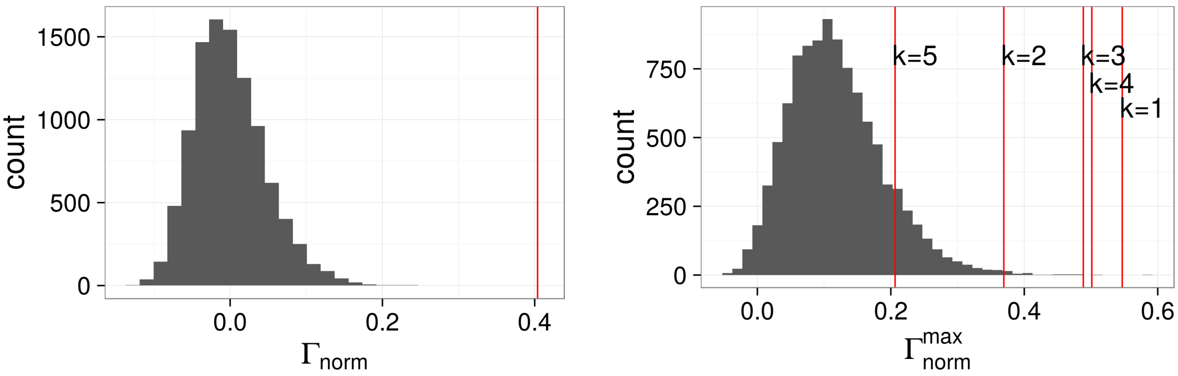
In this section, we continue our analysis of the HRS big five personality traits questionnaire described in Section 3.2.

Table 3.8: Partial block diagonal scenario: power (overall and blocks 1, 2, 3) and type I error rate (block 4) using Γ_{norm} in a permutation test for significance levels of $\alpha = 0.01$ and 0.05. 1,000 simulations were run for each sample size.

$\alpha = 0.01$					
	Block				
n	Overall	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$\alpha = 0.01$					
10	0.83	0.32	0.33	0.72	0.0010
50	1.0	0.94	0.95	1.0	0.0
100	1.0	0.99	1.0	1.0	0.0
1,000	1.0	1.0	1.0	1.0	0.0
$\alpha = 0.05$					
10	0.91	0.47	0.49	0.83	0.010
50	1.0	0.98	0.99	1.0	0.0020
100	1.0	1.0	1.0	1.0	0.0
1,000	1.0	1.0	1.0	1.0	0.0010

Table 3.9: GOF false alarm rate for the partial block diagonal scenario: percent of simulation results above the cutoff value (CFI and TLI above the cutoff indicate good model fit)

		Cutoff		
Fit index	n	0.95	0.9	0.8
CFI	10	0.93	0.94	0.98
	50	0.0038	0.13	0.83
	100	0.0	0.12	0.90
	1,000	0.0	0.018	0.93
TLI	10	0.93	0.94	0.97
	50	0.0025	0.089	0.75
	100	0.0	0.061	0.83
	1,000	0.0	0.0084	0.84



(a) Overall

(b) Block-specific

Figure 3.11: Distribution of Γ_{norm} and $\Gamma^{\max}_{\text{norm}}$ from $B = 10,000$ MC resamples with absolute Spearman's correlation coefficients for the HRS big five personality traits questionnaire. Solid red lines: observed $\Gamma^{0}_{\text{norm},k}$, $k = 1$ is neuroticism, $k = 2$ is extroversion, $k = 3$ is agreeableness, $k = 4$ is openness to experience, and $k = 5$ is conscientiousness.

Table 3.10: Results for HRS big five personality traits questionnaire with $B = 10,000$ MC resamples, controlling for family-wise error rate

Block k	Interpretation	$\Gamma^{0}_{\text{norm},k}$	p-value
–	Overall	0.40	< 0.0001
1	Neuroticism	0.55	0.0002
2	Extroversion	0.37	0.0025
3	Agreeableness	0.49	0.0003
4	Openness to experience	0.50	0.0002
5	Conscientousness	0.21	0.11

3.6.1 Permutation Test With Γ_{norm}

Figure 3.11 and Table 3.10 show the results from a permutation test with $B = 10,000$ MC resamples.

As seen in Figure 3.11 Table 3.10, the permutation test provides evidence in favor of validating the extroversion, agreeableness, neuroticism, and openness blocks, but not the conscientiousness block. However, based on Figure 3.1, the agreeableness, conscientiousness, and neuroticism blocks appear to be highly correlated with each other. In this case, we would recommend further discussions based on content area knowledge to better understand whether these blocks measure distinct underlying constructs in the HRS population. These results could potentially also help to inform future versions of the questionnaire.

3.6.2 Pattern Hypothesis Test and CFA

The p-value from the pattern hypothesis test with X_2 gave a p-value of $< 10^{-16}$, providing evidence against validating the construct. However, the large sample size in the HRS study leads to low type I power, making it unlikely that the pattern hypothesis test would provide evidence in favor of validating the big five personality traits.

Using the `lavaan` package for R (Rosseel, 2012), we fit a CFA model with five latent factors (one for each of the five constructs, with each item loading onto its hypothesized factor). This gave a CFI of 0.91 and a TLI of 0.90. Based on the recommendations of Hu and Bentler (1999) and Hooper et al. (2008), it is unclear whether these values provide evidence for or against validating the construct. If we strictly followed the 0.95 cutoff recommended by Hu and Bentler (1999), then we would not find evidence in support of the constructs. However, as we found with Γ_{norm} and the permutation test, we likely have evidence in support of validating the constructs, with the possible exception of the "conscientiousness" block.

3.7 Proof of Convergence Rate

Existing results for the large sample behavior of permutation tests focus on the relationship between the conditional permutation distribution of a statistic and the unconditional limiting distribution as the number of observations increases (e.g. see Lehmann and Romano, 2005, Section 15.2.2). In particular, let $T(x_1, \dots, x_n)$ be a test statistic of the n observations $x_1, \dots, x_n$. Also, let $\hat{R}_n(t)$ be the permutation distribution of T, and let $R(t)$ be the unconditional asymptotic distribution of T. Then most existing results study the scenario in which $\hat{R}_n \to R(t)$ as $n \to \infty$, with the goal of understanding the large sample properties of the permutation test, such as power.

In this section, we address a related but different question. In our setup, we need to account for: 1) measurement error and 2) fixed number of inputs to the test statistic. Let $a_j = \rho_j + u_j(n)$, where ρ_j is the true population quantity, a_j is our estimate, and $u_j(n)$ is measurement error, which is a function of the number of respondents n. In our proposed method, we use a statistic of the form $T(\rho_1 + u_1, \dots, \rho_N + u_N)$, where the number of correlations $N = p(p-1)/2$ is fixed by the questionnaire, which contains p items. In our setting, instead of letting $N \to \infty$, N is constant and $u_j(n) \to 0$ as $n \to \infty$ assuming a_j are consistent estimators of ρ_j. Our goal is to understand the rate at which the p-value with the estimated quantities converges to the p-value that would be obtained with the true quantities.

As before, we denote the $N \times 1$ vector of upper triangular elements of A as $\boldsymbol{a} = (a_1, a_2, \dots, a_N)^T$. Let π be a permutation, or bijection, of the columns and rows of A, let Π be the set of all such permutations π, and let $|\Pi| = p!$ be the total number of permutations in Π. Let A_π be matrix A with the rows and columns permuted according to π, and let $\boldsymbol{a}_\pi$ be the $N \times 1$ vector of upper triangular elements of A_π. Let $\Gamma_{\text{norm}}(\boldsymbol{a})$ be Hubert's Γ computed with $\boldsymbol{a}$, and let $\boldsymbol{a}_0$ be the vector of correlation coefficients under the hypothesized ordering.

In data analyses, we use Monte Carlo methods to approximate the permutation p-value obtained with the estimated quantities $\boldsymbol{a}$. We denote the permutation p-value with the estimated quantities as $\hat{p}(\boldsymbol{a}) = |\Pi|^{-1} \sum_{\pi \in \Pi} \mathbb{1}\left[|\Gamma_{\text{norm}}(\boldsymbol{a}_\pi)| \geq |\Gamma_{\text{norm}}(\boldsymbol{a}_0)|\right]$. However, we would ideally approximate the permutation p-value obtained with the true population quantities, which we denote as $\hat{p}(\boldsymbol{\rho}) = |\Pi|^{-1} \sum_{\pi \in \Pi} \mathbb{1}\left[|\Gamma_{\text{norm}}(\boldsymbol{\rho}_\pi)| \geq |\Gamma_{\text{norm}}(\boldsymbol{\rho}_0)|\right]$. Fortunately, under general conditions specified in Theorem 3.1, if $|a_j - \rho_j| = O(g(n))$ for $j = 1, \ldots, N$, then we also have $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(g(n))$. In other words, the rate of convergence for the permutation p-value is the same as the rate of convergence of the underlying elements of $\boldsymbol{a}$. As shown in Corollary 3.1, when $\boldsymbol{a}$ are Pearson's or Spearman's correlation coefficients, we have $g(n) = O(1/\sqrt{n})$. As shown in Corollary 3.2, the same rate of convergence holds when using the absolute values of Pearson's or Spearman's correlation coefficients.

As shown in Section 3.4.1, $\Gamma_{\text{norm}}(\boldsymbol{a}) = (\hat{\sigma}_\delta/\hat{\sigma}_a)(\bar{a}_{\text{in}} - \bar{a}_{\text{out}})$, where $\bar{a}_{\text{in}}$ is the mean of the within block elements and $\bar{a}_{\text{out}}$ is the mean of the between block elements. Since $\hat{\sigma}_\delta$ and $\hat{\sigma}_a$ are constant conditional on the data, this shows that $\Gamma_{\text{norm}}(\boldsymbol{a})$ is permutationally equivalent to the difference in means, which we denote by $D(\boldsymbol{a}) = \bar{a}_{\text{in}} - \bar{a}_{\text{out}}$. Similarly, we denote the difference in means of the true population quantities as $D(\boldsymbol{\rho}) = \bar{\rho}_{\text{in}} - \bar{\rho}_{\text{out}}$.

In this section, we work with D instead of Γ_{norm}, because the former simplifies the derivations. Since D and Γ_{norm} are permutationally equivalent, they produce identical permutation p-values. Consequently, the convergence rate of the permutation p-value must be the same for D as for Γ_{norm}.

Before focusing on our primary interest, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})|$, we state an inequality in Lemma 3.1 that we will use to prove our main result in Theorem 3.1.

Lemma 3.1. *Suppose that with probability one, $|a_j - \rho_j| \leq \epsilon_j(n) < \infty$, $j = 1, \ldots, N$. For each $n \in \mathbb{N}$, let $\epsilon_{\max}(n) = \max_j \epsilon_j(n)$. Then with probability one, $|D(\boldsymbol{a}) - D(\boldsymbol{\rho})| \leq 2\epsilon_{\max}(n)$.*

Proof of Lemma 3.1. From these assumptions, it follows that with probability one,

$$\begin{aligned}
|D(\boldsymbol{a}) - D(\boldsymbol{\rho})| &= |\bar{a}_{\text{in}} - \bar{a}_{\text{out}} - (\bar{\rho}_{\text{in}} - \bar{\rho}_{\text{out}})| \\
&= |\bar{a}_{\text{in}} - \bar{\rho}_{\text{in}} + \bar{\rho}_{\text{out}} - \bar{a}_{\text{out}}| \\
&\leq |\bar{a}_{\text{in}} - \bar{\rho}_{\text{in}}| + |\bar{\rho}_{\text{out}} - \bar{a}_{\text{out}}| \\
&\leq 2\epsilon_{\max}(n)
\end{aligned}$$

□

We now turn to our primary interest, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})|$. To that end, for fixed $\epsilon > 0$, let $B_\epsilon = (|D(\boldsymbol{\rho}_0)| - \epsilon, |D(\boldsymbol{\rho}_0)| + \epsilon)$ be the ϵ-ball centered around $|D(\boldsymbol{\rho}_0)|$. Also, let

$$\begin{aligned}
\Pi_B(\epsilon) &= \{\pi \in \Pi : |D(\boldsymbol{\rho}_\pi)| \in B_\epsilon\} \\
\Pi_{\bar{B}}(\epsilon) &= \{\pi \in \Pi : |D(\boldsymbol{\rho}_\pi)| \notin B_\epsilon\}.
\end{aligned}$$

Note that for each ϵ, $\Pi_B(\epsilon)$ and $\Pi_{\bar{B}}(\epsilon)$ partition Π, i.e. $\Pi = \Pi_B(\epsilon) \cup \Pi_{\bar{B}}(\epsilon)$ and $\Pi_B(\epsilon) \cap \Pi_{\bar{B}}(\epsilon) = \emptyset$.

For fixed ϵ we have

$$\begin{aligned}
|\Pi|\,|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| &= \left|\sum_{\pi\in\Pi} \mathbb{1}\left(|D(\boldsymbol{a}_\pi)| \geq |D(\boldsymbol{a}_0)|\right) - \sum_{\pi\in\Pi} \mathbb{1}\left(|D(\boldsymbol{\rho}_\pi)| \geq |D(\boldsymbol{\rho}_0)|\right)\right| \\
&= \left|\sum_{\pi\in\Pi} \left\{\mathbb{1}\left(|D(\boldsymbol{a}_\pi)| \geq |D(\boldsymbol{a}_0)|\right) - \mathbb{1}\left(|D(\boldsymbol{\rho}_\pi)| \geq |D(\boldsymbol{\rho}_0)|\right)\right\}\right| \\
&\leq \underbrace{\left|\sum_{\pi\in\Pi_B(2\epsilon)} \left\{\mathbb{1}\left(|D(\boldsymbol{a}_\pi)| \geq |D(\boldsymbol{a}_0)|\right) - \mathbb{1}\left(|D(\boldsymbol{\rho}_\pi)| \geq |D(\boldsymbol{\rho}_0)|\right)\right\}\right|}_{C_B} && (3.6) \\
&+ \underbrace{\left|\sum_{\pi\in\Pi_{\bar{B}}(2\epsilon)} \left\{\mathbb{1}\left(|D(\boldsymbol{a}_\pi)| \geq |D(\boldsymbol{a}_0)|\right) - \mathbb{1}\left(|D(\boldsymbol{\rho}_\pi)| \geq |D(\boldsymbol{\rho}_0)|\right)\right\}\right|}_{C_{\bar{B}}}. && (3.7)
\end{aligned}$$

We further partition $\Pi_{\bar{B}}(2\epsilon)$ into

$$\begin{aligned}
\Pi^L_{\bar{B}}(2\epsilon) &= \{\pi \in \Pi_{\bar{B}}(2\epsilon) : |D(\boldsymbol{\rho}_\pi)| < |D(\boldsymbol{\rho}_0)|\} \\
\Pi^R_{\bar{B}}(2\epsilon) &= \{\pi \in \Pi_{\bar{B}}(2\epsilon) : |D(\boldsymbol{\rho}_\pi)| > |D(\boldsymbol{\rho}_0)|\}.
\end{aligned}$$

We proceed by bounding C_B (3.6) in Lemma 3.2 and $C_{\bar{B}}$ (3.7) in Lemma 3.3. We then combine these bounds with the ϵ given by Lemma 3.1 to prove our main result in Theorem 3.1.

Let $\hat{R}_N(t)$ be the permutation distribution of $|D(\boldsymbol{\rho})|$, and suppose that for N sufficiently large, $\hat{R}_N(t) \approx R(t)$, where R is a limiting distribution with a well-behaved density as specified in Lemma 3.2.

Lemma 3.2. *Let $\hat{R}_N(t)$ be the permutation distribution of $|D(\boldsymbol{\rho})|$. Suppose $\hat{R}_N(t) \approx R(t)$ for N sufficiently large, where $R(t)$ has density $f(t)$ such that $M = \sup_t f(t) < \infty$. Also let $\epsilon = \epsilon(n)$ and suppose $\epsilon(n) = O(g(n))$ for some strictly decreasing function g such that $g(n) \to 0$ as $n \to \infty$. Then for N sufficiently large, $C_B = O(g(n))$. In particular, $C_B/|\Pi| \leq 4M\epsilon(n)$.*

Proof of Lemma 3.2. In the following, we use the convention that $f(t) = 0$ for $t \notin \mathrm{supp}(f)$, where $\mathrm{supp}(f)$ is the support of f. We have

$$\begin{aligned}
\frac{C_B}{|\Pi|} &\leq \frac{|\Pi_B\left(2\epsilon\right)|}{|\Pi|} \\
&= \hat{R}_N\left(|D(\boldsymbol{\rho}_0)| + 2\epsilon\right) - \hat{R}_N\left(|D(\boldsymbol{\rho}_0)| - 2\epsilon\right) \\
&\approx R\left(|D(\boldsymbol{\rho}_0)| + 2\epsilon\right) - R\left(|D(\boldsymbol{\rho}_0)| - 2\epsilon\right) && \text{(for large } N\text{)} \\
&= \int_{|D(\boldsymbol{\rho}_0)|-2\epsilon}^{|D(\boldsymbol{\rho}_0)|+2\epsilon} f(s)ds \\
&\leq 4M\epsilon && (3.8)
\end{aligned}$$

Since $|\Pi|$ is a constant and $\epsilon = \epsilon(n) = O(g(n))$, this shows that $C_B = O(g(n))$. □

We note that the constraint on R in Lemma 3.2 precludes distributions that concentrate on sets of measure zero, such as the dirac delta function. In other words, the limiting distribution cannot be degenerate. We also note that in Lemma 3.2, we could set $\epsilon(n) = 2\epsilon_{\max}(n)$, where $\epsilon_{\max}(n)$ is given in Lemma 3.1. In this case, (3.8) becomes $8M\epsilon_{\max}(n)$.

The proof of Lemma 3.2 assumes that $N = p(p-1)/2$ is sufficiently large for the approximation $\hat{R}_N(t) \approx R(t)$ to hold, i.e. that the matrix A has many elements. In practice, N is determined by the number of items p on the questionnaire. Furthermore, since the total number of permutations $p!$ grows very quickly, we anticipate that $p > 10$ ($N > 45$) is sufficient in most applications for the permutation distribution to be approximated well by a limiting distribution for which the density exists and is bounded above. The bound on C_B is then a function of the number of subjects n who reply to the questionnaire.

We now turn to the $C_{\bar{B}}$ term (3.7).

Lemma 3.3. *For fixed $\epsilon > 0$, suppose that $\Pr(|D(\boldsymbol{a}) - D(\boldsymbol{\rho})| \leq \epsilon) = 1$. Then $C_{\bar{B}} = 0$ almost surely, i.e. $\Pr(C_{\bar{B}} = 0) = 1$.*

Proof of Lemma 3.3.

$$
\begin{aligned}
\Pr(C_{\bar{B}} = 0) &\geq \Pr\Bigg(|D(\boldsymbol{a}_0)| \in B_\epsilon, \bigcap_{\pi \in \Pi^L_{\bar{B}}(2\epsilon)} |D(\boldsymbol{a}_\pi)| \leq |D(\boldsymbol{\rho}_0)| - \epsilon, \\
&\qquad \bigcap_{\pi \in \Pi^R_{\bar{B}}(2\epsilon)} |D(\boldsymbol{a}_\pi)| \geq |D(\boldsymbol{\rho}_0)| + \epsilon \Bigg) \qquad (3.9) \\
&\geq \Pr\left(|D(\boldsymbol{a}_0)| \in B_\epsilon\right) + \sum_{\pi \in \Pi^L_{\bar{B}}(2\epsilon)} \Pr\left(|D(\boldsymbol{a}_\pi)| \leq |D(\boldsymbol{\rho}_0)| - \epsilon\right) \\
&\quad + \sum_{\pi \in \Pi^R_{\bar{B}}(2\epsilon)} \Pr\left(|D(\boldsymbol{a}_\pi)| \geq |D(\boldsymbol{\rho}_0)| + \epsilon\right) - |\Pi_{\bar{B}}(2\epsilon)| \qquad (3.10) \\
&= 1 \qquad (3.11)
\end{aligned}
$$

To see why the inequality in (3.9) holds, consider the set $\mathcal{D}^L = \left\{|D(\boldsymbol{\rho}_\pi)| : \pi \in \Pi^L_{\bar{B}}(2\epsilon)\right\}$ and let $\pi_{\max} = \arg\max_\pi \left\{|D(\boldsymbol{\rho}_\pi)| \in \mathcal{D}^L\right\}$. The right-hand side of (3.9) requires that $|D(\boldsymbol{a}_{\pi_{\max}})| \leq |D(\boldsymbol{\rho}_0)| - \epsilon$ and $|D(\boldsymbol{a}_0)| > |D(\boldsymbol{\rho}_0)| - \epsilon$. Furthermore, by the definition of $\Pi^L_{\bar{B}}(2\epsilon)$, we have $|D(\boldsymbol{\rho}_{\pi_{\max}})| \leq |D(\boldsymbol{\rho}_0)| - 2\epsilon$. Therefore, we have both $|D(\boldsymbol{a}_{\pi_{\max}})| < |D(\boldsymbol{a}_0)|$ and $|D(\boldsymbol{\rho}_{\pi_{\max}})| < |D(\boldsymbol{\rho}_0)|$. Consequently, $\mathbb{1}\left\{|D(\boldsymbol{a}_{\pi_{\max}})| \geq |D(\boldsymbol{a}_0)|\right\} - \mathbb{1}\left\{|D(\boldsymbol{\rho}_{\pi_{\max}})| \geq |D(\boldsymbol{\rho}_0)|\right\} = 0$. The same argument applies to all elements in $\mathcal{D}^L$. Similarly, let $\mathcal{D}^R = \left\{|D(\boldsymbol{\rho}_\pi)| : \pi \in \Pi^R_{\bar{B}}(2\epsilon)\right\}$ and $\pi_{\min} = \arg\min_\pi \left\{|D(\boldsymbol{\rho}_\pi)| \in \mathcal{D}^R\right\}$. Then an analogous argument as above applies to the elements in $\mathcal{D}^R$.

Line (3.10) is a direct application of Bonferroni's inequality (see Casella and Berger, 2002, p. 13). Line (3.11) holds, because all probabilities in (3.10) are equal to one. To see this, note that by assumption, for all $\pi \in \Pi^L_{\bar{B}}$, with probability one

$$
\begin{aligned}
|D(\boldsymbol{a}_\pi)| &< |D(\boldsymbol{\rho}_\pi)| + \epsilon \\
&\leq |D(\boldsymbol{\rho}_{\pi_{\max}})| + \epsilon \\
&\leq |D(\boldsymbol{\rho}_0)| - \epsilon
\end{aligned}
$$

with analogous results for $\pi \in \Pi^R_{\bar{B}}$. By assumption, we also have $|D(\boldsymbol{a}_0)| \in B_\epsilon$ with probability one. This completes the proof. □

We note that in Lemma 3.3, we could take ϵ to be any small positive value. In particular, for each $n \in \mathbb{N}$, we could set $\epsilon = 2\epsilon_{\max}(n)$, where $\epsilon_{\max}(n)$ is given in Lemma 3.1.

We now state our main result in Theorem 3.1 followed by Corollaries 3.1 and 3.2, which focus on the special case of Pearson's and Spearman's correlations.

Theorem 3.1. *Let a_j be the sample estimates of ρ_j, $j = 1, \ldots, N$, and suppose that for all j, $|a_j - \rho_j| = O(g(n))$ with probability one for some strictly decreasing function g, such that*

$g(n) \to 0$ as $n \to \infty$. Also suppose that the permutation distribution $\hat{R}_N(t)$ has limiting distribution $R(t)$ such that the density of $R(t)$, denoted as $f(t)$, exists and $\sup_t f(t) < \infty$. Then for N sufficiently large, with probability one, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(g(n))$.

Proof of Theorem 3.1. From (3.6) and (3.7), we have $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| \leq |\Pi|^{-1}\left(C_B + C_{\bar{B}}\right)$. By assumption, with probability one, $|a_j - \rho_j| \leq \epsilon_j(n)$ where $\epsilon_j(n) = O(g(n))$, $j = 1, \ldots, N$. For each $n \in \mathbb{N}$, let $\epsilon_{\max}(n) = \max_j \epsilon_j(n)$. Then by Lemma 3.1, $|D(\boldsymbol{a}) - D(\boldsymbol{\rho})| \leq 2\epsilon_{\max}(n)$ with probability one. Therefore, by setting $\epsilon = 2\epsilon_{\max}(n)$, Lemma 3.3 gives $\Pr(C_{\bar{B}} = 0) = 1$, and Lemma 3.2 gives $C_B = O(g(n))$. It follows that with probability one $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(g(n))$. □

Corollary 3.1. *Let $\boldsymbol{a}$ be Pearson's or Spearman's correlation coefficients estimated from n iid observations. Let $\tau_j^2 = Var(a_j)$ and assume $\tau_j^2 < \infty$ for $j = 1, \ldots, N$. Also suppose that the permutation distribution $\hat{R}_N(t)$ has limiting distribution $R(t)$ such that the density of $R(t)$, denoted as $f(t)$, exists and $\sup_t f(t) < \infty$. Then for N sufficiently large, with probability one, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(1/\sqrt{n})$.*

Proof of Corollary 3.1. Suppose that $\boldsymbol{a}$ are Pearson's correlation coefficients. Then under these assumptions and by the central limit theorem and delta method, $\sqrt{n}(a_j - \rho_j)$ is asymptotically normal for $j = 1, \ldots, N$ (Lehmann and Romano, 2005, p. 438). Then for n sufficiently large and $\epsilon > 0$,

$$\begin{aligned}
\Pr\left(|a_j - \rho_j| > \epsilon\right) &= \Pr\left(\frac{\sqrt{n}|a_j - \rho_j|}{\tau_j} > \frac{\sqrt{n}\epsilon}{\tau_j}\right) \\
&\approx \Pr\left(|Z| > \sqrt{n}\epsilon/\tau_j\right) \qquad\qquad (Z \sim N(0,1)) \\
&= 2\left[1 - \Phi\left(\frac{\sqrt{n}\epsilon}{\tau_j}\right)\right],
\end{aligned}$$

where Φ is the standard normal CDF. Setting $\delta = 2\left(1 - \Phi(\sqrt{n}\epsilon/\tau_j)\right)$ and solving for $\delta \in [0, 1]$, we get that with probability $1 - \delta$, $|a_j - \rho_j| \leq \tau_j \Phi^{-1}\left(1 - \delta/(2N)\right)/\sqrt{n}$. Setting $\delta = 0$, we get that with probability one,

$$|a_j - \rho_j| \leq \tau_j \Phi^{-1}(1)/\sqrt{n}. \tag{3.12}$$

Hence with probability one, $|a_j - \rho_j| = O(1/\sqrt{n})$, $j = 1, \ldots, N$. Then by Theorem 3.1, with probability one, $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(1/\sqrt{n})$.

Since Spearman's correlation is Pearson's correlation of the ranks, the above argument carries over to Spearman's correlation. □

Corollary 3.2. *Under the same conditions as Corollary 3.1, but with $\boldsymbol{a}$ and $\boldsymbol{\rho}$ replaced with absolute values of Pearson's or Spearman's correlations, we also have that with probability one $|\hat{p}(\boldsymbol{a}) - \hat{p}(\boldsymbol{\rho})| = O(1/\sqrt{n})$.*

Proof of Corollary 3.2. Let $\boldsymbol{a}_{\text{abs}}$ and $\boldsymbol{\rho}_{\text{abs}}$ be $N \times 1$ vectors of the absolute values of the estimated correlation coefficients and the true correlations, respectively. We have $|a_{\text{abs},j} - \rho_{\text{abs},j}| = \big||a_j| - |\rho_j|\big| \leq |a_j - \rho_j| \leq \tau_j \Phi^{-1}(1)/\sqrt{n}$, where the last inequality follows from (3.12) in the proof of Corollary 3.1. Hence with probability one, $|a_{\text{abs},j} - \rho_{\text{abs},j}| = O(1/\sqrt{n})$, $j = 1, \ldots, N$. Then by Theorem 3.1, with probability one, $|\hat{p}(\boldsymbol{a}_{\text{abs}}) - \hat{p}(\boldsymbol{\rho}_{\text{abs}})| = O(1/\sqrt{n})$. □

We believe that the regularity conditions in these proofs are sufficiently general to be applicable to most data encountered in practice. However, in future work, we plan to investigate alternative proofs that relax the constraint that $\hat{R}_N(t)$ has a limiting distribution $R(t)$. We also plan to extend these results to the block-specific tests, and provide corollaries for other common correlations.

3.8 Discussion

Directly testing hypotheses concerning the structure of A with the methods described in this chapter, as opposed to implicitly testing the structure of A through a model-based approach, such as CFA, has both advantages and disadvantages. The tests of matrix structure presented in this chapter allow for greater variety of matrices (e.g. A can be correlations or absolute correlations, in addition to covariances), have null hypotheses that are aligned with the scientific question, and make it possible to test each block separately in addition to the overall test. These nonparametric tests also address the challenge in CFA of determining whether poor fit (small GOF index) is due to incorrect assumptions on the distributions of the random variables (secondary interest), or an inaccurate attribution of test questions to latent variables (primary interest).

However, CFA, and more generally, SEMs, allow for more flexible latent variable structures, and can be used in subsequent analyses to study associations between latent variables and additional covariates. With this in mind, we view methods for directly testing the structure of A as being useful either by themselves when appropriate, or to check the robustness of model-based approaches.

The simulation results suggest that the permutation test with Γ_{norm} maintains high power while controlling the type I error rate. In particular, the p-values are uniformly distributed under the null hypothesis, so type I error rates can be estimated theoretically. In contrast, CLI and TLI behave differently depending on the scenario, so it is not possible to theoretically estimate error rates, such as the GOF false alarm rate (type II errors, see Table (3.1)). This has the consequence that the known behavior of CLI and TLI are restricted to simulation results, and may not generalize to other settings.

In this chapter, we focused on scenarios in which each observed variable loads onto no more than one latent factor, which implies a block diagonal structure in the covariance and correlation matrices. This constraint is commonly imposed on CFA models as well. However, the Γ_{norm} statistic and permutation test are not restricted to these scenarios, and in future work it could be beneficial to study the performance of these methods when testing more general matrix structures.

In future work, it may also be beneficial to investigate the use of the studentized difference in means (3.3) in place of Γ_{norm} in the permutation test. In our simulations, (3.3) gave nearly identical results as Γ_{norm} (results not shown), but due to the results of (Chung and Romano, 2013), we speculate that there may be scenarios in which (3.3) controls the type I error rate better than Γ_{norm}.

Finally, we note that we view these tests as single pieces of information that can be used in a larger decision-making process. This approach is consistent with the American Statistical Association's statement on p-values (Wasserstein and Lazar, 2016).

3.9 R Package and Code

We have implemented the methods described in this chapter in an R package `matrixTest` available at `https://github.com/bdsegal/matrixTest`. Code for reproducing all analyses in this chapter is available at `https://github.com/bdsegal/code-for-matrixTest-paper`.

Part II

Semiparametric Regression

Chapter 4

P-splines with an ℓ_1 Penalty for Repeated Measures

4.1 Introduction

Many nonparametric regression methods, including smoothing splines and regression splines, obtain point estimates by minimizing a penalized negative log-likelihood function of the form $l_{\text{pen}} = -l(\boldsymbol{\beta}) + P(\boldsymbol{\beta})$, where l is a log-likelihood, P is a penalty term, and $\boldsymbol{\beta}$ are the coefficients to be estimated. Typically, quadratic (ℓ_2 norm) penalties are used, which lead to straightforward computation and inference. In particular, ℓ_2 penalties typically lead to ridge estimators, which have both closed form solutions and are linear smoothers. The ℓ_2 penalty also has connections to mixed models, which allows the smoothing parameters to be estimated as variance components (Green, 1987, Speed, 1991, Wang, 1998, Zhang et al., 1998).

However, nonparametric regression methods that use an ℓ_1-type penalty, such as ℓ_1 trend filtering (Kim et al., 2009) and locally adaptive regression splines (Mammen et al., 1997), are better able to adapt to local differences in smoothness and achieve the minimax rate of convergence for weakly differentiable functions of bounded variation (Tibshirani, 2014a), whereas ℓ_2 penalized methods do not (Donoho and Johnstone, 1988). The trade-off is that ℓ_1 penalties generally lead to more difficult computation and inference because the objective function is convex but non-differentiable, and the fit is no longer a linear smoother.

In this chapter, we propose P-splines with an ℓ_1 penalty as a framework for generalizing ℓ_1 trend filtering within the context of repeated measures data and semiparametric (additive) models (Hastie and Tibshirani, 1986). In Section 4.2, we discuss connections between P-splines and ℓ_1 trend filtering which motivate the methodological development. In Section

4.3, we present our proposed model, and in Section 4.4, we discuss related work. In Section 4.5, we propose an estimation procedure using the alternating direction method of multipliers (ADMM) (see Boyd et al., 2011) and cross validation (CV). In Section 4.6, we derive the degrees of freedom and propose computationally fast approximations, and in Section 4.7, we develop approximate confidence bands based on a ridge approximation to the ℓ_1 fit. In Section 4.8, we study our method through simulations and evaluate its performance in fitting non-smooth functions. In section 4.9, we demonstrate our method in an application to electrodermal activity data collected as part of a stress study. We close with a discussion in Section 4.10.

4.2 P-splines and ℓ_1 Trend Filtering

In this section, we give brief background on P-splines and ℓ_1 trend filtering, and show the relation between them when the data are independent and identically distributed (iid) normal.

P-splines (Eilers and Marx, 1996) are penalized B-splines (see De Boor, 2001). B-splines are flexible bases that are notable in part because they have compact support, which leads to banded design matrices and faster computation. This compact support can be seen in Figure 4.1, which shows eight evenly spaced first degree and third degree B-spline bases on $[0, 1]$. We can define an order M (degree $M-1$) B-spline basis with $j = 1, \ldots, p$ basis functions recursively as (De Boor, 2001)

$$\begin{aligned} \phi_j^m(x) &= \frac{x - t_j}{t_{j+m-1} - t_j}\phi_j^{m-1}(x) + \frac{t_{j+m} - x}{t_{j+m} - t_{j+1}}\phi_{j+1}^{m-1}(x), \quad j = 1, \ldots, 2M + c - m, \quad 1 < m \leq M \\ \phi_j^1(x) &= \begin{cases} 1 & t_j \leq x < t_{j+1} \\ 0 & \text{otherwise} \end{cases}, \qquad\qquad j = 1, \ldots, 2M + c - 1 \end{aligned}$$

where t_j are the knots, division by zero is taken to be zero, and c is the number of internal knots. For order M B-splines defined on the interval $[a, b]$, in order to obtain $j = 1, \ldots, p$ basis functions, we set $2M$ boundary knots (M knots on each side) and $c = p - M$ interior knots. In general, one can set $t_1 \leq t_2 \leq \cdots \leq t_M = a < t_{M+1} < \cdots < t_{M+c} < b = t_{M+c+1} \leq t_{M+c+2} \leq \cdots \leq t_{2M+c}$. In order to ensure continuity at the boundaries, we set $t_1 < t_2 < \cdots < t_{M-1} < t_M = a$ and $b = t_{M+c+1} < t_{M+c+2} < \cdots < t_{2M+c}$. We also use equally spaced interior knots, which is important for the P-spline penalty, and drop the superscript on ϕ designating order when the order does not matter or is stated in the text.

B-spline bases can be used to fit nonparametric models of the form $y(x) = f(x) + \epsilon(x)$, where $y(x)$ is the outcome y at point x, $f(x)$ is the mean response function at x, and $\epsilon(x)$

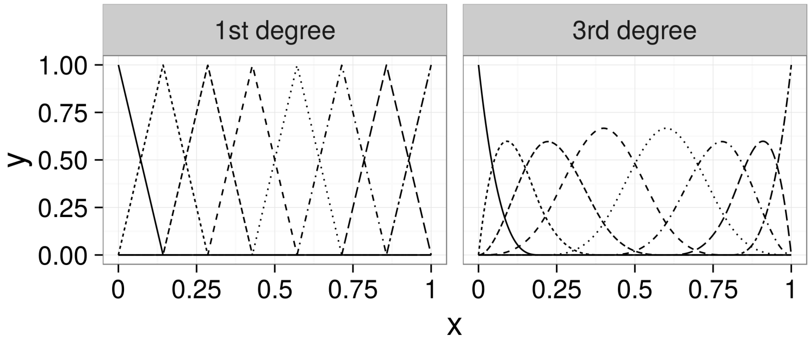


Figure 4.1: Eight evenly spaced B-spline bases on $[0, 1]$

is the error at x. To that end, let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be an $n \times 1$ vector of outcomes and $\boldsymbol{x} = (x_1 \ldots, x_n)^T$ be a corresponding $n \times 1$ vector of covariates. Also, let $\phi_1, \ldots, \phi_p$ be B-spline basis functions and let F be an $n \times p$ design matrix such that $F_{ij} = \phi_j(x_i)$, i.e., the j^{th} column of F is the j^{th} basis function evaluated at $x_1, \ldots, x_n$. Equivalently, the i^{th} row of F is the i^{th} data point evaluated by $\phi_1, \ldots, \phi_p$. For iid normal $\boldsymbol{y}$, a simple linear P-spline model with the standard ℓ_2 penalty can be written as

$$\hat{\beta}_0, \hat{\boldsymbol{\beta}} = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \beta_0 \mathbf{1} - F\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|D^{(k+1)}\boldsymbol{\beta}\|_2^2, \tag{4.1}$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameter estimates, $\mathbf{1}$ is an $n \times 1$ vector with each element equal to 1, $\lambda > 0$ is a smoothing parameter, and $D^{(k+1)} \in \mathbb{R}^{(p-k-1)\times p}$ is the $k+1$ order finite difference matrix. For example, for $k = 1$

$$D^{(2)} = \begin{bmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(p-2)\times p} \tag{4.2}$$

In general, as described by Tibshirani (2014a), $D^{(k+1)} = D^{(1)}D^{(k)}$ where $D^{(1)}$ is the $(p-k-1) \times (p-k)$ upper left matrix of:

$$D^{(1)} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(p-1)\times p}. \tag{4.3}$$

Our proposed model builds on one in which the ℓ_2 penalty in (4.1) is replaced with an ℓ_1 penalty:

$$\hat{\beta}_0, \hat{\boldsymbol{\beta}} = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \beta_0 \mathbf{1} - F\boldsymbol{\beta}\|_2^2 + \lambda \|D^{(k+1)}\boldsymbol{\beta}\|_1. \tag{4.4}$$

ℓ_1 trend filtering is similar to (4.4), and is based on the following objective function, in which it is assumed that $x_1 < x_2 < \cdots < x_n$ are unique and equally spaced:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + \lambda \|D^{(k+1)}\boldsymbol{\beta}\|_1. \tag{4.5}$$

Apart from requiring unique and equally spaced observations, (4.5) differs from (4.4) in that (4.5) has one parameter per data point, no intercept, and the design matrix is the identity matrix. $D^{(k+1)}$ is also resized appropriately by replacing p with n in the dimensions of (4.2) and (4.3). However, under certain conditions noted in Observation 4.1, (4.4) and (4.5) are identical.

Observation 4.1 (Continuous representation)**.** *For second order (first degree) B-splines with n basis functions, equally spaced data $x_1 < x_2 < \cdots < x_n$ with knots at $t_1 < x_1, t_2 = x_1, t_3 = x_2, \ldots, t_n = x_{n-1}, t_{n+1} = x_n, t_{n+2} > x_n$, and centered outcomes such that $y(0) = 0$, P-splines with an ℓ_1 penalty are a continuous analogue to ℓ_1 trend filtering.*

Proof of Observation 4.1. Under these conditions, for $i = 1, \ldots, n$

$$\phi_j^2(x_i) = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}.$$

To see this, note that

$$\begin{aligned} \phi_j^2(x_i) &= \frac{x_i - t_j}{t_{j+1} - t_j}\phi_j^1(x_i) + \frac{t_{j+2} - x_i}{t_{j+2} - t_{j+1}}\phi_{j+1}^1(x_i) \\ &= \frac{t_{i+1} - t_j}{t_{j+1} - t_j}\phi_j^1(t_{i+1}) + \frac{t_{j+2} - t_{i+1}}{t_{j+2} - t_{j+1}}\phi_{j+1}^1(t_{i+1}). \end{aligned} \tag{4.6}$$

Now,

$$\phi_j^1(t_{i+1}) = \begin{cases} 1 & t_j \le t_{i+1} < t_{j+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \phi_{j+1}^1(t_{i+1}) = \begin{cases} 1 & t_{j+1} \le t_{i+1} < t_{j+2} \\ 0 & \text{otherwise} \end{cases}.$$

We have $\phi_j^1(t_{i+1}) = 1$ for $i = j - 1$ and 0 otherwise, but for $i = j - 1$, we have $t_{i+1} - t_j = t_j - t_j = 0$. We also have $\phi_{j+1}^1(t_{i+1}) = 1$ for $i = j$ and 0 otherwise, and for $i = j$, we have $t_{j+2} - t_{i+1} = t_{j+2} - t_{j+1} > 0$. It follows that for $i = 1 \ldots, n$, (4.6) evaluates to 1 if $i = j$ and 0 otherwise.

Let F be the design matrix in (4.4), where $F_{ij} = \phi_j^2(x_i)$. Then from the previous result, we have $F = I_n$, where $I_{n \times n}$ is the $n \times n$ identity matrix. This, together with the assumption that $\beta_0 = y(0) = 0$, implies that the objective functions (4.4) and (4.5) are identical, which proves Observation 4.1. □

We note that Tibshirani (2014a) shows that ℓ_1 trend filtering has a continuous representation when expressed in the standard lasso form, and Observation 4.1 gives a continuous representation of ℓ_1 trend filtering when expressed in generalized lasso form.

Ramdas and Tibshirani (2016) developed an algorithm to extend ℓ_1 trend filtering to irregularly spaced data. It might also be possible to extend ℓ_1 trend filtering to repeated measures data to account for within-subject correlations. However, due to Observation 4.1, we think it is beneficial to view ℓ_1 trend filtering as a special case of P-splines with an ℓ_1 penalty. We think this approach has the potential to be a general framework, because higher order B-splines could be used in combination with different order difference matrices, just as can be done with P-splines that use the standard ℓ_2 penalty. Furthermore, expressing ℓ_1 trend filtering as P-splines with an ℓ_1 penalty may facilitate the development of confidence bands (see Section 4.7), which could help to fill a gap in the ℓ_1 penalized regression literature.

In addition, there are connections between P-splines with an ℓ_1 penalty and locally adaptive regression splines. In particular, as Tibshirani (2014a) shows, the continuous analogue of ℓ_1 trend filtering is identical to locally adaptive regression splines (Mammen et al., 1997) for $k = 0, 1$, and asymptotically equivalent for $k \geq 2$.

4.3 Proposed Model: Additive Mixed Model Using P-splines with an ℓ_1 Penalty

To introduce our model, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be an $n_i \times 1$ vector of responses for subject $i = 1, \ldots, N$, and let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T)^T$ be the stacked $n \times 1$ vector or responses for all N subjects, where $n = \sum_{i=1}^N n_i$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in_i})^T$ be a corresponding $n_i \times 1$ vector of covariates for subject i, and $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T)^T$ be the $n \times 1$ stacked vector of all covariate values. In many contexts, x is time. To account for the within-subject correlations of $\boldsymbol{y}_i$, we can incorporate random effects into the P-spline model. To that end, let $\tilde{Z}_i$ be an $n_i \times q_i$ design matrix for the random effects for subject i (possibly including a B-spline basis), and let $\tilde{\boldsymbol{b}}_i = (\tilde{b}_{i1}, \ldots, \tilde{b}_{iq_i})^T$ be the corresponding $q_i \times 1$ vector of random effect coefficients for subject i. Also, let

$$\tilde{Z} = \begin{bmatrix} \tilde{Z}_1 & & \\ & \ddots & \\ & & \tilde{Z}_N \end{bmatrix}$$

be the $n \times q$ block diagonal random effects design matrix for all subjects, where $q = \sum_{i=1}^N q_i$, and let $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}_1^T, \ldots, \tilde{\boldsymbol{b}}_N^T)^T$ be the $q \times 1$ stacked vector of random effects for all subjects. We propose an additive mixed model with $j = 1, \ldots, J$ smooths:

$$\underset{\beta_0 \in \mathbb{R}, \tilde{\boldsymbol{b}} \in \mathbb{R}^q, \tilde{\boldsymbol{\beta}}_j \in \mathbb{R}^{p_j}, j=1,\ldots,J}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{j=1}^J \tilde{F}_j \tilde{\boldsymbol{\beta}}_j - \tilde{Z}\tilde{\boldsymbol{b}}\|_2^2 + \sum_{j=1}^J \lambda_j \|\tilde{D}_j^{(k_j+1)} \tilde{\boldsymbol{\beta}}_j\|_1 + \tau \frac{1}{2} \tilde{\boldsymbol{b}}^T \tilde{S} \tilde{\boldsymbol{b}} \quad (4.7)$$

where $\tilde{F}_j$ is a $n \times p_j$ design matrix of B-spline bases for smooth j, $\tilde{D}_j^{(k_j+1)}$ is the $k_j + 1$ finite difference matrix, and $\sigma_b^2 \tilde{S}$ is the covariance matrix of the random effects $\tilde{\boldsymbol{b}}$. For example, if $\tilde{\boldsymbol{b}}$ are random intercepts, then $\tilde{S} = I_{N \times N}$ and $\tilde{Z}$ would be an $n \times N$ matrix such that $\tilde{Z}_{il} = 1$ if observation i belonged to subject l and zero otherwise. Alternatively, to obtain random curves using smoothing splines and a B-spline basis, we could set

$$\tilde{S} = \begin{bmatrix} \tilde{S}_1 & & \\ & \ddots & \\ & & \tilde{S}_N \end{bmatrix}$$

where $\tilde{S}_{j,il} = \int \phi''_{ji}(t) \phi''_{jl}(t) dt$, and $\phi''_{j1}, \ldots, \phi''_{jp_j}$ are the second derivatives of the B-spline basis functions for the j^{th} smooth. We would then set $\tilde{Z}$ to be the corresponding B-splines evaluated at the input points.

We note that (4.8) includes varying-coefficient models (Hastie and Tibshirani, 1993). For example, as pointed out by Wood (2006, p. 169), we could have $F_2 = \text{diag}(\boldsymbol{x'}) F_1$, where $\boldsymbol{x'}$ is another covariate vector, and $\text{diag}(\boldsymbol{x'})$ is a diagonal matrix with x'_i at the i^{th} leading diagonal position.

As written, (4.7) is not generally identifiable. To see this, suppose $\hat{y}(x) = \hat{\beta}_0 + \hat{f}_1(x) + \hat{f}_2(x)$, where neither f_1 nor f_2 are varying-coefficient terms. Then letting $\hat{f}'_1(x) = \hat{f}_1(x) + \delta$ and $\hat{f}'_2(x) = \hat{f}_2(X) - \delta$, we also have $\hat{y}(x) = \hat{\beta}_0 + \hat{f}'_1(x) + \hat{f}'_2(x)$. To make (4.7) identifiable, we follow Wood (2006, Section 4.2) and introduce a centering constraint on each non-varying coefficient smooth, i.e. $\int \hat{f}_j(x) dx = 0$ for all smooths $j = 1, \ldots, J$ such that $F_j \neq \text{diag}(\boldsymbol{x'}) F_l$ for some $\boldsymbol{x'}$ and $l \neq j$. To this end, let $\mathcal{E} = \{j \in \{1, \ldots, J\} : F_j \neq \text{diag}(\boldsymbol{x'}) F_l \text{ for some } \boldsymbol{x'}, l \neq j\}$ be the indices of the non-varying coefficient smooths, and let $\bar{\mathcal{E}} = \{j \in \{1, \ldots, J\} : j \notin \mathcal{E}\}$ be its complement. We constrain $\mathbf{1}^T \tilde{F}_j \tilde{\boldsymbol{\beta}}_j = \mathbf{0}$ for $j \in \mathcal{E}$ and $\mathbf{1}^T \tilde{Z} \tilde{\boldsymbol{b}} = \mathbf{0}$. We accomplish this by defining new $p_j \times (p_j - 1)$ orthonormal matrices Q_j, $j = 1, \ldots, J$, such that $\mathbf{1}^T \tilde{F}_j Q_j = \mathbf{0}$, and a $q \times (q-1)$ matrix Q_{J+1} such that $\mathbf{1}^T \tilde{Z} Q_{J+1} = \mathbf{0}$.

As Wood (2006, Section 1.8.1) shows, Q can be obtained by taking the QR decomposition of $\tilde{F}_j^T\mathbf{1}$ (or $\tilde{Z}^T\mathbf{1}$), and retaining the last $p_j - 1$ (or $q-1$) columns of the left orthonormal matrix.[1] We can then re-parameterize the p_j constrained parameters $\tilde{\boldsymbol{\beta}}_j$ in terms of the $p_j - 1$ unconstrained parameters $\boldsymbol{\beta}_j$, such that $\tilde{\boldsymbol{\beta}} = Q_j\boldsymbol{\beta}_j$. Similarly, we can re-parameterize the q constrained parameters $\tilde{\boldsymbol{b}}$ in terms of the $q-1$ unconstrained parameters $\boldsymbol{b}$, where $\tilde{\boldsymbol{b}} = Q_{J+1}\boldsymbol{b}$. For $j \in \mathcal{E}$, let $F_j = \tilde{F}_jQ_j$ and $D_j = \tilde{D}_j^{k_j+1}Q_j$. For $j \in \bar{\mathcal{E}}$, let $F_j = \tilde{F}_j$ and $D_j = \tilde{D}_j$. Also, let $S = Q_{J+1}^T\tilde{S}Q_{J+1}$ and $Z = \tilde{Z}Q_{J+1}$. Then we can re-write (4.7) in the identifiable form

$$\underset{\beta_0\in\mathbb{R},\boldsymbol{b}\in\mathbb{R}^{q-1},\boldsymbol{\beta}_j\in\mathbb{R}^{p'_j},j=1,\ldots,J}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_{j=1}^J F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^J \lambda_j\|D_j\boldsymbol{\beta}_j\|_1 + \tau\frac{1}{2}\boldsymbol{b}^TS\boldsymbol{b}. \quad (4.8)$$

where $p'_j = p_j - 1$ for $j \in \mathcal{E}$ and $p'_j = p_j$ for $j \in \bar{\mathcal{E}}$.

We note that the penalty matrix S given above for random subject-specific splines defines non-zero correlation between nearby random effect coefficients within subjects. This is in contrast to the approach of Ruppert et al. (2003) for estimating subject-specific random curves, which focuses on the case in which nearby within-subject coefficients are not correlated. Let $\hat{d}_i(x) = \sum_{j=1}^{q_i}\hat{b}_{ij}\phi_{ij}(x)$ be the estimated difference between the i^{th} subject-specific curve and the marginal mean at point x. The smoothing spline approach above constrains $\int(\hat{d}'')^2(x)dx = \boldsymbol{b}_i^TS_i\boldsymbol{b}_i < C$ for some constant $C > 0$, whereas the approach of Ruppert et al. (2003) constrains $\boldsymbol{b}_i^TI_{q_i\times q_i}\boldsymbol{b}_i = \sum_{j=1}^{q_i}\hat{b}_j^2 < C$. Whereas the non-diagonal penalty matrix S implies correlations between nearby coefficients, the identity matrix in the approach of Ruppert et al. (2003) implies zero correlation.

Similar to the equivalence between Bayesian models and ℓ_2 penalized smoothing splines (Wahba, 1990), there is an equivalence between Bayesian models and ℓ_1 penalized splines. In particular, (4.8) is equivalent to the following distributional assumptions, which we can use to obtain Bayesian estimates:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{b} &= \beta_0\mathbf{1} + \sum_{j=1}^J F_j\boldsymbol{\beta}_j + Z\boldsymbol{b} + \boldsymbol{\epsilon}\\
\boldsymbol{\epsilon} &\sim N\left(\mathbf{0}, \sigma_\epsilon^2 I\right)\\
\boldsymbol{b} &\sim N(\mathbf{0}, \sigma_b^2S^{-1}) \text{ for } Ass\sigma_b^2 = \sigma_\epsilon^2/\tau\\
\boldsymbol{\epsilon} &\perp \boldsymbol{b}\\
\left(D_j\boldsymbol{\beta}_j\right)_l &\sim \text{Laplace}(0, a_j) \text{ for } a_j = \sigma_\epsilon^2/(2\lambda_j), l = 1,\ldots,p_j - k_j - 1, j = 1,\ldots,J
\end{aligned}$$

[1] The matrices $\mathbf{1}^T\tilde{F}_j$ and $\mathbf{1}^T\tilde{Z}$ are of rank 1, so the remaining $p_j - 2$ (or $q-2$) columns are arbitrary orthonormal vectors. In R (R Core Team, 2017), when taking the QR decomposition of $\tilde{F}^T\mathbf{1}$, an appropriate matrix Q can be obtained as `Q <- qr.Q(qr(colSums(F_tilde)), complete = TRUE)[, -1]`.

The last distributional assumption is an element-wise Laplace prior on the $k_j + 1$ order differences in coefficients.

In some cases, the random effects penalty matrix S may be positive semidefinite but not invertible. For example, the smoothing spline random curves outlined above lead to a penalty matrix S that is not strictly positive definite, but that is still positive semidefinite. This does not cause problems for the ADMM algorithm, but some changes are required for other algorithms as well as for Bayesian estimation. Following Wood (2006, Section 6.6.1), let $S = U\Lambda U^T$ be the eigendecomposition of a positive semidefinite matrix S, where $UU^T = I_{(q-1)\times(q-1)}$ and Λ is a diagonal matrix with eigenvalues in descending order in the diagonal positions. Let $\breve{\boldsymbol{b}} = U^T\boldsymbol{b}$ and $\breve{Z} = ZU$, so that $\boldsymbol{b}^T S\boldsymbol{b} = \breve{\boldsymbol{b}}^T\Lambda\breve{\boldsymbol{b}}$ and $\breve{Z}\breve{\boldsymbol{b}} = Z\boldsymbol{b}$. Let q_r be the number of strictly positive eigenvalues of S, where $0 < q_r < q - 1$, and let Λ_r be the $q_r \times q_r$ upper left portion of Λ. We can partition $\breve{\boldsymbol{b}}$ as $\breve{\boldsymbol{b}} = (\breve{\boldsymbol{b}}_r^T, \breve{\boldsymbol{b}}_f^T)^T$, where $\breve{\boldsymbol{b}}_r^T$ is a $q_r \times 1$ vector of penalized coefficients and $\breve{\boldsymbol{b}}_f^T$ is a $q_f \times 1$ vector of unpenalized coefficients, where $q_r + q_f = q - 1$. Then $\breve{\boldsymbol{b}}^T\Lambda\breve{\boldsymbol{b}} = \breve{\boldsymbol{b}}_r^T\Lambda_r\breve{\boldsymbol{b}}_r$, and it follows that $\breve{\boldsymbol{b}}_r \sim N(\boldsymbol{0}, \sigma_b^2\Lambda_r^{-1})$ and $\breve{\boldsymbol{b}}_f \propto \boldsymbol{1}$.

However, allowing for unconstrained random effect parameters leads to identifiability issues. Therefore, in practice if $q_f > 0$, we recommend using a normal or Cauchy prior on $\breve{\boldsymbol{b}}_f$. In particular, $\breve{b}_{f,l} \sim N(0, \sigma_f)$ or $\breve{b}_{f,l} \sim \text{Cauchy}(0, \sigma_f)$, $l = 1, \ldots, q_f$ with either a diffuse prior on σ_f and constraints to ensure $\sigma_f > 0$, or a diffuse prior on $\log(\sigma_f)$ without constraints. The Cauchy prior may be a preferable first choice, as it provides a weaker penalty and is similar to the recommendations of Gelman et al. (2008) for logistic regression. However, in some cases, such as in Section 4.9, it is necessary to use a normal prior.

To further improve the computational efficiency of Monte Carlo sampling methods, we can partition $\breve{Z}$ into $\breve{Z} = [\breve{Z}_r, \breve{Z}_f]$ where $\breve{Z}_r$ contains the first q_r columns of $\breve{Z}$ and $\breve{Z}_f$ contains the remaining q_f columns. We then set $\breve{\boldsymbol{b}}_r = \Lambda_r^{-1/2}\breve{\boldsymbol{b}}$ and $\breve{Z}_r = \breve{Z}_r\Lambda_r^{1/2}$, so that $\breve{Z}_r\breve{\boldsymbol{b}}_r = \breve{Z}_r\breve{\boldsymbol{b}}_r$ and $\boldsymbol{b}_r \sim N(\boldsymbol{0}, \sigma_b^2 I)$, which allows for more efficient sampling.

4.4 Related Work

There are many nonparametric and semiparametric methods for analyzing repeated measures data. For an overview, please see Fitzmaurice et al. (2008, Part III). However, most existing methods use an ℓ_2 penalty (e.g. Rice and Wu, 2001, Guo, 2002, Chen and Wang, 2011, Scheipl et al., 2015).

Focusing on the optimization problem, our method puts a generalized lasso penalty (Tibshirani, 1996) on the fixed effects and a quadratic penalty on the random effects. Unlike the elastic net (Zou and Hastie, 2005), we do not mix the ℓ_1 and ℓ_2 penalties on the same parameters, though this could be done in the future.

While not developed for analyzing repeated measures, the fused lasso additive model (FLAM) (Petersen et al., 2016) is similar to ours. FLAM optimizes the following problem:

$$\underset{\theta_0 \in \mathbb{R}, \boldsymbol{\theta}_j \in \mathbb{R}^n, 1 \leq j \leq J}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \theta_0 \mathbf{1} - \sum_{j=1}^{J} \boldsymbol{\theta}_j\|_2^2 + \alpha\lambda \sum_{j=1}^{J} \|D^{(1)}\boldsymbol{\theta}_j\|_1 + (1-\alpha)\lambda \sum_{j=1}^{J} \|\boldsymbol{\theta}_j\|_2 \ , \qquad (4.9)$$

where $0 \leq \alpha \leq 1$ specifies the balance between fitting piecewise constant functions ($\alpha = 1$) and inducing sparsity on the selected smooths ($\alpha = 0$). From Observation 4.1, we see that (4.9) is equivalent to our model (4.8) when: $\alpha = 1$, there is $J = 1$ smooth, our design matrix has $p = n$ columns, our B-spline bases have appropriately chosen knots, and our model has no random effects. As Petersen et al. (2016) show, FLAM can be a very useful method for modeling additive phenomenon, and as with the fused lasso (Tibshirani et al., 2005), jumps in the piecewise linear fits have the advantage of being interpretable.

We also mention the sparse additive model (SpAM) (Ravikumar et al., 2009) and sparse partially linear additive model (SPLAM) (Lou et al., 2016). SpAM fits an additive model and uses a group lasso penalty (Yuan and Lin, 2006) to induce sparsity on the number of active smooths. SPLAM fits a partially linear additive model and uses a hierarchical group lasso penalty (Zhao et al., 2009) to induce sparsity in the selected predictors and to control the number of nonlinear features.

One notable difference between our model and FLAM, SpAM, and SPLAM, is that we allow for multiple smoothing parameters. In our applied experience with additive models and standard ℓ_2 penalties, we have found that in practice it can be important to allow for multiple smoothing parameters, particularly when the quantities of interest are the individual smooths as opposed to the overall prediction. This is equivalent to allowing each smooth to have different variance. However, this flexibility comes at a cost: fitting multiple smoothing parameters is currently the greatest challenge in fitting our proposed model. Perhaps due in part to these computational difficulties, several other authors also assume a single smoothing parameter in high-dimensional additive models (e.g. Lin et al., 2006, Meier et al., 2009).

There are fast and stable methods for fitting multiple smoothing parameters for ℓ_2 penalties paired with exponential family and quasilikelihood loss functions, notably the work of Wood (2004) using generalized cross-validation (GCV) and Wood (2011) using restricted maximum likelihood. Furthermore, Wood et al. (2015) extends these methods to larger datasets, and Wood et al. (2016) extends these methods to likelihoods outside the exponential family and quasilikelihood form. However, similarly computationally efficient methods do not yet exist for fitting multiple smoothing parameters for ℓ_1 penalties.

In addition to allowing for multiple smoothing parameters, we also propose approximate inferential methods, which is not typically provided for ℓ_1 penalized models. Yuan and

Lin (2006), Ravikumar et al. (2009), Lou et al. (2016), and Petersen et al. (2016) focus on prediction and provide bounds on the prediction risk and related quantities. These are important results, and we think that distributional results for individual parameters and smooths will also be useful to practitioners.

We also note that Eilers (2000) and Bollaerts et al. (2006) discuss a variant of P-splines for quantile regression, in which the ℓ_1 norm is used in both the loss and penalty function. However, we are not aware of existing P-spline methods that combine an ℓ_1 penalty with an ℓ_2 loss function.

4.5 Point Estimation

4.5.1 Regression Parameters and Random Effects

To fit (4.8), we use the alternating direction method of multipliers (ADMM) (see Boyd et al., 2011). ADMM has the advantage of being scalable to large datasets. To formulate (4.8) for ADMM, we introduce constraint terms $\boldsymbol{w}_j$ and re-write the optimization problem as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\left\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_{j=1}^{J} F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\right\|_2^2 + \sum_{j=1}^{J}\lambda_j\|\boldsymbol{w}_j\|_1 + \frac{\tau}{2}\boldsymbol{b}^T S\boldsymbol{b} \\ \text{subject to} \quad & D_j\boldsymbol{\beta}_j - \boldsymbol{w}_j = \mathbf{0},\ j = 1, \dots, J \end{aligned} \tag{4.10}$$

The augmented Lagrangian in scaled form (using $\boldsymbol{u}$ to denote the scaled dual variable) is

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{w}, \boldsymbol{u}) \propto \frac{1}{2}\left\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_j F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\right\|_2^2 + \sum_j \lambda_j\left\|\boldsymbol{w}_j\right\|_1 + \frac{\rho}{2}\sum_j \|D_j\boldsymbol{\beta}_j - \boldsymbol{w}_j + \boldsymbol{u}_j\|_2^2 \\ + \frac{\tau}{2}\boldsymbol{b}^T S\boldsymbol{b} \end{aligned}$$

where $\rho > 0$ is the penalty parameter. The dimensions are $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$, $\beta_0 \in \mathbb{R}$, $F_j \in \mathbb{R}^{n\times p'_j}$, $\boldsymbol{\beta}_j \in \mathbb{R}^{p'_j\times 1}$, $Z \in \mathbb{R}^{n\times(q-1)}$, $\boldsymbol{b} \in \mathbb{R}^{(q-1)\times 1}$, $D_j \in \mathbb{R}^{(p_j-k_j-1)\times p'_j}$, $\boldsymbol{w}_j \in \mathbb{R}^{(p_j-k_j-1)\times 1}$, $\boldsymbol{u}_j \in \mathbb{R}^{(p_j-k_j-1)\times 1}$, and $S \in \mathbb{R}^{(q-1)\times(q-1)}$, where $p'_j = p_j - 1$ if $j \in \mathcal{E}$ (non-varying coefficient smooths) and $p'_j = p_j$ if $j \in \bar{\mathcal{E}}$ (varying coefficient smooths).

ADMM is an iterative algorithm, and we re-estimate the parameters for updates $m = 1, 2, \dots$ until convergence.[2] It is straightforward to derive the $m+1$ updates (see Boyd et al.,

[2] We use m to denote the iteration of the ADMM algorithm. This is unrelated to our use of m in Section 4.2 to denote the order of the B-spline basis.

2011, Section 6.4.1):

$$
\begin{aligned}
\beta_0^{m+1} &= \frac{1}{n}\mathbf{1}^T\left(\boldsymbol{y} - \sum_j F_j\boldsymbol{\beta}_j^m - Z\boldsymbol{b}^m\right)\\
\boldsymbol{\beta}_j^{m+1} &:= \underset{\boldsymbol{\beta}_j}{\arg\min}\, L_\rho(\beta_0^{m+1}, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{l<j}^{m+1}, \boldsymbol{\beta}_{l>j}^m, \boldsymbol{b}^m, \boldsymbol{w}^m, \boldsymbol{u}^m)\\
&= \left(F_j^TF_j + \rho D_j^TD_j\right)^{-1}\left(F_j^T\boldsymbol{y}^{(j,m)} + \rho D_j^T(\boldsymbol{w}_j^m - \boldsymbol{u}_j^m)\right)\\
\boldsymbol{b}^{m+1} &:= \underset{\boldsymbol{b}}{\arg\min}\, L_\rho(\boldsymbol{\beta}_{j=1,\dots,J}^{m+1}, \boldsymbol{b}, \boldsymbol{w}^m, \boldsymbol{u}^m)\\
&= (Z^TZ + \tau S)^{-1}Z^T(\boldsymbol{y} - \beta_0^{m+1}\mathbf{1} - \sum_j F_j\boldsymbol{\beta}_j^{m+1})\\
\boldsymbol{w}_j^{m+1} &:= \underset{\boldsymbol{w}_j}{\arg\min}\, L_\rho(\boldsymbol{\beta}_{j=1,\dots,J}^{m+1}, \boldsymbol{b}^{m+1}, \boldsymbol{w}_j, \boldsymbol{u}^m)\\
&= \psi_{\lambda_j/\rho}(D_j\boldsymbol{\beta}_j^{m+1} + \boldsymbol{u}_j^m)\\
\boldsymbol{u}_j^{m+1} &:= \boldsymbol{u}_j^m + D_j\boldsymbol{\beta}_j^{m+1} - \boldsymbol{w}_j^{m+1}
\end{aligned}
$$

where $\boldsymbol{y}^{(j,m)} = \boldsymbol{y} - \beta_0^{m+1}\mathbf{1} - \sum_{l<j} F_l\boldsymbol{\beta}_l^{m+1} - \sum_{l>j} F_j\boldsymbol{\beta}_l^m - Z\boldsymbol{b}^m$ and $\psi_{\lambda/\rho}$ is the element-wise soft thresholding operator, where for a single scalar element x

$$
\psi_{\lambda/\rho}(x) = \begin{cases} x - \lambda/\rho & x > \lambda/\rho \\ 0 & |x| \le \lambda/\rho \\ x + \lambda/\rho & x < -\lambda/\rho \end{cases}
$$

For stopping criteria, we use the primal and dual residuals (r^m and s^m, respectively):

$$
\begin{aligned}
r^m &= \begin{bmatrix} D_1\boldsymbol{\beta}_1^m - \boldsymbol{w}_1^m \\ \vdots \\ D_J\boldsymbol{\beta}_J^m - \boldsymbol{w}_J^m \end{bmatrix} \in \mathbb{R}^{(p-k-J)\times 1}\\
s^m &= -\rho \begin{bmatrix} D_1^T\left(\boldsymbol{w}_1^m - \boldsymbol{w}_1^{m-1}\right) \\ \vdots \\ D_J^T\left(\boldsymbol{w}_J^m - \boldsymbol{w}_J^{m-1}\right) \end{bmatrix} \in \mathbb{R}^{p\times 1}
\end{aligned}
$$

where $k = \sum_{j=1}^J k_j$, $p = \sum_{j=1}^J p_j - |\mathcal{E}|$, and $|\mathcal{E}|$ is the cardinality of $\mathcal{E}$.

Following the guidance of Boyd et al. (2011), we stop when $\|r^m\|_2 \le \epsilon^{\text{pri}}$ and $\|s^m\|_2 \le \epsilon^{\text{dual}}$, where

$$\epsilon^{\text{pri}} = \epsilon^{\text{abs}} \sqrt{p - k - J} + \epsilon^{\text{rel}} \max \left\{ \left\| \begin{matrix} D_1 \boldsymbol{\beta}_1^m \\ \vdots \\ D_J \boldsymbol{\beta}_J^m \end{matrix} \right\|_2 , \left\| \begin{matrix} \boldsymbol{w}_1^m \\ \vdots \\ \boldsymbol{w}_J^m \end{matrix} \right\|_2 \right\}$$

$$\epsilon^{\text{dual}} = \epsilon^{\text{abs}} \sqrt{p} + \epsilon^{\text{rel}} \rho \left\| \begin{matrix} D_1^T \boldsymbol{u}_1^m \\ \vdots \\ D_J^T \boldsymbol{u}_J^m \end{matrix} \right\|_2 .$$

By default, we set $\epsilon^{\text{rel}} = 10^{-4}$ and $\epsilon^{\text{abs}} = 10^{-4}$, and the maximum number of iterations at $1,000$.

4.5.2 Smoothing Parameters

To estimate $\lambda_1, \ldots, \lambda_J$ and τ, we compute cross-validation (CV) error for a path of values one smoothing parameter at a time. In the CV, we split the sample at the subject level, as opposed to individual observations. First, we estimate a path for τ with $\lambda_1, \ldots, \lambda_J$ set to 0. Then we fix τ at the value that minimizes AIC and compute a path for λ_1, setting it to the value that minimizes CV, and so on.

We fit a path for each λ_j from $\lambda_j^{\max}$ to $10^{-5}\lambda_j^{\max}$ evenly spaced on the log scale, where $\lambda_j^{\max}$ is the smallest value at which $D_j \boldsymbol{\beta}_j = \mathbf{0}$. By taking the sub-differential of (4.8) with respect to $\boldsymbol{\beta}_j$ and setting $D_j \boldsymbol{\beta}_j$ to $\mathbf{0}$, we get $\lambda_j^{\max} = \|(D_j D_j^T)^{-1} D_j F^T \boldsymbol{y}\|_\infty$, where for a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_\infty = \max_j |a_j|$. We also use warm starts, passing starting values separately for each fold, though warm starts appear to be minimally beneficial with ADMM. We set $\rho = \min(\max(\lambda_1, \ldots, \lambda_J), c)$ at each iteration for some constant $c > 0$ (e.g. $c = 5$). When the number of smooths J is small (e.g. $J \leq 2$) a grid search is also feasible.

4.6 Degrees of Freedom

In this section, we obtain the degrees of freedom, with the primary goal of estimating variance. However, we note that degrees of freedom does not always align with a model's complexity in terms of its tendency to overfit the data (Janson et al., 2015).

4.6.1 Stein's Method and Estimate of Variance

Let $g(\boldsymbol{y}) = \hat{\boldsymbol{y}}$, where $g : \mathbb{R}^n \to \mathbb{R}^n$ is the model fitting procedure. For $\boldsymbol{y} \sim N(\mu, \sigma^2 I)$, the degrees of freedom is defined as (see Efron, 1986, Hastie and Tibshirani, 1990)

$$\mathrm{df}(g) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}(g_i(y), y_i). \tag{4.11}$$

As Tibshirani (2014a) notes, (4.11) is motivated by the fact that the risk $\mathrm{Risk}(g) = \mathbb{E}\|g(\boldsymbol{y}) - \boldsymbol{\mu}\|_2^2$ can be decomposed as $\mathrm{Risk}(g) = \mathbb{E}\|g(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 - n\sigma^2 + 2\sum_{i=1}^{n}\mathrm{Cov}(g_i(\boldsymbol{y}), y_i)$. Therefore, the degrees of freedom (4.11) corresponds to the difference between risk and expected training error. Furthermore, if g is continuous and weakly differentiable, then $\mathrm{df}(g) = \mathbb{E}[\nabla \cdot g(y)]$ (Stein, 1981) where $\nabla \cdot g = \sum_{i=1}^{n} \partial g_i / \partial y_i$ is the divergence of g. Therefore, an unbiased estimate of $\mathrm{df}(g)$ (also used in Stein's unbiased risk estimate (Stein, 1981)) is

$$\hat{\mathrm{df}}(g) = \sum_{i=1}^{n} \partial g_i / \partial y_i. \tag{4.12}$$

To obtain an estimate of degrees of freedom, we transform the generalized lasso component of our model to standard form, similar to the approach of Petersen et al. (2016). To do so, we use the following matrices described by Tibshirani (2014b). Let

$$\tilde{D}_j^* = \begin{bmatrix} \tilde{D}_{j,1}^{(0)} \\ \vdots \\ \tilde{D}_{j,1}^{(k_j)} \\ \tilde{D}_j^{(k_j+1)} \end{bmatrix} \in \mathbb{R}^{p_j' \times p_j'}$$

be an augmented finite difference matrix, where $\tilde{D}_{j,1}^{(i)}$ is the first row of the finite difference matrix $\tilde{D}_j^{(i)}$, and $\tilde{D}_j^{(0)} = I_{p_j' \times p_j'}$ is the identity matrix where as before, $p_j' = p_j - 1$ if $j \in \mathcal{E}$ (non-varying coefficient smooths) and $p_j' = p_j$ if $j \in \bar{\mathcal{E}}$ (varying coefficient smooths). As shown by Tibshirani (2014b), the inverse of $\tilde{D}_j^*$ is given by $M_j = M_j^{(0)} M_j^{(1)} \cdots M_j^{(k)}$ where[3]

$$M_j^{(i)} = \begin{bmatrix} I_{i \times i} & \\ & L_{(p_j'-i)\times(p_j'-i)} \end{bmatrix} \in \mathbb{R}^{p_j' \times p_j'},$$

where $L_{(p_j'-i)\times(p_j'-i)}$ is the $(p_j' - i) \times (p_j' - i)$ lower diagonal matrix of 1s.

Assuming our outcome $\boldsymbol{y}$ is centered, so that $\beta_0 = y(0) = 0$, and letting $V_j = F_j M_j$, $D_j^* = \tilde{D}_j^* Q_j$ for $j \in \mathcal{E}$ and $D_j^* = \tilde{D}_j^*$ for $j \in \bar{\mathcal{E}}$, and $\boldsymbol{\alpha}_j = D_j^* \boldsymbol{\beta}_j$, we can write the penalized

[3] We denote the inverse matrix as M_j. This is unrelated to our use of M in Section 4.2 to denote the order of the B-spline basis.

log likelihood (4.8) as

$$l_{\text{pen}} = \frac{1}{2}\|\boldsymbol{y} - \sum_j V_j \boldsymbol{\alpha}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{J} \lambda_j \sum_{l=k_j+2}^{p_j} |\alpha_{jl}| + \frac{1}{2}\tau \boldsymbol{b}^T S \boldsymbol{b}. \tag{4.13}$$

To avoid difficulties later differentiating with respect to the ℓ_1 norm, we remove the non-active ℓ_1 penalized coefficients from (4.13). We also form the concatenated design matrix $V = [V_1, \dots, V_J]$ and will need to index the active set of V. To these ends, let $\mathcal{A}_j = \{l \in \{k_j+2, \dots, p'_j\} : \hat{\alpha}_{j,l} \neq 0\}$ be the active set of the penalized coefficients for smooth j, and let $\mathcal{A}^*_j = \{1, \dots, k_j+1\} \cup \mathcal{A}_j$ be the active set for smooth j augmented with the unpenalized coefficients. Also, for a set $\mathcal{A}_j$ and constant $c \in \mathbb{R}$, let $\mathcal{A}_j + c = \{i + c : i \in \mathcal{A}_j\}$ be the set of elements in $\mathcal{A}_j$ shifted by c. Now let $\mathcal{A}^* = \bigcup_{j=1}^{J}(\mathcal{A}^*_j + \sum_{l=0}^{j-1} p'_l)$ be the augmented active set of V, where $p'_0 = 0$ and $p'_j, j = 1, \dots, J$ are the number of columns in V_j (equivalently F_j). Finally, let $V_{\mathcal{A}^*}$ be matrix V subset to retain only those columns indexed by $\mathcal{A}^*$. Similarly, let $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_1^T, \dots, \hat{\boldsymbol{\alpha}}_J^T)^T$ be the concatenated vector of estimated coefficients, and let $\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}$ be vector $\hat{\boldsymbol{\alpha}}$ subset to retain only elements indexed by $\mathcal{A}^*$. Then we can write the estimated penalized loss (4.13) as

$$\hat{l}_{\text{pen}} = \frac{1}{2}\left\|\boldsymbol{y} - [V_{\mathcal{A}^*}, Z]\begin{pmatrix}\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*} \\ \hat{\boldsymbol{b}}\end{pmatrix}\right\|_2^2 + \sum_{j=1}^{J} \lambda_j \sum_{l=k_j+2}^{p_j} |\hat{\alpha}_{jl}| + \frac{1}{2}\tau \hat{\boldsymbol{b}}^T S \hat{\boldsymbol{b}} \tag{4.14}$$

Taking the derivative of (4.14) and keeping in mind that the first $k_j + 1$ elements of each $\hat{\boldsymbol{\alpha}}_j$ are unpenalized and $|\hat{\alpha}_{jl}| > 0$ for all $l \in \mathcal{A}_j$, we have

$$\mathbf{0}_{(|\mathcal{A}^*|+q-1)\times 1} = \frac{\partial l_{\text{pen}}}{\partial(\hat{\boldsymbol{\alpha}}^T_{\mathcal{A}^*}, \hat{\boldsymbol{b}}^T)^T} = \begin{bmatrix} V^T_{\mathcal{A}^*} \\ Z^T \end{bmatrix}\left([V_{\mathcal{A}^*}, Z]\begin{pmatrix}\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*} \\ \hat{\boldsymbol{b}}\end{pmatrix} - \boldsymbol{y}\right) + \begin{pmatrix}\boldsymbol{\eta} \\ \tau S \hat{\boldsymbol{b}}\end{pmatrix} \tag{4.15}$$

where

$$\boldsymbol{\eta} = \begin{bmatrix} \mathbf{0}_{k_1+1} \\ \lambda_1 \operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_1}) \\ \mathbf{0}_{k_2+1} \\ \lambda_2 \operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_2+p_1}) \\ \vdots \\ \mathbf{0}_{k_J+1} \\ \lambda_J \operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_J+\sum_{j=1}^{J-1} p_j}) \end{bmatrix},$$

$\mathbf{0}_{k_j+1}$ is a $(k_j + 1) \times 1$ vector or zeros, and the sign operator is taken element-wise.

From Tibshirani and Taylor (2012, Lemmas 6 and 9), we know that within a small neighborhood of $\boldsymbol{y}$, the active set $\mathcal{A}$ and the sign of the fitted terms $\hat{\alpha}_{\mathcal{A}}$ are constant with respect to $\boldsymbol{y}$ except for $\boldsymbol{y}$ in a set of measure zero. Therefore, $\partial\boldsymbol{\eta}/\partial\boldsymbol{y} = 0_{|\mathcal{A}^*|\times n}$, where $0_{|\mathcal{A}^*|\times n}$

is an $|\mathcal{A}^*| \times n$ matrix of zeros and $|\mathcal{A}^*|$ is the cardinality of $\mathcal{A}^*$. Then taking the derivative of (4.15) with respect to $\boldsymbol{y}$, we have

$$0_{(|\mathcal{A}^*|+q-1)\times n} = \frac{\partial^2 l_{\text{pen}}}{\partial(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T \partial \boldsymbol{y}} = \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} [V_{\mathcal{A}^*}, Z] \begin{bmatrix} \partial \hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}/\partial \boldsymbol{y} \\ \partial \hat{\boldsymbol{b}}/\partial \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} + \begin{bmatrix} 0_{|\mathcal{A}^*|\times n} \\ \tau S(\boldsymbol{\partial \hat{b}/\partial y}) \end{bmatrix}.$$

Solving for the derivatives of the estimated coefficients, we have

$$\begin{bmatrix} \partial \hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}/\partial \boldsymbol{y} \\ \partial \hat{\boldsymbol{b}}/\partial \boldsymbol{y} \end{bmatrix} = \left(\begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix} [V_{\mathcal{A}^*}, Z] + \begin{bmatrix} 0_{|\mathcal{A}^*|\times|\mathcal{A}^*|} & 0_{|\mathcal{A}^*|\times(q-1)} \\ 0_{(q-1)\times|\mathcal{A}^*|} & \tau S \end{bmatrix} \right)^{-1} \begin{bmatrix} V_{\mathcal{A}^*}^T \\ Z^T \end{bmatrix}.$$

Now let $A = [V_{\mathcal{A}^*}, Z]$ and

$$\Omega = \begin{bmatrix} 0_{|\mathcal{A}^*|\times|\mathcal{A}^*|} & 0_{|\mathcal{A}^*|\times(q-1)} \\ 0_{(q-1)\times|\mathcal{A}^*|} & \tau S \end{bmatrix}.$$

Then since $\hat{\boldsymbol{y}} = A(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T$ we have

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{y}} &= \frac{\partial \hat{\boldsymbol{y}}}{\partial(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T} \frac{\partial(\hat{\boldsymbol{\alpha}}_{\mathcal{A}^*}^T, \hat{\boldsymbol{b}}^T)^T}{\partial \boldsymbol{y}} \\ &= A\left(A^T A + \Omega\right)^{-1} A^T. \end{aligned}$$

From Tibshirani and Taylor (2012, Lemmas 1 and 8), we know that $g(\boldsymbol{y}) = \hat{\boldsymbol{y}}$ is continuous and weakly differentiable. Also, $\nabla g = \text{tr}(\partial \hat{\boldsymbol{y}}/\partial \boldsymbol{y})$. Therefore, we can use Stein's formula (4.12) to estimate the degrees of freedom as

$$\hat{\text{df}} = 1 + \text{tr}\left(A(A^T A + \Omega)^{-1} A^T\right) = 1 + \text{tr}\left((A^T A + \Omega)^{-1} A^T A\right), \tag{4.16}$$

where we add 1 for the intercept. We note that this result is similar to the degrees of freedom for the elastic net (see the remark on page 18 of Tibshirani and Taylor, 2012) as well as for FLAM (Petersen et al., 2016).

To obtain degrees of freedom for individual smooths $j = 1, \ldots, J$, let E_j be an $(|\mathcal{A}^*|+q-1) \times (|\mathcal{A}^*|+q-1)$ matrix with 1s on the diagonal positions indexed by $\mathcal{A}_j^* + \sum_{l=0}^{j-1} |\mathcal{A}_l^*|$ and zero elsewhere, where $|\mathcal{A}_j^*|$ is the cardinality of $\mathcal{A}_j^*$ and $\mathcal{A}_0^* = \emptyset$. Also, let $\hat{f}_j = V_j \hat{\boldsymbol{\alpha}}_j$ be the estimate of the j^{th} smooth. Then as Ruppert et al. (2003) note, $\hat{f}_j = AE_j(A^T A + \Omega)^{-1} A^T \boldsymbol{y}$. Therefore,

$$\hat{\text{df}}_j = \text{tr}\left(AE_j(A^T A + \Omega)^{-1} A^T\right) = \text{tr}\left(E_j(A^T A + \Omega)^{-1} A^T A\right). \tag{4.17}$$

In other words, the degrees of freedom for smooth j is the sum of the diagonal elements of $(A^T A + \Omega)^{-1} A^T A$ indexed by $\mathcal{A}_j^* + \sum_{l=0}^{j-1} |\mathcal{A}_l^*|$.

We estimate the overall variance as $\hat{\sigma}_\epsilon^2 = \|\boldsymbol{r}\|_2^2 / \hat{\text{df}}_{\text{resid}}$, where $\hat{\text{df}}_{\text{resid}} = n - \hat{\text{df}}$ and $\boldsymbol{r} = \boldsymbol{y} - \sum_{j=1}^J F_j \hat{\boldsymbol{\beta}}_j - Z\hat{\boldsymbol{b}}$ is an $n \times 1$ vector of residuals. We note that there are alternative estimates for the residual degrees of freedom. In particular, letting $C = A(A^TA+\Omega)^{-1}A^T$, a common alternative estimate of the residual degrees of freedom is $\hat{\text{df}}_{\text{resid}} = n - \text{tr}(2C - CC^T)$ (Buja et al., 1989, Hastie and Tibshirani, 1990). However, in simulations we found that $\text{tr}(2C-CC^T) \approx \text{tr}(C)$ and the latter is easier to compute. Therefore, we set $\text{df}_{\text{resid}} = n - \text{tr}(C)$ as the residual degrees of freedom. This is also in keeping with the advice of Wood (2006, Section 4.4.1)

Finally, we note that when using the ADMM algorithm, or most likely any proximal algorithm, the fitted $D_j\hat{\boldsymbol{\beta}}_j$, or equivalently $\hat{\boldsymbol{\alpha}}_j$, will typically have several very small non-zero values, but will not typically be sparse. However, the vector $\hat{\boldsymbol{w}}_j$ is sparse, where in the ADMM algorithm we constrain $\boldsymbol{w}_j = D_j\boldsymbol{\beta}_j$. Therefore, in practice we use $\boldsymbol{w}_j$ to obtain the active set $\mathcal{A}_j$.

4.6.2 Stable and Fast Approximations

In some cases, it may be numerically instable or computationally expensive to compute (4.16) and (4.17). In this section we propose alternatives that are faster to compute and which use less memory.

Based on Restricted Derivatives

In this approach, we take derivatives of the fitted values restricted to individual smooths. In particular, from Section 4.6.1, we see that

$$\begin{aligned}\frac{\partial \hat{\boldsymbol{y}}}{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{A}_j^*}}\frac{\partial \hat{\boldsymbol{\alpha}}_{\mathcal{A}_j^*}}{\partial \boldsymbol{y}} &= V_{\mathcal{A}_j^*}(V_{\mathcal{A}_j^*}^T V_{\mathcal{A}_j^*})^{-1}V_{\mathcal{A}_j^*}^T \\ \frac{\partial \hat{\boldsymbol{y}}}{\partial \hat{\boldsymbol{b}}}\frac{\partial \hat{\boldsymbol{b}}}{\partial \boldsymbol{y}} &= Z(Z^TZ+\tau S)^{-1}Z^T.\end{aligned}$$

We can then approximate the degrees of freedom for each individual smooth and the random effects by

$$\tilde{\text{df}}_j = \begin{cases}\text{tr}\left((V_{\mathcal{A}_j^*}^T V_{\mathcal{A}_j^*})^{-1}V_{\mathcal{A}_j^*}^T V_{\mathcal{A}_j^*}\right) & j = 1, \ldots, J \\ \text{tr}\left((Z^TZ+\tau S)^{-1}Z^TZ\right) & j = J+1\end{cases} \tag{4.18}$$

We estimate the overall degrees of freedom as

$$\tilde{\text{df}} = 1 + \sum_{j=1}^{J+1}\tilde{\text{df}}_j \tag{4.19}$$

where we add 1 for the intercept.

This approach is similar to one described by Ruppert et al. (2003, p. 176), though in a different context and for a different purpose. In particular, whereas we use this approach to approximate the degrees of freedom after fitting the model, Ruppert et al. (2003) use it to set the degrees of freedom before fitting the model in the context of ℓ_2 penalized loss functions.

Based on ADMM Constraint Parameters

In this approach, we propose estimates of degrees of freedom specific to the ADMM algorithm. As in the previous section, this approach is based on estimates for the individual smooths. Consider the model with $J = 1$ smooth, no random effects, and centered $\boldsymbol{y}$:

$$\|\boldsymbol{y} - F\boldsymbol{\beta}\|_2^2 + \lambda\|D\boldsymbol{\beta}\|_1.$$

If we make the centering constraints described Section 4.3, i.e. for $j \in \mathcal{E}$, we have $F = \tilde{F}Q$ and $D = \tilde{D}^{(k+1)}Q$ for an $n \times p$ design matrix $\tilde{F}$, a $k+1$ order finite difference matrix $D^{(k+1)}$, and an orthonormal $p \times (p-1)$ matrix Q. Let $\mathcal{A} = \{l \in \{1, \ldots, p-k-1\} : (D\hat{\boldsymbol{\beta}})_l \neq 0\}$ be the active set, and let $|\mathcal{A}|$ be its cardinality. In our context, we expect the design matrices F to be full rank, in which case Theorem 3 of Tibshirani and Taylor (2012) (see the first Remark) states that the degrees of freedom is given by $\text{df} = \mathbb{E}[\text{nullity}(D_{-\mathcal{A}})]$. Here, $\text{nullity}(D)$ is the dimension of the null space of matrix D, and $D_{-\mathcal{A}}$ is matrix D with rows indexed by $\mathcal{A}$ removed. Now, D has dimensions $(p-k-1) \times (p-1)$, and we can see by inspection that for all $k < p-1$ the columns of D are linearly independent. Therefore, the rank of $D_{-\mathcal{A}}$ is equal to the number of rows $p-k-1-|\mathcal{A}|$, and the nullity is equal to the number of columns $p-1$ minus the number of rows. This gives $\hat{\text{df}} = \text{nullity}(D_{-\mathcal{A}}) = k + |\mathcal{A}|$ for centered smooths, i.e. the number of non-zero elements of $D\hat{\boldsymbol{\beta}}$ plus one less than the order of the difference penalty. This is similar to the result for ℓ_1 trend filtering, but we have lost one degree of freedom due to the constraint that $\mathbf{1}^T\tilde{F}\tilde{\boldsymbol{\beta}} = \mathbf{0}$. For uncentered smooths, D has dimensions $(p-k-1) \times p$, which gives $\hat{\text{df}} = \text{nullity}(D_{-\mathcal{A}})) = k + 1 + |\mathcal{A}|$.

As before, we note that in the ADMM algorithm, $D\hat{\boldsymbol{\beta}}$ will not generally be sparse, as ADMM is a proximal algorithm. However, the corresponding $\boldsymbol{w}$ is sparse, where in the optimization problem, we constrain $D\boldsymbol{\beta} = \boldsymbol{w}$. Suppose D is an order k finite difference matrix. Then for smooth $j = 1, \ldots, J$, a fast alternative to (4.17) is given by

$$\tilde{\text{df}}_j^{\text{ADMM}} = \mathbb{1}[j \in \mathcal{E}] + k_j + \sum_{l=1}^{p-k-1} \mathbb{1}\left[w_{jl} \neq 0\right]. \tag{4.20}$$

where $\mathcal{E}$ indexes the smooths subject to centering constraints and $\mathbb{1}$ is an indicator variable. We then combine (4.20) with the restricted derivative approximation for the degrees of freedom of $\hat{\boldsymbol{b}}$ given above to obtain the overall degrees of freedom

$$\tilde{\text{df}}^{\text{ADMM}} = 1 + \sum_{j=1}^{J} \hat{\text{df}}_j^{\text{ADMM}} + \text{tr}\left((Z^T Z + \tau S)^{-1} Z^T Z\right), \tag{4.21}$$

where we add 1 for the intercept.

4.6.3 Ridge Approximation

Let $U = [F_1, \ldots, F_J, Z]$ be the concatenated design matrix of fixed and random effects and

$$\Omega^{\text{ridge}} = \begin{bmatrix} \lambda_1 D_1^T D_1 & & & \\ & \ddots & & \\ & & \lambda_J D_J^T D_J & \\ & & & \tau S \end{bmatrix}$$

be the penalty matrix. Then the hat matrix from the linear smoother approximation (see Section 4.7) is given by $H = U(U^T U + \Omega^{\text{ridge}})^{-1} U^T$. Similar to before, we can get the overall degrees of freedom as

$$\hat{\text{df}}^{\text{ridge}} = 1 + \text{tr}\left((U^T U + \Omega^{\text{ridge}})^{-1} U^T U\right), \tag{4.22}$$

where we add 1 for the intercept. To obtain degrees of freedom for individual smooths $j = 1, \ldots, J$, let E_j be a $(p+q-1) \times (p+q-1)$ matrix with 1s on the diagonal positions indexed by the columns of F_j and zero elsewhere. Also, let $\hat{f}_j = F_j \hat{\boldsymbol{\beta}}_j$ be the estimate of the j^{th} smooth. Then the ridge approximation for smooth j is given by $\hat{f}_j \approx U E_j (U^T U + \Omega^{\text{ridge}})^{-1} U^T \boldsymbol{y}$. Therefore,

$$\hat{\text{df}}_j^{\text{ridge}} = \text{tr}\left(E_j (U^T U + \Omega^{\text{ridge}})^{-1} U^T U\right) \tag{4.23}$$

Similar to before, we also propose fast approximations to the ridge estimate of degrees of freedom based on restricted derivatives. In particular, let

$$\tilde{\text{df}}_j^{\text{ridge}} = \begin{cases} \text{tr}\left(({F_j}^T F_j + \lambda_j D_j^T D_j)^{-1} F_j^T F\right) & j = 1, \ldots, J \\ \text{tr}\left((Z^T Z + \tau S)^{-1} Z^T Z\right) & j = J+1 \end{cases} \tag{4.24}$$

Then we can estimate the overall degrees of freedom as

$$\tilde{\text{df}}^{\text{ridge}} = 1 + \sum_{j=1}^{J+1} \tilde{\text{df}}_j^{\text{ridge}} \tag{4.25}$$

where we add 1 for the intercept.

As noted above, this approach is similar to one described by Ruppert et al. (2003, p. 176), though for a different purpose. Whereas we use this approach to obtain the degrees of freedom after fitting the model, Ruppert et al. (2003) use it to set the degrees of freedom before fitting the model.

4.7 Approximate Inference

In this section, we discuss approximate inferential methods based on ridge approximations to the ℓ_1 penalized fit and conditional on the smoothing parameters $\lambda_j, j = 1, \ldots, J$ and τ. We use the ADMM algorithm to analyze the approximation. In particular, we note that we can write the ADMM update for $\boldsymbol{\beta}_j$ as

$$\boldsymbol{\beta}_j^{m+1} = \left(F_j^T F_j + \rho D_j^T D_j\right)^{-1} F_j^T \boldsymbol{y}^{(j,m)} + \boldsymbol{\delta}_j^m \tag{4.26}$$

where $\boldsymbol{\delta}_j^m = \rho(F_j^T F_j + \rho D_j^T D_j)^{-1} F_j^T D_j^T(\boldsymbol{w}_j^m - \boldsymbol{u}_j^m)$ and $\boldsymbol{y}^{(j,m)} = \boldsymbol{y} - \beta_0^{m+1} - \sum_{l<j} F_l \boldsymbol{\beta}_l^{m+1} - \sum_{l>j} F_l \boldsymbol{\beta}_l^m - Z\boldsymbol{b}^m$. As we note in Observation 1.10, $\boldsymbol{\delta}_j$ loosely represents the difference in the estimate of $\boldsymbol{\beta}_j$ obtained with the ℓ_1 and ℓ_2 penalties.

Observation 4.2. *With the ℓ_1 penalty, i.e. $\|D_j\boldsymbol{\beta}_j\|_1$, in general $\boldsymbol{\delta}_j^m \neq \mathbf{0}$. However, with the ℓ_2 penalty, i.e. $\|D_j\boldsymbol{\beta}_j\|_2^2$, and $\lambda_j = \rho$, we have $\boldsymbol{\delta}_j^m = \mathbf{0}$.*

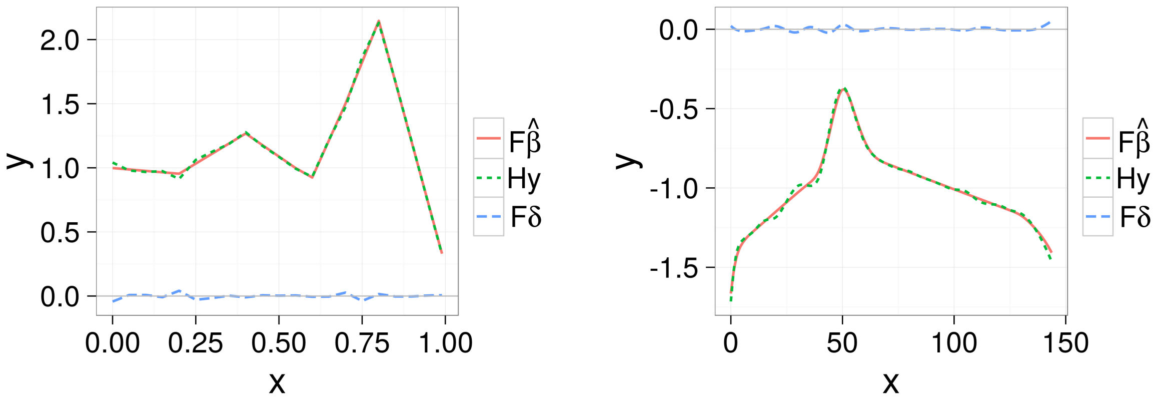
Proof of Observation 4.2. Similar to the ridge update for $\boldsymbol{b}$, if we changed $\lambda_j\|D_j\boldsymbol{\beta}_j\|_1$ to $(\lambda_j/2)\|D_j\boldsymbol{\beta}_j\|_2^2$ in (4.8) we could remove the $\boldsymbol{w}_j$ term and the constraint that $D_j\boldsymbol{\beta}_j^m = \boldsymbol{w}_j$ from (4.10) to obtain a ridge update $\boldsymbol{\beta}_j^{m+1} = \left(F_j^T F_j + \lambda_j D_j^T D_j\right)^{-1} F_j^T \boldsymbol{y}^{(j,m)}$. Then since we assumed $\lambda_j = \rho$, we have $\boldsymbol{\beta}_j^{m+1} = \left(F_j^T F_j + \rho D_j^T D_j\right)^{-1} F_j^T \boldsymbol{y}^{(j,m)}$. By comparison with (4.26), we see that $\boldsymbol{\delta}_j^m = \mathbf{0}$. □

Observation 4.2 motivates our approximate inferential strategy. Letting $\hat{\boldsymbol{f}}_j$ be the j^{th} fitted smooth, and letting $\boldsymbol{y}^{(j)} = \boldsymbol{y} - \hat{\beta}_0 - \sum_{l\neq j} F_l\hat{\boldsymbol{\beta}}_l - Z\hat{\boldsymbol{b}}$, we have

$$\begin{aligned}
\hat{\boldsymbol{f}}_j = F_j\hat{\boldsymbol{\beta}}_j &= F_j(F_j^T F_j + \rho D_j^T D_j)^{-1} F_j^T \boldsymbol{y}^{(j)} + F_j\hat{\boldsymbol{\delta}}_j && (4.27)\\
&\approx F_j(F_j^T F_j + \rho D_j^T D_j)^{-1} F_j^T \boldsymbol{y}^{(j)} && \text{(assuming } F_j\hat{\boldsymbol{\delta}}_j \approx \mathbf{0})\\
&\approx F_j(F_j^T F_j + \lambda_j D_j^T D_j)^{-1} F_j^T \boldsymbol{y}^{(j)} && \text{(assuming } \lambda_j \approx \rho)\\
&= H_j\boldsymbol{y}^{(j)} && (4.28)
\end{aligned}$$

where $H_j = F_j(F_j^T F_j + \lambda_j D_j^T D_j)^{-1} F_j^T$. We obtain confidence intervals for the linear smoother (4.28) centered around the estimated fit (4.27), ignore $F_j\boldsymbol{\delta}_j$ when estimating variance, and assume $\lambda_j \approx \rho$. We also condition on the smoothing parameters $\lambda_1, \ldots, \lambda_J$ and τ.

Figure 4.2 gives a visual demonstration of the approximation for the simulation presented in Section 4.8 and the application shown in Section 4.9. As seen in Figure 4.2, in these examples the ℓ_1 fit and ridge approximation are very similar. If this holds in general, then this would suggest that 1) the approximate inferential procedures we propose might have reliable coverage probabilities, and 2) there may be minimal practical advantage to using an ℓ_1 penalty instead of the standard ℓ_2 penalty. However, as shown in Section 4.8.3, the ℓ_1



(a) Simulation (Section 4.8)

(b) Application (Section 4.9)

Figure 4.2: Linear smoother approximation to the ℓ_1 penalized fit in the simulation (see Section 4.8) and application (see Section 4.9). The solid red line is the ℓ_1 penalized fit, the dotted green line is the linear smoother approximation, and the dashed blue line is the difference between the two.

penalty appears to perform noticeably better in certain situations, including the detection of change points.

4.7.1 Confidence Bands

In this section, we obtain confidence bands for typical subjects, i.e. for subjects for whom $\boldsymbol{b}_i = \mathbf{0}$. Since we assume a normal outcome, this is equivalent to the marginal population level response.

Frequentist Confidence Bands

Ignoring the distribution on $D_j\boldsymbol{\beta}_j$, $\boldsymbol{y}^{(j)}$ is normal with variance $\text{Var}(\boldsymbol{y}^{(j)}) = \sigma_\epsilon^2 I + \sigma_b^2 Z S^+ Z^T$, where S^+ is the Moore-Penrose generalized inverse of matrix S (as noted in Section 4.3, S may not be positive definite). Therefore, $\widehat{\text{Var}}(\hat{\boldsymbol{f}}_j) \approx H_j \widehat{\text{Var}}(\boldsymbol{y}^{(j)}) H_j^T$ where $\widehat{\text{Var}}(\boldsymbol{y}^{(j)})$ is an $n \times n$ estimate of $\text{Var}(\boldsymbol{y}^{(j)})$ with $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_b^2$ plugged in for σ_ϵ^2 and σ_b^2 respectively, and $\hat{\boldsymbol{f}}_j \dot{\sim} N(\hat{\boldsymbol{f}}_j, H_j \widehat{\text{Var}}(\boldsymbol{y}^{(j)}) H_j^T)$. The estimated variance of the fit at a single point x, which we denote as $\widehat{\text{Var}}(\hat{f}_j(x))$, is the corresponding diagonal element of $H_j \widehat{\text{Var}}(\boldsymbol{y}^{(j)}) H_j^T$. Therefore, asymptotic pointwise $1-\alpha$ confidence bands take the form $\hat{f}_j(x) \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{f}_j(x))}$ where $\Phi(z_a) = a$ and Φ is the standard normal CDF, e.g. $z_{1-\alpha/2} = 1.96$ for $\alpha = 0.05$.

For the purposes of interpretation, we include the intercept term in the confidence band for the $j = 1$ smooth, but not for the remaining smooths.

Bayesian Credible Bands

Many authors, including Wood (2006), recommend using Bayesian confidence bands for nonparametric and semiparametric models, because the point estimates are themselves biased. While Bayesian credible bands do not remedy the bias, they are self consistent.

To this end, we replace the element-wise Laplace prior with the (generally improper) joint normal prior that is equivalent to the standard ℓ_2 penalty: $\boldsymbol{\beta}_j \sim N\left(\mathbf{0}, (\lambda_j D_j^T D_j)^{-1}\right)$. This leads to the posterior

$$\boldsymbol{\beta}_j | \boldsymbol{y} \dot{\sim} N\left(\hat{\boldsymbol{\beta}}_j, (\underbrace{F_j^T \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)})^{-1} F_j + \lambda_j D_j^T D_j}_{W_j})^{-1}\right). \tag{4.29}$$

We can then form simultaneous Bayesian credible bands for $\boldsymbol{f}_j | \boldsymbol{y}$ by simulating from the posterior (4.29) and taking quantiles from $F_j \boldsymbol{\beta}_j^b, b = 1, \ldots, B$. Alternatively, for a faster approximation, we use frequentist confidence bands with $F_j W_j^{-1} F_j^T$ in place of $H_j \widehat{\mathrm{Var}}(\boldsymbol{y}^{(j)}) H_j^T$. In practice, we have found the simultaneous credible bands and the faster approximation to be nearly indistinguishable.[4]

As before, for the purposes of interpretation, we include the intercept term in the credible band for the $j = 1$ smooth, but not for the remaining smooths.

4.7.2 Preliminaries for Bounding the Error in Confidence Band Coverage Probabilities

In this section, we present preliminary work relevant to bounding the error in the coverage probabilities of the approximate confidence bands. Since the approximate confidence bands are based on a ridge approximation to the ℓ_1 penalized fit, we develop bounds for the difference between the ridge and ℓ_1 penalized fits. For simplicity, we focus on models with $J = 1$ smooth and no random effects. We use subscripts on parameters to denote the form of the penalty, i.e. $\boldsymbol{\beta}_1$ is obtained from a model with an ℓ_1 penalty, and $\boldsymbol{\beta}_2$ is obtained from a model with an ℓ_2 penalty. Throughout, we treat the smoothing parameter λ as constant.

First, consider the ℓ_1 penalized model

$$\underset{\boldsymbol{\beta}_1 \in \mathbb{R}^p}{\mathrm{minimize}} \frac{1}{2} \|\boldsymbol{y} - F\boldsymbol{\beta}_1\|_2^2 + \lambda \|D\boldsymbol{\beta}_1\|_1. \tag{4.30}$$

[4] It appears that the latter (faster) method is the default in the `mgcv` package (Wood, 2006). As in `mgcv`, we only need to compute the diagonal elements of $F_j W_j^{-1} F_j^T$ as `rowSums`$((F_j W_j^{-1}) \circ F_j)$, where $\circ$ is the Hadamard (element-wise) product.

From Section 4.5, the ADMM updates are

$$\begin{aligned}\boldsymbol{\beta}_1^{m+1} &= \left(F^TF + \rho D^TD\right)^{-1}\left(F^T\boldsymbol{y} + \rho D^T(\boldsymbol{w}_1^m - \boldsymbol{u}_1^m)\right) \\ \boldsymbol{w}_1^{m+1} &= \psi_{\lambda/\rho}(D\boldsymbol{\beta}_1^{m+1} + \boldsymbol{u}_1^m) \\ \boldsymbol{u}_1^{m+1} &= \boldsymbol{u}_1^m + D\boldsymbol{\beta}_1^{m+1} - \boldsymbol{w}_1^{m+1}\end{aligned} \tag{4.31}$$

where ψ is the element-wise soft thresholding operator defined in Section 4.5. Let M be the iteration at convergence of the ADMM algorithm. Then the estimate from the ADMM algorithm is given by

$$\begin{aligned}\hat{\boldsymbol{y}}_1 = F\boldsymbol{\beta}_1^M &= F(F^TF + \rho D^TD)^{-1}\left[F^T\boldsymbol{y} + \rho D^T(\boldsymbol{w}^M - \boldsymbol{u}^M)\right] \\ &= H_\rho\boldsymbol{y} + R(\boldsymbol{w}^M - \boldsymbol{u}^M)\end{aligned} \tag{4.32}$$

where $H_\rho = F(F^TF + \rho D^TD)^{-1}F^T$ and $R = \rho F(F^TF + \rho D^TD)^{-1}D^T$. We note that for $\boldsymbol{\delta}^m = \rho(F^TF + \rho D^TD)^{-1}D^T(\boldsymbol{w}^m - \boldsymbol{u}^m)$, we have $F\boldsymbol{\delta}^m = R(\boldsymbol{w}^m - \boldsymbol{u}^m)$.

We compare against the equivalent ℓ_2 penalized model

$$\underset{\boldsymbol{\beta}_2 \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2}\|\boldsymbol{y} - F\boldsymbol{\beta}_2\|_2^2 + \lambda\frac{1}{2}\|D\boldsymbol{\beta}_2\|_2^2. \tag{4.33}$$

For the ℓ_2 penalized model we can solve (4.33) directly to obtain parameter estimates

$$\hat{\boldsymbol{\beta}}_2 = (F^TF + \lambda D^TD)^{-1}F^T\boldsymbol{y} \tag{4.34}$$

and fitted values

$$\hat{\boldsymbol{y}}_2 = F\hat{\boldsymbol{\beta}}_2 = F(F^TF + \lambda D^TD)^{-1}F^T\boldsymbol{y} = H_\lambda\boldsymbol{y} \tag{4.35}$$

where $H_\lambda = F(F^TF + \lambda D^TD)^{-1}F^T$. We can then bound the difference between $\hat{\boldsymbol{y}}_1$ (4.32) and $\hat{\boldsymbol{y}}_2$ (4.35) as

$$\begin{aligned}\|\hat{\boldsymbol{y}}_1 - \hat{\boldsymbol{y}}_2\|_2^2 &= \|H_\rho\boldsymbol{y} + R(\boldsymbol{w}^M - \boldsymbol{u}^M) - H_\lambda\boldsymbol{y}\|_2^2 \\ &= \|(H_\rho - H_\lambda)\boldsymbol{y} + R(\boldsymbol{w}^M - \boldsymbol{u}^M)\|_2^2 \\ &\leq \|(H_\rho - H_\lambda)\boldsymbol{y}\|_2^2 + \|R(\boldsymbol{w}^M - \boldsymbol{u}^M)\|_2^2 \\ &= \boldsymbol{y}^TK_H\boldsymbol{y} + (\boldsymbol{w}^M - \boldsymbol{u}^M)^TK_R(\boldsymbol{w}^M - \boldsymbol{u}^M) \\ &\leq \xi_{H,1}\|\boldsymbol{y}\|_2^2 + \xi_{R,1}\|\boldsymbol{w}^M - \boldsymbol{u}^M\|_2^2,\end{aligned} \tag{4.36}$$

where $K_H = (H_\rho - H_\lambda)^T(H_\rho - H_\lambda)$, $K_R = R^TR$, and $\xi_{H,1}$ and $\xi_{R,1}$ are the top eigenvalues of K_H and K_R, respectively. We note that when $\lambda = \rho$, the first term in (4.36) is zero. The second term is a function of the ADMM parameters $\boldsymbol{w}$ and $\boldsymbol{u}$, and the matrix K_R, which is in turn a function of the design and penalty matrices. In future work, we plan to investigate whether further bounds can be put on these quantities under certain assumptions on the data.

As another potential strategy, we can bound the quantity $\boldsymbol{w}$ obtained with the ℓ_1 and ℓ_2 penalties. To that end, note that we can solve (4.33) using the ADMM algorithm by rewriting (4.33) as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\left\|\boldsymbol{y} - F\boldsymbol{\beta}_2\right\|_2^2 + \lambda\frac{1}{2}\|\boldsymbol{w}_2\|_2^2 \\ \text{subject to} \quad & D\boldsymbol{\beta}_2 - \boldsymbol{w}_2 = \boldsymbol{0}. \end{aligned} \tag{4.37}$$

The Lagrangian in scaled form for (4.37) is given by

$$L_\rho \propto \frac{1}{2}\left\|\boldsymbol{y} - F\boldsymbol{\beta}_2\right\|_2^2 + \lambda\frac{1}{2}\left\|\boldsymbol{w}_2\right\|_2^2 + \frac{\rho}{2}\left\|D\boldsymbol{\beta}_2 - \boldsymbol{w}_2 + \boldsymbol{u}_2\right\|_2^2,$$

which leads to the ADMM updates

$$\begin{aligned} \boldsymbol{\beta}_2^{m+1} &= \left(F^TF + \rho D^TD\right)^{-1}\left(F^T\boldsymbol{y} + \rho D^T(\boldsymbol{w}_2^m - \boldsymbol{u}_2^m)\right) \\ \boldsymbol{w}_2^{m+1} &= \frac{\rho}{\lambda+\rho}(D\boldsymbol{\beta}_2^{m+1} + \boldsymbol{u}_2^m) \\ \boldsymbol{u}_2^{m+1} &= \boldsymbol{u}_2^m + D\boldsymbol{\beta}_2^{m+1} - \boldsymbol{w}_2^{m+1}. \end{aligned} \tag{4.38}$$

By comparing the ADMM updates from the ℓ_1 and ℓ_2 penalized models, the only difference is the update for $\boldsymbol{w}$ in (4.31) and (4.38). If $D\boldsymbol{\beta}_1^{m+1} \in [-\lambda/\rho, \lambda/\rho]^{p-k-1}$, i.e. the absolute value of each component of $D\boldsymbol{\beta}_1^{m+1}$ is no more than λ/ρ, then the difference between the updated $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is

$$\left\|\boldsymbol{w}_1^{m+1} - \boldsymbol{w}_2^{m+1}\right\|_2^2 \leq (p-k-1)\left(\frac{\lambda}{\lambda+\rho}\right)^2. \tag{4.39}$$

This follows because $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are $(p-k-1) \times 1$ vectors, and under the assumption $D\boldsymbol{\beta}_1^{m+1} \in [-\lambda/\rho, \lambda/\rho]^{p-k-1}$, each term can be off by at most

$$\frac{\rho}{\lambda+\rho}\frac{\lambda}{\rho} = \frac{\lambda}{\lambda+\rho}.$$

However, further work is needed to incorporate (4.39) into bounds on the error in the coverage probabilities of the approximate confidence bands.

4.8 Simulation

We simulated data from a piecewise linear mean curve as shown in Figure 4.3. Each subject had a random intercept, and is observed over only a portion of the domain. There are 50

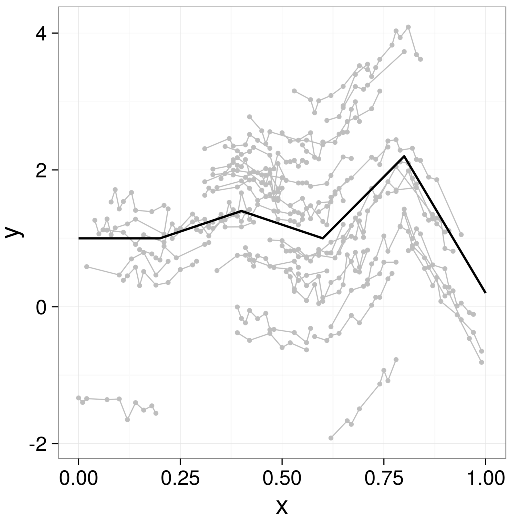


Figure 4.3: Simulated data: true marginal curve in black, observed (simulated) data in gray.

subjects, each with between 4 and 14 measurements (450 total observations). The random intercepts were normally distributed with variance $\sigma_b^2 = 1$, and the overall noise was normally distributed with variance $\sigma_\epsilon^2 = 0.01$.

In all models, we used order 2 (degree 1) B-splines with $p = 21$ basis functions.

4.8.1 Frequentist Estimation

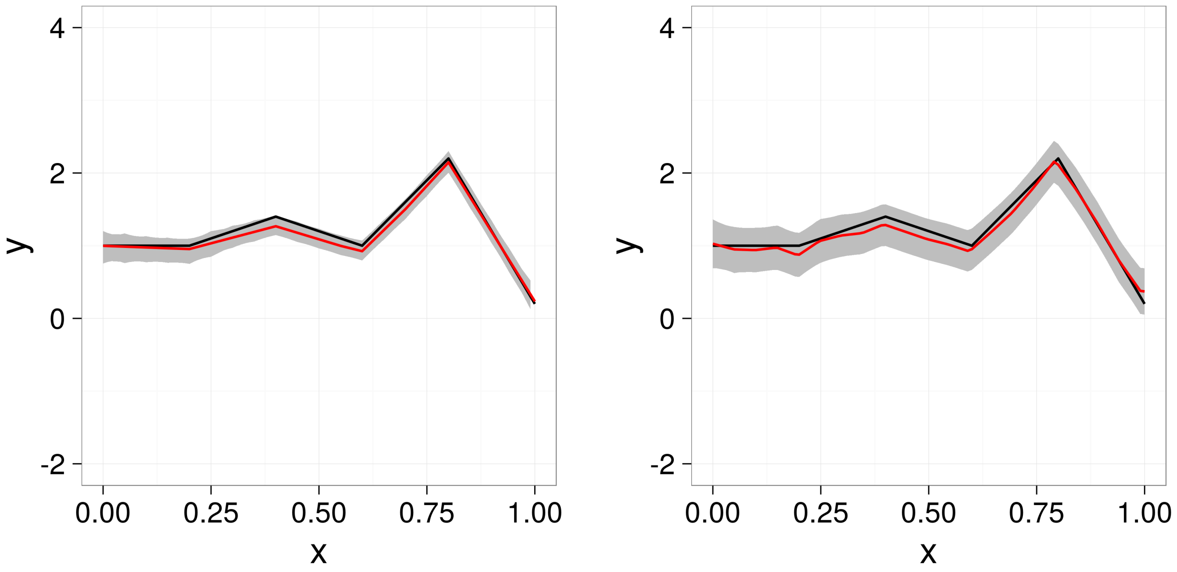
We fit models with $J = 1$ smooth term and random intercepts. To obtain estimates for the ℓ_1 penalized model, we used ADMM and 5-fold CV to minimize

$$\underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^{p-1}, \boldsymbol{b} \in \mathbb{R}^{N-1}}{\text{minimize}} \frac{1}{2}\|\boldsymbol{y} - \beta_0 \mathbf{1} - F\boldsymbol{\beta} - \boldsymbol{b}\|_2^2 + \lambda \|D^{(2)}\boldsymbol{\beta}\|_1 + \tau \boldsymbol{b}^T \boldsymbol{b}. \tag{4.40}$$

As noted above, we used order 2 (degree 1) B-splines with $p = 21$ basis functions, i.e. $F \in \mathbb{R}^{n \times (p-1)}$ where $n = 450$ and $p = 21$. We also fit an equivalent model with an ℓ_2 penalty using the `mgcv` package (Wood, 2006), i.e. with $(\lambda/2)\|D^{(2)}\boldsymbol{\beta}\|_2^2$ in place of $\lambda\|D^{(2)}\boldsymbol{\beta}\|_1$ in (4.40). Figure 4.4 shows the marginal mean with 95% credible intervals, and Figure 4.5 shows the subject-specific predicted curves.

As seen in Figures 4.4 and 4.5, the results from the ℓ_1 and ℓ_2 penalized models are very similar. For most purposes, we would recommend using the ℓ_2 penalized model. However, the ℓ_1 penalized model does slightly better at identifying the change points and the line segments. We explore this further in Section 4.8.3.

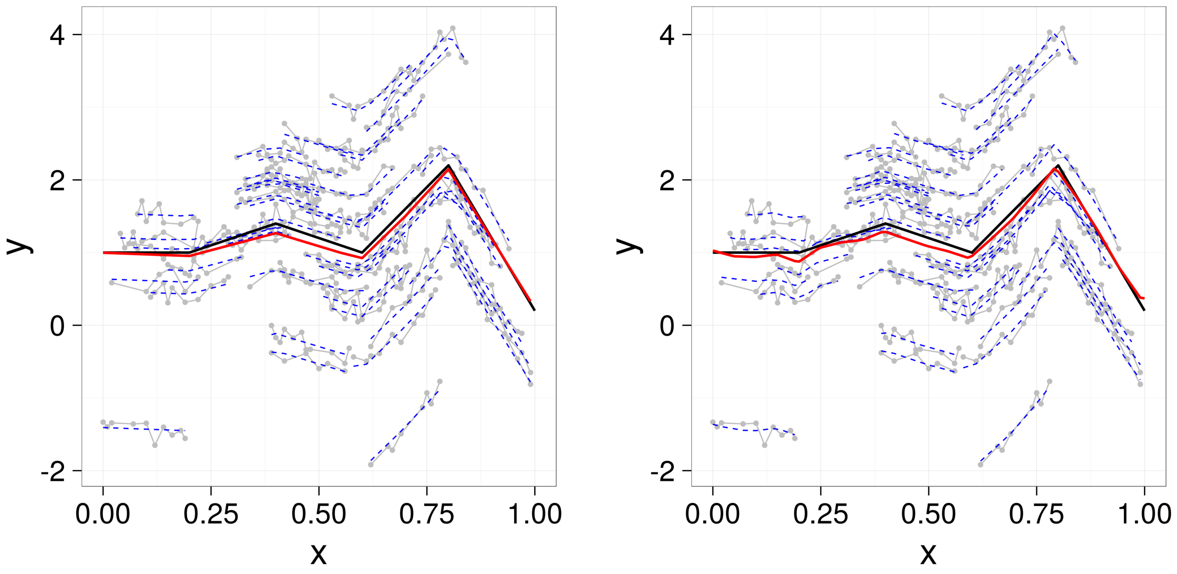
Table 4.1 compares the degrees of freedom and variance estimates from the ℓ_1 penalized fit against those from the ℓ_2 penalized fit. From Table 4.1, we see that the ridge degrees



(a) ℓ_1 fit with ADMM and CV

(b) ℓ_2 fit with mgcv (Wood, 2006)

Figure 4.4: Marginal mean and 95% credible intervals from frequentist estimation: black is true marginal mean, red is estimated marginal mean



(a) ℓ_1 fit with ADMM and CV

(b) ℓ_2 fit with mgcv (Wood, 2006)

Figure 4.5: Subject-specific predicted curves from frequentist estimation: black is true marginal mean, red is estimated marginal mean, blue is subject-specific curves

Table 4.1: Estimated degrees of freedom for smooth F and variance in ℓ_1 and ℓ_2 penalized models

	Penalty		
Estimator	ℓ_1	ℓ_2	Truth
$\hat{\text{df}}^{\text{ridge}}$	17.6	19.0	–
$\hat{\text{df}}$	9	–	–
$\hat{\sigma}^2_\epsilon$	0.0104	0.0106	0.01
$\hat{\sigma}^2_b$	0.207	1.05	1

Table 4.2: Comparison of degrees of freedom estimates for the ℓ_1 penalized model

		Smooth		
Estimator	Description	Overall	F	Z
$\hat{\text{df}}$	Stein (4.16) and (4.17)	58.6	9.00	48.6
$\tilde{\text{df}}$	Restricted (4.18) and (4.19)	58.7	9.00	48.7
$\tilde{\text{df}}^{\text{ADMM}}$	ADMM (4.20) and (4.21)	57.7	8.00	48.7
$\hat{\text{df}}^{\text{ridge}}$	Ridge (4.22) and (4.23)	67.2	17.6	48.5
$\tilde{\text{df}}^{\text{ridge}}$	Ridge restricted (4.24) and (4.25)	67.6	17.8	48.7

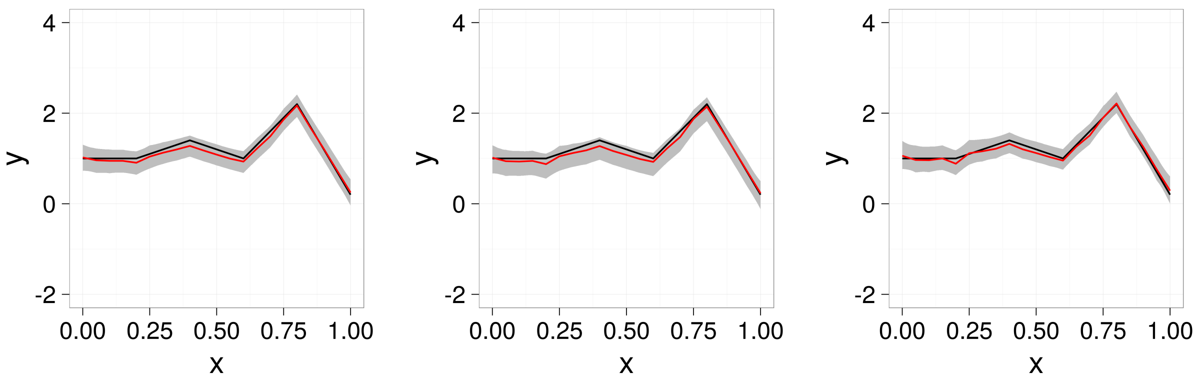
of freedom $\hat{\text{df}}^{\text{ridge}}$ appears reasonable, as it is near the estimate for the ℓ_2 penalized model. The true degrees of freedom $\hat{\text{df}}$ also seems reasonable. Ideally, the degrees of freedom should equal six, as there are four change points and we are using a second order difference penalty (see Section 4.6.2).

Table 4.2 compares the different estimates of degrees of freedom. In this simulation, the degrees of freedom based on the ridge approximation is larger than that from Stein's formula, and the approximations based on restricted derivatives are equal or near to the quantities they are estimating.

4.8.2 Bayesian Estimation

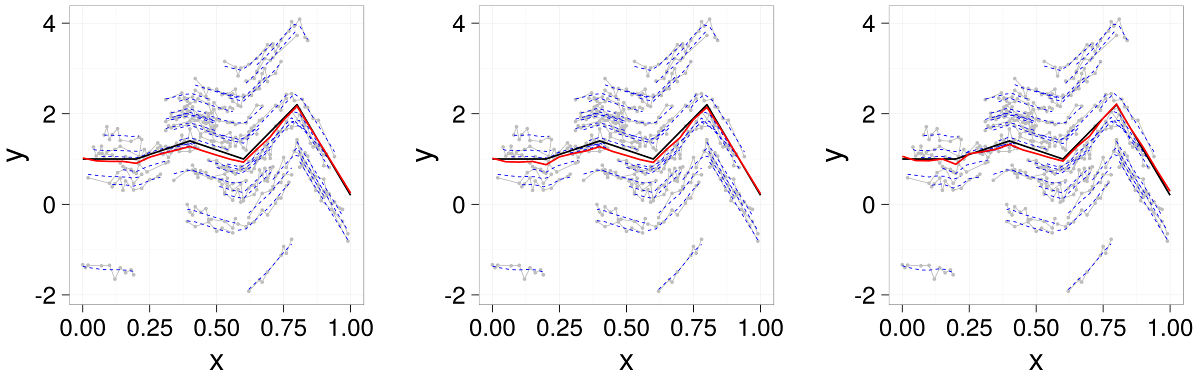
We modeled the data as $\boldsymbol{y}|\boldsymbol{b} = \beta_0\mathbf{1} + F\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\epsilon}$ where

$$\begin{aligned}
\boldsymbol{\epsilon} &\sim N(\mathbf{0}, \sigma^2_\epsilon I) \\
\boldsymbol{b} &\sim N(0, \sigma^2_b I) \\
D^{(2)}\boldsymbol{\beta} &\sim \text{Laplace}(\mathbf{0}, \sigma^2_\lambda I) \\
p(\sigma_\epsilon) &\propto 1 \\
p(\sigma_b) &\propto 1 \\
p(\log(\sigma_\lambda)) &\propto 1.
\end{aligned}$$



(a) $D\boldsymbol{\beta} \sim \text{Laplace}(\mathbf{0}, \sigma^2_\lambda I)$ (b) $D\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_\lambda I)$ (c) No prior on $D\boldsymbol{\beta}$

Figure 4.6: Credible bands for Bayesian models with order 2 (degree 1) B-splines. Black is true marginal mean, red dashed is estimated marginal mean, gray area is 95% credible interval



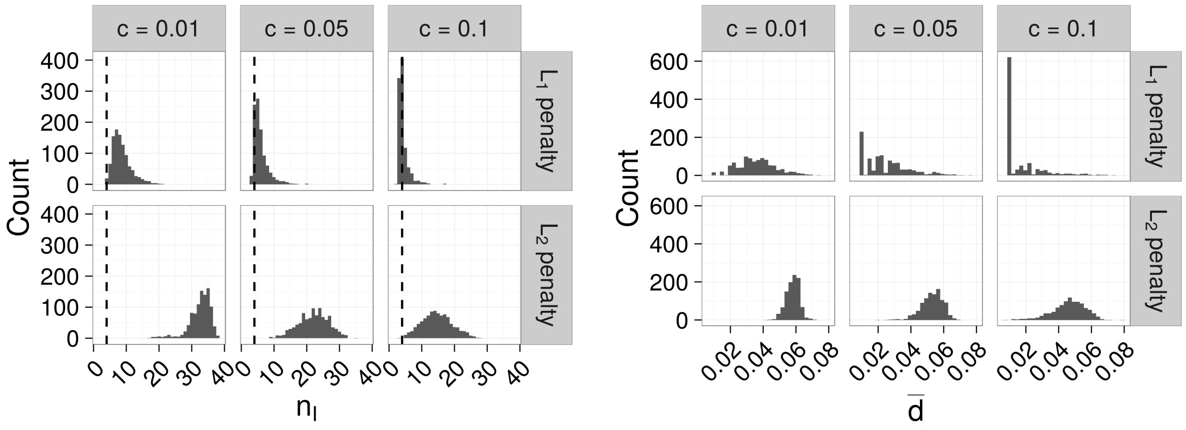
(a) $D\boldsymbol{\beta} \sim \text{Laplace}(\mathbf{0}, \sigma^2_\lambda I)$ (b) $D\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_\lambda I)$ (c) No prior on $D\boldsymbol{\beta}$

Figure 4.7: Subject-specific predicted curves from Bayesian models fit with order 2 (degree 1) B-splines. Gray is observed data, black is true marginal mean, red dashed is estimated marginal mean, and blue dashed is subject-specific predictions

We also fit models with normal and diffuse priors for $D^{(2)}\boldsymbol{\beta}$.

We fit all models with `rstan` (Stan Development Team, 2016), each with four chains of 2,000 iterations with the first 1,000 iterations of each chain used as warmup. The MCMC chains, not shown, appeared to be reasonably well mixing and stationary, and had $\hat{R}$ values under 1.1 (see Gelman et al., 2014). Figure 4.6 shows the marginal mean with 95% credible intervals, and Figure 4.7 shows point estimates.

As seen in Figures 4.6 and 4.7, all models performed well and gave similar fits as above. Similar to before, the Laplace prior appears to better enforce a piece-wise linear fit, particularly around $x = 0.2$.



(a) Number of estimated inflection points

(b) Mean absolute deviance

Figure 4.8: Results from 1,000 simulated datasets measuring ability of the models to detect inflection points

4.8.3 Change Point Detection

We simulated 1,000 datasets with the same generating mechanism used to produce the data shown in Figure 4.3 and measured the performance of the ℓ_1 and ℓ_2 penalized models on two criteria: 1) the number of inflection points found, and 2) the distance between the estimated inflection points and the closest true inflection point. To that end, let $\mathcal{T} = \{\tau_1, \ldots, \tau_4\}$ be the set of true inflection points, and $M = \max_{x \in \mathcal{X}} |\hat{f}''(x)|$ be the maximum absolute second derivative of the estimated function, where $\mathcal{X} = \{x_1, x_2, \ldots\}$ is the ordered set of unique simulated x values. We approximate $\hat{f}''$ by

$$\hat{f}''(x_i) \approx \frac{(\hat{f}(x_{i+1}) - \hat{f}(x_i))/(x_{i+1} - x_i) - (\hat{f}(x_i) - \hat{f}(x_{i-1}))/(x_i - x_{i-1})}{x_{i+1} - x_i}.$$

Then let $\mathcal{I} = \{x \in \mathcal{X} : |\hat{f}''(x)| \geq cM\}$ be the set of estimated inflection points, where $c \in (0, 1)$ is a cutoff value defining how large the second derivative must be to be counted as an inflection point. Also, let $n_{\mathcal{I}} = |\mathcal{I}|$ be the number of estimated inflection points, and $\bar{d} = n_{\mathcal{I}}^{-1} \sum_{x \in \mathcal{I}} \min_{\tau \in \mathcal{T}} |x - \tau|$ be the mean absolute deviance of the estimated inflection points.

Figure 4.8 shows the results from 1,000 simulated datasets. The ℓ_1 penalized model was better able to 1) find the correct number of inflection points, and 2) determine the location of the inflection points.

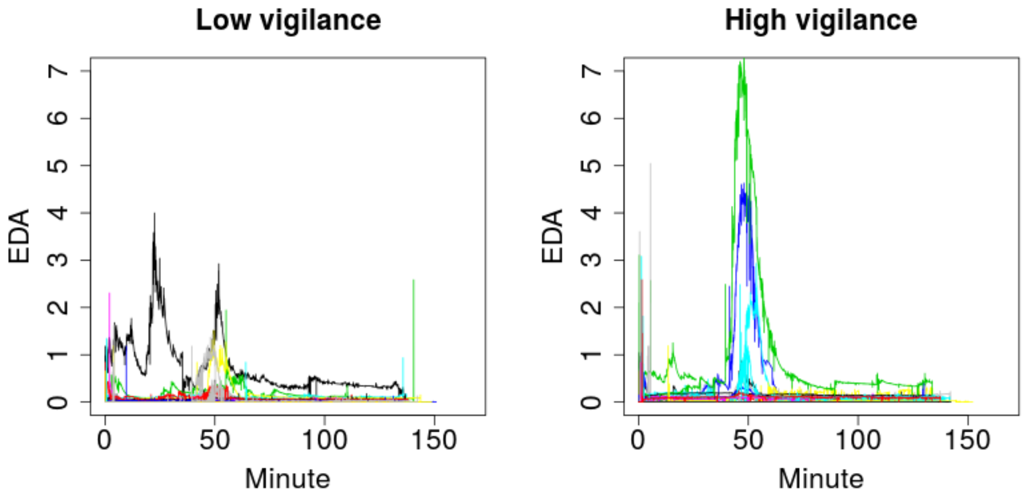


Figure 4.9: Raw electrodermal activity (EDA) data by experimental group

4.9 Application

4.9.1 Data Description and Preparation

In this section, we analyze electrodermal activity (EDA) data collected as part of a stress study. In brief, all subjects completed a written questionnaire prior to the study, which categorized the subjects as having either low vigilance or high vigilance personality types. During the study, all participants wore wristbands that collected EDA levels while undergoing stress-inducing activities, including giving a public speech and performing mental arithmetic in front of an audience. The scientific questions were: 1) Is EDA higher among high vigilance subjects, and 2) when did trends in stress levels change? In this section, we demonstrate how our methods could address both questions.

The raw EDA data are shown in Figure 4.9. After excluding subjects who had EDA measurements of essentially zero throughout the entire study, we were left with ten high vigilance subjects and seven low vigilance subjects.

To remove the extreme second-by-second fluctuations in EDA, which we believe are artifacts of the measurement process as opposed to real biological signals, we smoothed each curve separately with a Nadaraya–Watson kernel estimator using the `ksmooth` function in R. We then thinned the data to reduce computational burden, taking 100 evenly spaced measurements from each subject. Figure 4.10 shows the results of this process for a single subject, and Figure 4.11 shows the prepared data for all subjects. Because of the limited number of subjects, as well as issues of misalignment in the time series across individuals,

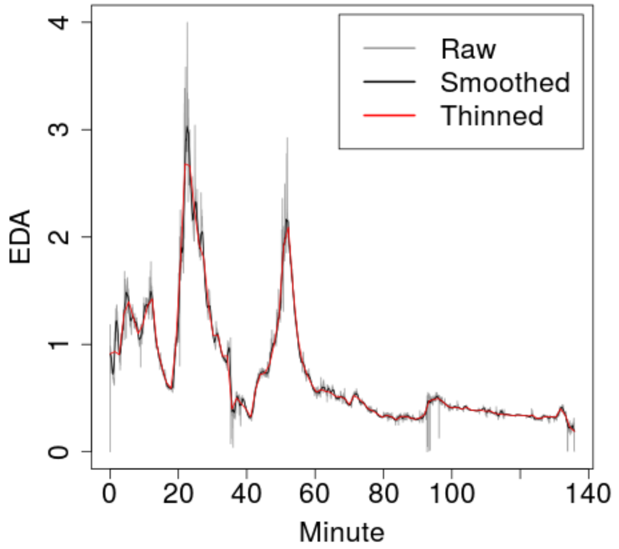


Figure 4.10: Raw, smoothed, and thinned electrodermal activity data for a single subject

the results presented here should be considered as illustrative rather than of full scientific validity.

4.9.2 Models

In all models, we fit the general structure

$$y_i(x) = \beta_0 + \beta_1(x) + \mathbb{1}_{\text{high}}[i]\beta_2(x) + b_i(x) + \epsilon_i(x)$$

where x represents time in minutes, $\mathbb{1}_{\text{high}}[i] = 1$ if subject i has high vigilance and $\mathbb{1}_{\text{high}}[i] = 0$ if subject i has low vigilance, $b_i(x)$ are random curves, and $\epsilon_i(x) \sim N(0, \sigma_\epsilon^2)$. For $\beta_1(x)$, $\beta_2(x)$, and $b_i(x)$, we used a fourth order B-spline basis with 31 basis functions each, and a second order difference penalty ($k = 1$).

Written in matrix notation, the ℓ_1 penalized model is

$$\min \frac{1}{2}\|\boldsymbol{y} - \beta_0\mathbf{1} - \sum_{j=1}^{2} F_j\boldsymbol{\beta}_j - Z\boldsymbol{b}\|_2^2 + \sum_{j=1}^{2} \lambda_j\|D^{(2)}\boldsymbol{\beta}_j\|_1 + \boldsymbol{b}^T S\boldsymbol{b} \tag{4.41}$$

where $\boldsymbol{y}$ is a stacked vector for subjects $i = 1, \ldots, 17$, F_1 is an $n \times p$ design matrix where $n = 1,700$ and $p = 31$, and $F_2 = \text{diag}(\mathbb{1}_{\text{high}}[\boldsymbol{i}])F_1$ where $\boldsymbol{i}$ is an $n \times 1$ vector of subject IDs. In other words, F_2 is equal to F_1, but with rows corresponding to low vigilance subjects zeroed

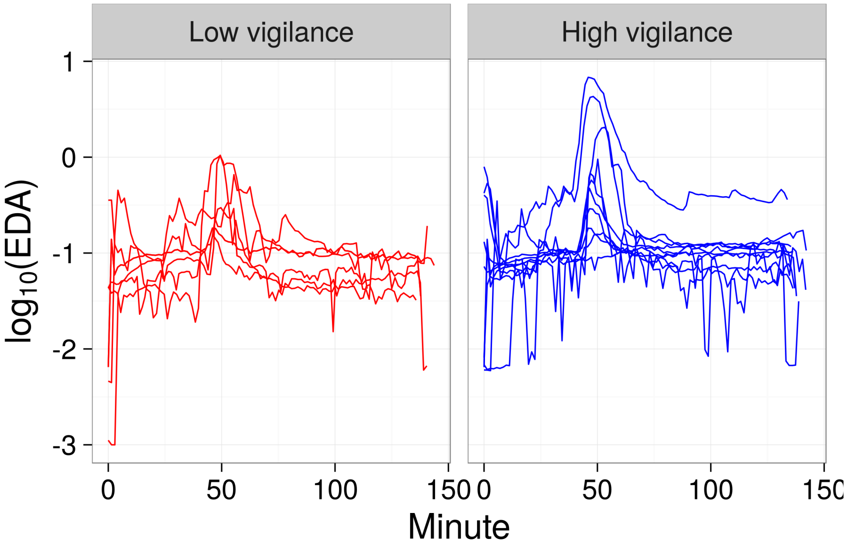


Figure 4.11: Electrodermal activity (EDA) data used in the analysis (seven low vigilance and ten high vigilance subjects). Note: subjects not aligned in time (x-axis).

out. We set

$$Z = \begin{bmatrix} Z_1 & & \\ & \ddots & \\ & & Z_{17} \end{bmatrix}$$

where each Z_i is an $n_i \times 31$ random effects design matrix of order 4 B-splines evaluated at the input points for subject i, and

$$S = \begin{bmatrix} S_1 & & \\ & \ddots & \\ & & S_{17} \end{bmatrix}$$

where $S_{i,jl} = \int \phi''_{ij}(t)\phi''_{il}(t)dt$ are smoothing spline penalty matrices. We also mean-centered F_1 and Z as described in Section 4.3, with the corresponding changes in dimensions.

To fit a comparable ℓ_2 penalized model, in which $\lambda_j \|D^{(2)}\boldsymbol{\beta}_j\|_1$ in (4.41) is replaced with $(\lambda_j/2)\|D^{(2)}\boldsymbol{\beta}_j\|_2^2$, we rotated the random effect design and penalty matrices Z and S as described in Section 4.3. To facilitate the use of existing software, we used a normal prior for the "unpenalized" random effect coefficients, i.e. $\breve{\boldsymbol{b}}_f \sim N(\mathbf{0}, \sigma_f^2 I)$.

We also fit a Bayesian model using the same rotations and equivalent penalties as above.

In particular, we modeled the data as $\boldsymbol{y}|\boldsymbol{b} = \beta_0\mathbf{1} + \sum_{j=1}^{J} F_j\boldsymbol{\beta}_j + \breve{Z}_r\breve{\boldsymbol{b}}_r + \breve{Z}_f\breve{\boldsymbol{b}}_f + \boldsymbol{\epsilon}$ where

$$\begin{aligned}
\breve{\boldsymbol{b}}_r &\sim N(\mathbf{0}, \sigma_r^2 I) \\
\breve{\boldsymbol{b}}_f &\sim N(\mathbf{0}, \sigma_f^2 I) \\
(D_j\boldsymbol{\beta}_j)_l &\sim \text{Laplace}(0, a_j) \text{ for } a_j = \sigma_\epsilon^2/(2\lambda_j), l = 1, \dots, p_j - k_j - 1, j = 1, \dots, J \\
\boldsymbol{\epsilon} &\sim N\left(\mathbf{0}, \sigma_\epsilon^2 I\right).
\end{aligned}$$

4.9.3 Results

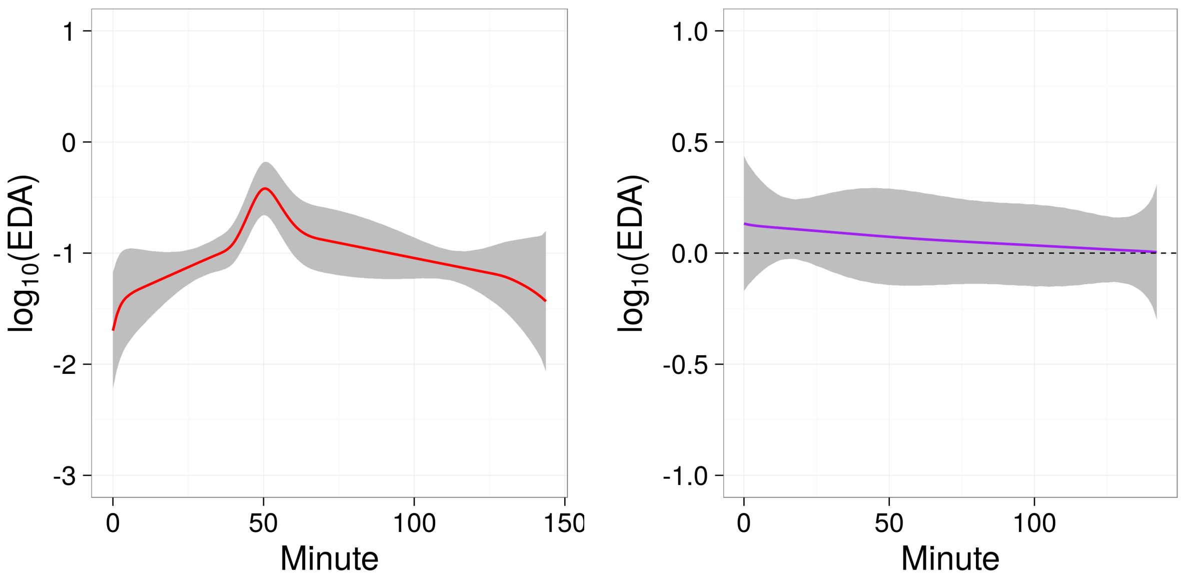
Frequentist Estimation

We tried to use CV to estimate the smoothing parameters for the ℓ_1 penalized model. However, with only 17 subjects split between two groups, we only did 3-fold CV. CV did not find a visually reasonable fit so we set the tuning parameters by hand.

Figure 4.12 shows the estimated marginal mean and 95% credible bands for the ℓ_1 penalized model, and Figure 4.13 shows the subject-specific predicted curves for the ℓ_1 penalized model. As seen in Figure 4.12a, our model identified a few inflection points, particularly near minutes 40, 50, and 60. From Figure 4.12b it appears that the difference in EDA levels between the low and high vigilance subjects was not statistically significant. Also, as seen in Figure 4.13, the subject-specific predicted curves are shrunk towards the mean, which is expected, because the predicted curves are analogous to best linear unbiased predictors (BLUPs), although they are not linear smoothers.

Figure 4.14 shows the estimated marginal mean and 95% credible bands for the ℓ_2 penalized model, and Figure 4.15 shows the subject-specific predicted curves for the ℓ_2 penalized model. The estimate shown in Figure 4.14a is similar to that shown in Figure 4.12a, though the inflection points are slightly less pronounced in Figure 4.14a. The results in Figure 4.14b are for the most part substantively the same as those in Figure 4.12b; the ℓ_2 penalized model does not show a statistically significant difference between the low and high vigilance subjects, with the possible exception of minutes 44 to 67. As seen in Figure 4.15, the predicted subject-specific curves from the ℓ_2 penalized model are also shrunk towards the mean.

Table 4.3 shows the estimated degrees of freedom for the ℓ_1 penalized model. Similar to the simulation, the restricted derivate approximations tend to be near the quantities they are estimating. In the ℓ_2 penalized model, smooth F_1 had 14.2 degrees of freedom, and smooth F_2 had 6.96 degrees of freedom.



(a) $\hat{\beta}_1(x)$ (low vigilance)

(b) $\hat{\beta}_2(x)$ (high – low vigilance)

Figure 4.12: ℓ_1 penalized model: parameter estimates with 95% confidence bands

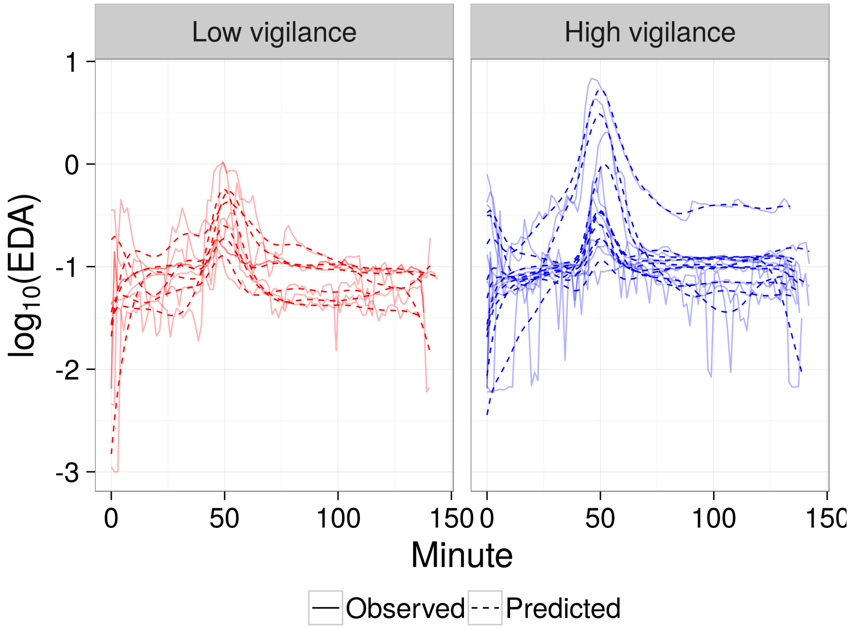
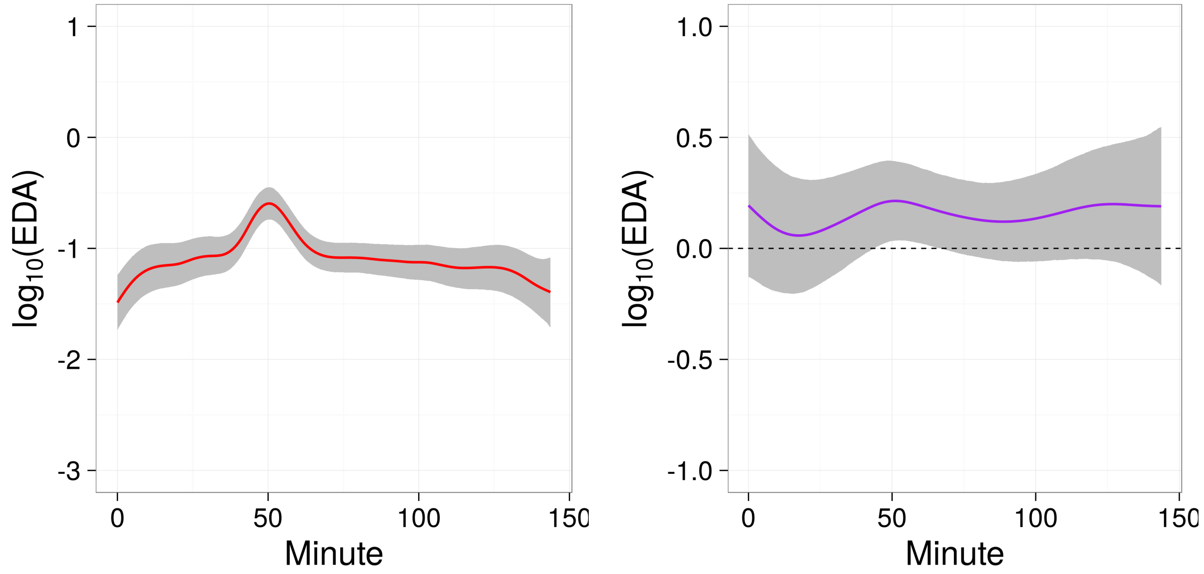


Figure 4.13: ℓ_1 penalized model: subject-specific predicted curves



(a) $\hat{\beta}_1(x)$ (low vigilance)

(b) $\hat{\beta}_2(x)$ (high − low vigilance)

Figure 4.14: ℓ_2 penalized model: parameter estimates with 95% confidence bands

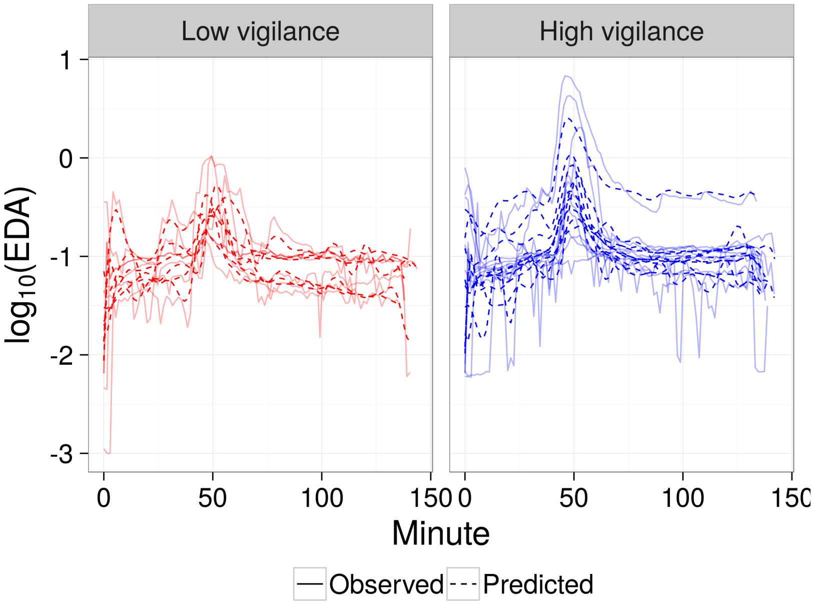
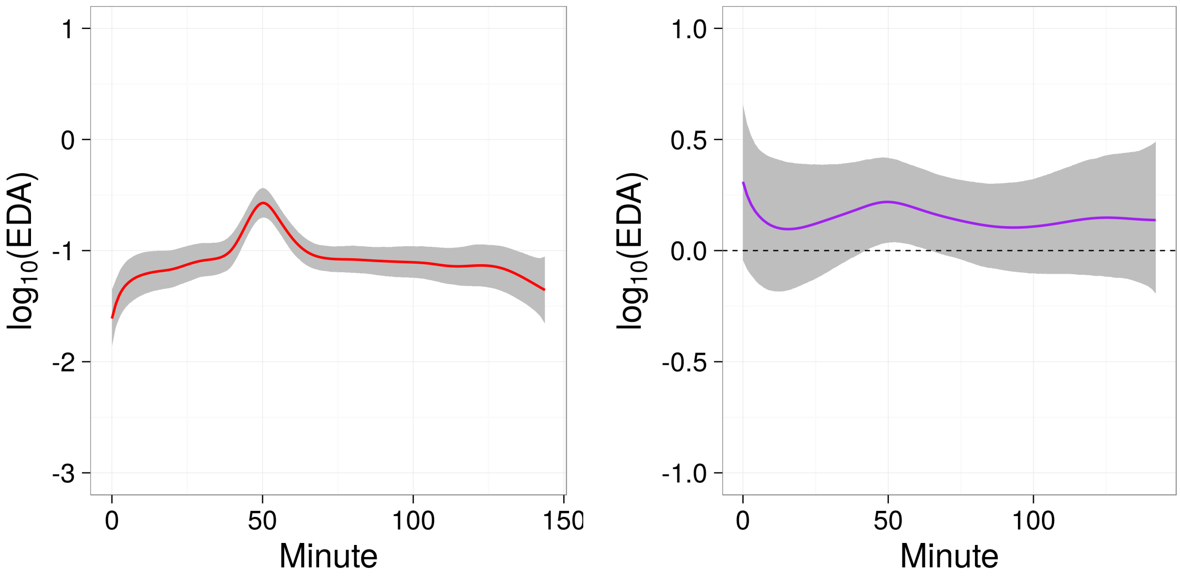


Figure 4.15: ℓ_2 penalized model: subject-specific predicted curves

Table 4.3: Comparison of degrees of freedom estimates for the ℓ_1 penalized model

		Smooth			
Estimator	Description	Overall	F_1	F_2	Z
$\hat{\mathrm{df}}$	Stein (4.16) and (4.17)	188	10.0	2.00	175
$\tilde{\mathrm{df}}$	Restricted (4.18) and (4.19)	193	10.0	2.00	180
$\tilde{\mathrm{df}}^{\mathrm{ADMM}}$	ADMM (4.20) and (4.21)	192	9.00	2.00	180
$\hat{\mathrm{df}}^{\mathrm{ridge}}$	Ridge (4.22) and (4.23)	196	19.5	8.09	167
$\tilde{\mathrm{df}}^{\mathrm{ridge}}$	Ridge restricted (4.24) and (4.25)	215	21.1	13.50	180



(a) $\hat{\beta}_1(x)$ (low vigilance subjects)

(b) $\hat{\beta}_2(x)$ (high – low vigilance subjects)

Figure 4.16: Bayesian model: parameter estimates with 95% confidence bands

Bayesian Estimation

We fit the model using `rstan` (Stan Development Team, 2016) with four chains of 5,000 iterations each, with the first 2,500 iterations of each chain used as warmup. The MCMC chains, not shown, appeared to be reasonably well mixing and stationary with $\hat{R}$ values under 1.1 (see Gelman et al., 2014). Figure 4.16 shows the marginal means with 95% credible intervals, and Figure 4.17 shows the subject-specific curves. Similar to the ℓ_2 penalized model, the Bayesian model found a slightly statistically significant difference between low and high vigilance between minutes 42 and 65.

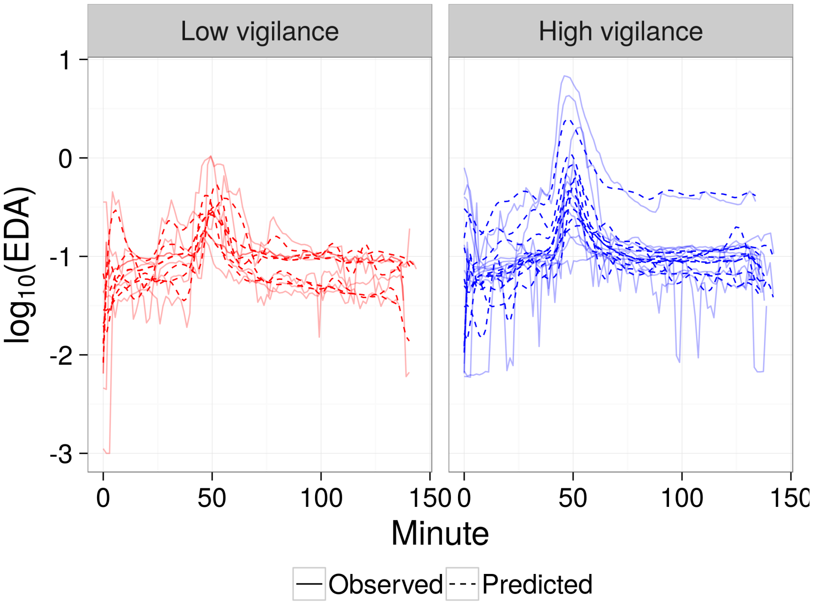


Figure 4.17: Bayesian model: subject-specific predicted curves

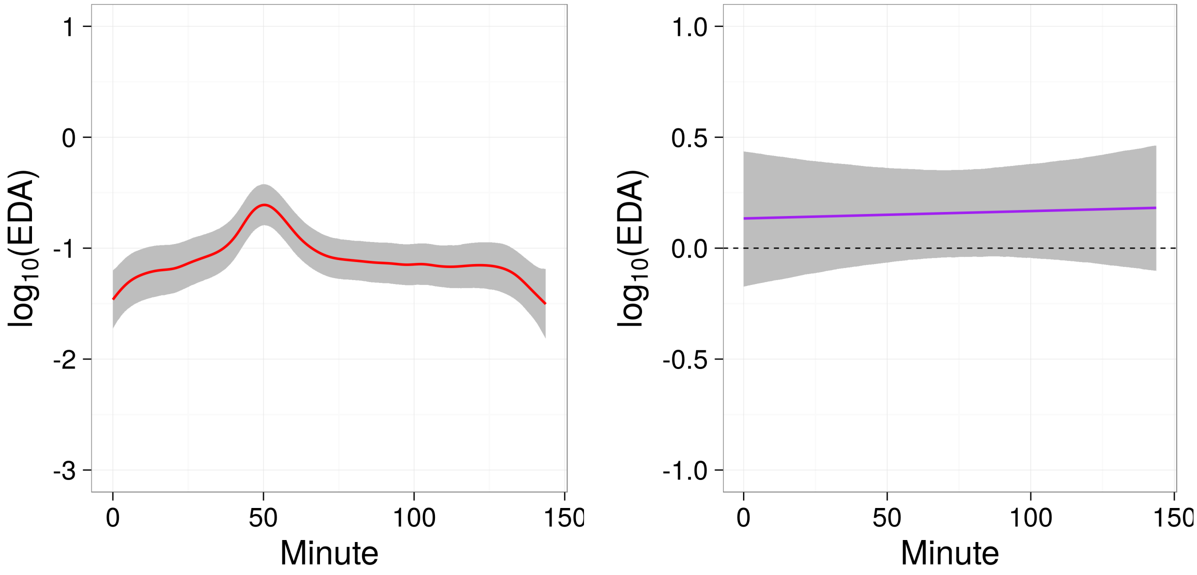
4.9.4 ℓ_2 Penalized Model with Alternative Correlation Structure

For comparison, we also fit an ℓ_2 penalized model with an alternative correlation structure similar to that recommended by Ruppert et al. (2003, p. 192). In place of the correlation structure implied by the penalty matrix S described above, we augmented each Z_i matrix on the left with the columns $[\mathbf{1}, \boldsymbol{x}_i]$, where $\boldsymbol{x}_i$ is an $n_i \times 1$ vector of measurement times for subject i. We then replaced $Z_i\boldsymbol{b}_i$ with $[\mathbf{1}, \boldsymbol{x}_i, Z_i](\boldsymbol{u}_i^T, \boldsymbol{b}_i^T)^T$, and assumed $(\boldsymbol{u}_i^T, \boldsymbol{b}_i^T)^T \sim N(0, \Sigma_i)$ where

$$\Sigma_i = \begin{bmatrix} \Sigma' & \\ & \sigma_b^2 I \end{bmatrix}$$

and Σ' is a common 2×2 unstructured positive definite matrix. To model the within-subject correlations, we used a continuous autoregressive process of order 1. In particular, $\mathrm{Cor}(y_i(x_{ij}), y_i(x_{ij'})) = \zeta^{|x_{ij} - x_{ij'}|}$ for a common parameter $\zeta > 0$.

Figure 4.18 shows the estimated marginal mean and 95% credible bands, and Figure 4.15 shows the subject-specific predicted curves. The estimates shown in Figure 4.18 are similar to that shown in Figure 4.14. While estimates of the difference between low and high vigilance subjects differs between this model and the ℓ_2 penalized model in Section 4.9.3, the more notable difference is in the subject-specific predicted curves. As seen in Figure 4.19, the predicted subject-specific curves are not shrunk towards the mean as much as in Figure 4.15.



(a) $\hat{\beta}_1(x)$ (low vigilance subjects)

(b) $\hat{\beta}_2(x)$ (high – low vigilance subjects)

Figure 4.18: ℓ_2 penalized model with alternative correlation structure: parameter estimates with 95% confidence bands

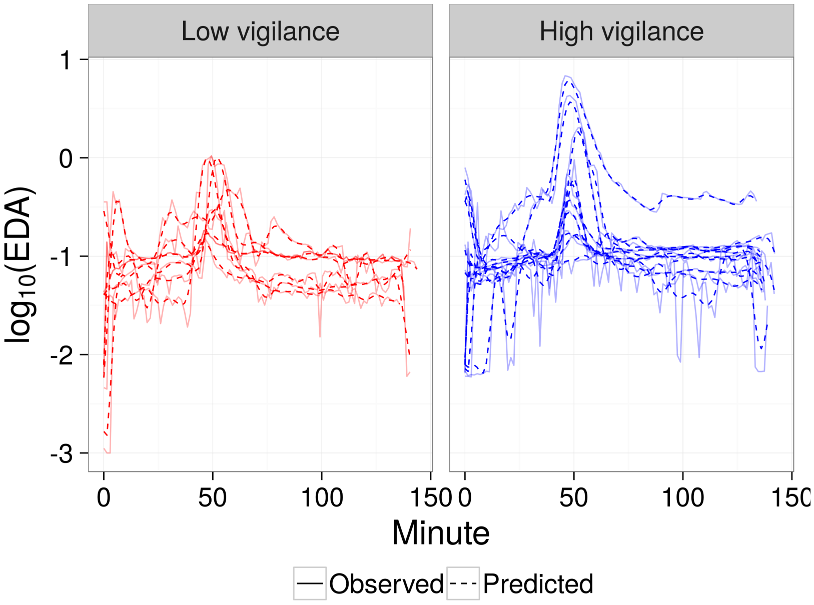


Figure 4.19: ℓ_2 penalized model with alternative correlation structure: subject-specific predicted curves

4.10 Discussion and Potential Extensions

As demonstrated in this chapter, P-splines with an ℓ_1 penalty can be useful for analyzing repeated measures data. Compared to related work with ℓ_1 penalties, our model is ambitious in that we allow for multiple smoothing parameters and propose approximate inferential procedures that do not require Bayesian estimation. However, these are also the two aspects of our proposed approach that require additional future work.

Regarding estimation, our current approach of using ADMM and CV appears to work reasonably well for $J = 1$ smooth, but is not yet reliable for $J > 1$ smooths. In the future, we plan to develop more robust estimation techniques, particularly for smoothing parameters. As one possibility, we have done preliminary work to minimize quantities similar to GCV and AIC instead of the more computationally intensive CV, though these approaches do not seem as promising as their ℓ_2 counterparts. When possible, Bayesian estimation may be the most reliable way to currently fit these models. Bayesian estimation also opens the possibility of using other sparsity inducing priors, such as spike and slab models (Ishwaran and Rao, 2005).

Regarding inference, in future work we plan to use the $\boldsymbol{\delta}$ quantity to bound difference between ℓ_1 and ℓ_2 penalized fits under certain assumptions on the data (see Section 4.7.2), and to study coverage probability through simulations. We also plan to investigate the use of post-selection inference methods to develop confidence bands for linear combinations of the active set, and to further investigate through simulations the performance of our proposed estimates of degrees of freedom. However, we note that our primary use of the degrees of freedom estimate $\hat{\text{df}}$ is to obtain the residual degrees of freedom $\hat{\text{df}}_{\text{resid}} = n - \hat{\text{df}}$, which we then use to estimate the variances $\hat{\sigma}_\epsilon^2 = \|\boldsymbol{r}\|_2^2 / \hat{\text{df}}_{\text{resid}}$. Therefore, when $n \gg \hat{\text{df}}$, $\hat{\sigma}_\epsilon^2$ is not very sensitive to $\hat{\text{df}}$, in which case it is not critical for our purposes to obtain an exact estimate of degrees of freedom.

In addition, we think it could be beneficial to investigate the addition of random effects to locally adaptive regression splines, and to implement smoothing splines with an ℓ_1 penalty (replacing the $\|D\boldsymbol{\beta}\|_1$ penalty with $\|\Psi\boldsymbol{\beta}\|_1$ where $\Psi_{ij} = \int \phi_i''(s)\phi_j''(s)ds$ and ϕ are basis functions). We also plan to extend these results to a generalized model to allow for non-normal response distributions.

Regarding the rate of convergence, from Observation 4.1 and the work of Tibshirani (2014a), we know that for equally spaced data and $F = I_{n\times n}$, P-splines with an ℓ_1 penalty achieve the minimax rate of convergence for the class of weakly differentiable functions of bounded variation. Let $\|A\|_{\max} = \max_{ij} |a_{ij}|$ be the maximum element of a matrix A. We speculate that for a general $n \times p$ design matrix F of full rank (not necessarily of first degree B-splines), ℓ_1 penalized models achieve the minimax rate of convergence when $\|FF^T - I_{n\times n}\|_{\max}$ is small, but not when it is large. Proving this assertion and finding a cutoff value has been challenging. The framework of Tibshirani (2014a) for comparing the fits from two lasso problems with different design matrices may be promising, but extending the results of Tibshirani (2014a) to design matrices of different dimensions has been difficult, and would be required in our setting. It may also be possible to build on results regarding the optimal number and placement of spline knots. We leave this for future work.

4.11 R Package and Code

We have implemented our method in the R package `pSplinesL1` available at `https://github.com/bdsegal/psplinesl1`. All code for the simulations and analyses in this chapter are available at `https://github.com/bdsegal/code-for-psplinesl1-paper`.

Appendix A

Big Five Questionnaire Items

As described by Smith et al. (2013), selected respondents to the 2010 Health and Retirement Survey were asked to rate how well 31 items described them on the following four point scale: 1) A lot, 2) Some, 3) A little, 4) Not at all.

The items were as follows (letters match those shown in Figure 3.1): a) Outgoing, b) Helpful, c) Reckless, d) Moody, e) Organized, f) Friendly, g) Warm, h) Worrying, i) Responsible, j) Lively, k) Caring, l) Nervous, m) Creative, n) Hardworking, o) Imaginative, p) Softhearted, q) Calm, r) Self-disciplined, s) Intelligent, t) Curious, u) Active, v) Careless, w) Broad-minded, x) Impulsive, y) Sympathetic, z) Cautious, z2) Talkative, z3) Sophisticated, z4) Adventurous, z5) Thorough, and z6) Thrifty.

The items were grouped into five sub-dimensions:

1. Neuroticism: d, h, l, q

2. Extroversion: a, f, j, u, z2

3. Agreeableness: b, g, k, p, y

4. Openness to experience: m, o, s, t, w, z3, z4

5. Conscientiousness: c, e, i, n, r, v, x, z, z5, z6

All but c, q, v, and x were reverse coded.

Bibliography

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, 42:815–824.

Bartra, O., McGuire, J. T., and Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, 76:412–427.

Beauducel, A. and Wittmann, W. (2005). Simulation study on fit indices in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12(1):41–75.

Becker, M. and Klößner, S. (2016). *PearsonDS: Pearson Distribution System.* R package version 0.98.

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

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, 107(2):238–246.

Bock, R. D. and Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika*, 31(4):507–534.

Bollaerts, K., Eilers, P. H. C., and Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the L1-norm. *Statistical Modelling*, 6(3):189–207.

Booth, J. G. and Butler, R. W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika*, 77(4):787–796.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453 – 510.

Butler, R. W. (2007). *Saddlepoint approximations with applications.* Cambridge University Press, Cambridge, UK.

Casella, G. and Berger, R. L. (2002). *Statistical inference.* Duxbury, Pacific Grove, CA, 2nd edition.

Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics*, 67(3):861–870.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome biology*, 4(210):1–10.

De Boor, C. (2001). *A practical guide to splines.* Springer, New York, NY, revised edition.

Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–294.

Donoho, D. L. and Johnstone, I. M. (1988). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879 – 921.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, 81(394):461–470.

Eilers, P. H. C. (2000). Robust and quantile smoothing with P-splines and the L1 norm. In *Proceedings of the 15th International Workshop on Statistical Modelling, Bilbao.*

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11(2):89–121.

Fan, X. and Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: rationale of two-index strategy revisited. *Structural Equation Modeling*, 12(3):343–367.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):175–185.

Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis.* Chapman and Hall/CRC, Boca Raton, FL.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis.* Chapman and Hall/CRC, Boca Raton, FL, 3rd edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Good, P. (2000). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, New York, NY, 2nd edition.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55(3):245 – 259.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1):121–128.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics*, 32(2):506–523.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3):107–145.

Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4):1–13.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1:297–318.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1st edition.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55(4):757–796.

Hawkins, D. L. (1989). Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. *The American Statistician*, 43(4):235–237.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hooper, D., Coughlan, J., and Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, 6(1):53–60.

HRS (2016). Health and retirement study (core data release) public use dataset.

Hu, L.-T. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1):1–55.

Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British journal of mathematical and statistical psychology*, 29(2):190–241.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730 – 773.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice Hall, Englewood Cliffs, NJ.

Janson, L., Fithian, W., and Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika*, pages 1–8.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem. *Statistics & probability letters*, 36(1):9–21.

Jiang, H. and Salzman, J. (2012). Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika*, 99(4):973–980.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2. John Wiley & Sons, New York, NY, 2nd edition.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4):443–477.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM review*, 51(2):339–360.

Kimmel, G. and Shamir, R. (2006). A fast method for computing high-significance disease association in large population-based studies. *The American Journal of Human Genetics*, 79(3):481–492.

Klar, B. (2015). A note on gamma difference distributions. *Journal of Statistical Computation and Simulation*, 85(18):3708–3715.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* Guilford Press, New York, NY, 3rd edition.

Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., and Shmulevich, I. (2009). Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168.

Leemis, L. M. and McQueston, J. T. (2008). Univariate distribution relationships. *The American Statistician*, 62(1):45–53.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco, CA.

Lehmann, E. L. (1999). *Elements of large-sample theory.* Springer Science & Business Media, New York, NY.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* Springer Science & Business Media, New York, NY, 3rd edition.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(323):1–16.

Li, Q., Zheng, G., Li, Z., and Yu, K. (2008). Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies. *Annals of Human Genetics*, 72(3):397–406.

Liang, F., Liu, C., and Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, 102(477):305–320.

Lin, Y., Zhang, H. H., et al. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297.

Lou, Y., Bien, J., Caruana, R., and Gehrke, J. (2016). Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4).

Lugannani, R. and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in applied probability*, 12(02):475–490.

Mammen, E., van de Geer, S., et al. (1997). Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209–220.

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentlers (1999) findings. *Structural Equation Modeling*, 11(3):320–341.

Mathai, A. M. (1993). On noncentral generalized laplacianness of quadratic forms in normal variables. *Journal of Multivariate Analysis*, 45(2):239–246.

McDonald, R. P. (1974). Testing pattern hypotheses for covariance matrices. *Psychometrika*, 39(2):189–201.

Mehta, C. R. and Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in r×c contingency tables. *Journal of the American Statistical Association*, 78(382):427–434.

Meier, L., Van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747.

National Cancer Institute (2015). The cancer genome atlas.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 12(4):1760–1779.

Nichols, T. E. and Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25.

Olkin, I. and Finn, J. (1990). Testing correlated correlations. *Psychological Bulletin*, 108(2):330–333.

Olkin, I. and Finn, J. (1995). Correlations redux. *Psychological Bulletin*, 118(1):155–164.

Pahl, R. and Schäfer, H. (2010). PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics*, 26(17):2093–2100.

Pearson, K. and Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. iv. on the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical transactions of the Royal Society of London (A)*, 191:229–311.

Petersen, A., Witten, D., and Simon, N. (2016). Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025.

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M. N., Replogle, J. M., Feng, T., Lee, M., Asinovski, N., and Frohlich, I. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, 344(6183):519–523.

Ramdas, A. and Tibshirani, R. J. (2016). Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858.

Ravikumar, P., Lafferty, J. D., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5).

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1):253–259.

Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):91–101.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, New York, NY.

Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):447–501.

Simpson, S., Lyday, R., Hayasaka, S., Marsh, A., and Laurienti, P. (2013). A permutation testing framework to compare groups of brain networks. *Frontiers in Computational Neuroscience*, 7(171):1–11.

Smith, J., Fisher, G., Ryan, L., Clarke, P., House, J., and Weir, D. (2013). Psychosocial and lifestyle questionnaire 2006–2010: Documentation report core section LB. Technical report, University of Michigan.

Speed, T. (1991). Comment on “That BLUP is a good thing: The estimation of random effects”. *Statistical science*, 6(1):42–44.

Srivastava, J. N. (1966). On testing hypotheses regarding a class of covariance structures. *Psychometrika*, 31(2):147–164.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1.

Steiger, J. H. (1980a). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research*, 15(3):335–352.

Steiger, J. H. (1980b). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245–251.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42(5):893–898.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135 – 1151.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S. E., and Tavaré, S. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics*, 1(6):695–704.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., and Lee, C. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267 – 288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. J. (2014a). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Tibshirani, R. J. (2014b). Supplement to “adaptive piecewise polynomial estimation via trend filtering”.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, (2):1198–1232.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). Mapsplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178–e178.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):159–174.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2).

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, New York, NY.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, Boca Raton, FL.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548 – 1575.

Yu, K., Liang, F., Ciampa, J., and Chatterjee, N. (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics*, 12(3):582–593.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, 40(1):115–148.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49 – 67.

Zaki, M. J. and Jr., W. M. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY.

Zhan, C., Yan, L., Wang, L., Sun, Y., Wang, X., Lin, Z., Zhang, Y., Shi, Y., Jiang, W., and Wang, Q. (2015). Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma. *Journal of Thoracic Disease*, 7(8):1398–1405.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93(442):710 – 719.

Zhang, Y. and Liu, J. S. (2011). Fast and accurate approximation to significance tests in genome-wide association studies. *Journal of the American Statistical Association*, 106(495):846–857.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468 – 3497.

Zhou, Y.-H. (2014). *mcc: Moment Corrected Correlation*. R package version 1.0.

Zhou, Y.-H. and Wright, F. A. (2015). Hypothesis testing at the extremes: fast and robust association for high-throughput data. *Biostatistics*, 16(3):611–625.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301 – 320.