Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Advances in Nonasymptotic and Nonparametric Inference

Presented by: Robin Dunn

Accepted by: Department of Statistics & Data Science

Readers:

Aaditya Ramdas, Co-chair

Larry Wasserman, Co-chair

Sivaraman Balakrishnan

Ryan Martin, North Carolina State University

Susan Murphy, Harvard University

Approved by the Committee on Graduate Degrees:

Richard Scheines, Dean   Date

Carnegie Mellon University

Advances in Nonasymptotic and Nonparametric Inference

A Dissertation Submitted to the Graduate School

in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics

by

Robin Dunn

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

July 2021

Abstract

This thesis develops tools for hypothesis testing and predictive inference in nonasymptotic settings. The universal likelihood ratio test (LRT) constructs hypothesis tests that are valid in finite samples and without regularity conditions. We implement the universal LRT to test the population mean of d-dimensional Gaussian data and to test whether a density satisfies the nonparametric shape constraint of log-concavity. Conformal predictive inference produces valid prediction sets in finite samples without model assumptions, in the case where the data are exchangeable. We extend conformal prediction to the random effects setting.

The LRT based on the asymptotic chi-squared distribution of the log likelihood is one of the fundamental tools of statistical inference. A recent universal LRT approach based on sample splitting provides valid hypothesis tests and confidence sets in any setting for which we can compute the split likelihood ratio statistic (or, more generally, an upper bound on the null maximum likelihood). This test empowers statisticians to construct tests in settings for which no valid hypothesis test previously existed. Chapter 1 explains the universal LRT.

In Chapter 2, we consider the simple but fundamental case of testing the population mean of d-dimensional Gaussian data. This work presents the first in-depth exploration of the size, power, and relationships between several universal LRT variants. We show that a repeated subsampling approach is the best choice in terms of size and power. We observe reasonable performance even in a high-dimensional setting. We illustrate the benefits of the universal LRT through testing a non-convex doughnut-shaped null hypothesis, where a universal inference procedure can have higher power than a standard approach.

Chapter 3 investigates the use of universal LRTs to test whether a density is log-concave. The shape constraint of log-concavity imposes a nonparametric density estimation problem with favorable convergence properties. We propose and implement several universal LRT variants for this test. This provides the first test of log-concavity with finite sample validity. We evaluate the universal LRT to test log-concavity on two-component Gaussian mixture models and on the Beta family. We find that universal LRTs that convert the d-dimensional testing problem to a one-dimensional testing problem can have the best performance.

Chapter 4 reviews the method of conformal predictive inference. Conformal prediction methods construct valid prediction sets in finite samples even when the assumed model is incorrect, under the assumption that the data are exchangeable. In Chapter 5, we extend the conformal method so that it is valid with random effects, in which case the data are not exchangeable. We develop a CDF pooling approach, a single subsampling approach, and a repeated subsampling approach to construct conformal prediction sets in unsupervised and supervised settings. We compare these approaches in terms of coverage and average set size. We recommend the repeated subsampling approach that constructs a conformal set by sampling one observation from each distribution multiple times. Simulations show that this approach has the best balance between coverage and average conformal set size.

Contents

List of Tables

List of Figures

Part I

Universal Likelihood Ratio Testing

Chapter 1

Overview of Universal Likelihood Ratio Testing

Hypothesis testing is one of the primary tools that statisticians use to draw conclusions in data-driven investigations. As a few examples, statisticians use hypothesis tests to evaluate the outcomes of policies, to compare the effectiveness of treatments, and to understand properties of the distribution from which data arise. For practitioners to trust the results of a hypothesis test, the test should have guaranteed validity (or at least approximate validity). That is, the probability of falsely rejecting a null hypothesis (type I error level) should be no more than a pre-specified value α.

The likelihood ratio test (LRT) provides one common framework for testing statistical hypotheses. Classical approaches to likelihood ratio testing depend on an asymptotic χ^2 approximation to the log likelihood ratio. The statistical literature has repeatedly emphasized that the asymptotic χ^2 approximation may not be valid in small sample settings. Examples include Bartlett (1937), Lehmann (2012), and Medeiros and Ferrari (2017). Small sample sizes pose a recurrent problem across biological science research. For instance, researchers have noted the prevalence of low-powered studies in neuroscience (Button et al., 2013) and the need for clinical trial designs that account for the small sample sizes common to rare disease and pediatric population research (Ildstad et al., 2001; McMahon et al., 2016).

The universal inference approach developed by Wasserman et al. (2020) provides a new likelihood ratio testing framework that addresses situations where the classical LRT is not valid.

This new LRT relies on sample splitting to construct a test and confidence set that are valid in finite samples and without regularity conditions. This universal inference method allows one to construct valid tests in settings for which no hypothesis test with type I error control and finite sample guarantees previously existed.

First, we review how the classical LRT fits into a parametric hypothesis testing framework. Suppose we have data from an unknown distribution P_{θ^*} which belongs to some set of distributions $(P_\theta : \theta \in \Theta)$. Assume $\Theta_0 \subset \Theta$. We wish to test the composite null hypothesis $H_0 : \theta^* \in \Theta_0$. We use the observed data to construct a test statistic T_n and reject H_0 if $T_n > c_\alpha$, where c_α must satisfy

$$\sup_{\theta^* \in \Theta_0} P_{\theta^*}(T_n > c_\alpha) \leq \alpha.$$

Consider, for example, the alternative $H_1 : \theta \in \Theta \setminus \Theta_0$. The generalized likelihood ratio statistic is $\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)$, where $\widehat{\theta}$ is the maximum likelihood estimate (MLE) in Θ and $\widehat{\theta}_0$ is the MLE in Θ_0. We reject H_0 when $2\log\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)\} > c_{\alpha,d}$, where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution and $d = df(\Theta) - df(\Theta_0)$. This construction arises from Wilks' Theorem (Wilks, 1938), which states that $2\log\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)\}$ has an asymptotic χ^2_d distribution under certain regularity conditions. This will apply, for instance, when we have independent and identically distributed (iid) data from an exponential family, Θ_0 is a subset of Θ, and Θ and Θ_0 are linear subspaces in Euclidean space (Van der Vaart, 2000, Theorem 4.6). We can invert the LRT to produce an asymptotically valid $100(1-\alpha)\%$ confidence region of the following form:

$$C_n^{\mathrm{LRT}}(\alpha) = \left\{\theta \in \Theta : 2\log\left\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\theta)\right\} \leq c_{\alpha,d}\right\}.$$

We reject H_0 if and only if $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0 = \emptyset$, which is equivalent to rejecting H_0 if and only if $2\log\{\mathcal{L}(\widehat{\theta}) / \mathcal{L}(\widehat{\theta}_0)\} > c_{\alpha,d}$. We refer to this testing framework as the classical LRT. Some composite nulls are irregular, meaning that Wilks' theorem does not apply and calculating a threshold can be hard due to intractable asymptotics.

Wasserman et al. (2020) presented an alternative to the classical LRT that is valid in finite samples without requiring regularity conditions. In the parametric case, suppose we have n iid observations $Y_1, \ldots, Y_n \sim P_{\theta^*}$, where P_{θ^*} is from a family $(P_\theta : \theta \in \Theta)$. Each P_θ has a density denoted by p_θ. We denote the dataset by $\mathcal{D} = \{Y_1, \ldots, Y_n\}$. To implement the test, first partition the data into $\mathcal{D}_0$ and $\mathcal{D}_1$. Let $\widehat{\theta}_1$ be an estimator constructed from $\mathcal{D}_1$. The parameter $\widehat{\theta}_1$ could be the MLE, but any parameter that is fixed given $\mathcal{D}_1$ is valid. Certain choices of $\widehat{\theta}_1$ may be more efficient. Using the data in $\mathcal{D}_0$, the likelihood function is $\mathcal{L}_0(\theta) = \Pi_{Y_i \in \mathcal{D}_0} p_\theta(Y_i)$. Define the split LRT statistic as

$$T_n(\theta) = \mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\theta).$$

The universal confidence set for θ^* using the split LRT is

$$C_n^{\text{split}}(\alpha) = \{\theta \in \Theta : T_n(\theta) < 1/\alpha\}.$$

Theorem 1.0.1. *$C_n^{split}(\alpha)$ is a valid $100(1-\alpha)\%$ confidence set for θ^*. As a consequence (and equivalently), when testing an arbitrary composite null $H_0 : \theta^* \in \Theta_0$ versus $H_1 : \theta^* \in \Theta \setminus \Theta_0$, rejecting H_0 when $\Theta_0 \cap C_n^{split}(\alpha) = \emptyset$ provides a valid level α hypothesis test. This rule reduces to rejecting H_0 if $T_n(\widehat{\theta}_0) \geq 1/\alpha$, where $\widehat{\theta}_0 \in \arg\max_{\theta \in \Theta_0} \mathcal{L}_0(\theta)$ is the null MLE.*

Theorem 1.0.1 is due to Wasserman et al. (2020). The validity of the universal test does not depend on large samples or regularity conditions. The proof establishes that $E_{\theta^*}\{T_n(\theta^*)\} \leq 1$ and then invokes Markov's inequality. See Appendix A for more details. This property on the expectation makes $T_n(\theta^*)$ an e-variable (Grünwald et al., 2020).

The validity of $C_n^{\text{split}}(\alpha)$ only depends on the fact that $\mathbb{E}_{\theta^*}\{T_n(\theta^*)\} \leq 1$. If we consider multiple test statistics that each satisfy this condition, then the average of those test statistics will satisfy the condition as well. Therefore, the average of test statistics $T_n(\theta^*)$ across multiple data splits is also a valid test statistic.

The universal test applies more generally for testing nonparametric classes as well. Suppose $\mathcal{F}$ is some collection of densities, which can be nonparametric. Assume we have n iid observations $Y_1, \ldots, Y_n$ with some true density f^*. We wish to test $H_0 : f^* \in \mathcal{F}$ versus $H_1 : f^* \notin \mathcal{F}$. Again, we partition the sample into $\mathcal{D}_0$ and $\mathcal{D}_1$. The likelihood function evaluated on a density f over the data in $\mathcal{D}_0$ is $\mathcal{L}_0(f) = \prod_{Y_i \in \mathcal{D}_0} f(Y_i)$. We define $\widehat{f}_0 = \underset{f \in \mathcal{F}}{\arg\max}\, \mathcal{L}_0(f)$. Let $\widehat{f}_1$ be any density estimated on $\mathcal{D}_1$. Similar to the parametric case, the split LRT statistic is

$$T_n(f) = \mathcal{L}_0(\widehat{f}_1)/\mathcal{L}_0(f).$$

Theorem 1.0.2 extends Theorem 1.0.1 to the nonparametric case. The proof of Theorem 1.0.2 (Appendix A) is nearly identical to the proof of Theorem 1.0.1, with some notational changes.

Theorem 1.0.2. *When $f^* \in \mathcal{F}$, $C_n^{split}(\alpha) = \{f \in \mathcal{F} : T_n(f) < 1/\alpha\}$ is a valid $100(1-\alpha)\%$ confidence set for f^*. A valid level α hypothesis test of $H_0 : f^* \in \mathcal{F}$ versus $H_1 : f^* \notin \mathcal{F}$ rejects H_0 if $\mathcal{F} \cap C_n^{split}(\alpha) = \emptyset$. This hypothesis test is equivalent to rejecting H_0 if $T_n(\widehat{f}_0) \geq 1/\alpha$, where $\widehat{f}_0 = \arg\max_{f \in \mathcal{F}} \mathcal{L}_0(f)$ is the null MLE.*

For nonparametric classes $\mathcal{F}$, it may be difficult to construct the set of densities $C_n^{\text{split}}(\alpha)$. Nevertheless, as long as we are able to construct $\widehat{f}_0$, it is possible to perform the nonparametric hypothesis test described in Theorem 1.0.2. In addition, we can easily check whether a given density f is in $C_n^{\text{split}}(\alpha)$.

Wasserman et al. (2020) describe numerous settings in which the universal LRT is the first hypothesis test with finite sample validity, in both parametric and nonparametric settings. For instance, the universal approach can test the number of components in mixture models (Hartigan, 1985; McLachlan, 1987; Chen et al., 2009; Li and Chen, 2010) or test whether a density satisfies the shape constraint of log-concavity (Cule et al., 2010b; Axelrod et al., 2019).

Many basic questions remain unanswered about the universal LRT, since its power even in very simple settings remains unknown. These questions include the following: In settings where Wilks' Theorem is valid, how large is $C_n^{\text{split}}(\alpha)$ relative to $C_n^{\text{LRT}}(\alpha)$? What proportions of data should we assign to $\mathcal{D}_0$ and $\mathcal{D}_1$ to maximize the universal test's power? If we average test statistics over multiple data splits versus a single data split, by how much does the confidence set shrink? Chapter 2 addresses these questions when testing the mean of a Gaussian density. In Chapter 3, we shift to a nonparametric setting, in which we test whether a density belongs to the class of log-concavity densities. We consider the power of the log-concave universal test across several classes of underlying densities and multiple universal test statistic variants. Together, these studies offer insights into the construction of universal tests for both parametric and nonparametric classes.

Chapter 2

Gaussian Universal Likelihood Ratio Testing

2.1 Introduction

We first study the universal LRT in the fundamental case of constructing confidence regions (or hypothesis tests) for the population mean $\theta^* \in \Theta = \mathbb{R}^d$ when $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. In this setting — where the classical LRT is valid — our results showcase the reasonable performance of the universal LRT in comparison to the classical approach.

This work provides two main contributions: First, we provide a careful analysis of several variants of the universal LRT in the Gaussian case. We show that a repeated subsampling approach is the best choice in terms of size and power. We observe reasonable performance in a high-dimensional setting, where the expected squared radius of the best universal LRT confidence set is approximately 3/2 times the squared radius of the set constructed through the classical approach. Thus, in particular, the power of the universal approaches has the same behavior (in n, d, α) as the classical approach. Second, we show an example of a hypothesis test on normally distributed data where universal LRT methods have higher power than classical testing methods. Specifically, when testing the non-convex "doughnut" null $H_0 : \|\theta^*\| \in [0.5, 1]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1]$ on $N(\theta^*, I_d)$ data, a universal LRT approach can have higher power than a standard approach that uses the classical LRT confidence set. A test of this form could examine, for instance, whether trial outcomes or biomarker levels are within an acceptable range.

2.2 Universal LRT Confidence Sets

2.2.1 Classical Test in Normal Setting

Assume $Y_1, \ldots, Y_n$ are d-dimensional iid vectors drawn from $N(\theta^*, I_d)$ with $\theta^* \in \Theta = \mathbb{R}^d$. Where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution, the classical LRT confidence set for θ^* is

$$C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \leq c_{\alpha,d}/n\right\}. \tag{2.1}$$

See Appendix B.2 for a derivation of (2.1). In this case, $C_n^{\text{LRT}}(\alpha)$ is valid in finite samples, since $n\|\theta^* - \overline{Y}\|^2$ follows a χ^2_d distribution. We compare $C_n^{\text{LRT}}(\alpha)$ to the split LRT set and several universal confidence sets that are variants of the split LRT set.

2.2.2 Split, Cross-fit, and Subsampling Sets in Normal Setting

First, we consider two universal LRT variants based on a single split of the data. The validity of these approaches follows from Theorem 1.0.1. Assume we split the n observations in half, such that $\mathcal{D}_0$ and $\mathcal{D}_1$ each contain $n/2$ observations. Define $\overline{Y}_0 = (2/n)\sum_{Y_i \in \mathcal{D}_0} Y_i$ and $\overline{Y}_1 = (2/n)\sum_{Y_i \in \mathcal{D}_1} Y_i$. Then the confidence set for θ^* based on the split likelihood ratio is

$$\begin{aligned} C_n^{\text{split}}(\alpha) &= \left\{\theta \in \Theta : \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \theta\|^2\right) < \frac{1}{\alpha}\right\} \\ &= \left\{\theta \in \Theta : \|\theta - \overline{Y}_0\|^2 < (4/n)\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}. \end{aligned} \tag{2.2}$$

See Appendix B.2 for a derivation of (2.2). Using the same split, we define the cross-fit statistic as $S_n(\theta) = \{T_n(\theta) + T_n^{\text{swap}}(\theta)\}/2$, where $T_n^{\text{swap}}(\theta)$ is computed by switching the roles of $\mathcal{D}_0$ and $\mathcal{D}_1$. Then the cross-fit confidence set is a valid $100(1-\alpha)\%$ set given by

$$C_n^{\text{CF}}(\alpha) = \left\{\theta \in \Theta : \frac{1}{2}\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2\right)\left\{\exp\left(\frac{n}{4}\|\overline{Y}_0 - \theta\|^2\right) + \exp\left(\frac{n}{4}\|\overline{Y}_1 - \theta\|^2\right)\right\} < \frac{1}{\alpha}\right\}.$$

The split and cross-fit sets have both statistical randomness (due to the random sampling of observations) and algorithmic randomness (due to the randomness in splitting the sample into $\mathcal{D}_0$ and $\mathcal{D}_1$). In contrast, the classical LRT only has statistical randomness, since the test is deterministic for a given set of observations. We now consider a repeated subsampling approach.

This universal method attempts to mitigate the algorithmic randomness from the split and cross-fit LRTs by splitting the observations many times and averaging the test statistics. Algorithm 1 shows how to compute the subsampling test statistic $T_n(\theta)$ at a given $\theta \in \mathbb{R}^d$.

Algorithm 1 Compute the subsampling test statistic $T_n(\theta)$

```
Input: $n$ independent $d$-dimensional observations $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$ ($\theta^*$ unknown),
    a value of $\theta \in \mathbb{R}^d$, number of subsamples $B$.
Output: The subsampling test statistic $T_n(\theta)$.
for $b = 1, 2, \ldots, B$ do
    Randomly partition the data into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$, each containing $n/2$ values of $Y_i$.
    Let $\overline{Y}_{0,b} = (2/n) \sum_{Y_i \in \mathcal{D}_{0,b}} Y_i$ and let $\overline{Y}_{1,b} = (2/n) \sum_{Y_i \in \mathcal{D}_{1,b}} Y_i$.
    Compute $T_{n,b}(\theta) = \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right)$.
return the subsampling test statistic $T_n(\theta) = B^{-1} \sum_{b=1}^{B} T_{n,b}(\theta)$.
```

As noted in Chapter 1, this method is also valid. The $100(1-\alpha)\%$ subsampling confidence set is

$$C_n^{\text{subsplit}}(\alpha) = \left\{ \theta \in \Theta : \frac{1}{B} \sum_{b=1}^{B} \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right) < \frac{1}{\alpha} \right\}.$$

Figure 2.1 shows coverage regions of the classical LRT, split LRT, cross-fit LRT, and subsampling LRT ($B = 100$) from six simulations with $\theta^* = (0, 0)$. We generate 1000 observations from $N(\theta^*, I_2)$, and we use this sample for all simulations. Hence, the variation in the split, cross-fit, and subsampling LRTs across simulations is due to algorithmic randomness.

The coverage regions in Fig. 2.1 suggest several relationships that we will formalize. We see that the classical LRT provides the smallest confidence regions. This is not surprising since, even in finite samples, the classical LRT statistic follows a chi square distribution under $H_0 : \theta = \theta^*$ in the Gaussian case. The volume of the cross-fit LRT set is less than or equal to the volume of the split LRT set, although the cross-fit set is not entirely contained within the split set. The split and cross-fit approaches both use a single split of the data, but there is a notable improvement from cross-fitting. The subsampling set also has less volume than the split LRT set. Recall that we construct the subsampling test statistic by performing the split LRT over repeated splits of the data and then averaging the test statistics $T_{n,b}(\theta)$. While any individual split LRT region is guaranteed to be spherical, the subsampling set is not necessarily a spherical region. For large

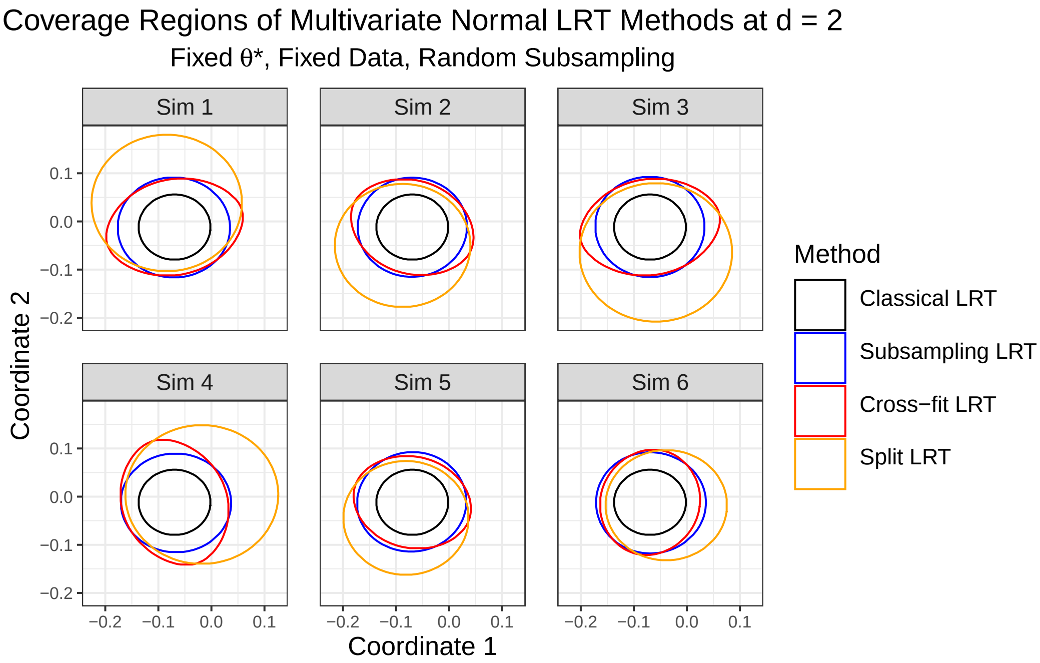


Figure 2.1: Coverage regions of classical LRT, subsampling LRT, cross-fit LRT, and split LRT at $\alpha = 0.1$. The six simulations use the same 1000 observations from $N(\theta^*, I_2)$ under $\theta^* = (0, 0)$.

B, however, we see that the subsampling region is approximately spherical. Thus, although the subsampling approach is computationally intensive, this hints that it may be possibly to derive a formulaic approximation to the limiting subsampling set.

2.2.3 Limit of Subsampling Region

We are particularly interested in the behavior of the subsampling confidence set as $B \to \infty$. Since $\lim_{B\to\infty} B^{-1} \sum_{b=1}^{B} T_{n,b}(\theta) = \mathbb{E}\{T_n(\theta) \mid \mathcal{D}\}$, the limiting subsampling set has no algorithmic randomness. We see hints of this in Fig. 2.1, where the subsampling set at $B = 100$ does not vary much across six simulations on the same data. Theorem 2.2.1 describes conditions for the convergence of the ratio of $\mathbb{E}\{T_n(\theta) \mid \mathcal{D}\}$ to an approximation. We have been suppressing the n subscript when it is clear we are working with a single dataset with n observations. Theorem 2.2.1 considers a sequence of datasets, so we use the n subscript to index the datasets.

Theorem 2.2.1. *Assume we have a sequence of datasets $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$, where $\mathcal{D}_n = \{Y_{n1}, \ldots, Y_{nn}\}$ and each Y_{ni} is an independent observation from $N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$, and let $\mathcal{D}_{1,n} = \mathcal{D}_n \backslash \mathcal{D}_{0,n}$. Define $\overline{Y}_n = (1/n)\sum_{i=1}^{n} Y_{ni}$, $\overline{Y}_{0,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{0,n}} Y_{ni}$, and*

$\overline{Y}_{1,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{1,n}} Y_{ni}$. *Let* $c > 0$, *and let* (θ_n) *be a sequence that satisfies* $\|\overline{Y}_n - \theta_n\| \leq c/\sqrt{n}$ *for all* n. *Then*

$$\mathbb{E}\{T_n(\theta_n) \mid \mathcal{D}_n\} \,/\, \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} = 1 + o_P(1). \tag{2.3}$$

In words, the subsampling statistic is approximately given by $R(\theta)^{3/5}(2/5)^{d/2}$ where $R(\theta) = \mathcal{L}(\widehat{\theta})/\mathcal{L}(\theta)$ is the classical likelihood ratio statistic.

Appendix B.1 contains a proof of Theorem 2.2.1. The proof relies critically on the finite sample central limit theorems from Hájek (1960) and Li and Ding (2017) and on the Portmanteau Theorem proof techniques from Van der Vaart (2000).

Since

$$\mathbb{E}\{T_n(\theta) \mid \mathcal{D}\} \approx \exp\left(\frac{3n}{10}\|\overline{Y} - \theta\|^2\right)\left(\frac{2}{5}\right)^{d/2}, \tag{2.4}$$

the subsampling confidence region is approximately

$$\begin{aligned} C_n^{\text{subsplit}}(\alpha) &= \left\{\theta \in \Theta : \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right) < \frac{1}{\alpha}\right\} \\ &\approx \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 < \frac{10}{3n}\log\left(\frac{(5/2)^{d/2}}{\alpha}\right)\right\}. \end{aligned} \tag{2.5}$$

Figure 2.2 validates (2.4) as a reasonable approximation. We simulate one sample $Y_1, \ldots, Y_n \sim N(0, I_d)$ at each combination of $d \in \{1, 20\}$ and $n \in \{10, 50, 200\}$. We consider θ values of the form $\theta = c\vec{1}$. Through $B = 100,000$ subsampling simulations at each (d, n, c) combination, we estimate

$$\mathbb{E}\{T_n(\theta) \mid \mathcal{D}\} \approx \frac{1}{B}\sum_{b=1}^{B} \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right).$$

The black dots represent this average at each combination of (d, n, c), and the red curve traces out $\exp((3n/10)\|\overline{Y} - \theta\|^2)(2/5)^{d/2}$ from (2.4). Except for the most difficult setting of $(d = 20, n = 10)$, the simulated and analytical estimates align well. At $\alpha = 0.1$, the confidence region includes all values of θ such that the test statistic is at most $1/0.1$. The horizontal dashed black line represents

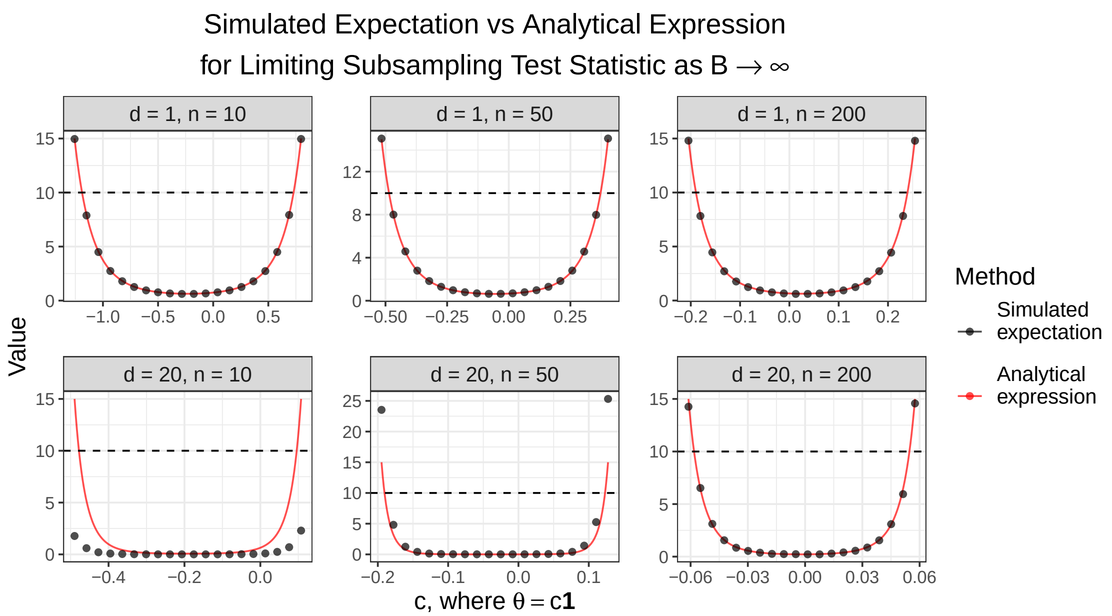


Figure 2.2: Analytical (red curve) and simulated (black dots) approximations of the limiting test statistic $\lim_{B\to\infty}\frac{1}{B}\sum_{b=1}^{B}T_{n,b}(\theta)$ at various dimensions d and numbers of observations n. The test points equal $\theta = c\vec{1}$ for various c. The horizontal dashed black line at $1/0.1$ is the cutoff for an $\alpha = 0.1$ confidence region.

this value. Thus, test statistics constructed from the simulated and analytical approaches would produce similar confidence regions.

2.3 Comparison of Universal LRT Sets

2.3.1 Optimal Split Proportions

We have been assuming that the universal LRTs place $n/2$ observations in $\mathcal{D}_0$ and $n/2$ observations in $\mathcal{D}_1$. The statement $\mathbb{E}_{\theta^*}\{T_n(\theta^*)\} \leq 1$ holds regardless of the proportion of observations in $\mathcal{D}_0$ versus $\mathcal{D}_1$, though. Let p_0 denote the proportion of observations that we place in $\mathcal{D}_0$.

Theorem 2.3.1. *Let $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. The splitting proportion that minimizes $\mathbb{E}[r^2\{C_n^{split}(\alpha)\}]$ is*

$$p_0^* = 1 - \frac{\sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right)} - d}{2\log\left(\frac{1}{\alpha}\right)}. \tag{2.6}$$

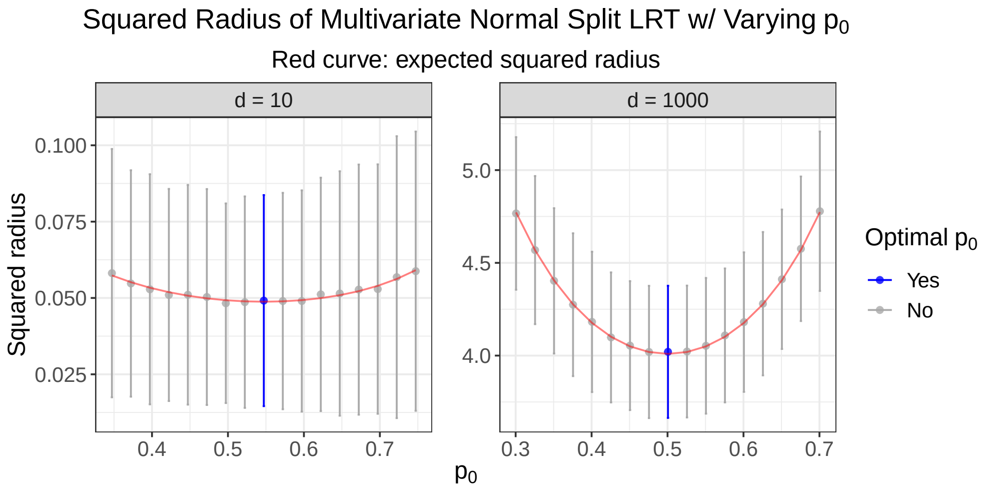


Figure 2.3: Squared radius of multivariate normal split LRT with varying p_0. We simulate $Y_1, \ldots, Y_{1000} \sim N(0, I_d)$ and compute the split LRT region at varying p_0. We repeat this simulation 1000 times. At each p_0, the circular point is the mean squared radius and the error bar represents the mean squared radius ± 1.96 standard deviations. Blue points/lines correspond to p_0^*. The red curve is the expected squared radius. (See Theorem 2.3.1 proof in Appendix B.1 for a derivation of the expected squared radius at p_0.)

As $d \to \infty$ for fixed α, the optimal split proportion p_0^* converges to 0.5. See Appendix B.1 for a proof of Theorem 2.3.1 and a derivation of this fact. Alternatively, as $\alpha \to 0$ for fixed d, the proportion p_0^* converges to 1, suggesting that one should use nearly all data for likelihood estimation. This is not an issue for reasonable α levels, though. For instance, at $d = 1$, one would need to set $\alpha < \exp(-40)$ to produce an optimal split proportion p_0^* that exceeds 0.90.

Figure 2.3 shows the average squared radius of the split LRT at p_0^* and at surrounding choices of p_0. The expected squared radius (red curve) is more sensitive to changes in p_0 at higher values of d. That is, use of the optimal p_0^* has a greater effect on the split LRT squared radius in higher dimensions. In high dimensions, though, p_0^* is close to 0.5. It is thus a reasonable choice to use $p_0 = 0.5$ in all dimensions. We use $p_0 = 0.5$ for all remaining analyses.

In the cross-fit case, we conjecture that $p_0 = 0.5$ minimizes the expected squared diameter. Simulations in Appendix B.3 support this claim. Intuitively, since the cross-fit approach uses both $\mathcal{D}_0$ and $\mathcal{D}_1$ once for parameter estimation and once for likelihood computation, we should not gain any efficiency by using unbalanced sets.

2.3.2 Split versus Cross-fit Volume

In Fig. 2.1, we see that the cross-fit LRT set volume is less than the split LRT set volume, but $C_n^{\text{CF}}(\alpha)$ is not a subset of $C_n^{\text{split}}(\alpha)$. Nevertheless, it holds that $\text{Volume}\{C_n^{\text{CF}}(\alpha)\} \leq \text{Volume}\{C_n^{\text{split}}(\alpha)\}$.

Theorem 2.3.2. *Suppose $Y_1, \ldots, Y_n$ are iid observations from $N(\theta^*, I_d)$. Split the sample such that $\mathcal{D}_0$ and $\mathcal{D}_1$ each contain $\frac{n}{2}$ observations. Use $\mathcal{D}_0$ and $\mathcal{D}_1$ to define the split and cross-fit sets. Then $Volume\{C_n^{CF}(\alpha)\} \leq Volume\{C_n^{split}(\alpha)\}$. Equality holds only when $\overline{Y}_0 = \overline{Y}_1$.*

Briefly, the proof of Theorem 2.3.2 constructs a spherical region centered at $\overline{Y}$ with radius equal to the split LRT radius. The cross-fit set is a subset of this re-centered split LRT region, so the volume of the cross-fit LRT set is bounded above by the volume of the split LRT set. If $\overline{Y}_0 = \overline{Y}_1$, then the split and cross-fit LRT sets are equivalent and have equal volume. The fact that equal volume holds only when $\overline{Y}_0 = \overline{Y}_1$ relies on the strict convexity of the squared L_2 norm and the exponential function. See Appendix B.1 for a complete proof.

Theorem 2.3.2 proves that the cross-fit LRT approach improves over the split LRT by constructing provably smaller confidence regions. Out of all universal methods, our simulations have shown that the subsampling approach tends to produce the smallest sets. Constructing a subsampling region can be computationally intensive, though, especially when the limiting subsampling test statistic is intractable. The cross-fit approach may be a reasonable compromise in settings where repeated subsampling is computationally prohibitive.

2.3.3 Comparative Size in High Dimensions

Figure 2.1 demonstrated the appearance of the four LRT regions in the $d = 2$ case at $\alpha = 0.1$. We observe that the classical LRT and the split LRT produce the smallest and largest confidence regions, respectively. While the split LRT region's radius appears to be approximately twice the classical LRT region's radius, we consider whether the ratio of their squared radii diverges in high dimensions or for very small α. We characterize the ratio of squared radii in terms of the expected ratio. The expected squared radius of $C_n^{\text{split}}(\alpha)$ is

$$\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}] = (4/n)\log(1/\alpha) + (4/n)d. \tag{2.7}$$

Thus, the expected ratio of the split LRT squared radius over the classical LRT radius is

$$\frac{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]}{r^2\{C_n^{\text{LRT}}(\alpha)\}} = \frac{(4/n)\log(1/\alpha) + (4/n)d}{c_{\alpha,d}/n} = \frac{4\log(1/\alpha) + 4d}{c_{\alpha,d}}. \tag{2.8}$$

For $d \geq 2$ and $\alpha \leq 0.17$,

$$\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \frac{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]}{r^2\{C_n^{\text{LRT}}(\alpha)\}} \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d - \frac{5}{2}}. \tag{2.9}$$

For $d = 1$ and $\alpha \leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right)$,

$$\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \frac{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]}{r^2\{C_n^{\text{LRT}}(\alpha)\}} \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + 9 - 4\sqrt{5 + 2\log(1/\alpha)}}. \tag{2.10}$$

See Appendix B.2 for derivations of (2.7), (2.9), and (2.10). The derivation of (2.9) relies on chi square quantile bounds from Theorem A and Proposition 5.1 of Inglot (2010). The derivation of the upper bound in (2.10) involves a bound from Section 2.1 of Polland (2015). The restrictions on α and d are necessary for the upper bounds to be valid. The lower bound of both (2.9) and (2.10) is valid for any $d \geq 1$ and $\alpha \in (0, 1)$. The upper and lower bounds both converge to 4 as $d \to \infty$. In addition, all bounds converge to 2 as $\alpha \to 0$. Figure 2.4 shows the true value of $\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}] \,/\, r^2\{C_n^{\text{LRT}}(\alpha)\}$ as well as the proved lower and upper bounds on this expectation at $d = 10$ and $d = 100,000$. We observe that the bounds converge to 2 for very small α relative to the dimension, and we observe that the bounds converge to 4 for high dimensions relative to α. Interestingly, we see that the expected value of the ratio is not monotone increasing in α.

Furthermore, this ratio of squared radii is less than 4 with probability approximately $1 - \alpha$ in high dimensions under a condition on α and d. Theorem 2.3.3 formalizes this result. See Appendix B.1 for a proof.

Theorem 2.3.3. *Assume $c_{\alpha,d} + \log(\alpha) > d - 2$. Let $f_d(x)$ be the probability density function of the χ^2_d distribution, and let $c_{\alpha,d}$ be the upper α quantile of the χ^2_d distribution. Then*

$$\begin{aligned} &\mathbb{P}\left[r^2\{C_n^{split}(\alpha)\}/r^2\{C_n^{LRT}(\alpha)\} \leq 4\right] \geq 1 - \alpha - \log(1/\alpha) f_d(c_{\alpha,d} + \log(\alpha)) \\ \textit{and} \quad &\mathbb{P}\left[r^2\{C_n^{split}(\alpha)\}/r^2\{C_n^{LRT}(\alpha)\} \leq 4\right] \leq 1 - \alpha - \log(1/\alpha) f_d(c_{\alpha,d}). \end{aligned}$$

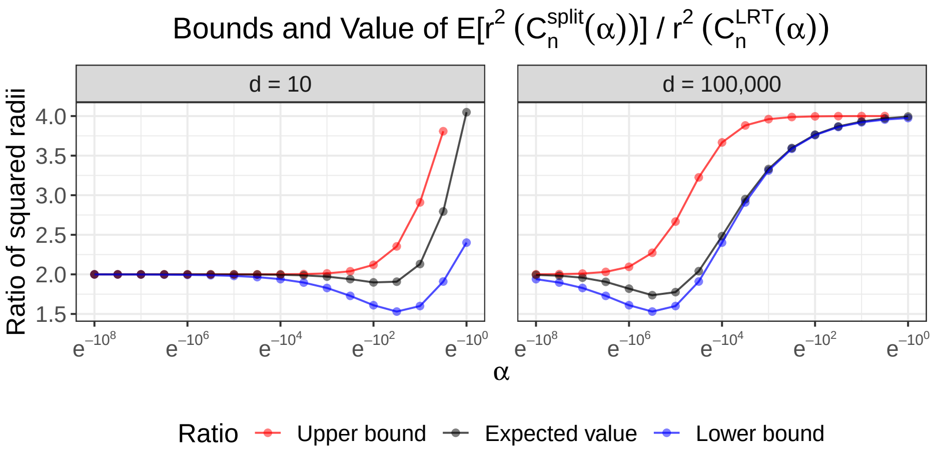


Figure 2.4: Expectation, lower bound, and upper bound of $\mathbb{E}\left[r^2\{C_n^{\text{split}}(\alpha)\}\right]/r^2\{C_n^{\text{LRT}}(\alpha)\}$. The expected value equals the expression from (2.8). The lower and upper bounds correspond to the bounds in (2.9). Data points correspond to values at $\alpha = \exp(-10^x)$ for x from 8 to 0 in increments of -0.5.

Figure 2.5 explores the bounds from Theorem 2.3.3. We see that the result from Theorem 2.3.3 is more informative in higher dimensions, where the upper and lower bounds are closer to each other. Both theoretically and empirically, the ratio of squared radii $r^2\{C_n^{\text{split}}(\alpha)\}/r^2\{C_n^{\text{LRT}}(\alpha)\}$ is less than 4 with probability slightly below $1-\alpha$ in higher dimensions.

From (2.5) and (2.7), we can see that

$$\frac{r^2\{C_n^{\text{subsplit}}(\alpha)\}}{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]} \approx \frac{5}{6}\left\{\frac{(d/2)\log(5/2)+\log(1/\alpha)}{d+\log(1/\alpha)}\right\}. \tag{2.11}$$

Combining (2.9) and (2.11), $r^2\{C_n^{\text{subsplit}}(\alpha)\}/r^2\{C_n^{\text{LRT}}(\alpha)\}$ is approximately $4(5/12)\log(5/2) \approx 3/2$ as $d \to \infty$, and the ratio is approximately $2(5/6) = 5/3$ as $\alpha \to 0$. Recall that the classical LRT cutoff is dimension dependent and uses the exact distribution's quantile, while the universal LRT cutoff is dimension independent. Regardless, in the extreme cases of $d \to \infty$ or $\alpha \to 0$, the ratio of the classical LRT region's radius to the subsampling universal LRT region's radius is less than 2.

2.3.4 Power

While the universal methods provide conservative confidence regions for θ^*, we establish that the universal tests can still have high power. Suppose we wish to test $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ at

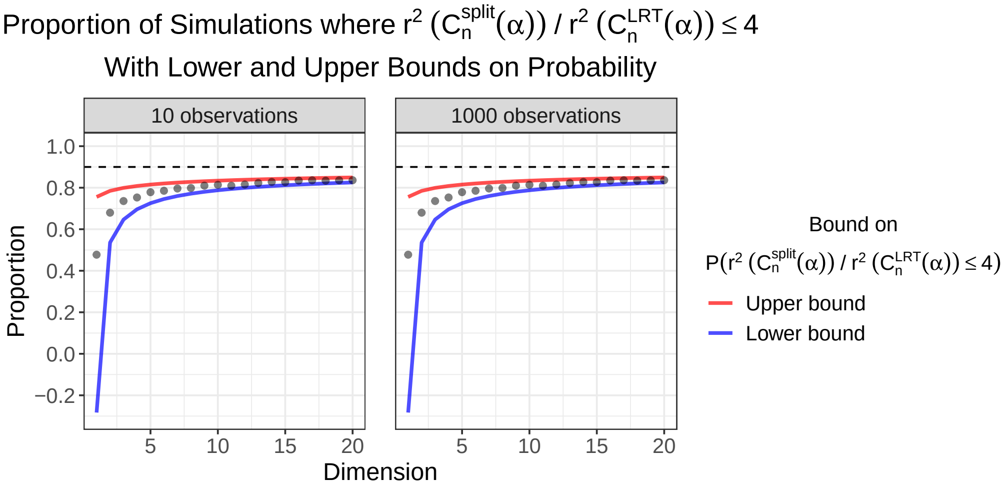


Figure 2.5: At each (d, n) combination, we perform 10,000 simulations. In each simulation, we generate a data sample $Y_1, \ldots, Y_{1000} \sim N(0, I_2)$, construct the split and classical LRT confidence sets, and compute the squared radii. The points represent the proportion of these simulations in which $r^2\{C_n^{\text{split}}(\alpha)\}/r^2\{C_n^{\text{LRT}}(\alpha)\} \leq 4$. The red and blue curves are the lower and upper bounds on $\mathbb{P}[r^2\{C_n^{\text{split}}(\alpha)\}\ /\ r^2\{C_n^{\text{LRT}}(\alpha)\} \leq 4]$ from Theorem 2.3.3 at $\alpha = 0.1$.

level $1 - \alpha$. We reject H_0 if $0 \notin C_n(\alpha)$, where $C_n(\alpha)$ is the confidence set defined by some likelihood ratio test. The power of the test at $\theta^* \neq 0$ is $\mathbb{P}_{\theta^*}\{0 \notin C_n(\alpha)\}$.

First, we consider the classical LRT, stated in (2.1). The power of the classical LRT at θ^* is

$$\text{Power}\{C_n^{\text{LRT}}(\alpha); \theta^*\} = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > c_{\alpha,d}/n\right) \approx \Phi\left\{\frac{d + n\|\theta^*\|^2 - c_{\alpha,d}}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right\}. \tag{2.12}$$

We can find a similar representation for the approximate power of the limiting subsampling LRT as $B \to \infty$:

$$\begin{aligned}\text{Power}\{C_n^{\text{subsplit}}(\alpha); \theta^*\} &\approx \mathbb{P}_{\theta^*}\left[n\|\overline{Y}\|^2 \geq \frac{10}{3}\log\left\{\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right\}\right] \\ &\approx \Phi\left(\frac{1}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\left[d + n\|\theta^*\|^2 - \frac{10}{3}\log\left\{\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right\}\right]\right). \end{aligned} \tag{2.13}$$

Since $n\|\overline{Y}\|^2 \sim \chi^2\left(df = d, \lambda = n\|\theta^*\|^2\right)$, (2.12) and (2.13) use the normal approximation to the non-central χ^2 distribution with a large noncentrality parameter λ (Chun and Shapiro, 2009). See Appendix B.2 for derivations of (2.12) and (2.13).

The power of the split LRT is

$$\mathrm{Power}\{C_n^{\mathrm{split}}(\alpha);\theta^*\} = \mathbb{P}_{\theta^*}\left\{\|\overline{Y}_0\|^2 \geq (4/n)\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\},$$

and the power of the cross-fit LRT is

$$\mathrm{Power}\{C_n^{\mathrm{CF}}(\alpha);\theta^*\} = \mathbb{P}_{\theta^*}\left[\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2\right)\left\{\exp\left(\frac{n}{4}\|\overline{Y}_0\|^2\right) + \exp\left(\frac{n}{4}\|\overline{Y}_1\|^2\right)\right\} \geq \frac{2}{\alpha}\right].$$

As $n\|\theta^*\|^2 \to \infty$ for fixed α, the power of the tests approaches 1. Importantly, this shows that although the universal methods are conservative, they will all have high power for sufficiently large n or for $\|\theta^*\|$ sufficiently far from 0. As $\alpha \to 0$, the power approaches 0.

Figure 2.6 plots the power of the LRTs against $\|\theta^*\|^2$. (Each vector θ^* has the form $c\vec{1}$.) This figure uses the standard normal CDF approximation to the non-central χ^2 CDF to plot the classical and subsampling LRT power. We use simulations to approximate the power of the split and cross-fit LRTs. For a given value of θ^*, we simulate $n = 1000$ observations $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. We construct split LRT and cross-fit LRT confidence sets from this sample. Then we test whether $\theta = 0$ is in each confidence set. We repeat this procedure 5000 times at each θ^*, and each procedure's estimated power at θ^* is the proportion of times that $0 \notin C_n(\alpha)$.

As we would expect, the power is higher when θ^* is farther from 0. In addition, the classical LRT has the highest power, followed in order by the subsampling LRT, the cross-fit LRT, and the split LRT. Interestingly, at $d = 1$ the subsampling and cross-fit LRT have nearly identical (approximate) power. As d increases, the difference between the subsampling and cross-fit LRT power increases.

2.4 Example: Hypothesis Testing a Doughnut Null Set

We present an example of a nontrivial testing problem that appears to be beyond the current reach of our mathematical analysis. Below, a procedure based on universal inference can have higher power than a more standard intersection approach using the classical, exact confidence set, motivating the need for further study of the pros and cons of such methods.

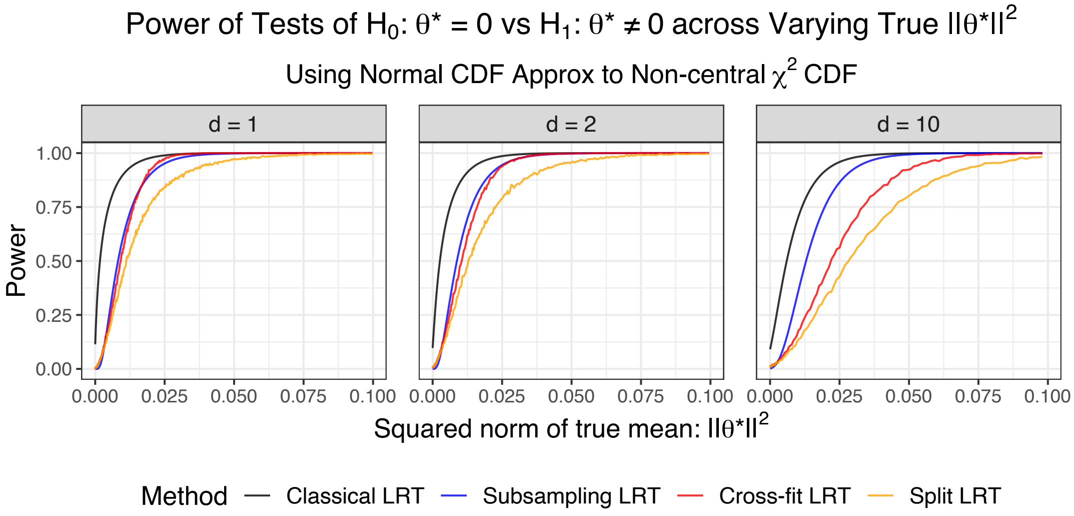


Figure 2.6: Estimated power of classical LRT, limiting subsampling LRT, cross-fit LRT, and split LRT. We are testing $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ across varying true $\|\theta^*\|^2$. We use the standard normal CDF approximation for the classical and subsampling LRT power calculations, and we use simulations to estimate the cross-fit and split LRT power.

Suppose we observe an iid sample $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$, and we wish to test

$$H_0 : \|\theta^*\| \in [0.5, 1.0] \text{ versus } H_1 : \|\theta^*\| \notin [0.5, 1.0].$$

Then $\Theta_0 = \{\theta \in \mathbb{R}^d : \|\theta\| \in [0.5, 1.0]\}$ and $\Theta_1 = \{\theta \in \mathbb{R}^d : \|\theta\| \notin [0.5, 1.0]\}$. The nonconvex structure of Θ_0 makes it unclear how to construct a valid test based on a limiting distribution. Nevertheless, we can use alternative methods, including universal inference tools, to construct valid hypothesis tests for $H_0 : \|\theta^*\| \in [0.5, 1.0]$. We compare three approaches to this test.

Approach 1: Intersect confidence set with Θ_0*.* $C_n^{\text{LRT}}(\alpha) = \{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \leq c_{\alpha,d}/n\}$ is a level $1 - \alpha$ confidence set for θ^*, where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution. Suppose we reject H_0 if and only if $C_n^{\text{LRT}}(\alpha) \cap \Theta_0 = \emptyset$. We can see that this test has valid type I error control. Assume $\theta^* \in \Theta_0$. Then

$$\begin{aligned}
\mathbb{P}_{\theta^*}\left\{C_n^{\text{LRT}}(\alpha) \cap \Theta_0 = \emptyset\right\} &\leq \mathbb{P}_{\theta^*}\left\{\theta^* \notin C_n^{\text{LRT}}(\alpha) \cup \theta^* \notin \Theta_0\right\} \\
&= \mathbb{P}_{\theta^*}\left\{\theta^* \notin C_n^{\text{LRT}}(\alpha)\right\} \\
&= \alpha.
\end{aligned}$$

To implement this test, we need to check whether the intersection $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0$ is empty. First, we set $\widehat{\theta}^{\mathrm{proj}}$ to the projection of $\overline{Y}$ onto Θ_0. That is,

$$\widehat{\theta}^{\mathrm{proj}} = \begin{cases} 0.5\,\overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| < 0.5 \\ \overline{Y} & : \|\overline{Y}\| \in [0.5, 1.0] \\ \overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| > 1. \end{cases}$$

Now $\widehat{\theta}^{\mathrm{proj}}$ minimizes $\|\theta - \overline{Y}\|^2$ out of all $\theta \in \Theta_0$. So $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0 = \emptyset$ if and only if $\widehat{\theta}^{\mathrm{proj}} \notin C_n^{\mathrm{LRT}}(\alpha)$.

Approach 2: Subsampled split LRT. To implement the subsampled split LRT, we repeatedly split the observations into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$. Let $\widehat{\theta}_{1,b}$ be any parameter estimated on the data in $\mathcal{D}_{1,b}$. Let $\widehat{\theta}_{0,b}^{\mathrm{split}}$ be the MLE under $H_0 : \|\theta^*\| \in [0.5, 1.0]$, estimated on the data in $\mathcal{D}_{0,b}$. Table 2.1 presents the chosen expression for $\widehat{\theta}_{1,b}$ and the MLE expression for $\widehat{\theta}_{0,b}^{\mathrm{split}}$. The subsampled split LRT rejects H_0 if $B^{-1} \sum_{b=1}^{B} U_{n,b} \geq 1/\alpha$, where

$$U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b}) \,/\, \mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\mathrm{split}}) = \prod_{Y_i \in \mathcal{D}_{0,b}} \{p_{\widehat{\theta}_{1,b}}(Y_i) \,/\, p_{\widehat{\theta}_{0,b}^{\mathrm{split}}}(Y_i)\}.$$

Approach 3: Subsampled hybrid LRT. As an alternative to the split LRT, Wasserman et al. (2020) establish a test based on the reversed information projection (RIPR); also see Grünwald et al. (2020). We first define the RIPR, following Definition 4.2 of the PhD thesis by Li (1999). Let Q be a distribution with density q, and let $\mathcal{P}_\Theta$ be a convex set of densities (or redefine it as its convex hull). Let $D_{\mathrm{KL}}(q \,\|\, p)$ be the Kullback-Leibler divergence of q from p. The RIPR of q onto $\mathcal{P}_\Theta$ is a (sub-)density p^* such that for arbitrary sequences p_n in $\mathcal{P}_\Theta$, $D_{\mathrm{KL}}(q\|p_n) \to \inf_{\theta \in \Theta} D_{\mathrm{KL}}(q\|p_\theta)$ implies $\log(p_n) \to \log(p^*)$ in $L^1(Q)$. Lemma 4.1 of Li (1999) proves that p^* exists and is unique; further, p^* satisfies $D_{\mathrm{KL}}(q \,\|\, p^*) = \inf_{\theta \in \Theta} D_{\mathrm{KL}}(q \,\|\, p_\theta)$, and if $Y \sim q$, then for all $\theta \in \Theta$, $\mathbb{E}_q\{p_\theta(Y)/p^*(Y)\} \leq 1$.

Using similar logic to Theorem 1.0.1, Wasserman et al. (2020) apply this property to construct a split RIPR LRT. Let $\mathcal{P}_{\Theta_0}$ be the set of all densities in H_0 (or its convex hull). Suppose $\widehat{\theta}_1$ is an estimator constructed on $\mathcal{D}_1$. Then conditioning on $\mathcal{D}_1$ fixes the value of $\widehat{\theta}_1$. Let p_0^* be the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$. Note that if the true $p_{\theta^*} \in \mathcal{P}_{\Theta_0}$, then $\mathbb{E}_{\theta^*}\{p_{\widehat{\theta}_1}(Y)/p_0^*(Y) \mid \mathcal{D}_1\} = \mathbb{E}_{\widehat{\theta}_1}\{p_{\theta^*}(Y)/p_0^*(Y) \mid$

$\mathcal{D}_1\} \leq 1$. Then a level α hypothesis test rejects H_0 if $R_n \geq 1/\alpha$, where

$$R_n = \prod_{Y_i \in \mathcal{D}_0} \{p_{\widehat{\theta}_1}(Y_i) \,/\, p_0^*(Y_i)\}.$$

This test is valid because if $\theta^* \in \Theta_0$, then $\mathbb{P}_{\theta^*}(R_n \geq 1/\alpha) \leq \alpha \mathbb{E}_{\theta^*}\{p_{\widehat{\theta}_1}(Y)/p_0^*(Y)\} \leq \alpha$. Furthermore, note that the RIPR test statistic will always exceed the split LRT statistic when the two tests use the same numerator, since the split LRT denominator maximizes the likelihood under H_0 on $\mathcal{D}_0$. Thus, the RIPR test will have higher power than the split LRT. (More generally, one can project $p_{\widehat{\theta}_1}^{|\mathcal{D}_0|}$ onto $\mathcal{P}_{\Theta_0}^{|\mathcal{D}_0|}$, but we omit this discussion for brevity.)

In the doughnut test setting, we let $\mathcal{P}_{\Theta_0}$ be the set of all convex combinations of $N(\theta, I_d)$ densities such that $\|\theta\| \in [0.5, 1]$. To implement the subsampled hybrid LRT for this test, we also repeatedly split the observations into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$. Depending on the value of $\|\overline{Y}_{1,b}\|$, we take one of three approaches:

1. If $\|\overline{Y}_{1,b}\| < 0.5$, use the split LRT on the b^{th} subsample. We define $\widehat{\theta}_{1,b}$ and $\widehat{\theta}_{0,b}^{\text{split}}$ as in Table 2.1, and the split LRT statistic is $U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b})/\mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\text{split}})$.

2. If $\|\overline{Y}_{1,b}\| \in [0.5, 1]$, set the b^{th} subsample's test statistic to 1.

3. If $\|\overline{Y}_{1,b}\| > 1$, use the RIPR LRT on the b^{th} subsample. We define $\widehat{\theta}_{1,b}$ and $\widehat{\theta}_{0,b}^{\text{RIPR}}$ as in Table 2.1, and the RIPR statistic is $R_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b})/\mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\text{RIPR}})$.

Theorem 2.4.1 defines a valid test based on this approach. See Appendix B.1 for a proof.

Theorem 2.4.1. *In the doughnut null hypothesis test setting, assume the subsampled test statistics* $U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b}) \,/\, \mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{split})$ *and* $R_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b})/\mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{RIPR})$, $1 \leq b \leq B$. *A valid level* α *test rejects* H_0 *when*

$$\frac{1}{B}\sum_{b=1}^{B}\Big\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5, 1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\Big\} \geq 1/\alpha.$$

To justify the hybrid approach, recall that the RIPR test will have higher power than the split LRT when it is possible to implement the RIPR. Based on the construction of $\widehat{\theta}_{1,b}$, if $\|\overline{Y}_{1,b}\| > 1$, then $\|\widehat{\theta}_{1,b}\| > 1$. In this setting, the proof of Theorem 2.4.1 shows that the density p_θ, with $\theta = \widehat{\theta}_{1,b}/\|\widehat{\theta}_{1,b}\|$, is the RIPR of $\widehat{\theta}_{1,b}$ onto $\mathcal{P}_{\Theta_0}$. On the other hand, it is unclear how to implement the RIPR when $\|\overline{Y}_{1,b}\| < 0.5$, in which case $\|\widehat{\theta}_{1,b}\| < 0.5$. The hybrid approach allows us to use the RIPR when it is implementable, and it relies on the split LRT to provide a valid test when the RIPR is not implementable.

Table 2.1: Requirements and choices for the numerator and denominator in a single subsample of the split LRT and RIPR LRT statistics

Method	Split LRT	RIPR LRT
Restrictions on use	None	$\|\overline{Y}_1\| > 1$. (Computational restriction. RIPR unknown for $\|\overline{Y}_1\| = \|\widehat{\theta}_1\| < 0.5$.)
Numerator	$p_{\widehat{\theta}_1}$, where $\widehat{\theta}_1$ is any parameter fit on $\mathcal{D}_1$.	$p_{\widehat{\theta}_1}$, where $\widehat{\theta}_1$ is any parameter fit on $\mathcal{D}_1$.
Fitted value	Choose $\widehat{\theta}_1 = \overline{Y}_1$.	Choose $\widehat{\theta}_1 = \overline{Y}_1$.
Denominator	$p_{\widehat{\theta}_0}$, where $\widehat{\theta}_0$ is the MLE under H_0, constructed from $\mathcal{D}_0$.	p_0^* is the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$.
Fitted value	No choices. $\widehat{\theta}_0^{\text{split}} = \begin{cases} 0.5\left(\overline{Y}_0/\|\overline{Y}_0\|\right) & : \|\overline{Y}_0\| < 0.5 \\ \overline{Y}_0 & : \|\overline{Y}_0\| \in [0.5, 1] \\ \overline{Y}_0/\|\overline{Y}_0\| & : \|\overline{Y}_0\| > 1 \end{cases}$	No choices. Since $\|\widehat{\theta}_1\| > 1$, $p_0^* = p_\theta$, where $\theta = \widehat{\theta}_0^{\text{RIPR}} = \widehat{\theta}_1/\|\widehat{\theta}_1\|$.

Figure 2.7 shows the simulated power of these three tests of $H_0 : \|\theta^*\| \in [0.5, 1.0]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1.0]$. The intersection method and the subsampled hybrid LRT have the highest power. Interestingly, out of those two methods, the test with higher power varies across dimensions. When $d = 2$ or $d = 1000$, the simulated power of the subsampled hybrid LRT is less than (or equal to) the power of the standard intersection approach. At the intermediate dimensions of $d = 10$ and $d = 100$, the simulated power of the subsampled hybrid LRT is greater than (or equal to) the power of the standard intersection approach. The latter two cases show that even in the Gaussian setting, hypothesis tests based on a universal LRT can have higher power than tests based on the exact confidence set. When $\|\theta^*\| < 0.5$, the hybrid test and the split test have approximately the

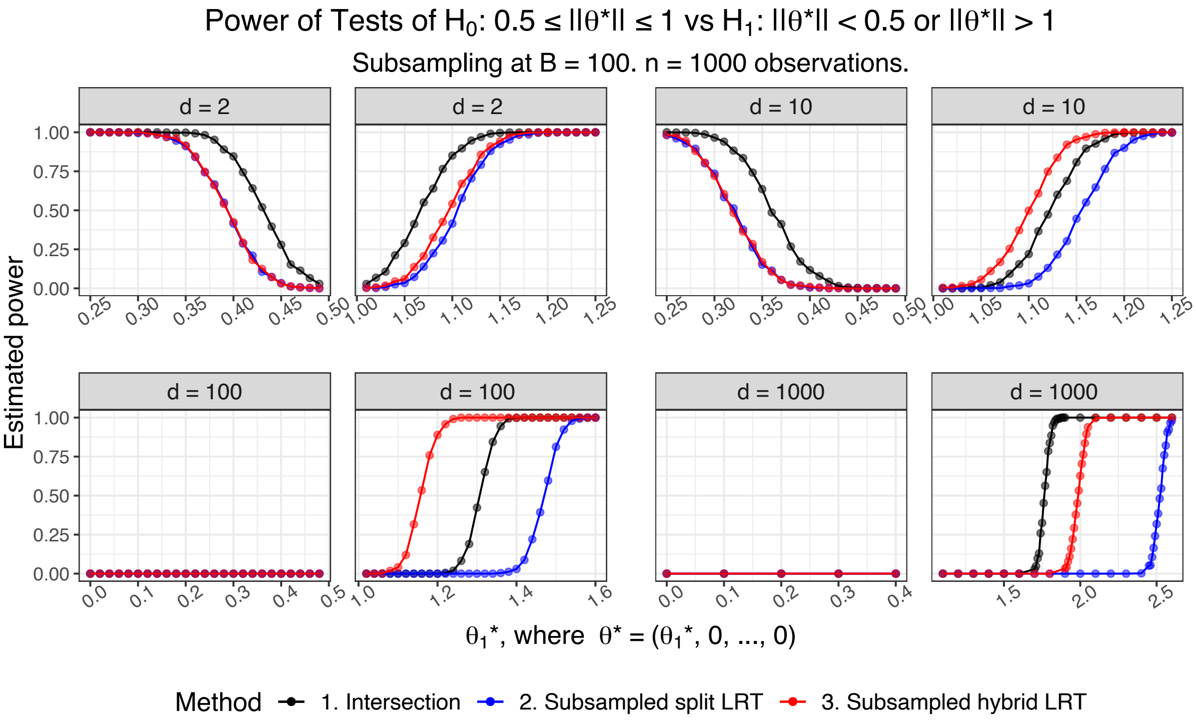


Figure 2.7: Estimated power of $H_0 : \|\theta^*\| \in [0.5, 1.0]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1.0]$ using the intersection, subsampled split LRT, and subsampled hybrid LRT methods. In these simulations, we set $\theta^* = (\theta_1^*, 0, \ldots, 0)$. The x-axis is the value of $\theta_1^* = \|\theta^*\|$ for each simulation. For each dimension, the left panel satisfies $\|\theta^*\| < 0.5$, and the right panel satisfies $\|\theta^*\| > 1$. We set $\alpha = 0.10$ and $n = 1000$, and we perform 1000 simulations at each value of $\|\theta^*\|$. We subsample $B = 100$ times.

same power. When $\|\theta^*\| > 1$, the hybrid test has higher power than the split test. We see that the intersection method always has higher power than the subsampled split LRT. One might consider whether we could combine the RIPR with the intersection method instead of combining the RIPR with the split LRT. It is unclear, though, how to construct a valid test from one approach that uses sample splitting and subsampling (RIPR) and a second approach that uses neither (intersection).

We can provide a partial theoretical justification for Fig. 2.7. For one, it is possible to derive an exact formula for the power of the intersection approach. Using the fact that $n\|\overline{Y}\|^2$ follows a non-central χ^2 distribution, we can write the power of the intersection method in terms of the non-central χ^2 CDF. When $d = 100$ or $d = 1000$, the hybrid method has no power at $\|\theta^*\| = 0$, though we would expect this case to have the highest power out of $\|\theta^*\| < 0.5$. At $d = 100$ and $\|\theta^*\| = 0$, the hybrid method satisfies $\|\overline{Y}_{1,b}\| < 0.5$ in most simulations, but the test statistic is too small to reject H_0. At $d = 1000$ and $\|\theta^*\| = 0$, $(n/2)\|\overline{Y}_{1,b}\|^2 \sim \chi^2_d$ is approximately d (Dasgupta and Schulman, 2007, Lemma 2). Hence $\|\overline{Y}_{1,b}\| \approx \sqrt{2}$, which means the hybrid approach selects the "incorrect" case of $\|\overline{Y}_{1,b}\| > 1$. This test also has approximately zero power. See Appendix B.4 for more details. In addition, for any given subsample, the hybrid LRT power is provably greater than or equal to the split LRT power. This holds because the RIPR test statistic is always larger than the split test statistic when both tests use the same numerator (Wasserman et al., 2020). The theoretical justification behind the relative power of the intersection and subsampled hybrid methods remains an open question, since the power of the latter method is not easily tractable.

2.5 Conclusion

The recent development of the universal LRT provides a hypothesis testing framework that is valid in finite samples and does not rely on regularity conditions. We have explored the performance of several universal LRT variants in the simple but fundamental case of testing for the mean θ^* when data arise from a $N(\theta^*, I_d)$ distribution. We have seen that even in high dimensions or for very small α, the ratio of the radius of the limiting subsampling universal LRT confidence set over an exact confidence set is less than 2. While the universal method tests the likelihood ratio against a dimension-independent cutoff, the universal LRT can still exhibit reasonable performance in high dimensions.

Future research directions may focus on settings where hypothesis tests were previously intractable or only asymptotically valid. Researchers can apply the universal LRT in any setting where it is possible to write a likelihood ratio or, more generally, upper bound the maximum likelihood under the null hypothesis. This allows for the development of valid tests for the number of components in mixture models and for log-concavity of the underlying density. Additionally, we have shown proof of concept that the universal LRT can be more powerful than existing valid tests. In the Gaussian setting, this phenomenon may apply more generally across other tests of non-convex null parameter spaces. Wasserman et al. (2020) also describe how the universal LRT can be used to test independence versus conditional independence in a Gaussian setting. Recent work by Guo and Richardson (2020) also provides a valid test in that setting, but the relative power of these two approaches is currently unknown.

Chapter 3

Universal Test for Log-concavity

3.1 Introduction

3.1.1 Log-concavity Definition and Properties

A log-concave density f has the form $f = e^g$ for some concave function g. This shape-constrained class of densities encompasses many common families of densities, such as the normal, uniform, exponential, logistic, and extreme value densities (e.g., Table 1 of Bagnoli and Bergstrom (2005)). Furthermore, specifying that the density is log-concave poses a middle ground between assumption-free density estimation and use of a parametric density family. As noted in Cule et al. (2010b), log-concave density estimation does not require the choice of a bandwidth, whereas kernel density estimation in d dimensions requires a $d \times d$ bandwidth matrix.

Bagnoli and Bergstrom (2005) describe applications of log-concavity across economics and reliability theory (or survival modeling). Suppose a survival density function f is defined on (a, b) and has a survival function (or reliability function) $\overline{F}(x) = \int_x^b f(t)dt$. If f is log-concave, then its survival function is log-concave as well. The failure rate associated with f is $r(x) = f(x)/\overline{F}(x) = -\overline{F}\,'(x)/\overline{F}(x)$. Corollary 2 of Bagnoli and Bergstrom (2005) states that if f is log-concave on (a, b), then the failure rate $r(x)$ is monotone increasing on (a, b). Proposition 12 of An (1997) states that if a survival function $\overline{F}(x)$ is log-concave, then for any pair of non-negative numbers x_1, x_2, the survival function satisfies $\overline{F}(x_1 + x_2) \leq \overline{F}(x_1)\overline{F}(x_2)$. This property is called the new-is-better-than-used property; it implies that the probability that a new unit will survive for time x_1 is greater than or equal to the probability that at time x_2, an existing unit will survive an additional time x_1.

An (1997) describes numerous properties of log-concave densities. As a few examples, log-concave densities are unimodal, they have at most exponential tails, all moments of the density exist, and they are closed under convolution. This means that if X and Y are independent random variables from log-concave densities, then the density of $X + Y$ is log-concave as well. A unimodal density f is strongly unimodal if the convolution of f with any unimodal density g is unimodal. Proposition 2 of An (1997) states that a density f is log-concave if and only if f is strongly unimodal.

Prior to Wasserman et al. (2020), there was no valid hypothesis test for $H_0 : f$ is log-concave versus $H_1 : f$ is not log-concave. We explore log-concave densities in more depth and examine universal tests of this hypothesis.

3.1.2 Solving for the Log-concave MLE

Suppose we observe an iid sample $X_1, \ldots, X_n \in \mathbb{R}^d$ from a d-dimensional density f^*, where $n \geq d+1$. Let $\mathcal{F}_d$ be the class of all log-concave densities in d dimensions. The log-concave maximum likelihood estimator is $\widehat{f}_n = \underset{f \in \mathcal{F}_d}{\arg\max} \sum_{i=1}^n \log\{f(X_i)\}$. Theorem 1 of Cule et al. (2010b) states that with probability 1, $\widehat{f}_n$ exists and is unique. Importantly, it is not necessary that $f^* \in \mathcal{F}_d$.

The construction of $\widehat{f}_n$ relies on the concept of a tent function $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$. For a given vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and given the sample $X_1, \ldots, X_n$, the tent function $\bar{h}_y$ is the least concave function that satisfies $\bar{h}_y(X_i) \geq y_i$ for $i = 1, \ldots, n$. Let C_n be the convex hull of the observations $X_1, \ldots, X_n$. Consider the objective function

$$\sigma(y_1, \ldots, y_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\{\bar{h}_y(x)\}dx.$$

Theorem 2 of Cule et al. (2010b) states that σ is a convex function, and σ has a unique minimum at the value $y^* \in \mathbb{R}^n$ that satisfies $\log(\widehat{f}_n) = \bar{h}_{y^*}$.

Thus, to find the tent function that defines the log-concave MLE, we need to minimize σ over $y \in \mathbb{R}^n$. σ is a convex function, but σ is not differentiable. Shor's algorithm (Shor, 2012) uses a subgradient method to optimize convex, non-differentiable functions (i.e., to find y^* that minimizes σ). This method is guaranteed to converge, but convergence can be slow. Shor's r-algorithm involves some computational speed-ups over Shor's algorithm, and Cule et al. (2010b) use this algorithm in their implementation. Shor's r-algorithm is not guaranteed to converge, but Cule et al. (2010b) state they agree with Kappel and Kuntsevich (2000) that the algorithm is "robust, efficient, and accurate." The `LogConcDEAD` package for log-concave density estimation in arbitrary dimensions implements this method (Cule et al., 2009).

As an alternative algorithm, the `logcondens` package implements an active set approach to solve for the log-concave MLE in one dimension (Dümbgen and Rufibach, 2011). This approach is based on solving for a vector that satisfies a set of active constraints and then using the tent function structure to compute the log-concave density associated with that vector. See Section 3.2 of Dümbgen et al. (2007) for more details.

Cule et al. (2010b) formalize the convergence of $\widehat{f}_n$. Let $D_{\mathrm{KL}}(g\|f)$ be the Kullback-Leibler divergence of g from f. Define $f^{\mathrm{LC}} = \underset{f\in\mathcal{F}_d}{\arg\min}\ D_{\mathrm{KL}}(f^*\|f)$. Hence, if $f^* \in \mathcal{F}_d$, then $f^{\mathrm{LC}} = f^*$. Regardless of whether $f^* \in \mathcal{F}_d$, suppose f^* satisfies the following conditions: $\int_{\mathbb{R}^d} \|x\| f^*(x)dx < \infty$, $\int_{\mathbb{R}^d} f^* \log_+(f^*) < \infty$ (where $\log_+(x) = \max\{\log(x), 0\}$), and the support of f^* contains an open set. By Lemma 1 of Cule et al. (2010a), there exists some $a_0 > 0$ and $b_0 \in \mathbb{R}$ such that $f^{\mathrm{LC}}(x) \le \exp(-a_0\|x\| + b_0)$ for any $x \in \mathbb{R}^d$. Theorem 3 of Cule et al. (2010b) states that for any $a < a_0$,

$$\int_{\mathbb{R}^d} \exp(a\|x\|)|\widehat{f}_n(x) - f^{\mathrm{LC}}(x)|dx \to 0 \qquad \text{almost surely.}$$

This means that the integrated difference between $\widehat{f}_n$ and f^{LC} converges to 0 even when we multiply the tails by some exponential weight. Furthermore, Theorem 3 states that if f^{LC} is continuous, then

$$\sup_{x\in\mathbb{R}^d} \left\{\exp(a\|x\|)|\widehat{f}_n(x) - f^{\mathrm{LC}}(x)|\right\} \to 0 \qquad \text{almost surely.}$$

In the case where $f^* \in \mathcal{F}_d$, it is possible to describe rates of convergence of the log-concave MLE in terms of the Hellinger distance. The Hellinger distance is given by

$$d_H^2(f, g) = \int_{\mathbb{R}^d} (f^{1/2} - g^{1/2})^2.$$

As stated in Chen et al. (2021) and shown in Kim et al. (2016) and Kur et al. (2019), the rate of convergence of $\widehat{f}_n$ to f^* in Hellinger distance is

$$\sup_{f^* \in \mathcal{F}_d} \mathbb{E}[d_H^2(\widehat{f}_n - f^*)] \leq K_d \cdot \begin{cases} n^{-4/5} & d = 1 \\ n^{-2/(d+1)} \log(n) & d \geq 2 \end{cases},$$

where $K_d > 0$ depends only on d.

3.2 Tests for Log-concavity

We have noted that the log-concave MLE $\widehat{f}_n$ has some favorable convergence properties. If one wishes to use the log-concave MLE as a density estimate, it is helpful to understand whether log-concavity is a reasonable assumption. To test $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$, we describe a permutation test as developed in Cule et al. (2010b), and we propose several universal tests. The universal tests are guaranteed to control the type I error at level α, while the permutation test is not guaranteed to be valid.

3.2.1 Permutation Test

Cule et al. (2010b) describe a permutation test of the hypothesis $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$. Algorithm 2 explains the permutation test.

Algorithm 2 Permutation test for $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$

```
    Input: $n$ iid $d$-dimensional observations $Y_1, \ldots, Y_n$, number of shuffles $B$, significance level $\alpha$.
    Output: Decision of whether to reject $H_0 : f \in \mathcal{F}_d$.
1: Fit the log-concave MLE $\widehat{f}_n$ on $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$.
2: Draw another sample $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_n^*\}$ from the log-concave MLE $\widehat{f}_n$.
3: Compute the test statistic $T = \sup_{A \in \mathcal{A}_0} |P_n(A) - P_n^*(A)|$, where $\mathcal{A}_0$ is the set of all balls
   centered at a point in $\mathcal{Y} \cup \mathcal{Y}^*$, $P_n(A)$ is the proportion of $\mathcal{Y}$ observations in $A$, and $P_n^*(A)$ is
   the proportion of $\mathcal{Y}^*$ observations in $A$.
4: for $b = 1, 2, \ldots, B$ do
5:     'Shuffle the stars' to randomly place $n$ observations from $\mathcal{Y} \cup \mathcal{Y}^*$ into $\mathcal{Y}_b$.
6:     Place the remaining $n$ observations in $\mathcal{Y}_b^*$.
7:     Using these new samples, compute $T_b^* = \sup_{A \in \mathcal{A}_0} |P_{n,b}(A) - P_{n,b}^*(A)|$. $P_{n,b}(A)$ is the
         proportion of $\mathcal{Y}_b$ observations in $A$, and $P_{n,b}^*(A)$ is the proportion of $\mathcal{Y}_b^*$ observations in $A$.
8: Arrange the test statistics $(T_1^*, T_2^*, \ldots, T_B^*)$ into the order statistics $(T_{(1)}^*, T_{(2)}^*, \ldots, T_{(B)}^*)$.
9: return Reject $H_0$ if $T > T^*_{(\lceil (B+1)(1-\alpha) \rceil)}$.
```

Intuitively, this test assumes that if H_0 is true, the samples $\mathcal{Y}$ and $\mathcal{Y}^*$ will be similar, so T will not be particularly large relative to $T_1^*, \ldots, T_B^*$. Alternatively, if H_0 is false, the samples $\mathcal{Y}$ and $\mathcal{Y}^*$ will be dissimilar, and the converse will hold. This approach is not guaranteed to control the type I error level. We will observe cases both where the permutation test performs well and where the permutation test's false positive rate is much higher than α.

We provide several computational notes on Algorithm 2. Steps 1 and 2 use functions from the `LogConcDEAD` library. To perform step 1, we can use the `mlelcd` function, which estimates the log-concave MLE density from a sample. To perform step 2, we can use the `rlcd` function, which samples from a fitted log-concave density. Where $\mathcal{A}_0$ is the set of all balls centered at a point in $\mathcal{Y} \cup \mathcal{Y}^*$, $|P_n(A) - P_n^*(A)|$ only takes on finitely many values over $A \in \mathcal{A}_0$. To see this, consider fixing a point at some value in $y \in \mathcal{Y} \cup \mathcal{Y}^*$, letting $A_r(y)$ be the sphere of radius r centered at y, and increasing r from 0 to infinity. As $r \to \infty$, $|P_n(A_r(y)) - P_n^*(A_r(y))|$ only changes when $A_r(y)$ expands to include an additional observation. This means there are only finitely many possible values of $|P_n(A) - P_n^*(A)|$ for $A \in \mathcal{A}_0$. Hence, it is possible to compute $\sup_{A \in \mathcal{A}_0} |P_n(A) - P_n^*(A)|$

by varying the radius of A over all distances between the center of A and another observation. For large n, it may be necessary to approximate the test statistics $T, T_1^*, T_2^*, \dots, T_B^*$ by varying the radius of A across a smaller set of fixed increments. In each of our simulations, we compute the test statistics exactly.

3.2.2 Universal Tests in d Dimensions

Alternatively, we can use universal approaches to test for log-concavity. Theorem 1.0.2 justifies the universal approach to testing whether the true density is in some potentially nonparametric class. Recall that the universal LRT provably controls the type I error level for any number of observations n. To implement the universal test on a single subsample, we partition the sample $\{Y_1, Y_2, \dots, Y_n\}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$. Let $\widehat{f}_0$ be the maximum likelihood log-concave density estimator fit on $\mathcal{D}_0$. Let $\widehat{f}_1$ be any density estimator fit on $\mathcal{D}_1$. The universal test rejects H_0 when

$$T_n = \prod_{Y_i \in \mathcal{D}_0} \{\widehat{f}_1(Y_i)/\widehat{f}_0(Y_i)\} \geq 1/\alpha.$$

Algorithm 3 explains how to compute a subsampled test statistic T_n for the test of $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$. In this case, we reject H_0 when $B^{-1} \sum_{b=1}^B T_{n,b} \geq 1/\alpha$.

Algorithm 3 Compute the subsampling test statistic T_n for $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$

```
Input: $n$ independent $d$-dimensional observations $Y_1, \dots, Y_n$ with density $f^*$,
  number of subsamples $B$, any density estimation approach.
Output: The subsampling test statistic $T_n$.
for $b = 1, 2, \dots, B$ do
    Randomly partition the data into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$, each containing $n/2$ values of $Y_i$.
    Where $\mathcal{L}_{0,b}(f) = \prod_{Y_i \in \mathcal{D}_{0,b}} f(Y_i)$, compute $\widehat{f}_{0,b} = \arg\max_{f \in \mathcal{F}_d} \mathcal{L}_{0,b}(f)$.
    Fit a density $\widehat{f}_{1,b}$ on $\mathcal{D}_1$, using the input density estimation approach.
    Compute $T_{n,b} = \mathcal{L}_{0,b}(\widehat{f}_{1,b})/\mathcal{L}_{0,b}(\widehat{f}_{0,b})$.
return the subsampling test statistic $T_n = B^{-1} \sum_{b=1}^B T_{n,b}$.
```

Both `logcondens` and `LogConcDEAD` provide functions to compute the log-concave MLE $\widehat{f}_0$. We have flexibility in the choice of $\widehat{f}_1$, which can be any density. We explore several choices of $\widehat{f}_1$.

Full Oracle

The full oracle approach uses the d-dimensional true density f^* in the numerator. In Algorithm 3, the input density estimation approach is to set $\widehat{f}_{1,b} = f^*$. This method is a helpful theoretical comparison, since it avoids the depletion in power that occurs when $\widehat{f}_{1,b}$ does not approximate f^* well. We would expect the power of this approach to be greater than or equal to the power of any approach that estimates a numerator on each subsample $\mathcal{D}_{1,b}$.

Partial Oracle

The partial oracle approach uses a d-dimensional parametric MLE density estimate in the numerator. Suppose we know (or we guess) that the true density is parameterized by some unknown real-valued vector $\theta^* \in \mathbb{R}^p$ such that $f^* = f_{\theta^*}$. Let $\mathcal{L}_{1,b}(f_\theta) = \prod_{Y_i \in \mathcal{D}_{1,b}} f_\theta(Y_i)$. In Algorithm 3, the input density estimation approach is to set $\widehat{f}_{1,b} = f_{\widehat{\theta}_{1,b}}$, where $\widehat{\theta}_{1,b} = \underset{\theta \in \mathbb{R}^p}{\arg\max}\, \mathcal{L}_{1,b}(f_\theta)$. If the true density is from the parametric family $(f_\theta : \theta \in \mathbb{R}^p)$, we would expect this method to have good power relative to other density estimation methods.

Fully Nonparametric

The fully nonparametric method uses a d-dimensional kernel density estimate (KDE) in the numerator. In Algorithm 3, the input density estimation approach is to set $\widehat{f}_{1,b}$ to the kernel density estimate computed on $\mathcal{D}_{1,b}$. Kernel density estimation involves the choice of a bandwidth. The `ks` package (Duong, 2021) in `R` can fit multidimensional KDEs and has several bandwidth computation procedures. These options include a plug-in bandwidth (Wand and Jones, 1994; Duong and Hazelton, 2003; Chacón and Duong, 2010), a least squares cross-validated bandwidth (Bowman, 1984; Rudemo, 1982), and a smoothed cross-validation bandwidth (Jones et al., 1991; Duong and Hazelton, 2005). We would not expect the fully nonparametric method to have as high of power as the full oracle method or as the partial oracle method in the parametric case. If we do not want to make assumptions about the true density, this may be a good choice.

3.2.3 Universal Tests with Dimension Reduction

Suppose each random variable $Y \in \mathbb{R}^d$ with associated density f^* can be written as $Y = (Y^{(1)}, \ldots, Y^{(d)})$. As noted in An (1997), if the density of Y is log-concave, then the marginal densities of $Y^{(1)}, \dots, Y^{(d)}$ are all log-concave. In the converse direction, if marginal densities of $Y^{(1)}, \dots, Y^{(d)}$ are all log-concave and $Y^{(1)}, \dots, Y^{(d)}$ are all independent, then Y is log-concave. More generally, Proposition 1(a) of Cule et al. (2010b) uses a result from Prékopa (1973) to deduce the following:

Theorem 3.1. *If V is a subspace of $\mathbb{R}^d$, $P_V(y)$ is the orthogonal projection of y onto V, and f^* is log-concave, then the marginal density of $P_V(Y)$ is log-concave.*

When considering how to test for log-concavity, An (1997) notes that univariate tests for log-concavity could be used in the multivariate setting. We use these results to develop new universal tests.

To reduce the data to one dimension, we take one of two approaches.

Dimension Reduction Approach 1: d Single Dimensions

We can represent any d-dimensional observation Y_i as $Y_i = (Y_i^{(1)}, Y_i^{(2)}, \dots, Y_i^{(d)})$. Algorithm 4 describes an approach that computes a test statistic for each of the d dimensions.

Algorithm 4 Compute d single dimension test statistics

```
Input: $n$ iid $d$-dimensional observations $Y_1, \dots, Y_n$, number of subsamples $B$.
Output: $d$ test statistics $T_n^{(j)}$, $j = 1, \dots, d$.
for $j = 1, 2, \dots, d$ do
    for $b = 1, 2, \dots, B$ do
        Randomly partition the $n$ observations $\{Y_1^{(j)}, Y_2^{(j)}, \dots, Y_n^{(j)}\}$ such that
          $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$ each contain $n/2$ values of $Y_i^{(j)}$.
        Estimate a one-dimensional density $\widehat{f}_{1,b}$ on $\mathcal{D}_{1,b}$.
        Estimate the log-concave MLE $\widehat{f}_{0,b}$ on $\mathcal{D}_{0,b}$.
    Compute the test statistic $T_n^{(j)} = B^{-1} \sum_{b=1}^{B} \prod_{Y_i^{(j)} \in \mathcal{D}_{0,b}} \{\widehat{f}_{1,b}(Y_i^{(j)}) / \widehat{f}_{0,b}(Y_i^{(j)})\}$
return the test statistics $T_n^{(j)}$, $j = 1, \dots, d$.
```

We reject H_0 if at least one of the d test statistics $T_n^{(1)}, \dots, T_n^{(d)}$ exceeds d/α. This rejection rule has valid type I error control because under H_0,

$$\mathbb{P}(T_n^{(1)} \geq d/\alpha \cup T_n^{(2)} \geq d/\alpha \cup \cdots \cup T_n^{(d)} \geq d/\alpha) \leq d \cdot P(T_n^{(1)} \geq d/\alpha) \leq d(\alpha/d) = \alpha.$$

Dimension Reduction Approach 2: Random Projections

We can also construct one-dimensional densities by projecting the data onto a vector draw uniformly from the unit sphere. Algorithm 5 shows how to compute the random projection test statistic T_n.

Algorithm 5 Compute the random projection test statistic T_n

```
Input: $n$ iid $d$-dimensional observations $Y_1, \ldots, Y_n$, number of subsamples $B$,
  number of random projections $n_{\text{proj}}$.
Output: The random projection test statistic $T_n$.
for $k = 1, 2, \ldots, n_{\text{proj}}$ do
    Draw a vector $V$ uniformly from the $d$-dimensional unit sphere. To obtain $V$,
      draw $X \sim N(0, I_d)$ and set $V = X/\|X\|$.
    Project each $Y$ observation onto $V$. The projection of $Y_i$ is $P_V(Y_i) = Y_i^T V$.
    for $b = 1, 2, \ldots, B$ do
        Randomly partition the $P_V(Y_i)$ observations such that
          $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$ each contain $n/2$ values of $P_V(Y_i)$.
        Estimate a one-dimensional density $\widehat{f}_{1,b}$ on $\mathcal{D}_{1,b}$.
        Estimate the log-concave MLE $\widehat{f}_{0,b}$ on $\mathcal{D}_{0,b}$.
    Compute the test statistic $T_{n,k} = B^{-1} \sum_{b=1}^{B} \prod_{P_V(Y_i) \in \mathcal{D}_{0,b}} \{\widehat{f}_{1,b}(P_V(Y_i))/\widehat{f}_{0,b}(P_V(Y_i))\}$
return the random projection test statistic $T_n = n_{\text{proj}}^{-1} \sum_{k=1}^{n_{\text{proj}}} T_{n,k}$.
```

T_n is an average of $B \cdot n_{\text{proj}}$ test statistics that are each e-variables. Hence, T_n is also an e-variable. A valid level $1 - \alpha$ test rejects H_0 if $T_n \geq 1/\alpha$.

In both dimension reduction approaches, we fit some one-dimensional density $\widehat{f}_{1,b}$ on $\mathcal{D}_1$. We consider two density estimation methods.

Density Estimation Method 1: Partial Oracle

This approach uses parametric knowledge about the underlying density. The numerator $\widehat{f}_{1,b}$ is the parametric MLE fit on $\mathcal{D}_{1,b}$.

Density Estimation Method 2: Fully Nonparametric

This approach does not use any prior knowledge about the underlying density. Instead, we use kernel density estimation (e.g., `ks` package with plug-in bandwidth) to fit $\widehat{f}_{1,b}$.

Thus, for the universal LRTs with dimension reduction, we consider four total combinations of two dimension reduction approaches and two density estimation methods.

3.3 Example 1: Testing Log-concavity of Normal Mixture

3.3.1 Setup

Suppose ϕ_d is the $N(0, I_d)$ density. Cule et al. (2010b) note that for $\pi \in (0, 1)$, the normal location mixture $f(x) = \pi\phi_d(x) + (1 - \pi)\phi_d(x - \mu)$ is log-concave only when $\|\mu\| \leq 2$. As some intuition for the one-dimensional case, log-concave densities are unimodal and have at most exponential tails. That is, $f(x) = o(\exp(-cx))$ for some $c > 0$ (An, 1997). A one-dimensional mixture of two Gaussians with means μ_1, μ_2 and standard deviations σ_1, σ_2 is unimodal if $|\mu_2 - \mu_1| \leq 2\min\{\sigma_1, \sigma_2\}$ (Sitek, 2016). We explore how the validity and power of the tests varies over $\|\mu\|$. We use $\pi = 0.5$ in all analyses.

3.3.2 Visualizing Log-concave MLEs

We begin by visualizing the log-concave MLEs of several samples from two-component Gaussian mixtures. The underlying density is

$$f^*(x) = 0.5\phi_d(x) + 0.5\phi_d(x - \mu).$$

These simulations help us to see the log-concave MLE outputs for both small and large sample sizes and for both log-concave ($\|\mu\| \leq 2$) and not log-concave ($\|\mu\| > 2$) true densities.

In the one-dimensional setting, we compute the log-concave MLEs $\widehat{f}_n$ on samples $\{x_1, \ldots, x_n\}$. Figures 3.1 and 3.2 show the true and log-concave MLE densities for samples with $n = 50$ and $n = 5000$, respectively. These simulations use both the `LogConcDEAD` and `logcondens` packages. `logcondens` only works in one dimension but is much faster than `LogConcDEAD`. Visually, we see that these two packages produce approximately the same densities. Furthermore, we include values of $n^{-1}\sum_{i=1}^{n}\log(f^*(x_i))$ on the true density plots and $n^{-1}\sum_{i=1}^{n}\log(\widehat{f}_n(x_i))$ on the log-concave MLE plots. The log likelihood is approximately the same for the two density estimation methods.

When $\|\mu\| = 0$ or $\|\mu\| = 2$, the true density is log-concave. As we increase from $n = 50$ to $n = 5000$, the log-concave MLE becomes a better approximation to the true density. We see this improvement both visually and numerically. That is, $n^{-1}\sum_{i=1}^{n}\log(\widehat{f}_n(x_i))$ is closer to $n^{-1}\sum_{i=1}^{n}\log(f^*(x_i))$ for larger n. When $\|\mu\| = 4$, the underlying density is not log-concave. The $\|\mu\| = 4$ log-concave MLE at $n = 5000$ seems to have normal tails, but it is nearly uniform in the middle.

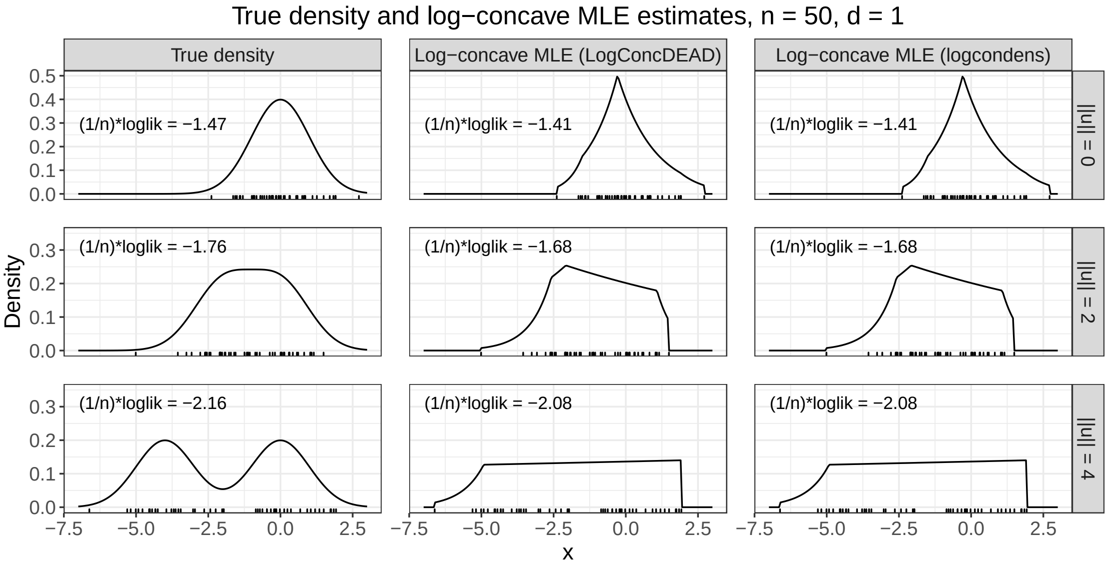


Figure 3.1: Density plots from fitting log-concave MLE on $n = 50$ observations. Tick marks represent the observations. The true density is the Normal mixture $f^*(x) = 0.5\phi_1(x)+0.5\phi_1(x-\mu)$.

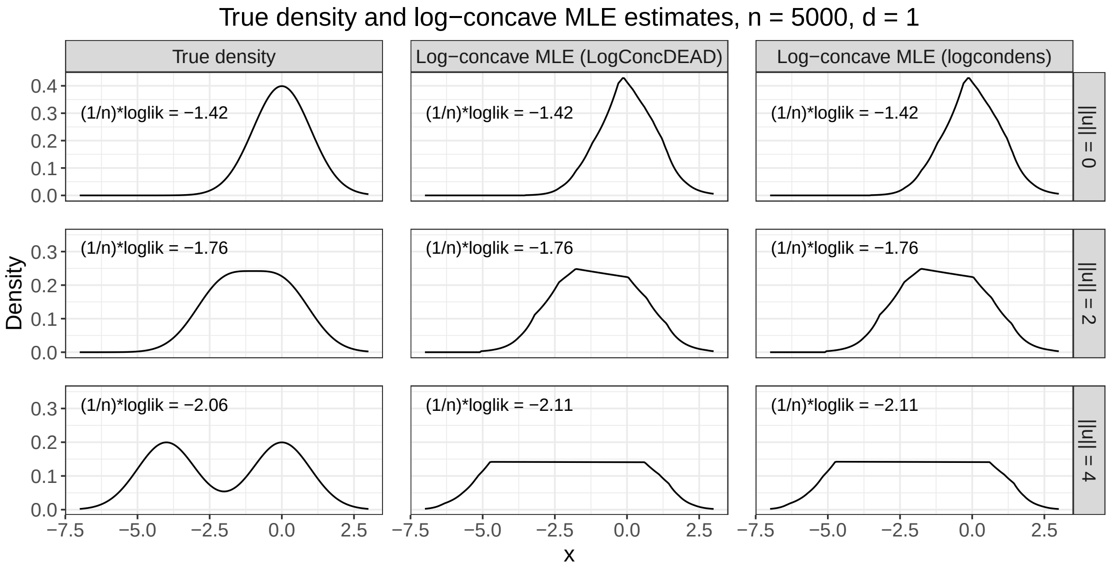


Figure 3.2: Density plots from fitting log-concave MLE on $n = 5000$ observations. The true density is the Normal mixture $f^*(x) = 0.5\phi_1(x) + 0.5\phi_1(x - \mu)$.

We observe similar behavior in the two-dimensional setting. In two dimensions, we use $\mu = (\|\mu\|, 0)$. Figures 3.3 and 3.4 show two-dimensional contour plots for the true and log-concave MLEs with $n = 50$ and $n = 500$. In Figure 3.3, we can clearly see that support of the log-concave MLE is the convex hull of the observed sample. For $\|\mu\| = 0$ and $\|\mu\| = 2$, the true density and log-concave MLE have more similar appearances when $n = 500$. In addition, $n^{-1}\sum_{i=1}^{n}\log(\widehat{f}_n(x_i))$ is closer to $n^{-1}\sum_{i=1}^{n}\log(f^*(x_i))$ for larger n. When $\|\mu\| = 4$, the log-concave MLE density is nearly flat in the center of the density.

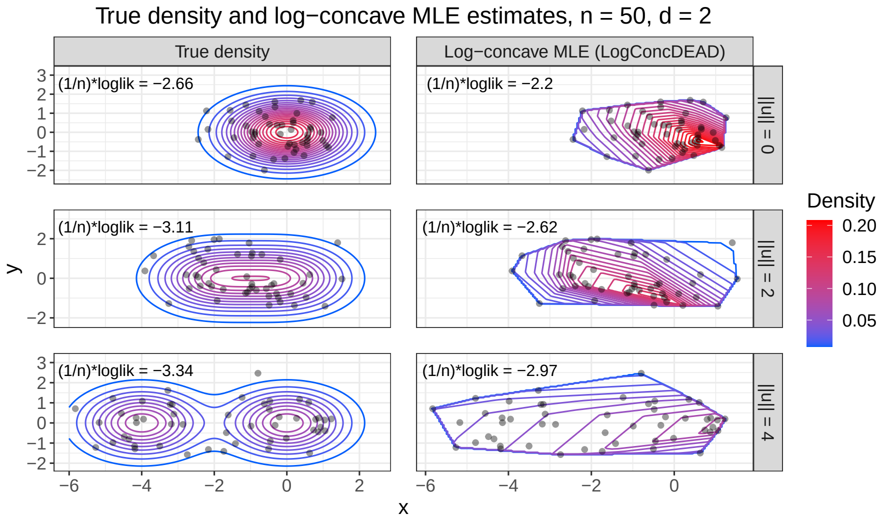


Figure 3.3: Contour plots from fitting log-concave MLE on $n = 50$ observations. Points represent the 50 observations. The true density is the Normal mixture $f^*(x) = 0.5\phi_2(x) + 0.5\phi_2(x - \mu)$.

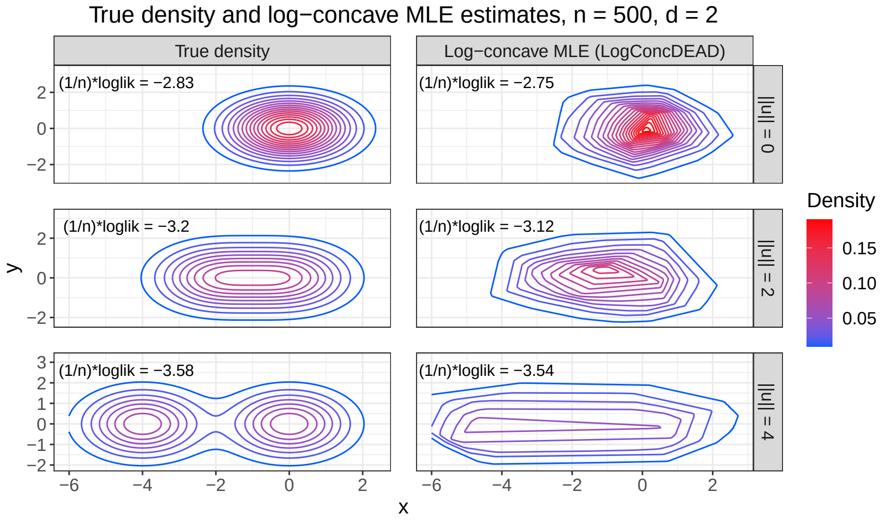


Figure 3.4: Contour plots from fitting log-concave MLE on $n = 500$ observations. The true density is the Normal mixture $f^*(x) = 0.5\phi_2(x) + 0.5\phi_2(x - \mu)$.

3.3.3 Permutation Test is Not Always Valid

To test $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$, we start by consider the permutation test. Cule et al. (2010b) simulate this test in the $d = 2$ case, using the underlying density $f^*(x) = 0.5\phi_d(x) + 0.5\phi_d(x - \mu)$. For $\|\mu\| \leq 2$ (when H_0 is true), they find that the test is valid but conservative. Setting $\alpha = 0.05$, the rejection proportion at $n = 200$ is 0.015, and the rejection proportion at

$n = 1000$ is 0.005. At $\|\mu\| = 4$ (when H_0 is false), the test has high power for large n. The rejection proportion at $n = 200$ is 0.475, and the rejection proportion at $n = 1000$ is 0.995.

We now simulate this test for $1 \leq d \leq 5$ and $n = 100$. We set $\mu = (\mu_1, 0, \ldots, 0)$, so that $\|\mu\| = \|\mu_1\|$. We use a significance level of $\alpha = 0.10$. Each point represents the proportion of times we reject H_0 over 200 simulations. We shuffle $B = 99$ times, which is the same choice of B as in Cule et al. (2010b). Figure 3.5 shows that the test is valid at $d = 1$ and $d = 2$ and approximately valid at $d = 3$. Alternatively, at $d = 4$ and $d = 5$, this test rejects H_0 at proportions much higher than α, even when the underlying density is log-concave ($\|\mu\| \leq 2$).

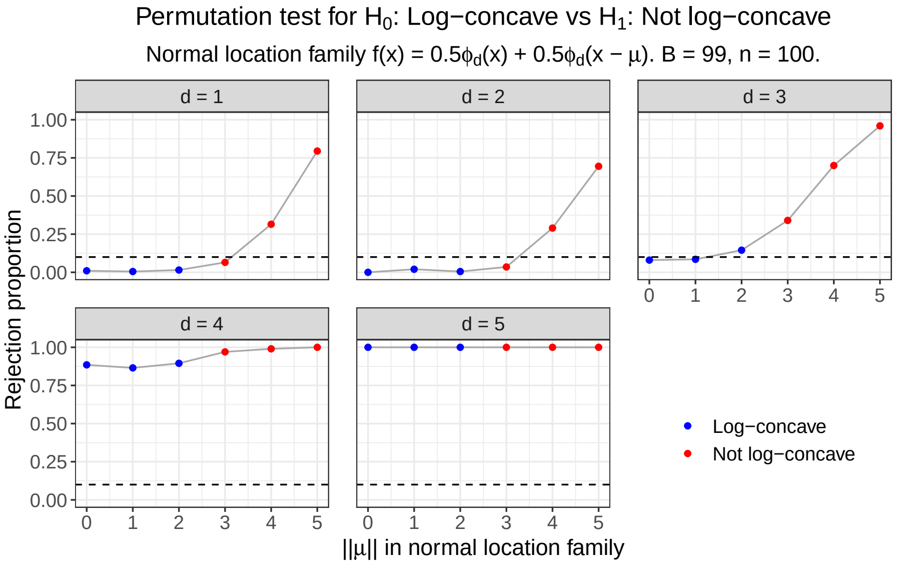


Figure 3.5: Rejection proportions for test of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave, using the permutation test implementation from Cule et al. (2010b). We set $\alpha = 0.10$ and $n = 100$, and we perform 200 simulations at each combination of $(d, \|\mu\|)$. The test permutes the observations $B = 99$ times.

We consider whether these results still hold with a larger sample size. Figure 3.6 repeats these simulations at $n = 250$. Compared to the $n = 100$ setting, this setting has slightly higher power at $\|\mu\| = 4$ and $\|\mu\| = 5$ when $d = 1$ or $d = 2$. We still see that the rejection proportion is much higher than 0.10 for $\|\mu\| \leq 2$ and $d = 4$ or $d = 5$.

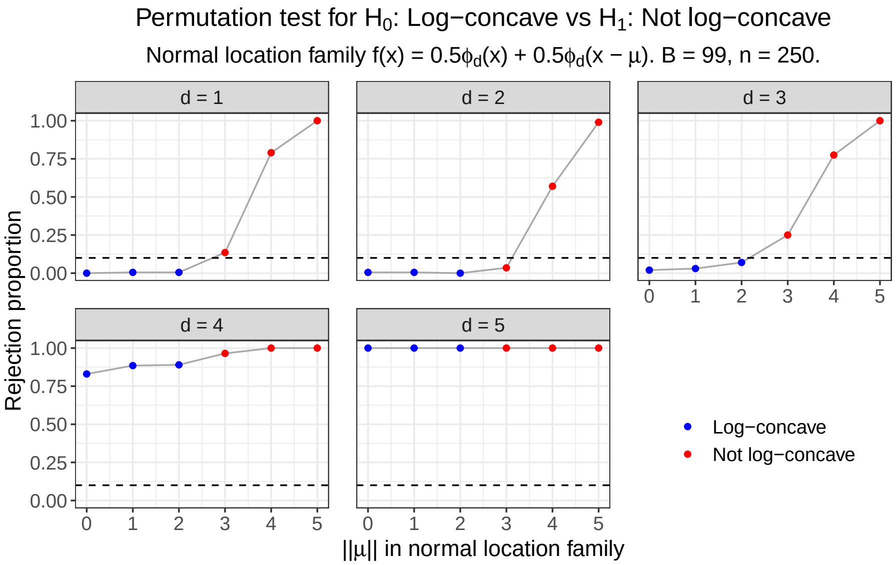


Figure 3.6: Rejection proportions for test of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave, using the permutation test implementation from Cule et al. (2010b). We set $\alpha = 0.10$ and $n = 250$, and we perform 200 simulations at each combination of $(d, \|\mu\|)$. The test permutes the observations $B = 99$ times.

Next, we consider whether the results hold if we increase B, the number of times that we shuffle the sample. In Figure 3.7, we show the results of simulations at $B \in \{100, 200, 300, 400, 500\}$ on $n = 100$ observations. Each row corresponds to the same set of simulations performed at five values of B. Looking across each row, we do not see an effect as B increases from 100 to 500. In these analyses, the lack of validity at $d = 4$ and $d = 5$ remains as we increase n or increase B.

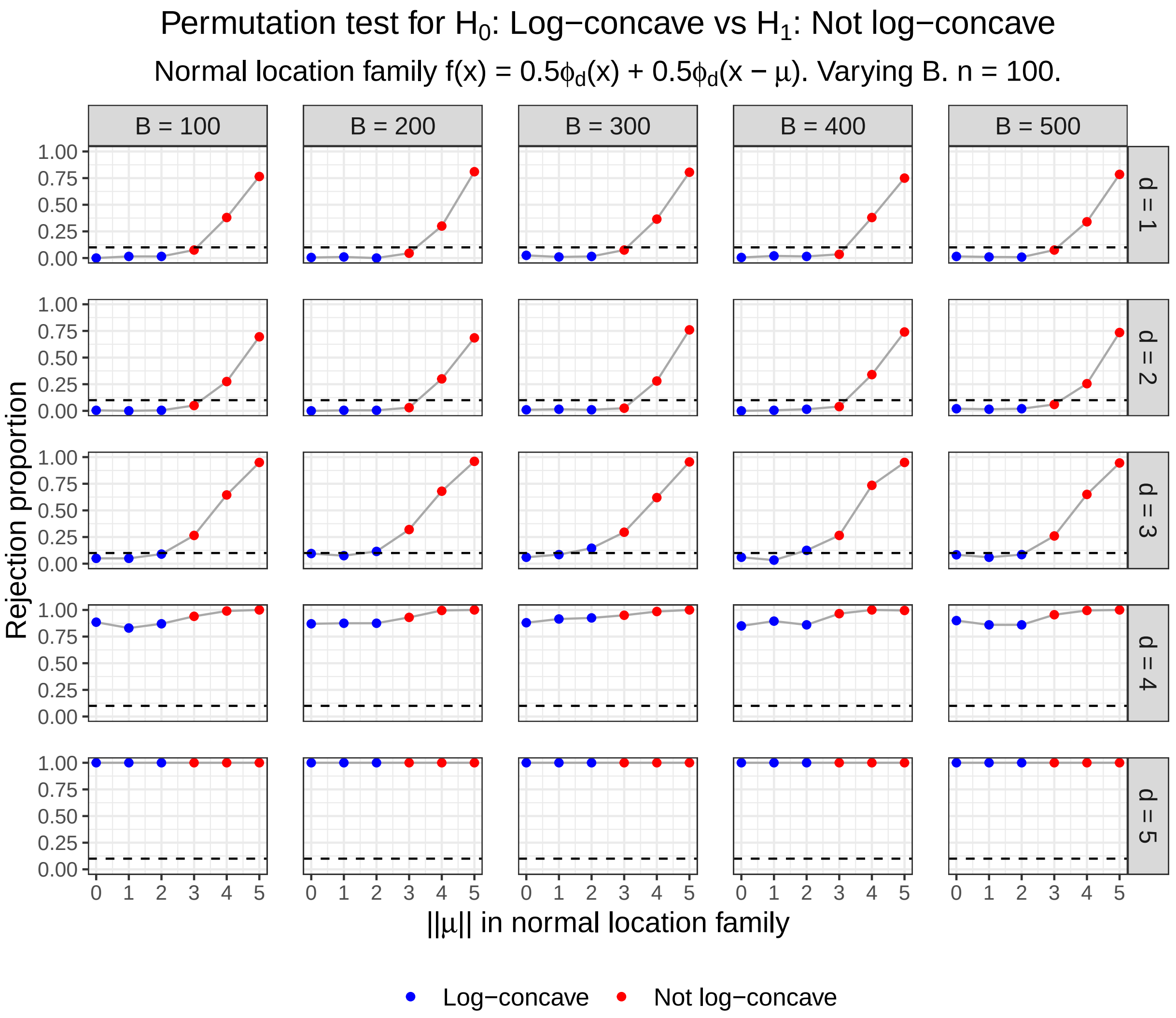


Figure 3.7: Rejection proportions for test of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave, using the permutation test implementation from Cule et al. (2010b). We set $\alpha = 0.10$ and $n = 100$, and we perform 200 simulations at each combination of $(B, d, \|\mu\|)$.

Recall that the test statistic is $T = \sup_{A \in \mathcal{A}_0} |P_n(A) - P_n^*(A)|$, and the test statistic on a shuffled sample is $T_b^* = \sup_{A \in \mathcal{A}_0} |P_{n,b}(A) - P_{n,b}^*(A)|$. Both P and P^* are proportions (out of n observations), so T and T_b^* can only take on finitely many values. We consider whether the conservativeness of the test (e.g., $d = 1$) or the lack of validity of test (e.g., $d = 5$) is due to this discrete nature. Figure 3.8 plots the distribution of shuffled test statistics T_b^* across eight simulations. The left panels consider the $d = 1$ case at all combinations of $\|\mu\| \in \{0, 2\}$ and $B \in \{100, 500\}$. We see that "bunching" of the quantiles is not responsible for the test being conservative in this case. (For instance, if the

90^{th} percentile were equal to the 99^{th} percentile, then it would make sense for the method to be conservative at $\alpha = 0.10$.) Instead, the 0.90, 0.95, and 0.99 quantiles (dashed blue lines) are all distinct, and the original test statistic (solid black line) is less than each of these values. We also consider the behavior in the $d = 5$ case (right panels). Again, these three quantiles are all distinct. In this case, though, the original test statistic is in the far right tail of the distribution of shuffled data test statistics.

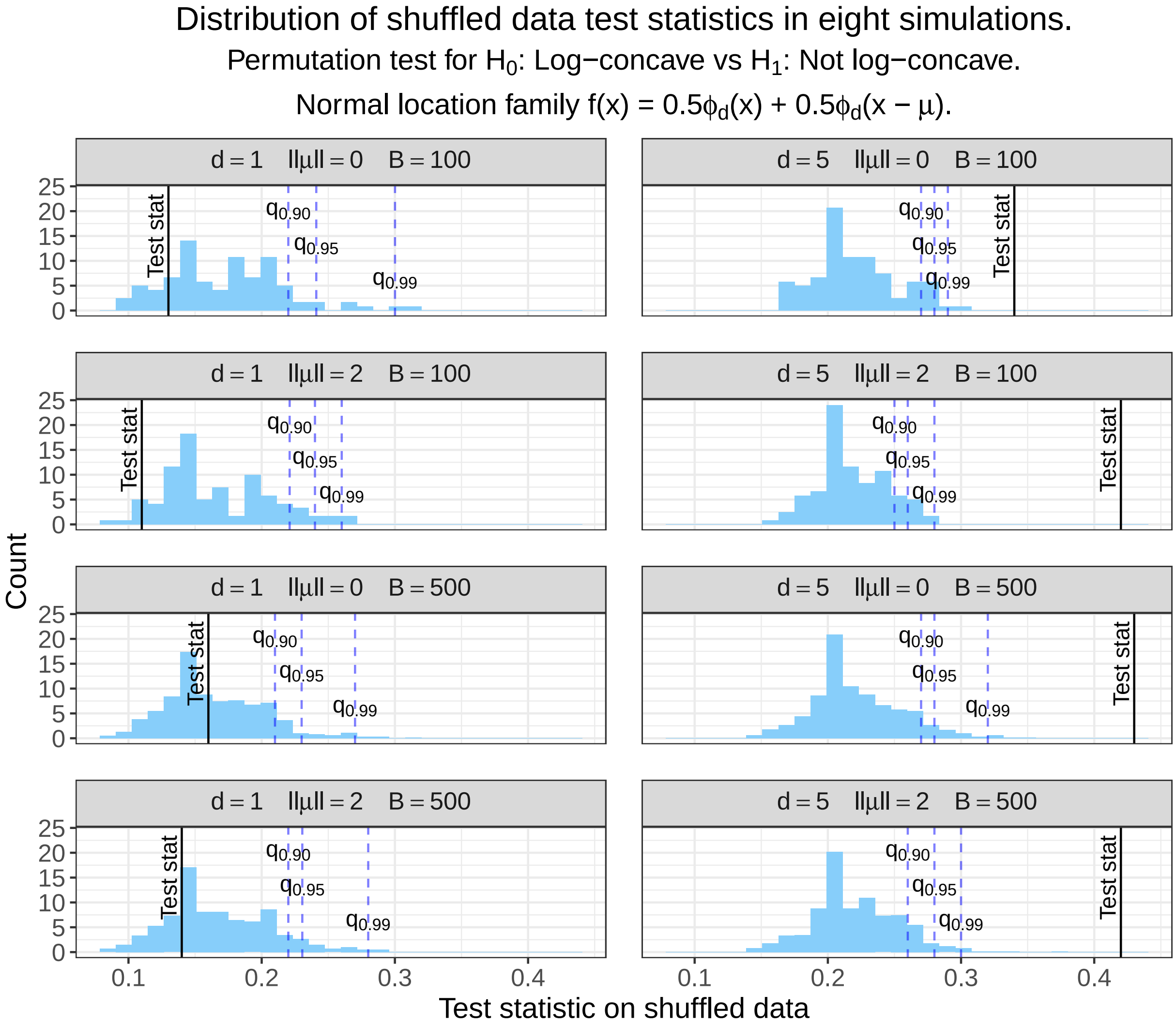


Figure 3.8: Distribution of $T^*_{n,b}$ across eight simulations. The dashed blue lines correspond to the quantiles of the distribution of shuffled data test statistics. The solid black lines correspond to the original test statistics in each simulation.

3.3.4 Full Oracle has Inadequate Power

We compare the permutation test to the full oracle universal test. Figure 3.9 shows the power for $d \in \{1, 2, 3, 4\}$ on $n = 100$ observations. For the universal test, we subsample $B = 100$ times. Each point is the rejection proportion over 200 simulations. For some $\|\mu\|$ values when $d = 1$, the full oracle test has higher power than the permutation test. For most $(d, \|\mu\|)$ combinations, though, the full oracle test has lower power than the permutation test. Unlike the permutation test, the universal test is valid for $d \geq 4$.

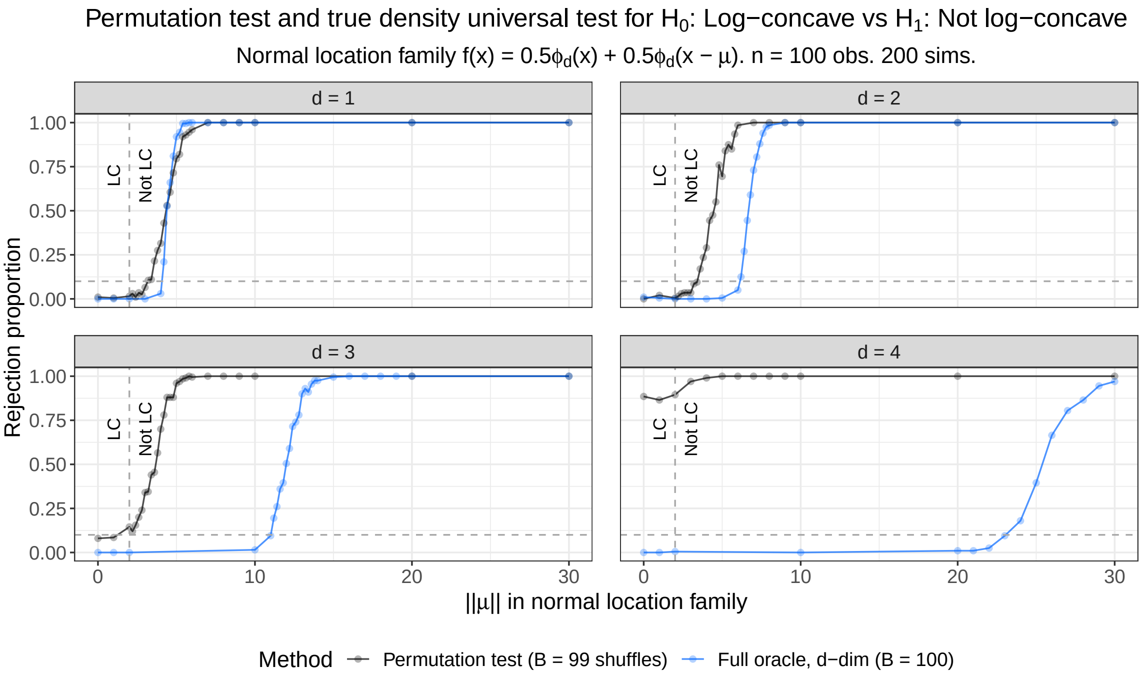


Figure 3.9: Rejection proportions for test of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave. The universal approach uses the true density in the numerator.

From Figure 3.9, we can see that as d increases, we need larger $\|\mu\|$ for the test to have power. Figure 3.10 formalizes this relationship, by exploring how $\|\mu\|$ needs to grow with d to maintain power of approximately 0.90. For each value of d, we vary $\|\mu\|$ in increments of 1 and estimate the power through 200 simulations. We choose the value of $\|\mu\|$ with power closest to 0.90. If none of the $\|\mu\|$ values have power in the range of $[0.88, 0.92]$ at a given d, then we use finer-grained values of $\|\mu\|$. From the best fit curve, it appears that $\|\mu\|$ needs to grow at an exponential rate in d to maintain the same power. Thus, while the full oracle approach offers an improvement in validity over the permutation test, the power becomes substantially worse in higher dimensions.

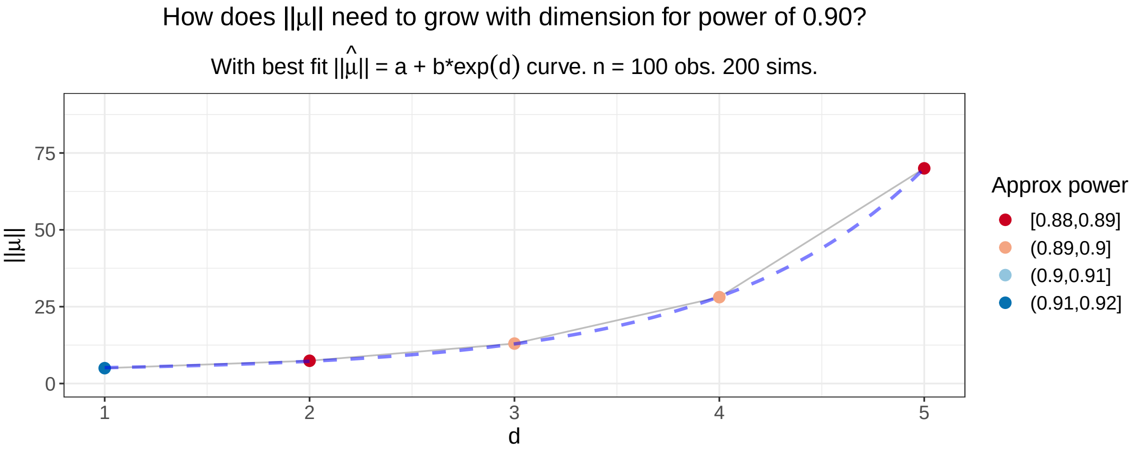


Figure 3.10: For test of H_0 : f^* is log-concave versus H_1 : f^* is not log-concave, how does $\|\mu\|$ grow with d to maintain power of 0.90? Universal test with true density numerator and $B = 100$ subsamples on $n = 100$ observations.

3.3.5 Superior Performance of Dimension Reduction Approaches

We have seen that the partial oracle universal LRT requires $\|\mu\|$ to grow exponentially to maintain power as d increases. We turn to the dimension reduction universal LRT approaches, and we find that they produce higher power for smaller $\|\mu\|$ values.

We implement all four combinations of the two dimension reduction approaches and the two density estimation methods. We compare them to three d-dimensional approaches: the permutation test, the full oracle test, and the partial oracle test. The full oracle d-dimensional approach uses the split LRT with the true density in the numerator and the d-dimensional log-concave MLE in the denominator. The partial oracle d-dimensional approach uses the split LRT with a two component Gaussian mixture in the numerator and the d-dimensional log-concave MLE in the denominator. We fit the Gaussian mixture using the EM algorithm, as implemented in the `mclust` package (Scrucca et al., 2016).

Figures 3.11 and 3.12 compare the four dimension reduction approaches and the d-dimensional approaches. The six universal approaches subsample at $B = 100$. The random projection approaches set $n_{\text{proj}} = 100$. The permutation test uses $B = 99$ permutations to determine the significance level of the original test statistic. Both figures use the normal location model $f^*(x) =$

$0.5\phi_d(x) + 0.5\phi_d(x - \mu)$ as the underlying model. However, Figure 3.11 uses $\mu = (\|\mu\|, 0, \ldots, 0)$, while Figure 3.12 uses $\mu = (\|\mu\|d^{-1/2}, \|\mu\|d^{-1/2}, \ldots, \|\mu\|d^{-1/2})$.

There are several key takeaways from Figures 3.11 and 3.12. The universal approaches that fit one-dimensional densities (d single dimensions and random projections) have higher power than the universal approaches that fit d-dimensional densities. (When $d = 1$, the "Partial oracle, d single dims" and "Partial oracle, d-dim" approaches are the same.) The permutation test is not always valid, especially for $d \geq 4$.

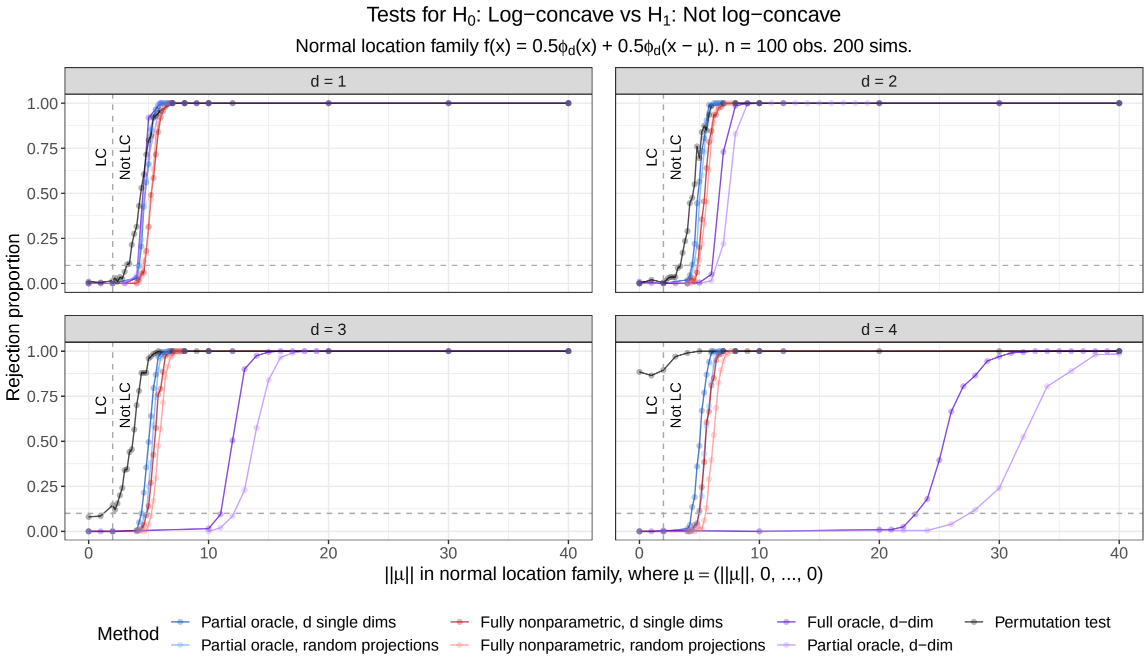


Figure 3.11: Power of tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave. μ vector for second component is $\mu = (\|\mu\|, 0, \ldots, 0)$.

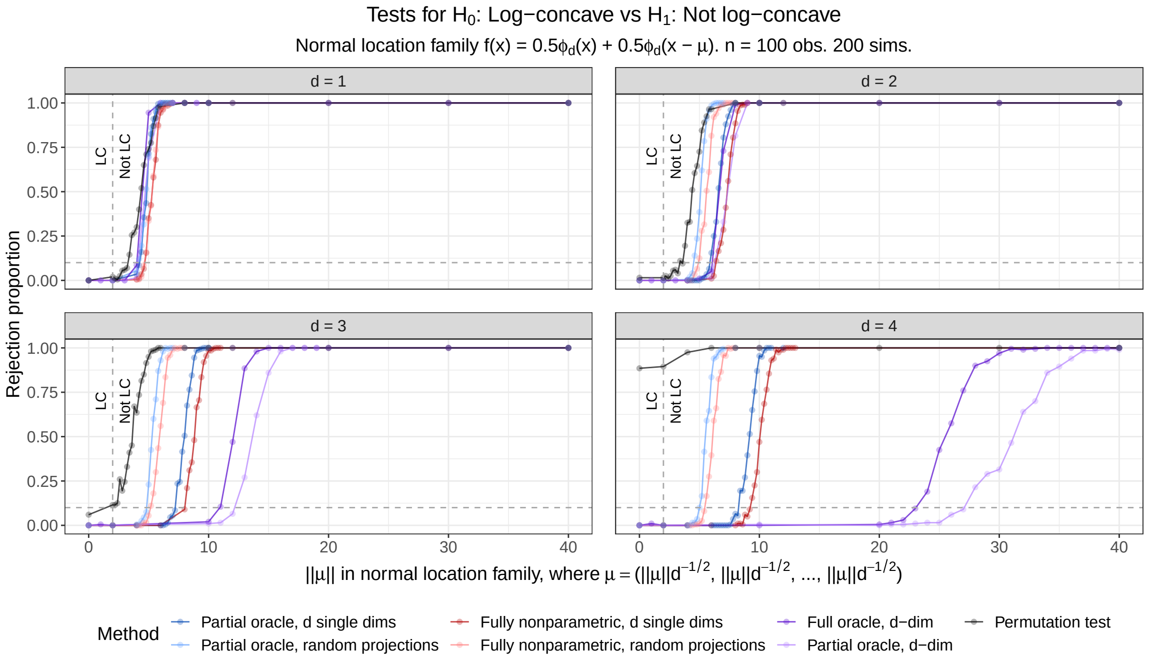


Figure 3.12: Power of tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave. μ vector for second component is $\mu = (\|\mu\|d^{-1/2}, \|\mu\|d^{-1/2}, \ldots, \|\mu\|d^{-1/2})$.

To compare the four universal approaches that fit one-dimensional densities, we consider Figures 3.13 and 3.14, which zoom in on a smaller range of $\|\mu\|$ values for those four methods. In both Figure 3.13 and 3.14, for a given dimension reduction approach (d single dimensions or random projections), the partial oracle approach has slightly higher power than the fully nonparametric approach. (That is, the dark blue curve has higher power than the dark red curve, and the light blue curve has higher power than the light red curve.) For a given density estimation approach, the dimension reduction approach with higher power changes based on the setting. When $d = 1$, the two partial oracle methods are equivalent, and the two fully nonparametric methods are equivalent. When $\mu = (\|\mu\|, 0, \ldots, 0)$ (Figure 3.13), the d single dimensions approach has higher power than the random projections approach. (That is, dark blue has higher power than light blue, and dark red has higher power than light red.) This makes sense because a single dimension contains all of the signal. When $\mu = (\|\mu\|d^{-1/2}, \|\mu\|d^{-1/2}, \ldots, \|\mu\|d^{-1/2})$ (Figure 3.14), the random projections approach has higher power than the d single dimensions approach. (That is, light blue has higher power than dark blue, and light red has higher power than dark red.) This makes sense because all directions have some evidence against H_0, and there are exist linear combinations of the coordinates that have higher power than any individual dimension.

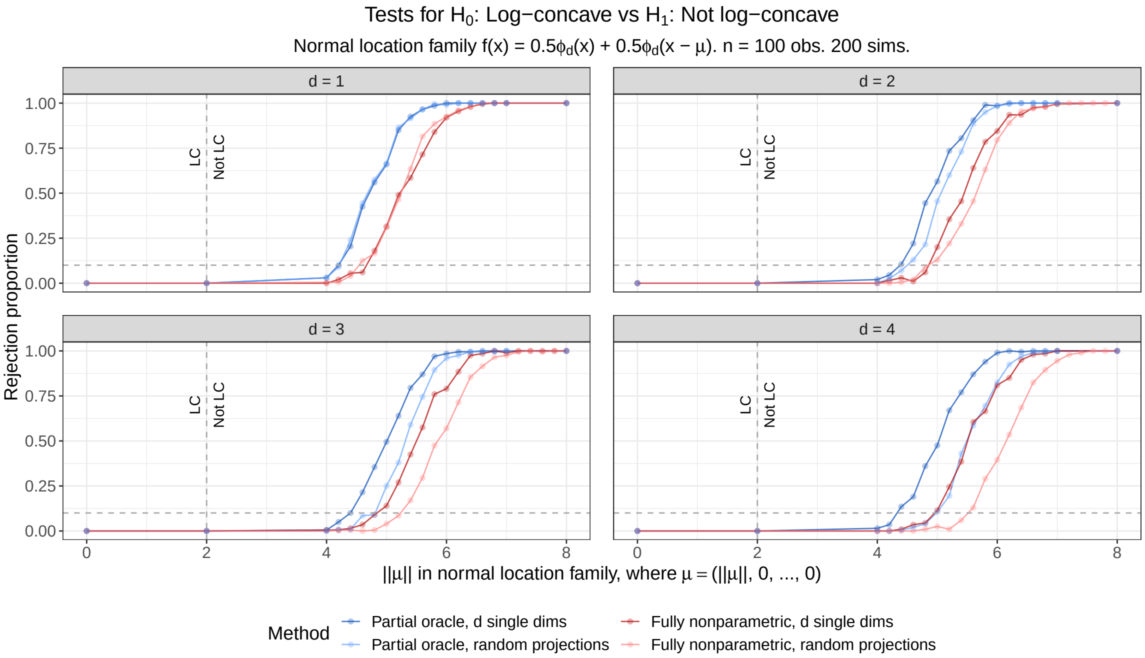


Figure 3.13: Power of four dimension-reduced tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave. μ vector for second component is $\mu = (\|\mu\|, 0, \ldots, 0)$.

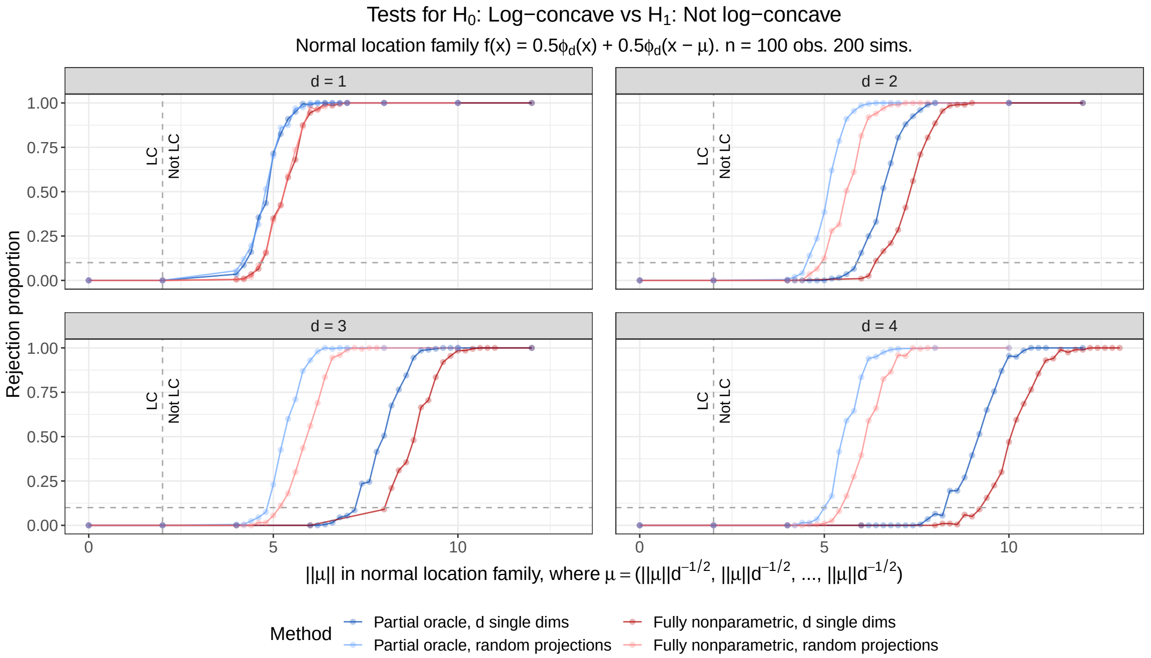


Figure 3.14: Power of four dimension-reduced tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave. μ vector for second component is $\mu = (\|\mu\|d^{-1/2}, \|\mu\|d^{-1/2}, \ldots, \|\mu\|d^{-1/2})$.

3.4 Example 2: Testing Log-concavity of Beta Density

3.4.1 Setup

In the one-dimensional normal mixture case, we saw that the full oracle universal test sometimes had higher power than the permutation test. We consider whether this holds in another one-dimensional setting.

The Beta(α, β) density has the form

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0, 1),$$

where $\alpha > 0$ and $\beta > 0$ are shape parameters.

As noted in Cule et al. (2010b), Beta(α, β) is log-concave if $\alpha \geq 1$ and $\beta \geq 1$. We can see this in a quick derivation:

$$\begin{aligned}
\frac{\partial^2}{\partial x^2} \log f(x; \alpha, \beta) &= \frac{\partial^2}{\partial x^2} \left[\log \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right) + (\alpha - 1)\log(x) + (\beta - 1)\log(1-x) \right] \\
&= \frac{\partial}{\partial x} \left[\frac{\alpha - 1}{x} + \frac{1-\beta}{1-x} \right] \\
&= \frac{1-\alpha}{x^2} + \frac{1-\beta}{x^2}.
\end{aligned}$$

This second derivative is less than or equal to 0 for all $x \in (0,1)$ only if both $\alpha \geq 1$ and $\beta \geq 1$. The Beta(α, β) distribution is hence log-concave when $\alpha \geq 1$ and $\beta \geq 1$. This means that tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave should reject H_0 if $\alpha < 1$ or $\beta < 1$.

3.4.2 Visualizing Log-concave MLEs

In general, it is non-trivial to solve for the limiting log-concave function $f^{\mathrm{LC}} = \underset{f \in \mathcal{F}_d}{\arg\min}\, D_{\mathrm{KL}}(f^* \| f)$. We try to determine f^{LC} in a few specific cases. In Figure 3.15, we consider two choices of shape parameters (α, β) such that the Beta(α, β) densities are not log-concave. On the left panels, we plot the Beta densities. For the right panels, we simulate 100,000 observations from the corresponding Beta(α, β) density, we fit the log-concave MLE on the sample using `logcondens`, and we plot this log-concave MLE density. Thus, the right panels should be good approximations to f^{LC} in these two settings.

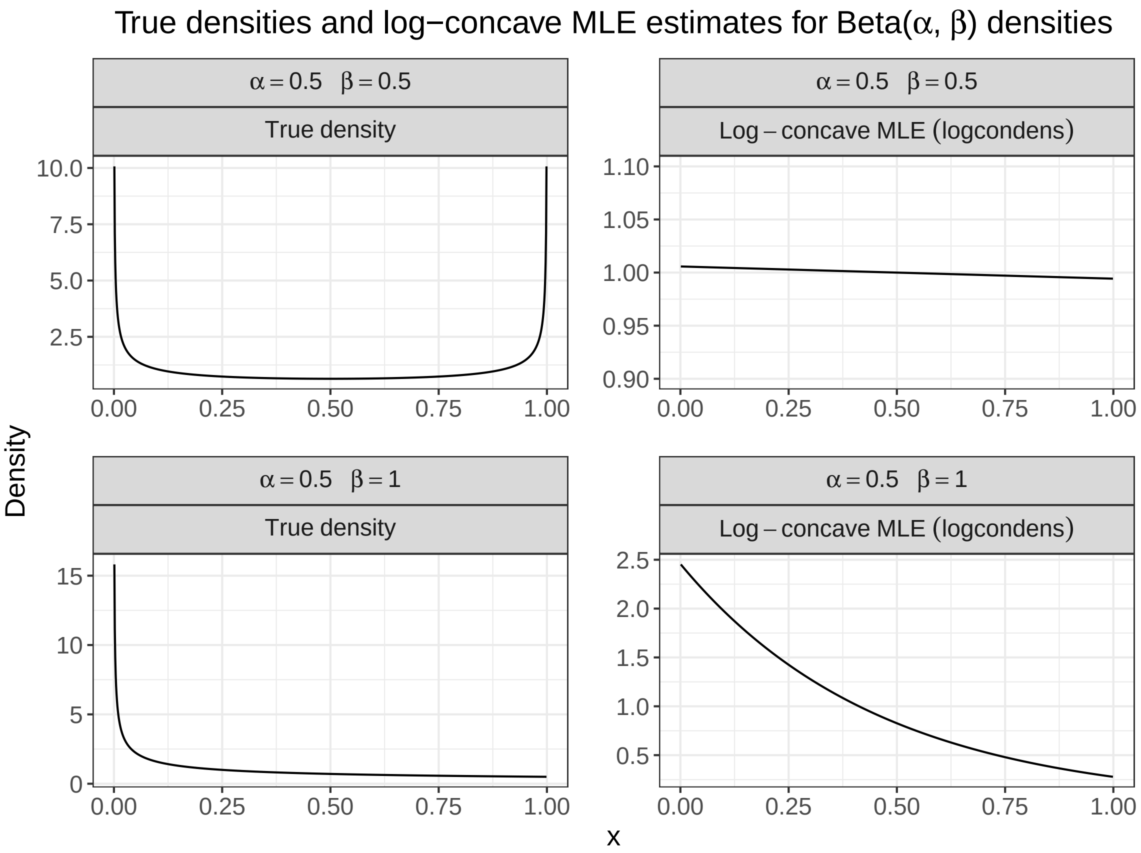


Figure 3.15: Two non-log-concave Beta densities and their corresponding log-concave MLEs, as estimated over $n = 100,000$ observations.

In the first setting ($\alpha = 0.5, \beta = 0.5$), it appears that the log-concave MLE is the Unif(0, 1) density. We consider the second setting ($\alpha = 0.5, \beta = 1$) in more depth. The density in row 2, column 2 looks similar to an exponential density, but x can only take on values between 0 and 1. The truncated exponential density is given by

$$f(x; \lambda, b) = \frac{\lambda \exp(-\lambda x)}{1 - \exp(-\lambda b)}, \quad 0 < x \leq b.$$

In this setting, we can try to fit a truncated exponential density with $b = 1$. In Figure 3.16, we see that a truncated exponential density with $\lambda = 2.18$ and $b = 1$ provides a good fit for the log-concave MLE.

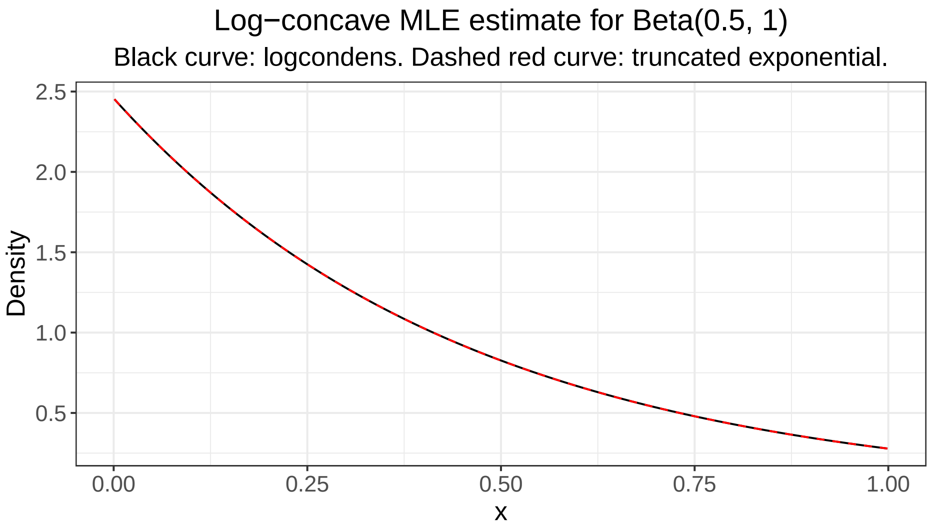


Figure 3.16: Log-concave MLE for Beta(0.5, 1) density (black solid) and truncated exponential density with $\lambda = 2.18$ and $b = 1$ (red dashed).

We can also see that the truncated exponential density is log-concave:

$$\begin{aligned}\frac{\partial^2}{\partial x^2} \log f(x; \lambda, b) &= \frac{\partial^2}{\partial x^2} [\log(\lambda) - \lambda x - \log(1 - \exp(-\lambda b))] \\ &= \frac{\partial}{\partial x}[-\lambda] \\ &= 0.\end{aligned}$$

3.4.3 Method Comparison

Figure 3.17 shows examples of both log-concave and not log-concave beta densities. We use similar α and β parameters in the simulations where we test for log-concavity. This shows that our simulations are capturing a variety of beta density shapes.

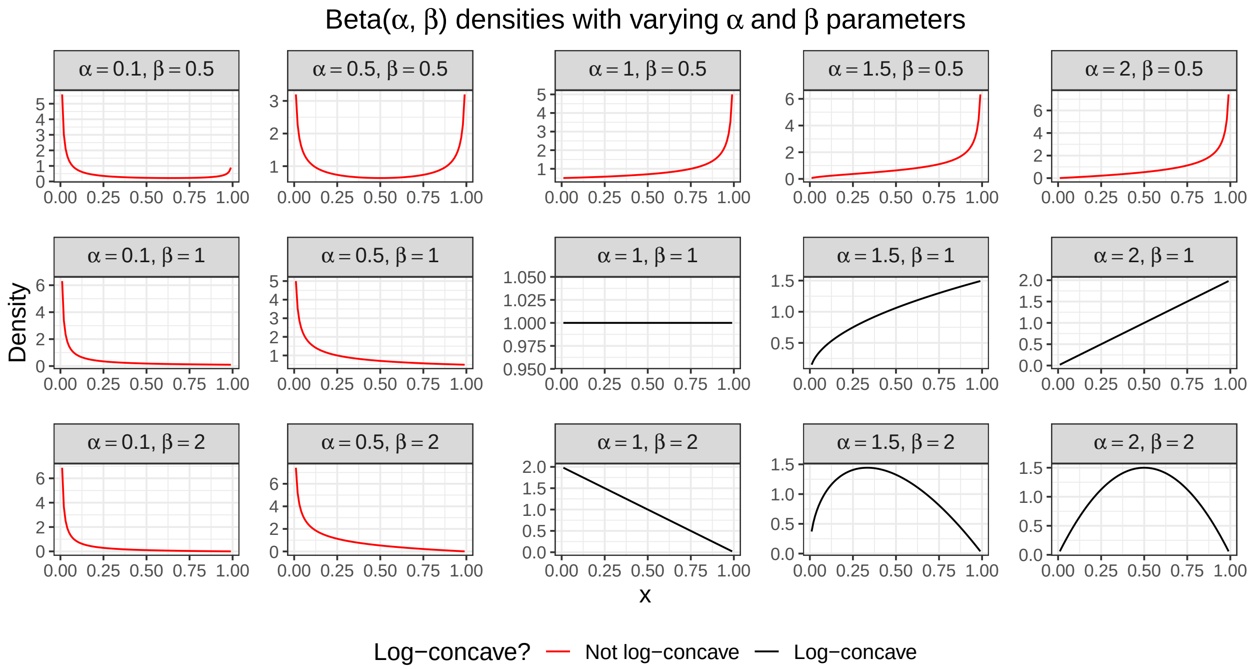


Figure 3.17: Beta densities across a variety of α and β parameters.

We now implement the full oracle LRT (universal), partial oracle LRT (universal), fully nonparametric LRT (universal), and permutation test. The full oracle LRT uses the true density in the numerator. The partial oracle LRT uses the knowledge that the true density comes from the Beta family. We use the `fitdist` function in the `fitdistrplus` library to find the MLE for α and β on $\mathcal{D}_1$ computationally (Delignette-Muller and Dutang, 2015). Then the numerator of the partial oracle LRT uses this Beta MLE density. The fully nonparametric approach fits a kernel density estimate on $\mathcal{D}_1$. In particular, we use the `kde1d` function from the `kde1d` library, and we restrict the support of the KDE to $[0, 1]$ (Nagler and Vatter, 2020). This is particularly important in the Beta family case, since some of the non-log-concave Beta densities assign high probability to observations near 0 or 1. (See Figure 3.17.) The numerator of the fully nonparametric approach uses the KDE.

Figure 3.18 compares the four tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave. We set $n = 100$, and we perform 200 simulations to determine each rejection proportion. The universal methods subsample at $B = 100$, and the permutation test uses $B = 99$ shuffles. In the first panel, $\beta = 0.5$, so the density is not log-concave for any choice of α. In the second and third panels, $\beta = 1$ and $\beta = 2$. In these cases, the density is log-concave only when $\alpha \geq 1$ as well.

We observe that the permutation test is valid in all settings, but the three universal tests often have higher power. As expected, out of the universal tests, the full oracle approach has the highest power, followed by the partial oracle approach and then the fully nonparametric approach. When $\beta = 0.5$, all of the universal LRTs have power greater than or equal to the permutation test. When $\beta \in \{1, 2\}$, the universal approaches have higher power for some values of α. Again, we see that even when the permutation test is valid, it is possible for universal LRTs to have higher power.

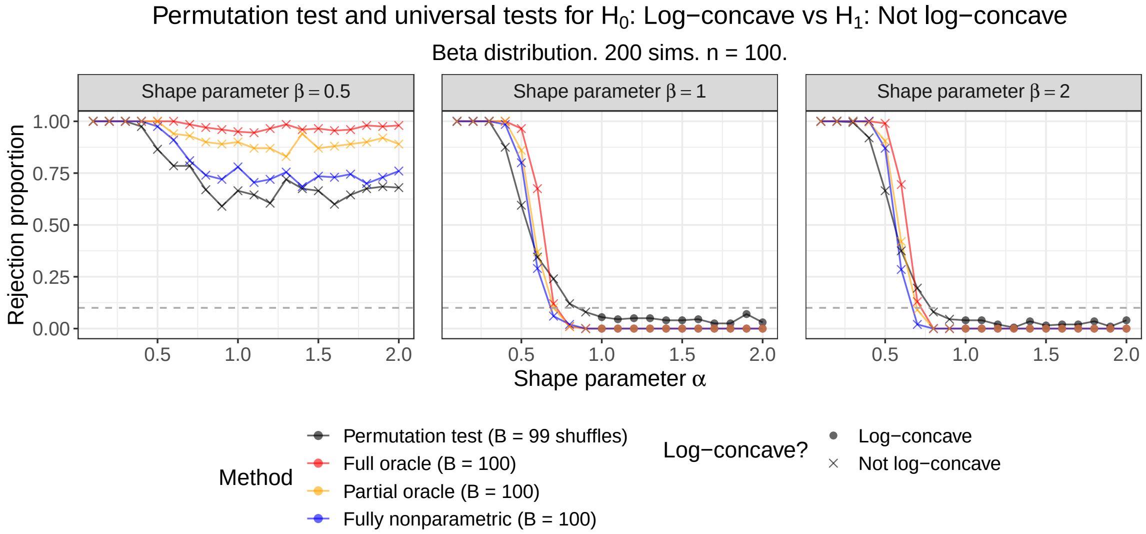


Figure 3.18: Rejection proportions for four tests of $H_0 : f^*$ is log-concave versus $H_1 : f^*$ is not log-concave, on Beta(α, β) density.

3.5 Theoretical Power of Log-concave Universal Tests

Through simulations, we have shown that the universal LRTs can have high power in tests of $H_0 : f^* \in \mathcal{F}_d$ versus $H_1 : f^* \notin \mathcal{F}_d$. We now prove a theoretical result that provides conditions under which the power of the universal tests converges to 1. First, we review or introduce some notation. Let $\widehat{f}_{1,n}$ be an estimate of f^*, fit on $\mathcal{D}_{1,n}$. Let $\widehat{f}_{0,n} = \underset{f \in \mathcal{F}_d}{\arg\max} \sum_{Y_i \in \mathcal{D}_{0,n}} \log(f(Y_i))$. In words, $\widehat{f}_{0,n}$ is the log-concave MLE fit on $\mathcal{D}_{0,n}$. The universal test statistic is

$$T_n = \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{\widehat{f}_{1,n}(Y_i)}{\widehat{f}_{0,n}(Y_i)}.$$

We reject H_0 if $T_n \geq 1/\alpha$. Let

$$f^{\mathrm{LC}} = \operatorname*{arg\,min}_{f \in \mathcal{F}_d} D_{KL}(f^* \| f) = \operatorname*{arg\,min}_{f \in \mathcal{F}_d} \int_{\mathbb{R}^d} f^*(x) \log\left(\frac{f^*(x)}{f(x)}\right) dx.$$

If $f^* \in \mathcal{F}_d$, then $f^{\mathrm{LC}} = f^*$. We write the empirical KL divergence between f^* and $\widehat{f}_{1,n}$, evaluated on $\mathcal{D}_{0,n}$ as

$$\widehat{D}_{KL}(f^* \| \widehat{f}_{1,n}) = \frac{1}{|\mathcal{D}_{0,n}|} \sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{f^*(Y_i)}{\widehat{f}_{1,n}(Y_i)}\right).$$

Then we can also see that $\widehat{f}_{0,n} = \operatorname*{arg\,min}_{f \in \mathcal{F}_d} \widehat{D}_{KL}(f^* \| f)$.

When $f^* \notin \mathcal{F}_d$, Theorem 3.5.1 establishes conditions under which the power of the universal test converges to 1 as $n \to \infty$.

Theorem 3.5.1. *We make five assumptions:*

1. *Suppose each $\widehat{f}_{1,n} \in \mathcal{C}$, where $\mathcal{C}$ is some (potentially nonparametric) class of functions that satisfies $\sup_{f \in \mathcal{C}} |\widehat{D}_{KL}(f^* \| f) - D_{KL}(f^* \| f)| = O_P(n^{-\beta_1})$ for some $\beta_1 > 0$.*

2. *$D_{KL}(f^* \| \widehat{f}_{1,n}) = O_P(n^{-\beta_2})$ for some $0 < \beta_2 \leq 1/2$.*

3. *$\int_{\mathbb{R}^d} \|x\| f^*(x) dx < \infty$.*

Suppose there is some set A with $P(A) = 1$ (e.g., the support of f^{LC} or its interior) that satisfies the following:

4. *For some $\ell > 0$, $\inf_{x \in A} f^{LC}(x) \geq \ell$.*

5. *$\sup_{x \in A} |\widehat{f}_{0,n}(x) - f^{LC}(x)| \xrightarrow{a.s.} 0$.*

Then $\lim_{n \to \infty} \mathbb{P}_{H_0}(T_n \geq 1/\alpha) = 1$.

We discuss several conjectures that might allow us to refine these assumptions. Assumption 4 may follow immediately from Lemma 3(b) of Cule et al. (2010a) if A is a compact and convex set. Suppose $X_1, X_2, \ldots$ is an iid sequence with density f^*, the support E is the smallest closed set with $\int_E f^* = 1$, and $\mathrm{conv}(S)$ is the convex hull of some compact subset S of the interior of E. Under our assumption 3, Lemma 3(b) states that there is a constant $c > 0$ such that, with probability one,

$$\liminf_{n \to \infty} \inf_{x \in \mathrm{conv}(S)} \widehat{f}_n(x) \geq c.$$

Assumption 5 may hold if f^{LC} is continuous over A. Under our assumption 3, as well as $\int_{\mathbb{R}^d} f^* \log_+(f^*) < \infty$, $\mathrm{int}(E) \neq \emptyset$, and continuity of f^{LC}, Theorem 3 of Cule et al. (2010b) states that $\sup_{x\in\mathbb{R}^d} \exp(a\|x\|)|\widehat{f}_n(x) - f^{\mathrm{LC}}(x)| \to 0$ almost surely as $n \to \infty$. Under assumption 4, it is impossible for f^{LC} to be continuous over $\mathbb{R}^d$. However, outside of A, the tails of $\widehat{f}_{0,n}$ and f^{LC} are both 0.

We provide a proof sketch to show how these five assumptions lead to the result. The full proof appears in Appendix C.

Proof sketch. We begin by separating T_n into a product of three components:

$$T_n = \underbrace{\left\{ \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{\widehat{f}_{1,n}(Y_i)}{f^*(Y_i)} \right\}}_{C_{1,n}} \underbrace{\left\{ \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^*(Y_i)}{f^{\mathrm{LC}}(Y_i)} \right\}}_{C_{2,n}} \underbrace{\left\{ \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^{\mathrm{LC}}(Y_i)}{\widehat{f}_{0,n}(Y_i)} \right\}}_{C_{3,n}}.$$

Define ϵ as

$$\epsilon = \|(f^{\mathrm{LC}})^{1/2} - (f^*)^{1/2}\|_2 = \left[\int \left((f^{\mathrm{LC}})^{1/2} - (f^*)^{1/2} \right)^2 d\mu \right]^{1/2}.$$

This choice of ϵ arises from Lemma 1 of Wong et al. (1995). We see that

$$\begin{aligned}
&\mathbb{P}(T_n < 1/\alpha) \\
&\leq \mathbb{P}\left(C_{2,n} < \exp\left(\frac{n}{8}\epsilon^2\right) \cup C_{1,n} < \frac{1}{\alpha}\exp\left(-\frac{n}{16}\epsilon^2\right) \cup C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right) \right) \\
&\leq \mathbb{P}\left(C_{2,n} < \exp\left(\frac{n}{8}\epsilon^2\right) \right) + \mathbb{P}\left(C_{1,n} < \frac{1}{\alpha}\exp\left(-\frac{n}{16}\epsilon^2\right) \right) + \mathbb{P}\left(C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right) \right).
\end{aligned}$$

We want to show that these three probabilities converge to 0.

Using Lemma 1 of Wong et al. (1995), we show that

$$\mathbb{P}\left(C_{2,n} < \exp\left(\frac{n}{8}\epsilon^2\right) \right) \leq \exp\left(-\frac{n}{8}\epsilon^2\right).$$

So $\lim_{n\to\infty} \mathbb{P}\left(C_{2,n} < \exp\left((n/8)\epsilon^2\right) \right) = 0$.

For the second probability, we use assumptions 1 and 2 to show that $\log(C_{1,n}) = O_P(n^{1-\beta})$, where $\beta = \min\{\beta_1, \beta_2\} \in (0, 1/2]$. This implies that $\lim_{n\to\infty} \mathbb{P}(C_{1,n} < (1/\alpha)\exp(-(n/16)\epsilon^2)) = 0$.

For the third probability, we use Markov's inequality to derive

$$\mathbb{P}\left(C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right)\right) \leq \frac{8}{\epsilon^2}\mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right)\right].$$

Let $\delta > 0$. Using ℓ from assumption 4, fix $\gamma > 0$ such that $\gamma < \ell(\exp(\delta) - 1)$. Note that this implies $\log((\gamma + \ell)/\ell) < \delta$. We derive

$$\mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right)\right] < \mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log(\gamma + f^{\mathrm{LC}}(Y_i))\right] - \mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log(f^{\mathrm{LC}}(Y_i))\right] +$$
$$\mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log(\widehat{f}_{0,n}(Y_i)) I\left(\sup_{x \in A}|\widehat{f}_{0,n}(x) - f^{\mathrm{LC}}(x)| \geq \gamma\right)\right].$$

We consider the difference of the first two expectations. The function $g(x) = \log(\gamma + x) - \log(x)$ is a decreasing function, so $g(x)$ is maximized at the smallest x. By assumption 4, we know that with probability 1, $f^{\mathrm{LC}}(Y_i) \geq \ell$. This tells us that

$$\mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log(\gamma + f^{\mathrm{LC}}(Y_i))\right] - \mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log(f^{\mathrm{LC}}(Y_i))\right] \leq \log(\gamma + \ell) - \log(\ell) < \delta.$$

By Lemma 3(a) of Cule et al. (2010a), assumption 3 implies that for some $u > 0$, $\limsup_{n\to\infty} \sup_{x \in \mathbb{R}^d} \widehat{f}_{0,n}(x) \leq u$ with probability 1. Using reverse Fatou's Lemma for conditional expectations in step (C.3), we determine

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log(\widehat{f}_{0,n}(Y_i)) I\left(\sup_{x \in A}|\widehat{f}_{0,n}(x) - f^{\mathrm{LC}}(x)| \geq \gamma\right)\right]$$
$$\leq u\mathbb{P}\left(\sup_{x \in A}|\widehat{f}_{0,n}(x) - f^{\mathrm{LC}}(x)| \geq \gamma\right) \qquad \text{with probability 1.}$$

Finally, by assumption 5, $\lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x) - f^{\mathrm{LC}}(x)| \geq \gamma\right) \to 0$. So with probability 1, for arbitrary $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right)\right] < \delta.$$

We conclude that $\lim_{n\to\infty} \mathbb{P}\left(C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right)\right) = 0$. Therefore, $\lim_{n\to\infty} \mathbb{P}(T_n < 1/\alpha) = 0$. □

3.6 Conclusion

We have implemented and evaluated several universal LRTs to test for log-concavity. These methods include a full oracle (true density) approach, a partial oracle (parametric) approach, a fully nonparametric approach, and several LRTs that reduce the d-dimensional test to a set of one-dimensional tests. We compare these tests to a permutation test that is not guaranteed to be valid. In one dimension, the universal tests can have higher power than the permutation test. In higher dimensions, we have seen that the permutation test can falsely reject H_0 at a rate much higher than α. In contrast, the universal tests are still valid in higher dimensions. As seen in the Gaussian mixture case, the dimension reduction universal approaches can have notably stronger performance than the universal tests that work with d-dimensional densities.

Part II

Conformal Prediction

Chapter 4

Overview of Conformal Prediction

4.1 Introduction

Predictive modeling helps statisticians to forecast new outcomes given previous outcomes and, in the supervised case, relevant predictor variables. Given a sample of data, statisticians may want to predict a new outcome such as the temperature for a given day, the traffic on a stretch of highway, or the level of antibodies following the administration of a vaccine. While a single point prediction gives the most likely outcome according to some model, the single prediction does not capture our level of confidence in the outcome. An alternative approach is to associate a probability to some set of predicted outcomes. A valid $100(1-\alpha)\%$ prediction set contains the new observation (unsupervised) or the outcome associated with a new observation (supervised) with probability of at least $1-\alpha$. The goal of *conformal predictive inference* (often called *conformal prediction* or *conformal inference*) is to construct valid prediction sets under the assumption that the data are iid, or at least exchangeable. Notably, conformal methods do not require knowledge about the form of the underlying density or model, and the confidence sets are valid in finite samples.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n iid pairs of observations from a distribution P. In set-valued, supervised prediction, we want to find a set-valued function C such that

$$P(Y \in C(X; \alpha)) \geq 1-\alpha \tag{4.1}$$

where $1-\alpha$ is the user-specified confidence level and (X, Y) denote a new pair drawn from P. (We should really write $P^{n+1}(Y \in C(X; \alpha)) \geq 1-\alpha$ since the randomness is over Y and the training data. We have suppressed the superscript $n+1$ for notational simplicity.) Vovk et al. (2005) created the method of conformal prediction to construct C such that (4.1) holds for all distributions P. In other words, conformal methods yield distribution-free prediction sets. Lei et al. (2013) connect conformal prediction to minimax density estimation, and Lei and Wasserman (2014) connect conformal prediction to minimax nonparametric regression.

Key early references on conformal prediction include Vovk et al. (2005) and Shafer and Vovk (2008). The literature on conformal prediction is quickly growing in several overlapping directions. Developments on conformal predictions include connections to traditional statistical methods, extensions to flexible settings, and implementations that are computationally efficient. Work in these directions includes interpolations between marginal and conditional coverage (Lei and Wasserman, 2014), extensions to multiclass set-valued classification (Sadinle et al., 2018), valid discretizations of conformal methods (Chen et al., 2018), anti-conservative bounds on coverage, methods for variable importance, and computationally efficient sample-splitting methods (Lei et al., 2018). Many open problems remain in extending conformal methods to new contexts.

Conformal prediction — introduced by Vovk et al. (2005) — is a general method for obtaining distribution-free prediction sets with confidence guarantees. Here, we review some background on conformal prediction.

4.2 The Unsupervised Case

Let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be iid observations from a distribution P where $Y_i \in \mathcal{Y}$, and let Y_{n+1} denote a new draw from P. The goal is to construct a set $C(\alpha)$ based on the training data $Y_1, \ldots, Y_n$ such that $P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha$ for every distribution P. When $d = 1$, Theorem 4.2.1 provides one valid construction based on order statistics. For a proof, see Appendix D.

Theorem 4.2.1. *We define a prediction interval $C(\alpha) = [Y_{(r)}, Y_{(s)}]$, where $r = \lfloor (n+1)(\alpha/2) \rfloor$ and $s = \lceil (n+1)(1-\alpha/2) \rceil$. Then for every distribution P, $P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha$. This interval is bounded if $n \geq 2/\alpha - 1$.*

The validity of Theorem 4.2.1 relies on the exchangeability of the original sample's order statistics. Alternative valid constructions rely on the exchangeability of conformal residuals constructed from the sample. For any $u \in \mathcal{Y}$, let $\mathcal{A}(u) = (Y_1, \ldots, Y_n, u)$, which can be thought of as the training data augmented with a guess that $Y_{n+1} = u$. Define the residual (or conformity score) $R_i(u) = \phi(Y_i, \mathcal{A}(u))$ where $\phi : \mathbb{R}^d \to \mathbb{R}$ is any function that is invariant under permutations of the elements of $\mathcal{A}(u)$. We wish to test the hypothesis $H_0 : Y_{n+1} = u$. The set of all u for which we do not reject H_0 at level $1 - \alpha$ will provide the $100(1-\alpha)\%$ prediction set. Assuming $Y_{n+1} = u$, we define

$$\pi(u) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i(u) \geq R_{n+1}(u)) \tag{4.2}$$

which is the p-value for testing this hypothesis. Intuitively, the p-value for a given u is small if the residuals at most of $Y_1, \ldots, Y_n$ are smaller than the residual at u (i.e., the p-value is small if $Y_{n+1} = u$ does not "conform" to the original sample). $\pi(u)$ is a valid p-value because under H_0, $\pi(u)$ follows a super-uniform distribution over $t \in [0,1]$. That is, $P(\pi(u) \leq t) = P(\pi(u) \leq \lfloor t(n+1) \rfloor/(n+1)) \leq \lfloor t(n+1) \rfloor/(n+1) \leq t$. Often P is a continuous distribution and $P(\phi(Y_i, \mathcal{A}(u)) = \phi(Y_j, \mathcal{A}(u))) = 0$ for $i \neq j$. In this case, $\pi(u)$ is uniformly distributed over the set $\{1/(n+1), 2/(n+1), \ldots, 1\}$.

We invert the test by defining

$$C(\alpha) = \{u : \ \pi(u) \geq \alpha\}. \tag{4.3}$$

Theorem 4.1. *For $C(\alpha)$ as given above, $P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha$ for every distribution P.*

For a proof, see Vovk et al. (2005). There is great flexibility in the choice of conformity score ϕ. Every choice leads to a valid prediction set, but different choices may lead to smaller or larger sets. Thus, the choice of ϕ can affect the efficiency of the prediction set but not its coverage; see Lei et al. (2013).

As an example, let $R_i(u) = |Y_i - \overline{Y}(u)|$ where $\overline{Y}(u) = (u + \sum_{i=1}^n Y_i)/(n+1)$ is the mean of the augmented data. Then $\pi(u) = (n+1)^{-1} \sum_{i=1}^{n+1} I\big(|Y_i - \overline{Y}(u)| \geq |u - \overline{Y}(u)|\big)$. Another useful conformity score is $R_i(u) = 1/\widehat{p}_u(Y_i)$ where $\widehat{p}_u$ is a density estimator based on the augmented data. Lei et al. (2013) showed that this choice is minimax optimal when some conditions hold.

4.3 The Supervised Case

In this case the data are $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$. Let $(X, Y) \sim P$ be a new observation. We then want a set $C(x; \alpha)$ such that $P(Y \in C(X; \alpha)) \geq 1 - \alpha$ for all P. In the supervised setting, the conformal set now depends on the X_i s as well. As one possibility, fix (x, y) and let $\widehat{m}_{(x,y)}$ be a regression estimator based on the augmented data $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ with $(X_{n+1}, Y_{n+1}) = (x, y)$. Let

$$C(x; \alpha) = \Big\{ y : \; \pi(x, y) \geq \alpha \Big\}$$

where

$$\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i(x, y) \geq R_{n+1}(x, y))$$

and $R_i(x, y) = |Y_i - \widehat{m}_{(x,y)}(X_i)|$. Then $\inf_P P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$.

A second useful choice of conformal residual is $R_i(x, y) = 1/\widehat{p}(X_i, Y_i)$ where $\widehat{p}$ is a joint density estimate based on the augmented data. See Lei et al. (2013) for more details.

4.4 Distribution-free Prediction from Working Models

It is possible to get distribution-free prediction from a relevant parametric model, even when the model is wrong. We will describe this idea in the unsupervised setting, but the same principle holds for the supervised case.

Let $\mathcal{Q} = (Q_\theta : \theta \in \Theta)$ denote a parametric model. Let $Y_1, \ldots, Y_n, Y \sim P$ where we do not assume that the true distribution P is in $\mathcal{Q}$. We can use the model to construct a conformal residual. For example, we could use $R_i(u) = 1/\widehat{q}_{\widehat{\theta}(u)}(Y_i)$ where $\widehat{\theta}(u)$ is the maximum likelihood estimate based on $(Y_1, \ldots, Y_n, u)$. Now let $C(\alpha)$ denote the conformal set constructed from the R_i s. It follows from Theorem 4.1 that $P(Y \in C(\alpha)) \geq 1 - \alpha$ for every P. This is true even though P may not be in the model. However, the size of $C(\alpha)$ is smaller if P is in, or close to, $\mathcal{Q}$.

4.5 Predicting a Batch of Observations

The coverage condition in (4.1) refers to the problem of predicting a single observation. Suppose we want to predict a new batch of observations $\mathcal{D} = \{Y_1', \ldots, Y_m'\}$. What coverage guarantee should we require? There are several possibilities:

1. Batch Coverage: find C such that $P(\mathcal{D} \in C) \geq 1 - \alpha$.

2. Marginal Coverage: find C so that $P(Y_i' \in C) \geq 1 - \alpha$ for $i = 1, \ldots, m$.

3. Probabilistic Coverage: find C so that for $\epsilon > 0$ and $\gamma > 0$,

$$P\left(\frac{1}{m}\sum_{i=1}^{m} I(Y_i' \in C) \geq 1 - \alpha - \epsilon\right) \geq 1 - \gamma.$$

The first requires a simultaneous prediction set for all of the future observations. This seems too strong since, when m is large, this will require a huge, high-dimensional prediction set. The second requires correct marginal coverage on each new observation. The third requires that the proportion of observations trapped in the prediction set is at least $1 - \alpha - \epsilon$ with high probability. In fact, by Hoeffding's inequality, marginal coverage implies probabilistic coverage if $m \geq \log(1/\gamma)/(2\epsilon^2)$. Hence, we focus on marginal coverage.

4.6 Simulation Example

We consider an example of conformal prediction in an unsupervised setting. Suppose we observe a sample of $n = 10,000$ iid observations $\{Y_1, \ldots, Y_n\}$ from a two-dimensional mixture of five Gaussians, given by

$$\begin{aligned} &0.2N((-0.5,-0.5),I_2) + 0.2N((-0.5,0.5),I_2) + 0.2N((0.5,-0.5),I_2) + \\ &\quad 0.2N((0.5,0.5),I_2) + 0.2N((0,0),I_2). \end{aligned}$$

When we obtain this data sample, we do not know the true underlying density. Figure 4.1 plots the sample as a two-dimensional hexagonal heatmap.

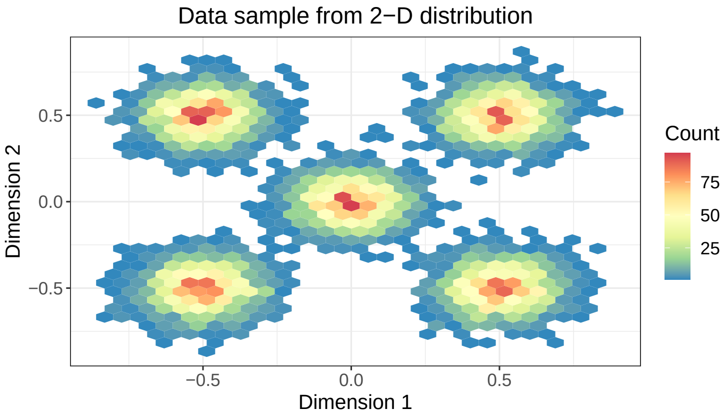


Figure 4.1: Sample of 10,000 iid observations from a mixture of five two-dimensional Normal distributions.

We use conformal prediction methods to construct a valid prediction set for new observations from this distribution. For any value $u \in \mathbb{R}^2$, suppose we wish to test $H_0 : Y_{n+1} = u$ to determine whether u lies within a $100(1-\alpha)\%$ conformal prediction set. Using one suggestion noted in Section 4.2, we fit a kernel density estimator $\widehat{p}_u$ on the augmented sample $\{Y_1, \ldots, Y_n, u\}$. We compute residuals $R_i(u) = 1/\widehat{p}_u(Y_i)$, $i = 1, \ldots, n+1$. Then we compute the p-value $\pi(u)$ from (4.2). As stated in (4.3), we include u in the conformal set $C(\alpha)$ if $\pi(u) \geq \alpha$.

Under the marginal coverage property, for a new sample $\{Y_1', \ldots, Y_m'\}$ drawn from the same distribution, each observation should satisfy $P(Y_i' \in C) \geq 1 - \alpha$, $i = 1, \ldots, m$. We test this by drawing $m = 1000$ new observations from this five component Gaussian mixture. We check whether each of these observations are in $C(\alpha)$ for $\alpha = 0.1$. Figure 4.2 plots the new sample, with each observation colored by whether it is contained within the 90% conformal prediction set. In alignment with the theoretical coverage, the conformal set contains 903 out of 1000 new observations.

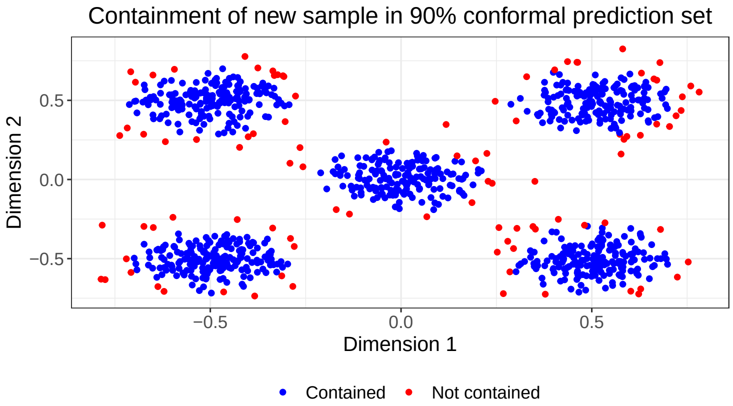


Figure 4.2: New sample of 1000 iid observations from the same distribution as Figure 4.1. 90.3% of these observations are contained within the 90% conformal prediction set.

Chapter 5

Distribution-Free Prediction Sets with Random Effects

5.1 Introduction

A fundamental assumption of the usual conformal method is that the data are iid (or, at least, exchangeable). In this paper, we extend conformal methods to the following random effects models where the iid assumption fails. Let $P_1, P_2, \ldots, P_k \sim \Pi$ be random distributions drawn from Π and let

$$\mathcal{D}_j = \{(X_{j1}, Y_{j1}), \ldots, (X_{jn_j}, Y_{jn_j})\}$$

be n_j iid observations drawn from P_j for $j = 1, \ldots, k$. It is helpful to imagine that we have k subjects and $\mathcal{D}_j$ represents n_j observations on subject j.

There are two tasks to consider:

1. Task 1: Predicting a new subject. Let $P_{k+1} \sim \Pi$ denote a new draw from Π (a new subject) and let $(X, Y) \sim P_{k+1}$. The goal is to construct a prediction set for Y using X and the training data $\mathcal{D}_1, \ldots, \mathcal{D}_k$.

2. Task 2: Predicting a new observation on one of the current subjects. Let (X, Y) denote a new draw from one of the distributions P_j. We want a prediction for Y based on X and the training data.

In Task 1, we have to deal with the complication that we have never observed any data from the future distribution P_{k+1}. Task 2 is easier since we already have data on subject j. This is where familiar tools such as shrinkage and borrowing strength can be used.

A familiar example of a parametric random effects working model is $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$. Recall that P_j denotes the true underlying distribution of (X, Y) for group j. Suppose this random effects model represents the true relationship between X and Y, and suppose $X \sim N(0, 1)$. Then drawing $(X_j, Y_j) \sim P_j$ amounts to drawing $X_j \sim N(0, 1)$ and $Y_j \sim N(\beta_{0j} + \beta_{1j}X_j, \sigma^2)$. Furthermore, suppose that P_β represents the distribution of (β_{0j}, β_{1j}) over the full population. Then drawing a distribution $P_j \sim \Pi$ reduces to drawing $(\beta_{0j}, \beta_{1j}) \sim P_\beta$. We will show that it is possible to use a parametric working model to get prediction sets with valid coverage even if the model is wrong.

5.1.1 Related Work

The literature on random effects models is vast. Most of the work focuses on estimation. Laird and Ware (1982) provide foundational work on the structure and estimation of random effects models for repeated-measures data. The authors note that random effects allow researchers to model both within- and between-subject variation, often using parameters that have natural interpretations. For instance, random effects models frequently are defined by within-subject and across-subject means and variances (DerSimonian and Laird, 1986). We incorporate this conceptualization in our simulations. Random effects models have been used for prediction by some researchers in parametric settings (Calvin and Sedransk, 1991; Booth and Hobert, 1998; Schofield et al., 2015). Claggett et al. (2014) develop methods for inference on the quantiles of study-level parameters without distributional assumptions on these parameters. Thus, random effects researchers have developed some approaches for inference and prediction in parametric settings and for inference without distributional assumptions. To the best of our knowledge, there are no papers on valid distribution-free prediction for random effects models.

5.1.2 Paper Outline

Section 5.2 presents our new methods for unsupervised and supervised conformal prediction sets in the random effects setting. Section 5.3 contains simulation studies for Tasks 1 and 2. Section 5.4

implements our supervised prediction methods on data from a sleep deprivation study. Section 5.5 provides concluding remarks. Appendix E contains proofs.

5.2 Prediction for Random Effects Models

We have shown two unsupervised prediction set methods that are valid under minimal assumptions. In the one-dimensional case, the order statistic approach is valid under the assumption that the data are exchangeable. This method is a simpler construction that does not require data augmentation or the choice of a conformity score. Alternatively, the residual approach relies on the exchangeability of the residuals; this condition holds for any permutation-invariant ϕ when the underlying data are exchangeable. The residual method affords more flexibility through the construction of a conformal score, and it extends beyond the one-dimensional setting. To construct prediction sets for a new observation on a new subject (task 1), we use methods based on the original sample's order statistics in the unsupervised setting, and we use the residual method in the supervised setting. To construct prediction sets for a new observation on an existing subject (task 2), we use the residual method. In the task 2 setting, the residual method allows us to implement a conformity score based on a shrinkage estimator.

5.2.1 Unsupervised Prediction for a New Distribution

We start with the unsupervised version. Recall that the data come in groups $\mathcal{D}_1, \ldots, \mathcal{D}_k$ and each group has iid data

$$\mathcal{D}_j = \{Y_{j1}, \ldots, Y_{jn_j}\} \sim P_j$$

where $P_1, \ldots, P_k \sim \Pi$. Assuming a new distribution $P_{k+1} \sim \Pi$ and $Y \sim P_{k+1}$, we want a prediction region for Y.

Method 0: Double Conformal

We note that there are two levels of randomness in this set-up. At the level of group j, we have independent observations from a distribution P_j. At the distribution level, each distribution is sampled from a larger superpopulation Π. If we draw a single new observation Y_i from a new distribution, we say that Y_i is a draw from $\overline{\Pi}$, where $\overline{\Pi} = \int P\Pi(dP)$. A "double conformal" method is one natural approach that incorporates this hierarchical structure. This method first uses conformal prediction within each group and then uses those results to construct a conformal prediction set across groups.

At the group level, let $C_j = [\ell_j, u_j]$ be the $100(1-\alpha/2)\%$ prediction set obtained by applying the method in Theorem 4.2.1 at level $\alpha/2$ to group j, $j = 1, \ldots, k$. Thus, we have constructed a sample of k lower bounds $\{\ell_1, \ldots, \ell_k\}$ and k upper bounds $\{u_1, \ldots, u_k\}$. Using the order statistics from those samples, we set $C^{\text{dbl}}(\alpha) = [\ell_{(r)}, u_{(s)}]$, where $r = \lfloor (k+1)(\alpha/4) \rfloor$ and $s = \lceil (k+1)(1-\alpha/4) \rceil$.

Theorem 5.2.1. *For $C^{dbl}(\alpha)$ as defined above, $\overline{\Pi}(Y \in C^{dbl}(\alpha)) \geq 1-\alpha$. This set is bounded if $k \geq 4/\alpha - 1$ and $n \geq 4/\alpha - 1$.*

By Theorem 5.2.1, $C^{\text{dbl}}(\alpha)$ is a valid $100(1-\alpha)\%$ prediction set for a new observation Y from a new group. For a proof, see the Appendix. While this method is valid, our results show that this method overcovers. Thus, we turn to several methods that are better choices.

Method 1: Pooling CDFs

As one approach, we construct an empirical CDF within each group. We average these CDFs across groups, and we determine the prediction set bounds based on the quantiles of the average of CDFs. This method is asymptotically valid as $k \to \infty$.

Formally, for any group j, the empirical CDF is defined as

$$\widehat{F}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(Y_{ji} \leq t).$$

We set

$$\widehat{q}_k(\alpha) = \inf \left\{ t \in \mathbb{R} : \frac{1}{k} \sum_{j=1}^{k} \widehat{F}_j(t) \geq \alpha \right\}.$$

Then $C^{\text{poolCDF}}(\alpha) = [\widehat{q}_k(\alpha/2), \widehat{q}_k(1-\alpha/2)]$. For a proof of Theorem 5.2.2, see the Appendix.

Theorem 5.2.2. *Assume that $F : \mathbb{R} \to [0,1]$, defined as $F(t) = \overline{\Pi}(Y \leq t)$, is strictly increasing. For $C^{poolCDF}(\alpha)$ as defined above, $\overline{\Pi}(Y \in C^{poolCDF}(\alpha)) \to 1-\alpha$ as $k \to \infty$.*

Method 2: Subsampling Once

Draw one observation randomly from each group. For instance, Y_1 can equal any element in $\mathcal{D}_1 = \{Y_{11}, Y_{12}, \ldots, Y_{1n_1}\}$ with probability $1/n_1$. Then the data consist of k iid observations $Y_1, Y_2, \ldots, Y_k$ from $\overline{\Pi}(\cdot) = \int P(\cdot)d\Pi(P)$. We define a prediction set

$$C^{\text{sub}}(\alpha) = [Y_{(r)}, Y_{(s)}],$$

where $r = \lfloor (k+1)(\alpha/2) \rfloor$ and $s = \lceil (k+1)(1-\alpha/2) \rceil$. This method has exact $100(1-\alpha)\%$ coverage since the k observations in the subsample are k iid observations from $\overline{\Pi}$. The validity of this method follows from Theorem 4.2.1.

Method 3: Repeated Subsampling

We now modify Method 2 to incorporate B subsamples of a single observation from each of the k groups. (Again, each subsample now contains k iid observations.) Gupta et al. (2020) developed the method of constructing conformal prediction sets through repeated subsampling in the case of exchangeable data. Suppose $Y^b_{(1)}, Y^b_{(2)}, \ldots, Y^b_{(k)}$ are the ordered observations from the b^{th} subsample. Conformal prediction is implicitly testing $H_0 : Y_{k+1} = u$ versus $H_1 : Y_{k+1} \neq u$, and the level $1-\alpha$ conformal prediction set is the set of values at which we would not reject H_0 under the given construction. Thus, within the b^{th} subsample, the p-value at $u \in \mathbb{R}$ is

$$\pi_b(u) = \inf\{\alpha : u \notin [Y^b_{(r)}, Y^b_{(s)}]\},$$

where $r = \lfloor (k+1)(\alpha/2) \rfloor$ and $s = \lceil (k+1)(1-\alpha/2) \rceil$.

We define a prediction set

$$C^{\text{rep}}(\alpha) = \left\{ u : \frac{1}{B} \sum_{b=1}^{B} \pi_b(u) \geq \alpha \right\}.$$

Theorem 5.1. *For $C^{rep}(\alpha)$ as defined above, $\overline{\Pi}(Y \in C^{rep}(\alpha)) \geq 1-2\alpha$.*

Theorem 5.1 holds because $\frac{2}{B}\sum_{b=1}^{B} \pi_b(u)$ (double the test statistic) is a valid p-value for the stated test (Rüschendorf, 1982; Meng, 1994; Barber et al., 2021; Vovk and Wang, 2020). In practice, however, $C^{\text{rep}}(\alpha)$ has close to $100(1-\alpha)\%$ coverage. The guaranteed level $1-2\alpha$ coverage and empirical level $1-\alpha$ coverage is analogous to the coverage of the jackknife+ method (Barber et al., 2021), which constructs conformal sets through leave-one-out prediction.

5.2.2 Supervised Prediction for a New Distribution

In the supervised setting, each group $\mathcal{D}_1, \ldots, \mathcal{D}_k$ has iid data

$$\mathcal{D}_j = \{(X_{j1}, Y_{j1}), \ldots, (X_{jn_j}, Y_{jn_j})\} \sim P_j$$

where $P_1, \ldots, P_k \sim \Pi$. Suppose we have a new distribution $P_{k+1} \sim \Pi$ and $(X, Y) \sim P_{k+1}$. Assuming that we only observe $X = x$, we want a prediction region for $Y = y$.

Method 1: Pooling CDFs

Similar to the unsupervised setting, we consider a method that averages empirical CDFs across groups. We first consider a method that is asymptotically valid as $k \to \infty$, regardless of the choice of model. Let $[k] = \{1, \ldots, k\}$. We start by pooling the observations from some strict subset $k_0 \subset [k]$ of the k groups to fit a model $\widehat{\mu}(X)$ as an estimator of $\mathbb{E}[Y \mid X]$. We use the remaining groups to fit the residuals $R_{ji} = |Y_{ji} - \widehat{\mu}(X_{ji})|$, $j \in [k] \backslash k_0$, $i = 1, \ldots, n_j$. Now for each $j \in [k] \backslash k_0$, we define group j's empirical CDF of the residuals

$$\widehat{F}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(R_{ji} \leq t).$$

We define

$$\widehat{q}_k(\alpha) = \inf \left\{ t \in \mathbb{R} : \frac{1}{|[k] \backslash k_0|} \sum_{j \in [k] \backslash k_0} \widehat{F}_j(t) \geq \alpha \right\}.$$

The $1-\alpha$ conformal prediction set is $C^{\text{poolCDF}}(x; \alpha) = [\widehat{\mu}(x) - \widehat{q}_k(1-\alpha), \widehat{\mu}(x) + \widehat{q}_k(1-\alpha)]$. For a proof of Theorem 5.2.3, see the Appendix.

Theorem 5.2.3. *Fit a model $\widehat{\mu}(X)$ as an estimator of $\mathbb{E}[Y \mid X]$ using the observations in groups $k_0 \subset [k]$. (Hence, this model stays fixed as k grows.) For $(X, Y) \sim \overline{\Pi}$, assume $\overline{\Pi}(|Y - \widehat{\mu}(X)| \leq t)$ is strictly increasing in t. Then $\overline{\Pi}(Y \in C^{poolCDF}(X; \alpha)) \xrightarrow{p} 1 - \alpha$ as $k \to \infty$.*

Under stronger parametric assumptions, we consider a second asymptotically valid approach ($k \to \infty$) that does not require sample splitting. Suppose (X, Y) arise from a parametric model

$Y = \mu(X;\theta) + \epsilon$, where $\mu(\cdot)$ is known, θ is unknown, and ϵ has a zero-mean distribution. We pool all of the $m = \sum_{j=1}^{k} n_j$ observations to fit $\widehat{\theta}$, using the true parametric model $\mu(\cdot)$. Thus, at any X, our point prediction is $\widehat{Y} = \mu(X;\widehat{\theta})$. We have the following residuals under the true θ and the estimated $\widehat{\theta}$:

$$\begin{aligned} R_{ji}(\theta) &= |\mu(X_{ji};\theta) - Y_{ji}| \\ R_{ji}(\widehat{\theta}) &= \left|\mu(X_{ji};\widehat{\theta}) - Y_{ji}\right|. \end{aligned}$$

The empirical CDFs of these residuals are

$$\begin{aligned} \widehat{F}_{j,\theta}(t) &= \frac{1}{n_j}\sum_{i=1}^{n_j} I(R_{ji}(\theta) \leq t) \\ \widehat{F}_{j,\widehat{\theta}}(t) &= \frac{1}{n_j}\sum_{i=1}^{n_j} I(R_{ji}(\widehat{\theta}) \leq t). \end{aligned}$$

Averaging the $\widehat{F}_{j,\widehat{\theta}}$ values, we obtain a sample quantile

$$\widehat{q}_k(\widehat{\theta};\alpha) = \inf\left\{ t \in \mathbb{R} : \frac{1}{k}\sum_{j=1}^{k} \widehat{F}_{j;\widehat{\theta}}(t) \geq \alpha \right\}.$$

Under the parametric assumptions in Theorem 5.2.4 (proved in the Appendix), $C^{\text{param}}(x;\alpha) = [\mu(x;\widehat{\theta}) - \widehat{q}_k(1-\alpha), \mu(x;\widehat{\theta}) + \widehat{q}_k(1-\alpha)]$ is an asymptotically valid prediction set as $k \to \infty$.

Theorem 5.2.4. *Suppose (X,Y) arises from a parametric model $Y = \mu(X;\theta) + \epsilon$, where $\mu(\cdot)$ is known, θ is unknown, and ϵ has a zero-mean distribution. Assume $\overline{\Pi}(Y - \mu(X;\theta) \leq t)$ is strictly increasing in t. Assume the true θ satisfies*

$$\frac{1}{k}\sum_{j=1}^{k} \sup_t |\widehat{F}_{j,\widehat{\theta}}(t) - \widehat{F}_{j,\theta}(t)| \xrightarrow{p} 0$$

as $k \to \infty$, and assume that for $\delta > 0$, $\lim_{k\to\infty} \overline{\Pi}(|\mu(X;\theta) - \mu(X;\widehat{\theta})| > \delta) = 0$. For $C^{param}(x;\alpha)$ as defined above, $\lim_{k\to\infty} \overline{\Pi}(Y \in C^{param}(x;\alpha)) = 1 - \alpha$.

Method 2: Subsampling Once

For the single subsample method, we randomly select one observation from each of the k groups. This creates a sample of k pairs of iid observations (X, Y). Suppose that we now have a new observation $(X_{k+1}, Y_{k+1}) \sim P_{k+1}$, but we only observe X_{k+1}. Letting $X_{k+1} = x$, we have an augmented X sample $(X_1, \ldots, X_k, X_{k+1})$. For each possible y, we test $H_0 : Y_{k+1} = y$ at a $1 - \alpha$ confidence level using the following procedure: Assume $Y_{k+1} = y$, giving an augmented Y sample of $(Y_1, \ldots, Y_k, Y_{k+1})$. Using the sample augmented with (x, y) as training data, fit a model $\widehat{\mu}_{(x,y)}(X)$ as an estimator of $\mathbb{E}[Y \mid X]$. Then compute conformity scores $R_i(x, y) = |\widehat{\mu}_{(x,y)}(X_i) - Y_i|$, $i = 1, \ldots, k+1$. The p-value for the test of $H_0 : Y_{k+1} = y$ is $\pi(x, y) = (k + 1)^{-1} \sum_{i=1}^{k+1} I(R_i(x, y) \geq R_{k+1}(x, y))$. The $1 - \alpha$ conformal prediction set is $C^{\text{sub}}(x; \alpha) = \{y \in \mathbb{R} : \pi(x, y) \geq \alpha\}$. Since the subsample of k observations is an iid sample, this method is justified by Section **??**.

Method 3: Repeated Subsampling

We modify the supervised subsampling method to incorporate B subsamples of a single observation from each of the k groups. For the b^{th} subsample, $(X_{1b}, Y_{1b}), \ldots, (X_{kb}, Y_{kb})$ contains one observed pair from each of the k groups. Suppose $X_{k+1} = x$ is a new observed covariate from a new group. Conformal prediction is implicitly testing $H_0 : Y_{k+1} = y$ versus $H_1 : Y_{k+1} \neq y$, and the level $1 - \alpha$ conformal prediction set is the set of values at which we would not reject H_0 under the given construction. Using the b^{th} subsample augmented with (x, y), we construct residuals $R_{b,i}(x, y)$ in the same manner as Section 5.2.2. Then $\pi_b(x, y) = (k + 1)^{-1} \sum_{i=1}^{k+1} I(R_{b,i}(x, y) \geq R_{b,k+1}(x, y))$ is a valid p-value for the stated test. We construct $\pi_b(x, y)$ for B subsamples. We define a conformal prediction set

$$C^{\text{rep}}(x; \alpha) = \left\{ y : \frac{1}{B} \sum_{b=1}^{B} \pi_b(x, y) \geq \alpha \right\}.$$

Theorem 5.2. *For $C^{rep}(x; \alpha)$ as defined above, $\overline{\Pi}(Y \in C^{rep}(X; \alpha)) \geq 1 - 2\alpha$.*

Similar to Theorem 5.1 in the unsupervised case, Theorem 5.2 is true because $(2/B) \sum_{b=1}^{B} \pi_b(x, y)$ is a valid p-value for the stated test. As in the unsupervised case, $C^{\text{rep}}(x; \alpha)$ has empirical coverage of approximately $1 - \alpha$.

5.2.3 Unsupervised Prediction for an Observed Group

Task 2 considers the question of predicting a new observation on an existing subject rather than on a future subject. We assume without loss of generality that we wish to predict a new observation from subject 1. We explore two methods for creating conformal intervals to capture the new observation in the unsupervised setting. The first method is a standard conformal procedure using subject 1's data. The second method "borrows" information from other subjects to obtain a shrinkage estimator of the mean of subject 1's data. Then it performs conformal prediction using this shrinkage estimator. The validity of either method follows from the usual theory described in Section **??**, but we expect that using a shrinkage estimator will lead to smaller prediction sets.

Method 1: Isolate Single Group

For the first method, we only use subject 1's data to construct a $1-\alpha$ conformal prediction set. We propose a new value of y, and we wish to test $H_0 : Y_{1,n_1+1} = y$ at a $1-\alpha$ confidence level. Letting $Y_{1,n_1+1} = y$, we have an augmented data vector $(Y_{1,1}, \dots, Y_{1,n_1}, Y_{1,n_1+1})$ for subject 1. We define $\overline{Y}_1 = \frac{1}{n_1+1}\sum_{i=1}^{n_1+1} Y_{1,i}$. Then we calculate conformity scores $R_i = \left|Y_{1,i} - \overline{Y}_1\right|$, $i = 1, \dots, n_1+1$. The p-value for the test of $H_0 : Y_{1,n_1+1} = y$ is $\pi(y) = \frac{1}{n_1+1}\sum_{i=1}^{n_1+1} I(R_i \geq R_{n_1+1})$. We invert this test to obtain a $1-\alpha$ conformal prediction set $C^{\text{isolate}}(\alpha) = \{y : \pi(y) \geq \alpha\}$.

Method 2: James-Stein Shrinkage

The second method uses all data from $\mathcal{D}_1, \dots, \mathcal{D}_k$ and performs shrinkage to predict a new observation from subject 1. Again, we construct a $1-\alpha$ conformal prediction set for a new observation from subject 1. In this method, however, we use data from all subjects. We propose a new value of y, and we wish to test $H_0 : Y_{1,n_1+1} = y$ at a $1-\alpha$ confidence level. We define $\overline{Y}_1 = \frac{1}{n_1+1}\sum_{i=1}^{n_1+1} Y_{1,i}$. Then for $j = 2, \dots, k$, we define $\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{j,i}$. We also estimate $\widehat{\sigma}_1^2$ as the sample variance of $(Y_{1,1}, \dots, Y_{1,n_1}, Y_{1,n_1+1})$ and $\nu = \frac{1}{k}\sum_{j=1}^{k} \overline{Y}_j$.

Now in place of $\overline{Y}_1$ in the conformity scores, we use the James-Stein shrinkage estimator:

$$\widetilde{Y}_1 = \left(1 - \frac{(k-2)\frac{\widehat{\sigma}_1^2}{n_1}}{\sum_j (\overline{Y}_j - \nu)^2}\right)_+ (\overline{Y}_1 - \nu) + \nu,$$

where $(x)_+ = \max(x, 0)$. The rest of the procedure mirrors Method 1. We calculate conformity scores $R_i = |Y_{1,i} - \widetilde{Y}_1|$, $i = 1, \ldots, n_1 + 1$. For the proposed value of y, we obtain a p-value $\pi(y) = \frac{1}{n_1+1} \sum_{i=1}^{n_1+1} I(R_i \geq R_{n_1+1})$. The $1 - \alpha$ conformal prediction set is $C^{\text{shrinkage}}(\alpha) = \{y : \pi(y) \geq \alpha\}$.

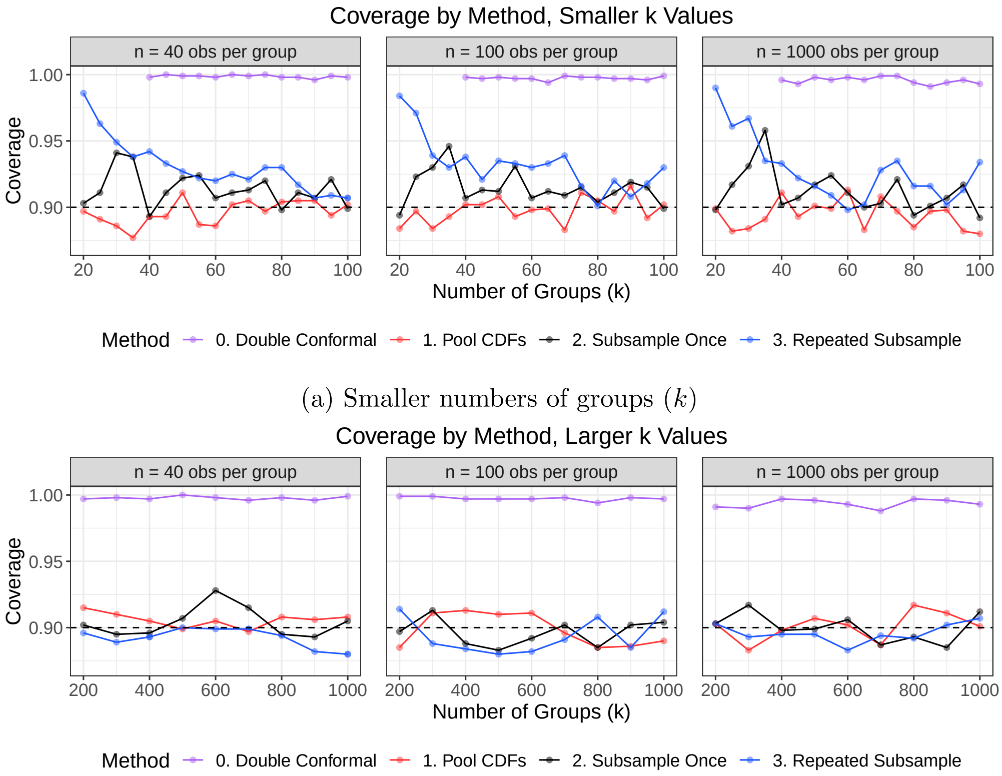
5.3 Simulations

We present simulations on the unsupervised and supervised methods for predicting a new observation from a new group, as well as the unsupervised methods for predicting a new observation from an observed group.

5.3.1 The Unsupervised Case

We begin by generating data from k distributions. We draw $\theta_1, \ldots, \theta_k \sim N(\mu, \tau^2)$, with $\mu = 0$ and $\tau^2 = 1$. Then we simulate $Y_{j1}, \ldots, Y_{jn_j} \sim N(\theta_j, 1)$. We allow the number of observations per group (n_j) to equal 40, 100, and 1000. We vary the number of groups (k) from 20 to 100 in increments of 5 and from 200 to 1000 in increments of 100. We construct prediction sets following the procedures from Section 5.2.1. The repeated subsampling intervals (Method 3) use $B = 100$ subsamples. Each simulation generates a data sample, draws a new Y from a new distribution, constructs prediction sets $C(\alpha)$, determines the size of each prediction set, and checks whether $Y \in C(\alpha)$. We set $\alpha = 0.1$, and we perform 1000 simulations at each combination of k and n_j.

Figure 5.1 displays the empirical coverage of Method 0 (double conformal), Method 1 (CDF pooling), Method 2 (subsampling once), and Method 3 (repeated subsampling) from Section 5.2.1. The coverage is the proportion of simulations for which $Y \in C(\alpha)$. The double conformal method consistently overcovers, with coverage close to 1. The CDF pooling method undercovers at small values of k (e.g., $k \leq 35$) but has approximate $1-\alpha$ coverage for larger k. Single subsampling tends to overcover for small k and has approximately $1 - \alpha$ coverage for large k. Repeated subsampling sometimes overcovers for small k and has coverage of about $1 - \alpha$ for large k. Across simulations, the number of observations per group does not appear to affect the relationship between coverage and number of groups.



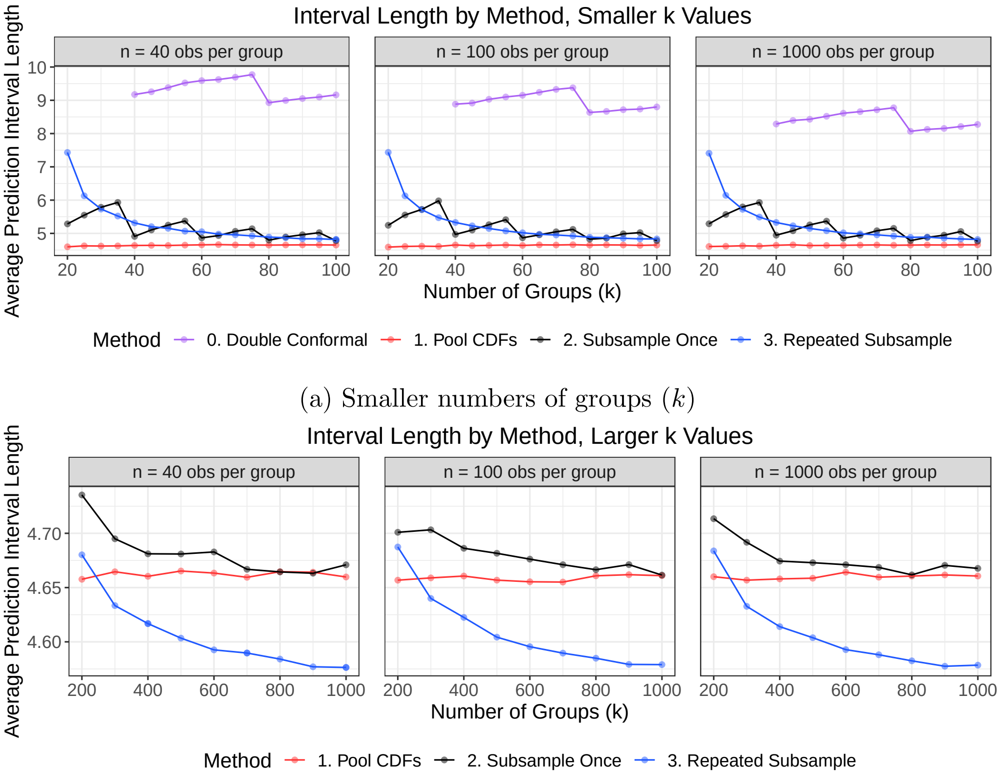
(a) Smaller numbers of groups (k)

(b) Larger numbers of groups (k)

Figure 5.1: Coverage of unsupervised prediction sets for a new group's observation.

Figure 5.2 shows the average prediction set lengths. For small k, the pooling method has the smallest intervals, the single subsampling and repeated subsampling methods have the next largest intervals (mostly on par), and double conformal has the largest intervals. For large k, repeated subsampling has the smallest intervals, followed by pooling and then single subsampling. Figure 5.2b excludes the length of the double conformal intervals. For $k \geq 200$ and $n \in \{40, 100, 1000\}$, the double conformal intervals have average lengths between 7.8 and 8.6. In this case, it is difficult to distinguish between the pooling, single subsample, and repeated subsample lengths on a scale that includes the double conformal lengths.

We recommend the repeated subsampling method. This method has guaranteed coverage at level $1 - 2\alpha$, but in practice it tends to cover at level $1 - \alpha$. Furthermore, its prediction sets are about the same size as the single subsample method for small k. For large k, its prediction sets are often smaller than both the pooling and the single subsample methods.



(a) Smaller numbers of groups (k)

(b) Larger numbers of groups (k)

Figure 5.2: Average unsupervised prediction set length for a new group's observation.

The examples in Figures 5.1 and 5.2 have considered cases with balanced numbers of observations in each group. We now consider a highly unbalanced case: one group has 200 times as many observations as each of the other groups, and the between-group variation exceeds the within-group variation by two orders of magnitude. We take $Y_{ij} \sim N(\theta_j, \sigma^2 = 0.1)$ where $\theta_j \sim N(0, \tau^2 = 10)$, $n_1 = 1000$, and $n_j = 5$ for $2 \leq j \leq k$. We let k vary from 20 to 100 in increments of 5. Figure 5.3 shows that the CDF pooling method undercovers by about 0.05 at most, while the single subsample and repeated subsample methods typically have at least nominal coverage. The CDF pooling method produces the smallest prediction sets, and the single and repeated subsampling methods have similar average prediction set lengths. Thus, the behavior we observe in this highly unbalanced case is similar to the balanced case.

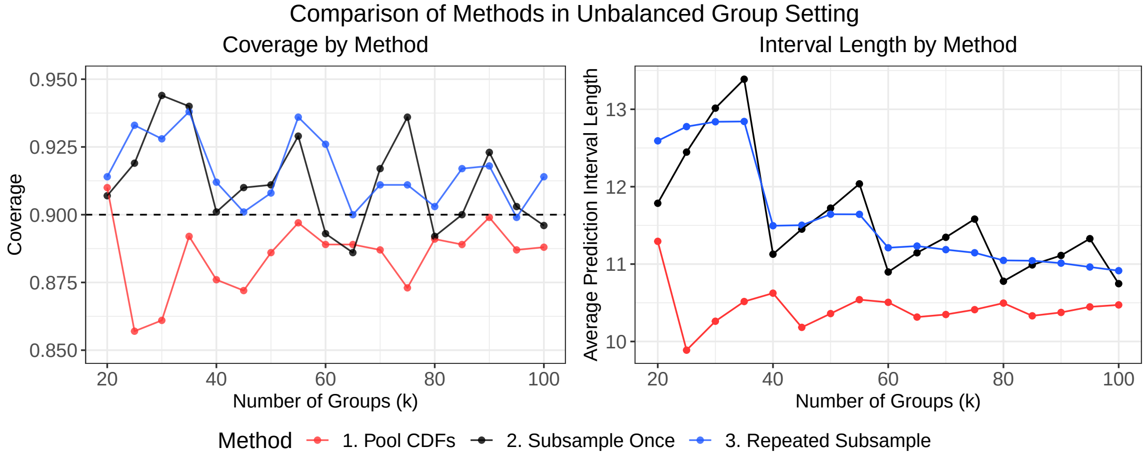


Figure 5.3: In this unbalanced setting, one group has 1000 observations, and the remaining $k-1$ groups have 5 observations. Intervals constructed at level $\alpha = 0.1$.

5.3.2 The Supervised Case

Now the data consist of k groups where $\mathcal{D}_j = \{(X_{j1}, Y_{j1}), \ldots, (X_{jn_j}, Y_{jn_j})\}$. For (X, Y) from a new distribution, we are given X, and we want to predict Y. For each simulation, we generate data from k distributions. We draw

$$\begin{aligned} \theta_1, \ldots, \theta_k &\sim N(\mu, \tau^2) \\ X_{j1}, \ldots, X_{jn_j} &\sim N(0,1) \\ \epsilon_{j1}, \ldots, \epsilon_{jn_j} &\sim N(0,1). \end{aligned}$$

We let $Y_{j\ell} = \theta_j X_{j\ell} + \epsilon_{j\ell}$, $j = 1, \ldots, k$, $\ell = 1, \ldots, n_j$.

Now we draw a new $X \sim N(0,1)$, $\theta_{k+1} \sim N(\mu, \tau^2)$, and response $Y \sim N(\theta_{k+1} X, 1)$. Treating θ_{k+1} and Y as unknown, we wish to predict Y from the observed $X = x$. We construct prediction sets following the procedures from Section 5.2.2. For the CDF pooling method (Method 1), we use the approach justified by Theorem 5.2.3. We pool the observations from $k_0 = \lfloor k/2 \rfloor$ groups to fit a one-parameter linear regression model $\widehat{\mu}(X) = \widehat{\theta} X$. Then we use the remaining groups for quantile estimation. For the subsampling methods (Methods 2 and 3), we fit $\widehat{\mu}_{(x,y)}(X; \widehat{\theta}) = \widehat{\theta} X$ using subsamples of one observation per group, augmented with (x, y). (Recall that x is the new observed value, and y is a hypothesized value.) For Method 3, we use $B = 100$ subsamples.

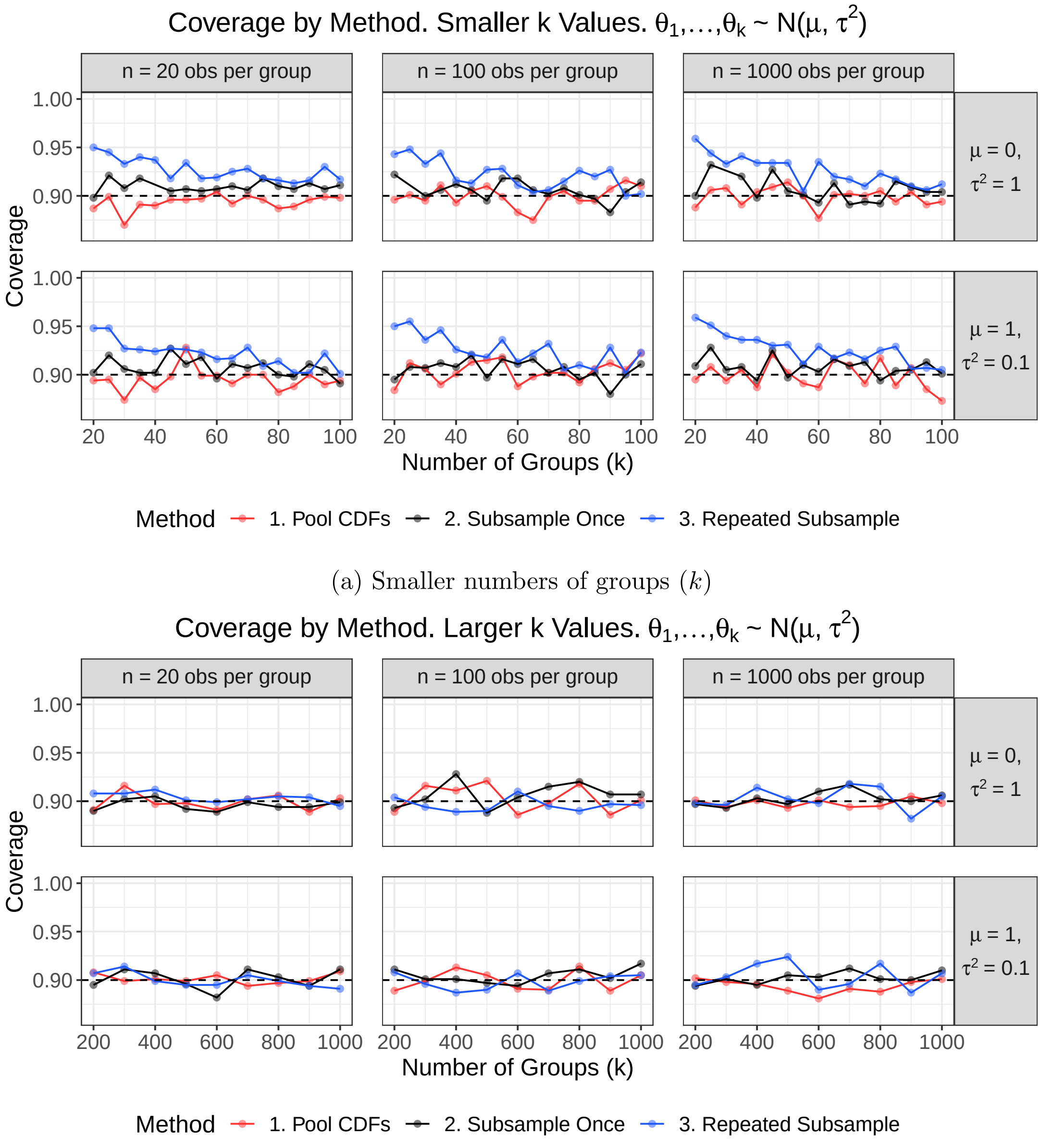
We let the number of observations per group (n_j) equal 20, 100, and 1000. We vary the number of groups (k) from 20 to 100 in increments of 5 and from 200 to 1000 in increments of 100. To draw the θ parameters, we try $\mu = 0$, $\tau^2 = 1$ and $\mu = 1$, $\tau^2 = 0.1$. The first pair of parameters represents a case where the relationships between X and Y may be quite different across groups. The second pair of parameters is a case where the groups have similar trends that relate X and Y. We perform 1000 simulations at each combination of k, n_j, and (μ, τ^2). We set $\alpha = 0.1$. Each simulation generates a data sample, draws a new (X, Y) from a new distribution, constructs prediction sets $C(X; \alpha)$, determines the size of each prediction set, and checks whether $Y \in C(X; \alpha)$.

Figure 5.4 shows the coverage of the supervised methods in these two settings. The coverage is the proportion of simulations for which $Y \in C(X; \alpha)$. At both $(\mu = 0, \tau^2 = 1)$ and $(\mu = 1, \tau^2 = 0.1)$, all three methods have coverage close to $1 - \alpha$ for all n and k. For small k, the repeated subsample method often overcovers by up to 0.05. The pooling method undercovers more often than the other two methods.

Figure 5.5 shows the average length of the prediction sets from these supervised methods. In almost all cases, the pooling intervals are the smallest, followed by the single subsampling intervals, and the repeated subsampling intervals are the largest. Overall, the single subsampling method appears to be the best choice in this setting. The single subsampling method has coverage of approximately $1 - \alpha$ even for small k. In addition, this method produces smaller prediction sets than the repeated subsampling method, which also has coverage at or above $1 - \alpha$ for small k.

5.3.3 Unsupervised Prediction for an Observed Group

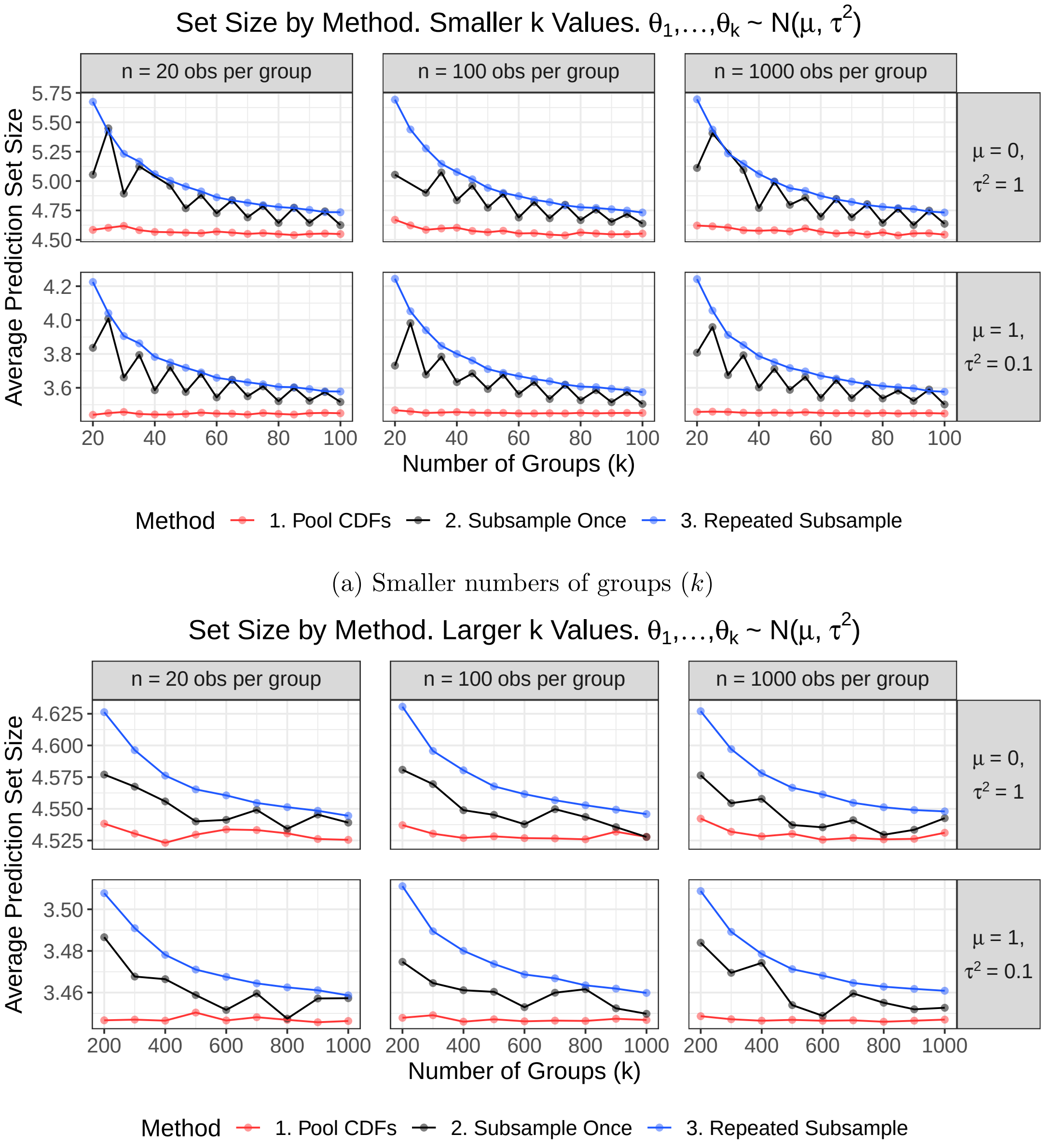
To construct a conformal prediction set for a new observation from an observed group, we proposed two approaches. Method 1 constructs a prediction set using only the observations from the distribution of interest, and Method 2 uses a residual based on the James-Stein shrinkage estimator to borrow strength across distributions. We compare the results of these methods under two data generation processes. We draw subject-specific means $\theta_1, \ldots, \theta_k \sim N(0, 1)$. Then for $j = 1, \ldots, k$, we generate $Y_{j1}, Y_{j2}, \ldots, Y_{jn_j} \sim N(\theta_j, \sigma^2)$. We consider $\sigma^2 = 1$ and $\sigma^2 = 100$. Across all simulations, we set $n_j = 20$. We vary k from 5 to 1000 in increments of 5. At each choice of k, we perform 1000 simulations at $\alpha = 0.1$. Each simulation generates a data sample, draws



(a) Smaller numbers of groups (k)

(b) Larger numbers of groups (k)

Figure 5.4: Coverage of supervised conformal prediction sets for an outcome from a new group.



(a) Smaller numbers of groups (k)

(b) Larger numbers of groups (k)

Figure 5.5: Average size of supervised conformal prediction sets for an outcome from a new group.

another observation $Y \sim N(\theta_1, \sigma^2)$ from subject 1's distribution, constructs prediction sets $C(\alpha)$ for subject 1, determines the size of each prediction set, and checks whether $Y \in C(\alpha)$.

Figures 5.6 and 5.7 show the results of shrinkage Methods 1 and 2 for predicting a new observation from subject 1. The left panels use $\sigma^2 = 1$, and the right panels use $\sigma^2 = 100$. The simulations confirm that basing the conformal residuals on a shrinkage estimator can lead to smaller predictive sets.

- Figure 5.6 shows the empirical coverage of shrinkage Methods 1 and 2 with the coverage level fixed at $1 - \alpha = 0.9$. The coverage is typically about 0.9, and there is no clear difference in performance between shrinkage Methods 1 and 2.
- Figure 5.7 plots the average size of conformal sets from shrinkage Methods 1 and 2 in both data set-ups. When $\sigma^2 = 1$, Methods 1 and 2 produce conformal sets with similar length. When $\sigma^2 = 100$, Method 2 consistently produces smaller conformal sets than Method 1. This shows that shrinkage is especially beneficial when the within-group variance is high, relative to the between-group variance. There does not appear to be a trend in conformal set size as the number of groups increases.

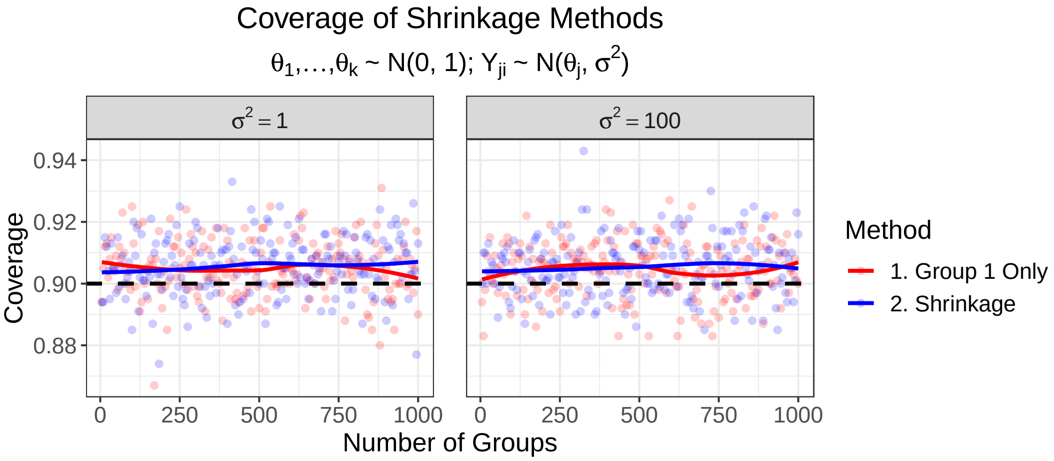


Figure 5.6: Coverage of conformal methods for predicting a new observation from an observed group. 20 observations per group. Loess smoothing for trend visualization.

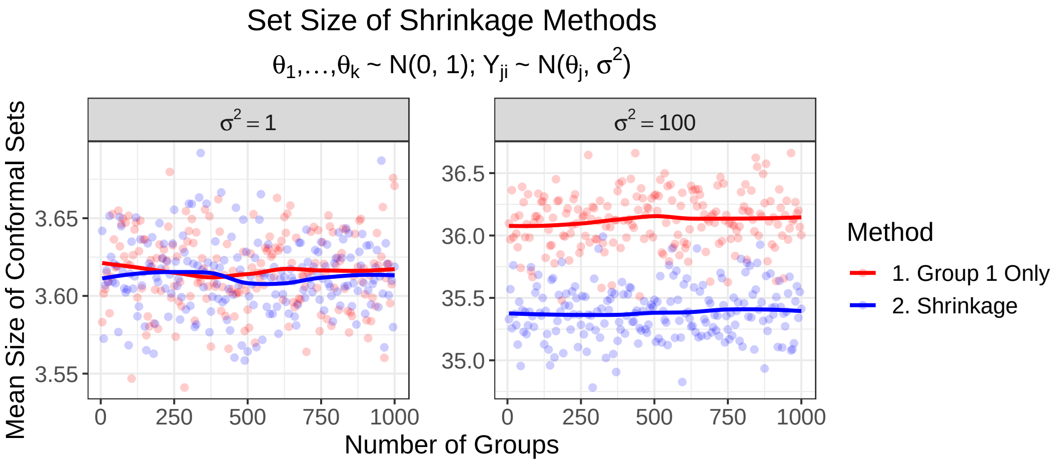


Figure 5.7: Average size of conformal sets for predicting a new observation from an observed group. 20 observations per group. Loess smoothing for trend visualization.

5.4 Data Example

We now consider a data example from a sleep deprivation study (Balkin et al., 2000; Belenky et al., 2003). This study evaluates 18 commercial vehicle drivers on a series of tests after $0, 1, 2, \ldots, 9$ nights of restriction to 3 hours of sleep. On each day, each subject takes a series of reaction time tests, and the experimenters record each subject's average reaction time. The data are available in the `sleepstudy` dataset of `R`'s `lme4` package (Bates et al., 2015).

We restructure the data to fit regressions that predict average sleep-deprived reaction time (Y) from number of days of sleep deprivation (X_1) and the subject's baseline (Day 0) average reaction time under their normal sleep amount (X_2). For each individual j, we observe nine triplets (X_{1j}, X_{2j}, Y_j). For the purpose of this demonstration, we treat each (X_{1j}, X_{2j}, Y_j) as a random draw from a subject-specific distribution P_j. (Alternatively, we could treat X_{1j} as fixed, X_{2j} as random, and Y_j as a random draw from $P_{j,Y|X}$. These methods are valid as long as the conformity scores are exchangeable, as discussed below.) The variable X_{1j} ranges from 1 to 9 days, and the baseline time X_{2j} is measured once for each subject j. Across subjects, X_2 ranges from 199 to 322 milliseconds, and Y ranges from 194 to 466 milliseconds. Our fitted regression models have the form $\widehat{Y} = \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2$. We have also considered a model that includes an intercept. This does not make much of a difference when assessing whether the residuals appear to be exchangeable.

Suppose we observe $(X_1, X_2) = \mathbf{x}$ on a nineteenth individual, and we want to predict the associated Y. We construct conformal prediction sets $C(X_1, X_2; \alpha)$ such that $P(Y \in C(X_1, X_2; \alpha)) \geq 1 - \alpha$. We use the constructions from Section 5.2.2, and we use conformity scores of $R_i(\mathbf{x}, y) = |Y_i - \widehat{Y}_i|$. The CDF pooling method (Method 1) uses the process justified by Theorem 5.2.3. We fit the regression model on the pooled observations of 9 of the 18 individuals, and we estimate the quantiles from the remaining individuals' residuals. The one subsample method (Method 2) randomly selects one observation per individual. We augment the subsample with (X_1, X_2, y) for the observed (X_1, X_2) and some proposed y. We fit the regression model on this augmented sample of size 19. The repeated subsampling method (Method 3) averages p-values across $B = 100$ repetitions of Method 2 using the same (X_1, X_2, y). Method 1 is asymptotically valid ($k \to \infty$) if the conformity scores are exchangeable across all observations used for quantile estimation. Methods 2 and 3 are valid if the conformity scores are exchangeable for any subsample of one observation per subject. From visual inspection (not shown), the Method 1 exchangeability assumption may not be met. Several subjects have particularly high or particularly low absolute residuals on all of their observations. In addition, we only have data on $k = 18$ subjects. The Method 2/3 exchangeability assumption seems reasonable, from plots of the absolute residuals when we fit and evaluate the model on one observation per subject.

Figure 5.8 shows the output and size of the conformal methods at $\alpha = 0.10$. The left panel shows the prediction sets at $X_1 = \{1, 5, 9\}$ and at $X_2 = \{200, 230, 260, 290, 320\}$. For most (X_1, X_2) combinations, all three intervals have similar centers. The right panel compares the length of the three intervals over an expanded set of (X_1, X_2) combinations. The CDF pooling method produces the smallest predictions in most cases, but this method is only asymptotically valid ($k \to \infty$). The repeated subsample method has smaller intervals than the single subsample method in about half of the cases. Favorably, the repeated subsample method has the least variation in interval lengths across X_2 values for a given X_1.

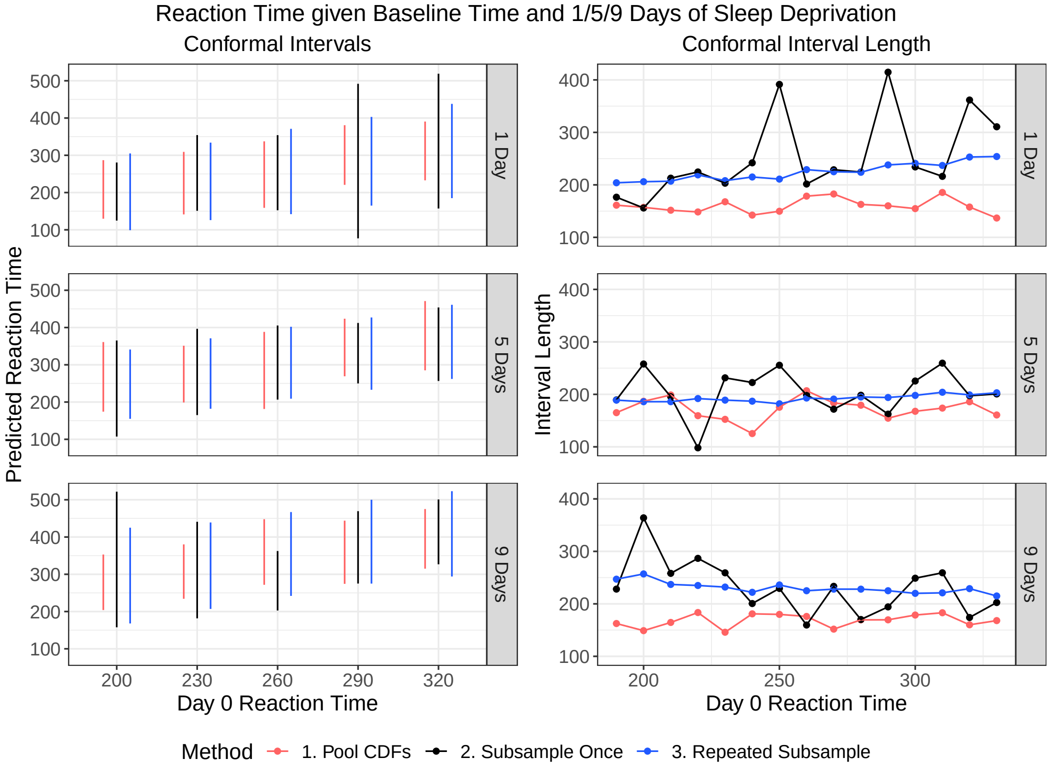


Figure 5.8: Conformal intervals for predicting sleep-deprived reaction time given baseline reaction time and days of sleep deprivation. The left panel shows 90% prediction sets at all combinations of $X_1 = \{1, 5, 9\}$ and $X_2 = \{200, 230, 260, 290, 320\}$. The right panel shows 90% prediction set lengths at all combinations of $X_1 = \{1, 5, 9\}$ and X_2 ranging from 190 to 330 in increments of 10.

We also consider the estimated coverage of these three methods. The pooling method is only asymptotically valid ($k \to \infty$) at level $1 - \alpha$, the single sample method is valid at level $1 - \alpha$ but has more variation, and the repeated subsample method only has guaranteed coverage at level $1 - 2\alpha$. We evaluate coverage by holding out 1 of the 18 individuals, selecting a triplet (X_1, X_2, Y) from the held-out individual, fitting a prediction set C on the remaining 17 subjects, and checking whether $Y \in C(X_1, X_2; \alpha)$. We perform this procedure $18 \times 9 = 162$ times, holding out each of the 9 observations from each of the 18 subjects once. The proportion of simulations in which $Y \in C(X_1, X_2; \alpha)$ for held-out (X_1, X_2, Y) is an estimate of the coverage of these methods. Method 1 has algorithmic randomness in the individuals selected for model fitting (8 individuals) versus quantile estimation (9 individuals). Methods 2 and 3 have algorithmic randomness in the observations selected for each subsample. Thus, we repeat this coverage estimation procedure 1000 times.

Table 5.1 shows the coverage proportions at $\alpha \in \{0.10, 0.15, 0.20\}$. For each method, Table 5.1 displays the average coverage, the 2.5^{th} percentile, and the 97.5^{th} percentile over 1000 simulations. On average, we see that CDF pooling undercovers by about 0.02 to 0.03, and the subsampling methods overcover by about 0.03 to 0.06. The repeated subsampling method has slightly higher coverage than the single subsampling method, but repeated subsampling also has lower variation in coverage. Overall, the repeated subsampling method is the best choice in this setting. This method achieves coverage of at least $1 - \alpha$ and has lower variation in set size and coverage than the other two methods.

Table 5.1: Estimated coverage of conformal intervals on sleep deprivation data. At varying α, we show the average coverage and (2.5^{th} %ile, 97.5^{th} %ile) coverage intervals of each method over 1000 simulations.

Method	$\alpha = 0.10$	$\alpha = 0.15$	$\alpha = 0.20$
1. CDF Pooling	0.87 (0.84, 0.90)	0.83 (0.80, 0.86)	0.78 (0.75, 0.81)
2. Subsample Once	0.94 (0.92, 0.97)	0.89 (0.86, 0.92)	0.83 (0.80, 0.87)
3. Repeated Subsample	0.95 (0.94, 0.96)	0.91 (0.90, 0.92)	0.84 (0.83, 0.85)

5.5 Conclusion

We have proposed and compared several methods for constructing distribution-free prediction sets for random effects models. We believe these are the first such methods. We consider a CDF pooling method that is asymptotically valid as $k \to \infty$, a single subsample method that uses one observation per group, and a repeated subsample method that repeatedly selects one observation per group and averages p-values over subsamples. The single subsample method is valid at level $1 - \alpha$. The repeated subsample method has guaranteed coverage at level $1 - 2\alpha$ but tends to have coverage of at least $1 - \alpha$ in practice.

Based on our simulations and data example, we recommend the repeated subsample method. In the unsupervised simulations, this method has coverage close to $1 - \alpha$ and has the smallest prediction sets for large k. In the sleep data example, this method has coverage of at least $1 - \alpha$ and has more stable size and coverage than the other methods. Pooling CDFs often produces small prediction sets but is only asymptotically valid. Single subsampling is valid at level $1 - \alpha$ but requires throwing away most of the data. Repeated subsampling has less algorithmic variation than single subsampling, which makes this method more stable and more reproducible. In the supervised simulations, repeated subsampling is a reasonable choice, but pooling and single subsampling both produce smaller sets with approximately nominal coverage. It is a curiosity that single subsampling (which ignores most of the data) produces smaller prediction sets than repeated subsampling in this case.

The main focus of this paper has been the prediction of a new observation on a new subject. In the unsupervised setting, we also considered the simpler problem of predicting a future observation on an existing subject. Future work may consider alternatives to the James-Stein shrinkage-based residual or may incorporate repeated subsampling into the shrinkage approach. In addition, random effects conformal prediction for an existing subject in the supervised setting remains an open problem. Space does not permit a thorough investigation of these problems here, but we hope to report more on this problem in a future paper.

Bibliography

An, M. Y. (1997). Log-concave probability distributions: Theory and statistical testing. *Duke University Dept of Economics Working Paper*, (95-03). 24, 25, 30, 31, 33

Axelrod, B., Diakonikolas, I., Stewart, A., Sidiropoulos, A., and Valiant, G. (2019). A polynomial time algorithm for log-concave maximum likelihood via locally exponential families. In *Advances in Neural Information Processing Systems*, pages 7723–7735. 5

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Economic Theory*, 26(2):445–469. 24

Balkin, T., Thome, D., Sing, H., Thomas, M., Redmond, D., Wesensten, N., Williams, J., Hall, S., and Belenky, G. (2000). Effects of Sleep Schedules on Commercial Motor Vehicle Driver Performance. Technical report, United States. Department of Transportation. Federal Motor Carrier Safety Administration. 81

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive Inference with the Jackknife+. *The Annals of Statistics*, 49(1):486–507. 68, 69

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282. 2

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48. 81

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., and Balkin, T. J. (2003). Patterns of Performance Degradation and Restoration during

Sleep Restriction and Subsequent Recovery: A Sleep Dose-Response Study. *Journal of Sleep Research*, 12(1):1–12. 81

Booth, J. G. and Hobert, J. P. (1998). Standard Errors of Prediction in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 93(441):262–272. 65

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360. 30

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376. 2

Calvin, J. A. and Sedransk, J. (1991). Bayesian and Frequentist Predictive Inference for the Patterns of Care Studies. *Journal of the American Statistical Association*, 86(413):36–48. 65

Chacón, J. E. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19(2):375–398. 30

Chen, J., Li, P., et al. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542. 5

Chen, W., Chun, K.-J., and Barber, R. F. (2018). Discretized Conformal Prediction for Efficient Distribution-free Inference. *Stat*, 7(1):e173. 58

Chen, W., Mazumder, R., and Samworth, R. (2021). A new computational framework for log-concave density estimation. *arXiv preprint arXiv:2105.11387*. 27

Chun, S. Y. and Shapiro, A. (2009). Normal versus noncentral chi-square asymptotics of misspecified models. *Multivariate Behavioral Research*, 44(6):803–827. 16, 124

Claggett, B., Xie, M., and Tian, L. (2014). Meta-analysis with Fixed, Unknown, Study-specific Parameters. *Journal of the American Statistical Association*, 109(508):1660–1671. 65

Cule, M., Gramacy, R., and Samworth, R. (2009). LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density. *Journal of Statistical Software*, 29(2). 26

Cule, M., Samworth, R., et al. (2010a). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics*, 4:254–270. 26, 52, 54, 139

Cule, M., Samworth, R., and Stewart, M. (2010b). Maximum likelihood estimation of a multidimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607. xi, 5, 24, 25, 26, 27, 31, 33, 36, 37, 38, 39, 47, 53

Dasgupta, S. and Schulman, L. J. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226. 23

Delignette-Muller, M. L. and Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34. 50

DerSimonian, R. and Laird, N. (1986). Meta-analysis in Clinical Trials. *Controlled clinical trials*, 7(3):177–188. 65

Dümbgen, L., Hüsler, A., and Rufibach, K. (2007). Active set and em algorithms for log-concave densities based on complete and censored data. *arXiv preprint arXiv:0707.4643*. 26

Dümbgen, L. and Rufibach, K. (2011). logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, 39(6):1–28. 26

Duong, T. (2021). *ks: Kernel Smoothing*. R package version 1.12.0. 30

Duong, T. and Hazelton, M. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15(1):17–30. 30

Duong, T. and Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506. 30

Grünwald, P., de Heide, R., and Koolen, W. M. (2020). Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE. 4, 19

Guo, F. R. and Richardson, T. S. (2020). On testing marginal versus conditional independence. *arXiv preprint arXiv:1906.01850v2*. 23

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. K. (2020). Nested Conformal Prediction and Quantile Out-of-Bag Ensemble Methods. *arXiv preprint arXiv:1910.10562v2*. 68

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:361–374. 10, 97, 98, 99, 100

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Barkeley Conference in Honor of J. Neyman and J. Kiefer*, volume 2, pages 807–810. 5

Ildstad, S. T., Evans Jr, C. H., et al. (2001). *Small clinical trials: Issues and challenges*. National Academies Press. 2

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351. 14, 120

Jones, M., Marron, J. S., and Park, B. U. (1991). A simple root n bandwidth selector. *The Annals of Statistics*, pages 1919–1932. 30

Kappel, F. and Kuntsevich, A. V. (2000). An implementation of Shor's r-algorithm. *Computational Optimization and Applications*, 15(2):193–205. 26

Kim, A. K., Samworth, R. J., et al. (2016). Global rates of convergence in log-concave density estimation. *Annals of Statistics*, 44(6):2756–2779. 27

Kur, G., Dagan, Y., and Rakhlin, A. (2019). Optimality of maximum likelihood for log-concave density estimation and bounded convex regression. *arXiv preprint arXiv:1903.05315*. 27

Laird, N. M. and Ware, J. H. (1982). Random-effects Models for Longitudinal Data. *Biometrics*, pages 963–974. 65

Lehmann, E. L. (2012). On likelihood ratio tests. In *Selected Works of EL Lehmann*, pages 209–216. Springer. 2

Lei, J., G?Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free Predictive Inference for Regression. *Journal of the American Statistical Association*, pages 1–18. 58

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free Prediction Sets. *Journal of the American Statistical Association*, 108(501):278–287. 58, 59, 60

Lei, J. and Wasserman, L. (2014). Distribution-free Prediction Bands for Non-parametric Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96. 58

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491):1084–1092. 5

Li, Q. (1999). *Estimation of mixture models*. Yale University. `http://www.stat.yale.edu/~arb4/students_files/JonathanLiThesis.pdf`. 19

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769. 10, 97, 100

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324. 5

McMahon, A. W., Watt, K., Wang, J., Green, D., Tiwari, R., and Burckart, G. J. (2016). Stratification, hypothesis testing, and clinical trial simulation in pediatric drug development. *Therapeutic innovation & regulatory science*, 50(6):817–822. 2

Medeiros, F. M. and Ferrari, S. L. (2017). Small-sample testing inference in symmetric and log-symmetric linear regression models. *Statistica Neerlandica*, 71(3):200–224. 2

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22(3):1142–1160. 68

Nagler, T. and Vatter, T. (2020). *kde1d: Univariate Kernel Density Estimation.* R package version 1.0.3. 50

Polland, D. (2015). A few good inequalities. In *Mini-empirical (draft)*. 14, 123

Prékopa, A. (1973). On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343. 31

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78. 30

Rüschendorf, L. (1982). Random Variables with Maximum Sums. *Advances in Applied Probability*, pages 623–632. 68

Sadinle, M., Lei, J., and Wasserman, L. (2018). Least Ambiguous Set-valued Classifiers with Bounded Error Levels. *Journal of the American Statistical Association*, pages 1–12. 58

Schofield, L. S., Junker, B., Taylor, L. J., and Black, D. A. (2015). Predictive Inference Using Latent Variables With Covariates. *Psychometrika*, 80(3):727–747. 65

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317. 42

Shafer, G. and Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(Mar):371–421. 58

Shor, N. Z. (2012). *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media. 25

Sitek, G. (2016). The modes of a mixture of two normal distributions. *Silesian Journal of Pure and Applied Mathematics*, 6(1):59–67. 33

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. 3, 10, 101

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World.* Springer Science & Business Media. 58, 59

Vovk, V. and Wang, R. (2020). Combining p-values via Averaging. *Biometrika.* 68

Wand, M. P. and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116. 30

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890. 2, 3, 4, 5, 19, 23, 25, 94, 95

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62. 3

Wong, W. H., Shen, X., et al. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362. 53, 136

Appendix

Appendix A

Proofs from Chapter 1

Theorem 1.0.1. $C_n^{split}(\alpha)$ *is a valid* $100(1-\alpha)\%$ *confidence set for* θ^**. As a consequence (and equivalently), when testing an arbitrary composite null* $H_0 : \theta^* \in \Theta_0$ *versus* $H_1 : \theta^* \in \Theta \setminus \Theta_0$*, rejecting* H_0 *when* $\Theta_0 \cap C_n^{split}(\alpha) = \emptyset$ *provides a valid level* α *hypothesis test. This rule reduces to rejecting* H_0 *if* $T_n(\widehat{\theta}_0) \geq 1/\alpha$*, where* $\widehat{\theta}_0 \in \arg\max_{\theta \in \Theta_0} \mathcal{L}_0(\theta)$ *is the null MLE.*

Proof. This result is due to Wasserman et al. (2020). To prove this fact, we show that $\mathbb{E}_{\theta^*}\left[T_n(\theta^*) \mid \mathcal{D}_1\right] \leq 1$. First, we use only the data in $\mathcal{D}_1$ to fit a parameter $\widehat{\theta}_1$. Let $\mathcal{M}(\theta) = support(P_\theta)$. We see

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[T_n(\theta^*) \mid \mathcal{D}_1\right] &= \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta^*)} \,\middle|\, \mathcal{D}_1\right] = \mathbb{E}_{\theta^*}\left[\prod_{Y_i \in \mathcal{D}_0} \frac{p_{\widehat{\theta}_1}(Y_i)}{p_{\theta^*}(Y_i)} \,\middle|\, \mathcal{D}_1\right] \overset{iid}{=} \prod_{Y_i \in \mathcal{D}_0} \mathbb{E}_{\theta^*}\left[\frac{p_{\widehat{\theta}_1}(Y_i)}{p_{\theta^*}(Y_i)} \,\middle|\, \mathcal{D}_1\right] \\ &= \prod_{i=1}^{|\mathcal{D}_0|}\left\{\int_{\mathcal{M}(\theta^*)} \frac{p_{\widehat{\theta}_1}(y_i)}{p_{\theta^*}(y_i)} p_{\theta^*}(y_i) dy_i\right\} = \prod_{i=1}^{|\mathcal{D}_0|}\left\{\int_{\mathcal{M}(\theta^*)} p_{\widehat{\theta}_1}(y_i) dy_i\right\} \leq \prod_{i=1}^{|\mathcal{D}_0|}\left\{\int_{\mathcal{M}(\widehat{\theta}_1)} p_{\widehat{\theta}_1}(y_i) dy_i\right\} = 1.\end{aligned}$$

Applying Markov's inequality and the above fact,

$$\mathbb{P}_{\theta^*}\left(\theta^* \notin C_n^{\text{split}}(\alpha)\right) = \mathbb{P}_{\theta^*}\left(T_n(\theta^*) \geq 1/\alpha\right) \leq \alpha \mathbb{E}_{\theta^*}[T_n(\theta^*)] = \alpha \mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left[T_n(\theta^*) \mid \mathcal{D}_1\right]\right] \leq \alpha.$$

This shows that $\theta^* \in C_n^{\text{split}}(\alpha)$ with probability at least $1 - \alpha$.

Alternatively, suppose we want to test $H_0 : \theta^* \in \Theta_0$ versus $H_1 : \theta^* \in \Theta \setminus \Theta_0$. We see that rejecting H_0 when $\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset$ provides a valid level α hypothesis test. Under H_0,

$$\mathbb{P}_{\theta^*}\left\{\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset\right\} \leq \mathbb{P}_{\theta^*}\left\{\theta^* \notin \Theta_0 \cap C_n^{\text{split}}(\alpha)\right\} = \mathbb{P}_{\theta^*}\left\{\theta^* \notin C_n^{\text{split}}(\alpha)\right\} \leq \alpha.$$

We now show that (1) rejecting H_0 when $\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset$ is equivalent to (2) rejecting H_0 when $T_n(\widehat{\theta}_0) \geq 1/\alpha$.

($1 \Rightarrow 2$) Suppose $\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset$. Then for any $\theta \in \Theta_0$, $T_n(\theta) \geq 1/\alpha$. Since $\widehat{\theta}_0 \in \Theta_0$, that means $T_n(\widehat{\theta}_0) \geq 1/\alpha$.

($2 \Rightarrow 1$) Suppose $T_n(\widehat{\theta}_0) = \mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\widehat{\theta}_0) \geq 1/\alpha$. Then for any $\theta \in \Theta_0$, $\mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\theta) \geq \mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\widehat{\theta}_0) \geq 1/\alpha$. So $\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset$. □

Theorem 1.0.2. *When $f^* \in \mathcal{F}$, $C_n^{split}(\alpha) = \{f \in \mathcal{F} : T_n(f) < 1/\alpha\}$ is a valid $100(1-\alpha)\%$ confidence set for f^*. A valid level α hypothesis test of $H_0 : f^* \in \mathcal{F}$ versus $H_1 : f^* \notin \mathcal{F}$ rejects H_0 if $\mathcal{F} \cap C_n^{split}(\alpha) = \emptyset$. This hypothesis test is equivalent to rejecting H_0 if $T_n(\widehat{f}_0) \geq 1/\alpha$, where $\widehat{f}_0 = \arg\max_{f \in \mathcal{F}} \mathcal{L}_0(f)$ is the null MLE.*

Proof. This result is also due to Wasserman et al. (2020). The proof is similar to the proof of Theorem 1.0.1. First, we use only the data in $\mathcal{D}_1$ to fit a density $\widehat{f}_1$. Let $\mathcal{M}^*$ be the support of the distribution P^* with density f^*, and let $\widehat{\mathcal{M}_1}$ be the support of the distribution with density $\widehat{f}_1$. We see

$$\begin{aligned}\mathbb{E}_{P^*}\left[T_n(f^*) \mid \mathcal{D}_1\right] &= \mathbb{E}_{P^*}\left[\left.\frac{\mathcal{L}_0(\widehat{f}_1)}{\mathcal{L}_0(f^*)}\right| \mathcal{D}_1\right] = \mathbb{E}_{P^*}\left[\left.\prod_{Y_i \in \mathcal{D}_0} \frac{\widehat{f}_1(Y_i)}{f^*(Y_i)}\right| \mathcal{D}_1\right] \overset{iid}{=} \prod_{Y_i \in \mathcal{D}_0} \mathbb{E}_{P^*}\left[\left.\frac{\widehat{f}_1(Y_i)}{f^*(Y_i)}\right| \mathcal{D}_1\right] \\ &= \prod_{i=1}^{|\mathcal{D}_0|}\left\{\int_{\mathcal{M}^*} \frac{\widehat{f}_1(y_i)}{f^*(y_i)} f^*(y_i) dy_i\right\} = \prod_{i=1}^{|\mathcal{D}_0|}\left\{\int_{\mathcal{M}^*} \widehat{f}_1(y_i) dy_i\right\} \leq \prod_{i=1}^{|\mathcal{D}_0|}\left\{\int_{\widehat{\mathcal{M}_1}} \widehat{f}_1(y_i) dy_i\right\} = 1.\end{aligned}$$

Applying Markov's inequality and the above fact,

$$\mathbb{P}_{P^*}\left(f^* \notin C_n^{\text{split}}(\alpha)\right) = \mathbb{P}_{P^*}\left(T_n(f^*) \geq 1/\alpha\right) \leq \alpha \mathbb{E}_{P^*}[T_n(f^*)] = \alpha \mathbb{E}_{P^*}\left[\mathbb{E}_{P^*}\left[T_n(f^*) \mid \mathcal{D}_1\right]\right] \leq \alpha.$$

This shows that under H_0, $f^* \in C_n^{\text{split}}(\alpha)$ with probability at least $1 - \alpha$.

Alternatively, suppose we want to test $H_0 : f^* \in \mathcal{F}$ versus $H_1 : f^* \notin \mathcal{F}$. We see that rejecting H_0 when $\mathcal{F} \cap C_n^{\text{split}}(\alpha) = \emptyset$ provides a valid level α hypothesis test. Under H_0,

$$\mathbb{P}_{P^*}\left\{\mathcal{F} \cap C_n^{\text{split}}(\alpha) = \emptyset\right\} \leq \mathbb{P}_{P^*}\left\{f^* \notin \mathcal{F} \cap C_n^{\text{split}}(\alpha)\right\} = \mathbb{P}_{P^*}\left\{f^* \notin C_n^{\text{split}}(\alpha)\right\} \leq \alpha.$$

We now show that (1) rejecting H_0 when $\mathcal{F} \cap C_n^{\text{split}}(\alpha) = \emptyset$ is equivalent to (2) rejecting H_0 when $T_n(\widehat{f}_0) \geq 1/\alpha$.

($1 \Rightarrow 2$) Suppose $\mathcal{F} \cap C_n^{\text{split}}(\alpha) = \emptyset$. Then for any $p \in \mathcal{F}$, $T_n(p) \geq 1/\alpha$. Since $\widehat{f}_0 \in \mathcal{F}$, that means $T_n(\widehat{f}_0) \geq 1/\alpha$.

($2 \Rightarrow 1$) Suppose $T_n(\widehat{f}_0) = \mathcal{L}_0(\widehat{f}_1)/\mathcal{L}_0(\widehat{f}_0) \geq 1/\alpha$. Then for any $p \in \mathcal{F}$, $\mathcal{L}_0(\widehat{f}_1)/\mathcal{L}_0(p) \geq \mathcal{L}_0(\widehat{f}_1)/\mathcal{L}_0(\widehat{f}_0) \geq 1/\alpha$. So $\mathcal{F} \cap C_n^{\text{split}}(\alpha) = \emptyset$. □

Appendix B

Proofs and Additional Explorations from Chapter 2

B.1 Proofs of Theorems

Before proving Theorem 2.2.1, we establish Lemma B.0.1 and Lemma B.0.2. We draw heavily on finite population central limit theorem results from Hájek (1960) and Li and Ding (2017). Lemma B.0.1 combines key results from these two papers and adapts them to our setting.

Lemma B.0.1. *Let $(\mathcal{D}_n)_{n \in 2\mathbb{N}}$ be a sequence of datasets, where $\mathcal{D}_n = \{Y_{n1}, \ldots, Y_{nn}\}$ and each Y_{ni} is an independent observation from $N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$. Define $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_{ni}$ and $\overline{Y}_{0,n} = \frac{2}{n}\sum_{Y_{ni} \in \mathcal{D}_{0,n}} Y_{ni}$. As $n \to \infty$, $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)$ converges in distribution to $N(0, I_d)$ with probability 1.*

Proof. We show a highlight of the proof of Lemma B.0.1, in five steps.

Step 1 (Hájek, 1960): Show that simple random sampling and Poisson sampling approaches produce the same limiting distributions.

In the notation of Hájek (1960), suppose we have an infinite sequence of simple random sample experiments indexed by ν. Experiment ν draws a simple random sample of size n_ν from a population of size N_ν given by $\{Y_{\nu 1}, \ldots, Y_{\nu N_\nu}\}$. We assume that $n_\nu \to \infty$ and $N_\nu - n_\nu \to \infty$. In the simple random sampling set-up, a subset s_k of indices $\{1, \ldots, N_\nu\}$ is chosen with probability

$$P(s_k) = \begin{cases} \binom{N_\nu}{n_\nu}^{-1} & : \quad |s_k| = n_\nu \\ 0 & : \quad \text{else.} \end{cases}$$

In contrast, in a Poisson sampling approach with mean sample size n_ν, a subset s_k is chosen with probability

$$P(s_k) = \left(\frac{n_\nu}{N_\nu}\right)^k \left(1 - \frac{n_\nu}{N_\nu}\right)^{N_\nu - k}.$$

We say that each experiment produces a simple random sample (SRS) s_n and a Poisson sample s_k such that $s_n \subseteq s_k$ or $s_k \subseteq s_n$. To construct these samples, we take two steps:

(i) Draw $k \sim \text{Binom}(N_\nu, n_\nu/N_\nu)$.

(ii) If $k = n$, choose SRS s_n, and set $s_k = s_n$.
If $k > n$, choose SRS s_k, and then let s_n be an SRS of size n from s_k.
If $k < n$, choose SRS s_n, and then let s_k be an SRS of size k from s_n.

Using the two samples, we define two random variables:

$$\eta_\nu = \sum_{i \in s_n} (Y_{\nu i} - \overline{Y}_\nu) \qquad \text{and} \qquad \eta_\nu^* = \sum_{i \in s_k} (Y_{\nu i} - \overline{Y}_\nu).$$

We can show that the variance of η_ν^* is

$$D\eta_\nu^* = \text{var}(\eta_\nu^*) = \frac{n_\nu}{N_\nu}\left(1 - \frac{n_\nu}{N_\nu}\right) \sum_{i=1}^{N_\nu} (Y_{\nu i} - \overline{Y}_\nu)^2.$$

Under the assumption that $n_\nu \to \infty$ and $N - n_\nu \to \infty$, we can then show that

$$\lim_{\nu \to \infty} \frac{\mathbb{E}[(\eta_\nu - \eta_\nu^*)^2]}{D\eta_\nu^*} = 0. \tag{B.1}$$

Remark 2.1 of Hájek (1960) states that (B.1) implies that the limiting distributions of $\eta_\nu/\sqrt{D\eta_\nu^*}$ and $\eta_\nu^*/\sqrt{D\eta_\nu^*}$ are the same if they exist, and they exist under the same conditions. To see this, we use Chebyshev's inequality. For $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\eta_\nu}{\sqrt{D\eta_\nu^*}} - \frac{\eta_\nu^*}{\sqrt{D\eta_\nu^*}}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\text{var}\left(\frac{\eta_\nu - \eta_\nu^*}{\sqrt{D\eta_\nu^*}}\right) = \frac{1}{\epsilon^2}\frac{\mathbb{E}[(\eta_\nu - \eta_\nu^*)^2]}{D\eta_\nu^*} \overset{\nu\to\infty}{\to} 0.$$

This means that $\left|\eta_\nu/\sqrt{D\eta_\nu^*} - \eta_\nu^*/\sqrt{D\eta_\nu^*}\right| \overset{p}{\to} 0$. Under this condition, for any distribution W, $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow W$ if and only if $\eta_\nu^*/\sqrt{D\eta_\nu^*} \rightsquigarrow W$.

Since η_ν^* is a sum of independent random variables, it will be easier to work with $\eta_\nu^*/D\eta_\nu^*$ than to work with $\eta_\nu/D\eta_\nu^*$.

Step 2 (Hájek, 1960): Find conditions such that $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$. (We can think of η_ν as $(n/2)(\overline{Y}_{0,n} - \overline{Y}_n)$ and $D\eta_\nu^*$ as $\text{var}(\sum_{i=1}^n B_i(Y_{ni} - \overline{Y}_n))$ for $B_i \overset{iid}{\sim} \text{Bernoulli}(1/2)$.)

Theorem 3.1 in Hájek (1960) is the key result for asymptotic normality. We present an intermediate result from the proof of Theorem 3.1.

Let $\xi_\nu = \sum_{i\in s_{n,\nu}} Y_{\nu,i}$. (So $\eta_\nu = \xi_\nu - n_\nu\overline{Y}_\nu$.) Let $D\xi_\nu$ be the variance of ξ_ν. Let $S_{\nu\tau}$ be the subset of $S_\nu = \{1, \ldots, N_\nu\}$ on which the inequality

$$|Y_{\nu i} - \overline{Y}_\nu| > \tau\sqrt{D\xi_\nu}$$

holds. Suppose that $n_\nu \to \infty$ and $N_\nu - n_\nu \to \infty$. If

$$\lim_{\nu\to\infty} \frac{\sum_{i\in S_{\nu\tau}}(Y_{\nu i} - \overline{Y}_\nu)^2}{\sum_{i\in S_\nu}(Y_{\nu i} - \overline{Y}_\nu)^2} = 0 \quad \textit{for any } \tau > 0, \tag{B.2}$$

then $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$.

We will show that $\eta_\nu^*/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$, and then we can appeal to Step 1's result. η_ν^* is the centered sum of the Poisson sampling terms. We can write η_ν^* as

$$\eta_\nu^* = \sum_{i=1}^{N_\nu} \zeta_{\nu i}, \text{ where } \zeta_{\nu i} = \begin{cases} Y_{\nu i} - \overline{Y}_\nu & \text{with probabilty } n_\nu/N_\nu \\ 0 & \text{with probabilty } 1 - n_\nu/N_\nu. \end{cases}$$

In this setting, Lindeberg's condition for $\eta_\nu^*/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$ is for all $\tau > 0$,

$$\lim_{\nu\to\infty} \frac{1}{D\eta_\nu^*}\sum_{i=1}^{N_\nu} \mathbb{E}\left[(\zeta_{\nu i} - \mathbb{E}[\zeta_{\nu i}])^2 \cdot \mathbb{1}\left(|\zeta_{\nu i} - \mathbb{E}[\zeta_{\nu i}| > \tau\sqrt{D\eta_\nu^*}\right)\right] = 0.$$

We can show that (B.2) implies that the Lindeberg condition is satisfied. Since Step 1 implies that the limiting distribution of $\eta_\nu/\sqrt{D\eta_\nu^*}$ must be the same as the limiting distribution of $\eta_\nu^*/\sqrt{D\eta_\nu^*}$, we conclude that $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$.

Step 3: If $d = 1$, show that $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$ implies $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n) \rightsquigarrow N(0,1)$.

This is mostly a matter of adapting Step 2's result to our setting. When $n_\nu/N_\nu = 1/2$, η_ν is the same random variable as $(n/2)(\overline{Y}_{0,n} - \overline{Y}_n)$. Using the formula for $D\eta_\nu^*$,

$$\frac{\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_{ni} - \overline{Y}_n)^2}} = \frac{(n/2)(\overline{Y}_{0,n} - \overline{Y}_n)}{\sqrt{\frac{1}{4}\sum_{i=1}^n (Y_{ni} - \overline{Y}_n)^2}} \overset{d}{=} \frac{\eta_\nu}{\sqrt{D\eta_\nu^*}} \rightsquigarrow N(0,1).$$

In addition, $\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_{ni} - \overline{Y}_n)^2}/\sqrt{\operatorname{var}(Y_{ni})} \xrightarrow{p} 1$. By Slutsky's Theorem, $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n) \rightsquigarrow N(0,1)$.

Step 4 (Li and Ding, 2017): If $Y_{n1}, \dots, Y_{nn} \sim N(\theta^*, 1)$, show that the condition of Step 2 is satisfied with probability 1.

These results come from page 2 of the appendix of Li and Ding (2017). The authors show that if the Y_{ni}s are iid draws from a superpopulation with $2 + \epsilon$ ($\epsilon > 0$) absolute moments and nonzero variance, then $(1/n)\max_{1 \le i \le n}(Y_{ni} - \overline{Y}_n)^2 \equiv m_n/n \to 0$ with probability 1. Furthermore, they show that $m_n/n \to 0$ implies their condition (A2), which is a rewriting of Hájek (1960)'s condition (B.2).

Since $N(\theta^*, 1)$ satisfies the superpopulation conditions, condition (B.2) is satisfied with probability 1. Then following Steps 2 and 3, $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n) \rightsquigarrow N(0,1)$.

Step 5 (Hájek, 1960): Extend results to $d > 1$.

In d dimensions, suppose $Y_{n1}, \dots, Y_{nn} \sim N(\theta^*, I_d)$. Remark 3.2 of Hájek (1960) notes that we can user the Cramér-Wold device to extend the results to the multivariate case. Let $Z = (Z^{(1)}, \dots, Z^{(d)})$ represent the $N(0, I_d)$ distribution. Then for each component, $Z^{(j)} \sim N(0,1)$. By the Cramér-Wold device, we can say that $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n) \rightsquigarrow Z$ if and only if for any $\lambda \in \mathbb{R}^d$, $\sum_{j=1}^d \lambda^{(j)}\sqrt{n}(\overline{Y}_{0,n}^{(j)} - \overline{Y}_n^{(j)}) \rightsquigarrow \sum_{j=1}^d \lambda^{(j)} Z^{(j)}$.

For any dimension j, we can think of $Y_{n1}^{(j)}, \dots, Y_{nn}^{(j)}$ as draws from a $N(\theta^{*(j)}, 1)$ superpopulation. So the superpopulation conditions from Step 4 are satisfied, which means $\sqrt{n}(\overline{Y}_{0,n}^{(j)} - \overline{Y}_n^{(j)}) \rightsquigarrow Z^{(j)}$. We conclude that $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n) \rightsquigarrow N(0, I_d)$. □

Lemma B.0.2. *Assume $(\mathcal{D}_n)_{n \in 2\mathbb{N}}$ is a sequence of data sets such that $\mathcal{D}_n = \{Y_{n1}, Y_{n2}, \dots, Y_{nn}\}$ with observations $Y_{nj} \overset{iid}{\sim} N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$. Define $\overline{Y}_n = (1/n)\sum_{i=1}^n Y_{ni}$ and $\overline{Y}_{0,n} = (2/n)\sum_{Y_{ni} \in \mathcal{D}_{0,n}} Y_{ni}$. Let $c > 0$, and let (θ_n) be a sequence that satisfies $\|\overline{Y}_n - \theta_n\| \le c/\sqrt{n}$ for all n. Define $X_n \equiv \sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)$. Let Z denote a $N(0, I_d)$ random* *variable. Then*

$$\mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^T X_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] - \mathbb{E}\left[\exp\left(-\frac{3}{4}Z^T Z + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] = o_P(1).$$

Proof. Since (θ_n) is chosen such that $\|\overline{Y}_n - \theta_n\| \leq c/\sqrt{n}$, we can re-write $\theta_n = \overline{Y}_n + (c/\sqrt{n})v_n$, where $v_n \in \mathbb{R}^d$ satisfies $\|v_n\| \leq 1$ for all n.

Define a function f by

$$f(x_n, v_n) \equiv \exp\left(-\frac{3}{4}x_n^T x_n - \frac{c}{2}x_n^T v_n\right).$$

f is clearly a continuous function. We can also show that f is bounded. Define

$$g(x_n, v_n) \equiv -\frac{3}{4}x_n^T x_n - \frac{c}{2}x_n^T v_n$$

so that $f(x_n, v_n) = \exp(g(x_n, v_n))$. We can see that

$$\frac{\partial}{\partial x_n}g(x_n, v_n) = -\frac{3}{2}x_n - \frac{c}{2}v_n \overset{set}{=} \vec{0}$$

is solved by $x_n = -(c/3)v_n$. Since $g(x_n, v_n)$ is concave in x_n, $g(x_n, v_n)$ is maximized at $x_n = -(c/3)v_n$ for any v_n. Since $f(x_n, v_n) = \exp(g(x_n, v_n))$, $f(x_n, v_n)$ is also maximized at this value of x_n for any v_n. Under the assumption that $\|v_n\| \leq 1$, we see

$$\begin{aligned} f(x_n, v_n) &\leq \exp\left(-\frac{3}{4}\left(-\frac{c}{3}\right)^2 v_n^T v_n - \frac{c}{2}\left(-\frac{c}{3}\right)v_n^T v_n\right) \\ &= \exp\left(-\frac{c^2}{12}\|v_n\|^2 + \frac{c^2}{6}\|v_n\|^2\right) \\ &\leq \exp\left(\frac{c^2}{12}\right). \end{aligned}$$

Thus, $f(x_n, v_n)$ is a continuous and bounded function.

The claim of Lemma B.0.2 is equivalent to $\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n] = o_P(1)$. The Portmanteau Theorem provides several equivalent definitions of convergence in distribution, including that $X_n \rightsquigarrow Z$ if and only if $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(Z)]$ for every continuous, bounded function h. We prove the result on $f(X_n, v_n)$ by modifying the Van der Vaart (2000), Chapter 2, proof of this Portmanteau Theorem result.

Let $\gamma > 0$. Fix $\epsilon > 0$ such that

$$\epsilon < \gamma \,/\, (3 + 3\exp(c^2/12)). \tag{B.3}$$

Choose a large enough compact rectangle I such that

$$\mathbb{P}(Z \notin I) < \epsilon. \tag{B.4}$$

Let $\mathcal{B}_1(0)$ be the d-dimensional ball of radius 1 centered at 0. By construction, each $v_n \in \mathcal{B}_1(0)$. Since f is continuous and $I \times \mathcal{B}_1(0)$ is compact, $f(x_n, v_n)$ is uniformly continuous on $I \times \mathcal{B}_1(0)$. We can thus partition $I \times \mathcal{B}_1(0)$ into J compact regions $I_j \times V_j$ where $I \times \mathcal{B}_1(0) = \cup_{j=1}^{J}(I_j \times V_j)$ such that for any j and for any $(x_{n1}, v_{n1}), (x_{n2}, v_{n2}) \in I_j \times V_j$, $|f(x_{n1}, v_{n1}) - f(x_{n2}, v_{n2})| < \epsilon$. (For instance the I_j regions may be rectangles and the V_j regions may be rectangles truncated at the boundaries of $\mathcal{B}_1(0)$. These rectangular regions may be appropriately sized such that within a region $I_j \times V_j$, $d((x_{n1}, v_{n1}), (x_{n2}, v_{n2}))$ is small enough that $|f(x_{n1}, v_{n1}) - f(x_{n2}, v_{n2})| < \epsilon$.)

Select a point (x'_j, v'_j) from each $I_j \times V_j$. Define

$$f_\epsilon(x, v) = \sum_{j=1}^{J} f(x'_j, v'_j)\mathbb{1}((x, v) \in I_j \times V_j).$$

For a given sample $\mathcal{D}_n$, we note that there are $\binom{n}{n/2}$ possible values of X_n, since there are $\binom{n}{n/2}$ possible values of $\overline{Y}_{0,n}$. We denote the sum over all possible values of X_n as $\sum_{X_n}$.

Note that

$$
\begin{aligned}
&|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n]| \\
&= \left| \binom{n}{n/2}^{-1} \sum_{X_n} f(X_n, v_n) - \binom{n}{n/2}^{-1} \sum_{X_n} f_\epsilon(X_n, v_n) \right| \\
&= \left| \binom{n}{n/2}^{-1} \sum_{X_n} \left[(f(X_n, v_n) - f_\epsilon(X_n, v_n)) \mathbb{1}(X_n \in I) + (f(X_n, v_n) - f_\epsilon(X_n, v_n)) \mathbb{1}(X_n \notin I) \right] \right| \\
&\leq \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)| \mathbb{1}(X_n \in I) + \\
&\quad \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)| \mathbb{1}(X_n \notin I) \\
&= \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)| \mathbb{1}(X_n \in I, v_n \in \mathcal{B}_1(0)) + \\
&\quad \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)| \mathbb{1}(X_n \notin I) \\
&< \binom{n}{n/2}^{-1} \sum_{X_n} \epsilon + \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n)| \mathbb{1}(X_n \notin I) \\
&\leq \epsilon + \exp\left(c^2/12\right) \mathbb{P}(X_n \notin I \mid \mathcal{D}_n). \qquad \text{(B.5)}
\end{aligned}
$$

Similarly, we show that

$$
\begin{aligned}
&\Big|\mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n]\Big| \\
&= \Big|\mathbb{E}\big[(f(Z, v_n) - f_\epsilon(Z, v_n)) \mathbb{1}(Z \in I) + (f(Z, v_n) - f_\epsilon(Z, v_n)) \mathbb{1}(Z \notin I)\big]\Big| \\
&\leq \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big| \mathbb{1}(Z \in I) \mid \mathcal{D}_n\right] + \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big| \mathbb{1}(Z \notin I) \mid \mathcal{D}_n\right] \\
&= \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big| \mathbb{1}(Z \in I, v_n \in \mathcal{B}_1(0)) \mid \mathcal{D}_n\right] + \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big| \mathbb{1}(Z \notin I) \mid \mathcal{D}_n\right] \\
&< \epsilon + \exp(c^2/12) \mathbb{P}(Z \notin I \mid \mathcal{D}_n) \\
&= \epsilon + \exp(c^2/12) \mathbb{P}(Z \notin I) \\
&< \epsilon + \epsilon \exp(c^2/12). \qquad \text{(B.6)}
\end{aligned}
$$

In addition, we see that

$$
\begin{aligned}
&\left|\mathbb{E}\left[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n\right] - \mathbb{E}\left[f_\epsilon(Z, v_n) \mid \mathcal{D}_n\right]\right| \\
&= \left|\binom{n}{n/2}^{-1} \sum_{X_n} f_\epsilon(X_n, v_n) - \mathbb{E}[f_\epsilon(Z, v_n)]\right| \\
&= \left|\binom{n}{n/2}^{-1} \sum_{X_n} \sum_{j=1}^{J} f(x'_j, v'_j)\mathbb{1}((X_n, v_n) \in I_j \times V_j) - \sum_{j=1}^{J} f(x'_j, v'_j)\mathbb{P}(Z \in I_j)\mathbb{1}(v_n \in V_j)\right| \\
&\leq \sum_{j=1}^{J} \left|\binom{n}{n/2}^{-1} \sum_{X_n} f(x'_j, v'_j)\mathbb{1}(X_n \in I_j)\mathbb{1}(v_n \in V_j) - f(x'_j, v'_j)\mathbb{P}(Z \in I_j)\mathbb{1}(v_n \in V_j)\right| \\
&\leq \sum_{j=1}^{J} \left|\binom{n}{n/2}^{-1} \sum_{X_n} f(x'_j, v'_j)\mathbb{1}(X_n \in I_j) - f(x'_j, v'_j)\mathbb{P}(Z \in I_j)\right| \\
&= \sum_{j=1}^{J} \left|f(x'_j, v'_j)\left[\binom{n}{n/2}^{-1} \sum_{X_n} \mathbb{1}(X_n \in I_j) - \mathbb{P}(Z \in I_j)\right]\right| \\
&\leq \sum_{j=1}^{J} \left|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\right| \times \left|f(x'_j, v'_j)\right|. \qquad \text{(B.7)}
\end{aligned}
$$

For the sequence of datasets $(\mathcal{D}_n)_{n \in 2\mathbb{N}}$, Lemma B.0.1 establishes that $X_n \rightsquigarrow N(0, I_d)$ with probability 1. This tells us that with probability 1 over the randomness in sequences $(\mathcal{D}_n)_{n \in 2\mathbb{N}}$, $\lim_{n\to\infty} \mathbb{P}(X_n \in I \mid \mathcal{D}_n) = \mathbb{P}(Z \in I)$. Since almost sure convergence implies convergence in probability, for any $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|\mathbb{P}(X_n \in I \mid \mathcal{D}_n) - \mathbb{P}(Z \in I)| > \delta\big) = 0 \qquad \text{(B.8)}$$

$$\text{and } \lim_{n\to\infty} \mathbb{P}\big(|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)| > \delta\big) = 0 \text{ for } 1 \leq j \leq J. \qquad \text{(B.9)}$$

The outer probability is over the randomness in the sequences $(\mathcal{D}_n)_{n \in 2\mathbb{N}}$.

Now we see

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma)\\
&\le \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n]\big| + \\
&\qquad\qquad \big|\mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n]\big| + \\
&\qquad\qquad \big|\mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma\big)\\
&\le \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n]\big| > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma/3\big)\\
&\le \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \exp(c^2/12)\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) > \gamma/3\big) + \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \epsilon\exp(c^2/12) > \gamma/3) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\left(\sum_{j=1}^{J} \big|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\big| \times |f(x'_j, v'_j)| > \gamma/3\right) \text{ by (B.5), (B.6), and (B.7)}\\
&= \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \exp(c^2/12)\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\left(\sum_{j=1}^{J} \big|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\big| \times |f(x'_j, v'_j)| > \gamma/3\right) \text{ by (B.3)}\\
&\le \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \exp(c^2/12)\left(\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) - \mathbb{P}(Z \notin I)\right) > \gamma/3 - \exp(c^2/12)\mathbb{P}(Z \notin I)\big) + \\
&\quad \lim_{n\to\infty} \sum_{j=1}^{J} \mathbb{P}\left(\big|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\big| > (\gamma/3)|f(x'_j, v_j)|^{-1}\right)\\
&\le \lim_{n\to\infty} \mathbb{P}\left(\epsilon + \exp(c^2/12)(\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) - \mathbb{P}(Z \notin I)) > \gamma/3 - \epsilon\exp(c^2/12)\right) \text{ by (B.4) and (B.9)}\\
&= \lim_{n\to\infty} \mathbb{P}\left(\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) - \mathbb{P}(Z \notin I) > \frac{\gamma - 3\epsilon - 3\epsilon\exp(c^2/12)}{3\exp(c^2/12)}\right)\\
&= 0 \text{ by (B.3) and (B.8).}
\end{aligned}
$$

We have shown that for arbitrary $\gamma > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma) = 0.$$

We conclude that $\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n] = o_P(1)$. □

Theorem 2.2.1. *Assume we have a sequence of datasets $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$, where $\mathcal{D}_n = \{Y_{n1}, \ldots, Y_{nn}\}$ and each Y_{ni} is an independent observation from $N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$, and let $\mathcal{D}_{1,n} = \mathcal{D}_n \backslash \mathcal{D}_{0,n}$. Define $\overline{Y}_n = (1/n)\sum_{i=1}^n Y_{ni}$, $\overline{Y}_{0,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{0,n}} Y_{ni}$, and $\overline{Y}_{1,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{1,n}} Y_{ni}$. Let $c > 0$, and let (θ_n) be a sequence that satisfies $\|\overline{Y}_n - \theta_n\| \leq c/\sqrt{n}$ for all n. Then*

$$\mathbb{E}\{T_n(\theta_n) \mid \mathcal{D}_n\} \;/\; \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} = 1 + o_P(1). \tag{2.3}$$

Proof. Define $X_n \equiv \sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)$ and let $Z \sim N(0, I_d)$. In addition, define $\mu_n \equiv (\sqrt{n}/5)(\overline{Y}_n - \theta_n)$ and $\Omega \equiv (2/5)I_d$. Then

$$\begin{aligned}
&\mathbb{E}[T_n(\theta_n) \mid \mathcal{D}_n] \;/\; \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} \\
&= \mathbb{E}\left[\exp\left(-\frac{n}{4}\|\overline{Y}_{0,n} - \overline{Y}_{1,n}\|^2 + \frac{n}{4}\|\overline{Y}_{0,n} - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \;/\; \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} \\
&= \mathbb{E}\left[\exp\left(-\frac{n}{4}\|2\overline{Y}_{0,n} - 2\overline{Y}_n\|^2 + \frac{n}{4}\|\overline{Y}_{0,n} - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \exp\left(-\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-n\|\overline{Y}_{0,n} - \overline{Y}_n\|^2 + \frac{n}{4}\|\overline{Y}_{0,n} - \overline{Y}_n + \overline{Y}_n - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \exp\left(-\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-\frac{3n}{4}\|\overline{Y}_{0,n} - \overline{Y}_n\|^2 + \frac{n}{2}(\overline{Y}_{0,n} - \overline{Y}_n)^T(\overline{Y}_n - \theta_n) + \frac{n}{4}\|\overline{Y}_n - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \times \\
&\qquad \exp\left(-\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^T X_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] \exp\left(-\frac{n}{20}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^T X_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] / \mathbb{E}\left[\exp\left(-\frac{3}{4}Z^T Z + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right]
\end{aligned} \tag{B.10}$$

$$= 1 + o_P(1). \tag{B.11}$$

Step (B.10) holds because

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ+\frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n-\theta_n\right)\right)\mid\mathcal{D}_n\right]\\
&=\int_{\mathbb{R}^d}\left[\frac{1}{(2\pi)^{d/2}|I_d|^{1/2}}\exp\left(-\frac{1}{2}z^Tz\right)\exp\left(-\frac{3}{4}z^Tz+\frac{\sqrt{n}}{2}z^T\left(\overline{Y}_n-\theta_n\right)\right)\right]dz\\
&=\int_{\mathbb{R}^d}\left[\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{5}{4}z^Tz+\frac{\sqrt{n}}{2}z^T\left(\overline{Y}_n-\theta_n\right)\right)\right]dz\\
&=|\Omega|^{1/2}\int_{\mathbb{R}^d}\left[\frac{1}{(2\pi)^{d/2}|\Omega|^{1/2}}\exp\left(-\frac{1}{2}(z-\mu_n)^T\Omega^{-1}(z-\mu_n)+\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2\right)\right]dz \qquad \text{(B.12)}\\
&=\exp\left(\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2\right)|\Omega|^{1/2}\\
&=\exp\left(\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}.
\end{aligned}
$$

Step (B.12) uses the following equality:

$$
\begin{aligned}
&-\frac{5}{4}z^Tz+\frac{\sqrt{n}}{2}z^T(\overline{Y}_n-\theta_n)\\
&\quad=-\frac{5}{4}\left[z^Tz-\frac{2\sqrt{n}}{5}z^T(\overline{Y}_n-\theta_n)+\frac{n}{25}(\overline{Y}_n-\theta_n)^T(\overline{Y}_n-\theta_n)-\frac{n}{25}(\overline{Y}_n-\theta_n)^T(\overline{Y}_n-\theta_n)\right]\\
&\quad=-\frac{5}{4}\left(z-\frac{\sqrt{n}}{5}(\overline{Y}_n-\theta_n)\right)^T\left(z-\frac{\sqrt{n}}{5}(\overline{Y}_n-\theta_n)\right)+\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2\\
&\quad=-\frac{1}{2}\left(z-\frac{\sqrt{n}}{5}(\overline{Y}_n-\theta_n)\right)^T\left(\frac{5}{2}I_d\right)\left(z-\frac{\sqrt{n}}{5}(\overline{Y}_n-\theta_n)\right)+\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2\\
&\quad=-\frac{1}{2}(z-\mu_n)^T\Omega^{-1}(z-\mu_n)+\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2.
\end{aligned}
$$

To justify step (B.11), note that $\mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ+\frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n-\theta_n\right)\right)\mid\mathcal{D}_n\right]$, which equals $\exp\left(\frac{n}{20}\|\overline{Y}_n-\theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}$, is bounded between $(2/5)^{d/2}$ and $\exp(c^2/20)(2/5)^{d/2}$ under the assumption that $\|\overline{Y}_n-\theta_n\|\leq c/\sqrt{n}$. By Lemma B.0.2,

$$
\mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^TX_n+\frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n-\theta_n\right)\right)\mid\mathcal{D}_n\right]-\mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ+\frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n-\theta_n\right)\right)\mid\mathcal{D}_n\right]=o_P(1).
$$

Combining these two facts, we conclude that

$$
\mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^TX_n+\frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n-\theta_n\right)\right)\mid\mathcal{D}_n\right]/\mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ+\frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n-\theta_n\right)\right)\mid\mathcal{D}_n\right]=1+o_P(1).
$$

□

Theorem 2.3.1. *Let* $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$*. The splitting proportion that minimizes* $\mathbb{E}[r^2\{C_n^{split}(\alpha)\}]$ *is*

$$p_0^* = 1 - \frac{\sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right)} - d}{2\log\left(\frac{1}{\alpha}\right)}. \tag{2.6}$$

Proof. Recall that p_0 represents the proportion of observations that we place in $\mathcal{D}_0$.

We know that

$$\begin{aligned} \overline{Y}_0 &\sim N\left(\theta^*,\ Var = \frac{1}{np_0} I_d\right) \\ \overline{Y}_1 &\sim N\left(\theta^*,\ Var = \frac{1}{n(1-p_0)} I_d\right). \end{aligned}$$

Since all observations in $\mathcal{D}_0$ and $\mathcal{D}_1$ are mutually independent, this implies

$$\overline{Y}_0 - \overline{Y}_1 \sim N\left(0, \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right) I_d\right) \tag{B.13}$$

and, hence,

$$\left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right)^{-1/2} \left(\overline{Y}_0 - \overline{Y}_1\right) \sim N\left(0, I_d\right).$$

We now see

$$\begin{aligned} \|\overline{Y}_0 - \overline{Y}_1\|^2 &= \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right) \left\| \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right)^{-1/2} (\overline{Y}_0 - \overline{Y}_1) \right\|^2 \\ &\overset{d}{=} \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right) \chi_d^2. \end{aligned} \tag{B.14}$$

When $p_0 = 1/2$, this expression is $(4/n)\chi_d^2$, as shown in the derivation of equation 2.7.

Setting $\widehat{\theta}_1 = \overline{Y}_1$, at $\theta \in \mathbb{R}^d$ we construct the test statistic. The derivation of equation 2.2 justifies the equality of the first and second lines.

$$\begin{aligned}
T_n(\theta) &= \frac{\prod_{Y_i \in \mathcal{D}_0} \exp\left(-\frac{1}{2}(Y_i - \widehat{\theta}_1)^T (Y_i - \widehat{\theta}_1)\right)}{\prod_{Y_i \in \mathcal{D}_0} \exp\left(-\frac{1}{2}(Y_i - \theta)^T (Y_i - \theta)\right)} \\
&= \exp\left(\sum_{Y_i \in \mathcal{D}_0} \left(-\frac{1}{2}(\overline{Y}_0 - \overline{Y}_1)^T (\overline{Y}_0 - \overline{Y}_1) + \frac{1}{2}(\overline{Y}_0 - \theta)^T (\overline{Y}_0 - \theta)\right)\right) \\
&= \exp\left(-\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2\right).
\end{aligned}$$

Using a split proportion of p_0, the split LRT confidence set is now

$$\begin{aligned}
C_n^{\text{split}} &= \left\{\theta \in \Theta : \exp\left(-\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2\right) \leq \frac{1}{\alpha}\right\} \\
&= \left\{\theta \in \Theta : -\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2 \leq \log\left(\frac{1}{\alpha}\right)\right\} \\
&= \left\{\theta \in \Theta : \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2 \leq \log\left(\frac{1}{\alpha}\right) + \frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2\right\} \\
&= \left\{\theta \in \Theta : \|\overline{Y}_0 - \theta\|^2 \leq \frac{2}{np_0}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}.
\end{aligned}$$

The squared radius is thus $R^2(C_n^{\text{split}}) = (2/(np_0))\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2$. By (B.14), the expected squared radius at a given value of p_0 is

$$r(p_0) = \frac{2}{np_0}\log\left(\frac{1}{\alpha}\right) + \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right)d.$$

We can now minimize this function:

$$\begin{aligned}
0 &\overset{set}{=} \frac{\partial}{\partial p_0} r(p_0) = \frac{-2}{np_0^2}\log\left(\frac{1}{\alpha}\right) - \frac{d}{np_0^2} + \frac{d}{n(1-p_0)^2} \\
&\Updownarrow \\
0 &= -2(1-p_0)^2\log\left(\frac{1}{\alpha}\right) - d(1-p_0)^2 + dp_0^2 \\
&= -2(1-2p_0+p_0^2)\log\left(\frac{1}{\alpha}\right) - d(1-2p_0+p_0^2) + dp_0^2 \\
&= -2\log\left(\frac{1}{\alpha}\right) + 4p_0\log\left(\frac{1}{\alpha}\right) - 2p_0^2\log\left(\frac{1}{\alpha}\right) - d + 2dp_0 - dp_0^2 + dp_0^2 \\
&= p_0^2\left(-2\log\left(\frac{1}{\alpha}\right)\right) + p_0\left(4\log\left(\frac{1}{\alpha}\right) + 2d\right) + \left(-2\log\left(\frac{1}{\alpha}\right) - d\right).
\end{aligned}$$

This is now a quadratic expression in p_0. Thus, this formula is solved by

$$\begin{aligned}
p_0 &= \frac{-4\log\left(\frac{1}{\alpha}\right) - 2d \pm \sqrt{\left(4\log\left(\frac{1}{\alpha}\right)+2d\right)^2 - 4\left(-2\log\left(\frac{1}{\alpha}\right)\right)\left(-2\log\left(\frac{1}{\alpha}\right)-d\right)}}{2\left(-2\log\left(\frac{1}{\alpha}\right)\right)} \\
&= \frac{4\log\left(\frac{1}{\alpha}\right) + 2d \pm \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)}}{4\log\left(\frac{1}{\alpha}\right)} \\
&= 1 + \frac{d \pm \sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right)}}{2\log\left(\frac{1}{\alpha}\right)}.
\end{aligned}$$

We now consider the ± choice. In the + direction, we have

$$p_0 = 1 + \frac{d + \sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right)}}{\log\left(\frac{1}{\alpha}\right)} > 1.$$

However, in the − direction, we can show that $p_0 \in \left(\frac{1}{2}, 1\right)$. We note that

$$\begin{aligned}
d &< \sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right)} < \sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right) + \left(\log\left(\frac{1}{\alpha}\right)\right)^2} \\
&= \sqrt{\left(d + \log\left(\frac{1}{\alpha}\right)\right)^2} = d + \log\left(\frac{1}{\alpha}\right).
\end{aligned}$$

So

$$p_0 = 1 + \frac{d - \sqrt{d^2 + 2d \log\left(\frac{1}{\alpha}\right)}}{2 \log\left(\frac{1}{\alpha}\right)} < 1 + \frac{d - d}{2\log\left(\frac{1}{\alpha}\right)} = 1$$

and

$$p_0 = 1 + \frac{d - \sqrt{d^2 + 2d \log\left(\frac{1}{\alpha}\right)}}{2 \log\left(\frac{1}{\alpha}\right)} > 1 + \frac{d - d - \log\left(\frac{1}{\alpha}\right)}{2\log\left(\frac{1}{\alpha}\right)} = 1 - \frac{1}{2} = \frac{1}{2}.$$

This means that

$$p_0^* = 1 - \frac{\sqrt{d^2 + 2d \log\left(\frac{1}{\alpha}\right)} - d}{2 \log\left(\frac{1}{\alpha}\right)}$$

optimizes $r(p_0)$, and $p_0^* \in \left(\frac{1}{2}, 1\right)$. Furthermore, this optimum must be a minimum, since for any $p_0 \in (0, 1)$,

$$\frac{\partial^2}{\partial p_0^2} r(p_0) = \frac{4}{np_0^3} \log\left(\frac{1}{\alpha}\right) + \frac{2d}{np_0^3} + \frac{2d}{n(1-p_0)^3} > 0.$$

We can use L'Hôpital's Rule to show that $p_0^* \to \frac{1}{2}$ as $d \to \infty$:

$$\begin{aligned}
\lim_{d\to\infty} p_0^* &= 1 - \lim_{d\to\infty} \frac{\sqrt{d^2 + 2d\log\left(\frac{1}{\alpha}\right)} - d}{2\log\left(\frac{1}{\alpha}\right)} \\
&= 1 - \lim_{d\to\infty} \frac{\sqrt{1 + (2/d)\log(1/\alpha)} - 1}{(2/d)\log(1/\alpha)} \\
&= 1 - \lim_{d\to\infty} \frac{\frac{1}{2}\left(1 + (2/d)\log(1/\alpha)\right)^{-1/2}(-2/d^2)\log(1/\alpha)}{(-2/d^2)\log(1/\alpha)} \\
&= 1 - \frac{1}{2}\lim_{d\to\infty}\left(1 + (2/d)\log(1/\alpha)\right)^{-1/2} \\
&= \frac{1}{2}.
\end{aligned}$$

We conclude that as $d \to \infty$ for fixed α, the optimal choice of $p_0^* \to 0.5$. □

Theorem 2.3.2. *Suppose $Y_1, \ldots, Y_n$ are iid observations from $N(\theta^*, I_d)$. Split the sample such that $\mathcal{D}_0$ and $\mathcal{D}_1$ each contain $\frac{n}{2}$ observations. Use $\mathcal{D}_0$ and $\mathcal{D}_1$ to define the split and cross-fit sets. Then Volume$\{C_n^{CF}(\alpha)\} \leq$ Volume$\{C_n^{split}(\alpha)\}$. Equality holds only when $\overline{Y}_0 = \overline{Y}_1$.*

Proof. Let $\theta \in C_n^{\mathrm{CF}}(\alpha)$. Then

$$\begin{aligned}
&\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y} - \theta\|^2\right) \\
&\quad = \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\left\|\frac{1}{2}(\overline{Y}_0 - \theta) + \frac{1}{2}(\overline{Y}_1 - \theta)\right\|^2\right) \\
&\quad \leq \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{8}\|\overline{Y}_0 - \theta\|^2 + \frac{n}{8}\|\overline{Y}_1 - \theta\|^2\right) && \text{(B.15)} \\
&\quad = \exp\left(-\frac{n}{8}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{8}\|\overline{Y}_0 - \theta\|^2 - \frac{n}{8}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{8}\|\overline{Y}_1 - \theta\|^2\right) \\
&\quad \leq \frac{1}{2}\left[\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \theta\|^2\right) + \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_1 - \theta\|^2\right)\right] && \text{(B.16)} \\
&\quad < \frac{1}{\alpha}.
\end{aligned}$$

Line (B.15) holds because $\|\cdot\|^2$ is convex. Line (B.16) holds because $\exp(\cdot)$ is convex. Thus, $C_n^{\mathrm{CF}}(\alpha) \subseteq \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 < \frac{4}{n}\log(\frac{1}{\alpha}) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$, which has the same volume as $C_n^{\mathrm{split}}(\alpha) = \left\{\theta \in \Theta : \|\overline{Y}_0 - \theta\|^2 < \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$. Hence, $Vol\left(C_n^{\mathrm{CF}}(\alpha)\right) \leq Vol\left(C_n^{\mathrm{split}}(\alpha)\right)$.

Furthermore, since $\|\cdot\|^2$ and $\exp(\cdot)$ are strictly convex, equality holds in (B.15) and (B.16) only when $\overline{Y}_0 = \overline{Y}_1$. If $\overline{Y}_0 = \overline{Y}_1$, then $C_n^{\mathrm{CF}}(\alpha) = \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 \leq \frac{4}{n}\log(\frac{1}{\alpha}) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$, which means $Vol\left(C_n^{\mathrm{CF}}(\alpha)\right) = Vol\left(C_n^{\mathrm{split}}(\alpha)\right)$. □

Theorem 2.3.3. *Assume $c_{\alpha,d} + \log(\alpha) > d - 2$. Let $f_d(x)$ be the probability density function of the χ_d^2 distribution, and let $c_{\alpha,d}$ be the upper α quantile of the χ_d^2 distribution. Then*

$$\begin{aligned}
&\mathbb{P}\left[r^2\{C_n^{split}(\alpha)\}/r^2\{C_n^{LRT}(\alpha)\} \leq 4\right] \geq 1 - \alpha - \log(1/\alpha)f_d(c_{\alpha,d} + \log(\alpha)) \\
\textit{and}\quad &\mathbb{P}\left[r^2\{C_n^{split}(\alpha)\}/r^2\{C_n^{LRT}(\alpha)\} \leq 4\right] \leq 1 - \alpha - \log(1/\alpha)f_d(c_{\alpha,d}).
\end{aligned}$$

Proof. We use the fact that $r^2(C_n^{\mathrm{split}}(\alpha)) = \frac{4}{n}\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2$. As established in the proof of Theorem 2.3.1 and the derivation of equation 2.7, we know that $\|\overline{Y}_0 - \overline{Y}_1\|^2 \overset{d}{=} (4/n)\chi_d^2$. Let

$X \sim \chi^2_d$. Note that $\log(\alpha) < 0$. Then

$$\begin{aligned}
\mathbb{P}\left(r^2(C_n^{\text{split}}(\alpha)) \,/\, r^2(C_n^{\text{LRT}}(\alpha)) \leq 4\right) &= \mathbb{P}\left(r^2(C_n^{\text{split}}(\alpha)) \leq \frac{4}{n}c_{\alpha,d}\right) \\
&= \mathbb{P}\left(\frac{4}{n}\log(1/\alpha) + \frac{4}{n}X \leq \frac{4}{n}c_{\alpha,d}\right) \\
&= \mathbb{P}\left(\log(1/\alpha) + X \leq c_{\alpha,d}\right) \\
&= \mathbb{P}(X \leq c_{\alpha,d} + \log(\alpha)) \\
&= \mathbb{P}(X \leq c_{\alpha,d}) - \mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d}) \\
&= 1 - \alpha - \mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d}).
\end{aligned}$$

Now we need to bound $\mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d})$. Under the assumed conditions, we show that the χ^2_d pdf is decreasing on $[c_{\alpha,d} + \log(\alpha), c_{\alpha,d}]$. Let $f_d(x)$ be the χ^2_d pdf. The following five statements are equivalent:

$$\begin{aligned}
0 &> \frac{\partial}{\partial x} f_d(x) \\
0 &> \frac{1}{2^{d/2}\,\Gamma(d/2)}\left[\left(\frac{d}{2} - 1\right)x^{d/2-2}e^{-x/2} + x^{d/2-1}\left(-\frac{1}{2}e^{-x/2}\right)\right] \\
x^{d/2-1}\left(\frac{1}{2}e^{-x/2}\right) &> \left(\frac{d}{2} - 1\right)x^{d/2-2}e^{-x/2} \\
\frac{x}{2} &> \frac{d}{2} - 1 \\
x &> d - 2
\end{aligned}$$

By our initial assumption, $c_{\alpha,d} + \log(\alpha) > d - 2$. Thus, $f_d(x)$ is decreasing on $[c_{\alpha,d} + \log(\alpha), c_{\alpha,d}]$. Since the interval has length $\log(1/\alpha)$,

$$\log(1/\alpha) f_d(c_{\alpha,d}) \;\leq\; \mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d}) \;\leq\; \log(1/\alpha) f_d(c_{\alpha,d} + \log(\alpha)).$$

The bounds on $\mathbb{P}\left(r^2(C_n^{\text{split}}(\alpha)) \,/\, r^2(C_n^{\text{LRT}}(\alpha)) \leq 4\right)$ follow immediately. □

Before proving Theorem 2.4.1, we establish Lemma B.0.3 and Lemma B.0.4.

Lemma B.0.3. *Assume the doughnut null test setting. Let $\mathcal{P}_{\Theta_0}$ be the set of all convex combinations of $N(\theta, I_d)$ densities such that $\|\theta\| \in [0.5, 1]$. When $\|\overline{Y}_1\| > 1$ and $\widehat{\theta}_1 = \overline{Y}_1$, the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$ is $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}$.*

Proof. Suppose $\|\overline{Y}_1\| > 1$. Defining $\widehat{\theta}_1 = \overline{Y}_1$ as in Table 2.1, $\|\widehat{\theta}_1\| > 1$. The RIPR of $\widehat{\theta}_1$ onto the convex set $\mathcal{P}_{\Theta_0}$ minimizes $D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \| p_0)$ out of all densities $p_0 \in \mathcal{P}_{\Theta_0}$. Suppose $p_0 \in \mathcal{P}_{\Theta_0}$. Then we can write p_0 as a mixture of $N(\theta_k, I_d)$ densities. We write $p_0 = \sum_{k=1}^K w_k p_{\theta_k}$, where $K \in \mathbb{N}$, $\sum_{k=1}^K w_k = 1$, and for each $k = 1, \ldots, K$, $0 < w_k < 1$ and $\|\theta_k\| \in [0.5, 1]$. Note that $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|} \in \mathcal{P}_0$. To prove that $D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}) = \inf_{p_0 \in \mathcal{P}_{\Theta_0}} D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel p_0)$, we show $D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}) \leq D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel \sum_{k=1}^K w_k p_{\theta_k})$.

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(p_{\widehat{\theta}_1} \,\Big\|\, \sum_{k=1}^K w_k p_{\theta_k}\right) - D_{\mathrm{KL}}\left(p_{\widehat{\theta}_1} \parallel p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}\right) \\
&= \int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{p_{\widehat{\theta}_1}(y)}{\sum_{k=1}^K w_k p_{\theta_k}(y)}\right) dy - \int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{p_{\widehat{\theta}_1}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right) dy \\
&= \int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}{\sum_{k=1}^K w_k p_{\theta_k}(y)}\right) dy \\
&= -\int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{\sum_{k=1}^K w_k p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right) dy \\
&= -\mathbb{E}_{\widehat{\theta}_1}\left[\log\left\{\frac{\sum_{k=1}^K w_k p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\}\right] \\
&\geq -\log \mathbb{E}_{\widehat{\theta}_1}\left\{\frac{\sum_{k=1}^K w_k p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\} && \text{(B.17)} \\
&= -\log\left[\sum_{k=1}^K w_k \mathbb{E}_{\widehat{\theta}_1}\left\{\frac{p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\}\right] \\
&\geq -\log\left\{\sum_{k=1}^K w_k(1)\right\} && \text{(B.18)} \\
&= 0.
\end{aligned}
$$

(B.17) holds by Jensen's inequality. (B.18) holds by the following derivation:

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\theta}_1}\left\{\frac{p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\}\\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2\right) \frac{\exp\left(-\frac{1}{2}\|y-\theta_k\|^2\right)}{\exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right)} dy\\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2 - \frac{1}{2}\|y-\widehat{\theta}_1+\widehat{\theta}_1-\theta_k\|^2 + \frac{1}{2}\|y-\widehat{\theta}_1+\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right) dy\\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\Bigg(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2 - (y-\widehat{\theta}_1)^T(\widehat{\theta}_1-\theta_k) - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 + (y-\widehat{\theta}_1)^T(\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|)+\\
&\qquad\qquad \frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\Bigg)dy\\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2\right)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2+(y-\widehat{\theta}_1)^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right)dy\\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2\right)\mathbb{E}_{\widehat{\theta}_1}\left[\exp\left\{(y-\widehat{\theta}_1)^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right\}\right]\\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 - \widehat{\theta}_1^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right)\mathbb{E}_{\widehat{\theta}_1}\left[\exp\left\{(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)^Ty\right\}\right]\\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 - \widehat{\theta}_1^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right)\exp\left\{\widehat{\theta}_1^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)+\frac{1}{2}\|\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right\}\\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 + \frac{1}{2}\|\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right)\\
&= \exp\Bigg(\frac{1}{2}\|\widehat{\theta}_1\|^2 - \widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\| + \frac{1}{2}\widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\|^2 - \frac{1}{2}\|\widehat{\theta}_1\|^2 + \widehat{\theta}_1^T\theta_k - \frac{1}{2}\|\theta_k\|^2+\\
&\qquad \frac{1}{2}\|\theta_k\|^2 - \theta_k^T\widehat{\theta}_1/\|\widehat{\theta}_1\| + \frac{1}{2}\widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\|^2\Bigg)\\
&= \exp\left(\widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\|^2 - \widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k^T\widehat{\theta}_1/\|\widehat{\theta}_1\| + \widehat{\theta}_1^T\theta_k\right)\\
&= \exp\left\{(\widehat{\theta}_1/\|\widehat{\theta}_1\|-\widehat{\theta}_1)^T(\widehat{\theta}_1/\|\widehat{\theta}_1\|-\theta_k)\right\}\\
&\le \exp(0) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{B.19})\\
&= 1.
\end{aligned}
$$

To justify (B.19), note that

$$(\widehat{\theta}_1/\|\widehat{\theta}_1\|-\widehat{\theta}_1)^T(\widehat{\theta}_1/\|\widehat{\theta}_1\|-\theta_k) = \left\|\widehat{\theta}_1/\|\widehat{\theta}_1\|-\widehat{\theta}_1\right\|\left\|\widehat{\theta}_1/\|\widehat{\theta}_1\|-\theta_k\right\|\cos(\gamma),$$

where γ is the angle between $\widehat{\theta}_1/\|\widehat{\theta}_1\|-\widehat{\theta}_1$ and $\widehat{\theta}_1/\|\widehat{\theta}_1\|-\theta_k$. Recall that Θ_0 has a spherical outer border, $\|\theta_k\| \in [0.5, 1]$, $\|\widehat{\theta}_1\| > 1$, and $\widehat{\theta}_1/\|\widehat{\theta}_1\|$ is on the outer border of Θ_0. Thus, γ will always be between 90° and 270°. (See Fig. B.1.) This implies that $(\widehat{\theta}_1/\|\widehat{\theta}_1\|-\widehat{\theta}_1)^T(\widehat{\theta}_1/\|\widehat{\theta}_1\|-\theta_k) \le 0$. □

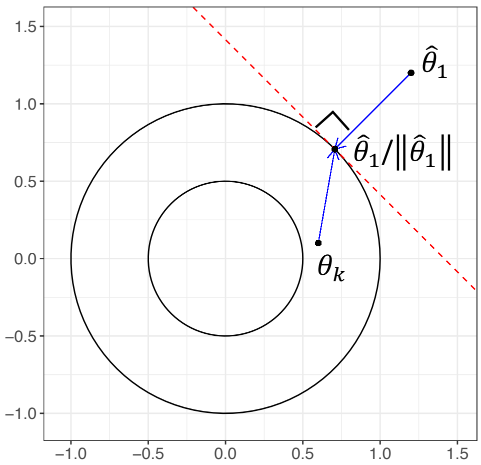


Figure B.1: Lemma B.0.3 companion diagram. The angle between $\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1$ and $\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k$ must be between 90° and 270°.

Lemma B.0.4. *Assume the doughnut null test setting. Let* $R_n = \prod_{Y_i \in \mathcal{D}_0} \{p_{\widehat{\theta}_1}(Y_i)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y_i)\}$. *If* $\theta^* \in \Theta_0$, *then* $\mathbb{E}_{\theta^*}\{R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\} \leq \mathbb{1}(\|\overline{Y}_1\| > 1)$.

Proof. If $\mathcal{D}_1$ satisfies $\|\overline{Y}_1\| \leq 1$, then

$$\mathbb{E}_{\theta^*}\{R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\} = 0 = \mathbb{1}(\|\overline{Y}_1\| > 1).$$

Now suppose $\mathcal{D}_1$ satisfies $\|\overline{Y}_1\| > 1$. Then $\|\widehat{\theta}_1\| > 1$, and $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}$ is the RIPR of $p_{\widehat{\theta}_1}$ onto the convex set of densities $\mathcal{P}_{\Theta_0}$, as proved in Lemma B.0.3. Since $\theta^* \in \Theta_0$, $\widehat{\theta}_1 \in \Theta_1$, and $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}$ is the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$, we know $\mathbb{E}_{\theta^*}\{p_{\widehat{\theta}_1}(Y)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y) \mid \mathcal{D}_1\} \leq 1$, as explained under *Approach 3: Subsampled hybrid LRT*. So

$$\begin{aligned}
\mathbb{E}_{\theta^*}\{R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\} &= \mathbb{E}_{\theta^*}\left[\prod_{Y_i \in \mathcal{D}_0} \{p_{\widehat{\theta}_1}(Y_i)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y_i)\} \mid \mathcal{D}_1\right] \\
&\overset{iid}{=} \prod_{i=1}^{n/2} \mathbb{E}_{\theta^*}\left\{p_{\widehat{\theta}_1}(Y_i)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y_i) \mid \mathcal{D}_1\right\} \\
&\leq 1 \\
&= \mathbb{1}(\|\overline{Y}_1\| > 1).
\end{aligned}$$

□

Theorem 2.4.1. *In the doughnut null hypothesis test setting, assume the subsampled test statistics* $U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b}) \,/\, \mathcal{L}_{0,b}(\widehat{\theta}^{split}_{0,b})$ *and* $R_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b})/\mathcal{L}_{0,b}(\widehat{\theta}^{RIPR}_{0,b})$, $1 \le b \le B$. *A valid level* α *test rejects* H_0 *when*

$$\frac{1}{B}\sum_{b=1}^{B}\Big\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5,1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\Big\} \ge 1/\alpha.$$

Proof. Assume $\theta^* \in \Theta_0$. The probability of falsely rejecting H_0 is

$$\begin{align*}
&\mathbb{P}_{\theta^*}\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5,1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\Big\} \ge 1/\alpha\right] \\
&\quad \le \alpha\mathbb{E}_{\theta^*}\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5,1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\Big\}\right] \\
&\quad \le \alpha\mathbb{E}_{\theta^*}\left[\frac{1}{B}\sum_{b=1}^{B}\Big\{T_{n,b}(\theta^*)\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5,1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\Big\}\right] \tag{B.20} \\
&\quad = \alpha\mathbb{E}_{\theta^*}\Big\{T_n(\theta^*)\mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5,1]) + R_n\mathbb{1}(\|\overline{Y}_1\| > 1)\Big\} \\
&\quad = \alpha\mathbb{E}_{\theta^*}\Big[\mathbb{E}_{\theta^*}\Big\{T_n(\theta^*)\mathbb{1}(\|\overline{Y}_1\| < 0.5) \mid \mathcal{D}_1\Big\}\Big] + \alpha\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5,1]) + \alpha\mathbb{E}_{\theta^*}\Big[\mathbb{E}_{\theta^*}\Big\{R_n\mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\Big\}\Big] \\
&\quad \le \alpha\mathbb{E}_{\theta^*}\Big[\mathbb{1}(\|\overline{Y}_1\| < 0.5)\mathbb{E}_{\theta^*}\left\{T_n(\theta^*) \mid \mathcal{D}_1\right\}\Big] + \alpha\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5,1]) + \alpha\mathbb{E}_{\theta^*}\{\mathbb{1}(\|\overline{Y}_1\| > 1)\} \tag{B.21} \\
&\quad \le \alpha\mathbb{E}_{\theta^*}\{\mathbb{1}(\|\overline{Y}_1\| < 0.5)\} + \alpha\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5,1]) + \alpha\mathbb{P}_{\theta^*}\{\mathbb{1}(\|\overline{Y}_1\| > 1)\} \tag{B.22} \\
&\quad = \alpha\Big\{\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| < 0.5) + \mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5,1]) + \mathbb{P}_{\theta^*}(\|\overline{Y}_1\| > 1)\Big\} \\
&\quad = \alpha.
\end{align*}$$

(B.20) holds because $\widehat{\theta}^{\mathrm{split}}_{0,b} = \arg\max\limits_{\theta\in\Theta_0} \mathcal{L}_{0,b}(\theta)$. Since $\theta^* \in \Theta_0$,

$$U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_1)/\mathcal{L}_{0,b}(\widehat{\theta}^{\mathrm{split}}_{0,b}) \le \mathcal{L}_{0,b}(\widehat{\theta}_1)/\mathcal{L}_{0,b}(\theta^*) = T_{n,b}(\theta^*).$$

(B.21) holds by Lemma B.0.4. (B.22) holds because $\mathbb{E}_{\theta^*}\{T_n(\theta^*) \mid \mathcal{D}_1\} \le 1$, as established by Theorem 1.0.1. □

B.2 Derivations of Equations

Derivation of Equation 2.1. The classical likelihood ratio confidence set for $\theta^* \in \mathbb{R}^d$ is given by

$$C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : 2\log\frac{\mathcal{L}(\overline{Y})}{\mathcal{L}(\theta)} \leq c_{\alpha,d}\right\},$$

where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution. $\overline{Y}$ is the sample mean of the Y_i observations, and it is also the MLE estimate for θ^*. We re-write this confidence set such that the squared radius of the set is apparent.

$$\begin{aligned}
2\log\frac{\mathcal{L}(\overline{Y})}{\mathcal{L}(\theta)} &= 2\log\left(\frac{\Pi_{i=1}^n \exp\left(-\frac{1}{2}(Y_i-\overline{Y})^T(Y_i-\overline{Y})\right)}{\Pi_{i=1}^n \exp\left(-\frac{1}{2}(Y_i-\theta)^T(Y_i-\theta)\right)}\right)\\
&= 2\log\left(\exp\left(-\frac{1}{2}\sum_{i=1}^n (Y_i-\overline{Y})^T(Y_i-\overline{Y}) + \frac{1}{2}\sum_{i=1}^n (Y_i-\theta)^T(Y_i-\theta)\right)\right)\\
&= -\sum_{i=1}^n (Y_i-\overline{Y})^T(Y_i-\overline{Y}) + \sum_{i=1}^n (Y_i-\theta)^T(Y_i-\theta)\\
&= \sum_{i=1}^n \left(-(Y_i-\overline{Y})^T(Y_i-\overline{Y}) + (Y_i-\overline{Y}+\overline{Y}-\theta)^T(Y_i-\overline{Y}+\overline{Y}-\theta)\right)\\
&= \sum_{i=1}^n \Bigg(-(Y_i-\overline{Y})^T(Y_i-\overline{Y}) + (Y_i-\overline{Y})^T(Y_i-\overline{Y}) +\\
&\qquad 2(Y_i-\overline{Y})^T(\overline{Y}-\theta) + (\overline{Y}-\theta)^T(\overline{Y}-\theta)\Bigg)\\
&= n\|\overline{Y}-\theta\|^2.
\end{aligned}$$

The final step holds because the first two terms cancel and the summation over the third term equals 0. Therefore,

$$C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \leq c_{\alpha,d}/n\right\}.$$

□

Derivation of Equation 2.2. Let $\widehat{\theta}_1 = \overline{Y}_1$ be the sample mean of the $n/2$ observations in $\mathcal{D}_1$. Where $\mathcal{L}_0(\theta) = \prod_{Y_i \in \mathcal{D}_0} p_\theta(Y_i)$ and $T_n(\theta) = \mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\theta)$, the universal confidence set using the split likelihood ratio statistic is

$$C_n^{\text{split}}(\alpha) = \{\theta \in \Theta : T_n(\theta) < 1/\alpha\}\,.$$

We also re-write this confidence set such that the squared radius of the set is apparent.

$$\begin{aligned}
T_n(\theta) &= \frac{\Pi_{Y_i\in\mathcal{D}_0} \exp\left(-\frac{1}{2}(Y_i-\widehat{\theta}_1)^T(Y_i-\widehat{\theta}_1)\right)}{\Pi_{Y_i\in\mathcal{D}_0} \exp\left(-\frac{1}{2}(Y_i-\theta)^T(Y_i-\theta)\right)}\\
&= \exp\left(\sum_{Y_i\in\mathcal{D}_0}\left(-\frac{1}{2}(Y_i-\overline{Y}_1)^T(Y_i-\overline{Y}_1)+\frac{1}{2}(Y_i-\theta)^T(Y_i-\theta)\right)\right)\\
&= \exp\Bigg(\sum_{Y_i\in\mathcal{D}_0}\Bigg(-\frac{1}{2}(Y_i-\overline{Y}_0+\overline{Y}_0-\overline{Y}_1)^T(Y_i-\overline{Y}_0+\overline{Y}_0-\overline{Y}_1)+\\
&\qquad\qquad \frac{1}{2}(Y_i-\overline{Y}_0+\overline{Y}_0-\theta)^T(Y_i-\overline{Y}_0+\overline{Y}_0-\theta)\Bigg)\Bigg)\\
&= \exp\Bigg(\sum_{Y_i\in\mathcal{D}_0}\Bigg(-\frac{1}{2}\left[(Y_i-\overline{Y}_0)^T(Y_i-\overline{Y}_0)+2(Y_i-\overline{Y}_0)^T(\overline{Y}_0-\overline{Y}_1)+(\overline{Y}_0-\overline{Y}_1)^T(\overline{Y}_0-\overline{Y}_1)\right]+\\
&\qquad\qquad \frac{1}{2}\left[(Y_i-\overline{Y}_0)^T(Y_i-\overline{Y}_0)+2(Y_i-\overline{Y}_0)^T(\overline{Y}_0-\theta)+(\overline{Y}_0-\theta)^T(\overline{Y}_0-\theta)\right]\Bigg)\Bigg) \qquad \text{(B.23)}\\
&= \exp\left(\sum_{Y_i\in\mathcal{D}_0}\left(-\frac{1}{2}(\overline{Y}_0-\overline{Y}_1)^T(\overline{Y}_0-\overline{Y}_1)+\frac{1}{2}(\overline{Y}_0-\theta)^T(\overline{Y}_0-\theta)\right)\right)\\
&= \exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\theta\|^2\right). \qquad \text{(B.24)}
\end{aligned}$$

The first and fourth terms of (B.23) cancel, and the cross-product terms equal 0 upon taking the summation. (B.24) holds because $\mathcal{D}_0$ contains $\frac{n}{2}$ elements. Therefore,

$$\begin{aligned}
C_n^{\text{split}}(\alpha) &= \left\{\theta\in\Theta : T_n(\theta) < \frac{1}{\alpha}\right\}\\
&= \left\{\theta\in\Theta : \exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\theta\|^2\right) < \frac{1}{\alpha}\right\}\\
&= \left\{\theta\in\Theta : -\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\theta\|^2 < \log\left(\frac{1}{\alpha}\right)\right\}\\
&= \left\{\theta\in\Theta : \frac{n}{4}\|\overline{Y}_0-\theta\|^2 < \log\left(\frac{1}{\alpha}\right)+\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2\right\}\\
&= \left\{\theta\in\Theta : \|\overline{Y}_0-\theta\|^2 < \frac{4}{n}\log\left(\frac{1}{\alpha}\right)+\|\overline{Y}_0-\overline{Y}_1\|^2\right\}.
\end{aligned}$$

□

Derivation of Equation 2.7. We have partitioned $\mathcal{D}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$. Let us rewrite the observations in these data sets as $\mathcal{D}_0 = \{Y_{0i}\}_{i=1}^{n/2}$ and $\mathcal{D}_1 = \{Y_{1i}\}_{i=1}^{n/2}$. From the presentation of $C_n^{\text{split}}(\alpha)$ in (2.2), we see that $r^2(C_n^{\text{split}}(\alpha)) = \frac{4}{n}\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2$. Note that

$$\|\overline{Y}_0 - \overline{Y}_1\|^2 = \left\| \frac{2}{n}\sum_{i=1}^{n/2}(Y_{0i} - Y_{1i}) \right\|^2 = \frac{4}{n}\left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n/2}(Y_{0i} - Y_{1i}) \right\|^2 \overset{d}{=} \frac{4}{n}\chi_d^2.$$

To see why the last step holds, note that $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\theta^*, I_d)$. So for any i, $Y_{0i} - Y_{1i} \overset{iid}{\sim} N(0, 2I_d)$. Then $\sum_{i=1}^{n/2}(Y_{0i} - Y_{1i}) \overset{iid}{\sim} N\left(0, \frac{n}{2}(2I_d)\right)$, and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n/2}(Y_{0i} - Y_{1i}) \overset{iid}{\sim} N(0, I_d)$. This implies that $r^2(C_n^{\text{split}}(\alpha)) \overset{d}{=} \frac{4}{n}\log(1/\alpha) + \frac{4}{n}\chi_d^2$. Therefore, $\mathbb{E}[r^2(C_n^{\text{split}}(\alpha))] = \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \frac{4}{n}d$.

Derivation of Equation 2.9. From equation 2.8, we know that

$$\frac{\mathbb{E}[r^2(C_n^{\text{split}}(\alpha))]}{r^2(C_n^{\text{LRT}}(\alpha))} = \frac{4\log(1/\alpha) + 4d}{c_{\alpha,d}}.$$

For $d \geq 1$ and $\alpha \in (0, 1)$, Inglot (2010) shows the upper bound

$$c_{\alpha,d} \leq d + 2\log\left(\frac{1}{\alpha}\right) + 2\sqrt{d\log\left(\frac{1}{\alpha}\right)}.$$

Also, for $d \geq 2$ and $\alpha \leq 0.17$, Inglot (2010) shows the lower bound

$$c_{\alpha,d} \geq d + 2\log\left(\frac{1}{\alpha}\right) - \frac{5}{2}.$$

Combining these facts, we see that for $d \geq 2$ and $\alpha \leq 0.17$,

$$\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \frac{\mathbb{E}[r^2(C_n^{\text{split}}(\alpha))]}{r^2(C_n^{\text{LRT}}(\alpha))} \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d - \frac{5}{2}}.$$

□

Derivation of Equation 2.10. From equation 2.8, we know that

$$\frac{\mathbb{E}[r^2(C_n^{\text{split}}(\alpha))]}{r^2(C_n^{\text{LRT}}(\alpha))} = \frac{4\log(1/\alpha) + 4d}{c_{\alpha,d}}.$$

The lower bound of equation 2.10 is the same as the lower bound from equation 2.9. We consider the upper bound. Suppose $d = 1$ and $\alpha \leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right)$. Let $t = -2 + \sqrt{5 + 2\log(1/\alpha)}$. We will show that $c_{\alpha,1} \geq t^2$ in several steps:

Step 1: Show that $t^2 + 4t - 2 < 2\log(1/\alpha)$.

$$\begin{aligned}
t^2 + 4t - 2 &= \left(-2 + \sqrt{5 + 2\log(1/\alpha)}\right)^2 + 4(-2 + \sqrt{5 + 2\log(1/\alpha)}) - 2 \\
&= 4 - 4\sqrt{5 + 2\log(1/\alpha)} + 5 + 2\log(1/\alpha) - 8 + 4\sqrt{5 + 2\log(1/\alpha)} - 2 \\
&= 2\log(1/\alpha) - 1 \\
&< 2\log(1/\alpha).
\end{aligned}$$

Step 2: Show that $\log(1/\alpha) > \frac{t^2}{2} + 2\log(t) + \log(\sqrt{2\pi})$. Starting with the result from Step 1,

$$\begin{aligned}
\log(1/\alpha) &> \frac{t^2}{2} + 2t - 1 \\
&\geq \frac{t^2}{2} + 2(\log(t) + 1) - 1 \qquad \text{since } t \geq \log(t) + 1 \text{ for } t > 0 \\
&= \frac{t^2}{2} + 2\log(t) + 1 \\
&> \frac{t^2}{2} + 2\log(t) + \log(\sqrt{2\pi}).
\end{aligned}$$

Step 3: Show that $t^2 - 1 \geq t$. We start by showing that $t \geq \frac{1}{2}(1 + \sqrt{5})$ follows from our definitions of t and α:

$$\begin{aligned}
& & \alpha &\leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right) \\
&\iff & \frac{1}{\alpha} &\geq \exp\left(\frac{5(1+\sqrt{5})}{4}\right) \\
&\iff & 8\log(1/\alpha) &\geq 10(1+\sqrt{5}) \\
&\iff & 20 + 8\log(1/\alpha) &\geq 30 + 10\sqrt{5} \\
&\iff & 4(5 + 2\log(1/\alpha)) &\geq 25 + 10\sqrt{5} + 5 \\
&\iff & 2\sqrt{5 + 2\log(1/\alpha)} &\geq 5 + \sqrt{5} \\
&\iff & -4 + 2\sqrt{5 + 2\log(1/\alpha)} &\geq 1 + \sqrt{5} \\
&\iff & -2 + \sqrt{5 + 2\log(1/\alpha)} &\geq \frac{1}{2}(1+\sqrt{5}) \\
&\iff & t &\geq \frac{1}{2}(1+\sqrt{5}).
\end{aligned}$$

The roots of the convex function $t^2 - t - 1$ are at $t = (1 \pm \sqrt{5})/2$. At $t \geq (1/2)(1+\sqrt{5})$, we know $t^2 - 1 \geq t$.

Step 4: Show that $t^2 \leq c_{\alpha,1}$. Starting with the results of steps 2 and 3,

$$\begin{aligned}
\log(t^2 - 1) - t^2/2 - \log(\sqrt{2\pi}) &\geq \log(t) + 2\log(t) + \log(\alpha) \\
&= 3\log(t) + \log(\alpha).
\end{aligned}$$

Exponentiating,

$$\left(t^2 - 1\right) \exp\left(-t^2/2\right) \left(\frac{1}{\sqrt{2\pi}}\right) \geq t^3 \alpha.$$

So

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \exp\left(-t^2/2\right) \left(\frac{1}{\sqrt{2\pi}}\right) \geq \alpha.$$

If $Z \sim N(0,1)$ and $X = Z^2 \sim \chi^2_1$, then using an inequality on $\mathbb{P}(Z \geq t)$ from Polland (2015),

$$\mathbb{P}(X \geq t^2) = 2\mathbb{P}(Z \geq t) > \mathbb{P}(Z \geq t) \geq \left(\frac{1}{t} - \frac{1}{t^3}\right) \exp\left(-t^2/2\right) \left(\frac{1}{\sqrt{2\pi}}\right) \geq \alpha.$$

This implies that $c_{\alpha,1} \geq t^2 = 2\log(1/\alpha) + 9 - 4\sqrt{5 + 2\log(1/\alpha)}$. We conclude that for $d = 1$ and $\alpha \leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right)$,

$$\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \frac{\mathbb{E}[r^2(C_n^{\text{split}}(\alpha))]}{r^2(C_n^{\text{LRT}}(\alpha))} \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + 9 - 4\sqrt{5 + 2\log(1/\alpha)}}.$$

□

Derivation of Equation 2.12. The classical LRT set is

$$C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 \leq c_{\alpha,d} \,/\, n\right\},$$

where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution. Suppose we are testing $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$. The power of the classical LRT at the true θ^* is thus

$$\text{Power}(C_n^{\text{LRT}}(\alpha); \theta^*) = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > c_{\alpha,d}/n\right).$$

We can express the power function of the classical LRT in terms of the CDF of a noncentral χ^2 distribution. Let us denote $\theta^* = (\theta^*_1, \ldots, \theta^*_d)$. We see that

$$n\|\overline{Y}\|^2 = \left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i\right\|^2 = \sum_{j=1}^d \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_{ij}\right)^2.$$

For each dimension j, $n^{-1/2}\sum_{i=1}^n Y_{ij} \sim N(\theta^*_j\sqrt{n}, 1)$. So this follows a non-central χ^2 distribution given by

$$n\|\overline{Y}\|^2 \overset{d}{=} \chi^2\left(df = d, \lambda = \sum_{j=1}^d n(\theta^*_j)^2\right) \overset{d}{=} \chi^2\left(df = d, \lambda = n\|\theta^*\|^2\right).$$

Let $\Phi(\cdot)$ represent the standard normal CDF. Suppose $X \sim \chi^2(df = d, \lambda = n\|\theta^*\|^2)$. As $d \to \infty$ or as $\lambda \to \infty$, it holds that

$$\frac{X - (d + n\|\theta^*\|^2)}{\sqrt{2(d + 2n\|\theta^*\|^2)}} \approx N(0,1).$$

See Chun and Shapiro (2009). Using the Normal approximation to the non-central chi-squared CDF, the power of the classical LRT is

$$\begin{aligned}
\text{Power}(C_n^{\text{LRT}}(\alpha); \theta^*) &= \mathbb{P}_{\theta^*}\left(\|\bar{Y}\|^2 > c_{\alpha,d}/n\right) \\
&= \mathbb{P}_{\theta^*}\left(n\|\bar{Y}\|^2 > c_{\alpha,d}\right) \\
&= \mathbb{P}_{\theta^*}\left(\frac{n\|\bar{Y}\|^2 - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}} > \frac{c_{\alpha,d} - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right) \\
&\approx 1 - \Phi\left(\frac{c_{\alpha,d} - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right) \\
&= \Phi\left(\frac{d + n\|\theta^*\|^2 - c_{\alpha,d}}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right).
\end{aligned}$$

□

Derivation of Equation 2.13. Using methods from the derivation of equation 2.12, we can find an approximation to the power of the limiting subsampling LRT as $B \to \infty$. From equation 2.4,

$$C_n^{\text{subsplit}}(\alpha) \approx \left\{\theta \in \Theta : \|\bar{Y} - \theta\|^2 < \frac{10}{3n} \log\left(\left(\frac{5}{2}\right)^{d/2} \frac{1}{\alpha}\right)\right\}$$

So the power of the limit of subsampling LRT for large B is

$$\begin{aligned}
\text{Power}(C_n^{\text{subsplit}}(\alpha); \theta^*) &\approx \mathbb{P}_{\theta^*}\left(n\|\bar{Y}\|^2 \geq \frac{10}{3} \log\left(\left(\frac{5}{2}\right)^{d/2} \frac{1}{\alpha}\right)\right) \\
&= \mathbb{P}_{\theta^*}\left(\frac{n\|\bar{Y}\|^2 - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}} \geq \frac{(10/3)\log\left((5/2)^{d/2}(1/\alpha)\right) - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right) \\
&\approx \Phi\left(\frac{1}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\left(d + n\|\theta^*\|^2 - \frac{10}{3}\log\left(\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right)\right)\right).
\end{aligned}$$

□

B.3 Simulated Cross-fit Sets with Varying p_0

In the split LRT case, the optimal split proportion p_0^* (established in Theorem 2.3.1) converges to 0.5 as $d \to \infty$. This optimal split proportion minimizes the expected squared radius. Under general p_0, the cross-fit set is defined as

$$C_n^{\mathrm{CF}}(\alpha) = \Bigg\{ \theta \in \Theta : \frac{1}{2}\Bigg[\exp\left(-\frac{np_0}{2}\|\bar{Y}_0 - \bar{Y}_1\|^2 + \frac{np_0}{2}\|\bar{Y}_0 - \theta\|^2\right) + \\ \exp\left(-\frac{n(1-p_0)}{2}\|\bar{Y}_0 - \bar{Y}_1\|^2 + \frac{n(1-p_0)}{2}\|\bar{Y}_1 - \theta\|^2\right)\Bigg] < \frac{1}{\alpha} \Bigg\}.$$

Noting the symmetry of the set $C_n^{\mathrm{CF}}(\alpha)$, we conjecture that $p_0 = 0.5$ will minimize the expected squared diameter of the cross-fit set. Figure B.2 presents examples of cross-fit sets at varying p_0 on a single sample of 1000 observations simulated from a $N(\vec{0}, I_2)$ distribution. We see that the regions with $p_0 \in \{0.5, 0.7\}$ have smaller diameters than the regions with $p_0 \in \{0.1, 0.3, 0.9\}$.

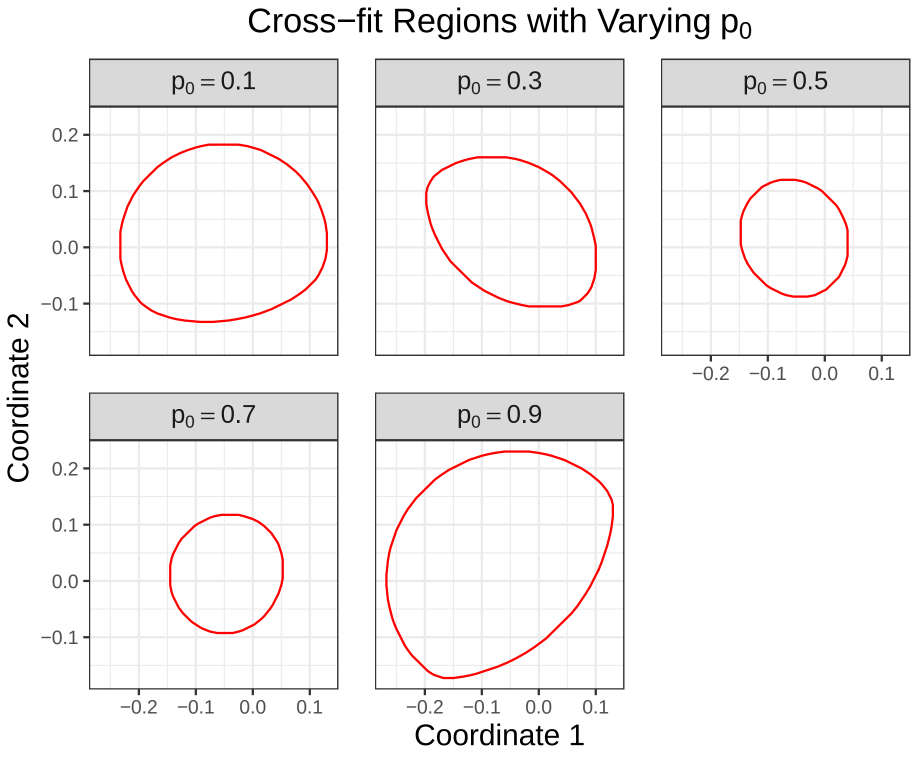


Figure B.2: Simulated cross-fit regions at varying p_0, using a single data sample.

B.4 Power of Tests of $H_0 : \|\theta^*\| \in [0.5, 1]$

B.4.1 Exact Formula for Power of Intersection Test

In section 2.4, we present hypothesis tests for $H_0 : \|\theta^*\| \in [0.5, 1]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1]$. The power of the intersection method that we present is tractable. We derive a formula for the intersection method's power at θ^*. From the intersection method's description, we reject H_0 if and only if $C_n^{\text{LRT}}(\alpha) \cap (\mathcal{S}_1 \backslash \mathcal{S}_{0.5}) = \emptyset$, where $C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \leq c_{\alpha,d}/n\right\}$. This is equivalent to rejecting H_0 if and only if $\widehat{\theta}^{\text{proj}} \notin C_n^{\text{LRT}}(\alpha)$, where

$$\widehat{\theta}^{\text{proj}} = \begin{cases} 0.5\,\overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| < 0.5 \\ \overline{Y} & : \|\overline{Y}\| \in [0.5, 1.0] \\ \overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| > 1 \end{cases} \cdot$$

In **Case 2**, we have $\|\overline{Y}\| \in [0.5, 1]$. In this setting, it is always true that $\widehat{\theta}^{\text{proj}} = \overline{Y} \in C_n^{\text{LRT}}(\alpha)$. So we will never reject H_0 in this case. We consider **Case 1** ($\|\overline{Y}\| < 0.5$) and **Case 3** ($\|\overline{Y}\| > 1$). For $\|\theta^*\| \notin [0.5, 1.0]$, the power is given by

$$\begin{aligned} \text{Power}(\theta^*) = \mathbb{P}_{\theta^*}&\left(\left\|\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| > 1\right) + \\ &\mathbb{P}_{\theta^*}\left(\left\|0.5\,\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| < 0.5\right). \end{aligned}$$

We know that $n\|\overline{Y}\|^2 \sim \chi^2(df = d, \lambda = n\|\theta^*\|^2)$. We will use this fact to write Power(θ^*) in terms of this non-central χ^2 CDF.

Case 1. Note that

$$\left\|0.5\,\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 = \frac{\overline{Y}^T\overline{Y}}{4\|\overline{Y}\|^2} - 2\frac{\overline{Y}^T\overline{Y}}{2\|\overline{Y}\|} + \|\overline{Y}\|^2 = \frac{1}{4} - \|\overline{Y}\| + \|\overline{Y}\|^2 = \left(\|\overline{Y}\| - \frac{1}{2}\right)^2.$$

Then we write

$$
\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left\|0.5\,\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| < 1/2\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\left(\|\overline{Y}\| - 1/2\right)^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| < 1/2\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(1/2 - \|\overline{Y}\| > (c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\| < 1/2\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\| < 1/2 - (c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\| < 1/2\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\| < 1/2 - (c_{\alpha,d}/n)^{1/2}\right) \\
&\quad = \mathbb{1}\left(c_{\alpha,d}/n < 1/4\right)\mathbb{P}_{\theta^*}\left(\|\overline{Y}\| < 1/2 - (c_{\alpha,d}/n)^{1/2}\right) \\
&\quad = \mathbb{1}\left(n > 4c_{\alpha,d}\right)\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 < 1/4 - \sqrt{c_{\alpha,d}/n} + c_{\alpha,d}/n\right) \\
&\quad = \mathbb{1}\left(n > 4c_{\alpha,d}\right)\mathbb{P}_{\theta^*}\left(n\|\overline{Y}\|^2 < n/4 - \sqrt{nc_{\alpha,d}} + c_{\alpha,d}\right) \\
&\quad = \mathbb{1}\left(n > 4c_{\alpha,d}\right)F_{d,n\|\theta^*\|^2}\left(n/4 - \sqrt{nc_{\alpha,d}} + c_{\alpha,d}\right), \qquad \text{(B.25)}
\end{aligned}
$$

where $F_{d,n\|\theta^*\|^2}$ is the non-central $\chi^2(df = d, \lambda = n\|\theta^*\|^2)$ CDF.

Case 3. Note that

$$
\left\|\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 = \frac{\overline{Y}^T\overline{Y}}{\|\overline{Y}\|^2} - 2\frac{\overline{Y}^T\overline{Y}}{\|\overline{Y}\|} + \|\overline{Y}\|^2 = 1 - 2\|\overline{Y}\| + \|\overline{Y}\|^2 = \left(\|\overline{Y}\| - 1\right)^2.
$$

Then we write

$$
\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left\|\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| > 1\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\left(\|\overline{Y}\| - 1\right)^2 > c_{\alpha,d}/n,\ \|\overline{Y}\|^2 > 1\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\| - 1 > (c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\|^2 > 1\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\| > 1 + (c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\|^2 > 1\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > 1 + (2/\sqrt{n})c_{\alpha,d}^{1/2} + c_{\alpha,d}/n,\ \|\overline{Y}\|^2 > 1\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > 1 + (2/\sqrt{n})c_{\alpha,d}^{1/2} + c_{\alpha,d}/n\right) \\
&\quad = \mathbb{P}_{\theta^*}\left(n\|\overline{Y}\|^2 > n + 2\sqrt{nc_{\alpha,d}} + c_{\alpha,d}\right) \\
&\quad = 1 - F_{d,n\|\theta^*\|^2}(n + 2\sqrt{nc_{\alpha,d}} + c_{\alpha,d}),
\end{aligned}
\tag{B.26}
$$

where $F_{d,n\|\theta^*\|^2}$ is the non-central $\chi^2(df = d, \lambda = n\|\theta^*\|^2)$ CDF.

For a given $\|\theta^*\| \notin [0.5, 1]$, our calculation of Power(θ^*) is given by (B.26) + (B.25). That is,

$$
\begin{aligned}
\text{Power}(\theta^*) = 1 - F_{d,n\|\theta^*\|^2}(n + 2\sqrt{nc_{\alpha,d}} + c_{\alpha,d}) + \\
\mathbb{1}\left(n > 4c_{\alpha,d}\right) F_{d,n\|\theta^*\|^2}\left(n/4 - \sqrt{nc_{\alpha,d}} + c_{\alpha,d}\right).
\end{aligned}
$$

Figure B.3 compares this calculated power to the simulated power of the intersection method from Figure 2.7. The points correspond to the simulated power, and the curves trace out the calculated power. The calculated and simulated powers align.

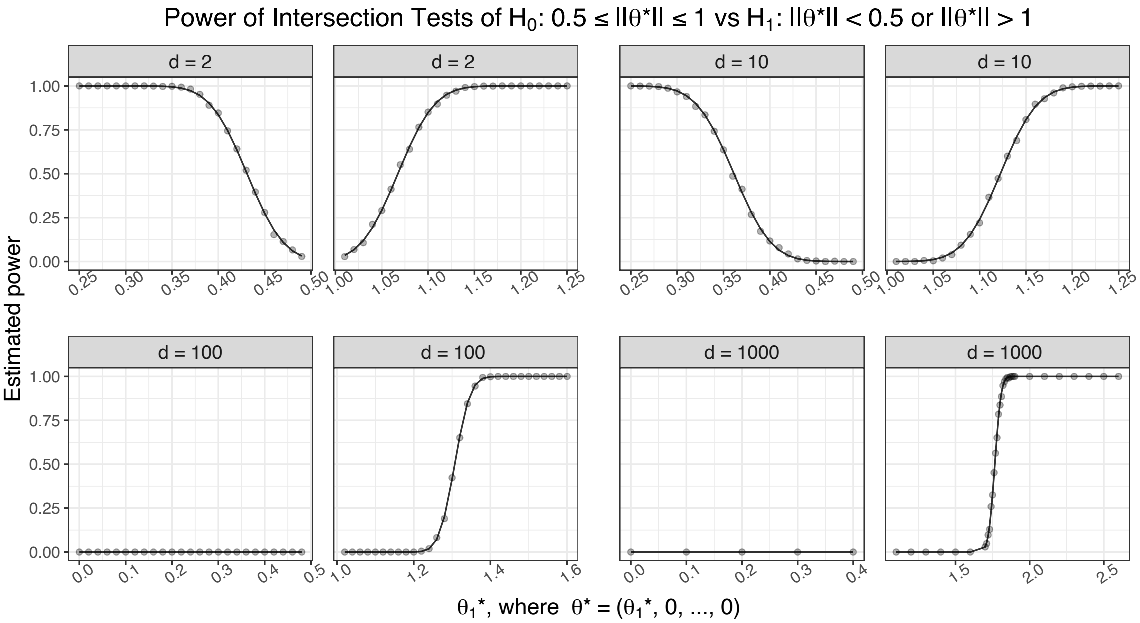


Figure B.3: Calculated power of $H_0 : \|\theta^*\| \in [0.5, 1.0]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1.0]$ using the intersection method. We compare the simulated power to the calculation (B.26) + (B.25). The points correspond to the simulated power, and the curves trace out the calculated power.

B.4.2 Cases of the Subsampled Hybrid LRT

The subsampled hybrid test of $H_0 : \|\theta^*\| \in [0.5, 1]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1]$ takes one of three approaches within each repeated subsample:

1. If $\|\overline{Y}_{1,b}\| < 0.5$, use the split LRT statistic U_n on the b^{th} subsample.

2. If $\|\overline{Y}_{1,b}\| \in [0.5, 1]$, set the b^{th} subsample's test statistic to 1.

3. If $\|\overline{Y}_{1,b}\| > 1$, use the RIPR LRT statistic R_n on the b^{th} subsample.

Figure B.4 shows the proportion of these three cases that make up the hybrid test. We consider all $\|\theta^*\|$ values from Fig. 2.7 of the main paper, as well as cases where $\|\theta^*\|$ is within the null region. At any given value of d and $\|\theta^*\|$, the three proportions sum to 1. Interestingly, although $\|\overline{Y}_{1,b}\| < 0.5$ approximately 95% of the time when $\|\theta^*\| = 0$ and $d = 100$, the hybrid test has approximately zero power at that choice of parameters. We derive this fact in section B.4.3. In addition, when $d = 1000$ we see that $\|\overline{Y}_{1,b}\| > 1$ in all simulations, even at $\theta^* = 0$. In section B.4.4, we see why this setting has approximately zero power as well.

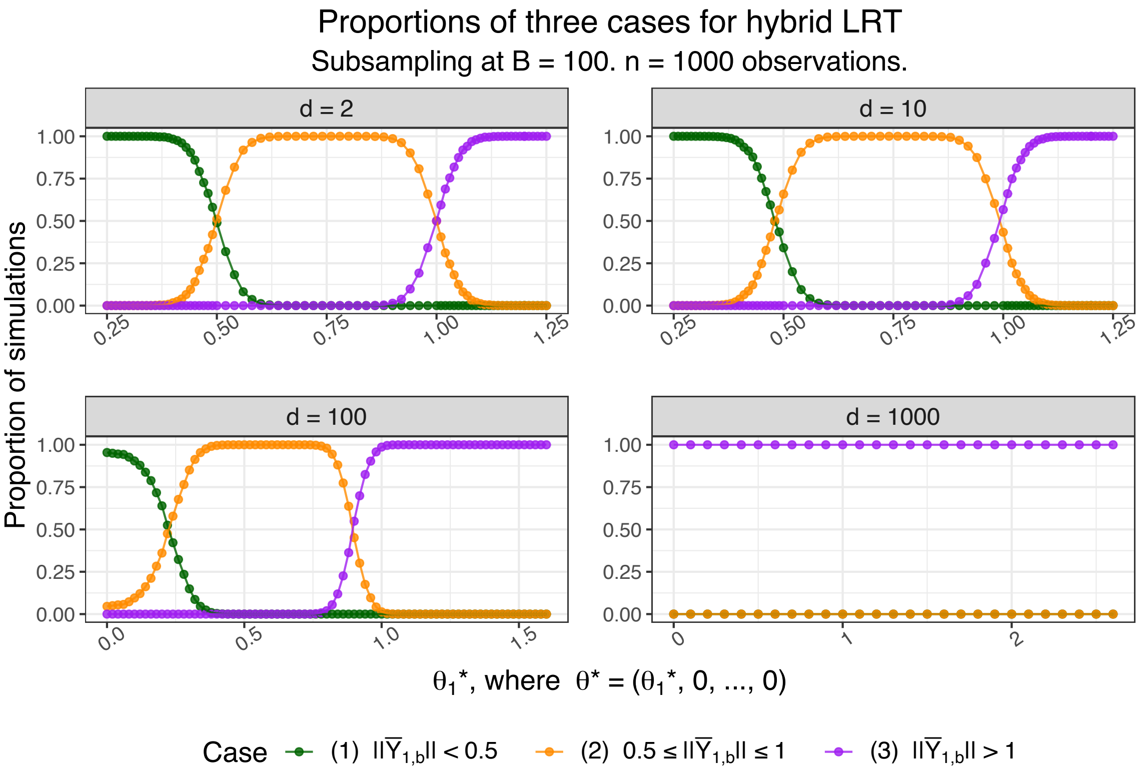


Figure B.4: Proportions of three cases that compose the hybrid LRT. We set $\alpha = 0.10$ and $n = 1000$, and we perform 1000 simulations at each value of $\|\theta^*\|$. We subsample $B = 100$ times.

B.4.3 Hybrid Power when $\theta^* = 0$, $d = 100$, and $n = 1000$

When $\theta^* = 0$, $d = 100$, and $n = 1000$, Fig. B.4 shows that $\|\overline{Y}_{1,b}\| < 0.5$ (case 1) occurs with probability of approximately 0.95, and $\|\overline{Y}_{1,b}\| \in [0.5, 1]$ (case 2) occurs with probability of approximately 0.05. At these parameters, the hybrid method has power of approximately 0, as shown in Fig. 2.7 in the main paper. We consider the power of the hybrid method at a single split of the data:

$$
\begin{aligned}
&\mathbb{P}_{\theta^*=0}(U_n \mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \geq 1/\alpha) \\
&= \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| < 0.5)}_{A_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - 0.5\overline{Y}_0/\|\overline{Y}_0\|\|^2\right) \geq \frac{1}{\alpha} \;\Big|\; \|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| < 0.5\right)}_{A_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| \in [0.5, 1])}_{B_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_0\|^2\right) \geq \frac{1}{\alpha} \;\Big|\; \|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| \in [0.5, 1]\right)}_{B_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| > 1)}_{C_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_0/\|\overline{Y}_0\|\|^2\right) \geq \frac{1}{\alpha} \;\Big|\; \|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| > 1\right)}_{C_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| \in [0.5, 1])}_{D_1} \underbrace{\mathbb{P}_{\theta^*=0}(1 \geq 1/\alpha \mid \|\overline{Y}_1\| \in [0.5, 1])}_{D_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| > 1)}_{E_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1/\|\overline{Y}_1\|\|^2\right) \geq \frac{1}{\alpha} \;\Big|\; \|\overline{Y}_1\| > 1\right)}_{E_2}.
\end{aligned}
$$

The probabilities B_2 and D_2 equal 0. In addition,

$$
\begin{aligned}
\mathbb{P}_{\theta^*=0}(\|\overline{Y}_0\| > 1) &= \mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| > 1) \\
&= \mathbb{P}_{\theta^*=0}\left(\frac{n}{2}\|\overline{Y}_1\|^2 > \frac{n}{2}\right) \\
&= \mathbb{P}(\chi^2_{df=100} > 1000/2) \\
&\approx 0.
\end{aligned}
$$

So C_1 and E_1 are also approximately 0. That means we only need to consider A_1A_2. It will be easier to work with the product of these two probabilities:

$$
\begin{aligned}
A_1A_2 &= \mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - 0.5\overline{Y}_0/\|\overline{Y}_0\|\|^2\right) \geq \frac{1}{\alpha},\ \|\overline{Y}_1\| < 0.5,\ \|\overline{Y}_0\| < 0.5\right) \\
&\leq \mathbb{P}_{\theta^*=0}\left(\|\overline{Y}_0 - \overline{Y}_1\|^2 < \|\overline{Y}_0 - 0.5\overline{Y}_0/\|\overline{Y}_0\|\|^2, \|\overline{Y}_0\| < 0.5\right) \\
&\leq \mathbb{P}_{\theta^*=0}(\|\overline{Y}_0 - \overline{Y}_1\|^2 < 0.25) \\
&= \mathbb{P}((4/n)\chi^2_{df=100} < 1/4) \\
&= \mathbb{P}(\chi^2_{df=100} < 1000/16) \\
&\approx 0.001.
\end{aligned}
$$

This means that at a single split of the data, the power at $\|\theta^*\| = 0$, $d = 100$, and $n = 1000$ is

$$\mathbb{P}_{\theta^*=0}(U_n \mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \geq 1/\alpha) \leq 0.001.$$

B.4.4 Hybrid Power when $\theta^* = 0$, $d = 1000$, and $n = 1000$

When $\theta^* = 0$, $d = 1000$, and $n = 1000$, we see that the hybrid method selects case 3 ($\|\overline{Y}_{1,b}\| > 1$) in all simulations. This is essentially choosing the wrong case, since $\|\theta^*\| = 0 < 0.5$. Numerically, we can show that the hybrid method will have power of approximately 0 at these parameters. Again, we consider a single split of the data.

$$\begin{aligned}
&\mathbb{P}_{\theta^*=0}(U_n \mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \geq 1/\alpha) \\
&\quad = \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5)}_{A_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(U_n \geq 1/\alpha \,\middle|\, \|\overline{Y}_1\| < 0.5\right)}_{A_2} + \\
&\qquad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| \in [0.5, 1])}_{B_1} \underbrace{\mathbb{P}_{\theta^*=0}(1 \geq 1/\alpha \mid \|\overline{Y}_1\| \in [0.5, 1])}_{B_2} + \\
&\qquad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| > 1)}_{C_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1/\|\overline{Y}_1\|\|^2\right) \geq \frac{1}{\alpha} \,\middle|\, \|\overline{Y}_1\| > 1\right)}_{C_2}.
\end{aligned}$$

The probability B_2 equals 0. In addition, A_1 is approximately 0 because

$$\begin{aligned}
\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5) &= \mathbb{P}_{\theta^*=0}\left((n/2)\|\overline{Y}_1\|^2 < n/8\right) \\
&= \mathbb{P}(\chi^2_{df=1000} < 1000/8) \\
&\approx 0.
\end{aligned}$$

So the probability of rejecting H_0 at this choice of parameters is approximately

$$\begin{aligned}
C_1C_2 &= \mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1/\|\overline{Y}_1\|\|^2\right)\geq\frac{1}{\alpha},\|\overline{Y}_1\|>1\right)\\
&\leq \mathbb{P}_{\theta^*=0}\left(\|\overline{Y}_0-\overline{Y}_1\|^2<\|\overline{Y}_0-\overline{Y}_1/\|\overline{Y}_1\|\|^2,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(\|\overline{Y}_0\|^2-2\overline{Y}_0^T\overline{Y}_1+\|\overline{Y}_1\|^2<\|\overline{Y}_0\|^2-2\overline{Y}_0^T\overline{Y}_1/\|\overline{Y}_1\|+1,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1(1/\|\overline{Y}_1\|-1)+\|\overline{Y}_1\|^2<1,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1(1-\|\overline{Y}_1\|)/\|\overline{Y}_1\|<1-\|\overline{Y}_1\|^2,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1(1-\|\overline{Y}_1\|)/\|\overline{Y}_1\|<(1-\|\overline{Y}_1\|)(1+\|\overline{Y}_1\|),\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1>\|\overline{Y}_1\|(1+\|\overline{Y}_1\|),\|\overline{Y}_1\|>1\right)\\
&\leq \mathbb{P}_{\theta^*=0}\left(\overline{Y}_0^T\overline{Y}_1>1\right).
\end{aligned}$$

Let $\sigma=1/\sqrt{500}$. Since $\overline{Y}_0$ and $\overline{Y}_1$ are averages of 500 $N(0,I_d)$ random variables, we see that $\overline{Y}_0\sim N(0,\sigma^2 I_d)$ and $\overline{Y}_1\sim N(0,\sigma^2 I_d)$. Let $\lambda=-d/2+(1/2)\sqrt{d^2+4/\sigma^4}$. (This choice of λ minimizes $\mathbb{E}[\exp(\lambda\overline{Y}_0^T\overline{Y}_1)]/\exp(\lambda)$ out of $\lambda>0$.) Let $\nu=\sigma/(1-\sigma^4\lambda^2)^{1/2}$. We derive

$$\begin{aligned}
&\mathbb{P}_{\theta^*=0}\left(\overline{Y}_0^T\overline{Y}_1>1\right)\\
&\quad= \mathbb{P}_{\theta^*=0}\left(\exp\left(\lambda\overline{Y}_0^T\overline{Y}_1\right)>\exp(\lambda)\right)\\
&\quad\leq \mathbb{E}_{\theta^*=0}\left[\exp\left(\lambda\overline{Y}_0^T\overline{Y}_1\right)\right]/\exp(\lambda)\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^d|\sigma^2 I_d|}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_0\|^2-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2+\lambda\overline{Y}_0^T\overline{Y}_1\right)d\overline{Y}_0 d\overline{Y}_1\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2\right)\left\{\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_0\|^2+\lambda\overline{Y}_0^T\overline{Y}_1\right)d\overline{Y}_0\right\}d\overline{Y}_1\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2\right)\left\{\mathbb{E}\left[\exp((\lambda\overline{Y}_1)^T\overline{Y}_0)\mid\overline{Y}_1\right]\right\}d\overline{Y}_1\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2\right)\exp\left(\frac{1}{2}\lambda^2\sigma^2\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}-\sigma^2\lambda^2\right)\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2}\left(\frac{1-\sigma^4\lambda^2}{\sigma^2}\right)\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&\quad= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2 I_d|^{1/2}}\exp\left(-\frac{1}{2\nu^2}\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&\quad= \exp(-\lambda)\frac{|\nu^2 I_d|^{1/2}}{|\sigma^2 I_d|^{1/2}}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\nu^2 I_d|^{1/2}}\exp\left(-\frac{1}{2\nu^2}\|\overline{Y}_1\|^2\right)d\overline{Y}_1
\end{aligned}$$

$$
\begin{aligned}
&= \exp(-\lambda)(\nu/\sigma)^d \\
&\approx \exp(-207)(1.1)^{1000} \\
&\approx 0.
\end{aligned}
$$

At a single split of the data, the power at $\|\theta^*\| = 0$, $d = 1000$, and $n = 1000$ is approximately 0 because

$$
\begin{aligned}
&\mathbb{P}_{\theta^*=0}(U_n \mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \geq 1/\alpha) \\
&\qquad \approx \mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1/\|\overline{Y}_1\|\|^2\right) \geq \frac{1}{\alpha}, \|\overline{Y}_1\| > 1\right) \\
&\qquad \leq \mathbb{P}_{\theta^*=0}\left(\overline{Y}_0^T \overline{Y}_1 > 1\right) \\
&\qquad \approx 0.
\end{aligned}
$$

Appendix C

Proofs from Chapter 3

Theorem 3.5.1. *We make five assumptions:*

1. *Suppose each* $\widehat{f}_{1,n} \in \mathcal{C}$*, where* $\mathcal{C}$ *is some (potentially nonparametric) class of functions that satisfies* $\sup_{f \in \mathcal{C}} |\widehat{D}_{KL}(f^* \| f) - D_{KL}(f^* \| f)| = O_P(n^{-\beta_1})$ *for some* $\beta_1 > 0$.
2. $D_{KL}(f^* \| \widehat{f}_{1,n}) = O_P(n^{-\beta_2})$ *for some* $0 < \beta_2 \leq 1/2$.
3. $\int_{\mathbb{R}^d} \|x\| f^*(x) dx < \infty$.

Suppose there is some set A *with* $P(A) = 1$ *(e.g., the support of* f^{LC} *or its interior) that satisfies the following:*

4. *For some* $\ell > 0$, $\inf_{x \in A} f^{LC}(x) \geq \ell$.
5. $\sup_{x \in A} |\widehat{f}_{0,n}(x) - f^{LC}(x)| \overset{a.s.}{\to} 0$.

Then $\lim_{n \to \infty} \mathbb{P}_{H_0}(T_n \geq 1/\alpha) = 1$.

Proof. We begin by separating T_n into a product of three components:

$$T_n = \underbrace{\left\{ \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{\widehat{f}_{1,n}(Y_i)}{f^*(Y_i)} \right\}}_{C_{1,n}} \underbrace{\left\{ \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^*(Y_i)}{f^{\mathrm{LC}}(Y_i)} \right\}}_{C_{2,n}} \underbrace{\left\{ \prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^{\mathrm{LC}}(Y_i)}{\widehat{f}_{0,n}(Y_i)} \right\}}_{C_{3,n}}.$$

Define ϵ as

$$\epsilon = \|(f^{\mathrm{LC}})^{1/2} - (f^*)^{1/2}\|_2 = \left[\int \left((f^{\mathrm{LC}})^{1/2} - (f^*)^{1/2} \right)^2 d\mu \right]^{1/2}.$$

We see that

$$\begin{aligned}
&\mathbb{P}(T_n < 1/\alpha) \\
&\quad \leq \mathbb{P}\left(C_{2,n} < \exp\left(\frac{n}{8}\epsilon^2\right) \cup C_{1,n} < \frac{1}{\alpha}\exp\left(-\frac{n}{16}\epsilon^2\right) \cup C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right)\right) \\
&\quad \leq \mathbb{P}\left(C_{2,n} < \exp\left(\frac{n}{8}\epsilon^2\right)\right) + \mathbb{P}\left(C_{1,n} < \frac{1}{\alpha}\exp\left(-\frac{n}{16}\epsilon^2\right)\right) + \mathbb{P}\left(C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right)\right).
\end{aligned}$$

We want to show that each of these three probabilities converge to 0.

By Lemma 1 of Wong et al. (1995), we can see

$$\begin{aligned}
\mathbb{P}\left(C_{2,n} < \exp\left(\frac{n}{8}\epsilon^2\right)\right) &= \mathbb{P}\left(\prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^*(Y_i)}{f^{\mathrm{LC}}(Y_i)} < \exp\left(\frac{n}{8}\epsilon^2\right)\right) \\
&= \mathbb{P}\left(\prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^{\mathrm{LC}}(Y_i)}{f^*(Y_i)} > \exp\left(-\frac{n}{8}\epsilon^2\right)\right) \\
&\leq \exp\left(-\frac{n}{8}\epsilon^2\right).
\end{aligned}$$

So $\lim_{n\to\infty} \mathbb{P}\left(C_{2,n} < \exp\left((n/8)\epsilon^2\right)\right) = 0$.

For the second probability, we use assumptions 1 and 2 to see that

$$\begin{aligned}
\log(C_{1,n}) &= \log\left(\prod_{Y_i \in \mathcal{D}_{0,n}} \frac{\widehat{f}_{1,n}(Y_i)}{f^*(Y_i)}\right) \\
&= \sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{1,n}(Y_i)}{f^*(Y_i)}\right) \\
&= -\frac{n}{2} \cdot \frac{2}{n} \sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{f^*(Y_i)}{\widehat{f}_{1,n}(Y_i)}\right) \\
&= -\frac{n}{2} \widehat{D}_{KL}(f^* \| \widehat{f}_{1,n}) \\
&= -\frac{n}{2}\left(\widehat{D}_{KL}(f^* \| \widehat{f}_{1,n}) - D_{KL}(f^* \| \widehat{f}_{1,n})\right) - \frac{n}{2} D_{KL}(f^* \| \widehat{f}_{1,n}) \\
&= -\frac{n}{2} O_P(n^{-\beta_1}) - \frac{n}{2} O_P(n^{-\beta_2}) \\
&= O_P(n^{1-\beta}),
\end{aligned}$$

where $\beta = \min\{\beta_1, \beta_2\} \in (0, 1/2]$. We fix $M > 0$ such that $\lim\limits_{n\to\infty} \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > M) = 0$. We see that

$$\begin{aligned}
&\mathbb{P}(C_{1,n} < (1/\alpha)\exp(-(n/16)\epsilon^2)) \\
&\quad = \mathbb{P}(\log(C_{1,n}) < \log(1/\alpha) - (n/16)\epsilon^2) \\
&\quad = \mathbb{P}(\log(C_{1,n}) < \log(1/\alpha) - (n/16)\epsilon^2, \log(C_{1,n}) \geq 0) + \\
&\qquad\quad \mathbb{P}(\log(C_{1,n}) < \log(1/\alpha) - (n/16)\epsilon^2, \log(C_{1,n}) < 0) \\
&\quad \leq \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) + \mathbb{P}(-|\log(C_{1,n})| < \log(1/\alpha) - (n/16)\epsilon^2) \\
&\quad = \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) + \mathbb{P}(|\log(C_{1,n})| > (n/16)\epsilon^2 - \log(1/\alpha)) \\
&\quad = \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) + \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > (n^{\beta}/16)\epsilon^2 - n^{\beta-1}\log(1/\alpha)) \\
&\quad = \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) + \\
&\qquad\quad \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > (n^{\beta}/16)\epsilon^2 - n^{\beta-1}\log(1/\alpha),\ (n^{\beta}/16)\epsilon^2 - n^{\beta-1}\log(1/\alpha) > M) + \\
&\qquad\quad \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > (n^{\beta}/16)\epsilon^2 - n^{\beta-1}\log(1/\alpha),\ (n^{\beta}/16)\epsilon^2 - n^{\beta-1}\log(1/\alpha) \leq M) \\
&\quad \leq \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) + \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > M) + \mathbb{P}((n^{\beta}/16)\epsilon^2 \leq n^{\beta-1}\log(1/\alpha) + M) \\
&\quad \leq \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) + \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > M) + \mathbb{P}((n^{\beta}/16)\epsilon^2 \leq \log(1/\alpha) + M).
\end{aligned}$$

We know that $\lim_{n\to\infty} \mathbb{P}(\log(1/\alpha) - (n/16)\epsilon^2 > 0) = 0$, $\lim_{n\to\infty} \mathbb{P}(|\log(C_{1,n})/n^{1-\beta}| > M) = 0$, and $\lim_{n\to\infty} \mathbb{P}((n^{\beta}/16)\epsilon^2 \leq \log(1/\alpha) + M) = 0$. We conclude that

$$\lim_{n\to\infty} \mathbb{P}(C_{1,n} < (1/\alpha)\exp(-(n/16)\epsilon^2)) = 0.$$

For the third probability, we see that

$$\begin{aligned}\mathbb{P}\left(C_{3,n} < \exp\left(-\frac{n}{16}\epsilon^2\right)\right) &= \mathbb{P}\left(\log\left(\prod_{Y_i \in \mathcal{D}_{0,n}} \frac{f^{\mathrm{LC}}(Y_i)}{\widehat{f}_{0,n}(Y_i)}\right) < -\frac{n}{16}\epsilon^2\right) \\ &= \mathbb{P}\left(\sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right) > \frac{n}{16}\epsilon^2\right) \\ &= \mathbb{P}\left(\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right) > \frac{1}{8}\epsilon^2\right) \\ &\leq \frac{8}{\epsilon^2}\mathbb{E}\left[\frac{2}{n}\sum_{Y_i \in \mathcal{D}_{0,n}} \log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right)\right].\end{aligned}$$

Note that Markov's inequality applies in the final step because $\widehat{f}_{0,n}$ and f^{LC} are both log-concave densities, and $\widehat{f}_{0,n}$ maximizes the likelihood over all log-concave densities on $\mathcal{D}_{0,n}$. This means that the quantity within the expectation is nonnegative. Using ℓ from assumption 4, fix $\gamma > 0$ such that $\gamma < \ell(\exp(\delta) - 1)$. Note that this implies $\log((\gamma + \ell)/\ell) < \delta$. Then

$$
\begin{aligned}
&\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right)\right]\\
&=\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i))\right]-\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(f^{\mathrm{LC}}(Y_i))\right]\\
&=\mathbb{E}\left[\left\{\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i))\right\}\left\{I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|<\gamma\right)+I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\right\}\right]-\\
&\qquad\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(f^{\mathrm{LC}}(Y_i))\right]\\
&=\mathbb{E}\left[\left\{\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i)-f^{\mathrm{LC}}(Y_i)+f^{\mathrm{LC}}(Y_i))\right\}I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|<\gamma\right)\right]+\\
&\qquad\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i))I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\right]-\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(f^{\mathrm{LC}}(Y_i))\right]\\
&<\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\gamma+f^{\mathrm{LC}}(Y_i))\right]-\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(f^{\mathrm{LC}}(Y_i))\right]+\\
&\qquad\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i))I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\right] \qquad\qquad (\mathrm{C}.1)\\
&\leq\log(\gamma+\ell)-\log(\ell)+\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i))I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\right] \qquad (\mathrm{C}.2)\\
&<\delta+\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log(\widehat{f}_{0,n}(Y_i))I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\right].
\end{aligned}
$$

Step (C.1) to step (C.2) holds because $g(x)=\log(\gamma+x)-\log(x)$ is a decreasing function, so $g(x)$ is maximized at the smallest x. By assumption 4, we know that with probability 1, $f^{\mathrm{LC}}(Y_i)\geq\ell$, so setting $x=\ell$ provides an upper bound.

By Lemma 3(a) of Cule et al. (2010a), assumption 3 implies that for some $u>0$, $\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^d}\widehat{f}_{0,n}(x)\leq u$ with probability 1. Using reverse Fatou's Lemma for conditional expectations in step (C.3), we see

$$
\begin{aligned}
&\limsup_{n\to\infty} \mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}} \log(\widehat{f}_{0,n}(Y_i)) I\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\right] \\
&= \limsup_{n\to\infty} \mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}} \log(\widehat{f}_{0,n}(Y_i)) \,\middle|\, \sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right] \mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right) \\
&\leq \limsup_{n\to\infty} \mathbb{E}\left[\sup_{x\in\mathbb{R}^d}(\log(\widehat{f}_{0,n}(x))) \,\middle|\, \sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right] \mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right) \\
&\leq \limsup_{n\to\infty} \mathbb{E}\left[\sup_{x\in\mathbb{R}^d}(\widehat{f}_{0,n}(x)) \,\middle|\, \sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right] \mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right) \\
&\leq \mathbb{E}\left[\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^d}(\widehat{f}_{0,n}(x)) \,\middle|\, \sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right] \mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right) \qquad \text{(C.3)} \\
&\leq u\mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right) \qquad \text{with probability 1.}
\end{aligned}
$$

Finally, by assumption 5, $\lim_{n\to\infty}\mathbb{P}\left(\sup_{x\in A}|\widehat{f}_{0,n}(x)-f^{\mathrm{LC}}(x)|\geq\gamma\right)\to 0$. So with probability 1, for arbitrary $\delta>0$,

$$
\lim_{n\to\infty}\mathbb{E}\left[\frac{2}{n}\sum_{Y_i\in\mathcal{D}_{0,n}}\log\left(\frac{\widehat{f}_{0,n}(Y_i)}{f^{\mathrm{LC}}(Y_i)}\right)\right]<\delta.
$$

We conclude that $\lim_{n\to\infty}\mathbb{P}\left(C_{3,n}<\exp\left(-\frac{n}{16}\epsilon^2\right)\right)=0$. Therefore, $\lim_{n\to\infty}\mathbb{P}(T_n<1/\alpha)=0$. □

Appendix D

Proofs from Chapter 4

We recall **Theorem 4.2.1**. Let $Y_1, \ldots, Y_n$ be iid observations from a distribution P where $Y_i \in \mathcal{Y}$, and let Y_{n+1} denote a new draw from P.

Theorem 4.2.1. *We define a prediction interval $C(\alpha) = [Y_{(r)}, Y_{(s)}]$, where $r = \lfloor (n+1)(\alpha/2) \rfloor$ and $s = \lceil (n+1)(1-\alpha/2) \rceil$. Then for every distribution P, $P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha$. This interval is bounded if $n \geq 2/\alpha - 1$.*

Proof of Theorem 4.2.1. We show that this method produces valid $100(1-\alpha)\%$ prediction intervals. Furthermore, if $n \geq 2/\alpha - 1$, we show that the lower and upper bounds of the interval will be finite.

Suppose the data arise from a continuous distribution such that ties occur with probability 0. (This is helpful for intuition, but the inequalities that follow are valid without this assumption.) We can define the following $n+1$ intervals:

$$\underbrace{(-\infty, Y_{(1)}]}_{A_0},\ \underbrace{(Y_{(1)}, Y_{(2)}]}_{A_1},\ \ldots,\ \underbrace{(Y_{(n-1)}, Y_{(n)}]}_{A_{n-1}},\ \underbrace{(Y_{(n)}, \infty)}_{A_n}.$$

A new observation $Y_{n+1} \sim P$ is equally likely to fall in any of those $n+1$ intervals. To see this, consider the augmented sample $(Y_1, Y_2, \ldots, Y_n, Y_{n+1})$ with updated order statistics $(Y'_{(1)}, Y'_{(2)}, \ldots, Y'_{(n)}, Y'_{(n+1)})$. The new observation Y_{n+1} is equally likely to be any of those order statistics. That means

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in A_0) &= \mathbb{P}(Y_{n+1} \leq Y_{(1)}) = \mathbb{P}(Y_{n+1} = Y'_{(1)}) = 1/(n+1), \\
\mathbb{P}(Y_{n+1} \in A_1) &= \mathbb{P}(Y_{(1)} < Y_{n+1} \leq Y_{(2)}) = \mathbb{P}(Y_{n+1} = Y'_{(2)}) = 1/(n+1),
\end{aligned}$$

and so forth. Allowing for ties, for $m \in \{1, \ldots, n\}$ we have

$$\begin{aligned}
\mathbb{P}(Y_{n+1} < Y_{(m)}) &\leq \frac{m}{n+1} \\
\mathbb{P}(Y_{n+1} > Y_{(m)}) &\leq \frac{n-m+1}{n+1}.
\end{aligned}$$

These inequalities are equalities when P is a continuous distribution.

We construct a prediction interval $[Y_{(r)}, Y_{(s)}]$ where $r = \lfloor (n+1)(\alpha/2) \rfloor$ and $s = \lceil (n+1)(1-\alpha/2) \rceil$. We see that

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \notin [Y_{(r)}, Y_{(s)}]) &= \mathbb{P}(Y_{n+1} < Y_{(r)}) + \mathbb{P}(Y_{n+1} > Y_{(s)}) \\
&\leq \frac{r}{n+1} + \frac{n-s+1}{n+1} \\
&= 1 + \frac{r-s}{n+1} \\
&= 1 + \frac{\lfloor (n+1)(\alpha/2) \rfloor - \lceil (n+1)(1-\alpha/2) \rceil}{n+1} \\
&= 1 + \frac{\lfloor (n+1)(\alpha/2) \rfloor - ((n+1) + \lceil -(n+1)(\alpha/2) \rceil)}{n+1} \\
&= \frac{\lfloor (n+1)(\alpha/2) \rfloor - \lceil -(n+1)(\alpha/2) \rceil}{n+1} \\
&= \frac{\lfloor (n+1)(\alpha/2) \rfloor + \lfloor (n+1)(\alpha/2) \rfloor}{n+1} \\
&\leq \frac{(n+1)\alpha}{n+1} \\
&= \alpha.
\end{aligned}$$

So $\mathbb{P}(Y_{n+1} \in [Y_{(r)}, Y_{(s)}]) \geq 1 - \alpha$.

If $n \geq 2/\alpha - 1$, then the lower and upper bounds will both be finite. Under this condition on n, the lower bound is greater than or equal to $Y_{(1)}$ because

$$\begin{aligned} \lfloor (n+1)(\alpha/2) \rfloor &\geq \lfloor (2/\alpha)\,(\alpha/2) \rfloor \\ &= 1. \end{aligned}$$

Applying this result, we also see that the upper bound is less than or equal to $Y_{(n)}$ because

$$\begin{aligned} \lceil (n+1)\,(1-\alpha/2) \rceil &= \lceil n+1-(n+1)(\alpha/2) \rceil \\ &= n+1+\lceil -(n+1)(\alpha/2) \rceil \\ &= n+1-\lfloor (n+1)(\alpha/2) \rfloor \\ &\leq n+1-1 \\ &= n. \end{aligned}$$

□

Appendix E

Proofs from Chapter 5

We recall **Theorem 5.2.1**. The data come in groups $\mathcal{D}_1, \ldots, \mathcal{D}_k$ and each group has iid data

$$\mathcal{D}_j = \{Y_{j1}, \ldots, Y_{jn_j}\} \sim P_j$$

where $P_1, \ldots, P_k \sim \Pi$. Assuming a new distribution $P_{k+1} \sim \Pi$ and $Y \sim P_{k+1}$, we want a prediction region for Y.

At the group level, let $C_j = [\ell_j, u_j]$ be the $100(1-\alpha/2)\%$ prediction set obtained by applying the method in Theorem 4.2.1 at level $\alpha/2$ to group j, $j = 1, \ldots, k$. Assume $C_{k+1} = [\ell_{k+1}, u_{k+1}]$ is the $100(1-\alpha/2)\%$ prediction set for group $k+1$ if we had n observations from P_{k+1}. At the distribution level, we want to construct an interval $C^{\mathrm{dbl}}(\alpha)$ such that $\Pi(C_{k+1} \subseteq C^{\mathrm{dbl}}(\alpha)) \geq 1 - \alpha/2$. Note that we have constructed a sample of k lower bounds $\{\ell_1, \ldots, \ell_k\}$ and k upper bounds $\{u_1, \ldots, u_k\}$. Using the order statistics from those samples, we set $C^{\mathrm{dbl}}(\alpha) = [\ell_{(r)}, u_{(s)}]$, where $r = \lfloor (k+1)(\alpha/4) \rfloor$ and $s = \lceil (k+1)(1-\alpha/4) \rceil$.

Theorem 5.2.1. *For $C^{dbl}(\alpha)$ as defined above, $\overline{\Pi}(Y \in C^{dbl}(\alpha)) \geq 1 - \alpha$. This set is bounded if $k \geq 4/\alpha - 1$ and $n \geq 4/\alpha - 1$.*

Proof of Theorem 5.2.1. For a new $Y \sim \overline{\Pi}$, we show that $\overline{\Pi}(Y \in C^{\mathrm{dbl}}(\alpha)) \geq 1 - \alpha$. Similar to the proof of Theorem 4.2.1, we can see that

$$\begin{aligned}\Pi(\ell_{k+1} < \ell_{(r)}) &\leq \lfloor (k+1)(\alpha/4) \rfloor \frac{1}{k+1} \\ &\leq \alpha/4\end{aligned}$$

and

$$\begin{aligned}\Pi(u_{k+1} > u_{(r)}) &\leq \frac{k - \lceil (k+1)(1-\alpha/4) \rceil + 1}{k+1} \\ &\leq \alpha/4.\end{aligned}$$

This implies that

$$\begin{aligned}\Pi(C_{k+1} \nsubseteq C^{\text{dbl}}(\alpha)) &= \Pi(\ell_{k+1} < \ell_{(r)} \cup u_{k+1} > u_{(s)}) \\ &\leq \Pi(\ell_{k+1} < \ell_{(r)}) + \Pi(u_{k+1} > u_{(s)}) \\ &\leq \alpha/4 + \alpha/4 \\ &= \alpha/2. \qquad \text{(E.1)}\end{aligned}$$

Let A denote the event that $C_{k+1} \subseteq C^{\text{dbl}}(\alpha)$. We can now show the main result:

$$\begin{aligned}\Pi(Y \notin C^{\text{dbl}}(\alpha)) &= \Pi(Y \notin C^{\text{dbl}}(\alpha), A) + \Pi(Y \notin C^{\text{dbl}}(\alpha), A^c) \qquad \text{(E.2)} \\ &\leq \Pi(Y \notin C_{k+1}) + \Pi(A^c) \qquad \text{(E.3)} \\ &\leq \alpha/2 + \alpha/2 \qquad \text{(E.4)} \\ &= \alpha.\end{aligned}$$

To get from (E.2) to (E.3), we note that $Y \notin C^{\text{dbl}}(\alpha)$ and $C_{k+1} \subseteq C^{\text{dbl}}(\alpha)$ implies that $Y \notin C_{k+1}$. The second two terms use the fact that $\Pi(Y \notin C^{\text{dbl}}(\alpha), A^c) \leq \Pi(A^c)$. To get from (E.3) to (E.4), the first probability uses the fact that $C_1, \ldots, C_{k+1}$ are $100(1-\alpha/2)\%$ prediction sets for groups $j = 1, \ldots, k+1$. The second probability holds because $\Pi(A^c) = \Pi(C_{k+1} \nsubseteq C^{\text{dbl}}(\alpha)) < \alpha/2$ from (E.1).

Furthermore, we can show that $C^{\text{dbl}}(\alpha)$ is bounded if $n \geq 4/\alpha - 1$ and $k \geq 4/\alpha - 1$. We use n observations to construct each of the $100(1-\alpha/2)\%$ intervals $C_1, \ldots, C_k$. Thus, by a similar argument as Theorem 4.2.1, these k intervals are all bounded if $n \geq 4/\alpha - 1$. In addition, since $C^{\text{dbl}}(\alpha)$ uses the $\lfloor (k+1)(\alpha/4) \rfloor$ order statistic of $\{\ell_1, \ldots, \ell_k\}$ and the $\lceil (k+1)(1-\alpha/4) \rceil$ order statistic of $\{u_1, \ldots, u_k\}$, $C^{\text{dbl}}(\alpha)$ is bounded if $k \geq 4/\alpha - 1$. □

The unsupervised pooling method referenced in **Theorem 5.2.2** pools the empirical CDFs across the k groups. For any group j with observations $Y_{j1}, \ldots, Y_{jn_j}$, the empirical CDF is defined as

$$\widehat{F}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(Y_{ji} \leq t).$$

We set

$$\widehat{q}_k(\alpha) = \inf \left\{ t \in \mathbb{R} : \frac{1}{k} \sum_{j=1}^{k} \widehat{F}_j(t) \geq \alpha \right\}.$$

Then $C^{\text{poolCDF}}(\alpha) = [\widehat{q}_k(\alpha/2), \widehat{q}_k(1 - \alpha/2)]$.

Theorem 5.2.2. *Assume that $F : \mathbb{R} \to [0, 1]$, defined as $F(t) = \overline{\Pi}(Y \leq t)$, is strictly increasing. For $C^{poolCDF}(\alpha)$ as defined above, $\overline{\Pi}(Y \in C^{poolCDF}(\alpha)) \to 1 - \alpha$ as $k \to \infty$.*

Proof. Let $\widehat{G}_k(t) = (1/k) \sum_{j=1}^{k} \widehat{F}_j(t)$. That means that for $\alpha \in (0, 1)$, the sample quantiles $\widehat{q}_k(\alpha)$ and the true quantiles $q(\alpha)$ are

$$\widehat{q}_k(\alpha) = \inf \left\{ t \in \mathbb{R} : \widehat{G}_k(t) \geq \alpha \right\}$$
$$q(\alpha) = \inf \left\{ t \in \mathbb{R} : F(t) \geq \alpha \right\}.$$

We prove this theorem in three steps:

1. For $t \in \mathbb{R}$, $\widehat{G}_k(t) \xrightarrow{p} F(t)$ as $k \to \infty$.

2. For $\alpha \in (0, 1)$, $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$ as $k \to \infty$.

3. For $Y \sim \Pi$, $\Pi\left(Y \in C^{\text{poolCDF}}(\alpha)\right) \to 1 - \alpha$ as $k \to \infty$.

Step 1. Fix $t \in \mathbb{R}$. We write

$$\widehat{G}_k(t) = \frac{1}{k} \sum_{j=1}^{k} \widehat{F}_j(t) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{i=1}^{n_j} I(Y_{ji} \leq t).$$

We see that $\mathbb{E}_{\Pi}[I(Y_{ji} \leq t)] = F(t)$, so

$$\mathbb{E}_{\Pi}[\widehat{G}_k(t)] = F(t).$$

In addition, note that the k distributions $P_1, \ldots, P_k$ are independently drawn from Π. Since $\widehat{F}_j(t)$ is bounded between $[0, 1]$, we determine

$$\mathrm{Var}_{\Pi}(\widehat{G}_k(t)) = \mathrm{Var}_{\Pi}\left(\frac{1}{k}\sum_{j=1}^{k}\widehat{F}_j(t)\right) = \frac{1}{k^2}\sum_{j=1}^{k}\mathrm{Var}_{\Pi}\left(\widehat{F}_j(t)\right) \le \frac{1}{k^2}\sum_{j=1}^{k}\frac{1}{4} = \frac{1}{4k} \to 0$$

as $k \to \infty$. We conclude that $\widehat{G}_k(t) \xrightarrow{p} F(t)$ as $k \to \infty$.

Step 2. Fix $\alpha \in (0, 1)$. Let $\epsilon > 0$ and $\delta > 0$. To show that $\lim_{k\to\infty} \Pi\left(|\widehat{q}_k(\alpha) - q(\alpha)| > \epsilon\right) = 0$, we will show that there exists $K \in \mathbb{N}$ such that for $k \ge K$, $\Pi\left(|\widehat{q}_k(\alpha) - q(\alpha)| > \epsilon\right) < \delta$.

We are assuming that $F(t)$ is strictly increasing. Thus, $F(q(\alpha) + \epsilon) - F(q(\alpha)) > 0$. Since $\widehat{G}_k(t) \xrightarrow{p} F(t)$ as $k \to \infty$, we can fix $K_1 \in \mathbb{N}$ such that for $k \ge K_1$,

$$\Pi\left(\left|\widehat{G}_k(q(\alpha) + \epsilon) - F(q(\alpha) + \epsilon)\right| > F(q(\alpha) + \epsilon) - F(q(\alpha))\right) < \frac{\delta}{2}.$$

In addition, $F(q(\alpha)) - F(q(\alpha) - \epsilon) > 0$. We can fix $K_2 \in \mathbb{N}$ such that for $k \ge K_2$,

$$\Pi\left(\left|\widehat{G}_k(q(\alpha) - \epsilon) - F(q(\alpha) - \epsilon)\right| \ge F(q(\alpha)) - F(q(\alpha) - \epsilon)\right) < \frac{\delta}{2}.$$

Now let $K = \max\{K_1, K_2\}$. Assume $k \ge K$.

From the definition of $\widehat{q}_k(\alpha)$, it holds that $\widehat{q}_k(\alpha) > q(\alpha) + \epsilon$ if and only if $\widehat{G}_k(q(\alpha) + \epsilon) < \alpha$. It also holds that $\widehat{q}_k(\alpha) < q(\alpha) - \epsilon$ if and only if $\widehat{G}_k(q(\alpha) - \epsilon) \geq \alpha$. We see

$$\begin{aligned}
\Pi\left(|\widehat{q}_k(\alpha) - q(\alpha)| > \epsilon\right) &= \Pi\left(\widehat{q}_k(\alpha) > q(\alpha) + \epsilon\right) + \Pi(\widehat{q}_k(\alpha) < q(\alpha) - \epsilon) \\
&= \Pi\left(\widehat{G}_k(q(\alpha) + \epsilon) < \alpha\right) + \Pi\left(\widehat{G}_k(q(\alpha) - \epsilon) \geq \alpha\right) \\
&= \Pi\left(\widehat{G}_k(q(\alpha) + \epsilon) < F(q(\alpha))\right) + \Pi\left(\widehat{G}_k(q(\alpha) - \epsilon) \geq F(q(\alpha))\right) \\
&= \Pi\left(\widehat{G}_k(q(\alpha) + \epsilon) < F(q(\alpha) + \epsilon) - (F(q(\alpha) + \epsilon) - F(q(\alpha)))\right) \\
&\quad + \Pi\left(\widehat{G}_k(q(\alpha) - \epsilon) \geq F(q(\alpha) - \epsilon) + (F(q(\alpha)) - F(q(\alpha) - \epsilon))\right) \\
&\leq \Pi\left(\left|\widehat{G}_k(q(\alpha) + \epsilon) - F(q(\alpha) + \epsilon)\right| > F(q(\alpha) + \epsilon) - F(q(\alpha))\right) \\
&\quad + \Pi\left(\left|\widehat{G}_k(q(\alpha) - \epsilon) - F(q(\alpha) - \epsilon)\right| \geq F(q(\alpha)) - F(q(\alpha) - \epsilon)\right) \\
&< \frac{\delta}{2} + \frac{\delta}{2} \\
&= \delta.
\end{aligned}$$

We conclude that $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$ as $k \to \infty$.

Step 3. Since $F(t)$ is a strictly increasing CDF, F is continuous. For Y randomly drawn from the superpopulation Π, we apply the continuous mapping theorem to see

$$\begin{aligned}
\lim_{k\to\infty} \Pi\left(Y \in C^{\text{poolCDF}}(\alpha)\right) &= \lim_{k\to\infty}\left[F(\widehat{q}_k(1 - \alpha/2)) - F(\widehat{q}_k(\alpha/2))\right] \\
&= F(q(1 - \alpha/2)) - F(q(\alpha/2)) \\
&= (1 - \alpha/2) - \alpha/2 \\
&= 1 - \alpha.
\end{aligned}$$

We conclude that $C^{\text{poolCDF}}(\alpha)$ is asymptotically valid as $k \to \infty$. □

We recall the setup for the supervised CDF pooling method referenced in **Theorem 5.2.3**. Let $[k] = \{1, \ldots, k\}$. We start by pooling the observations from some strict subset $k_0 \subset [k]$ of the k groups to fit a model $\widehat{\mu}(X)$ as an estimator of $\mathbb{E}[Y \mid X]$. We use the remaining groups to fit the residuals $R_{ji} = |Y_{ji} - \widehat{\mu}(X_{ji})|$, $j \in [k]\backslash k_0$, $i = 1, \ldots, n_j$. Now for each $j \in [k]\backslash k_0$, we define group

j's empirical CDF of the residuals

$$\widehat{F}_j(t) = \frac{1}{n_j}\sum_{i=1}^{n_j} I(R_{ji} \leq t).$$

We define

$$\widehat{q}_k(\alpha) = \inf\left\{t \in \mathbb{R} : \frac{1}{|[k]\backslash k_0|}\sum_{j\in[k]\backslash k_0} \widehat{F}_j(t) \geq \alpha\right\}.$$

The $1-\alpha$ conformal prediction set is $C^{\text{poolCDF}}(x;\alpha) = [\widehat{\mu}(x) - \widehat{q}_k(1-\alpha), \widehat{\mu}(x) + \widehat{q}_k(1-\alpha)]$.

Theorem 5.2.3. *Fit a model $\widehat{\mu}(X)$ as an estimator of $\mathbb{E}[Y \mid X]$ using the observations in groups $k_0 \subset [k]$. (Hence, this model stays fixed as k grows.) For $(X,Y) \sim \overline{\Pi}$, assume $\overline{\Pi}(|Y - \widehat{\mu}(X)| \leq t)$ is strictly increasing in t. Then $\overline{\Pi}(Y \in C^{poolCDF}(X;\alpha)) \xrightarrow{p} 1-\alpha$ as $k \to \infty$.*

Proof. The proof of Theorem 5.2.3 is similar to the proof of Theorem 5.2.2. We explain how to modify the argument to prove the supervised result. Let $\widehat{G}_k(t) = (|[k]\backslash k_0|)^{-1}\sum_{j\in[k]\backslash k_0}\widehat{F}_j(t)$. Let $F(t) = \Pi(|Y - \widehat{\mu}(X)| \leq t)$. For $\alpha \in (0,1)$, the sample quantiles $\widehat{q}_k(\alpha)$ and the true quantiles $q(\alpha)$ are

$$\widehat{q}_k(\alpha) = \inf\left\{t \in \mathbb{R} : \widehat{G}_k(t) \geq \alpha\right\}$$
$$q(\alpha) = \inf\left\{t \in \mathbb{R} : F(t) \geq \alpha\right\}.$$

Similar to Theorem 5.2.2, we prove this theorem in three steps:

1. For $t \in \mathbb{R}$, $\widehat{G}_k(t) \xrightarrow{p} F(t)$ as $k \to \infty$.
2. For $\alpha \in (0,1)$, $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$ as $k \to \infty$.
3. For $(X,Y) \sim \Pi$, $\Pi\left(Y \in C^{\text{poolCDF}}(X;\alpha)\right) \to 1-\alpha$ as $k \to \infty$.

Step 1. Fix $t \in \mathbb{R}$. We write

$$\widehat{G}_k(t) = \frac{1}{|[k]\backslash k_0|}\sum_{j\in[k]\backslash k_0}\widehat{F}_j(t) = \frac{1}{|[k]\backslash k_0|}\sum_{j\in[k]\backslash k_0}\frac{1}{n_j}\sum_{i=1}^{n_j} I(|Y_{ji} - \widehat{\mu}(X_{ji})| \leq t).$$

Again, we assume that $\widehat{\mu}(\cdot)$ is fixed, using the observations in the groups indexed by k_0. We see that $\mathbb{E}_\Pi[I(|Y_{ji} - \widehat{\mu}(X_{ji})| \leq t)] = \Pi(|Y - \widehat{\mu}(X)| \leq t)$, so

$$\mathbb{E}_\Pi[\widehat{G}_k(t)] = \Pi(|Y - \widehat{\mu}(X)| \leq t).$$

Since $\widehat{F}_j(t)$ is bounded between $[0,1]$, we see

$$\begin{aligned}\mathrm{Var}_\Pi(\widehat{G}_k(t)) &= \mathrm{Var}_\Pi\left(\frac{1}{|[k]\backslash k_0|}\sum_{j\in[k]\backslash k_0}\widehat{F}_j(t)\right) = \left(\frac{1}{|[k]\backslash k_0|}\right)^2\sum_{j\in[k]\backslash k_0}\mathrm{Var}_\Pi(\widehat{F}_j(t))\\ &\leq \frac{1}{4|[k]\backslash k_0|} \to 0\end{aligned}$$

as $k \to \infty$. We conclude that $\widehat{G}_k(t) \xrightarrow{p} F(t)$ as $k \to \infty$.

Step 2. We can show that for $\alpha \in (0,1)$, $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$ as $k \to \infty$ using the same steps as in the proof of Theorem 5.2.2. The only modification is that $\widehat{G}_k(t)$ and $F(t)$ have different definitions in the supervised case.

Step 3. Recall that $C^{\text{poolCDF}}(x;\alpha) = [\widehat{\mu}(x) - \widehat{q}_k(1-\alpha), \widehat{\mu}(x) + \widehat{q}_k(1-\alpha)]$. For (X,Y) randomly drawn from Π, we apply the continuous mapping theorem to see

$$\begin{aligned}\lim_{k\to\infty}\Pi\left(Y \in C^{\text{poolCDF}}(X;\alpha)\right) &= \lim_{k\to\infty}\Pi\left(|Y - \widehat{\mu}(X)| \leq \widehat{q}_k(1-\alpha)\right)\\ &= \Pi(|Y - \widehat{\mu}(X)| \leq q(1-\alpha))\\ &= 1 - \alpha.\end{aligned}$$

□

We recall the setup for the supervised parametric CDF pooling method referenced in **Theorem 5.2.4**. We also introduce some additional parameters for the proof. Suppose (X,Y) arise from a known parametric model $Y = \mu(X;\theta) + \epsilon$, where $\mu(\cdot)$ is known, θ is unknown, and ϵ has a zero-mean distribution. We pool all of the $m = \sum_{j=1}^k n_j$ observations to fit $\widehat{\theta}$, using the true parametric model $\mu(\cdot)$. Thus, at any X, our point prediction is $\widehat{Y} = \mu(X;\widehat{\theta})$. We have the following residuals under the true θ and the estimated $\widehat{\theta}$:

$$
\begin{aligned}
R_{ji}(\theta) &= |\mu(X_{ji};\theta) - Y_{ji}| \\
R_{ji}(\widehat{\theta}) &= \left|\mu(X_{ji};\widehat{\theta}) - Y_{ji}\right|.
\end{aligned}
$$

The empirical CDFs of these residuals are

$$
\begin{aligned}
\widehat{F}_{j,\theta}(t) &= \frac{1}{n_j}\sum_{i=1}^{n_j} I(R_{ji}(\theta) \leq t) \\
\widehat{F}_{j,\widehat{\theta}}(t) &= \frac{1}{n_j}\sum_{i=1}^{n_j} I(R_{ji}(\widehat{\theta}) \leq t).
\end{aligned}
$$

The true CDF of the residuals is

$$
F(t) = \Pi(|Y - \mu(X;\theta)| \leq t).
$$

We obtain sample quantiles $\widehat{q}_k(\widehat{\theta};\alpha)$ and true quantiles $q(\alpha)$:

$$
\begin{aligned}
\widehat{q}_k(\widehat{\theta};\alpha) &= \inf\left\{ t \in \mathbb{R} : \frac{1}{k}\sum_{j=1}^{k} \widehat{F}_{j;\widehat{\theta}}(t) \geq \alpha \right\} \\
q(\alpha) &= \inf\left\{ t \in \mathbb{R} : F(t) \geq \alpha \right\}.
\end{aligned}
$$

The $1-\alpha$ prediction set is $C^{\text{param}}(x;\alpha) = [\mu(x;\widehat{\theta}) - \widehat{q}_k(1-\alpha), \mu(x;\widehat{\theta}) + \widehat{q}_k(1-\alpha)]$.

Theorem 5.2.4. *Suppose (X,Y) arises from a parametric model $Y = \mu(X;\theta) + \epsilon$, where $\mu(\cdot)$ is known, θ is unknown, and ϵ has a zero-mean distribution. Assume $\overline{\Pi}(Y - \mu(X;\theta) \leq t)$ is strictly increasing in t. Assume the true θ satisfies*

$$
\frac{1}{k}\sum_{j=1}^{k} \sup_t |\widehat{F}_{j,\widehat{\theta}}(t) - \widehat{F}_{j,\theta}(t)| \xrightarrow{p} 0
$$

as $k \to \infty$, and assume that for $\delta > 0$, $\lim_{k\to\infty} \overline{\Pi}(|\mu(X;\theta) - \mu(X;\widehat{\theta})| > \delta) = 0$. For $C^{param}(x;\alpha)$ as defined above, $\lim_{k\to\infty} \overline{\Pi}(Y \in C^{param}(x;\alpha)) = 1 - \alpha$.

Proof. The proof of Theorem 5.2.4 is similar to the proof of Theorem 5.2.3. Define

$$\widehat{G}_{k,\widehat{\theta}}(t) = (1/k)\sum_{j=1}^{k} \widehat{F}_{j,\widehat{\theta}}(t)$$
$$\widehat{G}_{k,\theta}(t) = (1/k)\sum_{j=1}^{k} \widehat{F}_{j,\theta}(t)$$
$$F(t) = \Pi(|Y - \mu(X;\theta)| \leq t).$$

For $\alpha \in (0,1)$, the sample quantiles $\widehat{q}_k(\alpha)$ and the true quantiles $q(\alpha)$ are

$$\widehat{q}_k(\widehat{\theta};\alpha) = \inf\left\{t \in \mathbb{R} : \frac{1}{k}\sum_{j=1}^{k} \widehat{F}_{j;\widehat{\theta}}(t) \geq \alpha\right\}$$
$$q(\alpha) = \inf\left\{t \in \mathbb{R} : F(t) \geq \alpha\right\}.$$

Similar to Theorem 5.2.3, we prove this theorem in three steps:

1. For $t \in \mathbb{R}$, $\widehat{G}_{k,\widehat{\theta}}(t) \xrightarrow{p} F(t)$ as $k \to \infty$.

2. For $\alpha \in (0,1)$, $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$ as $k \to \infty$.

3. For $(X,Y) \sim \Pi$, $\Pi\left(Y \in C^{\text{param}}(X;\alpha)\right) \to 1-\alpha$ as $k \to \infty$.

Step 1. Fix $t \in \mathbb{R}$. By the assumption that $(1/k)\sum_{j=1}^{k} \sup_t |\widehat{F}_{j,\widehat{\theta}}(t) - \widehat{F}_{j,\theta}(t)| \xrightarrow{p} 0$ as $k \to \infty$, we know that

$$|\widehat{G}_{k,\widehat{\theta}}(t) - \widehat{G}_{k,\theta}(t)| \leq \frac{1}{k}\sum_{j=1}^{k} |\widehat{F}_{j,\widehat{\theta}}(t) - \widehat{F}_{j,\theta}(t)| \xrightarrow{p} 0.$$

Next, we write

$$\widehat{G}_{k,\theta}(t) = \frac{1}{k}\sum_{j=1}^{k} \widehat{F}_{j,\theta}(t) = \frac{1}{k}\sum_{j=1}^{k} \frac{1}{n_j}\sum_{i=1}^{n_j} I(|Y_{ji} - \mu(X;\theta)| \leq t).$$

We see that $\mathbb{E}_\Pi[I(|Y_{ji} - \mu(X;\theta)| \leq t)] = \Pi(|Y - \mu(X;\theta)| \leq t)$, so

$$E_\Pi[\widehat{G}_{k,\theta}(t)] = \Pi(|Y - \mu(X;\theta)| \leq t) = F(t).$$

Since $\widehat{F}_{j,\theta}(t)$ is bounded between $[0,1]$, we see

$$\mathrm{Var}_{\Pi}(\widehat{G}_{k,\theta}(t)) = \mathrm{Var}_{\Pi}\left(\frac{1}{k}\sum_{j=1}^{k}\widehat{F}_{j,\theta}(t)\right) = \frac{1}{k^2}\sum_{j=1}^{k}\mathrm{Var}_{\Pi}\left(\widehat{F}_{j,\theta}(t)\right) \leq \frac{1}{4k} \to 0.$$

That means that $\widehat{G}_{k,\theta}(t) \xrightarrow{p} F(t)$ as $k \to \infty$. Combining these two convergence statements,

$$|\widehat{G}_{k,\widehat{\theta}}(t) - F(t)| = |\widehat{G}_{k,\widehat{\theta}}(t) - \widehat{G}_{k,\theta}(t) + \widehat{G}_{k,\theta}(t) - F(t)| \leq |\widehat{G}_{k,\widehat{\theta}}(t) - \widehat{G}_{k,\theta}(t)| + |\widehat{G}_{k,\theta}(t) - F(t)| \xrightarrow{p} 0.$$

We conclude that $\widehat{G}_{k,\widehat{\theta}}(t) \xrightarrow{p} F(t)$ as $k \to \infty$.

Step 2. We can show that for $\alpha \in (0,1)$, $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$ as $k \to \infty$ using the same steps as in the proof of Theorem 5.2.2. As modifications, we replace $\widehat{G}_k(t)$ with $\widehat{G}_{k,\widehat{\theta}}(t)$, and we use $F(t) = \Pi(|Y - \mu(X;\theta)| \leq t)$. In addition, Theorem 5.2.2 uses the fact that $F(t)$ is strictly increasing. We are assuming that $\Pi(Y - \mu(X;\widehat{\theta}) \leq t)$ is strictly increasing, which implies that $F(t) = \Pi(|Y - \mu(X;\widehat{\theta})| \leq t)$ is strictly increasing.

Step 3. Recall that $C^{\mathrm{param}}(x;\alpha) = [\mu(x;\widehat{\theta}) - \widehat{q}_k(1-\alpha), \mu(x;\widehat{\theta}) + \widehat{q}_k(1-\alpha)]$. We are assuming that $\Pi(Y - \mu(X;\theta) \leq t)$ is strictly increasing in t, so $\Pi(Y - \mu(X;\theta) \leq t)$ is continuous. We have also shown that $\widehat{q}_k(\alpha) \xrightarrow{p} q(\alpha)$, and we are assuming that $|\mu(X;\widehat{\theta}) - \mu(X;\theta)| \xrightarrow{p} 0$ over Π. For

(X, Y) randomly drawn from Π, we apply the continuous mapping theorem to see

$$\begin{aligned}
&\lim_{k\to\infty} \Pi\left(Y \in C^{\text{param}}(X;\alpha)\right) \\
&\qquad = \lim_{k\to\infty} \Pi\left(|Y - \mu(X;\widehat{\theta})| \le \widehat{q}_k(1-\alpha)\right) \\
&\qquad = \lim_{k\to\infty} \Pi\left(-\widehat{q}_k(1-\alpha) \le Y - \mu(X;\widehat{\theta}) \le \widehat{q}_k(1-\alpha)\right) \\
&\qquad = \lim_{k\to\infty} \Pi\left(-\widehat{q}_k(1-\alpha) + \mu(X;\widehat{\theta}) - \mu(X;\theta) \le Y - \mu(X;\theta) \le \widehat{q}_k(1-\alpha) + \mu(X;\widehat{\theta}) - \mu(X;\theta)\right) \\
&\qquad = \lim_{k\to\infty} \Pi\left(Y - \mu(X;\theta) \le \widehat{q}_k(1-\alpha) + \mu(X;\widehat{\theta}) - \mu(X;\theta)\right) - \\
&\qquad\qquad \lim_{k\to\infty} \Pi\left(Y - \mu(X;\theta) \le -\widehat{q}_k(1-\alpha) + \mu(X;\widehat{\theta}) - \mu(X;\theta)\right) \\
&\qquad = \Pi\left(Y - \mu(X;\theta) \le q(1-\alpha)\right) - \Pi\left(Y - \mu(X;\theta) \le -q(1-\alpha)\right) \\
&\qquad = \Pi\left(-q(1-\alpha) \le Y - \mu(X;\theta) \le q(1-\alpha)\right) \\
&\qquad = \Pi\left(|Y - \mu(X;\theta)| \le q(1-\alpha)\right) \\
&\qquad = 1 - \alpha.
\end{aligned}$$

□