# TOPICS IN TWO-SAMPLE TESTING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nelson C. Ray
2013

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Susan P. Holmes) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Persi W. Diaconis)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Bradley Efron)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jerome H. Friedman)

Approved for the University Committee on Graduate Studies

# Abstract

Driven by recent advances in the collection of biological data, many such studies draw from heterogeneous datasources. We develop an idea of Jerome Friedman's to conduct two-sample testing using supervised learning procedures. In special cases, this technique generalizes the randomization $t$-test, for which an asymptotic normality result is known. Using Stein's method of exchangeable pairs, we produce Berry–Esseen-type bounds for the permutation $t$-statistic for the purpose of statistical inference. We demonstrate the use of kernel methods in two-sample testing on non-vectorial data (text and images), and apply multiple kernel learning (MKL) to the heterogeneous data domain. We show that these techniques can effectively synthesize signals from multiple datasources and produce interpretable weights that highlight the role of each component.

# Acknowledgments

This work would not have been possible without the teachings, contributions, and support from many parties. I would like to thank:

- my adviser, Susan Holmes, for her endless encouragement, keen advice, and for equipping me with all that was necessary to succeed;
- my academic benefactors, Gunnar Carlsson, Persi Diaconis, Brad Efron, Jerry Friedman, and Sourav Chatterjee for valuable feedback;
- my friends, especially Kelly, Noah, Leo, Dennis, Omkar, and Zhen for their fellowship;
- my family, for their unwavering backing and love;
- the entire Stanford Statistics Department, for providing a comfortable academic environment in which to grow.

This work was partially supported by a VIGRE fellowship from the National Science Foundation under Grant DMS-0502385.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The two-sample testing problem arises whenever one collects data for two distinct groups and wishes to know whether there are meaningful differences between such groups: could they have been sampled from the same underlying distribution? Throughout statistical history, many two-sample tests have been developed to address a multitude of scenarios. However, there is a dearth of scholarship in the heterogeneous data setting, where a single sample is comprised of data from different domains. For instance, samples may comprise vectorial measurements and images, coupled with graph structures.

We develop Jerome Friedman's method of conducting two-sample tests by leveraging techniques from regression and classification. In particular, we construct two-sample tests for non-vectorial data by way of kernel support vector machines. We extend these tests to deal with heterogeneous data by using multiple kernel learning to create an optimal single kernel as a convex combination of various input kernels. We show that these tests generalize the two-sample permutation $t$-test, for which asymptotic normality results are known but rate of convergence bounds remain lacking. Using Stein's method of exchangeable pairs, we produce Berry–Esseen-type bounds that allow us to quantify our error in approximating the permutation distribution of this statistic with the normal distribution.

Chapter 2 gives an overview of Stein's method, with an emphasis on exchangeable pairs. We review a similar CLT with dependence, Hoeffding's Combinatorial CLT, as

our proof proceeds in a similar fashion. Ancillary results can be found in Appendix A, generalizations of previous work on Stein's method in Appendix B, and proofs of the main propositions contributing to our rate of convergence bounds are in Appendix C

In Chapter 3, we review population- and permutation-based inference and prior related work on distributional approximations for the $t$-statistic, as well as its connection with self-normalized sums. We produce a conservative bound on rate of convergence of the permutation $t$-distribution to the standard normal distribution, in addition to a Berry–Esseen-type bound given an additional boundedness condition, appealing to some results that are proved in Appendix C.

Chapter 4 is a computational companion to Chapter 3. We collate the results from numerous computer simulations to experimentally verify the theoretical results and inform the development of further theory. We also describe some novel computational procedures in order to effect simulation over a wide range of conditions.

We introduce Friedman's idea for two-sample testing in Chapter 5 and apply it using kernel support vector machines, developing two-sample tests for situations for which kernel functions exist. We show that these tests generalize the permutation $t$-test and compare their performance against an alternative, kernel-based test utilizing the maximum mean discrepancy statistic in vectorial and non-vectorial settings.

Lastly, in Chapter 6, we construct two-sample tests on heterogeneous data using multiple kernel learning to identify the most salient discriminating features of the data. We demonstrate their efficacy and informativeness on wine data and are able to discriminate between two popular wine varietals by reliably fusing data from different domains.

# Chapter 2

# Stein's method

In this chapter, we present an introduction to Stein's method of exchangeable pairs, which we use to prove the core theoretical result of this thesis: a rate of convergence bound on the Kolmogorov distance between the randomization distribution of the $t$-statistic and the standard normal distribution.

Due to similarities between this problem and Hoeffding's combinatorial central limit theorem (HCCLT), we first review the HCCLT and related results.

## 2.1 Introduction

Stein's method provides a means of bounding the distance between two probability distributions in a given probability metric. When applied with the normal distribution as the target, this results in central-limit-type theorems. Traditional results of this type typically make use of the complex-valued characteristic function. For an account of the early history and many variations of the Central Limit Theorem, see Le Cam [51]. In [88], Stein introduced a groundbreaking approach with a real characterizing equation in order to more effectively tame various forms of dependence.

He soon refined these initial ideas into his monograph [89], which includes the Poisson approximation work of Chen [13], random graph results of Barbour and Eagleson [4], and binomial approximation of Diaconis [20]. Several flavors of Stein's method (e.g., the method of exchangeable pairs) proceed via auxiliary randomization.

Other compilations with many illuminating applications include those of Diaconis and Holmes [23], Barbour and Chen [2, 3], and Chen, Goldstein, and Shao [14].

It will be instructive to follow the proof of the HCCLT in [89] because our proof proceeds in a similar fashion but with the following generalizations: an approximate contraction property, less cancellation of terms due to separate estimation of various denominators, and non-unit variance of the random variable of interest. In Appendix B, we prove various generalizations of theorems in [14]. We also refer to auxiliary results in Appendix A.

## 2.2 Stein's Theorem

Theorem 2.1 bounds the Kolmogorov distance between the distribution of the random variable $W$ and the standard normal distribution in terms of functions of the difference of the exchangeable pair $(W, W')$. It is applied to the situation where $W$ is the sum of the random diagonal of a matrix to prove the HCCLT. Chen et al. [14] generalize Theorem 2.1 to allow for situations in which the regression condition does not hold exactly, and we later in addition relax the assumption that $W$ has unit variance.

**Theorem 2.1** (Stein)**.** *If $W$, $W'$ are mean 0, exchangeable random variables with variance 1 satisfying the exact regression condition*

$$\mathbb{E}[W' - W|W] = -\lambda W$$

*for some $\lambda \in (0, 1)$, then*

$$\sup_{w\in\mathbb{R}} |P(W \le w) - \Phi(w)| \le 2\sqrt{\mathbb{E}\left[1 - \frac{1}{2\lambda}E[(W'-W)^2|W]\right]} + (2\pi)^{-1/4}\sqrt{\frac{1}{\lambda}\mathbb{E}|W'-W|^3}$$
$$\le 2\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} + (2\pi)^{-1/4}\sqrt{\frac{1}{\lambda}\mathbb{E}|W'-W|^3}.$$

## 2.3 Hoeffding's Combinatorial CLT

Stein's proof of the HCCLT relies on an application of Theorem 2.1.

**Theorem 2.2** (Hoeffding's Combinatorial Central Limit Theorem)**.** *Let* $\{a_{ij}\}_{i,j}$ *be an* $n \times n$ *matrix of real-valued entries that is row- and column-centered and scaled such that the sums of the squares of its elements equals* $n-1$*:*

$$\sum_{j=1}^{n} a_{ij} = 0$$
$$\sum_{i=1}^{n} a_{ij} = 0$$
$$\sum_{i=1,j=1}^{n} a_{ij}^2 = n-1$$

*Let* $\Pi$ *be a random permutation of* $\{1, \dots, n\}$ *drawn uniformly at random from the set of all permutations:*

$$P(\Pi = \pi) = \frac{1}{n!}.$$

*Define*

$$W = \sum_{i=1}^{n} a_{i\Pi(i)}$$

*to be the sum of a random diagonal. Then*

$$|P(W \leq w) - \Phi(w)| \leq \frac{C}{\sqrt{n}} \left[ \sqrt{\sum_{i,j=1}^{n} a_{ij}^4} + \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^3} \right].$$

Originally proved by Hoeffding using the method of moments [42], Theorem 2.2 generalizes the non-parametric measure of rank correlation known as Spearman's footrule [87] [73]

$$D(\pi, \sigma) = \sum_{i=1}^{n} |\pi(i) - \sigma(i)|,$$

where $\pi$ and $\sigma$ are elements of $S_n$, the set of permutations of $n$ letters. Considering $D$ as a metric on $S_n$, Diaconis and Graham [21] related it to other common non-parametric measures and derived various asymptotic results. In fact, Spearman's footrule has found contemporary applications such as in ranking search queries and microarray experiments [78], where consistency of the top $k$ out of $n$ objects is desired.

Theorem 2.2 was extended by Motoo [61] and Schneller [74], generalizing the work

on the permutation-based tests of Wald and Wolfowitz [96] and Noether [64]. These results build further on the problem of devising exact tests of significance when the underlying probability distribution is unknown, originating from Fisher [28, 30].

Given fixed data $a_{ij}$, the HCCLT provides a bound in terms of a universal constant $C$, a sample-size-dependent term $\frac{1}{\sqrt{n}}$, and a function of the data $\sqrt{\sum_{i,j=1}^n a_{ij}^4} + \sqrt{\sum_{i,j=1}^n |a_{ij}|^3}$. Thus, given $C$, we can calculate an explicit bound on the Kolmogorov distance.

Consider a sequence of matrices, $a_{ij}^{(n)}$, of dimension $n \times n$. In order to achieve an $\mathcal{O}(n^{-1/2})$ rate of convergence, we require that the function of the data be bounded. However, this will not typically be the case.

Bolthausen [8] proved the following result:

**Theorem 2.3** (Bolthausen)**.** *Under the conditions of Theorem 2.2, there is an absolute constant $K > 0$ such that*

$$|P(W \leq w) - \Phi(w)| \leq K \frac{\sqrt{\sum_{i,j=1}^n |a_{ij}|^3}}{n}.$$

Then, given the sequence of matrices $a_{ij}^{(n)}$, the theorem yields a convergence rate of $\mathcal{O}(n^{-1/2})$ as long as $\sqrt{\sum_{i,j=1}^n |a_{ij}|^3}/\sqrt{n}$ remains bounded.

von Bahr [93] and Ho and Chen [41] also obtained the same rate under some boundedness conditions on random-valued entries of the matrix. For example, von Bahr obtained the following result:

**Theorem 2.4** (von Bahr)**.** *Let $\{X_{ij}\}_{i,j=1}^n$ be a square matrix of random variables with independent row vectors. With $\mu(i,j) = \mathbb{E}X_{ij}$, define*

$$\bar{\mu}(i,\cdot) = n^{-1} \sum_j \mu(i,j), \quad \bar{\mu}(\cdot,j) = n^{-1} \sum_i \mu(i,j), \quad \bar{\mu}(\cdot,\cdot) = n^{-2} \sum_i \sum_j \mu(i,j).$$

*Set $Y_{ij} = X_{ij} - \bar{\mu}(i,\cdot) - \bar{\mu}(\cdot,j) + \bar{\mu}(\cdot,\cdot)$, $\tau_{ij}^2 = \mathbb{E}Y_{ij}^2$, and $\tau^2 = n^{-2} \sum_{i,j} \tau_{ij}^2$.*

*Let $\Pi$ be a uniformly at random permutation, independent of $X_{ij}$, and $W = \sum_{i=1}^n X_{i\Pi(i)}$ be the sum of a random diagonal. Then there exists an absolute constant*

*$C$ such that for all $w \in \mathbb{R}$,*

$$\left| P\left( \frac{W - \mu_W}{\sigma_W} \leq w \right) - \Phi(w) \right| \leq \frac{C}{\sqrt{n}} \gamma,$$

*where $\mu_W$ denotes the mean of $W$, $\sigma_W$ its standard deviation, $\Phi(w)$ the standard normal CDF, and*

$$\gamma = \max \left[ \max_{i,j} \mathbb{E}|Y_{ij}|/\tau, \max_i \sum_j \mathbb{E}Y_{ij}^2/n\tau^2, \max_j \sum_i \mathbb{E}Y_{ij}^2/n\tau^2, \max_{i,j} \sum_{i,j} \mathbb{E}|Y_{ij}|^3/n^2\tau^3 \right].$$

Now, we return to generalizing Theorem 2.1.

## 2.4 Generalized Stein's Theorems

Here, we treat the situation where the regression condition fails to hold exactly. Chen et al. [14] serves as an excellent reference for results of this type.

**Definition 2.5** (Approximate Stein Pair)**.** *Let $(W, W')$ be an exchangeable pair. If the pair satisfies the "approximate linear regression condition"*

$$\mathbb{E}[W - W'|W] = \lambda(W - R), \tag{2.1}$$

*where $R$ is a random variable and $\lambda \in (0, 1)$, then we call $(W, W')$ an approximate Stein pair.*

Typically, $R$ must be of "small order" for the resulting bounds to be meaningful due to a resulting $\mathbb{E}|R|$ term such as in Theorems 2.7 and 2.8.

Here we generalize Lemma 5.1 from [14] to the setting of non-unit variance:

**Theorem 2.6.** *If $W$, $W'$ are mean 0 exchangeable random variables with variance $\mathbb{E}W^2$ satisfying*

$$\mathbb{E}[W' - W|W] = -\lambda(W - R)$$

*for some* $\lambda \in (0,1)$ *and some random variable* $R$*, then for any* $z \in \mathbb{R}$ *and* $a > 0$*,*

$$\mathbb{E}[(W'-W)^2 \mathbf{1}_{\{-a \le W'-W \le 0\}} \mathbf{1}_{\{z-a \le W \le z\}}] \le 3a\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|)$$

*and*

$$\mathbb{E}[(W'-W)^2 \mathbf{1}_{\{0 \le W'-W \le a\}} \mathbf{1}_{\{z-a \le W \le z\}}] \le 3a\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|).$$

The following theorem will allow us to prove an $\mathcal{O}(n^{-1/4})$ rate under mild conditions.

Here, we generalize Theorem 5.5 from [14]:

**Theorem 2.7.** *If* $W$*,* $W'$ *are mean 0 exchangeable random variables with variance* $\mathbb{E}W^2$ *satisfying*

$$\mathbb{E}[W'-W|W] = -\lambda(W-R)$$

*for some* $\lambda \in (0,1)$ *and some random variable* $R$*, then*

$$\sup_{z \in \mathbb{R}} |P(W \le z) - \Phi(z)| \le (2\pi)^{-1/4} \sqrt{\frac{\mathbb{E}|W'-W|^3}{\lambda}} + \frac{1}{2\lambda} \sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} + |\mathbb{E}W^2 - 1| + \mathbb{E}|WR| + \mathbb{E}|R|$$

The following result will let us achieve a rate of $\mathcal{O}(n^{-1/2})$ subject to the difference $|W'-W|$ being bounded. This condition has been similarly used by Rinott and Rotar [72] and Shao and Su [82]. It generalizes part of Theorem 5.3 from [14]:

**Theorem 2.8.** *If* $W$*,* $W'$ *are mean 0 exchangeable random variables with variance* $\mathbb{E}W^2$ *satisfying*

$$\mathbb{E}[W'-W|W] = -\lambda(W-R)$$

*for some* $\lambda \in (0,1)$ *and some random variable* $R$ *and* $|W'-W| \le \delta$*, then*

$$\sup_{z \in \mathbb{R}} |P(W \le z) - \Phi(z)| \le \frac{.41\delta^3}{\lambda} + 3\delta(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) + \frac{1}{2\lambda} \sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} + |\mathbb{E}W^2 - 1| + \mathbb{E}|WR| + \mathbb{E}|R|$$

# Chapter 3

# Kolmogorov Distance Bounds

In this chapter, we prove the core theoretical results of this thesis: rate of convergence bounds on the Kolmogorov distance between the randomization distribution of the $t$-statistic and the standard normal distribution, using Theorems 2.7 and 2.8 of Chapter 2.

## 3.1 Motivation

Motivated by concerns regarding normality assumptions in the hypothesis being tested, Fisher [28] proposed a nonparametric randomization test. Also known as a permutation test, Fisher applied this novel test to Charles Darwin's *Zea mays* data and noted that the achieved significance level was very similar to that observed in the parametric test. Indeed, Diaconis and Holmes [22] used efficient Gray-code-based calculations and Berry–Esseen bounds to show that the randomization distribution is well approximated by a normal distribution. For more history on the development of randomization procedures, see Zabell [97] or David [17]. Diaconis and Lehmann [24] in their comment on Zabell's paper further expanded on some properties of these randomization tests. Freedman and Lane [31] expounded an alternative interpretation of reported significance levels that is similar to other permutation-based approaches.

The $t$-test is one of the most-studied statistical procedures. For a good survey, see Zabell [97]. The $t$-test also possesses various robustness properties. Efron [27] studied

the distribution of the $t$-statistic under departures from normality; Logan, Mallows, Rice, and Shepp [58] suggested an asymptotic normality result; and Giné, Götze, and Mason [35] proved a necessary and sufficient condition for such a result.

Ludbrook and Dudley [60] have written about the advantages of permutation tests, especially in biomedical research, and outlined two models of statistical inference: the so-called population model, formally introduced by Neyman and Pearson [63], and Fisher's randomization model [28].

Neyman and Pearson's population model, formally proposed in 1928 [63], assumes that there has been a random sampling from a population or populations, on which a statistical test has been performed. The level of statistical significance corresponds to rejections of the null hypothesis under repeated random samplings from these populations. Because it is typically unfeasible to repeatedly sample, the sampling distribution is taken to conform to a theoretical distribution such as the $t$ or $F$ distributions. Ludbrook and Dudley argued that the preference for controlling Type I error rates in biomedical research is due to the strong desire to avoid introducing valueless new therapies. Neyman [62] also first proposed using confidence intervals as an alternative to hypothesis testing.

By contrast, Fisher's randomization model [28] did not involve sampling from a population. A fixed sample of experimental units is divided into groups, for which a test statistic is calculated. The unique sampling distribution corresponding to the fixed sample is compiled exactly by permutation. The population model has implications for the outcome of statistical tests on future samples, whereas the permutation-based test possesses no such guarantees. For a history of Neyman and Fisher's contributions to statistics, see Lehmann [53].

Under the randomization model and using the language of triangular arrays, Lehmann [52] proved a weak convergence result of the randomization distribution of the $t$-statistic to the standard normal distribution, however, at present there is no known Berry–Esseen-type bound for this rate of convergence.

Bentkus and Götze [6] first obtained a Berry–Esseen bound for the one-sample $t$-statistic under the population model. Up to an unspecified absolute constant, their result coincides with the classical Berry–Esseen bound for the mean, given a bound

on the third moment. In this setting, Shao [81] followed up with a similar result with a specified constant. Given sampling from a finite population, Bloznelis [7] also produced a Berry–Esseen bound for the one sample $t$-statistic, with additional results from Pinelis [66].

We use Stein's method of exchangeable pairs to prove a conservative bound of $\mathcal{O}(n^{-1/4})$ on the rate of convergence of the randomization $t$-distribution to the standard normal distribution. With an additional condition on the data, we are able to obtain an $\mathcal{O}(n^{-1/2})$ rate. An alternative approach to bounding this rate of convergence involves calculating an Edgeworth expansion. For an example of this approach for general linear rank statistics, see Schneller's [75] proof via Stein's method.

## 3.2 Self-Normalized Sums

Student's one-sample $t$-statistic

$$\begin{aligned} T_n &= \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n X_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}} \\ &= S_n(2)\sqrt{\frac{n-1}{n - S_n^2(2)}} \end{aligned}$$

is closely related to the self-normalized sum

$$S_n(p) = \frac{\sum_{i=1}^n X_i}{\left(\sum_{i=1}^n |X_i|^p\right)^{1/p}},$$

for which many limiting results exist. Logan et al. [59] studied $S_n(p)$ for general $p$, and Efron [27] showed that the limiting distributions of $T_n$ and $S_n(2)$ coincide and related the $t$-test to its non-parametric, one-sample counterparts.

More generally, self-normalized processes take the form $A_t/B_t$, where $B_t$ is a random variable that measures some form of dispersion of $A_t$ [18]. For the one-sample $t$-statistic, $A_t$ represents the standardized sample mean and $B_t$ the sample standard deviation. Since self-normalized processes are unit free, they are invariant to changes

of scale. Importantly, by normalizing $A_t$ by nonrandom $b_t$, we obviate the requirement of moment conditions such as in Shao's [80] work on self-normalized large deviations. For more information on the theory of self-normalized processes, see the book by de la Peña, Lai, and Shao [19].

## 3.3 Set-up

We observe two samples of equal size: $\mathcal{S}_1 = \{x_i\}_{i=1}^n$ and $\mathcal{S}_2 = \{x_i\}_{i=n+1}^{2n}$. Since we consider the $t$-statistic under different permutations, it will be convenient to re-write the sample values relative to the null permutation $\pi_0$: $\mathcal{S}_1 = \{x_{\pi_0(i)}\}_{i=1}^n$ and $\mathcal{S}_2 = \{x_{\pi_0(i)}\}_{i=n+1}^{2n}$, where $\pi_0(i) = i$.

The Behrens–Fisher [5, 29] problem concerns the setting when $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent random variables and the problem is to test $\mathcal{H}_A : \mu_1 \neq \mu_2$ against $\mathcal{H}_0 : \mu_1 = \mu_2$ when the variances of the two distributions are not assumed to be equal. Consider the class of statistics $u = (\bar{Y} - \bar{X})^2/(a_1 S_X^2 + a_2 S_Y^2)$ , where $\bar{X}$ and $S_X^2$ indicate the sample mean of the first group and sum of squared differences from such, respectively. Let $N = n + m$. Then when $a_1 = a_2 = N/mn(N-2)$, $u$ reduces to the square of Student's $t$ statistic, and to the Behrens–Fisher statistic when $A_1 = 1/m(m-1)$ and $A_2 = 1/n(n-1)$. Many researchers have made important contributions to the Behrens–Fisher problem including Hsu [44], Chapman [12], Prokof'yev and Shishkin [70], and Dudewicz and Ahmed [25, 26].

Under the randomization distribution, where $\Pi$ is a uniformly chosen permutation, Student's two-sample $t$-statistic is given by $T_\Pi = T_\Pi(\{x_{\Pi(i)}\}_{i=1}^n, \{x_{\Pi(i)}\}_{i=n+1}^{2n})$,

notationally suppressing the dependence on the data, where

$$\begin{aligned} T_\Pi &= \frac{\bar{x}_{1,\Pi} - \bar{x}_{2,\Pi}}{\sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{\Pi(i)}-\bar{x}_{1,\Pi})^2}{n} + \frac{\frac{1}{n-1}\sum_{i=n+1}^{2n}(x_{\Pi(i)}-\bar{x}_{2,\Pi})^2}{n}}} \\ &= \frac{1}{\sqrt{\frac{n}{n-1}}} \frac{\sum_{i=1}^{n} x_{\Pi(i)} - \sum_{i=n+1}^{2n} x_{\Pi(i)}}{\sqrt{\sum_{i=1}^{n}(x_{\Pi(i)} - \bar{x}_{1,\Pi})^2 + \sum_{i=n+1}^{2n}(x_{\Pi(i)} - \bar{x}_{2,\Pi})^2}} \\ &= \sqrt{\frac{n-1}{n}} \frac{q_\Pi}{d_\Pi}, \end{aligned}$$

where

$$\begin{aligned} q_\Pi &= \left( \sum_{i=1, i\neq I}^{n} x_{\Pi(i)} + x_{\Pi(I)} - \sum_{i=n+1, i\neq J}^{2n} x_{\Pi(i)} - x_{\Pi(J)} \right) \\ d_\Pi &= \sqrt{\sum_{i=1}^{n}(x_{\Pi(i)} - \bar{x}_{1,\Pi})^2 + \sum_{i=n+1}^{2n}(x_{\Pi(i)} - \bar{x}_{2,\Pi})^2} \\ \bar{x}_{1,\Pi} &= \frac{1}{n}\sum_{i=1}^{n} x_{\Pi(i)} \text{ and } \bar{x}_{2,\Pi} = \frac{1}{n}\sum_{i=n+1}^{2n} x_{\Pi(i)}. \end{aligned}$$

In order to perform hypothesis testing, we compute the observed value of $T_{\Pi=\pi_0}$, which we compare with the randomization distribution of $T_\Pi$. We shall create an exchangeable pair $(T_\Pi, T'_\Pi)$ by considering a uniformly random transposition $(I, J)$.

Without loss of generality, take $I \leq J$. We apply this transposition to the group labels. Note that if $I, J \in \{1, \dots, n\}$ or $I, J \in \{n+1, \dots, 2n\}$ then $T'_\Pi = T_\Pi$, where $T'_\Pi$ is the $t$-statistic under this random transposition. That is, the $t$-statistic is invariant to within-group transpositions: the only changes occur when $1 \leq I \leq n$ and $n+1 \leq J \leq 2n$.

With this in mind, let's redefine our transposition to be uniformly at random over

the $n^2$ cases where $1 \leq I \leq n$ and $n+1 \leq J \leq 2n$. Thus,

$$\begin{aligned} T'_\Pi(\{x_{\Pi(i)}\}_{i=1}^n, \{x_{\Pi(i)}\}_{i=n+1}^{2n}) &= T_{\Pi\circ(I,J)}(\{x_{\Pi\circ(I,J)(i)}\}_{i=1}^n, \{x_{\Pi\circ(I,J)(i)}\}_{i=n+1}^{2n}) \\ &= \sqrt{\frac{n-1}{n}}\frac{q'_\Pi}{d'_\Pi} \\ q'_\Pi &= \left(\sum_{i=1, i\neq I}^{n} x_{\Pi(i)} + x_{\Pi(J)} - \sum_{i=n+1, i\neq J}^{2n} x_{\Pi(i)} - x_{\Pi(I)}\right) \\ &= q_\Pi - 2x_{\Pi(I)} + 2x_{\Pi(J)} \\ d'_\Pi &= \sqrt{\sum_{i=1}^{n}(x_{\Pi(i)} - \bar{x}'_{1,\Pi})^2 + \sum_{i=n+1}^{2n}(x_{\Pi(i)} - \bar{x}'_{2,\Pi})^2}. \end{aligned}$$

## 3.4 Assumptions

Recall that the $t$-statistic is invariant up to sign under affine transformations, so we can mean-center and scale so that $\sum_{i=1}^{2n} x_i = 0$ and $\sum_{i=1}^{2n} x_i^2 = 2n$. The transformation that achieves this centering and scaling is given by

$$x_i \leftarrow \sqrt{\frac{2n}{\sum_{i=1}^{2n}(x_i - \bar{x})^2}}(x_i - \bar{x}), \tag{3.1}$$

where $\leftarrow$ means "replace by." Therefore, we just assume that the $x_i$'s have already been transformed. This can be seen as a very mild assumption: only $x_i = c$ for all $i$ cannot be scaled in this way.

We also assume that the pooled sample standard deviation is non-zero for all permutations:

$$d_\Pi = \sqrt{\sum_{i=1}^{n}(x_{\Pi(i)} - \bar{x}_{1,\Pi})^2 + \sum_{i=n+1}^{2n}(x_{\Pi(i)} - \bar{x}_{2,\Pi})^2} > 0 \tag{3.2}$$

This estimate is zero if and only if there exists a grouping that is constant in each group. The condition also implies that the sample mean for any group is strictly less than 1 in absolute value. In fact, this assumption subsumes the former.

The mean-centering assumption implies that $\sum_{i=1}^{n} x_{\Pi(i)} = -\sum_{i=n+1}^{2n} x_{\Pi(i)}$ and

hence that $\bar{x}_{1,\Pi} = -\bar{x}_{2,\Pi}$ for all $\Pi$.

Here we establish an equality with $d_\Pi$ that will prove easier to work with:

$$\begin{aligned} d_\Pi^2 &= \sum_{i=1}^{n}(x_{\Pi(i)} - \bar{x}_{1,\Pi})^2 + \sum_{i=n+1}^{2n}(x_{\Pi(i)} - \bar{x}_{2,\Pi})^2 \\ &= \sum_{i=1}^{2n} x_{\Pi(i)}^2 - n\bar{x}_{1,\Pi}^2 - n\bar{x}_{2,\Pi}^2 \\ &= 2n - n\bar{x}_{2,\Pi}^2 - n\bar{x}_{2,\Pi}^2 \\ &= 2n(1 - \bar{x}_{2,\Pi}^2) \end{aligned}$$

Since $d_\Pi > 0$, it follows that $|\bar{x}_{2,\Pi}| < 1$. Define

$$B = \max_\Pi |\bar{x}_{2,\Pi}| < 1. \tag{3.3}$$

## 3.5 Preliminaries

Here we collect useful bounds and other results. We include them here rather than in Appendix A because in Chapter 4 we compare the theoretical bounds with simulated results.

In order to bound various moments of $\bar{x}_{2,\Pi}$ under the permutation distribution, we use a result of Serfling's [79]:

**Proposition 3.1.** *Consider sampling without replacement from a finite list of values* $x_1, \ldots, x_{2n}$. *Let* $x_\Delta := \max_i x_i - \min_i x_i$. *Then for* $p > 0$,

$$\begin{aligned} \mathbb{E}[\bar{x}_{2,\Pi}^p] &\le \frac{\Gamma(p/2+1)}{2^{p/2+1}}\left[\frac{n+1}{2n}x_\Delta^2\right]^{p/2}(2n)^{-p/2} \\ &\le \frac{\Gamma(p/2+1)}{2^{p/2+1}}\left[\frac{n+1}{4n}x_\Delta^2\right]^{p/2}n^{-p/2} \\ &:= f_{c_1}(p)n^{-p/2}. \end{aligned} \tag{3.4}$$

By Assumption (3.3),

$$(d_\Pi)^{-p} = \frac{1}{(2n(1-\bar{x}_{2,\Pi}^2))^{p/2}} \leq \frac{1}{(2n(1-B^2))^{p/2}} := f_{c_2}(p)n^{-p/2}. \tag{3.5}$$

The transposition $(I, J)$ also affects the denominator of $T'_\Pi$, and we need to quantify the difference between the denominators of $T_\Pi$ and $T'_\Pi$. Letting $\bar{x}'^2_{2,\Pi}$ denote the sample mean of the second group after the transposition,

$$\begin{aligned}\bar{x}'^2_{2,\Pi} &= \left(\bar{x}_{2,\Pi} - \frac{1}{n}x_{\Pi(J)} + \frac{1}{n}x_{\Pi(I)}\right)^2 \\ &= \bar{x}^2_{2,\Pi} + 2\bar{x}_{2,\Pi}\left(-\frac{1}{n}x_{\Pi(J)} + \frac{1}{n}x_{\Pi(I)}\right) + \frac{1}{n^2}(x_{\Pi(I)} - x_{\Pi(J)})^2\end{aligned}$$

We consider the difference

$$\begin{aligned}h_\Pi &= d_\Pi^2 - d'^2_\Pi \\ &= 2n - 2n\bar{x}^2_{2,\Pi} - 2n + 2n\bar{x}'^2_{2,\Pi} \\ &= 4\bar{x}_{2,\Pi}(x_{\Pi(I)} - x_{\Pi(J)}) + \frac{2}{n}(x_{\Pi(I)} - x_{\Pi(J)})^2\end{aligned}$$

Therefore, by the $c_r$-inequality,

$$\begin{aligned}\mathbb{E}[h_\Pi^p] &= \mathbb{E}\left|4\bar{x}_{2,\Pi}(x_{\Pi(I)} - x_{\Pi(J)}) + \frac{2}{n}(x_{\Pi(I)} - x_{\Pi(J)})^2\right|^p \\ &\leq 2^{p-1}\left(\mathbb{E}|4\bar{x}_{2,\Pi}(x_{\Pi(I)} - x_{\Pi(J)})|^p + \mathbb{E}\left|\frac{2}{n}(x_{\Pi(I)} - x_{\Pi(J)})^2\right|^p\right) \\ &\leq 2^{p-1}\left[(4x_\Delta)^p\mathbb{E}\left|\bar{x}_{2,\Pi}\right|^p + \left(\frac{2}{n}x_\Delta^2\right)^p\right] \\ &\leq 2^{p-1}(4x_\Delta)^p f_{c_1}(p)n^{-p/2} + 2^{p-1}(2x_\Delta^2)^p n^{-p} \\ &:= f_{c_3}(p)n^{-p/2}.\end{aligned} \tag{3.6}$$

Now we establish a bound on the difference $d_\Pi - d'_\Pi$ via a bound on the remainder

of a zeroth-order Taylor approximation. Let

$$d'_\Pi = \sqrt{d_\Pi^2 - h_\Pi} = f(h_\Pi) = f(0) + R_0(h_\Pi) = d_\Pi + R_0(h_\Pi).$$

By Taylor's theorem, the remainder of the zeroth-order expansion takes the form

$$R_0(h_\Pi) = \frac{f'(\xi_L)}{1} h_\Pi = \frac{-h_\Pi}{2\sqrt{d_\Pi^2 - \xi_L}}, \quad \text{where } \xi_L \in [0, h_\Pi].$$

We are approximating $d'_\Pi$ by a constant and bounding the error via a function of the first derivative. This is a sufficient approximation because the squared difference $h_\Pi$ is not so big relative to the flattening out of the square root function. Now

$$|d_\Pi - d'_\Pi| \leq |R_0(h_\Pi)| \leq \frac{|h_\Pi|}{2\sqrt{d_\Pi^2 - \xi_L}} \leq \frac{|h_\Pi|}{2\sqrt{d_\Pi^2 - \max(0, h_\Pi)}}.$$

Recall that $h_\Pi = d_\Pi^2 - d_\Pi'^2$, so

$$d_\Pi^2 - \max(0, d_\Pi^2 - d_\Pi'^2) = \begin{cases} d_\Pi^2 & \text{if } d_\Pi^2 - d_\Pi'^2 \leq 0 \\ d_\Pi'^2 & \text{if } d_\Pi^2 - d_\Pi'^2 > 0. \end{cases}$$

Therefore,

$$|d_\Pi - d'_\Pi| \leq \frac{|h_\Pi|}{2\min(d_\Pi, d'_\Pi)} \leq \max\left(\frac{|h_\Pi|}{2d_\Pi}, \frac{|h_\Pi|}{2d'_\Pi}\right) \leq \frac{|h_\Pi|}{2d_\Pi} + \frac{|h_\Pi|}{2d'_\Pi}.$$

The important thing to do is to isolate $|h_\Pi|$, which is small in expectation, but

not absolutely. By the $c_r$-inequality,

$$\begin{aligned}
\mathbb{E}|d_\Pi - d'_\Pi|^p &\le 2^{p-1}\left(\mathbb{E}\left|\frac{h_\Pi}{2d_\Pi}\right|^p + \mathbb{E}\left|\frac{h_\Pi}{2d'_\Pi}\right|^p\right) \\
&\le 2^{-1}\left(\sqrt{\mathbb{E}[h_\Pi^{2p}]\mathbb{E}[d_\Pi^{-2p}]} + \sqrt{\mathbb{E}[h_\Pi^{2p}]\mathbb{E}[d_\Pi'^{-2p}]}\right) \\
&\le \sqrt{f_{c_3}(2p)n^{-2p/2} f_{c_2}(2p)n^{-2p/2}} \quad \text{by (3.5) and (3.6)} \\
&:= f_{c_4}(p)n^{-p}. \qquad (3.7)
\end{aligned}$$

With

$$q_\Pi = n\bar{x}_{1,\Pi} - n\bar{x}_{2,\Pi} = -2n\bar{x}_{2,\Pi}, \qquad (3.8)$$

(3.4), and noting that $q_\Pi$ and $q'_\Pi$ are exchangeable,

$$\mathbb{E}[q_\Pi'^p] = \mathbb{E}[q_\Pi^p] = \mathbb{E}[(-2n\bar{x}_{2,\Pi})^p] \le 2^p n^p f_{c_1}(p)n^{-p/2} := f_{c_5}(p). \qquad (3.9)$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^p\right] &\le \sqrt{\mathbb{E}|q'_\Pi|^{2p}\mathbb{E}|d_\Pi d'_\Pi|^{-2p}} \\
&\le \sqrt{\mathbb{E}|q_\Pi|^{2p}\sqrt{\mathbb{E}|d_\Pi|^{-4p}\mathbb{E}|d'_\Pi|^{-4p}}} \\
&= \sqrt{\mathbb{E}|q_\Pi|^{2p}\mathbb{E}|d_\Pi|^{-4p}} \\
&\le \sqrt{f_{c_5}(2p)n^{2p/2} f_{c_2}(4p)n^{-4p/2}} \quad \text{from (3.5) and (3.9)} \\
&:= f_{c_6}(p)n^{-p/2}. \qquad (3.10)
\end{aligned}$$

## 3.6 Proof

We proceed to verify the conditions of Theorems 2.7 and 2.8. $T_\Pi$ and $T'_\Pi$ are exchangeable by construction:

$$\begin{aligned}
P(\Pi = \pi, \Pi' = \pi') &= P(\Pi' = \pi' | \Pi = \pi) P(\Pi = \pi) \\
&= \frac{1}{n^2} \mathbb{1}_{\{\pi' = \pi \circ (i,j), 1 \le i \le n, n+1 \le j \le 2n\}} P(\Pi = \pi') \\
&= \frac{1}{n^2} \mathbb{1}_{\{\pi = \pi' \circ (i,j), 1 \le i \le n, n+1 \le j \le 2n\}} P(\Pi = \pi') \\
&= P(\Pi' = \pi | \Pi = \pi') P(\Pi = \pi') \\
&= P(\Pi = \pi', \Pi' = \pi)
\end{aligned}$$

Since $(\Pi, \Pi')$ are exchangeable, $(T_\Pi, T'_\Pi) = (T(\Pi), T(\Pi'))$ are exchangeable as well. $T_\Pi$, and thus $T'_\Pi$ by exchangeability, have mean zero by symmetry. Let $\pi^*$ identify the permutation that reverses the order of the indices after applying the original permutation $\pi$. That is, $\pi^* = (2n, \dots, 1) \circ \pi$. Since indices 1 to $n$ correspond to the first group and $n+1$ to $2n$ to the second, $\pi^*$ reverses the groups after $\pi$, so $T_{\pi^*} = -T_\pi$.

$$\begin{aligned}
P(T_\Pi = t) &= \sum_{\pi : T_\pi = t} P(\Pi = \pi) \\
&= \sum_{\pi : T_\pi = t} P(\Pi = \pi^*) \quad \text{by exchangeability} \\
&= \sum_{\pi^* : T_{\pi^*} = -t} P(\Pi = \pi^*) \quad \text{since } T_{\pi^*} = -T_\pi \text{ and } \pi | \to \pi^* \text{ is bijective} \\
&= P(T_\Pi = -t)
\end{aligned}$$

To show the approximate regression condition, the difference of our exchangeable

pair is given by

$$\begin{aligned} T'_\Pi - T_\Pi &= \sqrt{\frac{n-1}{n}}\left(\frac{q'_\Pi}{d'_\Pi} - \frac{q_\Pi}{d_\Pi}\right) \\ &= \sqrt{\frac{n-1}{n}}\frac{1}{d_\Pi}\left(q'_\Pi - q_\Pi + q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\right) \\ &= \sqrt{\frac{n-1}{n}}\frac{1}{d_\Pi}\left(2x_{\Pi(J)} - 2x_{\Pi(I)} + q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\right). \end{aligned} \tag{3.11}$$

Note that

$$\begin{aligned} \sqrt{\frac{n-1}{n}}\mathbb{E}\left[\frac{1}{d_\Pi}(2x_{\Pi(J)} - 2x_{\Pi(I)})\Big|\Pi = \pi\right] &= \sqrt{\frac{n-1}{n}}\frac{2}{d_\Pi}\frac{1}{n^2}\sum_{I=1}^{n}\sum_{I=n+1}^{2n}(x_{\Pi(J)} - x_{\Pi(I)}) \\ &= -\frac{2}{n}T_\Pi. \end{aligned}$$

Therefore,

$$\sqrt{\frac{n-1}{n}}\mathbb{E}\left[\frac{1}{d_\Pi}(2x_{\Pi(J)} - 2x_{\Pi(I)})\Big|\Pi = \pi\right] = \sqrt{\frac{n-1}{n}}\mathbb{E}\left[\frac{1}{d_\Pi}(2x_{\Pi(J)} - 2x_{\Pi(I)})\Big|T_\Pi\right]$$

and

$$\lambda = \frac{2}{n}.$$

$$\begin{aligned} \mathbb{E}[T'_\Pi - T_\Pi|T_\Pi] &= -\lambda T_\Pi + \sqrt{\frac{n-1}{n}}\mathbb{E}\left[\frac{q'_\Pi}{d_\Pi}\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\Big|T_\Pi\right] \\ &= -\lambda\left(T_\Pi - \left(\frac{n}{2}\right)\sqrt{\frac{n-1}{n}}\mathbb{E}\left[\frac{q'_\Pi}{d_\Pi}\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\Big|T_\Pi\right]\right) \end{aligned}$$

so

$$R_\Pi = \left(\frac{n}{2}\right)\sqrt{\frac{n-1}{n}}\frac{1}{d_\Pi}\mathbb{E}\left[q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\Big|T_\Pi\right]. \tag{3.12}$$

For convenience, we restate Theorem 2.7 of Chapter 2, taking our random variables $W$ to be the randomization $t$-statistic $T_\Pi$ and $W'$ to be its coupled counterpart $T'_\Pi$:

**Theorem 1.8.** *If $T_\Pi$, $T'_\Pi$ are mean 0 exchangeable random variables with variance $\mathbb{E}T_\Pi^2$ satisfying*

$$\mathbb{E}[T'_\Pi - T_\Pi | T_\Pi] = -\lambda(T_\Pi - R_\Pi)$$

*for some $\lambda \in (0,1)$ and some random variable $R_\Pi$, then*

$$\sup_{t\in\mathbb{R}} |P(T_\Pi \le t) - \Phi(t)| \le (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi - T_\Pi|^3}{\lambda}} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} + |\mathbb{E}T_\Pi^2 - 1| + \mathbb{E}|T_\Pi R_\Pi| + \mathbb{E}|R_\Pi|$$

With the preliminaries in place, we proceed to provide bounds on each term in Theorem 2.7, the proofs of which we defer to Appendix C.

**Proposition 3.2.** $(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi - T_\Pi|^3}{\lambda}} < (2\pi)^{-1/4} c_9 n^{-1/4}$.

**Proposition 3.3.** $\frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} \le n^{-1} c_3 \sqrt{20 + 16\frac{\sum_{i=1}^{2n} x_i^4}{n^2}}$

**Proposition 3.4.** $|\mathbb{E}T_\Pi^2 - 1| \le c_2 n^{-1}$

**Proposition 3.5.** $\mathbb{E}|T_\Pi R| \le \frac{1}{2}(f_{c_6}(4) f_{c_4}(4))^{1/4}\sqrt{2 + 2c_1} n^{-1/2}$.

**Proposition 3.6.** $\mathbb{E}|R| \le \frac{1}{2}\sqrt{f_{c_6}(2) f_{c_4}(2)} n^{-1/2}$.

The bound in Proposition 3.2 is suboptimal, as it will only allow us to obtain a rate of $\mathcal{O}(n^{-1/4})$. In Section 3.7, we introduce an additional condition to improve upon this rate.

Collecting the results of Propositions 3.2, 3.3, 3.4, 3.5, and 3.6, we have

$$\begin{aligned}
\sup_{t\in\mathbb{R}} |P(T_\Pi \le t) - \Phi(t)| &\le (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi - T_\Pi|^3}{\lambda}} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} \\
&\quad + |\mathbb{E}T_\Pi^2 - 1| + \mathbb{E}|T_\Pi R_\Pi| + \mathbb{E}|R_\Pi| \\
&\le (2\pi)^{-1/4} c_9 n^{-1/4} + n^{-1} c_3 \sqrt{20 + 16\frac{\sum_{i=1}^{2n} x_i^4}{n^2}} + c_2 n^{-1} \\
&\quad + \frac{1}{2}(f_{c_6}(4) f_{c_4}(4))^{1/4}\sqrt{2 + 2c_1} n^{-1/2} + \frac{1}{2}\sqrt{f_{c_6}(2) f_{c_4}(2)} n^{-1/2}
\end{aligned}$$

Note that since $\|x\|_4 \leq \|x\|_2$,

$$\sum_{i=1}^{2n} x_i^4 \leq \left( \sum_{i=1}^{2n} x_i^2 \right)^{4/2} = (2n)^2 = 4n^2.$$

This result is similar to the HCCLT. Given fixed data, we can obtain an explicit upper bound on the Kolmogorov distance between the randomization distribution of our statistic of interest and the standard normal distribution.

## 3.7 Better Rate

Here, we use Theorem 2.8 to establish a rate of $\mathcal{O}(n^{-1/2})$ with the condition that $|T_\Pi - T'_\Pi| \leq \delta$ is $\mathcal{O}(n^{-1/2})$.

From Proposition 3.4, $\mathbb{E}T_\Pi^2 \leq c_2 n^{-1} + 1$, and from Proposition 3.6, $\mathbb{E}|R| \leq \frac{1}{2}\sqrt{f_{c_6}(2) f_{c_4}(2)} n^{-1/2}$. If $\delta < c_{10} n^{-1/2}$ for $n$ sufficiently large, applying Theorem 2.8, we see

$$\begin{aligned}
\sup_{t \in \mathbb{R}} |P(T_\Pi \leq t) - \Phi(t)| &\leq \frac{.41\delta^3}{\lambda} + 3\delta \left( \sqrt{\mathbb{E}T_\Pi^2} + \mathbb{E}|R| \right) + \frac{1}{2\lambda} \sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2 | T_\Pi])} \\
&\quad + |\mathbb{E}T_\Pi^2 - 1| + \mathbb{E}|T_\Pi R| + \mathbb{E}|R| \\
&\leq .205 c_{10} n^{-1/2} + 3 c_{10} n^{-1/2} \left( c_2 n^{-1} + 1 + \frac{1}{2} \sqrt{f_{c_6}(2) f_{c_4}(2)} n^{-1/2} \right) \\
&\quad + n^{-1} c_3 \sqrt{20 + 16 \frac{\sum_{i=1}^{2n} x_i^4}{n^2}} + c_2 n^{-1} \\
&\quad + \frac{1}{2} (f_{c_6}(4) f_{c_4}(4))^{1/4} \sqrt{2 + 2c_1} n^{-1/2} + \frac{1}{2} \sqrt{f_{c_6}(2) f_{c_4}(2)} n^{-1/2}.
\end{aligned}$$

Again, this result is conditional on the data. We can consider a sequence of vectors $\{x_i^{(2n)}\}$, where each $x_i^{(j)}$ is drawn from some distribution $p$. As long as all data-dependent functions of the bound are "well-behaved," we shall have the desired rates of convergence, such as in [8].

To determine whether $\delta = |T_\Pi - T'_\Pi|$ is $\mathcal{O}(n^{-1/2})$ for reasonable classes of data

$\{x_i\}$, recall that

$$T_\Pi(\{x_{\Pi(i)}\}_{i=1}^n, \{x_{\Pi(i)}\}_{i=n+1}^{2n}) = \frac{\bar{x}_{1,\Pi} - \bar{x}_{2,\Pi}}{\sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^n (x_{\Pi(i)} - \bar{x}_{1,\Pi})^2}{n} + \frac{\frac{1}{n-1}\sum_{i=n+1}^{2n} (x_{\Pi(i)} - \bar{x}_{2,\Pi})^2}{n}}}.$$

We need to set $\delta = \max_{\pi,i,j} |T_\pi - T_{\pi \circ (i,j)}|$ so that the bound is tight. This appears to be a daunting optimization problem. There are $(2n)!$ permutations and $n^2$ possible transpositions $(i, j)$ for each permutation. Because the $t$-statistic is invariant to permutations within groups and due to symmetry, there are $\binom{2n}{n}/2$ permutations to consider.

We have to solve the maximization problem jointly over $T$ and $T'$. We can attempt to first maximize over $T$ and then $T'$. Note that these sequential approaches do not work for general optimization problems.

If we sort the data in ascending order such that the two groups are $\{x_{(i)}\}_{i=1}^n$ and $\{x_{(i)}\}_{i=n+1}^{2n}$, then it seems like we will have maximized $|T|$. The absolute difference between the sample means of the two groups is maximized, while the pooled sample standard deviation is minimized.

The transposition that should then maximize $|T - T'|$ is $(1, 2n)$ since it swaps the most different points, decreasing the difference in sample means and increasing the pooled sample standard deviation.

Let $\pi^*$ be the permutation that sorts the data in ascending order such that $x_{\pi^*(i)} = x_{(i)}$, where $x_{(i)}$ are the order statistics of $\{x_i\}$. Let $i^* = 1$ and $j^* = 2n$.

**Conjecture 3.7.** *$\delta = \max_{\pi,i,j} |T_\pi - T_{\pi \circ (i,j)}|$ is maximized at $\pi = \pi^*$, $i = i^*$, and $j = j^*$.*

This conjecture has held true under many simulations. We can show that when $x_i = i$,

$$\lim_{n \to \infty} \delta\sqrt{n} = 16\sqrt{6}.$$

# Chapter 4

# Simulations

This chapter is a computational companion to Chapter 3.

## 4.1 Preliminaries

First, we provide simulations accompanying Section 3.5. We generate independently and identically distributed samples $\{x_i\}_{i=1}^{n} \sim \mathcal{N}(-1, 1)$ and $\{x_i\}_{i=n+1}^{2n} \sim \mathcal{N}(1, 1)$ for exponentially-spaced values of $n$. The $x_i$ are scaled and centered, and for each $n$, we perform $10,000$ permutations.

In Figure 4.1, we plot Monte Carlo estimates of the means of each term, scaled by the rate of our bound, along with 95th percentile bootstrap confidence intervals for different values of $p \in \{2, 4, 6, 8\}$.

Due to the flatness of the curves, we conclude that the bounds we have proved are of the correct rate. In addition, we can observe the behavior of the constants as functions of $p$. For instance, our $f_{c_3}(p)$ constant for $\mathbb{E}h_{\Pi}^{p}$ appears to be an exponential function of $p$.

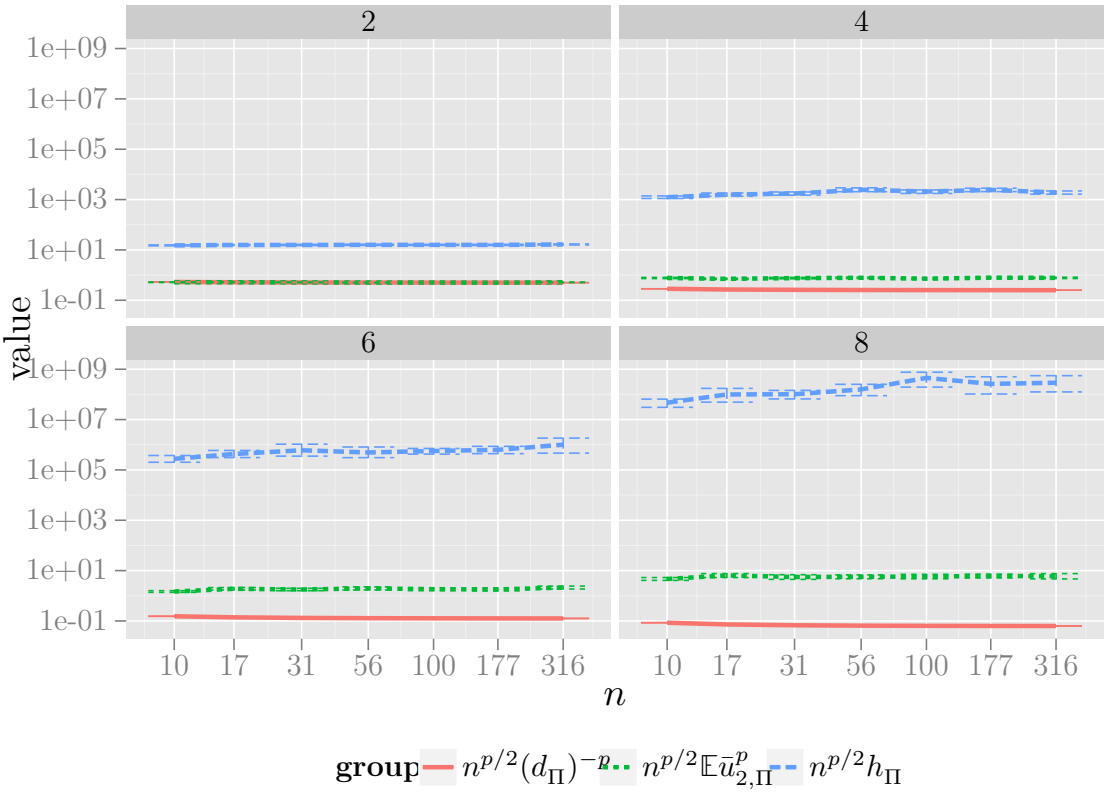


Figure 4.1: Log-log plots of values scaled by proven upper bounds of rates, faceted on $p$.

In Figure 4.2, to compute the corresponding "prime" random variables in the coupled pair, in each permutation we pick a transposition uniformly at random among transpositions that switch groups.

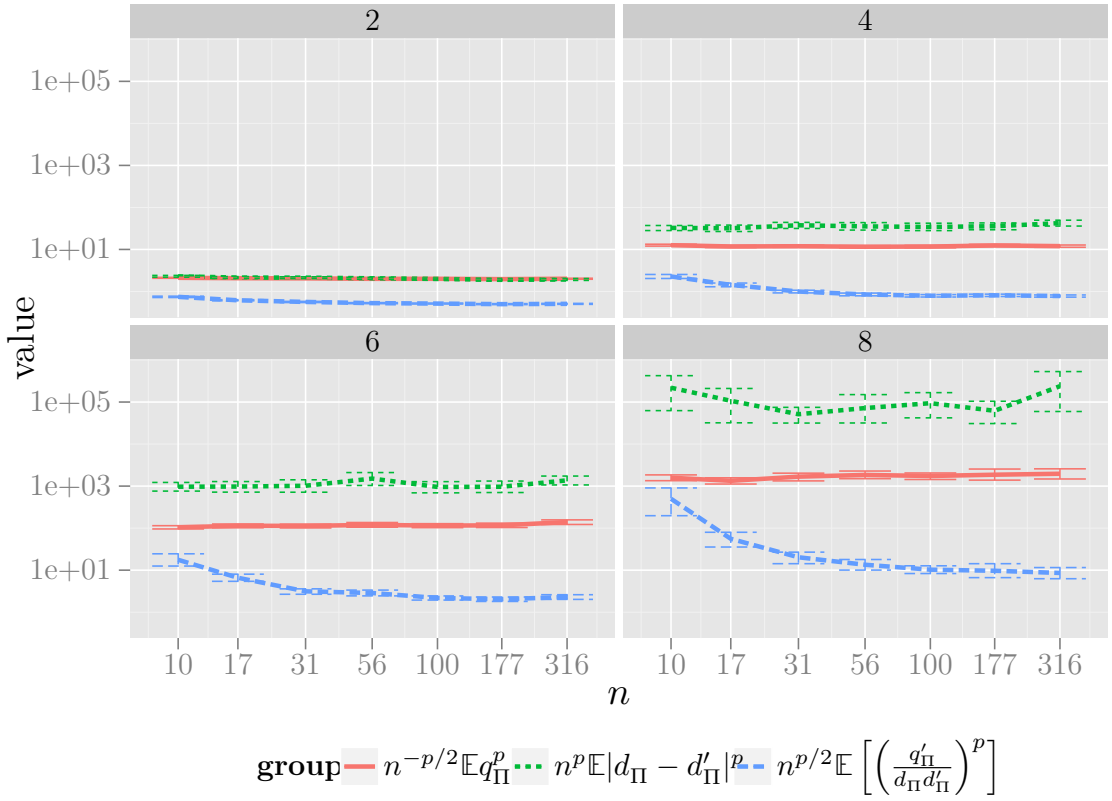


Figure 4.2: Log-log plots of values scaled by proven upper bounds of rates, faceted on $p$.

It is possible that the bound of rate $n^{-p/2}$ on $\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^p\right]$ is a bit conservative.

## 4.2 Approximate Regression Condition

From the approximate regression condition

$$\mathbb{E}[T'_\Pi - T_\Pi | T_\Pi] = -\lambda(T_\Pi - R_\Pi),$$

we get

$$\mathbb{E}[T'_\Pi | T_\Pi] = (1-\lambda)T_\Pi - \lambda R_\Pi.$$

That is, the conditional expectation of $T'_\Pi$ on $T_\Pi$ is expected to lie near the line $(1-\lambda)T_\Pi$ with a small perturbation of order $1/n$ (recall that $\lambda = 2/n$).

For various values of $n$, we compute 20 permutations that correspond to 20 values of $T_\Pi$. For each $T_\Pi$, we draw a transposition $(I, J)$ uniformly at random from the space of our allowable transpositions, repeating this 50 times, each producing a value of $T'_\Pi$

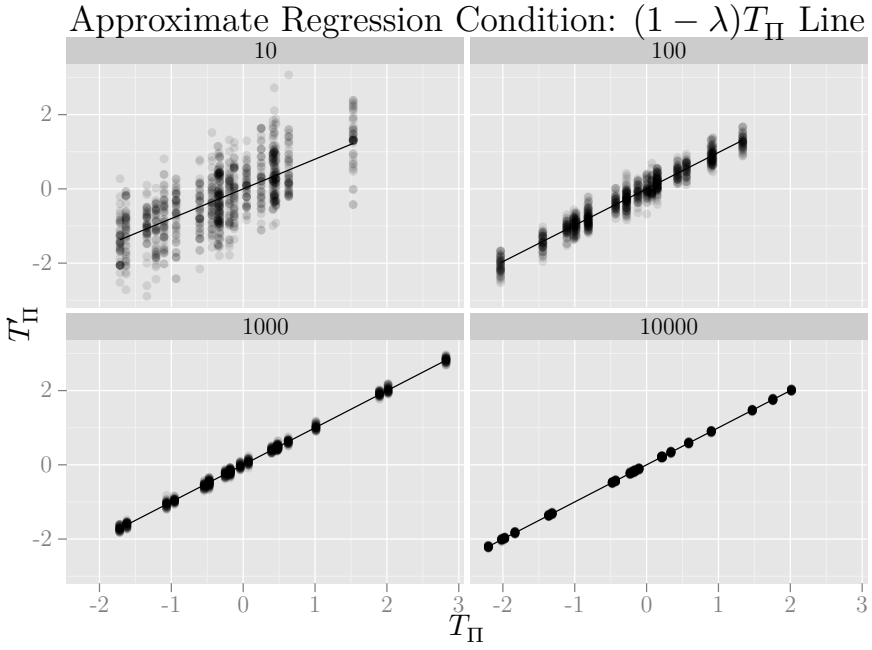


Figure 4.3: Faceted on per-group sample size, $n$.

The approximate regression condition appears to hold visually.

## 4.3 Main Bounds

Here we simulate the main bounds under the same conditions as the previous section.

### 4.3.1 Failure of Monte Carlo

Again, we simulate the conditional expectations of the form $\mathbb{E}[f(T'_\Pi, T_\Pi)|T_\Pi]$ with 1,000 draws from the uniform distribution on all group-switching transpositions $(I, J)$.

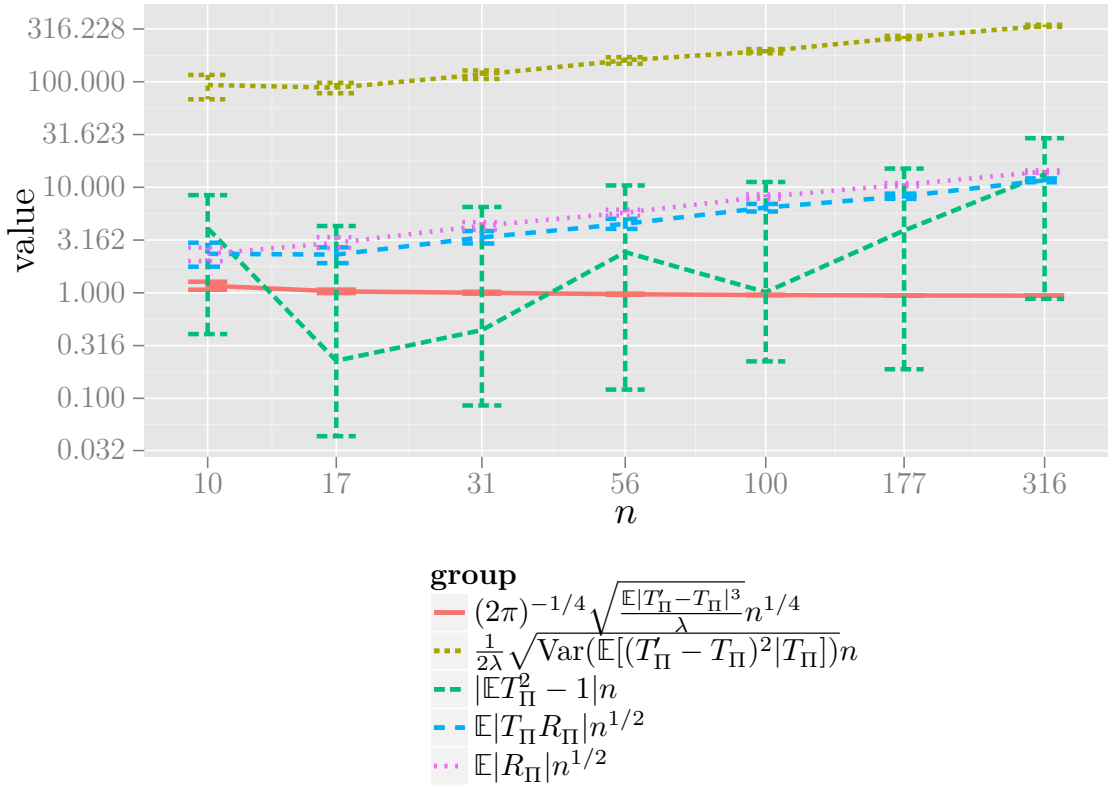


Figure 4.4: Log-log plot of values for each term in the bound, simulating the conditional expectation using Monte Carlo methods.

The Monte Carlo error is too large, and we see some scaled bounds actually increase.

### 4.3.2 Exact Conditional Expectation Calculations

Motivated by the high Monte Carlo error in estimating the conditional expectations $\mathbb{E}[f(T'_\Pi, T_\Pi)|T_\Pi]$, we describe an efficient procedure to calculate all values $T'$ corresponding to a given $T$ exactly in Section 4.4. We still use Monte Carlo simulations to calculate a subset of all the permutations $\Pi$, otherwise the computational cost would be prohibitive for large sample sizes.

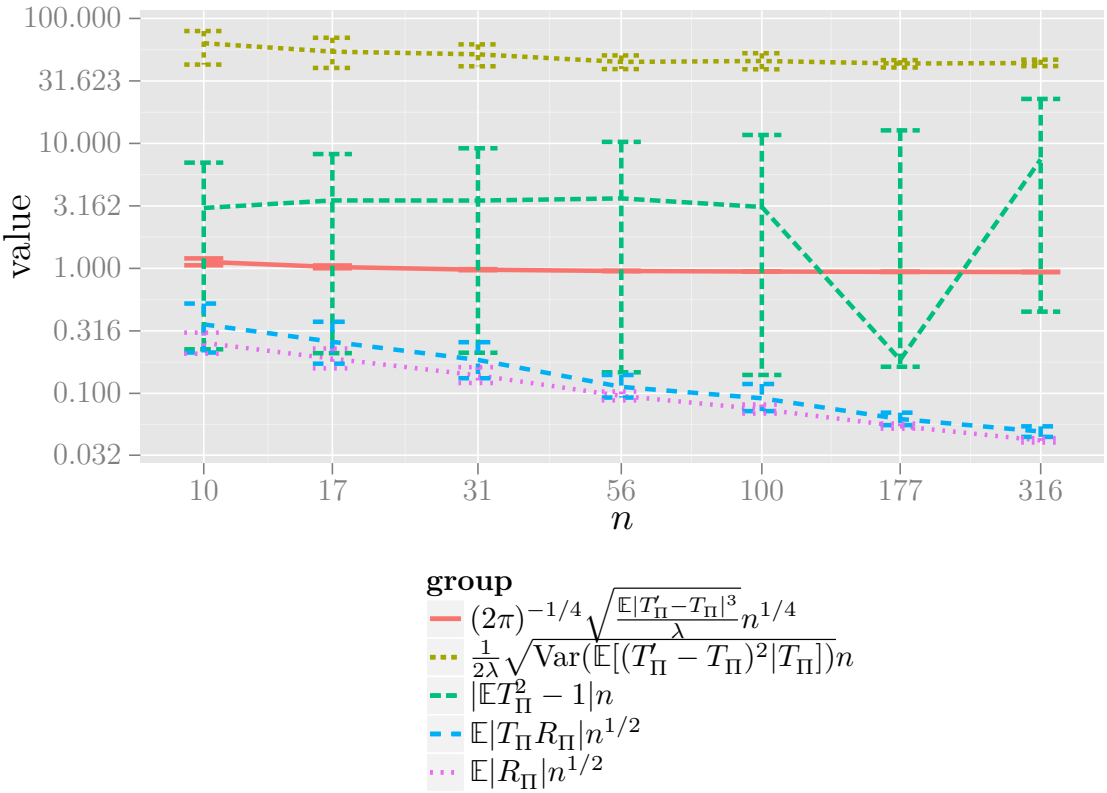


Figure 4.5: Log-log plot of values for each term in the bound, calculating the conditional expectation exactly ($10n$ permutations each).

Our bounds appear to be of the correct order or slightly conservative in some cases. The bounds on the remainder terms ($\mathbb{E}|R_\Pi|$ and $\mathbb{E}|T_\Pi R_\Pi|$) are of order $n^{1/2}$, but the true rates are probably lower.

### 4.3.3 Better Rate

All terms except the term involving $\delta^3$ appear to be of or better than the proven rate.

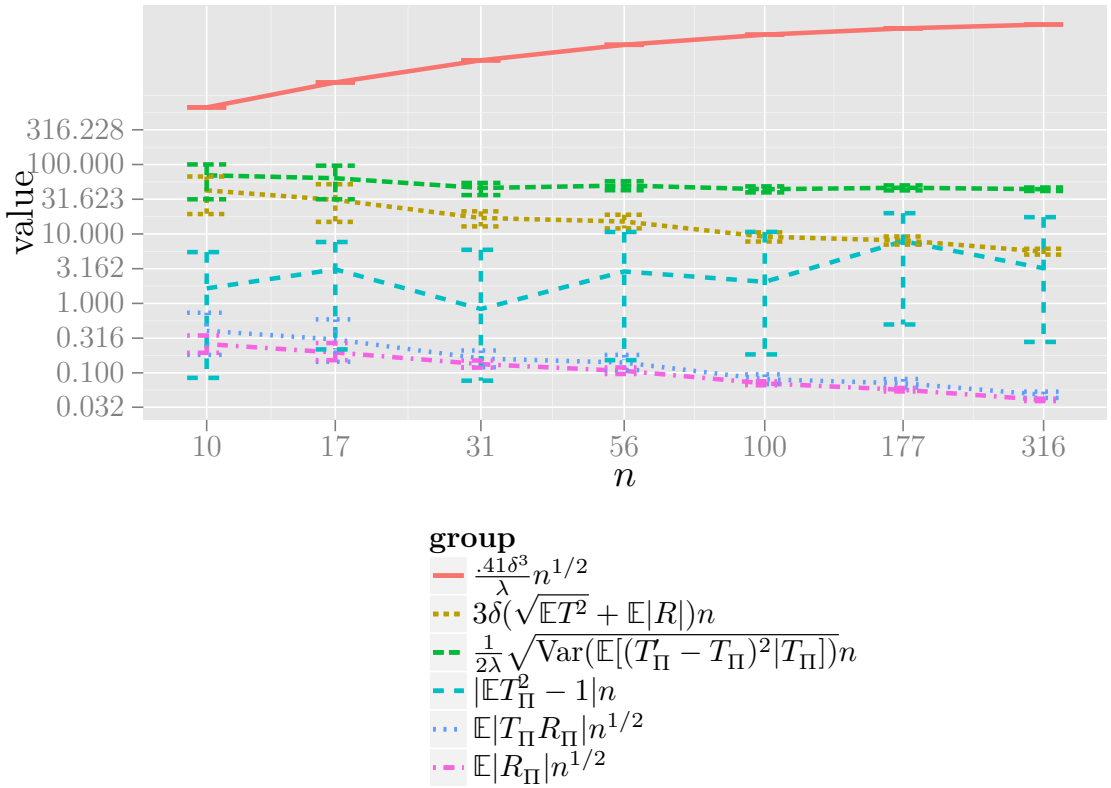


Figure 4.6: Log-log plot of values for each term in the bound, calculating the conditional expectation exactly ($10n$ permutations each).

Figure 4.7 shows that $\frac{.41\delta^3}{\lambda}n^{1/2}$ is in fact bounded in certain settings.

Below, we see a plot of $\frac{.41\delta^3}{\lambda}n^{1/2}$ on $n$ when $u_i = i$.

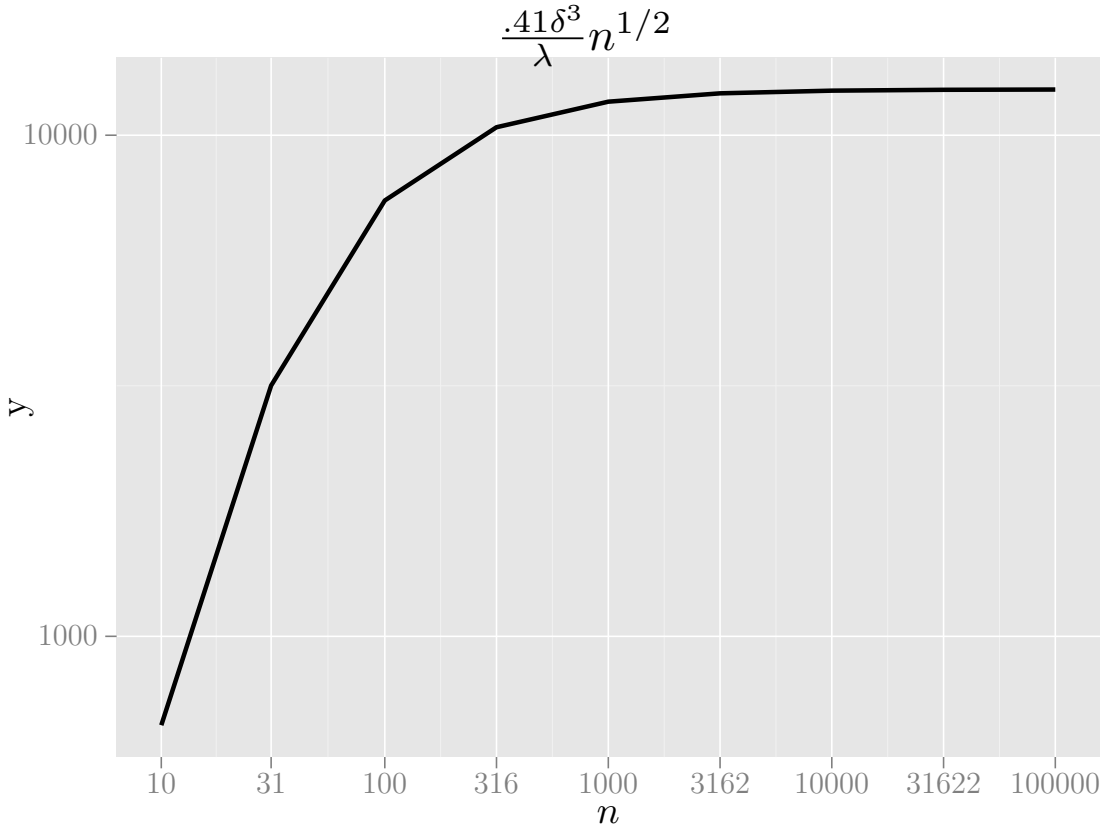


Figure 4.7: $\frac{.41\delta^3}{\lambda}n^{1/2}$ on $n$.

Recall that

$$\lim_{n\to\infty} \delta\sqrt{n} = 16\sqrt{6},$$

so the contribution of this term to the bound of Theorem 2.8 is not terribly useful for small $n$.

## 4.4 Efficient Updates

Instead of conditioning on the value of $T_\Pi$, we condition on the observed permutation $\pi$. For $n$ observations in each group, there are $n^2\, T'_\Pi$ values that come from swapping

one value in the first group with one value in the second. $T'_\Pi$ should not differ much from $T_\Pi$, and calculating the $t$-statistics anew is inefficient.

We use an efficient $t$-statistic update rule to easily calculate millions of $t$-statistics. The two sample $t$-statistic is given by

$$T_\Pi = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{2}{n}}\sqrt{\frac{1}{2}(S_X^2 + S_Y^2)}},$$

where $S_X^2 = \frac{1}{n-1}(\sum_{i=1}^n x_i^2 - n\bar{x}^2)$.

Let $T_{x_i,y_j}$ be the result of $T'$ by swapping $x_i$ with $y_j$:

$$\begin{aligned}
\Delta &= y_j - x_i \\
\bar{x}_{x_i,y_j} &= \bar{x} - \frac{1}{n}x_i + \frac{1}{n}y_j = \bar{x} + \frac{\Delta}{n} \\
\bar{y}_{x_i,y_j} &= \bar{y} + \frac{1}{n}x_i - \frac{1}{n}y_j = \bar{y} - \frac{\Delta}{n} \\
S^2_{X_{x_i,y_j}} &= \frac{1}{n-1}\left(\sum_{k=1}^n x_k^2 - x_i^2 + y_j^2\right) - \frac{n}{n-1}\bar{x}^2_{x_i,y_j} \\
S^2_{Y_{x_i,y_j}} &= \frac{1}{n-1}\left(\sum_{k=1}^n y_k^2 + x_i^2 - y_j^2\right) - \frac{n}{n-1}\bar{y}^2_{x_i,y_j} \\
\bar{x}^2_{x_i,y_j} &= \bar{x}^2 + \frac{2\Delta}{n}\bar{x} + \frac{\Delta^2}{n} \\
\bar{y}^2_{x_i,y_j} &= \bar{y}^2 - \frac{2\Delta}{n}\bar{y} + \frac{\Delta^2}{n}
\end{aligned}$$

Then

$$\begin{aligned}
T_{x_i,y_j} &= \frac{\bar{x}_{x_i,y_j} - \bar{y}_{x_i,y_j}}{\sqrt{\frac{2}{n}}\sqrt{\frac{1}{2}(S^2_{X_{x_i,y_j}} + S^2_{Y_{x_i,y_j}})}} \\
&= \frac{\bar{x} - \bar{y} + \frac{2\Delta}{n}}{\sqrt{\frac{2}{n}}\sqrt{\frac{1}{2(n-1)}[\sum_{k=1}^n (x_k^2 + y_k^2) - n(\bar{x}^2 + \bar{y}^2 + \Delta(\frac{2\bar{x}}{n} - \frac{2\bar{y}}{n}) + \frac{2}{n^2}\Delta^2)]}}.
\end{aligned}$$

Only the terms involving $\Delta$ need to be recomputed for each of the $n^2$ swaps.

# Chapter 5

# Friedman's Test

In this chapter we describe Friedman's approach to the two-sample problem, provide examples using a kernel support vector machine (KSVM), and explain the connection between the KSVMs and the theory developed in Chapter 3.

## 5.1 Motivation

The two-sample problem addresses the issue of comparing samples from two possibly different probability distributions. They range from simple parametric, location alternative tests on univariate data such as the $t$-test to more general non-parametric, asymptotically consistent tests, which have power against all alternatives. Many options exist for vectorial data, and kernels provide an enticing avenue for extensions to more general data types.

The two-sample problem is also widely prevalent: ensuring cross-platform compability of microarray data allows for the merging samples to achieve larger sample sizes. Biologists would like to know whether gene expression levels on a set of genes differ between cancer and control groups. Further uses for two-sample testing include authorship validation: Given two sets of documents, is the hypothesis of a single author consistent with the data?

## 5.2 Two-Sample Tests

The two-sample problem is generally posed in the following fashion: $\{\mathbf{x}_i\}_1^n$ are drawn from $p(\mathbf{x})$ and $\{\mathbf{y}_i\}_1^m$ are drawn from $q(\mathbf{y})$, where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^p$. The goal is to test $H_0 : p(\mathbf{x}) = q(\mathbf{y})$ against $H_A : p(\mathbf{x}) \neq q(\mathbf{y})$. An ideal test should have power against all alternatives. That is, as $n, m \to \infty$, the test will always reject when $p \neq q$ for any non-zero significance level $\alpha$.

The $t$-test possesses a storied history, with its genesis at Guinness Brewery in Dublin [10]. At the time, Guinness made a habit of employing the finest chemistry graduates of Oxford and Cambridge as brewers and managers, among whom was W.S. Gosset, better known as Student. Gosset, appointed brewer in 1899, became Brewer-in-Charge of the Experimental Brewery and was tasked with analyzing the results of barley-breeding experiments. After consulting with Karl Pearson, Gosset worked on determining the probable error of the mean and tabulated these results for various sample sizes.

Gosset successfully analyzed the factors that influenced barley yield and quality. As a result, Guinness purchased as much Danish Archer barley seed as possible and allowed Gosset to publish his methodology under the pseudonym Student in [91] and [90].

Fisher later developed a non-parametric, permutation-based two-sample test [28]. Hotelling [43] devised a parametric, multivariate generalization of the $t$-test. Friedman and Rafsky [33] used minimum spanning trees to extend the non-parametric Kolmogorov-Smirnov test [83] to a multivariate setting. In Section 5.6, we describe the Maximum Mean Discrepancy test of Gretton et al. [36], the first kernel-based two-sample test.

## 5.3 The Friedman Two-Sample Test

Friedman proposed the following approach to the two-sample problem [32]:

For $\{\mathbf{x}_i\}_1^n$ drawn from $p(\mathbf{x})$ and $\{\mathbf{x}_i\}_{n+1}^{n+m}$ drawn from $q(\mathbf{x})$, we would like to test $\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$.

1. Assign a response value $l_i = 1$ to the observations from the first sample ($1 \leq i \leq n$) and $l_i = -1$ to the observations from the second sample ($n+1 \leq i \leq n+m$).

2. Apply a binary classification learning machine to the training data to produce a scoring function $f(\mathbf{x})$ to score each of the observations $\{s_i = f(\mathbf{x}_i)\}_1^{n+m}$.

3. Calculate a univariate two-sample test statistic $\hat{t} = T(\{s_i\}_1^n, \{s_i\}_{n+1}^{n+m})$.

4. Determine the permutation null distribution of the above statistic to yield a p-value.

5. The test rejects $\mathcal{H}_0$ at significance level $\alpha$ if $p < \alpha$.

Note that in Step 2, for a given learning machine, there can still be some choice in the scoring function $f(\mathbf{x})$. In Section 5.5, we shall consider the merits of different scoring functions $f$ producing correspondingly different scores $s_i$. Although Friedman suggested using the $t$-statistic, other statistics are valid as well such as the Kolmogorov–Smirnov statistic [48, 84]. Given that the $t$-statistic has no power to detect differences in variance, sensitivity to such a change must come from the score mapping $f$. For instance, the mapping $f(x) = x^2$ would imbue the $t$-statistic with the power to detect variance shifts.

The Friedman test (FT) is a simple, elegant idea that leverages the many advancements made over the past several decades in the fields of prediction and classification and applies them to the problem of two-sample testing. In short, as long as there exists a learning machine for the problem at hand, the Friedman test provides a recipe for turning that learning machine into a two-sample test. This immediately yields two-sample tests for many kinds of data, including all types for which kernels have been defined. But there still remains some choice in the scoring function $f(\mathbf{x})$. It must be flexible enough to discriminate between the potential distributional differences of the problem at hand. The operating characteristics of the new two-sample test is *solely* a function of the paired learning algorithm. For an excellent overview of many statistical learning algorithms, see Hastie et al. [39].

By virtue of its permutation construction, the test has level $\alpha$—the probability that we reject the null hypothesis given that the null hypothesis is true, also known

as type I error. Given a threshold $\alpha$, we wish to minimize the type II error, accepting the null hypothesis given that the alternative hypothesis is true. Equivalently, we wish to maximize the power, one minus the type II error [54]. The downside of the permutation design is, of course, that any computational cost is naïvely multiplied by the number of permutations. However, there are many situations for which the cost is sublinear in the number of permutations. For instance, caching the computation of the kernel matrix yields substantial savings when re-using it for permutation based inference. This is especially true when computation of the kernel matrix is expensive relative to finding the SVM parameters via quadratic programming.

The exact randomization distribution will be a complicated, discrete distribution parametrized by the observed data. If, however, we can approximate this distribution with a simpler one and derive error bounds on the difference between the two distributions in some probability metric, then we can use the target distribution as a basis for inference. We will gain in computational efficiency by only having to compute the test statistic once.

## 5.4 Kernel Methods

There exist many two-sample tests for vectorial data $\mathbf{x}_i \in \mathbb{R}^p$. Increasingly, data collected for many applications is heterogeneous in nature and include non-vectorial components such as text, audio, or graph structures for which the mathematical and geometric operations required of many learning algorithms are not defined. Kernel methods allow us to identify a mapping of the data from a general set into a Hilbert space in which we can apply certain classes of algorithms. For instance, Haussler [40] developed ways of constructing kernels between discrete structures such as strings, trees, and graphs. There is much literature on kernel methods, but one particularly comprehensive treatment is the monograph by Schölkopf and Smola [76].

Given $n$ observed datapoints in some general set, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, kernelized learning algorithms depend only on the pairwise "similarities" between any two observations by way of the kernel function. Thus, kernel methods effectively decouple the algorithm (e.g. a support vector machine) from the representation of the data (e.g.

a particular kernel).

**Definition 5.1** (Positive Semidefinite Kernel)**.** *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive semidefinite kernel iff it is symmetric ($K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$) and positive semidefinite:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

*for any $n > 0$, any choice of $n$ objects $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, and any $c_1, \ldots, c_n \in \mathbb{R}$.*

Because inner products are symmetric, positive semidefinite functions, they satisfy Definition 5.1 and are valid kernels. When $\mathcal{X} = \mathbb{R}^p$, the linear kernel is defined as

$$K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$$

for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

For objects in a general set $\mathcal{X}$, we can define a mapping $\phi$ into a Hilbert space $\mathcal{H}$ for which an inner product exists. As there are many such mappings, one challenge is to choose one that is maximally useful in exploiting the structure of the data for the task at hand. In Chapter 6, we shall explore a technique to identify the most useful mapping given some parametrized space of mappings.

In fact, for every kernel $K$ we can identify a feature mapping into a Hilbert space, where the kernel can be expressed as the inner product of the mapped features [77]:

**Theorem 5.2.** *For any kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a Hilbert space $\mathcal{H}$ and a feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that*

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$$

*for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, where $\langle u, v \rangle_{\mathcal{H}}$ represents the inner product in $\mathcal{H}$.*

In the coming sections and in Chapter 6, we shall see examples of nonvectorial spaces $\mathcal{X}$, kernels $K$, feature mappings $\phi$, and Hilbert spaces $\mathcal{H}$.

## 5.5 Support Vector Machines

A Support Vector Machine (SVM) [16] is a supervised learning technique that seeks to find a hyperplane that maximizes the margin between points of different classes. In the case that there exists no separating hyperplane, a regularization term can be added that controls the effect of misclassified points.

Although SVMs find linear decision boundaries, the algorithm depends only on inner products between its datapoints. The "kernel trick" [1] allows us to replace the inner products with kernel function evaluations, thus effectively finding a linear decision boundary in the Hilbert space $\mathcal{H}$ identified by the feature mapping $\phi(x) : \mathcal{X} \to \mathcal{H}$. Although linear in the typically-higher-dimensional space $\mathcal{H}$, the decision boundary can be nonlinear in $\mathcal{X}$.

Since kernel methods divorce the representation of the data with the learning algorithm, development on both fronts can proceed independently. For instance, faster optimization algorithms for solving the general SVM problem can proceed in parallel with the problem-specific designing of new kernels to more efficiently or effectively exploit the structure of the data.

Consider the $\ell_1$-regularized (soft margin) support vector classification problem [76] in its primal form:

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H}, b\in\mathbb{R}, \xi\in\mathbb{R}^{n+m}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n+m}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^t\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \qquad\qquad \text{for all } i = 1, \ldots, n+m. \end{aligned} \tag{5.1}$$

There are three obvious possibilities for the Friedman scoring function $f$:

1. The predicted class label $f_1(\mathbf{x}_i) = \text{sign}(\mathbf{w}^t\mathbf{x}_i + b)$.

2. An estimate of the posterior class probability, such as by a sigmoid (Platt [67, 57]) or a logistic link function (Wahba [94, 95]) $f_2(\mathbf{x}_i) = \frac{1}{1+\exp(-(\mathbf{w}^t\mathbf{x}_i+b))}$.

3. The margin $f_3(\mathbf{x}_i) = \mathbf{w}^t\mathbf{x}_i + b$.

Because the predicted class label is simply the sign of the margin, we lose information about how likely it is for a given observation to belong to a particular class. Moreover, given constant within-class predicted class labels, the sample standard deviation is 0 and hence the $t$-statistic is unbounded.

Using, for instance, a logistic link function, $f_2$ has the interpretability of being a posterior class probability–if one believes in the probability model–and yields no information loss since it is simply an invertible function of the margin. However, it is typically not a linear function of the margin.

The margin $f_3$ has the advantage of being an affine function of the data. In one dimension, $f_3(x_i) = wx_i + b$, and since the $t$-statistic is invariant (up to sign) to affine transformations of the data, we can see that using the margin generalizes the permutation $t$-test in some sense.

In the univariate setting, whether or not the $t$-statistic computed on the scores $\{f_3(x_i)\}$ agrees with that on the raw data $\{x_i\}$ depends on sign$(w)$. Due to symmetry, in the permutation null distribution, $w$ is negative with probability .5 and positive with probability .5. Thus, the permutation null in both settings appears to be $t$ and hence asymptotically normal.

### 5.5.1 Kernelized Form

It is advantageous to treat the dual [11] of Problem (5.1). Because of strong duality, the primal and dual solutions are equivalent. Although both optimization problems are quadratic programs, there exist fast algorithms such as the sequential minimal optimization algorithm [68] that exploit the special structure of the dual problem.

In addition, the dual problem is expressed only in terms of inner products, $\langle \mathbf{x_i}, \mathbf{x_j} \rangle$. The kernel trick [1] amounts to replacing these inner products with kernel function

evaluations, $K(\mathbf{x_i}, \mathbf{x_j})$. The dual optimization problem is given by

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^{n+m}}{\text{minimize}} \quad & \sum_{i=1}^{n+m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n+m} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \quad \text{for all } i = 1, \dots, n+m \\ \text{and} \quad & \sum_{i=1}^{n+m} \alpha_i y_i = 0. \end{aligned} \tag{5.2}$$

The Karush–Kuhn–Tucker (KKT) [49] conditions for optimality imply that

$$\mathbf{w} = \sum_{i=1}^{n+m} \alpha_i y_i \mathbf{x}_i.$$

Therefore,

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^{n+m} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b,$$

where $\alpha_i$ are the dual variables.

In fact, there is another view of the SVM problem in the framework of regularized empirical risk minimization [92] that will be useful for Chapter 6. Define the hinge loss function

$$L(y, f(\mathbf{x})) := (1 - yf(\mathbf{x}))_+ := \max(0, 1 - yf(\mathbf{x})).$$

Then the following optimization problem is equivalent to Problem (5.1):

$$\min_{b \in \mathbb{R}, \mathbf{w} \in \mathcal{H}} \sum_{i=1}^{n+m} (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))_+ + \frac{1}{2C} \|\mathbf{w}\|_2^2. \tag{5.3}$$

It turns out that regularized risk minimization problems admit particularly elegant solutions, a result owing to Kimeldorf and Wahba [46]. We present a slightly generalized representer theorem from [76]:

**Theorem 5.3** (Representer Theorem)**.** *Let $\Omega : [0, \infty] \to \mathbb{R}$ be a strictly monotonic increasing function, $\mathcal{X}$ be a set, and $c : (\mathcal{X} \times \mathbb{R}^2)^{n+m} \to \mathbb{R} \cup \{\infty\}$ be an arbitrary loss*

*function. Then each minimizer $f' \in \mathcal{H}$ of the regularized risk*

$$c((\mathbf{x}_1, y_1, f'(\mathbf{x}_1)), \dots, (\mathbf{x}_{n+m}, y_{n+m}, f'(\mathbf{x}_{n+m}))) + \Omega(\|f'\|_{\mathcal{H}}) \tag{5.4}$$

*admits a representation of the form*

$$f'(\mathbf{x}) = \sum_{i=1}^{n+m} \alpha'_i K(\mathbf{x}_i, \mathbf{x}).$$

Although Problem (5.4) has a feasible set $\mathcal{H}$ that is possibly infinite dimensional, Theorem 5.3 guarantees that the solution lies in the span of the $n + m$ particular kernels centered on the observations $\mathbf{x}_i$.

We apply Theorem 5.3 to Problem (5.3), noting that it holds for all fixed $b$, to conclude that

$$f'(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n+m} \alpha'_i K(\mathbf{x}_i, \mathbf{x}).$$

Thus, setting $\alpha'_i = y_i\alpha_i$, we again find

$$f(\mathbf{x}) = f'(\mathbf{x}) + b = \sum_{i=1}^{n+m} y_i\alpha_i K(\mathbf{x}, \mathbf{x}_i) + b.$$

We list the kernelized representations of possible Friedman scoring functions:

1. The predicted class label $f_1(\mathbf{x}_i) = \text{sign}(\sum_{i=1}^{n+m} y_i\alpha_i K(\mathbf{x}, \mathbf{x}_i) + b)$.
2. An estimate of the posterior class probability, such as by a sigmoid (Platt [67, 57]) or a logistic link function (Wahba [94, 95]) $f_2(\mathbf{x}_i) = \frac{1}{1+\exp(-(\sum_{i=1}^{n+m} y_i\alpha_i K(\mathbf{x},\mathbf{x}_i)+b)}$.
3. The margin $f_3(\mathbf{x}_i) = \sum_{i=1}^{n+m} y_i\alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$.

### 5.5.2 Equivalence to the Permutation $t$-test

**Theorem 5.4.** *The Friedman test paired with support vector regression or support vector classification (using the margin as a score) with the appropriate kernel generalizes the two-sample permutation t-test. In particular, the two procedures are equivalent with univariate data and a linear kernel.*

*Proof.*

$$f(x) = \sum_{i=1}^{n+m} y_i \alpha_i K(x_i, x) + b = \left( \sum_{i=1}^{n+m} y_i \alpha_i x_i \right) x + b = wx + b,$$

since we have univariate data and an affine kernel. Therefore, the SVM score is simply an affine transformation of the data. Welch's $t$-statistic is given by

$$T(\{x_i\}_1^n, \{x_i\}_{n+1}^{n+m}) = \frac{\bar{x} - \bar{x}'}{\sqrt{\frac{s_X^2}{n} + \frac{s_{X'}^2}{m}}}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Let $z = f(x) = wx + b$ and note that

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{w}{n} \sum_{i=1}^{n} x_i + b = w\bar{x} + b$$

and

$$s_Z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (wx_i + b - w\bar{x} + b)^2 = w^2 s_X^2.$$

Therefore,

$$T(\{f(x_i)\}_1^n, \{f(x_i)\}_{n+1}^{n+m}) = \frac{w\bar{x} + b - w\bar{x}' + b}{|w|\sqrt{\frac{s_X^2}{n} + \frac{s_{X'}^2}{m}}} = \text{sign}(w) T(\{x_i\}_1^n, \{x_i\}_{n+1}^{n+m}).$$

Since we are interested in two-sided testing, we consider

$$|T(\{f(x_i)\}_1^n, \{f(x_i)\}_{n+1}^{n+m})| = |T(\{x_i\}_1^n, \{x_i\}_{n+1}^{n+m})|.$$

Thus, the $t$-statistics are identical, and since the permutation procedure is the same, the tests are equivalent.

To see the result for support vector regression, recall that support vector regression

solves the following problem [76]:

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H},\boldsymbol{\xi}^{(*)}\in\mathbb{R}^{n+m},b\in\mathbb{R}}{\text{minimize}} \quad & \tau\left(\mathbf{w},\boldsymbol{\xi}^{(*)}\right) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n+m}(\xi_i+\xi_i^*) \\ \text{subject to} \quad & f(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i \\ & y_i - f(\mathbf{x}_i) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \qquad\qquad \text{for all } i = 1, \ldots, n+m. \end{aligned}$$

with solution is given by

$$f(x) = \sum_{i=1}^{n+m}(\alpha_i^* - \alpha_i)K(x_i, x) + b.$$

□

Using the bounds derived in Chapter 3, we can conduct statistical inference with the normal distribution rather than trying to compute the randomization distribution.

We have shown that for univariate data, some kernels generalize the permutation $t$-test. Is it possible to characterize all such kernels? For what kernels $K$ do we have

$$\sum_{i=1}^{n+m} y_i\alpha_i K(x, x_i) + b = cx + d? \tag{5.5}$$

A sufficient condition is for $K(x, x_i) = \langle\Phi(x), \Phi(x_i)\rangle = f(x_i)x$. The linear kernel satisfies this condition with $f(x_i) = x_i$. The RBF kernel $K(x, x_i) = \exp(-\sigma(x - x_i)^2)$ does not yield an affine function of the data, as

$$\sum_{i=1}^{n+m} y_i\alpha_i\exp(-\sigma(x - x_i)^2) + b \tag{5.6}$$

cannot be written as $cx + d$.

We use Support Vector Machine (SVM) classification as implemented in the **ksvm** function of the **R** [71] package **kernlab** [45].

The cost parameter $C$ controls the complexity of the prediction function. It is typically chosen via cross-validation over a grid of choices.

## 5.6 Maximum Mean Discrepancy

Gretton et al. [36, 38, 37, 9] introduced a kernel-based approach for the two-sample problem based on the Maximum Mean Discrepancy (MMD) statistic, an integral probability metric. MMD provides good performance in practice, strong theoretical guarantees, and is the first two-sample test for comparing distributions over graphs.

**Definition 5.5.** *With $\mathfrak{F}$ a class of functions $f : \mathcal{X} \to \mathbb{R}$, $p$ and $q$ probability distributions, and $X \sim p$ and $Z \sim q$ random variables, the maximum mean discrepancy (MMD) and an empirical estimate are defined as*

$$MMD[\mathfrak{F}, p, q] := \sup_{f \in \mathfrak{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{z \sim q}[f(z)]),$$

$$MMD[\mathfrak{F}, X, Z] := \sup_{f \in \mathfrak{F}} \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \right).$$

An unbiased empirical estimate [37] estimate of the statistic is given by

$$\mathrm{MMD}_u^2[\mathfrak{F}, X, Z] := \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} K(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K(z_i, z_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} K(x_i, z_j).$$

The function class $\mathfrak{F}$ is typically taken to be the unit ball in a reproducing kernel Hilbert space (RKHS), however, well-known metrics can be obtained over other function classes. Although Gretton et al. provide several distribution-free tests based on MMD theory, we instead compare the Friedman test (FT) against the permutation-based MMD so as to compare statistic with statistic. In this way, the theory is dissociated from the comparison. We feel that this is the most fair comparison of the two tests because many of the theoretical results are inexact. We also do not have large enough sample sizes in our real datasets to ensure low error in theoretical approximations. Even if we did, the power for the tests would be very nearly one, making comparisons on non-simulated data difficult.

The Kernel MMD (KMMD) test seeks "smooth" functions that maximize the difference between the two classes of points, where smoothness is defined in terms

of the Hilbert norm. This allows for nonlinear functions $f$ in the feature space, as opposed to the hyperplanes learned by the SVM algorithm.

## 5.7 Null Distributions

The null distribution plays a fundamental role in frequentist statistical inference. Hotelling's $T^2$-statistic has null distribution that corresponds to a scaled central $F_{(p,n+m-1-p)}$ distribution, where $p$ is the dimensionality of the data and $n, m$ are the sample sizes of the two groups. As its name suggests, the $T^2$-test is a generalization of Student's $t$-test, and for $T \sim t(n+m-2)$, we have that $T^2 \sim F_{(1,n+m-2)}$. As a consequence of Theorem 5.4, the Friedman statistic in the univariate data, linear kernel setting is equal to the $|T|$. In Figure 5.2 we simulate 200 standard multivariate normal draws from each class with dimension $D \in \{1, 5, 10\}$. We compare the null distributions of the $T^2$-statistic, KMMD, and Friedman statistics with a linear kernel and RBF kernel with width parameter 1. We draw 5,000 samples from each permutation null distribution and apply a kernel density smoother to the results. It appears that many of the null distributions are very close to normal (or, $t(398)$, rather).

The Friedman Statistic null distributions appear to be consistent with a standard normal distribution. In Figure 5.1 we examine the relationship between the Friedman Statistic and the regularization parameter, $C$. We have proven that in the univariate setting with a linear kernel, the statistic is independent of $C$. Empirically, it appears that for higher dimensional data, $C$ has a minor effect on the Friedman statistic for the linear kernel. With an RBF kernel with width 1, it appears that $C$ has a small effect.

The $T^2$ densities correspond to a parametrized family of $F$-distributions. It is not surprising that the MMD linear kernel null distributions shift rightward as a function of dimension: the higher dimensionality allows the function in the RKHS to better find discrepancies between the two empirical distributions. The same rationale holds true for the FS when thinking of separating hyperplanes. Interestingly, there are marked differences between the MMD and FS for the RBF kernel. Note also that the support of the KMMD statistic is $\mathbb{R}^+$.

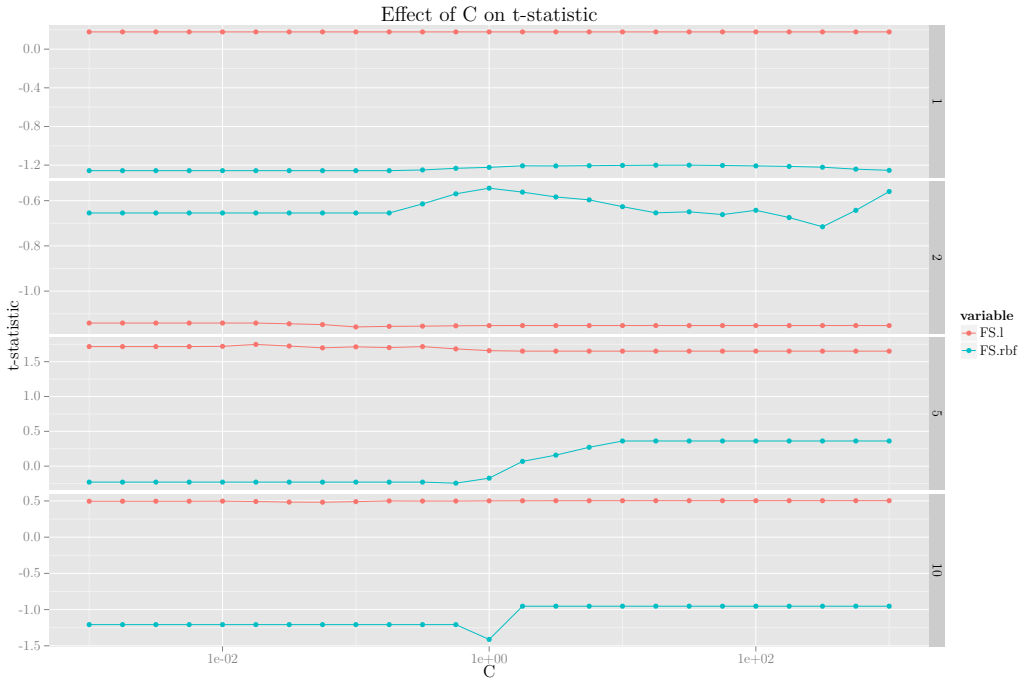


Figure 5.1: FS.l: FS with a linear kernel; FS.rbf: FS with RBF kernel. We vary the dimension of the data: 1, 2, 5, 10.

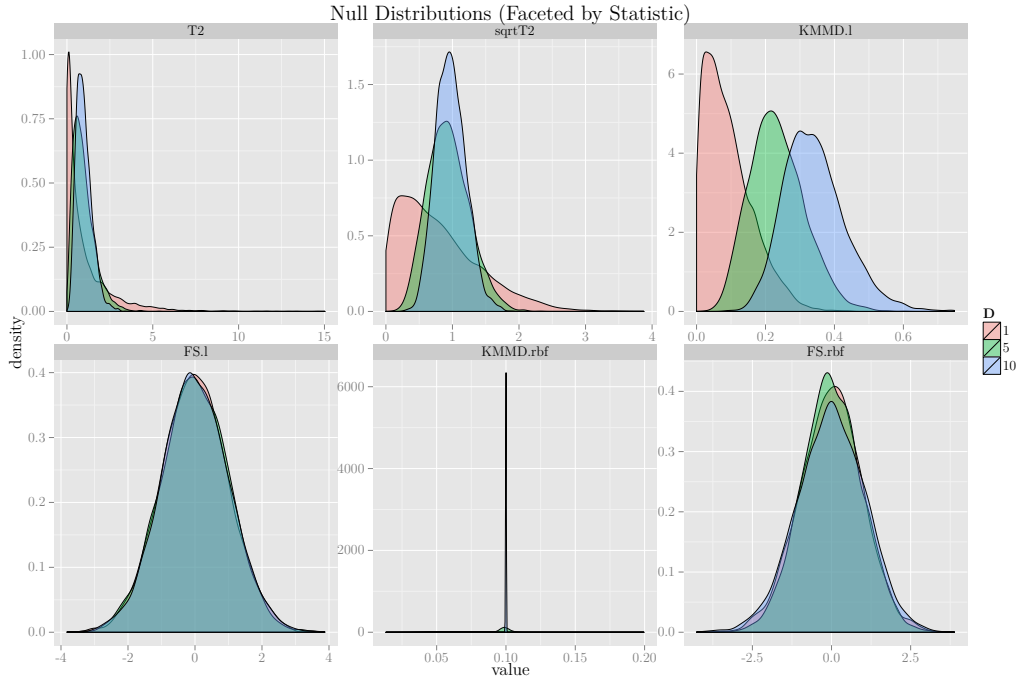


Figure 5.2: T2: Hotelling's $T^2$-statistic; sqrtT2: $|T|$; KMMD.l: kernel MMD with a linear kernel; FS.l: FS with a linear kernel; KMMD.rbf: kernel MMD with a radial basis function (RBF) kernel; FS.rbf: FS with RBF kernel

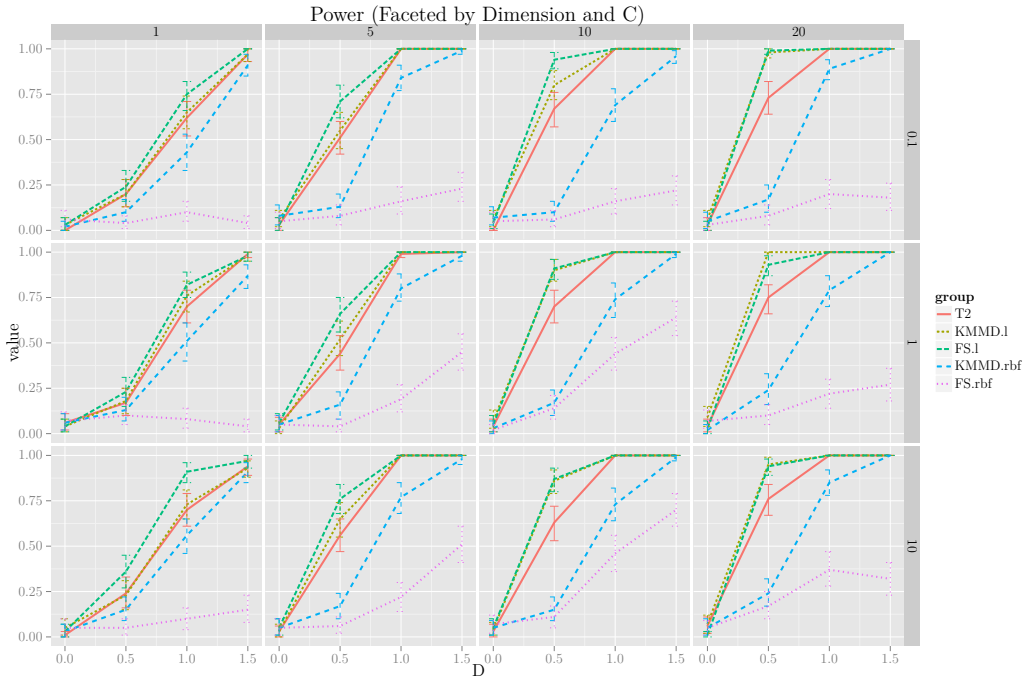


Figure 5.3: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; T2: Hotelling's $T^2$-statistic; Error bars indicate 95% bootstrap confidence intervals.

## 5.8 Experiments

### 5.8.1 Vectorial Data

We consider $\{x_i\}_{i=1}^{20} \sim \text{MVN}_d(\mathbf{0}, \mathbf{I})$ and $\{y_i\}_{i=1}^{20} \sim \text{MVN}_d(\Delta\mathbf{1}, \mathbf{I})$ where our dimensionality $d \in \{1, 5, 10, 20\}$ and mean difference $\Delta \in \{0, .5, \ldots, 1.5\}$ in Figure 5.3.

For FS and MMD, we use the the RBF kernel with a width of 1. The methods perform similarly with the exception of the kernel methods using the RBF kernel. This suggests that either a width of 1 is ineffective or the RBF kernel is unsuitable for these data.

### 5.8.2 String Data

For a string data comparison using data from Twitter, we collected the latest 1,000 tweets from Barack Obama (@BarackObama) and Sarah Palin (@SarahPalinUSA) obtained from the **R** package **twitteR** [34]. We pre-process and transform the tweets into an alphabet comprising only lowercase English letters and spaces. For simplicity, we choose the $k$-spectrum kernel [55] with $k \in \{1, 2, 3\}$ as our kernels for both the FT and MMD.

Here is an example of the raw tweets:

```
"BarackObama: We need to reward education reforms that are
driven not by Washington, but by principals and teachers and
parents. http://OFA.BO/6p2EMy"
"SarahPalinUSA: You betcha!! MT \"@AlaskaAces: Alaska Aces
are 2011 Kelly Cup Champs w/ 5-3 win over Kalamazoo Wings!
Aces win  ECHL Championship series 4-1\""
```

And pre-processed tweets:

```
"we need to reward education reforms that are driven not by
washington but by principals and teachers and parents "
"you betcha mt alaskaaces alaska aces are  kelly cup champs
w  win over kalamazoo wings aces win  echl championship
series "
```

For the $k$-spectrum kernel,

- $\mathcal{X}$ = set of all finite-length sequences from an alphabet $\mathcal{A}$.
- $\phi(\mathbf{x})$ = the number of length $k$ contiguous subsequences ($k$-mers) in $\mathbf{x}$.
- $\mathcal{H} = \mathbb{N}^{|\mathcal{A}|^k}$.
- $K_k(\mathbf{x}, \mathbf{x}') = \langle \phi_k(\mathbf{x}), \phi_k(\mathbf{x}') \rangle$.

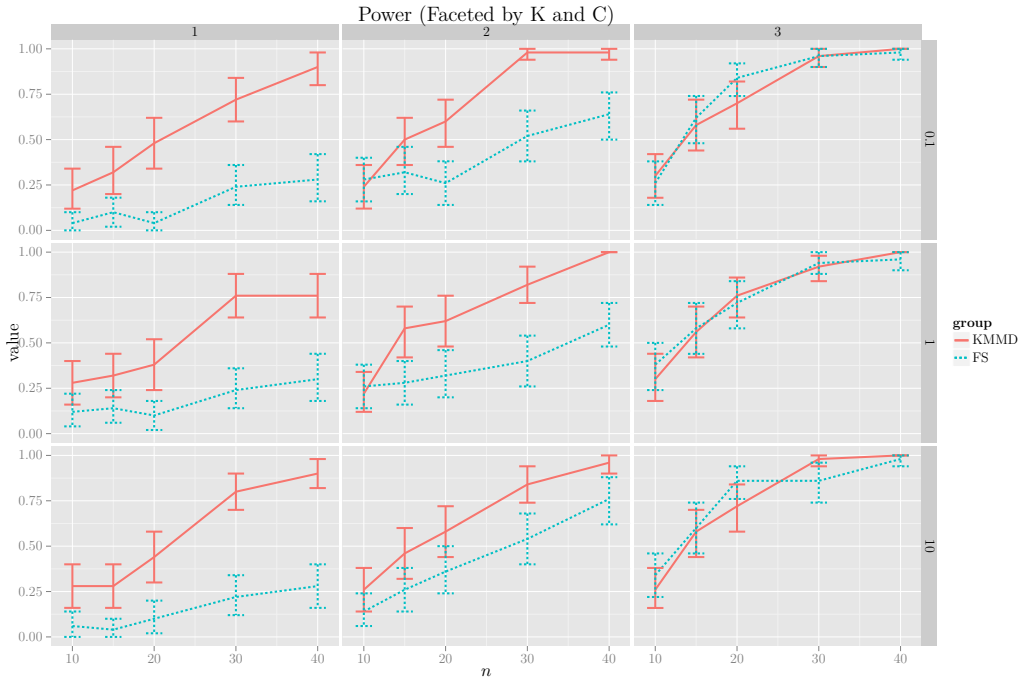


Figure 5.4: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; Error bars indicate 95% bootstrap confidence intervals.

Suffix trees allow for efficient kernel calculations, computing $K_k(\mathbf{x}, \mathbf{x}')$ in $\mathcal{O}(kn)$ time.

We draw samples of various sizes from both the Barack Obama tweets and Sarah Palin tweets in order to empirically determine the power, with results detailed in Figure 5.4.

The MMD test outperforms the Friedman test on this task for $k < 3$. Power increases as a function of $k$ for both tests, and it is somewhat surprising to see the strong performance from considering only frequencies of unigrams. The KMMD's strong performance is likely due to the greater flexibility in being able to learn a nonlinear function in the smaller feature spaces corresponding to $k = 1$ and 2. We see that the advantage largely disappears for $k = 3$.

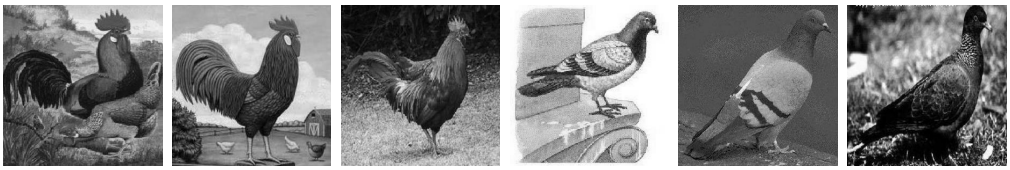

Figure 5.5: Images of roosters and pigeons for use in discrimination test.

### 5.8.3 Image Data

We consider the task of discriminating between images of roosters and pigeons from the Caltech 101 Object Categories dataset [56]. Samples of the birds are in Figure 5.5. We resize images to a common resolution of $300 \times 297$ and convert to a vector of 8 bit grayscale values. To correct for global differences in illumination and ensure that only local patterns would be used for discrimination, we centered and scaled each vector. Power comparisons can be seen in Figure 5.6.

For the polynomial kernel,

- $\mathcal{X} = \mathbb{R}^n$.
- $\phi_2([x_1, x_2]) = [x_1^2, 2x_1x_2, x_2^2, \sqrt{2c}x_1, \sqrt{2c}x_2, c]$.
- $\langle \phi_2(\mathbf{x}), \phi_2(\mathbf{x}') \rangle$ is $\mathcal{O}(n^2)$.
- $\mathcal{H} = \mathbb{R}^{d'}$, where $d' = \binom{n+d}{d}$.
- $K_d(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + c)^d$ is $\mathcal{O}(n)$.

Again, MMD performs better. However, the Friedman test's performance certainly improves when considering higher degree polynomials.

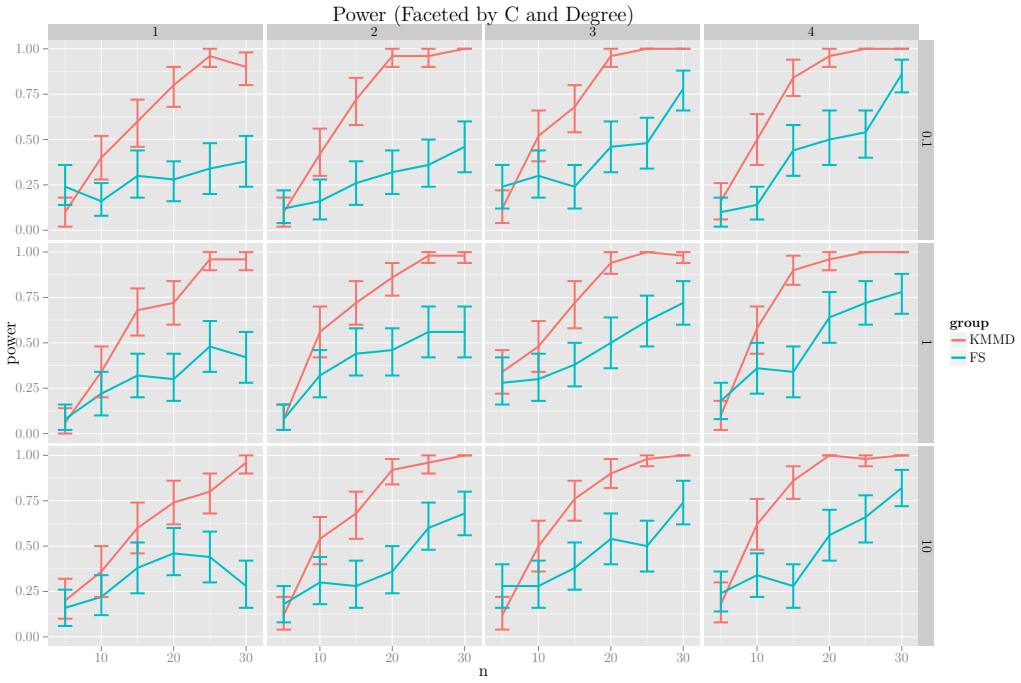


Figure 5.6: KMMD: the kernel Maximum Mean Discrepancy test; FS: the Friedman test; Columns indicate the degree of the polynomial kernel; Rows indicate the regularization parameter $C$; Error bars indicate 95% bootstrap confidence intervals.

# Chapter 6

# Multiple Kernels

In this chapter we introduce a framework for two-sample testing of heterogeneous data via multiple kernel learning (MKL).

## 6.1 Introduction

Given a set of kernels, it is possible to combine them in order to produce new kernels. This is a starting point for heterogeneous data analysis: we can define a kernel $K_i$ for each data domain and develop a kernel $K$ that operates on the union of the domains. We typically shall produce a parametrized family of kernels $K_\theta$ and seek the "best" choice of parameters $\theta$.

For example, the class of kernel functions on $\mathcal{X}$ is closed under pointwise products (also known as Schur products) of kernels,

$$K(\mathbf{x}, \mathbf{x}') = (K_1 K_2)(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}')K_2(\mathbf{x}, \mathbf{x}'),$$

tensor products of kernels,

$$K(\mathbf{x}, \mathbf{x}') = (K_1 \otimes K_2)(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1', \mathbf{x}_2') = K_1(\mathbf{x}_1, \mathbf{x}_1')K_2(\mathbf{x}_2, \mathbf{x}_2'),$$

and conic combinations of kernels,

$$K_\theta(\mathbf{x}, \mathbf{x}') = (\theta_1 K_1 + \ldots + \theta_m K_m)(\mathbf{x}, \mathbf{x}') = \theta_1 K_1(\mathbf{x}, \mathbf{x}') + \ldots + \theta_m K_m(\mathbf{x}, \mathbf{x}').$$

An unweighted sum of kernels is equivalent to concatenating the individual feature spaces.

## 6.2 Multiple Kernel Learning

In a landmark paper, Lanckriet et al. [50] showed that for various SVM objective functions, the problem of finding the optimal conic combination of kernels could be posed as a convex optimization problem. Although the initial approach involved solving a computationally expensive semidefinite program, this sparked a flurry of research on similar convex approaches to MKL.

Kloft et al. [47] conceived a general $\ell_p$-norm approach to MKL, unifying many special cases and further proposed a highly efficient algorithm. This can be seen as generalizing Problem (5.3) of Chapter 5.

Let $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss function, $\Omega : \mathcal{H} \to \mathbb{R}$ be a regularizer, and $\lambda > 0$ be a tradeoff parameter.

Kloft et al. consider linear models of the form

$$h_{\tilde{w},b,\theta}(\mathbf{x}) = \sum_{i=1}^{M} \sqrt{\theta_m} \langle \tilde{w}_m, \phi_m(\mathbf{x}) \rangle_{\mathcal{H}_m} + b = \langle \tilde{w}, \phi_\theta(\mathbf{x}) \rangle_{\mathcal{H}} + b,$$

where $\tilde{w} = [\tilde{w}_1^T, \ldots, \tilde{w}_m^T]^T$ and $\phi_\theta = \sqrt{\theta_1}\phi_1 \times \ldots \times \sqrt{\theta_m}\phi_m$.

The regularized risk minimization problem is the following:

$$\min_{\tilde{w},b,\theta:\theta\succeq 0} \frac{1}{n} \sum_{i=1}^{n} L\left( \sum_{m=1}^{M} \sqrt{\theta_m} \langle \tilde{w}_m, \phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m} + b, y_i \right) + \frac{\lambda}{2} \sum_{m=1}^{M} \|\tilde{w}_m\|_{\mathcal{H}_m}^2 + \tilde{\mu}\tilde{\Omega}(\theta), \tag{6.1}$$

for $\tilde{\mu} > 0$.

Problem (6.1) is not convex but can be transformed into a convex problem via

the substitution

$$w_m \leftarrow \sqrt{\theta_m}\tilde{w}_m.$$

Decoupling the regularization parameter from the sample size by letting $\tilde{C} = \frac{1}{n\lambda}$ and $\mu \leftarrow \frac{\tilde{\mu}}{\lambda}$, and using convex regularizers of the form $\tilde{\Omega}(\theta) = \|\theta\|_p^2$, we get

$$\min_{w,b,\theta:\theta\succeq 0} \tilde{C}\sum_{i=1}^{n} L\left(\sum_{m=1}^{M}\langle w_m, \phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m} + b, y_i\right) + \frac{1}{2}\sum_{m=1}^{M}\frac{\|w_m\|_{\mathcal{H}_m}^2}{\theta_m} + \mu\|\theta\|_p^2, \quad (6.2)$$

where $\frac{t}{0} = 0$ if $t = 0$ and $\infty$ otherwise.

Kloft et al. proved that the Tikhonov-regularized Problem (6.2) with two parameters is in fact equivalent to the following Ivanov-regularized formulation with one regularization parameter, $C$:

$$\begin{aligned} \underset{w,b,\theta:\theta\succeq 0}{\text{minimize}} \quad & C\sum_{i=1}^{n} L\left(\sum_{m=1}^{M}\langle w_m, \phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m} + b, y_i\right) + \frac{1}{2}\sum_{m=1}^{M}\frac{\|w_m\|_{\mathcal{H}_m}^2}{\theta_m} \\ \text{subject to} \quad & \|\theta\|_p^2 \leq 1. \end{aligned} \quad (6.3)$$

We use Problem (6.3) as implemented in SHOGUN [85].

## 6.3 Simulation

We generate heterogeneous data triplets $(\mathbf{x}_i, s_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^2$, $s_i \in \mathcal{S}$, the set of finite length sequences of {A, C, T, G}, and $y_i \in \{-1, 1\}$. $\mathbf{x}_i$ are drawn independently from the star example from [86].

The star has a single radius parameter, $r$. $p_{\mathbb{R}^2}$ takes $r = 4$, and $q_{\mathbb{R}^2}$ takes $r > 4$. Each $s_i$ has independent length $N \sim \text{Pois}(100)$ and is generated via a Markov chain. The start of the sequence is drawn from the stationary distribution $[.25, .25, .25, .25]$,

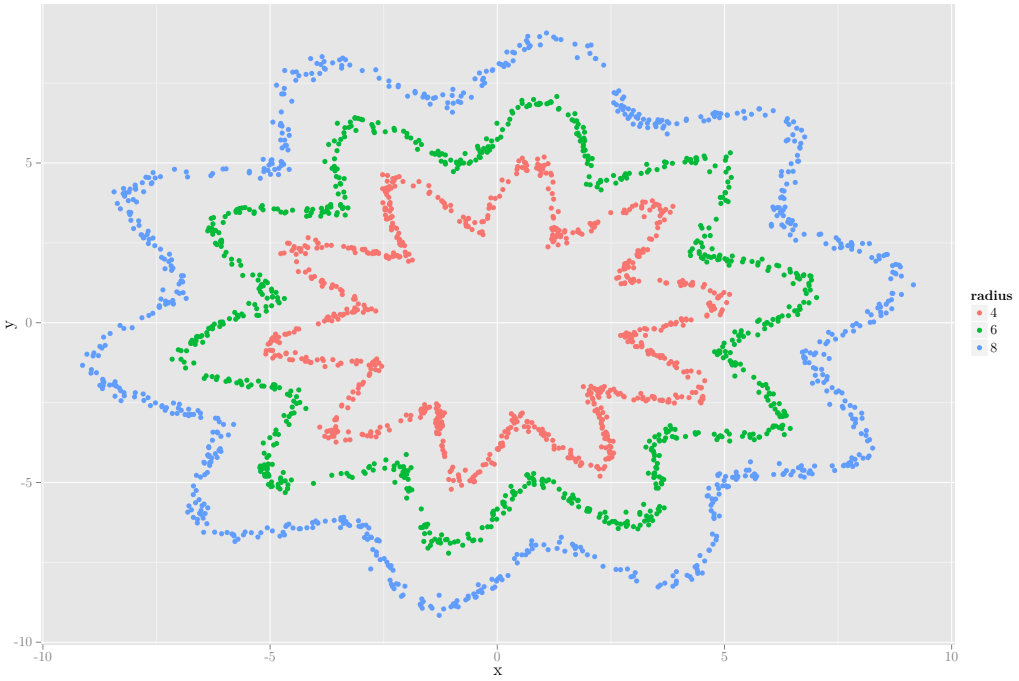


Figure 6.1: Star Distribution: radius 4 versus radius $> 4$.

and the following $N-1$ elements are drawn according to the transition probabilities

$$M(p^\star) = \begin{array}{c} \\ A \\ C \\ T \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & C & T & G \end{array} \\ \begin{pmatrix} \frac{1-p^\star}{3} & p^\star & \frac{1-p^\star}{3} & \frac{1-p^\star}{3} \\ \frac{1-p^\star}{3} & \frac{1-p^\star}{3} & p^\star & \frac{1-p^\star}{3} \\ \frac{1-p^\star}{3} & \frac{1-p^\star}{3} & \frac{1-p^\star}{3} & p^\star \\ p^\star & \frac{1-p^\star}{3} & \frac{1-p^\star}{3} & \frac{1-p^\star}{3} \end{pmatrix} \end{array}$$

$p_{\mathcal{S}}$ takes $p^\star_{\mathcal{S}} = .25$ and $q_{\mathcal{S}}$ takes $p^\star_{\mathcal{S}} > .25$. Note that $p_{\mathcal{S}}$ and $q_{\mathcal{S}}$ generate similar numbers of 1-mers, but $q_{\mathcal{S}}$ can generate more AC, CT, TG, GA 2-mers.

## 6.4 Kernel Normalization

Kernel normalization can have a large effect on the performance of kernel- and MKL-based algorithms. In the regularized regression setting, it is common to standardize variables to have mean zero and unit variance so that the results are invariant to differences in the unit of measurement. Normalization plays a similar role in MKL. For instance, the Gaussian RBF kernel has $K(\mathbf{x}, \mathbf{x}) = 1$. When considering long strings, it is common for the $k$-spectrum kernel to take on large values.

We pre-process all kernels used in MKL by ensuring that the vectors in the feature space lie on the unit hypersphere:

$$K_i(\mathbf{x}, \mathbf{x}') \leftarrow \frac{K_i(\mathbf{x}, \mathbf{x}')}{\sqrt{K_i(\mathbf{x}, \mathbf{x})}\sqrt{K_i(\mathbf{x}', \mathbf{x}')}}$$

## 6.5 MKL Weights

Typically, two-sample tests provide a single bit of information: accept or reject the null hypothesis that the two samples arose from the same distribution. The MKL algorithm–or any other learning procedure that generates interpretable weights–provides useful ancillary information in the kernel weights. Here we investigate the degree to which MKL is able to learn the structure of the data and identify the data

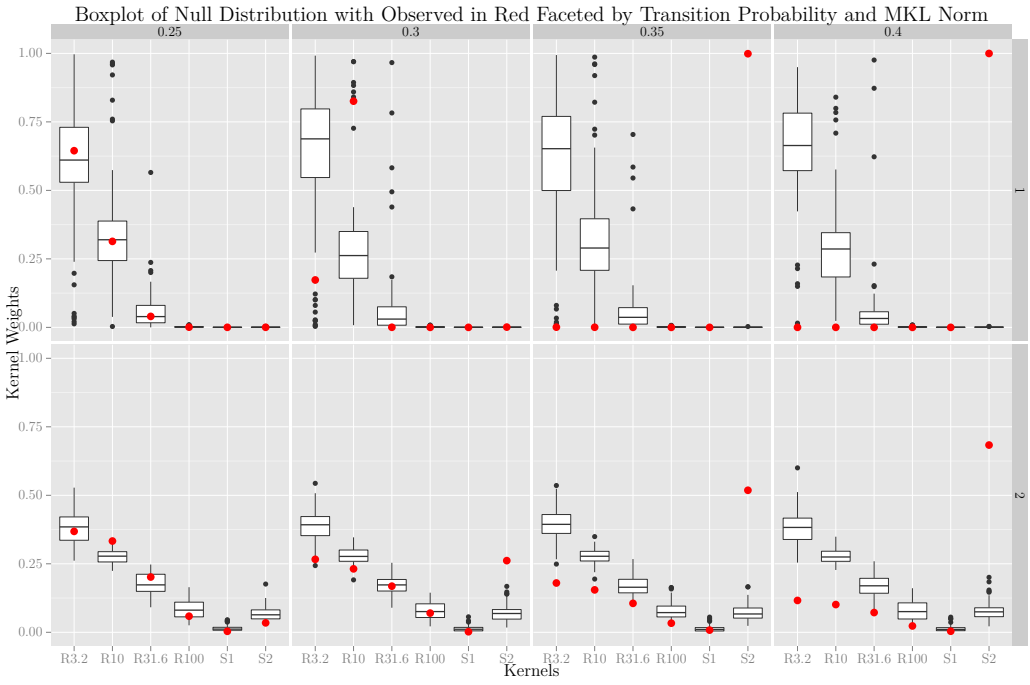


Figure 6.2: The MKL weights in the 1- (upper row) and 2-norm (lower row) cases shift progressively more to the 2-spectrum kernel as the DNA signal is increased.

domains with the highest signal in discrimininating between the two samples.

We include 4 Gaussian RBF kernels with widths $\{3.2, 10, 31.6, 100\}$ and 1- and 2-spectrum kernels. By design, only the 2-spectrum kernel should be discriminatory on the DNA string data. We perform both 1- and 2- norm MKL.

In Figure 6.2, we fix the radius of the inner star to be $r = 4$ and the outer star to be $r = 4.5$. We draw 200 samples from each distribution and vary the signal on the DNA string data by letting $p^\star$ vary in $\{.25, .3, .35, .4\}$. We fix the regularization parameter $C = 0.1$, and perform 100 permutations of the labels. The unpermuted weights are shown as red points, and the permuted labels give rise to boxplots of weights for both the 1-norm and 2-norm MKL.

We see that as we increase the difference between $p_{\mathcal{S}}$ and $q_{\mathcal{S}}$, more weight is being assigned to the 2-spectrum kernel. 1-norm MKL yields sparse weight vectors.

In Figure 6.3 we vary the radius of the outer star in $\{4, 7, 10, 13, 16\}$, while fixing

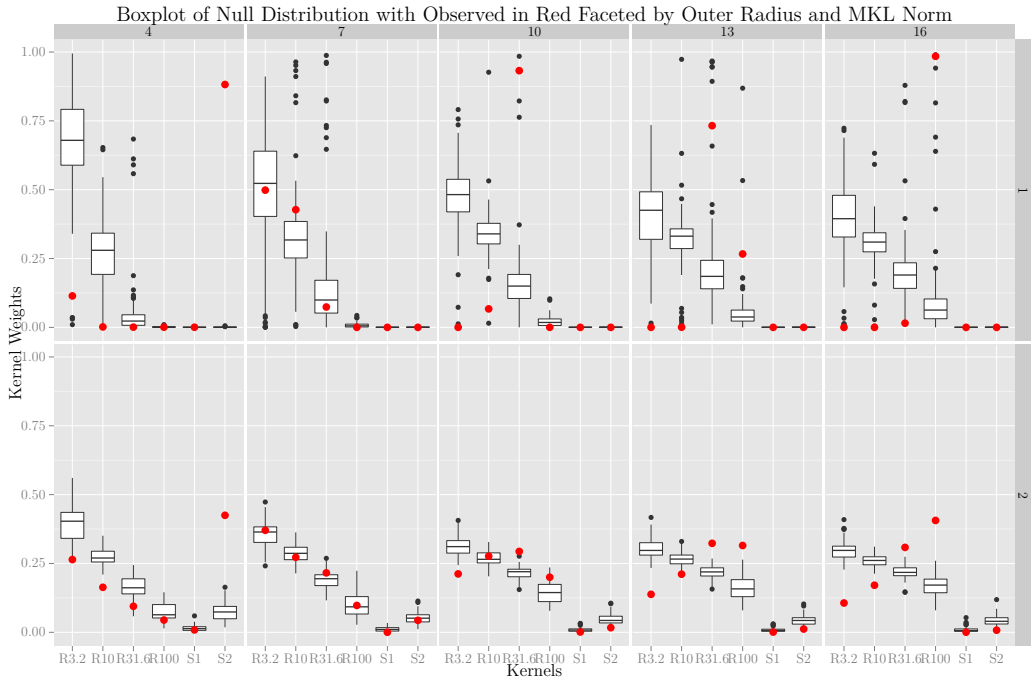


Figure 6.3: The MKL weights in the 1- (upper row) and 2-norm (lower row) cases shift progressively more to the higher-width RBF kernels as we increase the distance between the two stars.

the radius of the inner star to be 4 and the transition probability $p^\star = .3$. We see the dominant weight for the unpermuted case shift to higher-width kernels as we increase the radius of the outer star.

## 6.6 Power

Given that we have successfully learned the structure of the data, we now investigate the statistical power of these methods. For each simulation, we take 50 samples from each distribution and fix $C = 1$.

In Figure 6.4 we vary the radius of the outer star in the rows of the plot, taking values in $\{4, 4.3, 4.6\}$, and the $x$-axis sees the transition probability take values in $\{.25, .3, .35, .4, .45\}$. We compare the power of three RBF kernels with widths 5,

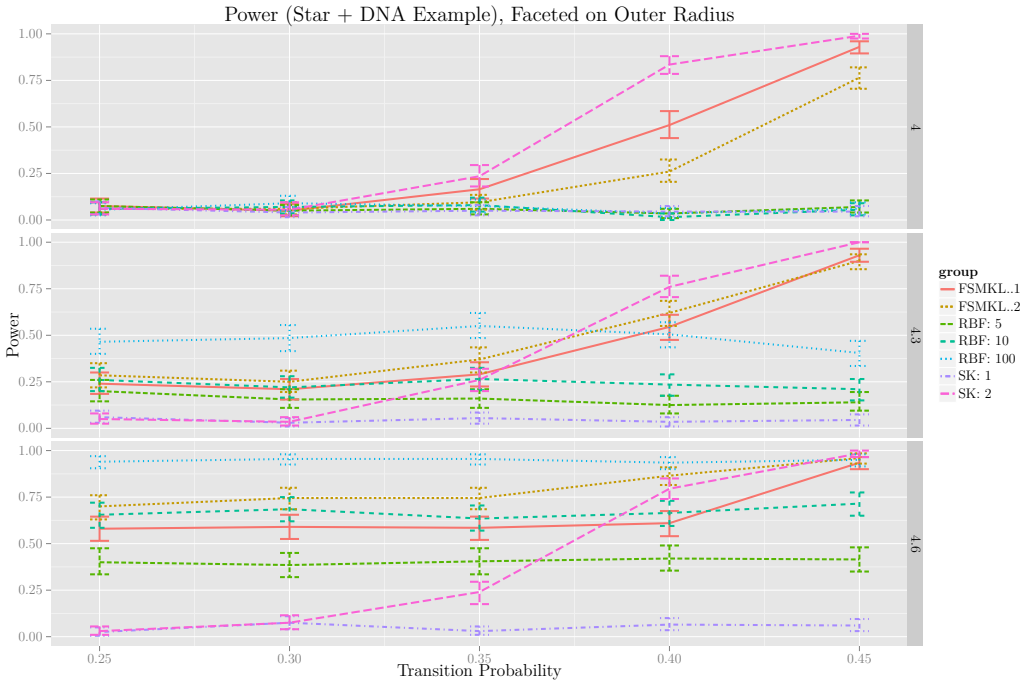


Figure 6.4: Increasing signal on the vectorial data down the rows and increasing signal on the string data across the x-axis.

10, and 100, the 1- and 2-spectrum kernels, and 1- and 2-norm MKL taking convex combinations of these five kernels.

We see that for a radius of 4 and transition probability of .25, $p = q$, and the power is equal to the level of the test, $\alpha = .05$. The MKL-based tests consistently perform second best, just behind the top-performing kernel in each setting. It appears that there is a minor penalty in performance to be paid for selecting the kernel weights versus a priori placing all weight on the best kernel for the job.

## 6.7 Null Distribution

Permutation-based tests exact an onerous computational burden, requiring computation proportional to the number of permutations in order to conduct meaningful

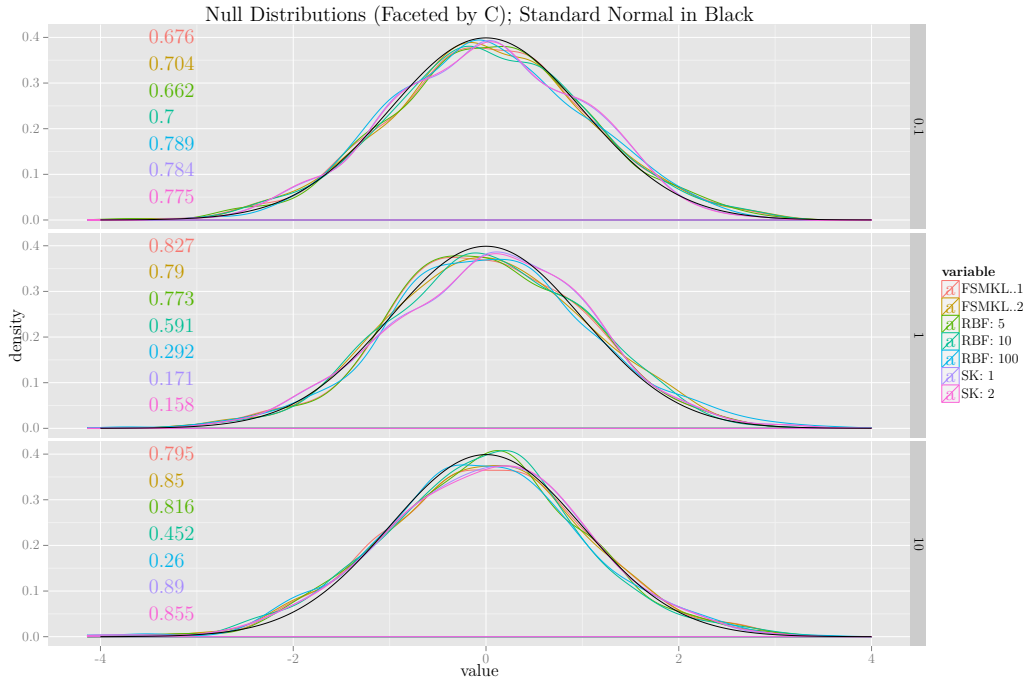


Figure 6.5: Permutation null samples are consistent with the standard normal distribution.

statistical inference. Thus, distributional approximations to these discrete, permutation null distributions are of great interest.

Here we compare the null distributions over 2000 permutations of the labels in each scenario, adjusting the regularization parameter $C \in \{.1, 1, 10\}$. We report $p$-values from the Anderson–Darling test for normality in Figure 6.5 along with density plots. In Figure 6.6, we display the same data in normal $Q$–$Q$ plots. Except for the situation with the highest emphasis on the loss function ($C = 10$), the permutation null samples are consistent with the standard normal distribution for all kernels and MKL statistics.

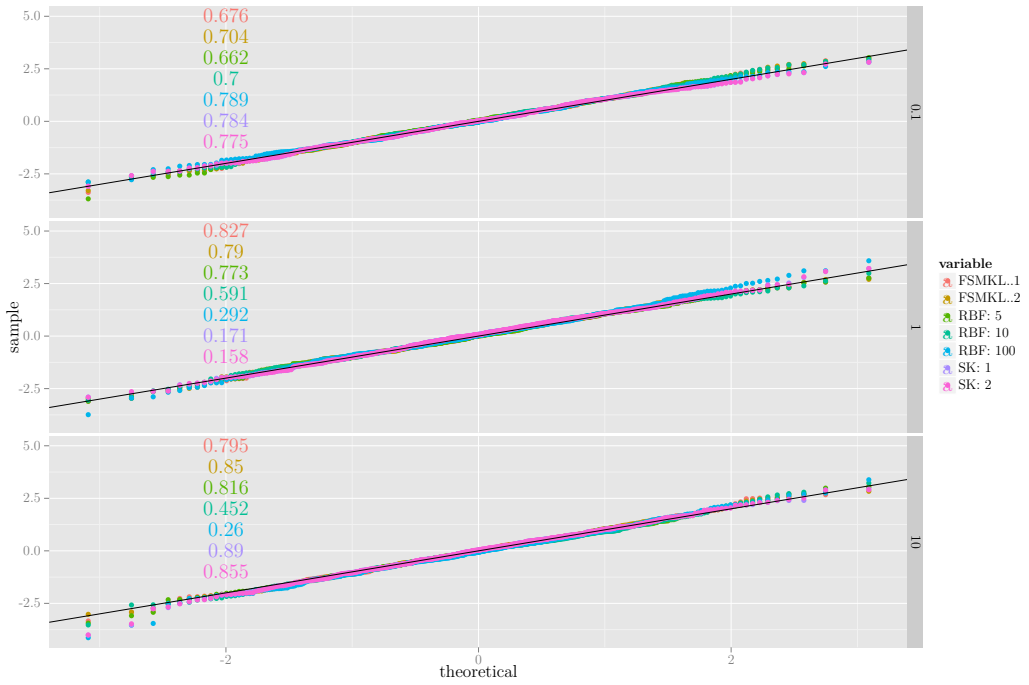


Figure 6.6: Permutation null samples are consistent with the standard normal distribution.

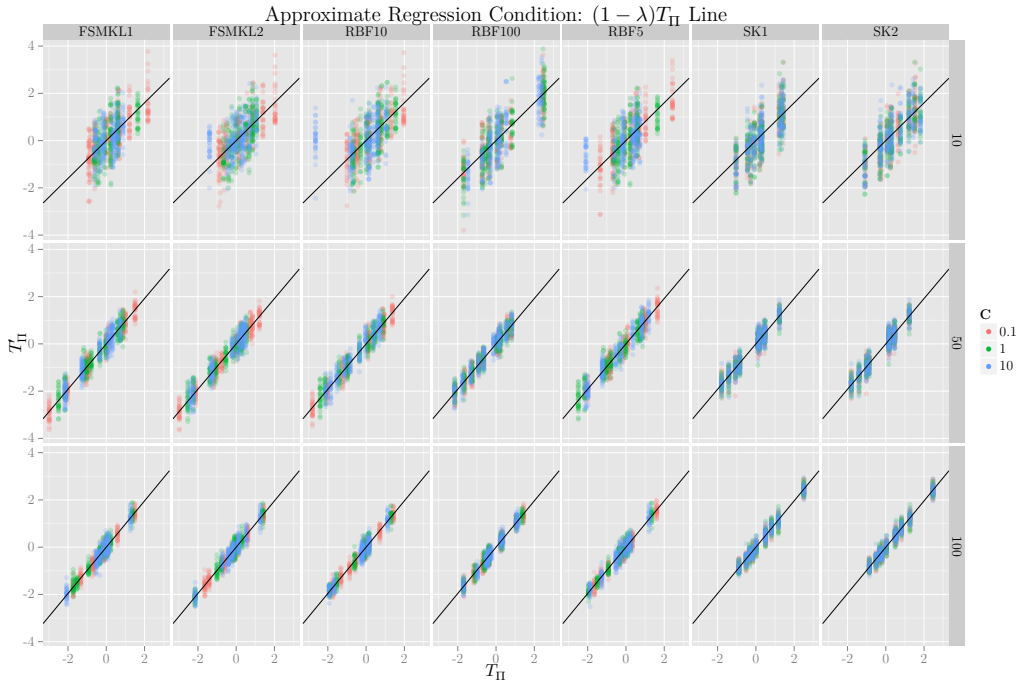


Figure 6.7: Different choices of kernels in the columns, and different sample sizes in the rows.

## 6.8 Approximate Regression Condition

Here we examine whether or not the approximate regression condition holds in the MKL setting. Taking the heterogeneous data example from before, we vary the per-group sample size $n \in \{10, 50, 100\}$. For each value of $T$, we perform 30 label swaps to compute the corresponding values $T'$ in Figure 6.7. The approximate regression condition appears to hold in all settings, and this suggests that we may be able to derive an error bound using the techniques from Chapters 2 and 3.

Given a large number of independent kernels, one worry is that our learning method may overfit to limited data. In order to explore this possibility, we generate $X \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$ data and associate to each entry $X_i$ a Gaussian RBF kernel of width 1. Because the distribution is isotropic, we have $d$ independent kernels. We estimate the randomization distribution over 200 permutations with $Q$–$Q$ plots against the standard normal quantiles and in addition report the Anderson–Darling $p$-values in

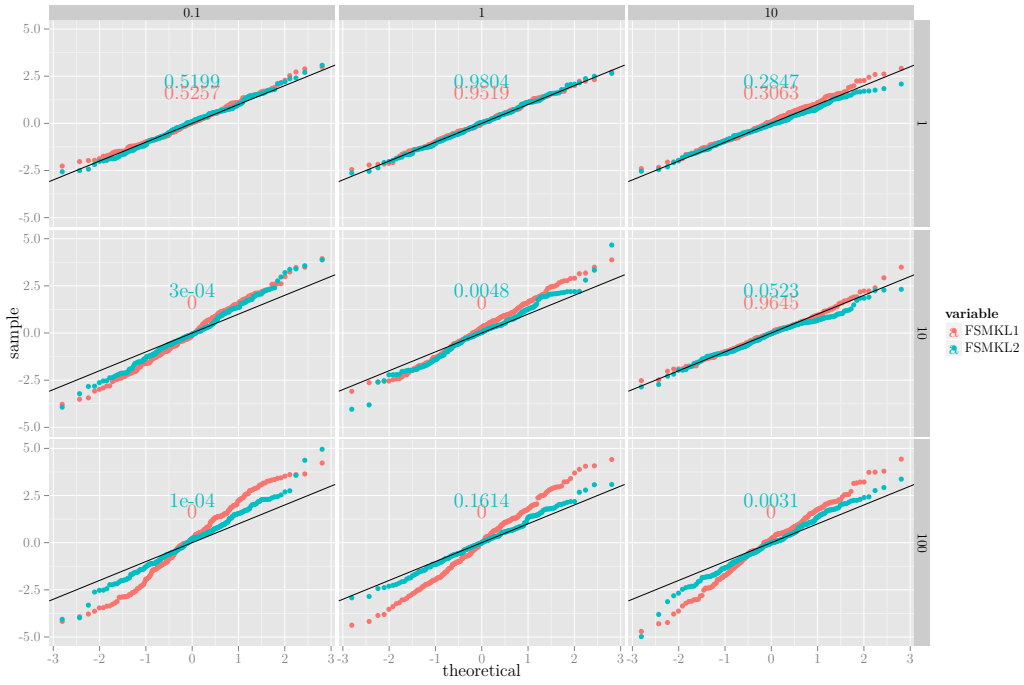


Figure 6.8: The regularization parameter $C$ in the columns, and the number of Gaussian RBF kernels in the rows.

Figure 6.8. It appears that the number of independent kernels has a large effect on the normality of the randomization distribution.

We verify that the approximate regression condition fails to hold in Figure 6.9. Given 100 independent kernels, the remainder $R$ in Equation (2.1) does not appear to be a random variable of small and decreasing order.

## 6.9 Wine Example

In July 2012, we collected 78,443 descriptions of wine from the K&L Wine Merchants' website: `http://www.klwines.com/`. Each description includes characteristics about the wine such as its name, price, varietal, critics' ratings and reviews, and winery location.

For example, here is the raw data for a 2002 Pinot Noir from Siduri Wines:

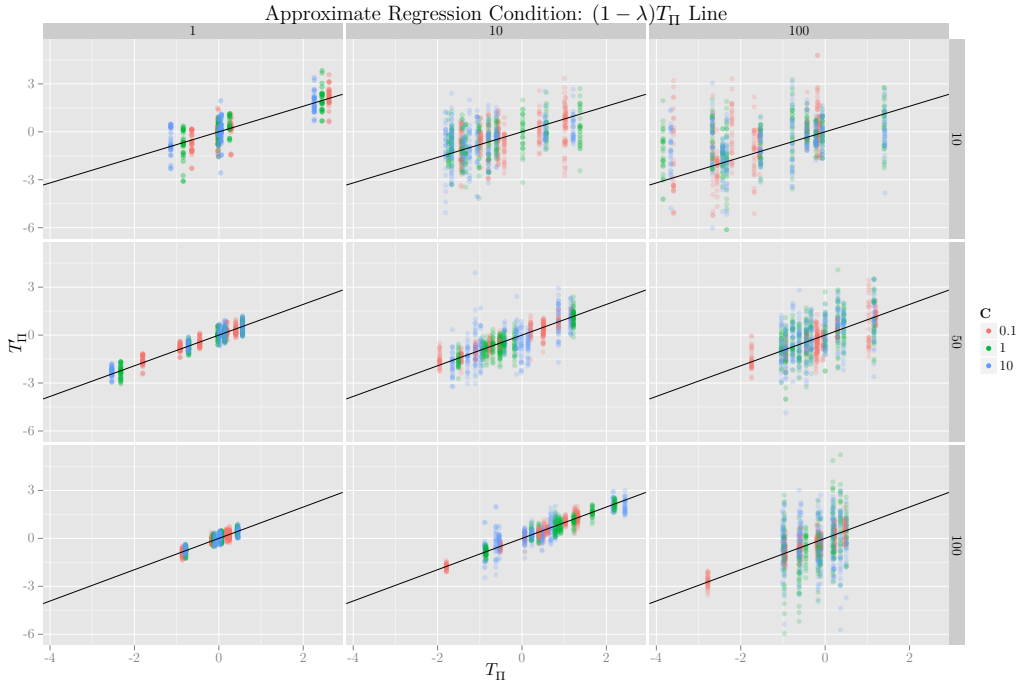


Figure 6.9: Sample size in the rows, and number of independent kernels in the columns.

```
{"sku": "1011975",
  "info": {
    "Country:": "United States",
    "Specific Appellation:": "Sonoma County",
    "Sub-Region:": "California",
    "Varietal:": "Pinot Noir"
  },
  "title": "2002 Siduri \"Hirsch Vineyard\" Sonoma Pinot Noir
  (Previously $49.99)",
  "price": "39.99",
  "ratings": {
    "ST": "90"
  },
  "desc": "90 points Stephen Tanzer's International Wine Cellar:
  \"Full medium red. Sappy aromas of cherry, plum, raspberry and
  minerals, with complicating floral and forest floor notes. Sweet, fat
  and full of fruit; very smooth flavors of black cherry, licorice,
  mocha and bitter chocolate. Nicely balanced pinot, and showing well
  today.\" (May/Jun 04)"
}
```


We'd like to know if our two-sample test can discriminate between two different wine varietals on the basis of some of these features. Moreover, we are also interested in understanding which are the most influential features. Because of the large number of samples of the two varietals, we focus on the set of Pinot Noir and Chardonnay wines.

We extract the vintage (year) of the wine into a new field and only consider letters and spaces in the wine titles and descriptions. After removing examples without a Robert Parker rating, we have 879 Chardonnay and 1134 Pinot Noir examples remaining. We use the 4-spectrum kernel for the string fields, title and description; the Gaussian RBF kernel with width 1 for the numeric fields, price, Robert Parker rating, and vintage; and the identity kernel for the subregion. Although simple, these

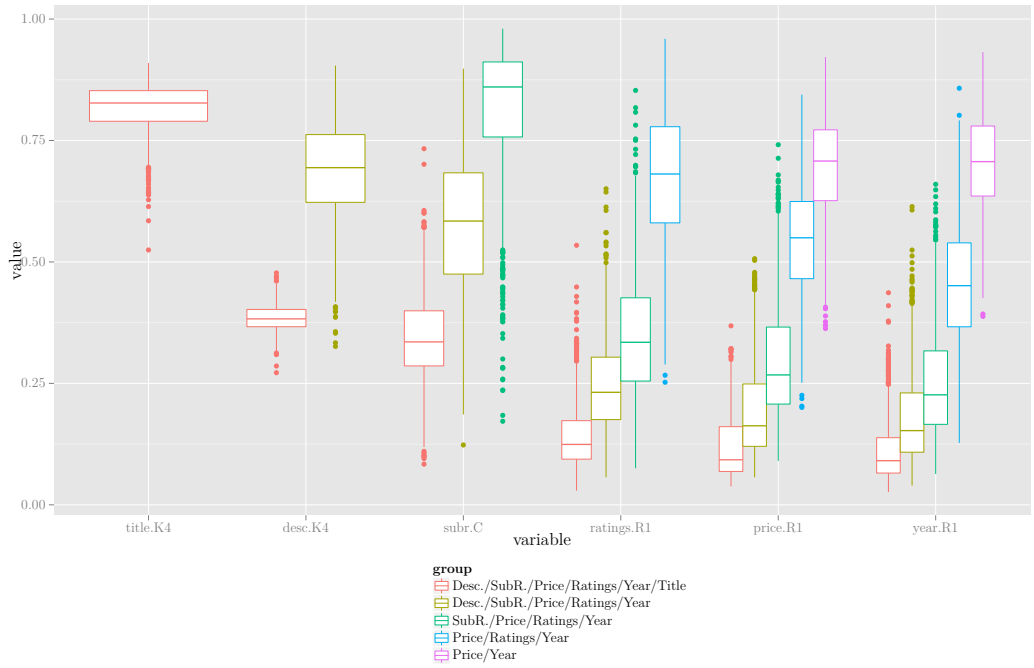


Figure 6.10: Boxplots of 2-norm MKL weights, in decreasing order of highest variable weight for the corresponding model.

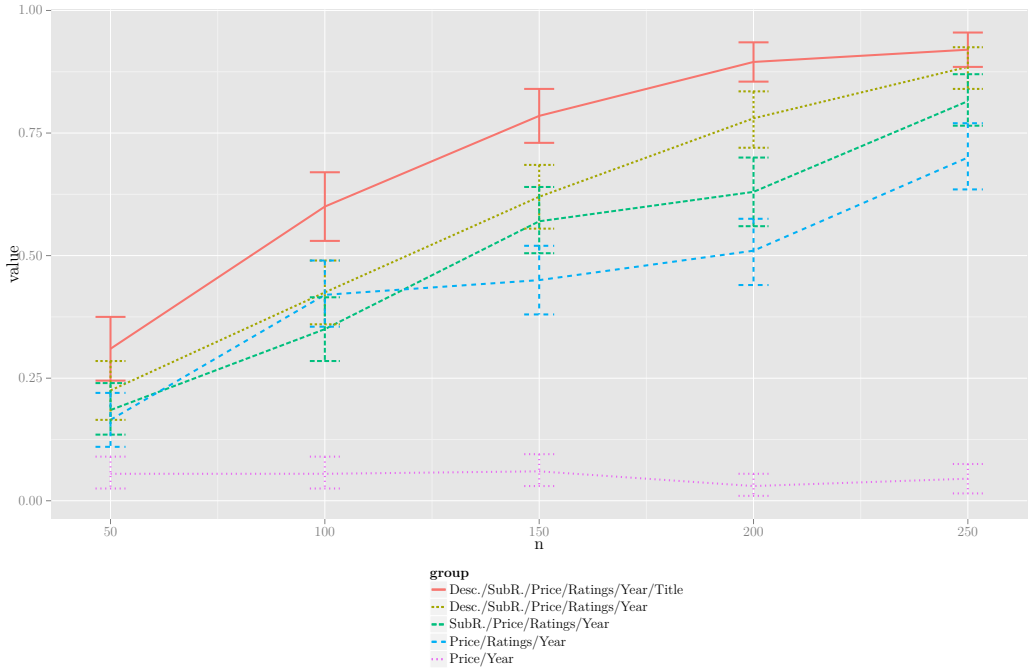


Figure 6.11: Power on sample size: MKL can make effective use of additional information.

kernel choices are also quite reasonable: the numeric fields are all roughly on the same scale, comparable with the RBF width. The identity kernel takes on a value of 1 only when the subregions agree entirely. Thus, we lose geographic information such as distances between subregions.

We train a 2-norm MKL with regularization parameter $C = .01$, initially on all 6 kernels, with varying sample sizes in $\{50, 100, 150, 200, 250\}$, repeating each test 200 times in order to estimate the distribution of the kernel weights and the statistical power of the tests. In Figure 6.10 we report boxplots of the kernel weights. For the 6-kernel model on all the features, the dominant weight lies on the 4-spectrum kernel associated with the name (title) of the wine. In Figure 6.11, we see that this test has the highest power. We then remove the kernel with the highest weight and fit a 5-kernel model on all of the features except for the name of the wine. We have ordered the boxplot x-axis in order of the highest variable weight corresponding to

the model under consideration. Thus, the title of the wine has the highest weight for the 6-kernel model, the description of the wine has the highest weight for the 5-kernel model, and so on.

The title of the wine sometimes contains information about the very thing we're trying to predict – the wine varietal. Additionally, many wineries specialize in particular varietals of wine. For example, Siduri only produces single vineyard Pinot Noirs from vineyards located in California and Oregon. Therefore, we should expect that models incorporating this information should perform very well.

Pinot Noir wines are typically lighter-colored reds that have notes of red fruits such as cherries, strawberries, and raspberries. The grape requires a cooler climate for best results and is associated with the Burgundy region of France and areas of California and Oregon. Chardonnay, by contrast, is a white wine with flavors of buttery oak and fruits such as apples and pears. Chardonnay is also commonly grown in Burgundy and California as well as Australia and New Zealand. The wine reviews use many adjectives and descriptors associated with each varietal, in addition to sometimes including the name of the varietal. By constructing views of each data type through kernels, we can pick up on such differences between the two varietals.

In Figure 6.11, by repeatedly removing the most-informative feature, we monotonically decrease the statistical power of the test. The final model including just the price and year of the wine has very poor performance. In fact, the distributions of the weights of the two variables are very similar, indicating that both features are equally poor predictors of the type of wine. For this wine example, MKL can effectively exploit additional data types for higher discriminatory power.

## 6.10 Future Work

We hope to better understand the Friedman test outside of the univariate data and linear kernel setting. That is, we would like to characterize the kernels for which the Friedman test generalizes the permutation $t$-test and understand the degree to which this relationship holds approximately given higher dimensional data and a wide range of kernels. We further hope to extend the theory of Chapters 2 and 3 to prove rates

of convergence bounds in these general settings. As seen in Figures 6.9 and 6.8, these bounds would likely incorporate some terms that represent the learning algorithm and numbers of independent kernels.

We can also potentially treat various kinds of missing data. If we consider entire missing modalities (e.g., one sample is missing some biometric reading), Poh et al. [69] developed the *neutral point substitution* technique to allow substitution of the missing modality with a new kernel that is *unbiased* with regard to the classification at hand. Panov et al. [65] modified the NPS method to allow for missing modalities in the test set. This allows for full use of both modalities that are present for all samples as well as those that are present only for a subset of the samples and effective utilization of all the data in the training set.

# Appendix A

# Auxiliary Results

The $c_r$-inequality and following corollary will provide useful bounds.

**Theorem A.1** (The $c_r$-inequality)**.** *Let $X$ and $Y$ be random variables and $r > 0$. Suppose that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^r < \infty$. Then*

$$\mathbb{E}|X+Y|^r < c_r(\mathbb{E}|X|^r + \mathbb{E}|Y|^r), \tag{A.1}$$

*where $c_r = 1$ when $r \leq 1$ and $c_r = 2^{r-1}$ when $r \geq 1$.*

**Corollary A.2.** *Suppose that $\mathrm{Var}(X) < \infty$ and $\mathrm{Var}(Y) < \infty$. Then*

$$\mathrm{Var}(X+Y) < 2(\mathrm{Var}(X) + \mathrm{Var}(Y)). \tag{A.2}$$

*Proof.* This follows immediately by applying Theorem A.1 to the centered random variables $X' = X - \mathbb{E}X$ and $Y' = Y - \mathbb{E}Y$. □

**Lemma A.3.** *If $(W, W')$ is an exchangeable pair, then $\mathbb{E}g(W, W') = 0$ for all anti-symmetric measurable functions such that the expected value exists.*

Here is a slight generalization of Lemma 2.7 from [14]:

**Lemma A.4.** *Let $(W, W')$ be an approximate Stein pair and $\Delta = W - W'$. Then*

$$\mathbb{E}W = \mathbb{E}R \quad \text{and} \quad \mathbb{E}\Delta^2 = 2\lambda\mathbb{E}W^2 - 2\lambda\mathbb{E}WR \quad \text{if } \mathbb{E}W^2 < \infty. \tag{A.3}$$

*Furthermore, when* $\mathbb{E}W^2 < \infty$*, for every absolutely continuous function* $f$ *satisfying* $|f(w)| \leq C(1 + |w|)$*, we have*

$$\mathbb{E}Wf(W) = \frac{1}{2\lambda}\mathbb{E}(W - W')(f(W) - f(W')) + \mathbb{E}f(W)R. \tag{A.4}$$

*Proof.* From (2.1) we have

$$\mathbb{E}[\mathbb{E}[W - W'|W]] = \mathbb{E}\lambda(W - R) = \lambda\mathbb{E}W - \lambda\mathbb{E}R.$$

We also have

$$\mathbb{E}[\mathbb{E}[W - W'|W]] = \mathbb{E}W - \mathbb{E}[\mathbb{E}[W'|W]] = \mathbb{E}W - \mathbb{E}W' = 0$$

using exchangeability. Equating the two expressions yields

$$\mathbb{E}W = \mathbb{E}R$$

As an intermediate computation,

$$\begin{aligned} \mathbb{E}W'W &= \mathbb{E}[\mathbb{E}[W'W|W]] \\ &= \mathbb{E}[W\mathbb{E}[W'|W]] \\ &= \mathbb{E}[W((1-\lambda)W + \lambda R)] \quad \text{from (2.1)} \\ &= (1-\lambda)\mathbb{E}W^2 + \lambda\mathbb{E}WR. \end{aligned} \tag{A.5}$$

Then

$$\begin{aligned} \mathbb{E}\Delta^2 &= \mathbb{E}(W - W')^2 \\ &= \mathbb{E}W^2 + \mathbb{E}W'^2 - 2\mathbb{E}W'W \\ &= 2\mathbb{E}W^2 - 2((1-\lambda)\mathbb{E}W^2 + \lambda\mathbb{E}WR) \quad \text{from (A.5)} \\ &= 2\lambda\mathbb{E}W^2 - 2\lambda\mathbb{E}WR. \end{aligned} \tag{A.6}$$

By the linear growth assumption on $f$, $\mathbb{E}g(W, W')$ exists for the antisymmetric

function $g(x,y) = (x-y)(f(y)+f(x))$. By Lemma A.3,

$$\begin{aligned}
0 &= \mathbb{E}(W-W')(f(W')+f(W)) \\
&= \mathbb{E}(W-W')(f(W')-f(W)) + 2\mathbb{E}f(W)(W-W') \\
&= \mathbb{E}(W-W')(f(W')-f(W)) + 2\mathbb{E}[f(W)\mathbb{E}[(W-W')|W]] \\
&= \mathbb{E}(W-W')(f(W')-f(W)) + 2\mathbb{E}f(W)(\lambda(W-R)).
\end{aligned}$$

Rearranging the expression yields

$$\mathbb{E}Wf(W) = \frac{1}{2\lambda}\mathbb{E}(W-W')(f(W)-f(W')) + \mathbb{E}f(W)R. \tag{A.7}$$

□

This is just a small part of Lemma 2.4 from [14]:

**Lemma A.5.** *For a given function $h : \mathbb{R} \to \mathbb{R}$, let $f_h$ be the solution to the Stein equation. If $h$ is absolutely continuous, then*

$$||f_h|| \leq 2||h'||. \tag{A.8}$$

Lemma 2.2 from [14]:

**Lemma A.6.** *For fixed $z \in \mathbb{R}$ and $\Phi(z) = P(Z \leq z)$, the unique bounded solution $f_z(w)$ of the equation*

$$f'(w) - wf(w) = \mathbf{1}_{\{w \leq z\}} - \Phi(z) \tag{A.9}$$

*is given by*

$$f_z(w) = \begin{cases} \sqrt{2\pi}e^{w^2/2}\Phi(w)[1-\Phi(z)] & \text{if } w \leq z \\ \sqrt{2\pi}e^{w^2/2}\Phi(z)[1-\Phi(w)] & \text{if } w > z. \end{cases} \tag{A.10}$$

Part of Lemma 2.3 from [14]:

**Lemma A.7.** *Let $z \in \mathbb{R}$ and let $f_z$ as in (A.10) Then*

$$|(w+u)f_z(w+u) - (w+v)f_z(w+v)| \leq (|w| + \sqrt{2\pi}/4)(|u|+|v|).$$

# Appendix B

# Stein's Method Proofs

## B.1 Proof of Lemma 2.6

*Proof.* Let

$$f(w) = \begin{cases} -3a/2 & w \leq w \leq z - 2a, \\ w - z + a/2 & z - 2a \leq w \leq z + a, \\ 3a/2 & w \geq z + a. \end{cases}$$

Since

$$\mathbb{E}Wf(W) \leq \mathbb{E}[|W||f(W)|] \leq \frac{3a}{2}\mathbb{E}|W| \leq \frac{3a}{2}\sqrt{\mathbb{E}W^2},$$

we have

$$\begin{aligned} 3a\lambda\sqrt{\mathbb{E}W^2} &\geq 2\lambda\mathbb{E}WF(W) \\ &= \mathbb{E}[(W - W')(f(W) - f(W'))] + 2\lambda\mathbb{E}f(W)R \quad \text{by } (A.4) \end{aligned}$$

We also bound the term involving the remainder

$$-2\lambda\mathbb{E}f(W)R \leq 2\lambda\mathbb{E}|f(W)||R| \leq 3a\lambda\mathbb{E}|R|$$

so that

$$\begin{aligned}
3a\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) &\geq \mathbb{E}(W - W')(f(W) - f(W')) \\
&= \mathbb{E}\left((W - W')\int_{W'-W}^{0} f'(W+t)dt\right) \\
&\geq \mathbb{E}\left((W - W')\int_{W'-W}^{0} \mathbf{1}_{\{|t|\leq a\}}\mathbf{1}_{\{z-a\leq W\leq z\}} f'(W+t)dt\right).
\end{aligned}$$

Since $f'(W+t) = \mathbf{1}_{\{z-2a\leq W+t\leq z+a\}}$,

$$\mathbf{1}_{\{|t|\leq a\}}\mathbf{1}_{\{z-a\leq W\leq z\}} f'(W+t) = \mathbf{1}_{\{|t|\leq a\}}\mathbf{1}_{\{z-a\leq W\leq z\}}.$$

Therefore,

$$\begin{aligned}
3a\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) &\geq \mathbb{E}\left((W - W')\int_{W'-W}^{0} \mathbf{1}_{\{|t|\leq a\}} dt \mathbf{1}_{\{z-a\leq W\leq z\}}\right) \\
&= \mathbb{E}(|W - W'|\min(a, |W - W'|)\mathbf{1}_{\{z-a\leq W\leq z\}}) \\
&\geq \mathbb{E}((W - W')^2\mathbf{1}_{\{0\leq W-W'\leq a\}}\mathbf{1}_{\{z-a\leq W\leq z\}}) \\
&= \mathbb{E}((W - W')^2\mathbf{1}_{\{-a\leq W'-W\leq 0\}}\mathbf{1}_{\{z-a\leq W\leq z\}}).
\end{aligned}$$

The proof of the second claim proceeds similarly:

$$\begin{aligned}
3a\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) &\geq \mathbb{E}(W - W')(f(W) - f(W')) \\
&= \mathbb{E}(W' - W)(f(W') - f(W)) \\
&= \mathbb{E}\left((W' - W)\int_0^{W'-W} f'(W+t)dt\right) \\
&\geq \mathbb{E}\left((W' - W)\int_0^{W'-W} \mathbf{1}_{\{|t|\leq a\}}\mathbf{1}_{\{z-a\leq W\leq z\}} f'(W+t)dt\right) \\
&= \mathbb{E}\left((W' - W)\int_0^{W'-W} \mathbf{1}_{\{|t|\leq a\}}dt\mathbf{1}_{\{z-a\leq W\leq z\}}\right) \\
&= \mathbb{E}(|W' - W|\min(a, |W' - W|)\mathbf{1}_{\{z-a\leq W\leq z\}}) \\
&\geq \mathbb{E}((W' - W)^2\mathbf{1}_{\{0\leq W-W'\leq a\}}\mathbf{1}_{\{z-a\leq W\leq z\}}).
\end{aligned}$$

□

## B.2 Proof of Theorem 2.7

*Proof.* For $z \in \mathbb{R}$ and $\alpha > 0$ let $f$ be the solution to the Stein equation

$$f'(w) - wf(w) = h_{z,\alpha}(w) - \Phi(z) \tag{B.1}$$

for the smoothed indicator

$$h_{z,\alpha}(w) = \begin{cases} 1 & w \leq z \\ 1 + \frac{z-w}{\alpha} & z < w \leq z + \alpha \\ 0 & w > z + \alpha. \end{cases} \tag{B.2}$$

Therefore,

$$\begin{aligned}
|P(W \le z) - \Phi(z)| &= |\mathbb{E}[(f'(W) - Wf(W))]| \\
&= \left|\mathbb{E}\left[f'(W) - \frac{(W'-W)(f(W')-f(W))}{2\lambda} + f(W)R\right]\right| \\
&= \left|\mathbb{E}\left[f'(W)\left(1 - \frac{(W'-W)^2}{2\lambda}\right)\right.\right. \\
&\quad \left.\left. + \frac{f'(W)(W'-W)^2 - (f(W')-f(W))(W'-W)}{2\lambda} + f(W)R\right]\right| \\
&:= |\mathbb{E}[J_1 + J_2 + J_3]| \\
&\le |\mathbb{E}J_1| + |\mathbb{E}J_2| + |\mathbb{E}J_3|.
\end{aligned} \tag{B.3}$$

From [15], we know that for all $w \in \mathbb{R}, 0 \le f(w) \le 1$ and $|f'(w)| \le 1$. Then

$$|\mathbb{E}J_3| \le \mathbb{E}|J_3| = \mathbb{E}|f(W)R| \le \mathbb{E}|R| \tag{B.4}$$

and

$$\begin{aligned}
|\mathbb{E}J_1| &= \left|\mathbb{E}\left[f'(W)\left(1 - \frac{(W'-W)^2}{2\lambda}\right)\right]\right| \\
&\le \mathbb{E}\left[\left|f'(W)\left(1 - \frac{(W'-W)^2}{2\lambda}\right)\right|\right] \\
&\le \mathbb{E}\left[\left|\left(1 - \frac{(W'-W)^2}{2\lambda}\right)\right|\right] \\
&= \frac{1}{2\lambda}\mathbb{E}[|2\lambda - \mathbb{E}[(W'-W)^2|W]|] \\
&= \frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}W^2 - \mathbb{E}WR) - \mathbb{E}[(W'-W)^2|W] + 2\lambda(1 - \mathbb{E}W^2 + \mathbb{E}WR)|] \\
&\le \frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}W^2 - \mathbb{E}WR) - \mathbb{E}[(W'-W)^2|W]|] + \mathbb{E}|1 - \mathbb{E}W^2 + \mathbb{E}WR|.
\end{aligned} \tag{B.5}$$

Note that

$$\mathbb{E}[\mathbb{E}[(W'-W)^2|W]] = \mathbb{E}\Delta^2 = 2\lambda(\mathbb{E}W^2 - \mathbb{E}WR), \tag{B.6}$$

so

$$\frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}W^2 - \mathbb{E}WR) - \mathbb{E}[(W' - W)^2|W]|] \leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])}. \quad \text{(B.7)}$$

Combining with (B.5),

$$\begin{aligned} |\mathbb{E}J_1| &\leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} + \mathbb{E}|1 - \mathbb{E}W^2 + \mathbb{E}WR| \\ &\leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} + \mathbb{E}|1 - \mathbb{E}W^2| + \mathbb{E}|WR| \end{aligned} \quad \text{(B.8)}$$

Lastly, we bound the second term,

$$\begin{aligned} J_2 &= \frac{1}{2\lambda}(W' - W)\int_W^{W'} (f'(W) - f'(t))dt \\ &= \frac{1}{2\lambda}(W' - W)\int_W^{W'}\int_t^W f''(u)dudt \\ &= \frac{1}{2\lambda}(W' - W)\int_W^{W'} (W' - u)f''(u)du. \end{aligned} \quad \text{(B.9)}$$

To show the final equality, consider separately the cases $W \leq W'$ and $W' \leq W$. For the former,

$$\begin{aligned} -\frac{1}{2\lambda}(W' - W)\int_W^{W'}\int_W^t f''(u)dudt &= -\frac{1}{2\lambda}(W' - W)\int_W^{W'}\int_u^{W'} f''(u)dtdu \\ &= -\frac{1}{2\lambda}(W' - W)\int_W^{W'} (W' - u)f''(u)du. \end{aligned}$$

For the latter,

$$\begin{aligned} \frac{1}{2\lambda}(W' - W)\int_W^{W'}\int_t^W f''(u)dudt &= -\frac{1}{2\lambda}(W' - W)\int_{W'}^{W}\int_t^W f''(u)dudt \\ &= -\frac{1}{2\lambda}(W' - W)\int_{W'}^{W}\int_{W'}^u f''(u)dtdu \\ &= -\frac{1}{2\lambda}(W' - W)\int_{W'}^{W} (u - W')f''(u)du. \end{aligned}$$

Since $W$ and $W'$ are exchangeable,

$$\begin{aligned}
|\mathbb{E}J_2| &= \left|\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)\int_W^{W'}(W'-u)f''(u)du\right]\right| \\
&= \left|\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)\int_W^{W'}\left(\frac{W+W'}{2}-u\right)f''(u)du\right]\right| \\
&\leq \left|\mathbb{E}\left[||f''||\frac{1}{2\lambda}|W'-W|\int_{\min(W,W')}^{\max(W,W')}\left|\frac{W+W'}{2}-u\right|du\right]\right| \\
&= \left|\mathbb{E}\left[||f''||\frac{1}{2\lambda}\frac{|W'-W|^3}{4}\right]\right| \\
&\leq \frac{\mathbb{E}|W'-W|^3}{4\alpha\lambda},
\end{aligned} \tag{B.10}$$

where the final inequality follows from the fact that $|h'_{z,\alpha}(x)| \leq 1/\alpha$ for all $x \in \mathbb{R}$ and Lemma A.5.

Collecting the bounds, we obtain

$$\begin{aligned}
P(W \leq z) &\leq \mathbb{E}h_{z,\alpha}(W) \\
&\leq Nh_{z,\alpha} + \frac{\mathbb{E}|W'-W|^3}{4\alpha\lambda} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} \\
&\quad + |1-\mathbb{E}W^2| + \mathbb{E}|WR| + \mathbb{E}|R| \\
&\leq \Phi(z) + \frac{\alpha}{\sqrt{2\pi}} + \frac{\mathbb{E}|W'-W|^3}{4\alpha\lambda} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} \\
&\quad + |\mathbb{E}W^2-1| + \mathbb{E}|WR| + \mathbb{E}|R|.
\end{aligned} \tag{B.11}$$

The minimizer of the expression is

$$\alpha = \frac{(2\pi)^{1/4}}{2}\sqrt{\frac{\mathbb{E}|W'-W|^3}{\lambda}}. \tag{B.12}$$

Plugging this in, we get the upper bound

$$P(W \leq z) - \Phi(z) \leq (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|W' - W|^3}{\lambda}} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} + |\mathbb{E}W^2 - 1| + \mathbb{E}|WR| + \mathbb{E}|R| \tag{B.13}$$

The corresponding lower bound can be derived in a similar fashion. □

## B.3 Proof of Theorem 2.8

*Proof.* Now we bound $|\mathbb{E}J_2|$ with $\delta \geq 0$. From (B.3),

$$\begin{aligned} 2\lambda J_2 &= f'(W)(W' - W)^2 - (f(W') - f(W))(W' - W) \\ &= (W' - W)\int_0^{W'-W} (f'(W) - f'(W + t))dt \\ &= (W' - W)\mathbf{1}_{|W'-W|\leq\delta}\int_0^{W'-W} (f'(W) - f'(W + t))dt. \end{aligned}$$

Using (A.9), $f'(W) = Wf(W) + \mathbf{1}_{\{w\leq z\}} - \Phi(z)$ and $f'(W+t) = (W+t)f(W+t) + \mathbf{1}_{\{w+t\leq z\}} - \Phi(z)$. Therefore,

$$\begin{aligned} 2\lambda J_2 &= (W' - W)\mathbf{1}_{|W'-W|\leq\delta}\int_0^{W'-W} (Wf(W) - (W + t)f(W + t))dt \\ &\quad + (W' - W)\mathbf{1}_{|W'-W|\leq\delta}\int_0^{W'-W} (\mathbf{1}_{\{W\leq z\}} - \mathbf{1}_{\{W+t\leq z\}})dt \\ &\equiv J_{21} + J_{22}. \end{aligned}$$

We apply (A.7) with $w = W$, $u = 0$, and $v = t$ to get

$$\begin{aligned}
|\mathbb{E}J_{21}| &\leq \left|(W' - W)\mathbf{1}_{|W'-W|\leq\delta} \int_0^{W'-W} \left(|W| + \frac{\sqrt{2pi}}{4}\right) |t| dt\right| \\
&\leq \mathbb{E}\left[\frac{1}{2}|W' - W|^3 \mathbf{1}_{|W'-W|\leq\delta} \left(|W| + \frac{\sqrt{2pi}}{4}\right)\right] \\
&\leq \frac{1}{2}\delta^3 \left(1 + \frac{\sqrt{2\pi}}{4}\right) \\
&\leq .82\delta^3.
\end{aligned}$$

Now for $J_{22}$, we consider the two cases according to the sign of $W' - W$. When $W' - W \leq 0$, we have

$$\begin{aligned}
\mathbb{E}J_{22}\mathbf{1}_{\{\delta\leq W'-W\leq 0\}} &= \mathbb{E}\left[(W' - W)\mathbf{1}_{\{\delta\leq W'-W\leq 0\}} \int_0^{W'-W} (\mathbf{1}_{\{W\leq z\}} - \mathbf{1}_{\{W+t\leq z\}}) dt\right] \\
&= \mathbb{E}\left[(W - W')\mathbf{1}_{\{\delta\leq W'-W\leq 0\}} \int_{W'-W}^0 (\mathbf{1}_{\{z\leq W\leq z-t\}}) dt\right] \\
&\leq \mathbb{E}\left[(W - W')^2 \mathbf{1}_{\{\delta\leq W'-W\leq 0\}} \mathbf{1}_{\{z-\delta\leq W\leq z\}}\right] \\
&\leq 3\delta\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) \quad \text{by (2.6)}
\end{aligned}$$

Similarly, when $W' - W > 0$,

$$\begin{aligned}
\mathbb{E}J_{22}\mathbf{1}_{\{0<W'-W\leq\delta\}} &= \mathbb{E}\left[(W' - W)\mathbf{1}_{\{0<W'-W\leq\delta\}} \int_0^{W'-W} (\mathbf{1}_{\{W\leq z\}} - \mathbf{1}_{\{W+t\leq z\}}) dt\right] \\
&= \mathbb{E}\left[(W' - W)\mathbf{1}_{\{0<W'-W\leq\delta\}} \int_0^{W'-W} \mathbf{1}_{\{z-t<W\leq z\}} dt\right] \\
&\leq \mathbb{E}\left[(W' - W)^2 \mathbf{1}_{\{0<W'-W\leq\delta\}} \mathbf{1}_{\{z-\delta\leq W\leq z\}}\right] \\
&\leq 3\delta\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) \quad \text{by (2.6)}
\end{aligned}$$

Therefore,

$$\begin{aligned} |\mathbb{E}J_2| &\leq \frac{1}{2\lambda}(|\mathbb{E}J_{21}| + |\mathbb{E}J_{22}|) \\ &\leq \frac{.41\delta^3}{\lambda} + 3\delta(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|). \end{aligned}$$

The result follows from (B.3), noting that $J_1$ and $J_3$ stay the same. □

# Appendix C

# Rate of Convergence Bounds

## C.1 Proof of Proposition 3.4

*Proof.*

$$\begin{aligned}
\mathbb{E}T_{\Pi}^2 &= \frac{n-1}{n}\mathbb{E}\left[\left(\frac{q_{\Pi}}{d_{\Pi}}\right)^2\right] && \text{(C.1)}\\
&= \frac{n-1}{n}\mathbb{E}\left[\frac{4n^2\bar{x}_{2,\Pi}^2}{2n-2n\bar{x}_{2,\Pi}^2}\right] \quad \text{from (3.8)}\\
&= 2(n-1)\mathbb{E}\left[\frac{\bar{x}_{2,\Pi}^2}{1-\bar{x}_{2,\Pi}^2}\right]\\
&= 2(n-1)\mathbb{E}g(\bar{x}_{2,\Pi}), && \text{(C.2)}
\end{aligned}$$

where $g(x) = \frac{x^2}{1-x^2}$. Now we proceed to calculate moments of $\bar{x}_{2,\Pi}$.

Mean-centering the $x_i$ has the effect of mean-centering $\bar{x}_{2,\Pi}$:

$$\mathbb{E}\bar{x}_{2,\Pi} = \frac{1}{n}\mathbb{E}\left[\sum_{i=n+1}^{2n} x_{\Pi(i)}\right] = \frac{1}{n}\sum_{i=n+1}^{2n}\mathbb{E}x_{\Pi(i)} = \frac{1}{n}\sum_{i=n+1}^{2n}\frac{1}{2n}\sum_{j=1}^{2n}x_j = 0$$

Under independence, $\mathrm{Var}(\bar{x}_{2,\Pi})$ would be $\frac{1}{n}$ given the scaling. However, the negative dependence induced by the permutation structure approximately halves this value.

The scaling is such that $\mathrm{Var}(x_{\Pi(i)}) = 1$. Under independence and with $i \neq j$, $\mathrm{Var}(x_{\Pi(i)} + x_{\Pi(j)}) = 2$. Summing only 2 (out of $2n$) values under permutation dependence, $\mathrm{Var}(x_{\Pi(i)} + x_{\Pi(j)}) = 2 - \frac{2}{2n-1}$.

We can't use Serfling's result here because we need more than just an upper bound.

$$
\begin{aligned}
\mathrm{Var}(\bar{x}_{2,\Pi}) &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=n+1}^{2n} x_{\Pi(i)}\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=n+1}^{2n} x_{\Pi(i)}^2 + \sum_{i=n+1}^{2n}\sum_{j=n+1, j\neq i}^{2n} x_{\Pi(i)}x_{\Pi(j)}\right] \\
&= \frac{1}{n^2}\sum_{i=n+1}^{2n}\frac{1}{2n}\sum_{j=1}^{2n} x_j^2 + \frac{1}{n^2}\sum_{i=n+1}^{2n}\sum_{j=n+1, j\neq i}^{2n}\mathbb{E}[x_{\Pi(i)}x_{\Pi(j)}] \\
&= \frac{1}{n} + \frac{1}{n^2}\sum_{i=n+1}^{2n}\sum_{j=n+1, j\neq i}^{2n}\frac{1}{2n}\frac{1}{2n-1}\sum_{k=1}^{2n}\sum_{l=1, l\neq k}^{2n} x_k x_l \\
&= \frac{1}{n} + \frac{1}{n^2}\sum_{i=n+1}^{2n}\sum_{j=n+1, j\neq i}^{2n}\frac{1}{2n}\frac{1}{2n-1}\left(\left(\sum_{k=1}^{2n} x_k\right)^2 - \sum_{k=1}^{2n} x_k^2\right) \\
&= \frac{1}{n} + \frac{1}{n^2}\sum_{i=n+1}^{2n}\sum_{j=n+1, j\neq i}^{2n}\frac{1}{2n}\frac{1}{2n-1}\left(0^2 - 2n\right) \\
&= \frac{1}{n} + \frac{1}{n}(n^2 - n)\left(-\frac{1}{2n-1}\right) \\
&= \frac{2n-1}{n(2n-1)} + \frac{1-n}{n(2n-1)} \\
&= \frac{1}{2n-1}
\end{aligned}
$$

Having established the first two moments, we compute the third degree Taylor expansion and bound the error in the approximation. By Taylor's theorem, we expand the function $g(\bar{x}_{2,\Pi}) = \frac{\bar{x}_{2,\Pi}^2}{1-\bar{x}_{2,\Pi}^2}$ around $\mathbb{E}[\bar{x}_{2,\Pi}] = 0$:

$$
g(\bar{x}_{2,\Pi}) = \frac{\bar{x}_{2,\Pi}^2}{1-\bar{x}_{2,\Pi}^2} = g(0) + g'(0)\bar{x}_{2,\Pi} + \frac{g''(0)}{2!}\bar{x}_{2,\Pi}^2 + \frac{g^{(3)}(0)}{3!}\bar{x}_{2,\Pi}^3 + R_3(\bar{x}_{2,\Pi}),
$$

where $R_3(\bar{x}_{2,\Pi}) = \frac{g^{(4)}(\xi_L)}{4!}\bar{x}_{2,\Pi}^4$, with $\xi_L \in [0, \bar{x}_{2,\Pi}]$.

From (C.2) and evaluating the Taylor series, we have

$$\mathbb{E}g(\bar{x}_{2,\Pi}) = \frac{\mathbb{E}T_{\Pi}^2}{2(n-1)} = \mathbb{E}[\bar{x}_{2,\Pi}^2 + R_3(\bar{x}_{2,\Pi})].$$

Therefore,

$$\begin{aligned}
\left|\frac{\mathbb{E}T_{\Pi}^2}{2(n-1)} - \mathbb{E}\bar{x}_{2,\Pi}^2\right| &= \left|\frac{\mathbb{E}T_{\Pi}^2}{2(n-1)} - \frac{1}{2n-1}\right| \\
&\leq \mathbb{E}|R_3(\bar{x}_{2,\Pi})| \\
&= \mathbb{E}\left|\frac{24(5\xi_L^4 + 10\xi_L^2 + 1)}{4!\,(\xi_L - 1)^5}\bar{x}_{2,\Pi}^4\right| \\
&\leq \mathbb{E}\left|\frac{24(5\bar{x}_{2,\Pi}^4 + 10\bar{x}_{2,\Pi}^2 + 1)}{4!\,(\bar{x}_{2,\Pi} - 1)^5}\bar{x}_{2,\Pi}^4\right| \\
&\leq \frac{5B^4 + 10B^2 + 1}{|B-1|^5}\mathbb{E}\bar{x}_{2,\Pi}^4 \\
&\leq \frac{5B^4 + 10B^2 + 1}{|B-1|^5} f_{c_1}(4) n^{-2} \quad \text{by (3.4)} \\
&:= c_1 n^{-2}
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}T_{\Pi}^2 - 1| - \frac{1}{2n-1} &\leq \left|\mathbb{E}T_{\Pi}^2 - 1 + \frac{1}{2n-1}\right| \\
&= \left|\mathbb{E}T_{\Pi}^2 - \frac{2(n-1)}{2n-1}\right| \\
&= 2(n-1)\left|\frac{\mathbb{E}T_{\Pi}^2}{2(n-1)} - \frac{1}{2n-1}\right| \\
&\leq c_1 2(n-1)n^{-2}
\end{aligned}$$

This implies that

$$|\mathbb{E}T_{\Pi}^2 - 1| \leq \frac{1}{2n-1} + c_1\frac{2n-2}{n^2} \leq \frac{1+2c_1}{n} := c_2 n^{-1}$$

□

## C.2 Proof of Proposition 3.3

*Proof.* With two applications of the $c_r$ inequality, we can bound the variance of the sum by a constant times the sum of the variances. Suppose $X$, $Y$, and $Z$ have finite variances. Then, with the centered random variables represented by $\tilde{X}, \tilde{Y}$, and $\tilde{Z}$, we have that

$$\begin{aligned}
\mathrm{Var}(X+Y+Z) &= \mathrm{Var}(\tilde{X}+\tilde{Y}+\tilde{Z}) \\
&= \mathbb{E}|(\tilde{X}+\tilde{Y})+\tilde{Z}|^2 \\
&\leq 2\mathbb{E}|\tilde{X}+\tilde{Y}|^2 + 2\mathbb{E}|\tilde{Z}|^2 \\
&\leq 2(2\mathbb{E}\tilde{X}^2 + 2\mathbb{E}\tilde{Y}^2) + 2\mathbb{E}\tilde{Z}^2 \\
&\leq 4(\mathrm{Var}(X)+\mathrm{Var}(Y)+\mathrm{Var}(Z))
\end{aligned}$$

From (3.11),

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|\Pi=\pi]) &= \mathrm{Var}\left(\frac{n-1}{n}\mathbb{E}\left[\left(\frac{2x_{\Pi(J)}-2x_{\Pi(I)}}{d_\Pi} + T'_\Pi\frac{d_\Pi - d'_\Pi}{d_\Pi}\right)^2\middle|\Pi=\pi\right]\right) \\
&\leq \mathrm{Var}\left(\mathbb{E}\left[\left(\frac{2x_{\Pi(J)}-2x_{\Pi(I)}}{d_\Pi} + T'_\Pi\frac{d_\Pi - d'_\Pi}{d_\Pi}\right)^2\middle|\Pi=\pi\right]\right) \\
&\leq 4(\mathrm{Var}(X)+\mathrm{Var}(Y)+\mathrm{Var}(Z))
\end{aligned}$$

where

$$\begin{aligned}
X &= \mathbb{E}\left[\left(\frac{2x_{\Pi(J)}-2x_{\Pi(I)}}{d_\Pi}\right)^2\middle|\Pi=\pi\right] \\
Y &= \mathbb{E}\left[\left(T'_\Pi\frac{d_\Pi - d'_\Pi}{d_\Pi}\right)^2\middle|\Pi=\pi\right] \\
Z &= 2\mathbb{E}\left[\left(\frac{2x_{\Pi(J)}-2x_{\Pi(I)}}{d_\Pi}T'_\Pi\frac{d_\Pi - d'_\Pi}{d_\Pi}\right)\middle|\Pi=\pi\right]
\end{aligned}$$

The $X$ term will dominate, so we can afford to use coarser methods on $Y$ and $Z$.

The $\mathbb{E}[x_{\Pi(J)} - x_{\Pi(I)}|\Pi=\pi]$ term is common to applications of Stein's method of

exchangeable pairs. However, there is a complication in the $d_\Pi$ random variable in the denominator. Our strategy will be to calculate the two variances separately with some necessary additional terms.

First, we prove an intermediate result regarding the variance of a product of random variables

$$W = (d_\Pi)^{-2} \text{ and } V = \mathbb{E}[(x_{\Pi(J)} - x_{\Pi(I)})^2 | \Pi = \pi].$$

Then $\mathrm{Var}(X) = 4\,\mathrm{Var}(WV)$ since $d_\Pi$ is $\sigma(\Pi)$-measurable and

$$\begin{aligned}
\mathrm{Var}(WV) &= \mathrm{Var}(W(V - \mathbb{E}V) + W\mathbb{E}V) \\
&\leq 2\,\mathrm{Var}(W(V - \mathbb{E}V)) + 2\,\mathrm{Var}(W\mathbb{E}V) \\
&\leq 2\mathbb{E}[W^2(V - \mathbb{E}V)^2] + 2(\mathbb{E}V)^2\,\mathrm{Var}(W) \\
&\leq 2(f_{c_2}(2))^2 n^{-2}\,\mathrm{Var}(V) + 2x_\Delta^4\,\mathrm{Var}(W). \qquad \text{(C.3)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(W) &= \mathrm{Var}((d_\Pi)^{-2}) \\
&= \mathrm{Var}\left(\frac{1}{2n(1 - \bar{x}_{2,\Pi}^2)}\right) \\
&= \frac{1}{4n^2}\left[\mathbb{E}\left[\left(\frac{1}{1 - \bar{x}_{2,\Pi}^2}\right)^2\right] - \left(\mathbb{E}\left[\frac{1}{1 - \bar{x}_{2,\Pi}^2}\right]\right)^2\right] \\
&= \frac{1}{4n^2}[\mathbb{E}h(\bar{x}_{2,\Pi}) - (\mathbb{E}\tilde{h}(\bar{x}_{2,\Pi}))^2],
\end{aligned}$$

where

$$h(x) = \left(\frac{1}{1 - x^2}\right)^2 = 1 + 2x^2 + 3x^4 + \ldots \text{ and } \tilde{h}(x) = \frac{1}{1 - x^2} = 1 + x^2 + x^4 + \ldots$$

By Taylor's theorem,

$$\mathbb{E}\left[\left(\frac{1}{1-\bar{x}_{2,\Pi}^2}\right)^2\right] = 1 + 2\left(\frac{1}{2n-1}\right) + \mathbb{E}R_3(\bar{x}_{2,\Pi}),$$

with

$$\mathbb{E}|R_3(\bar{x}_{2,\Pi})| \leq \frac{24(35B^4+42B^2+3)}{4!\,(B-1)^6} f_{c_1}(4) n^{-2} := c_4 n^{-2}$$

Re-arranging, we get

$$\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{x}_{2,\Pi}^2}\right)^2\right] - 1 - \frac{2}{2n-1}\right| \leq c_4 n^{-2}.$$

Applying Taylor's theorem to $\tilde{h}$:

$$\mathbb{E}\left[\frac{1}{1-\bar{x}_{2,\Pi}^2}\right] = 1 + \frac{1}{2n-1} + \mathbb{E}\tilde{R}_3(\bar{x}_{2,\Pi}),$$

with

$$\mathbb{E}|\tilde{R}_3(\bar{x}_{2,\Pi})| \leq \frac{24(5B^4+10B^2+1)}{4!\,(B-1)^5} f_{c_1}(4) n^{-2} := c_5 n^{-2}$$

Squaring, applying the bound, and re-arranging yields

$$\left|\left(\mathbb{E}\left[\frac{1}{1-\bar{x}_{2,\Pi}^2}\right]\right)^2 - \left(1+\frac{1}{2n-1}\right)^2\right| \leq 2\left(1+\frac{1}{2n-1}\right) c_5 n^{-2} + c_5^2 n^{-4}$$

now we combine bounds to get

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{x}_{2,\Pi}^2}\right)^2\right]-\left(\mathbb{E}\left[\frac{1}{1-\bar{x}_{2,\Pi}^2}\right]\right)^2\right| \\
&=\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{x}_{2,\Pi}^2}\right)^2\right]-\left(\mathbb{E}\left[\frac{1}{1-\bar{x}_{2,\Pi}^2}\right]\right)^2+\frac{1}{(2n-1)^2}-\frac{1}{(2n-1)^2}\right| \\
&\leq\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{x}_{2,\Pi}^2}\right)^2\right]-\left(\mathbb{E}\left[\frac{1}{1-\bar{x}_{2,\Pi}^2}\right]\right)^2+\frac{1}{(2n-1)^2}\right|+\left|\frac{1}{(2n-1)^2}\right| \\
&\leq\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{x}_{2,\Pi}^2}\right)^2\right]-1-\frac{2}{2n-1}-\left(\left(\mathbb{E}\left[\frac{1}{1-\bar{x}_{2,\Pi}^2}\right]\right)^2-\left(1+\frac{1}{2n-1}\right)^2\right)\right|+\left|\frac{1}{(2n-1)^2}\right| \\
&\leq c_4 n^{-2}+2\left(1+\frac{1}{2n-1}\right)c_5 n^{-2}+c_5^2 n^{-4}+\left|\frac{1}{(2n-1)^2}\right| \\
&\leq (c_4+3c_5+c_5^2+\frac{1}{4})n^{-2} \\
&:= c_6 n^{-2}
\end{aligned}
$$

Therefore, $\mathrm{Var}(W) \leq \frac{c_6}{4}n^{-4}$ and

$$\mathrm{Var}(X) \leq 8(f_{c_2}(2))^2 n^{-2}\,\mathrm{Var}(V) + 8x_\Delta^4 \frac{c_6}{4}n^{-4}$$

with

$$
\begin{aligned}
\mathrm{Var}(V) &= \mathrm{Var}(\mathbb{E}[(x_{\Pi(J)}-x_{\Pi(I)})^2|\Pi=\pi]) \\
&= \mathrm{Var}(\mathbb{E}[x_{\Pi(J)}^2+x_{\Pi(I)}^2-2x_{\Pi(J)}x_{\Pi(I)}|\Pi=\pi]) \\
&= \mathrm{Var}\left(\frac{1}{n^2}\sum_{I=1}^{n}\sum_{J=n+1}^{2n}(x_{\pi(J)}^2+x_{\pi(I)}^2-2x_{\pi(J)}x_{\pi(I)})\right) \\
&= \mathrm{Var}\left(\frac{1}{n^2}\left(n\sum_{K=1}^{2n}x_K^2-\sum_{I=1}^{n}\sum_{J=n+1}^{2n}2x_{\pi(J)}x_{\pi(I)}\right)\right) \\
&= \frac{4}{n^4}\sum_{I=1}^{n}\sum_{J=n+1}^{2n}\sum_{K=1}^{n}\sum_{L=n+1}^{2n}\mathrm{Cov}(x_{\pi(I)}x_{\pi(J)},x_{\pi(K)}x_{\pi(L)})
\end{aligned}
$$

since $\sum_{K=1}^{2n} x_K^2 = 2n$ is a constant. We proceed by calculating

$$\mathrm{Cov}(x_{\pi(I)}x_{\pi(J)}, x_{\pi(K)}x_{\pi(L)}) = \mathbb{E}[x_{\pi(I)}x_{\pi(J)}x_{\pi(K)}x_{\pi(L)}] - \mathbb{E}[x_{\pi(I)}x_{\pi(J)}]\mathbb{E}[x_{\pi(K)}x_{\pi(L)}].$$

The index sets for variables $I$ and $J$ (and $K$ and $L$) are disjoint, so

$$\mathbb{E}[x_{\pi(I)}x_{\pi(J)}] = \mathbb{E}[x_{\pi(K)}x_{\pi(L)}] = \frac{1}{2n}\frac{1}{2n-1}\sum_{I=1}^{2n} x_I \sum_{J=1,J\neq I}^{2n} x_J = -\frac{1}{2n-1}$$

for all values of $I, J, K, L$ in the sum. Therefore,

$$\mathbb{E}[x_{\pi(I)}x_{\pi(J)}] = \mathbb{E}[x_{\pi(K)}x_{\pi(L)}] = \frac{1}{(2n-1)^2}.$$

However, $K$ could equal $I$ and $L$ could equal $J$, which changes the mass assigned by the permutation distribution, necessitating a separate treatment for each case.

Case $I \neq J \neq K \neq L$:

$$\begin{aligned}
&\mathbb{E}[x_{\pi(I)}x_{\pi(J)}x_{\pi(K)}x_{\pi(L)}] \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\sum_{I=1}^{2n}\sum_{J=1,J\neq I}^{2n}\sum_{K=1,K\neq I,J}^{2n}\sum_{L=1,L\neq I,J,K}^{2n} x_I x_J x_K x_L \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\sum_{I=1}^{2n} x_I \sum_{J=1,J\neq I}^{2n} x_J \sum_{K=1,K\neq I,J}^{2n} x_K(-x_I - x_J - x_K) \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\sum_{I=1}^{2n} x_I \sum_{J=1,J\neq I}^{2n} x_J((-x_I - x_J)(-x_I - x_J) + (x_I^2 + x_J^2 - 2n)) \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\sum_{I=1}^{2n} x_I \sum_{J=1,J\neq I}^{2n} x_J(2x_I^2 - 2n + 2x_J^2 + 2x_I x_J) \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\sum_{I=1}^{2n} x_I \left((2x_I^2 - 2n)(-x_I) + 2\sum_{J=1,J\neq I}^{2n} x_J^3 + 2x_I(2n - x_I^2)\right) \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\sum_{I=1}^{2n} x_I \left(-4x_I^3 + 6nx_I + 2\left(\sum_{J=1}^{2n} x_J^3 - x_I^3\right)\right) \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\frac{1}{2n-3}\left(-6\sum_{I=1}^{2n} x_I^4 + 12n^2\right)
\end{aligned}$$

for $n^2(n-1)^2$ terms in the sum.

Case $I = K$ and $J = L$:

$$\begin{aligned}
\mathbb{E}[x^2_{\pi(I)}x^2_{\pi(J)}] &= \frac{1}{2n}\frac{1}{2n-1}\sum_{I=1}^{2n}\sum_{J=1,J\neq I}^{2n} x_I^2 x_J^2 \\
&= \frac{1}{2n}\frac{1}{2n-1}\sum_{I=1}^{2n} x_I^2(2n - x_I^2) \\
&= \frac{2n}{2n-1} - \frac{1}{2n}\frac{1}{2n-1}\sum_{I=1}^{2n} x_I^4
\end{aligned}$$

for $n^2$ terms in the sum.

Case $I = K, J \neq L$ or $I \neq K, J = L$:

$$\begin{aligned}
\mathbb{E}[x^2_{\pi(I)}x_{\pi(J)}x_{\pi(K)}] &= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\sum_{I=1}^{2n}\sum_{J=1,J\neq I}^{2n}\sum_{K=1,K\neq I,J}^{2n} x_I^2 x_J x_K \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\sum_{I=1}^{2n}\sum_{J=1,J\neq I}^{2n} x_I^2 x_J(0 - x_I - x_J) \\
&= -\frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\left(\sum_{I=1}^{2n} x_I^3 \sum_{J=1,J\neq I}^{2n} x_J + \sum_{I=1}^{2n} x_I^2 \sum_{J=1,J\neq I}^{2n} x_J^2\right) \\
&= -\frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\left(\sum_{I=1}^{2n} -x_I^4 + \sum_{I=1}^{2n} x_I^2(2n - x_I^2)\right) \\
&= \frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\left(2\sum_{I=1}^{2n} x_I^4 - 4n^2\right)
\end{aligned}$$

for $2n^2(n-1)$ terms in the sum.

Putting it all together, we have

$$
\begin{aligned}
&\mathrm{Var}(\mathbb{E}[(x_{\Pi(J)} - x_{\Pi(i)})^2]|\Pi = \pi) \\
&= \frac{4}{n^4}(n^2(n-1)^2)\left(\frac{1}{(2n)(2n-1)(2n-2)(2n-3)}\left(-6\sum_{i=1}^{2n} x_i^4 + 12n^2\right) - \frac{1}{(2n-1)^2}\right) \\
&+ \frac{4}{n^4}n^2\left(\frac{2n}{2n-1} - \frac{1}{2n}\frac{1}{2n-1}\sum_{i=1}^{2n} x_i^4 - \frac{1}{(2n-1)^2}\right) \\
&+ \frac{4}{n^4}(2n^2(n-1))\left(\frac{1}{2n}\frac{1}{2n-1}\frac{1}{2n-2}\left(2\sum_{i=1}^{2n} x_i^4 - 4n^2\right) - \frac{1}{(2n-1)^2}\right) \\
&\leq \frac{48}{4n^2} + \frac{8}{n^2} + \frac{16\sum_{i=1}^{2n} x_i^4}{n^4} \\
&= \left(20 + 16\left(\sum_{i=1}^{2n} x_i^4\right)n^{-2}\right)n^{-2}
\end{aligned}
$$

Therefore,

$$\mathrm{Var}(X) \leq 8(f_{c_2}(2))^2\left(20 + 16\left(\sum_{i=1}^{2n} x_i^4\right)n^{-2}\right)n^{-4} + 8x_\Delta^4\frac{c_6}{4}n^{-4}$$

Because the latter two terms are much smaller in order, we can apply coarser techniques. In particular, we use the following bound:

$$\mathrm{Var}(\mathbb{E}[U|V]) = \mathrm{Var}(U) - \mathbb{E}(\mathrm{Var}(U|V)) \leq E[U^2]$$

Applying to the second term,

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}\left(\mathbb{E}\left[\left(T'_\Pi \frac{d_\Pi - d'_\Pi}{d_\Pi}\right)^2 \middle| \Pi = \pi\right]\right) \\
&\leq \mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right)^4\right] \\
&\leq \sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^8\right]\mathbb{E}[(d_\Pi - d'_\Pi)^8]} \\
&\leq \sqrt{f_{c_6}(8)n^{-8/2} f_{c_4}(8)n^{-8}} \text{ from (3.10), (3.7)} \\
&= \sqrt{f_{c_6}(8) f_{c_4}(8)} n^{-6} \\
&:= c_7 n^{-6}
\end{aligned}$$

And to the third,

$$\begin{aligned}
\mathrm{Var}(Z) &= 4\,\mathrm{Var}\left(\mathbb{E}\left[\left(\frac{2x_{\Pi(J)} - 2x_{\Pi(I)}}{d_\Pi} T'_\Pi \frac{d_\Pi - d'_\Pi}{d_\Pi}\right)\middle| \Pi = \pi\right]\right) \\
&\leq 16 x_\Delta^2 \mathbb{E}\left[\left(\frac{1}{d_\Pi}\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right)^2\right] \\
&\leq 16 x_\Delta^2 f_{c_2}(2) n^{-2/2} \sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^4\right]\mathbb{E}[(d_\Pi - d'_\Pi)^4]} \text{ from (3.5)} \\
&\leq 16 x_\Delta^2 f_{c_2}(2) n^{-1} \sqrt{f_{c_6}(4) n^{-4/2} f_{c_4}(4) n^{-4}} \text{ from (3.10), (3.7)} \\
&\leq 16 x_\Delta^2 f_{c_2}(2) (f_{c_6}(4))^{-1/2} (f_{c_4}(4))^{-1/2} n^{-4} \\
&:= c_8 n^{-4}
\end{aligned}$$

$$
\begin{aligned}
&\frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} \\
&= n\sqrt{(\mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z))} \\
&\leq n\sqrt{8(f_{c_2}(2))^2\left(20 + 16\left(\sum_{i=1}^{2n} x_i^4\right)n^{-2}\right)n^{-4} + 8x_\Delta^4\frac{c_6}{4}n^{-4} + c_7 n^{-6} + c_8 n^{-4}} \\
&:= n^{-1}c_3\sqrt{20 + 16\frac{\sum_{i=1}^{2n} x_i^4}{n^2}}
\end{aligned}
$$

□

## C.3 Proof of Proposition 3.2

*Proof.* The strategy is to break apart the remainder term from the main piece. From (3.11),

$$
\begin{aligned}
\mathbb{E}|T'_\Pi - T_\Pi|^3 &= \left(\frac{n-1}{n}\right)^{3/2}\mathbb{E}\left[d_\Pi^{-3}\left|2x_{\Pi(J)} - 2x_{\Pi(I)} + q'_\Pi\frac{d_\Pi - d'_\Pi}{d'_\Pi}\right|^3\right] \\
&\leq 8\left(8x_\Delta^3\mathbb{E}[d_\Pi^{-3}] + \sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^6\right]\mathbb{E}[(d_\Pi - d'_\Pi)^6]}\right) \\
&\leq 64x_\Delta^3 f_{c_2}(3)n^{-3/2} + 8\sqrt{f_{c_6}(6)n^{-6/2}f_{c_4}(6)n^{-6}} \text{ from (3.5), (3.10), (3.7)} \\
&\leq \frac{c_9^2}{2}n^{-3/2}
\end{aligned}
$$

Therefore,

$$
(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi - T_\Pi|^3}{\lambda}} \leq (2\pi)^{-1/4}c_9 n^{-1/4}.
$$

□

## C.4 Proof of Proposition 3.6

*Proof.*

$$\begin{aligned}
\mathbb{E}|R| &= \mathbb{E}\left|\left(\frac{n}{2}\right)\sqrt{\frac{n-1}{n}}\frac{1}{d_\Pi}\mathbb{E}\left[q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\middle|T_\Pi\right]\right| \\
&\leq \frac{n}{2}\mathbb{E}\left|\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right| \\
&\leq \frac{n}{2}\sqrt{\mathbb{E}\left|\frac{q'_\Pi}{d_\Pi d'_\Pi}\right|^2\mathbb{E}[d_\Pi - d'_\Pi]^2} \\
&\leq \frac{n}{2}\sqrt{f_{c_6}(2)n^{-2/2}f_{c_4}(2)n^{-2}} \text{ from (3.10), (3.7)} \\
&= \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}n^{-1/2}
\end{aligned}$$

□

## C.5 Proof of Proposition 3.5

*Proof.*

$$
\begin{aligned}
\mathbb{E}|T_\Pi R| &= \mathbb{E}\left|T_\Pi\left(\frac{n}{2}\right)\sqrt{\frac{n-1}{n}}\frac{1}{d_\Pi}\mathbb{E}\left[q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\middle|T_\Pi\right]\right| \\
&\leq \frac{n}{2}\mathbb{E}\left|T_\Pi\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right| \\
&\leq \frac{n}{2}\sqrt{\mathbb{E}T_\Pi^2\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^2(d_\Pi - d'_\Pi)^2\right]} \\
&\leq \frac{n}{2}\sqrt{\mathbb{E}T_\Pi^2\sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^4\right]\mathbb{E}[(d_\Pi - d'_\Pi)^4]}} \\
&\leq \frac{n}{2}\sqrt{\mathbb{E}T_\Pi^2\sqrt{f_{c_6}(4)n^{-4/2}f_{c_4}(4)n^{-4}}} \text{ from (3.10), (3.7)} \\
&= \frac{n^{-1/2}}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{\mathbb{E}T_\Pi^2} \\
&\leq \frac{1}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{2+2c_1}n^{-1/2}
\end{aligned}
$$

because $\mathbb{E}T_\Pi^2 \leq 1 + \frac{1+2c_1}{n} \leq 2 + 2c_1$. □

# References


[1] A. Aizerman, E.M. Braverman, and L.I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] A.D. Barbour and L.H.Y. Chen, editors. *An introduction to Stein's method*, volume 4 of *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*. Singapore University Press, Singapore, 2005. Lectures from the Meeting on Stein's Method and Applications: a Program in Honor of Charles Stein held at the National University of Singapore, Singapore, July 28–August 31, 2003.

[3] A.D. Barbour and Louis H.Y. Chen, editors. *Stein's method and applications*, volume 5 of *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*. Published jointly by Singapore University Press, Singapore, 2005.

[4] A.D. Barbour and G.K. Eagleson. Multiple comparisons and sums of dissociated random variables. *Adv. in Appl. Probab.*, 17(1):147–162, 1985.

[5] W.U. Behrens. *Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen*. Arbeiten aus dem Institute für Pflanzenbau der Universität Königsberg i. Pr. Institut für Pflanzenbau, 1928.

[6] V. Bentkus and F. Götze. The Berry-Esseen bound for Student's statistic. *Ann. Probab.*, 24(1):491–503, 1996.

[7] M. Bloznelis. A Berry-Esseen bound for finite population Student's statistic. *Ann. Probab.*, 27(4):2089–2108, 1999.

[8] E. Bolthausen. An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. Verw. Gebiete*, 66(3):379–386, 1984.

[9] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.P. Kriegel, B. Schölkopf, and A.J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[10] J.F. Box. Guinness, Gosset, Fisher, and small samples. *Statist. Sci.*, 2(1):45–52, 1987.

[11] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[12] D.G. Chapman. Some two sample tests. *Ann. Math. Statistics*, 21:601–606, 1950.

[13] L.H.Y. Chen. Poisson approximation for dependent trials. *Ann. Probability*, 3(3):534–545, 1975.

[14] L.H.Y. Chen, L. Goldstein, and Q.M. Shao. *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg, 2011.

[15] L.H.Y. Chen and Q.M. Shao. Normal approximation under local dependence. *Ann. Probab.*, 32(3A):1985–2028, 2004.

[16] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[17] H.A. David. The beginnings of randomization tests. *Amer. Statist.*, 62(1):70–72, 2008.

[18] V.H. de la Peña, M.J. Klass, and T.L. Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Ann. Probab.*, 32(3A):1902–1933, 2004.

[19] V.H. de la Peña, T.L. Lai, and Q.M. Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.

[20] P. Diaconis. The distribution of leading digits and uniform distribution *mod* 1. *Ann. Probability*, 5(1):72–81, 1977.

[21] P. Diaconis and R.L. Graham. Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B*, 39(2):262–268, 1977.

[22] P. Diaconis and S. Holmes. Gray codes for randomization procedures. *Statistics and Computing*, 4(4):287–302, 1994.

[23] P. Diaconis and S. Holmes, editors. *Stein's method: expository lectures and applications*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 46. Institute of Mathematical Statistics, Beachwood, OH, 2004. Papers from the Workshop on Stein's Method held at Stanford University, Stanford, CA, 1998.

[24] P. Diaconis and E. Lehmann. Comment. *J. Amer. Statist. Assoc.*, 103(481):16–19, 2008.

[25] E.J. Dudewicz and S.U. Ahmed. New exact and asymptotically optimal solution to the Behrens–Fisher problem, with tables. *American Journal of Mathematical and Management Sciences*, 18(3-4):359–426, 1998.

[26] E.J. Dudewicz and S.U. Ahmed. New exact and asymptotically optimal heteroscedastic statistical procedures and tables, ii. *American Journal of Mathematical and Management Sciences*, 19(1-2):157–180, 1999.

[27] B. Efron. Student's $t$-test under symmetry conditions. *J. Amer. Statist. Assoc.*, 64:1278–1302, 1969.

[28] R.A. Fisher. *The design of experiments.* Oliver & Boyd, 1935.

[29] R.A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398, 1935.

[30] R.A. Fisher. *Statistical methods for research workers*. Hafner Publishing Co., New York, 1973. Fourteenth edition—revised and enlarged.

[31] D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

[32] J.H. Friedman. On Multivariate Goodness–of–Fit and Two–Sample Testing. *Proceedings of Phystat2003, http://www.slac.stanford.edu/econf/C*, 30908, 2003.

[33] J.H. Friedman and L.C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.*, 7(4):697–717, 1979.

[34] J. Gentry. *twitteR: R based Twitter client*, 2011. R package version 0.99.6.

[35] E. Giné, F. Götze, and D.M. Mason. When is the Student $t$-statistic asymptotically standard normal? *Ann. Probab.*, 25(3):1514–1531, 1997.

[36] A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, and A.J. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19:513–520, 2007.

[37] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[38] A. Gretton, K. Fukumizu, Z. Harchaoui, and B.K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, 22:673–681, 2010.

[39] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[40] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, UC Santa Cruz, 1999.

[41] S.T. Ho and L.H.Y. Chen. An $L_p$ bound for the remainder in a combinatorial central limit theorem. *Ann. Probability*, 6(2):231–249, 1978.

[42] W. Hoeffding. A combinatorial central limit theorem. *Ann. Math. Statistics*, 22:558–566, 1951.

[43] H. Hotelling. The generalization of Student's ratio. *Ann. Math. Statistics*, pages 360–378, 1931.

[44] P.L. Hsu. Contribution to the theory of "Student's" $t$-test as applied to the problem of two samples. *Statistical Research Memoirs*, 2:1–24, 1938.

[45] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[46] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971.

[47] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. $\ell_p$-norm multiple kernel learning. *J. Mach. Learn. Res.*, 12:953–997, 2011.

[48] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[49] H.W. Kuhn and A.W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 481–492, Berkeley and Los Angeles, 1951. University of California Press.

[50] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2003/04.

[51] L. Le Cam. The central limit theorem around 1935. *Statist. Sci.*, 1(1):78–96, 1986. With comments, and a rejoinder by the author.

[52] E.L. Lehmann. *Elements of large-sample theory*. Springer Texts in Statistics. Springer-Verlag, New York, 1999.

[53] E.L. Lehmann. *Fisher, Neyman, and the creation of classical statistics*. Springer, New York, 2011.

[54] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer Verlag, 2005.

[55] C. Leslie, E. Eskin, and W.S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 566–575. Hawaii, USA., 2002.

[56] F.F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, April 2007.

[57] H.T. Lin, C.J. Lin, and R.C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.

[58] B.F. Logan, C.L. Mallows, S.O. Rice, and L.A. Shepp. Limit distributions of self-normalized sums. *Ann. Probability*, 1:788–809, 1973.

[59] B.F. Logan, C.L. Mallows, S.O. Rice, and L.A. Shepp. Limit distributions of self-normalized sums. *Ann. Probability*, 1:788–809, 1973.

[60] J. Ludbrook and H. Dudley. Why permutation tests are superior to $t$- and $F$-tests in biomedical research. *American Statistician*, pages 127–132, 1998.

[61] M. Motoo. On the Hoeffding's combinatorial central limit theorem. *Ann. Inst. Statist. Math. Tokyo*, 8:145–154, 1957.

[62] J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Estadística*, 17:587–651, 1959.

[63] J. Neyman and E.S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20(1/2):175–240, 1928.

[64] G.E. Noether. On a theorem by Wald and Wolfowitz. *Ann. Math. Statistics*, 20(3):455–458, 1949.

[65] M. Panov, A. Tatarchuk, V. Mottl, and D. Windridge. A modified neutral point method for kernel-based fusion of pattern-recognition modalities with incomplete data sets. *Multiple Classifier Systems*, pages 126–136, 2011.

[66] I. Pinelis. On the Berry-Esseen bound for the Student statistic. *arXiv preprint arXiv:1101.3286*, 2011.

[67] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[68] J.C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[69] N. Poh, D. Windridge, V. Mottl, A. Tatarchuk, and A. Eliseyev. Addressing missing values in kernel-based multimodal biometric fusion using neutral point substitution. *Information Forensics and Security, IEEE Transactions on*, 5(3):461–469, 2010.

[70] V.N. Prokof'yev and A.D. Shishkin. Successive classification of normal sets with unknown variances. *Radio Engineering Electronics Physics*, 19:141–143, 1974.

[71] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[72] Y. Rinott and V. Rotar. On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted *U*-statistics. *Ann. Appl. Probab.*, 7(4):1080–1105, 1997.

[73] I.A. Salama and D. Quade. A note on Spearman's footrule. *Comm. Statist. Simulation Comput.*, 19(2):591–601, 1990.

[74] W. Schneller. A short proof of Motoo's combinatorial central limit theorem using Stein's method. *Probab. Theory Related Fields*, 78(2):249–252, 1988.

[75] W. Schneller. Edgeworth expansions for linear rank statistics. *Ann. Statist.*, 17(3):1103–1123, 1989.

[76] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[77] B. Schölkopf, K. Tsuda, and J.P. Vert. *Kernel Methods in Computational Biology*. MIT press, 2004.

[78] P.K. Sen, I.A. Salama, and D. Quade. Spearman's footrule: asymptotics in applications. *Chil. J. Stat.*, 2(1):3–20, 2011.

[79] R.J. Serfling. Probability inequalities for the sum in sampling without replacement. *Ann. Statist.*, 2:39–48, 1974.

[80] Q.M. Shao. Self-normalized large deviations. *Ann. Probab.*, 25(1):285–328, 1997.

[81] Q.M. Shao. An explicit Berry-Esseen bound for Student's $t$-statistic via Stein's method. In *Stein's Method and Applications*, volume 5 of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, pages 143–155. Singapore Univ. Press, Singapore, 2005.

[82] Q.M. Shao and Z.G. Su. The Berry-Esseen bound for character ratios. *Proc. Amer. Math. Soc.*, 134(7):2153–2159, 2006.

[83] N. Smirnoff. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):16, 1939.

[84] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statistics*, 19:279–281, 1948.

[85] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. Bona, A. Binder, C. Gehl, and V. Franc. The shogun machine learning toolbox. *J. Mach. Learn. Res.*, 11:1799–1802, 2010.

[86] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, 2006.

[87] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[88] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, Berkeley, Calif., 1972. Univ. California Press.

[89] C. Stein. *Approximate Computation of Expectations.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 7. Institute of Mathematical Statistics, Hayward, CA, 1986.

[90] Student. Probable error of a correlation coefficient. *Biometrika*, pages 302–310, 1908.

[91] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[92] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.

[93] B. von Bahr. Remainder term estimate in a combinatorial limit theorem. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 35(2):131–139, 1976.

[94] G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In Martin Casdagli and Stephen Eubank,

editors, *Nonlinear Modeling and Forecasting*, pages 95–112. Addison-Wesley Publishing Co Inc, 1992.

[95] G. Wahba et al. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

[96] A. Wald and J. Wolfowitz. Statistical tests based on permutations of the observations. *Ann. Math. Statistics*, 15:358–372, 1944.

[97] S. L. Zabell. On Student's 1908 article "The probable error of a mean". *J. Amer. Statist. Assoc.*, 103(481):1–20, 2008. With comments and a rejoinder by the author.