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.

Contents

Chapter 1

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 Stein's proof of the HCCLT via the method of exchangeable pairs.

1.1 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. Several flavors of Stein's method (e.g. the method of exchangeable pairs) proceed via auxiliary randomization. In Appendix B, we reproduce Stein's [37] proof of the HCCLT.

It will be instructive to follow the proof of the HCCLT 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. We also refer to auxiliary results in Appendix A.

1.2 Stein's Theorem

Theorem 1.1 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. generalize Theorem 1.1 to allow for situations in which the regression condition does not hold exactly, and we later in addition relax the assumption that the W has unit variance.

Theorem 1.1 (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

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

1.3 Hoeffding's Combinatorial CLT

Stein's proof of the HCCLT relies on an application of Theorem 1.1.

Theorem 1.2 (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 Π 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

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

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 a $\mathcal{O}(n^{-1/2})$ rate of convergence, we require the function of the data to be bounded. However, this will not typically be the case.

Bolthausen [3] was able to prove the following result:

Theorem 1.3 (Bolthausen)**.** *Under the conditions of Theorem 1.2, 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.

Now, we return to generalizing Theorem 1.1.

1.4 Generalized Stein's Theorems

Here, we treat the situation where the regression condition fails to hold exactly. Chen et al. [6] serves as an excellent reference for results of this type.

Definition 1.4 (Approximate Stein Pair)**.** *Let (W, W') be an exchangeable pair. If the pair satisfies the "approximate linear regression condition"*

$$\mathbb{E}[W - W'|W] = \lambda(W - R), \tag{1.1}$$

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

Here we generalize Lemma 5.1 from [6] to the setting of non-unit variance:

Theorem 1.5. *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 a $\mathcal{O}(n^{-1/4})$ rate under mild conditions. Generalization of Theorem 5.5 from [6]:

Theorem 1.6. *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*

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

The following result will let us achieve a rate of $\mathcal{O}(n^{-1/2})$ subject to an additional constraint on the data. Generalization of part of Theorem 5.3 from [6]:

Theorem 1.7. *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 *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 2

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 1.6 and 1.7 of Chapter 1.

2.1 Motivation

Motivated by concerns regarding normality assumptions in the hypothesis being tested, Fisher [12] 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 [9] used efficient Gray code based calculations to show that the randomization distribution looked remarkably normal. For more history on the development of randomization procedures, see Zabell [40] or David [8]. Diaconis and Lehmann [10] in their comment on Zabell's paper further expanded on some properties of these randomization tests.

Ludbrook and Dudley [26] 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 Newman and Pearson [27], and Fisher's randomization model [12]. Add some more on these two models...

Under the randomization model and using the language of triangular arrays, Lehmann [22] proved a weak convergence result of the randomization distribution of the t-statistic to the standard normal distribution, however, there is no known Berry-Esseen type bound for this rate of convergence.

Introduced by Stein [37], the eponymous technique provides a powerful means with which to handle dependencies among collections of random variables, a common criticism of classical Fourier analytic methods. In addition, one can easily obtain bounds on rates of convergence. Bentkus and Götze [2] first obtained a Berry-Esseen bound for Student's statistic in the independent but non-identically distributed setting with additional work by Shao [34].

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 a $\mathcal{O}(n^{-1/2})$ rate.

2.2 Set-up

We observe two samples with equal sample size: $S_1 = \{u_i\}_{i=1}^N$ and $S_2 = \{u_i\}_{i=N+1}^{2N}$. Since we consider the t-statistic under different permutations, it will be convenient to re-write the sample values relative to the null permutation π_0: $S_1 = \{u_{\pi_0(i)}\}_{i=1}^N$ and $S_2 = \{u_{\pi_0(i)}\}_{i=N+1}^{2N}$, where $\pi_0(i) = i$. Under the randomization distribution, where Π is a uniformly chosen permutation, Student's two-sample t-statistic is given by

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

where

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

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

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

2.3 Assumptions

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

$$u_i \leftarrow \sqrt{\frac{2N}{\sum_{i=1}^{2N}(u_i - \bar{u})^2}}(u_i - \bar{u}), \tag{2.1}$$

so we just assume that the u_i's have already been transformed. This can be seen as a very mild assumption: only $u_i = c$ for all i cannot be scaled in this way.

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

$$d_\Pi = \sqrt{\sum_{i=1}^{N}(u_{\Pi(i)} - \bar{u}_{1,\Pi})^2 + \sum_{i=N+1}^{2N}(u_{\Pi(i)} - \bar{u}_{2,\Pi})^2} > 0 \tag{2.2}$$

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 $\sum_{i=1}^{N} u_{\Pi(i)} = -\sum_{i=N+1}^{2N} u_{\Pi(i)}$ and hence that $\bar{u}_{1,\Pi} = -\bar{u}_{2,\Pi}$ for all Π.

Here we establish an equality with d_Π that will prove easier to work with:

$$\begin{aligned}
d_\Pi^2 &= \sum_{i=1}^{N}(u_{\Pi(i)} - \bar{u}_{1,\Pi})^2 + \sum_{i=N+1}^{2N}(u_{\Pi(i)} - \bar{u}_{2,\Pi})^2 \\
&= \sum_{i=1}^{2N} u_{\Pi(i)}^2 - N\bar{u}_{1,\Pi}^2 - N\bar{u}_{2,\Pi}^2 \\
&= 2N - N\bar{u}_{2,\Pi}^2 - N\bar{u}_{2,\Pi}^2 \\
&= 2N(1 - \bar{u}_{2,\Pi}^2)
\end{aligned}$$

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

$$B = \max_{\Pi} |\bar{u}_{2,\Pi}| < 1. \tag{2.3}$$

2.4 Preliminaries

Here we collect useful bounds and other results. We include them here rather than in Appendix A because in Chapter 3 we compare the theoretical bounds with simulated results.

In order to bound various moments of $\bar{u}_{2,\Pi}$ under the permutation distribution, we use a result of Serfling's [33]:

Proposition 2.1. *Consider sampling without replacement from a finite list of values $u_1, \dots, u_{2N}$. Let $u_\Delta = \max_i u_i - \min_i u_i$. Then for $p > 0$,*

$$\begin{aligned}
\mathbb{E}[\bar{u}^p_{2,\Pi}] &\leq \frac{\Gamma(p/2+1)}{2^{p/2+1}} \left[\frac{N+1}{2N} u^2_\Delta\right]^{p/2} (2N)^{-p/2} \\
&\leq \frac{\Gamma(p/2+1)}{2^{p/2+1}} \left[\frac{N+1}{4N} u^2_\Delta\right]^{p/2} N^{-p/2} \\
&:= f_{c_1}(p) N^{-p/2}.
\end{aligned} \tag{2.4}$$

By Assumption (2.3),

$$(d_\Pi)^{-p} = \frac{1}{(2N(1-\bar{u}^2_{2,\Pi}))^{p/2}} \leq \frac{1}{(2N(1-B^2))^{p/2}} := f_{c_2}(p) N^{-p/2}. \tag{2.5}$$

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

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

We consider the difference

$$
\begin{aligned}
h_\Pi &= d_\Pi^2 - d_\Pi'^2 \\
&= 2N - 2N\bar{u}_{2,\Pi}^2 - 2N + 2N\bar{u}_{2,\Pi}'^2 \\
&= 4\bar{u}_{2,\Pi}(u_{\Pi(I)} - u_{\Pi(J)}) + \frac{2}{N}(u_{\Pi(I)} - u_{\Pi(J)})^2
\end{aligned}
$$

Therefore, by the c_r-inequality,

$$
\begin{aligned}
\mathbb{E}[h_\Pi^p] &= \mathbb{E}\left|4\bar{u}_{2,\Pi}(u_{\Pi(I)} - u_{\Pi(J)}) + \frac{2}{N}(u_{\Pi(I)} - u_{\Pi(J)})^2\right|^p \\
&\leq 2^{p-1}\left(\mathbb{E}|4\bar{u}_{2,\Pi}(u_{\Pi(I)} - u_{\Pi(J)})|^p + \mathbb{E}\left|\frac{2}{N}(u_{\Pi(I)} - u_{\Pi(J)})^2\right|^p\right) \\
&\leq 2^{p-1}\left[(4u_\Delta)^p\mathbb{E}\left|\bar{u}_{2,\Pi}\right|^p + \left(\frac{2}{N}u_\Delta^2\right)^p\right] \\
&\leq 2^{p-1}(4u_\Delta)^p f_{c_1}(p)N^{-p/2} + 2^{p-1}(2u_\Delta^2)^p N^{-p} \\
&:= f_{c_3}(p)N^{-p/2}. \qquad (2.6)
\end{aligned}
$$

Now we establish a bound on the difference $d_\Pi - d_\Pi'$ via a bound on the remainder of a zeroth order Taylor approximation. Write

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

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_Π' by a constant and bounding the error via a function of the first derivative. This is a sufficient approximation because the squared difference h_Π 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,

$$\begin{aligned} \mathbb{E}|d_\Pi - d_\Pi'|^p &\leq 2^{p-1}\left(\mathbb{E}\left|\frac{h_\Pi}{2d_\Pi}\right|^p + \mathbb{E}\left|\frac{h_\Pi}{2d_\Pi'}\right|^p\right) \\ &\leq 2^{-1}\left(\sqrt{\mathbb{E}[h_\Pi^{2p}]\mathbb{E}[d_\Pi^{-2p}]} + \sqrt{\mathbb{E}[h_\Pi^{2p}]\mathbb{E}[d_\Pi'^{-2p}]}\right) \\ &\leq \sqrt{f_{c_3}(2p)N^{-2p/2} f_{c_2}(2p)N^{-2p/2}} \qquad \text{by (2.6) and (2.5)} \\ &:= f_{c_4}(p)N^{-p}. \end{aligned} \tag{2.7}$$

With

$$q_\Pi = N\bar{u}_{1,\Pi} - N\bar{u}_{2,\Pi} = -2N\bar{u}_{2,\Pi}, \tag{2.8}$$

(2.4), and noting that q_Π and q_Π' are exchangeable,

$$\mathbb{E}[q_\Pi'^p] = \mathbb{E}[q_\Pi^p] = \mathbb{E}[(-2N\bar{u}_{2,\Pi})^p] \leq 2^p N^p f_{c_1}(p) N^{-p/2} := f_{c_5}(p). \tag{2.9}$$

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^p\right] &\leq \sqrt{\mathbb{E}|q'_\Pi|^{2p}\mathbb{E}|d_\Pi d'_\Pi|^{-2p}} \\
&\leq \sqrt{\mathbb{E}|q_\Pi|^{2p}\sqrt{\mathbb{E}|d_\Pi|^{-4p}\mathbb{E}|d'_\Pi|^{-4p}}} \\
&= \sqrt{\mathbb{E}|q_\Pi|^{2p}\mathbb{E}|d_\Pi|^{-4p}} \\
&\leq \sqrt{f_{c_5}(2p)N^{2p/2}f_{c_2}(4p)N^{-4p/2}} \qquad \text{from (2.9) and (2.5)} \\
&:= f_{c_6}(p)N^{-p/2}. \qquad (2.10)
\end{aligned}
$$

2.5 Proof

We proceed to verify the conditions of Theorems 1.6 and 1.7. T_Π and T'_Π are exchangeable by construction:

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

Since (Π, Π') are exchangeable, $(T_\Pi, T'_\Pi) = (T(\Pi), T(\Pi'))$ are exchangeable as well. T_Π, and thus T'_Π by exchangeability, have mean zero by symmetry. Let π^* identify the permutation that reverses the order of the indices after applying the original permutation π. 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, π^* reverses the groups after π, so

$T_{\pi^*} = -T_\pi$.

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

To show the approximate regression condition, the difference of our exchangeable pair is given by

$$\begin{aligned}
T'_\Pi - T_\Pi &= \sqrt{\frac{N-1}{N}} \left(\frac{q'_\Pi}{d'_\Pi} - \frac{q_\Pi}{d_\Pi} \right) \\
&= \sqrt{\frac{N-1}{N}} \frac{1}{d_\Pi} \left(q'_\Pi - q_\Pi + q'_\Pi \frac{(d_\Pi - d'_\Pi)}{d'_\Pi} \right) \\
&= \sqrt{\frac{N-1}{N}} \frac{1}{d_\Pi} \left(2u_{\Pi(J)} - 2u_{\Pi(I)} + q'_\Pi \frac{(d_\Pi - d'_\Pi)}{d'_\Pi} \right). \qquad (2.11)
\end{aligned}$$

Note that

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

Therefore,

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

and

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

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

so

$$R_\Pi = \binom{N}{2}\sqrt{\frac{N-1}{N}}\frac{1}{d_\Pi}\mathbb{E}\left[q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\middle|T_\Pi\right]. \tag{2.12}$$

For convenience, we restate Theorem 1.6 of Chapter 1, taking our random variables W to be the randomization t-statistic T_Π and W' to be its coupled counterpart T'_Π:

Theorem 1.8. *If T_Π, T'_Π 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_Π, then

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

With the preliminaries in place, we proceed to provide bounds on each term in Theorem 1.6, the proofs of which we defer to Appendix C.

Proposition 2.2. $(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi - T_\Pi|^3}{\lambda}} < (2\pi)^{-1/4}c_9 N^{-1/4}$.

Proposition 2.3. $\frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} \le N^{-1}c_3\sqrt{20 + 16\frac{\sum_{i=1}^{2N} u_i^4}{N^2}}$

Proposition 2.4. $|\mathbb{E}T_\Pi^2 - 1| \le c_2 N^{-1}$

Proposition 2.5. $\mathbb{E}|T_\Pi R| \le \frac{1}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{2+2c_1}N^{-1/2}$.

Proposition 2.6. $\mathbb{E}|R| \le \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}$.

The bound in Proposition 2.2 is suboptimal, as it will only allow us to obtain a rate of $\mathcal{O}(n^{-1/4})$. In Section 2.6, we introduce an additional condition to improve upon this rate.

Collecting the results of Propositions 2.4, 2.3, 2.2, 2.6, and 2.5, we have

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

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

$$\sum_{i=1}^{2N} u_i^4 \le \left(\sum_{i=1}^{2N} u_i^2\right)^{4/2} = (2N)^2 = 4N^2.$$

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.

2.6 Better Rate

Here, we use Theorem 1.7 to establish a rate of $\mathcal{O}(n^{-1/2})$ with the condition that $|T_\Pi - T'_\Pi| \le \delta$ is $\mathcal{O}(n^{-1/2})$.

From Proposition 2.4, $\mathbb{E}T_\Pi^2 \le c_2N^{-1} + 1$, and from Proposition 2.6, $\mathbb{E}|R| \le \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}$. If $\delta < c_{10}N^{-1/2}$ for N sufficiently large, applying Theorem 1.7, we see

$$\begin{aligned}
\sup_{t\in\mathbb{R}} |P(T_\Pi \le t) - \Phi(t)| &\le \frac{.41\delta^3}{\lambda} + 3\delta\left(\sqrt{\mathbb{E}T_\Pi^2} + \mathbb{E}|R|\right) + \frac{1}{2\lambda}\sqrt{\operatorname{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} \\
&\quad + |\mathbb{E}T_\Pi^2 - 1| + \mathbb{E}|T_\Pi R| + \mathbb{E}|R| \\
&\le .205c_{10}N^{-1/2} + 3c_{10}N^{-1/2}\left(c_2N^{-1} + 1 + \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}\right) \\
&\quad + N^{-1}c_3\sqrt{20 + 16\frac{\sum_{i=1}^{2N} u_i^4}{N^2}} + c_2N^{-1} \\
&\quad + \frac{1}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{2+2c_1}N^{-1/2} + \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}.
\end{aligned}$$

Again, this result is conditional on the data. We can consider a sequence of vectors $\{u_i^{(2N)}\}$, where each $u_i^{(j)}$ 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 [3].

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

$$T_\Pi(\{u_{\Pi(i)}\}_{i=1}^N, \{u_{\Pi(i)}\}_{i=N+1}^{2N}) = \frac{\bar{u}_{1,\Pi} - \bar{u}_{2,\Pi}}{\sqrt{\frac{\frac{1}{N-1}\sum_{i=1}^N (u_{\Pi(i)} - \bar{u}_{1,\Pi})^2}{N} + \frac{\frac{1}{N-1}\sum_{i=N+1}^{2N}(u_{\Pi(i)} - \bar{u}_{2,\Pi})^2}{N}}}.$$

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. Well, because the t-statistic is invariant to permutations within groups, there are $\binom{2N}{N}$ (really, $\binom{2N}{N}/2$ because of symmetry) 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 $\{u_{(i)}\}_{i=1}^N$ and

$\{u_{(i)}\}_{i=N+1}^{2N}$, 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 (is this true? source?).

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

Conjecture 2.7. $\delta = \max_{\pi,i,j} |T_\pi - T_{\pi\circ(i,j)}|$ *is maximized at* $\pi = \pi^*$, $i = i^*$, *and* $j = j^*$.

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

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

Chapter 3

Simulations

This chapter is a computational companion to Chapter 2.

3.1 Preliminaries

First, we provide simulations accompanying Section 2.4. We generate i.i.d. samples $\{u_i\}_{i=1}^{N} \sim \mathcal{N}(-1,1)$ and $\{u_i\}_{i=N+1}^{2N} \sim \mathcal{N}(1,1)$ for exponentially-spaced values of N. The u_i are scaled and centered, and for each N, we perform $10,000$ permutations.

We plot Monte Carlo estimates of the means of each term, scaled by the rate of our bound, along with 95%ile 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.

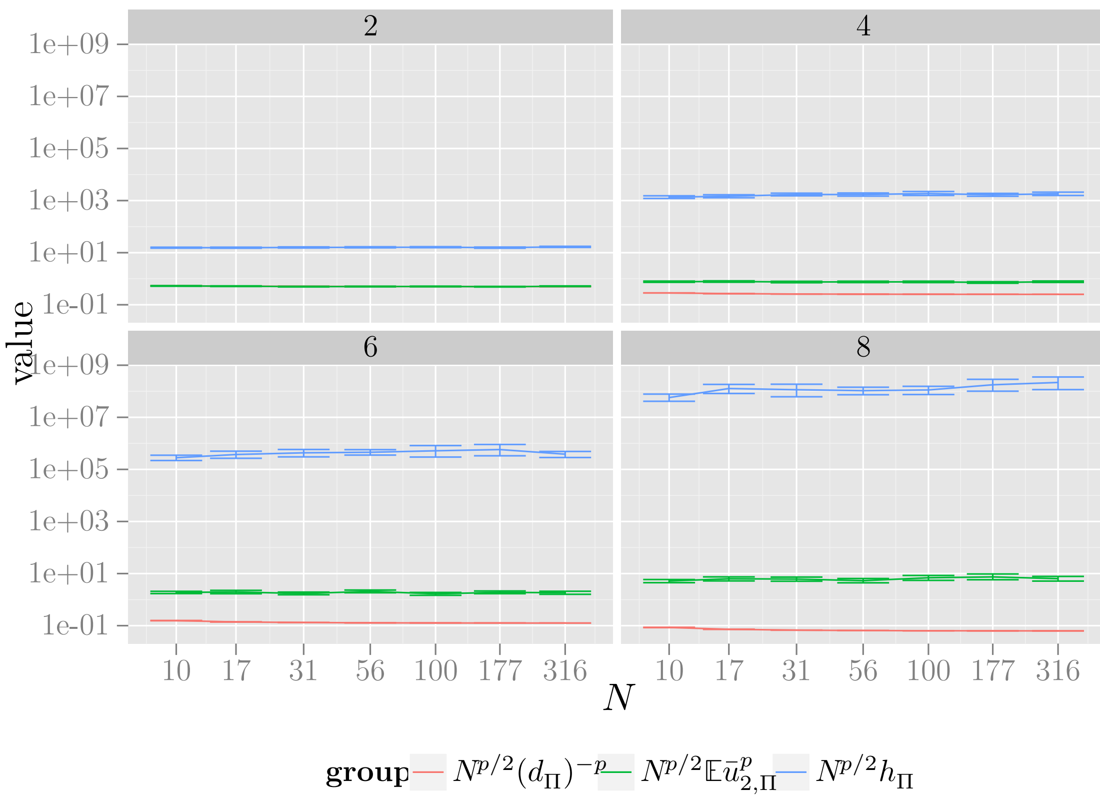

Figure 3.1: Log-log plots of values scaled by proven upper bounds of rates, faceted on p.

Here, 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.

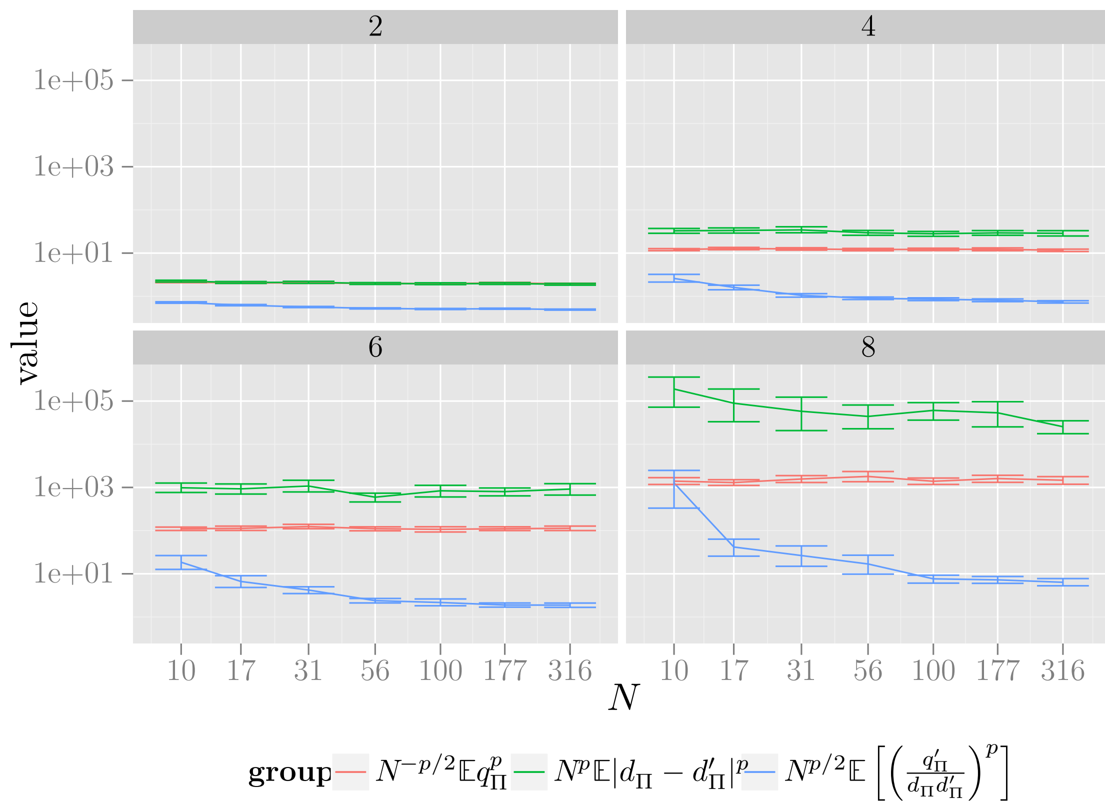

Figure 3.2: Log-log plots of values scaled by proven upper bounds of rates, faceted on p.

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.

3.2 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'_Π on T_Π 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_Π. For each T_Π, 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'_Π

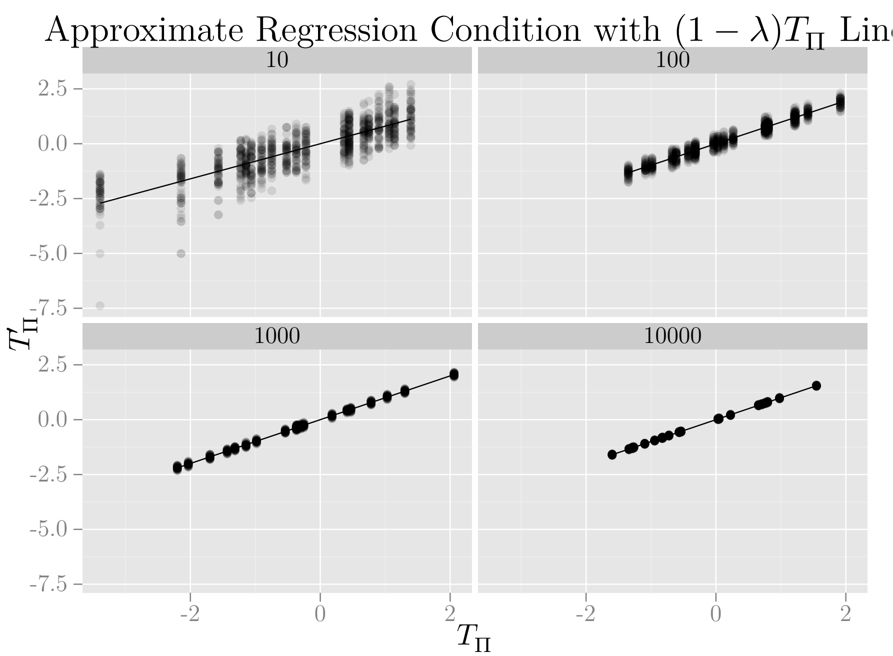

Figure 3.3: Faceted on per-group sample size, N.

The approximate regression condition appears to hold visually.

3.3 Main Bounds

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

3.3.1 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).

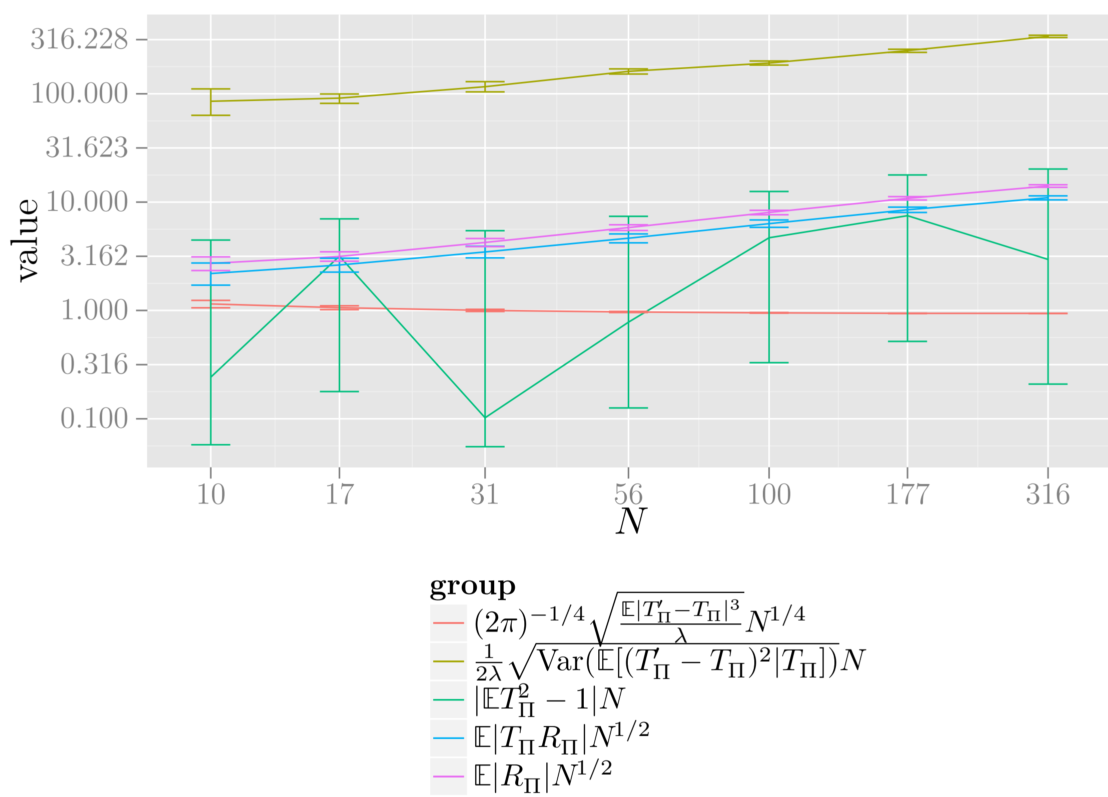

Figure 3.4: Log-log plot of values for each term in the bound, simulating the conditional by Monte Carlo.

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

3.3.2 Exact Conditional Expectation Calculations

Motivated by the 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 3.4. We still use Monte Carlo simulations to calculate a subset of all the permutations Π, otherwise the computational cost would be prohibitive for large sample sizes.

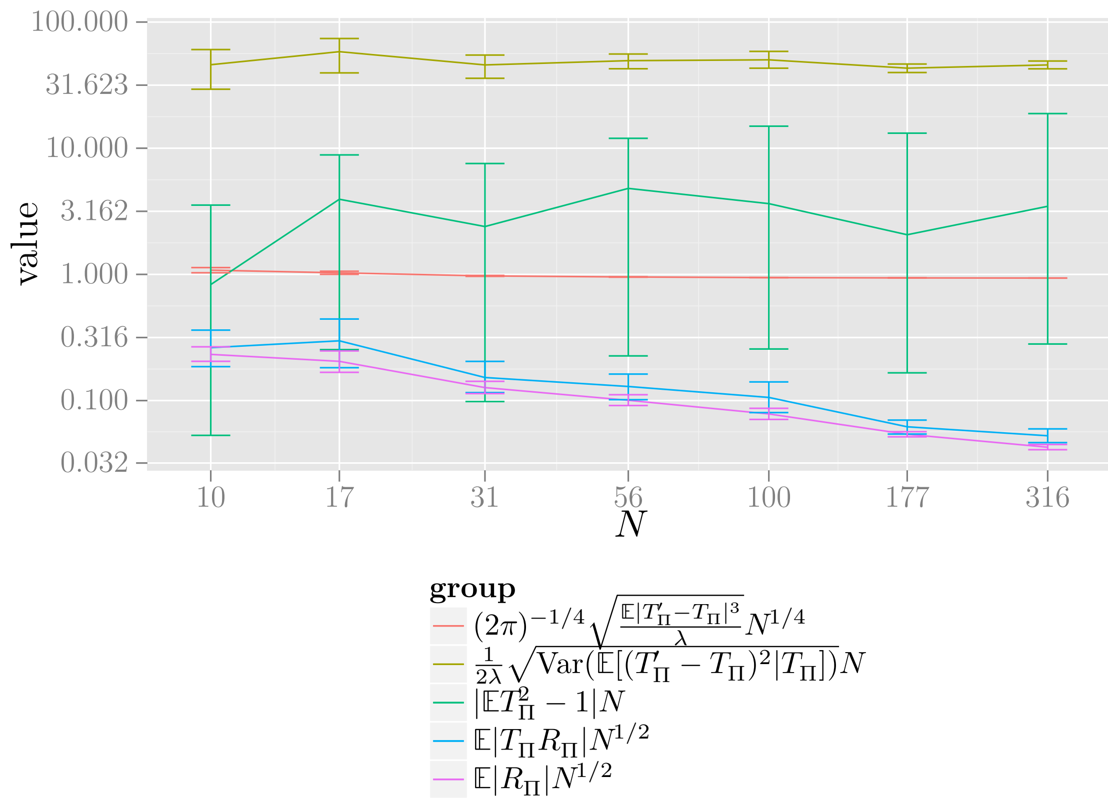

Figure 3.5: Log-log plot of values for each term in the bound, calculating the conditional expectation exactly ($10N$ permutations each).

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.

3.3.3 Better Rate

All terms except the one involving δ^3 appear to be of or better than the proven rate.

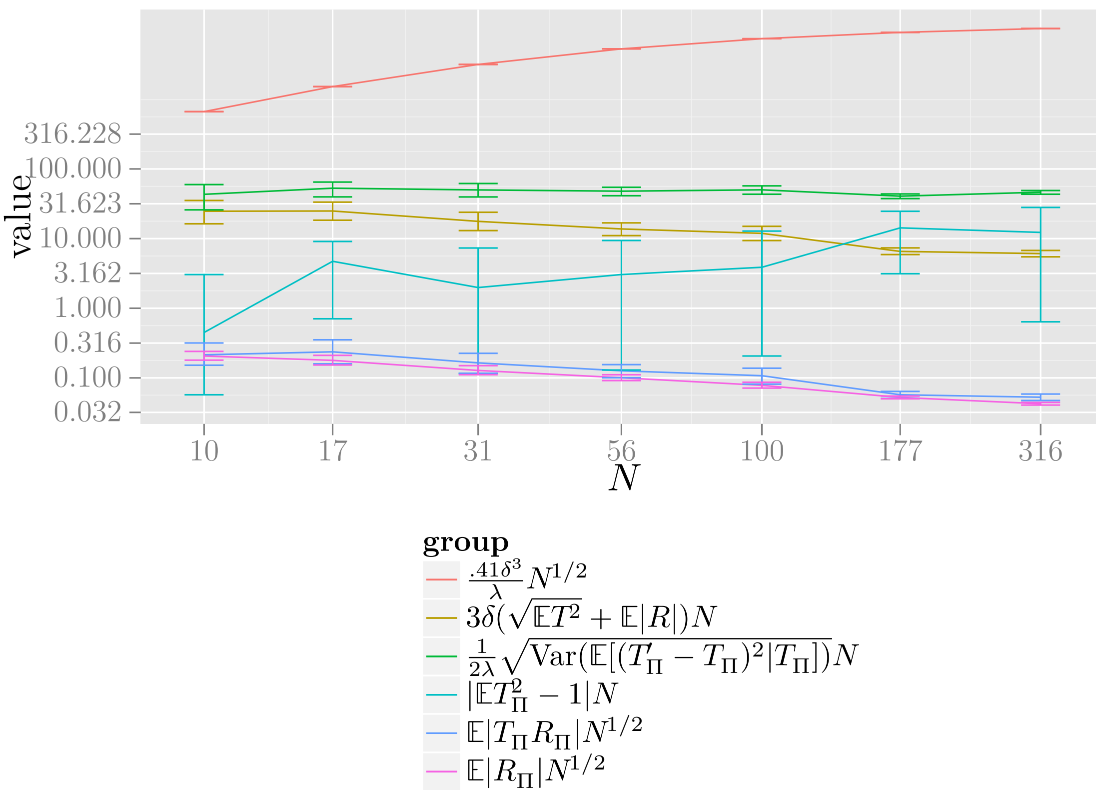

Figure 3.6: Log-log plot of values for each term in the bound, calculating the conditional expectation exactly ($10N$ permutations each).

We shall see that $\frac{.41\delta^3}{\lambda}N^{1/2}$ is bounded in some situations.

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

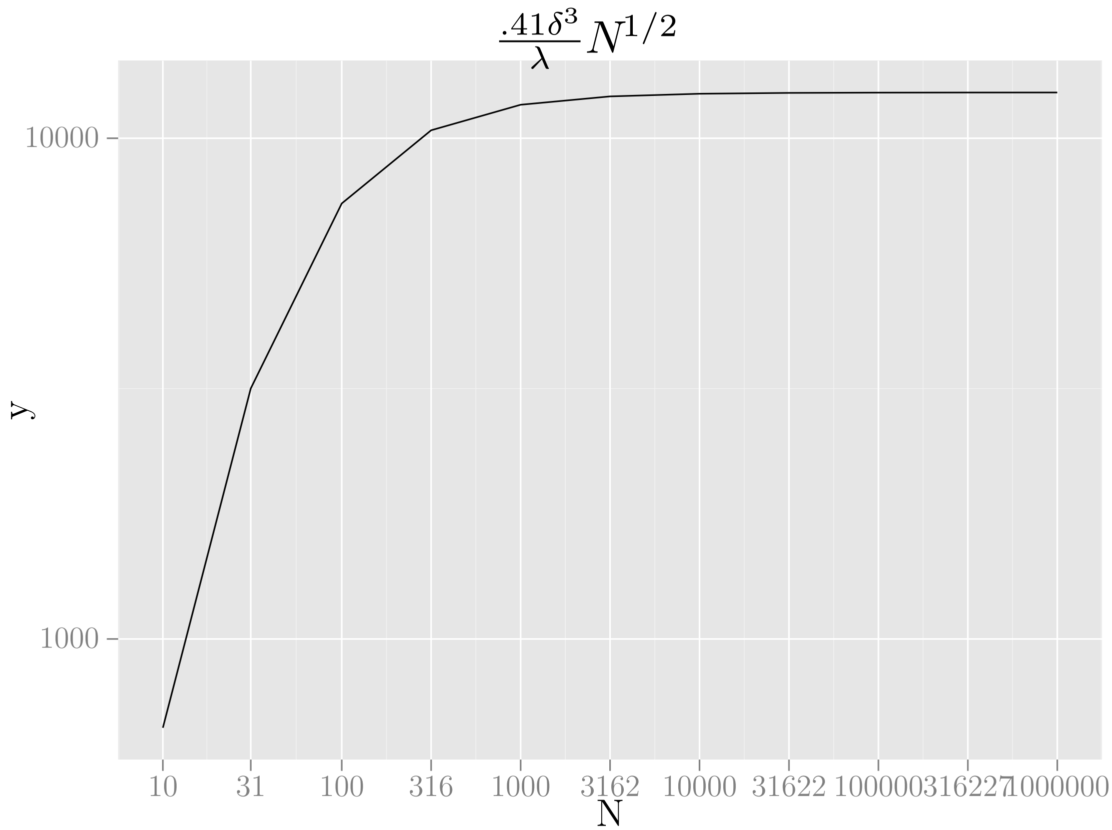

Figure 3.7: $\frac{.41\delta^3}{\lambda}N^{1/2}$ on N.

Recall that

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

3.4 Efficient Updates

Instead of conditioning on the value of T_Π, we condition on the observed permutation π. 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'_Π should not differ much from T_Π, and calculating the t-statistics from scratch 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

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

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

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

$$\begin{aligned}
\Delta &\equiv u_j - x_i \\
\bar{x}_{x_i,u_j} &= \bar{x} - \frac{1}{N}x_i + \frac{1}{N}u_j = \bar{x} + \frac{\Delta}{N} \\
\bar{u}_{x_i,u_j} &= \bar{u} + \frac{1}{N}x_i - \frac{1}{N}u_j = \bar{u} - \frac{\Delta}{N} \\
S^2_{X_{x_i,u_j}} &= \frac{1}{N-1}(\sum_{k=1}^{N} x_k^2 - x_i^2 + u_j^2) - \frac{N}{N-1}\bar{x}^2_{x_i,u_j} \\
S^2_{U_{x_i,u_j}} &= \frac{1}{N-1}(\sum_{k=1}^{N} u_k^2 + x_i^2 - u_j^2) - \frac{N}{N-1}\bar{u}^2_{x_i,u_j} \\
\bar{x}^2_{x_i,u_j} &= \bar{x}^2 + \frac{2\Delta}{N}\bar{x} + \frac{\Delta^2}{N} \\
\bar{u}^2_{x_i,u_j} &= \bar{u}^2 - \frac{2\Delta}{N}\bar{u} + \frac{\Delta^2}{N}
\end{aligned}$$

Then

$$\begin{aligned}
T_{x_i,u_j} &= \frac{\bar{x}_{x_i,u_j} - \bar{u}_{x_i,u_j}}{\sqrt{\frac{2}{N}}\sqrt{\frac{1}{2}(S^2_{X_{x_i,u_j}} + S^2_{U_{x_i,u_j}})}} \\
&= \frac{\bar{x} - \bar{u} + \frac{2\Delta}{N}}{\sqrt{\frac{2}{N}}\sqrt{\frac{1}{2(N-1)}[\sum_{k=1}^{N}(x_k^2 + u_k^2) - N(\bar{x}^2 + \bar{u}^2 + \Delta(\frac{2\bar{x}}{n} - \frac{2\bar{u}}{n}) + \frac{2}{n^2}\Delta^2)]}}.
\end{aligned}$$

Only the terms involving Δ need to be recomputed for each of the N^2 swaps.

3.5 A Different Exchangeable Pair

Rather than only consider transpositions that swap one element of the first group with one from the second group, we have a few different choices. Let's take the other extreme, where we consider all $(2N)^2$ transpositions, including null transpositions. There are N^2 transpositions within each group, for a total of $2N^2$. Each of these does not change the t-statistic. We previously only considered the N^2 transpositions where $I < J$. There are another N^2 with $I > J$. These transpositions have exactly the same effect as the previous group $(I, J) = (J, I)$, and all within-group transpositions have no effect.

The only changes should be to adjust the weights when taking conditional expectations (the weights should be $1/2$) and to divide λ by 2. The new λ is N^{-1}.

However, every term involving the conditional expectation also has a division by lambda, so any decrease in the c.e. is cancelled out by a corresponding decrease in λ, so there is no change in any of the simulations.

It's nice that the calculations are invariant to change in the exchangeable pair. Whether that holds true for more drastic changes (e.g. swapping more than 2 elements) is not known.

Chapter 4

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 2.

4.1 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?

	parametric	non-parametric
univariate	t-test	permutation t-test; Kolmogorov-Smirnov test; Wilcoxon rank-sum test
multivariate	Hotelling's T^2 test	Friedman-Rafsky test
non-vectorial	Maximum Mean Discrepancy (asymptotic)	Friedman's test (KSVM); MMD (distribution-free)
heterogeneous		Friedman's test (MKL)

Table 4.1: Two-sample tests.

4.2 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 α.

There are many two-sample tests in the literature, as Table 4.1 illustrates.

4.3 The Friedman Two-Sample Test

Friedman proposed the following approach to the two-sample problem [13]:

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

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ to create a predictor variable training set.

2. Assign a response value $y_i = 1$ to the observations from the first sample ($1 \leq i \leq N$) and $y_i = -1$ to the observations from the second sample ($N{+}1 \leq i \leq N{+}M$).

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

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

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

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

Note that in Step 3, for a given learning machine, there can still be some choice in the scoring function $f(\mathbf{u})$.

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 $F(\mathbf{u})$. 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.

By virtue of its permutation construction, the test has level α—the probability that we reject the null hypothesis given that the null hypothesis is true, also known as type I error. Given a threshold α, 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 [23]. 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.

4.4 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. There is much literature on kernel methods, but one particularly comprehensive treatment is the monograph by Schölkopf and Smola [31].

Given n observed datapoints in some general set, $x_1, \dots, x_n \in \mathcal{X}$, 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 4.1 (Positive Semidefinite Kernel). *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive semidefinite kernel iff it is symmetric ($K(x, x') = K(x', x)$ for any $x, x' \in \mathcal{X}$) and positive semidefinite:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

for any $n > 0$, any choice of n objects $x_1, \dots, x_n \in \mathcal{X}$, and any $c_1, \dots, c_n \in \mathbb{R}$.

Because inner products are symmetric, positive semidefinite functions, they satisfy Definition 4.1 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 general sets $\mathcal{X}$, we can define a mapping ϕ 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 5, 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 [32]:

Theorem 4.2. *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*

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

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

In the coming sections and in Chapter 5, we shall see examples of nonvectorial spaces $\mathcal{X}$, kernels K, feature mappings ϕ, and Hilbert spaces $\mathcal{H}$.

4.5 Support Vector Machines

A Support Vector Machine (SVM) [7] 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 ℓ_1-regularized (soft margin) support vector classification problem [31] in its primal form:

$$\begin{aligned}
&\underset{\mathbf{w}\in\mathcal{H},b\in\mathbb{R},\xi\in\mathbb{R}^m}{\text{minimize}} && \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\
&\text{subject to} && y_i(\mathbf{w}^t\mathbf{x}_i + b) \geq 1 - \xi_i \\
& && \xi_i \geq 0 \qquad\qquad \text{for all } i = 1, \dots, m.
\end{aligned} \tag{4.1}$$

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

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

2. An estimate of the posterior class probability, such as by a sigmoid (Platt [28, 25]) or a logistic link function (Wahba [38, 39]) $f_2(\mathbf{x}_i) = \frac{1}{1+\exp(-(\mathbf{w}^t\mathbf{x}_i+b))}$.

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 you believe 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 $\operatorname{sign}(w)$. 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).

4.5.1 Kernelized Form

It is advantageous to treat the dual [5] of Problem (4.1). 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 [29] 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

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^m}{\text{minimize}} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \quad \text{for all } i = 1, \ldots, m \\ \text{and} \quad & \sum_{i=1}^m \alpha_i y_i = 0. \end{aligned} \tag{4.2}$$

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

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

Therefore,

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

where α_i are the dual variables.

In fact, there is another view of the SVM problem in the framework of regularized empirical risk minimization that will be useful for Chapter 5. Define the hinge loss function

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

Then the following optimization problem is equivalent to Problem (4.1):

$$\min_{b\in\mathbb{R},\mathbf{w}\in\mathcal{H}} \sum_{i=1}^{m}(1-y_i(\mathbf{w}^T\mathbf{x}_i+b))_+ + \frac{1}{2C}\|\mathbf{w}\|_2^2. \tag{4.3}$$

It turns out that regularized risk minimization problems admit particularly elegant solutions, a result owing to Kimeldorf and Wahba [19]. We present a slightly generalized representer theorem from [31]:

Theorem 4.3 (Representer Theorem)**.** *Let $\Omega : [0,\infty] \to \mathbb{R}$ be a strictly monotonic increasing function, $\mathcal{X}$ be a set, and $c : (\mathcal{X}\times\mathbb{R}^2)^m \to \mathbb{R}\cup\{\infty\}$ be an arbitrary loss function. Then each minimizer $f' \in \mathcal{H}$ of the regularized risk*

$$c((\mathbf{x_1}, y_1, f'(\mathbf{x_1})), \dots, (\mathbf{x_m}, y_m, f'(\mathbf{x_m}))) + \Omega(\|f'\|_{\mathcal{H}}) \tag{4.4}$$

admits a representation of the form

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

Although Problem (4.4) has a feasible set $\mathcal{H}$ that is possibly infinite dimensional, Theorem 4.3 guarantees that the solution lies in the span of the m particular kernels centered on the observations $\mathbf{x}_i$.

We apply Theorem 4.3 to Problem (4.3), noting that it holds for all fixed b, to conclude that

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

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

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

We list the kernelized representations of possible Friedman scoring functions:

1. The predicted class label $f_1(\mathbf{x}_i) = \text{sign}(\sum_{i=1}^{m} y_i\alpha_i K(\mathbf{x}, \mathbf{x}_i) + b)$.

2. An estimate of the posterior class probability, such as by a sigmoid (Platt [28, 25]) or a logistic link function (Wahba [38, 39]) $f_2(\mathbf{x}_i) = \frac{1}{1+\exp(-(\sum_{i=1}^m y_i\alpha_i K(\mathbf{x},\mathbf{x}_i)+b)}$.

3. The margin $f_3(\mathbf{x}_i) = \sum_{i=1}^m y_i\alpha_i K(\mathbf{x},\mathbf{x}_i) + b$.

4.5.2 Equivalence to the Permutation t-test

Theorem 4.4. *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.

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

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

$$T(\{x_i\}_1^N, \{z_i\}_1^M) = \frac{\bar{x} - \bar{z}}{\sqrt{\frac{s_X^2}{N} + \frac{s_z^2}{M}}}$$

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,

$$T(\{f(x_i)\}_1^N, \{f(z_i)\}_1^M) = \frac{w\bar{x} + b - w\bar{z} + b}{|w|\sqrt{\frac{s_X^2}{N} + \frac{s_z^2}{M}}} = \operatorname{sign}(w) T(\{x_i\}_1^N, \{z_i\}_1^M).$$

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

$$|T(\{f(x_i)\}_1^N, \{f(z_i)\}_1^M)| = |T(\{x_i\}_1^N, \{z_i\}_1^M)|.$$

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 [31]:

$$\begin{array}{lll}
\underset{\mathbf{w}\in\mathcal{H},\boldsymbol{\xi}^{(*)}\in\mathbb{R}^m,b\in\mathbb{R}}{\text{minimize}} & \tau(\mathbf{w}, \boldsymbol{\xi}^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) & \\
\text{subject to} & f(\mathbf{x}_i) - y_i \le \epsilon + \xi_i & \\
 & y_i - f(\mathbf{x}_i) \le \epsilon + \xi_i^* & \\
 & \xi_i, \xi_i^* \ge 0 & \text{for all } i = 1, \dots, m.
\end{array}$$

with solution is given by

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

□

Using the bounds derived in Chapter 2, 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

$$\sum_{i=1}^{m} y_i \alpha_i K(x, x_i) + b = cx + d? \tag{4.5}$$

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:

$$\sum_{i=1}^{m} y_i \alpha_i \exp(-\sigma(x - x_i)^2) + b \tag{4.6}$$

cannot be written as $cx + d$.

We use Support Vector Machine (SVM) classification as implemented in the **ksvm** function of the **R** [30] package **kernlab** [18].

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

4.6 Maximum Mean Discrepancy

Gretton et al. [15, 17, 16, 4] introduce 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 4.5. *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*

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

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

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 big 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.

4.7 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 4.4, the Friedman Statistic in the univariate data, linear kernel setting is equal to the $|T|$. In Figure 4.2 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 4.1 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 the for higher dimensional data, C has a miniscule effect on the Friedman statistic for the linear kernel. With an RBF kernel (width 1), it appears that C has a small effect.

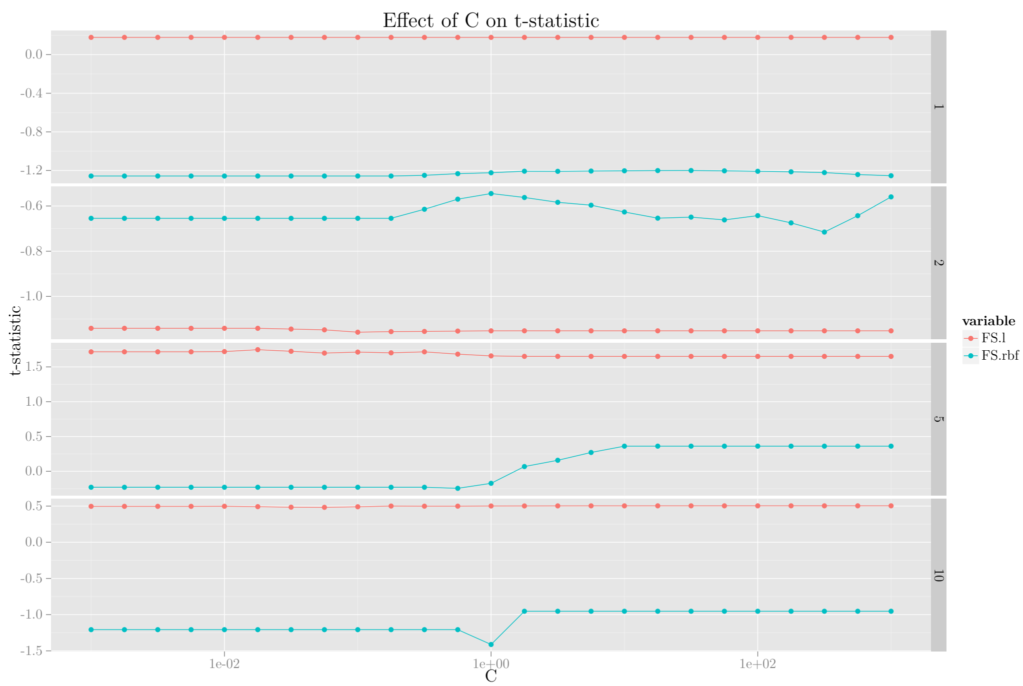

Figure 4.1: FS.l: FS with a linear kernel; FS.rbf: FS with RBF kernel. We vary the dimension of the data: 1, 2, 5, 10.

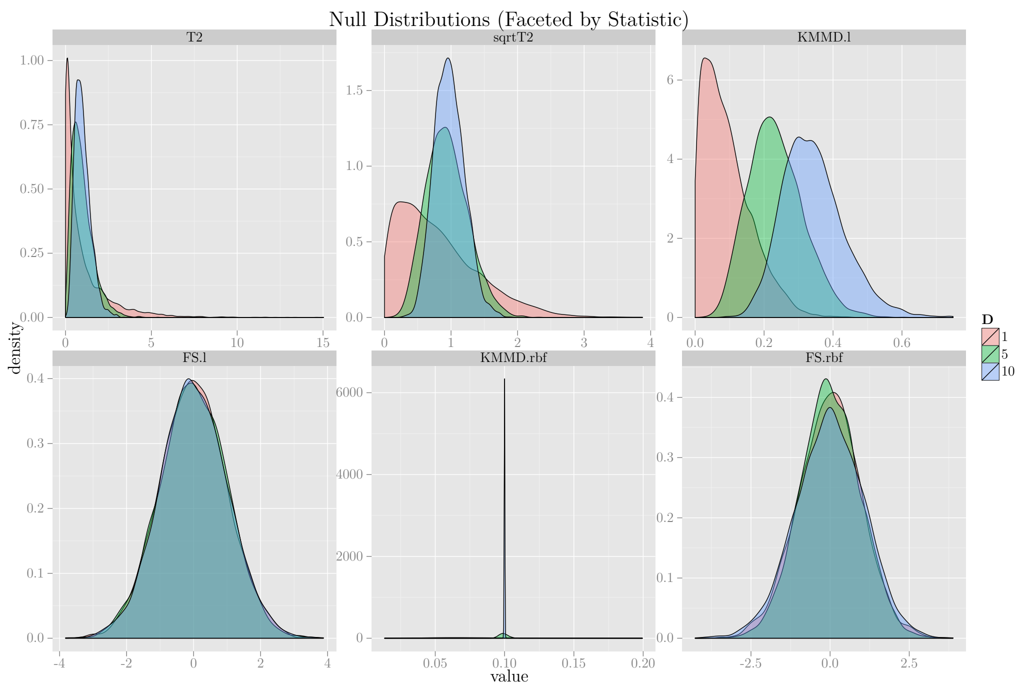

Figure 4.2: T2: Hotelling's T^2-statistic; sqrtT2: $|T|$; KMMD.l: kernel MMD with a linear kernel; FS.l: FS with a linear kernel; KMMD.rbf: kernel MMD with a radial basis function (RBF) kernel; FS.rbf: FS with RBF kernel

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 affords 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 $\mathbb{R}^+$.

4.8 Experiments

4.8.1 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 $\Delta \in \{0, .5, \dots, 1.5\}$ in Figure 4.3. The width parameter in the RBF kernel is fixed at 1.

For FS and MMD, we used 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 (probably the former).

4.8.2 String Data

For a string data comparison, we consider Twitter data and look at the latest 1,000 tweets from Barack Obama (@BarackObama) and Sarah Palin (@SarahPalinUSA) obtained from the **R** package **twitteR** [14]. We pre-process each tweet by removing all hyperlinks and anything that is neither a letter nor a space. Finally, we convert all letters to lowercase. For simplicity, we choose the k-spectrum kernel [24] with $k \in \{1, 2, 3\}$ as our kernels for both the FT and MMD.

For the k-spectrum kernel,

- $\mathcal{X}$ = set of all finite-length sequences from an alphabet $\mathcal{A}$.
- $\phi(x)$ = the number of length k contiguous subsequences (k-mers) in x.

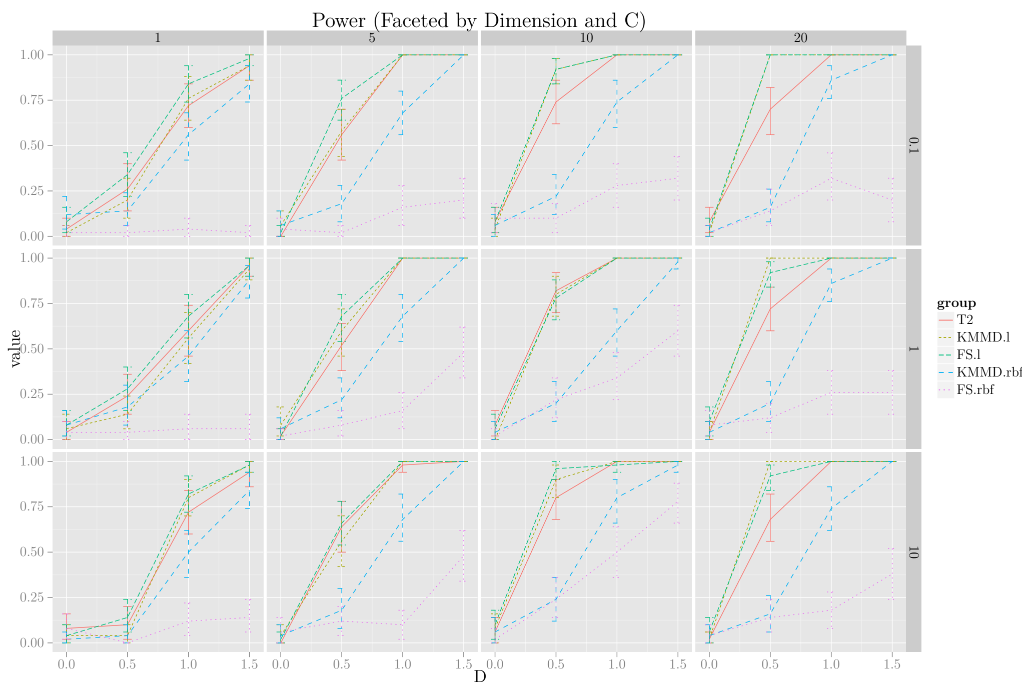

Figure 4.3: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; T2: Hotelling's T^2-statistic; Error bars indicate 95% bootstrap confidence intervals. The tests perform similarly, and the kernel-based tests use a linear kernel.

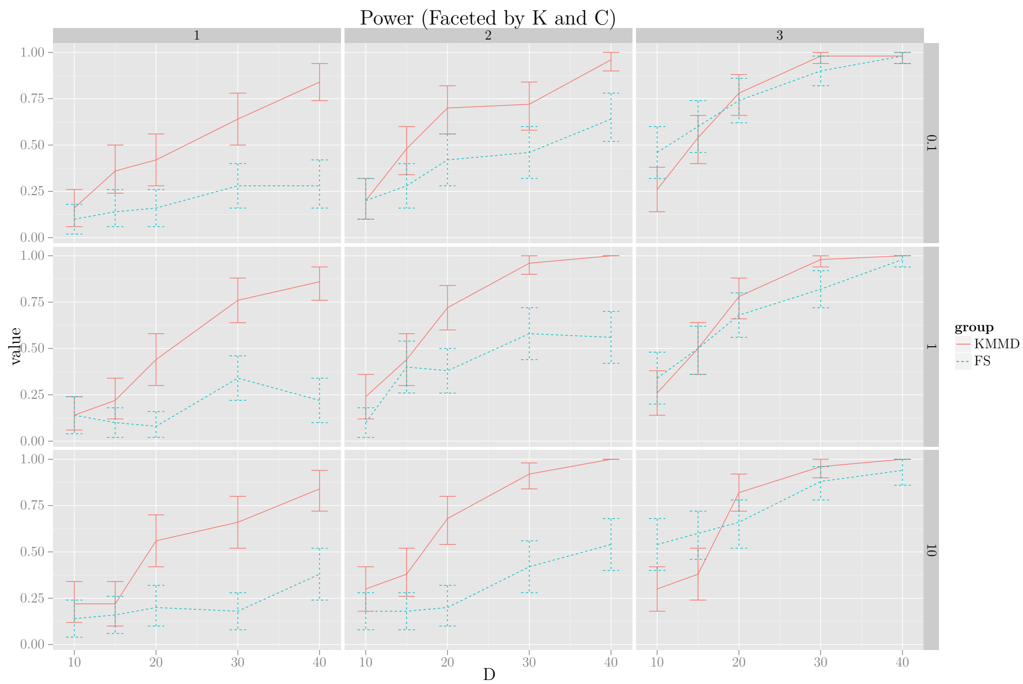

Figure 4.4: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; Error bars indicate 95% bootstrap confidence intervals.

- $\mathcal{H} = \mathbb{N}^{|\mathcal{A}|^k}$.
- $K_k(x, y) = \langle \phi_k(x), \phi_k(y) \rangle$.

Suffix trees allow for efficient kernel calculations, computing $K_k(x, y)$ 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 4.4.

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

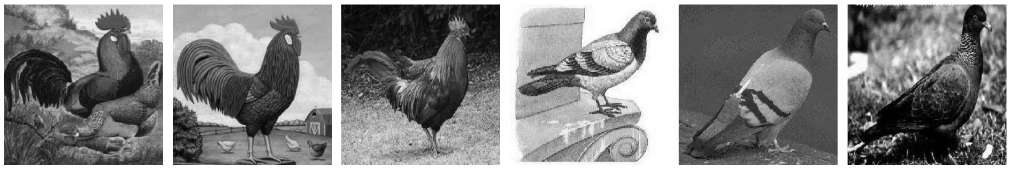

Figure 4.5: Images of roosters and pigeons for use in discrimination test.

see that the advantage largely disappears for $k = 3$.

4.8.3 Image Data

We consider the task of discriminating between images of roosters and pigeons from the Caltech 101 Object Categories dataset [11]. Samples of the birds are in Figure 4.5. We resize images to a common resolution of 300×297 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 center and scale each vector. Power comparisons can be seen in Figure 4.6.

For the polynomial kernel,

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

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

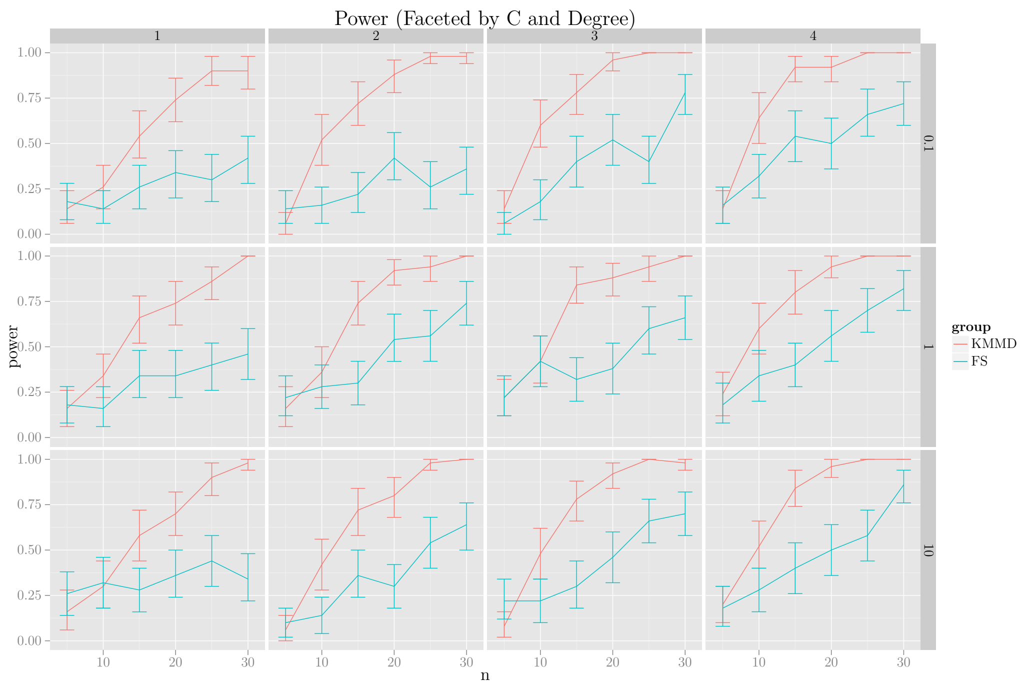

Figure 4.6: p1: linear kernel; p2: inhomogeneous degree 2 polynomial kernel; rbf: radial basis function kernel; Error bars indicate 95% bootstrap confidence intervals.

Chapter 5

Multiple Kernels

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

5.1 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_θ and seek the "best" choice of parameters θ.

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

$$K(x, x') = (K_1 K_2)(x, x') = K_1(x, x')K_2(x, x'),$$

tensor products of kernels,

$$K(x, x') = (K_1 \otimes K_2)(x_1, x_2, x'_1, x'_2) = K_1(x_1, x'_1)K_2(x_2, x'_2),$$

and conic combinations of kernels,

$$K_\theta(x,x') = (\theta_1 K_1 + \dots + \theta_m K_m)(x,x') = \theta_1 K_1(x,x') + \dots + \theta_m K_m(x,x').$$

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

5.2 Multiple Kernel Learning

In a landmark paper, Lanckriet et al. [21] 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. [20] conceived a general ℓ_p-norm approach to MKL, unifying many special cases and further proposed a highly efficient algorithm. This can be seen as generalizing Problem (4.3) of Chapter 4.

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

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

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

The regularized risk minimization problem is the following:

$$\min_{\tilde{w},b,\theta:\theta\succeq 0} \frac{1}{n} \sum_{i=1}^{n} L\left(\sum_{m=1}^{M} \sqrt{\theta_m} \langle \tilde{w}_m, \phi_m(x_i) \rangle_{\mathcal{H}_m} + b, y_i \right) + \frac{\lambda}{2} \sum_{m=1}^{M} \|\tilde{w}_m\|_{\mathcal{H}_m}^2 + \tilde{\mu}\tilde{\Omega}(\theta), \tag{5.1}$$

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

Problem (5.1) 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

$$\min_{w,b,\theta:\theta\succeq 0} \tilde{C}\sum_{i=1}^{n} L\left(\sum_{m=1}^{M}\langle w_m, \phi_m(x_i)\rangle_{\mathcal{H}_m} + b, y_i\right) + \frac{1}{2}\sum_{m=1}^{M}\frac{\|w_m\|_{\mathcal{H}_m}^2}{\theta_m} + \mu\|\theta\|_p^2, \quad (5.2)$$

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

Kloft et al. prove that the Tikhonov-regularized Problem (5.2) with two parameters is in fact equivalent to the following Ivanov-regularized formulation with one regularization parameter, C:

$$\begin{aligned} \underset{w,b,\theta:\theta\succeq 0}{\text{minimize}} \quad & C\sum_{i=1}^{n} L\left(\sum_{m=1}^{M}\langle w_m, \phi_m(x_i)\rangle_{\mathcal{H}_m} + b, y_i\right) + \frac{1}{2}\sum_{m=1}^{M}\frac{\|w_m\|_{\mathcal{H}_m}^2}{\theta_m} \\ \text{subject to} \quad & \|\theta\|_p^2 \le 1. \end{aligned} \quad (5.3)$$

We use Problem (5.3) as implemented in SHOGUN [35].

5.3 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 [36].

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 $N \sim$ Pois(100) and is generated via a Markov chain. The start of the sequence is drawn from the stationary distribution [.25, .25, .25, .25],

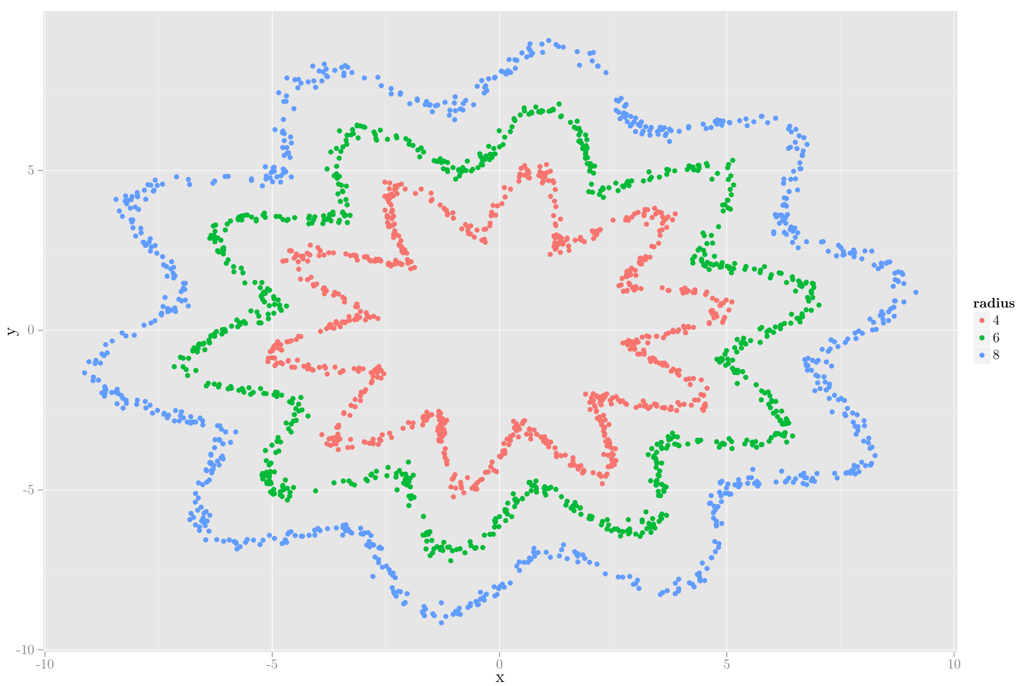

Figure 5.1: Star Distribution: radius 4 versus radius > 4.

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

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

$p_{\mathcal{S}}$ 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.

5.4 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 $K(x,x) = 1$. 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:

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

5.5 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

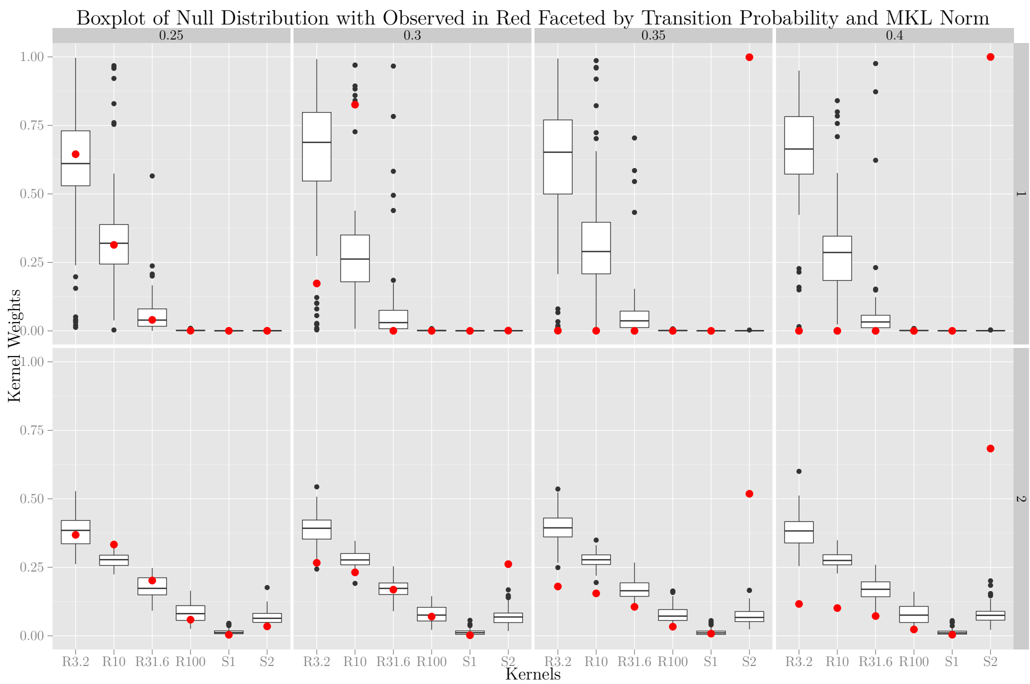

Figure 5.2: The MKL weights in the 1- (upper row) and 2-norm (lower row) cases shift progressively more to the 2-spectrum kernel as the DNA signal is increased.

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 5.2, 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 5.3 we vary the radius of the outer star in $\{4, 7, 10, 13, 16\}$, while fixing

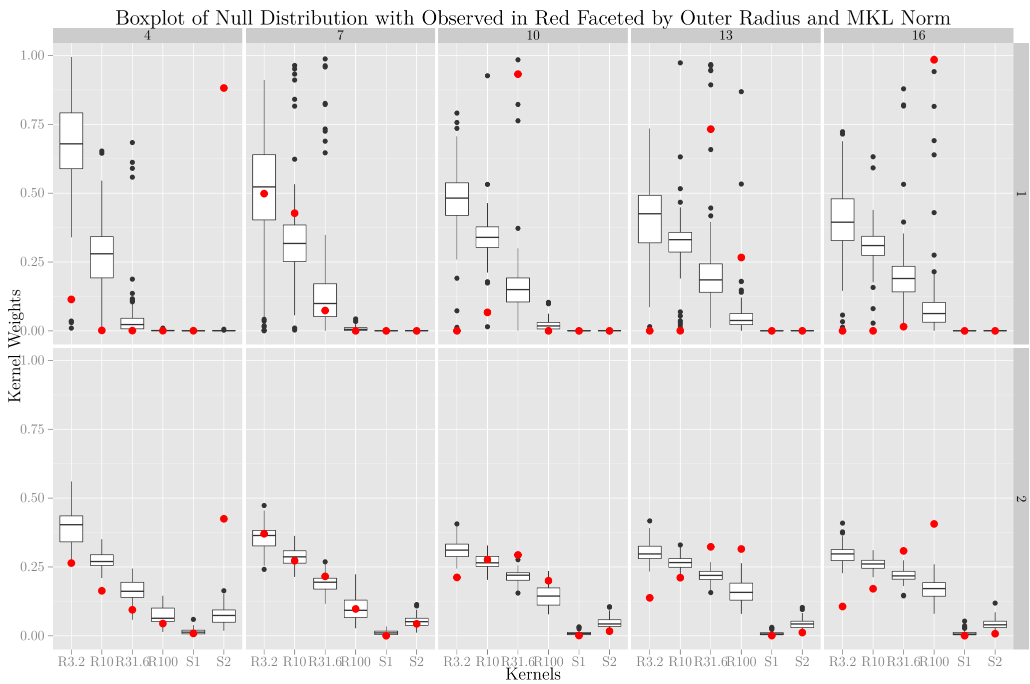

Figure 5.3: The MKL weights in the 1- (upper row) and 2-norm (lower row) cases shift progressively more to the higher-width RBF kernels as we increase the distance between the two stars.

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.

5.6 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 5.4 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,

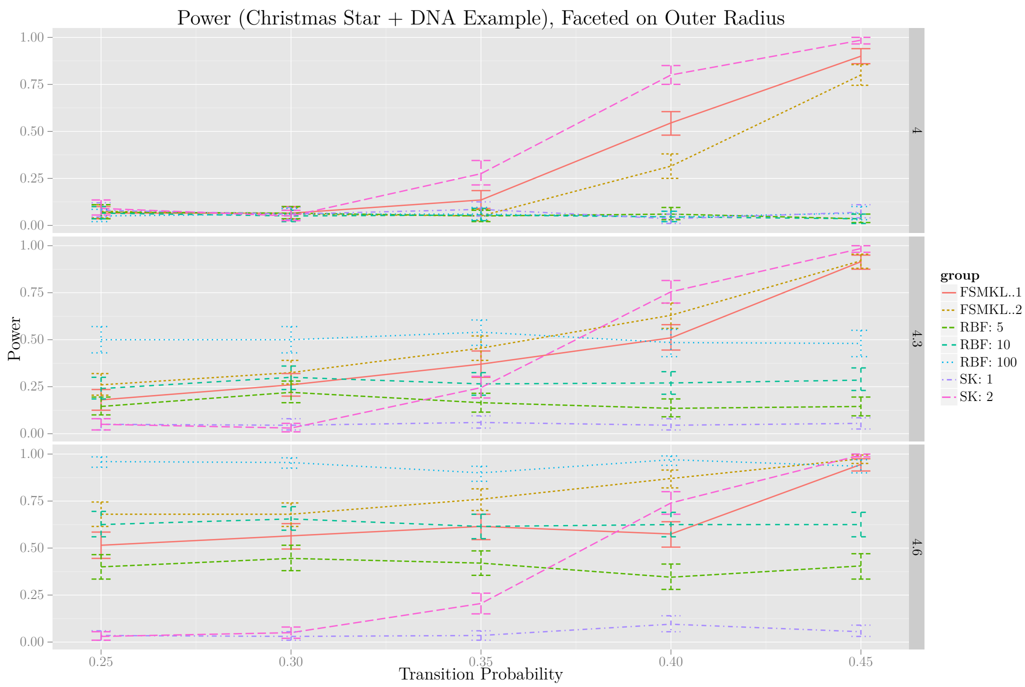

Figure 5.4: The MKL weights in the 1- (upper row) and 2-norm (lower row) cases shift progressively more to the higher-width RBF kernels as we increase the distance between the two stars.

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.

5.7 Null Distribution

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

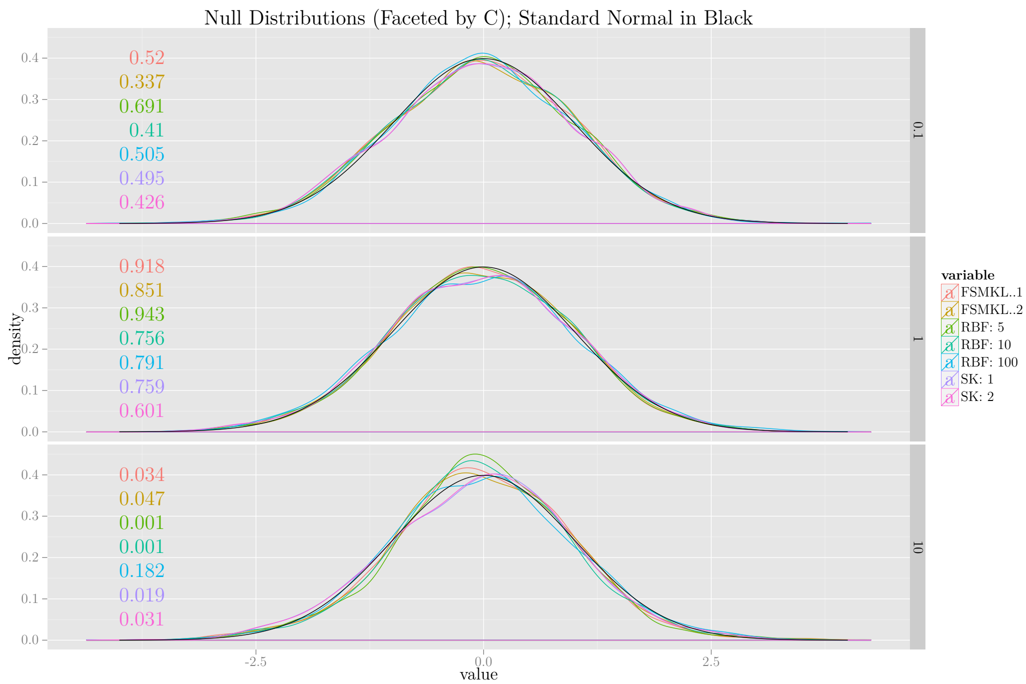

Figure 5.5: Except for $C = 10$, permutation null samples are consistent with the standard normal distribution.

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 5.5. 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.

Appendix A

Auxiliary Results

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

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 $Y' = Y - \mathbb{E}Y$. $\square$

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 [6]:

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

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

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 (1.1) 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,

$$\begin{aligned} \mathbb{E}W'W &= \mathbb{E}[\mathbb{E}[W'W|W]] \\ &= \mathbb{E}[W\mathbb{E}[W'|W]] \\ &= \mathbb{E}[W((1 - \lambda)W + \lambda R)] \quad \text{from (1.1)} \\ &= (1 - \lambda)\mathbb{E}W^2 + \lambda\mathbb{E}WR. \end{aligned} \tag{A.5}$$

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 [6]:

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 [6]:

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 [6]:

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.0.1 Proof of Theorem 1.2

Proof. In order to construct our exchangeable pair, we introduce the ordered pair of random variables (I, J) independent of Π that represents a uniformly at random draw from the set of all non-null transpositions:

$$P(I = i, J = j) = \frac{1}{n(n-1)} \quad i, j \in \{1, \dots, n\}, i \neq j. \tag{B.1}$$

Define the random permutation Π' by

$$\Pi'(i) = \Pi \circ (I, J) = \begin{cases} \Pi(J) & i = I \\ \Pi(I) & i = J \\ \Pi(i) & \text{else.} \end{cases} \tag{B.2}$$

We construct our exchangeable pair by defining

$$W' = \sum_{i=1}^{n} a_{i\Pi'(i)} = W - a_{I\Pi(I)} + a_{I\Pi(J)} - a_{J\Pi(J)} + a_{J\Pi(I)}. \tag{B.3}$$

We now verify the contraction property:

$$\begin{aligned}
\mathbb{E}[W - W'|\Pi] &= \mathbb{E}[a_{I\Pi(I)} - a_{I\Pi(J)} + a_{J\Pi(J)} - a_{J\Pi(I)}|\Pi] \\
&= \frac{2}{n}\sum_{i=1}^{n} a_{i\Pi(i)} - \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)} \\
&= \frac{2}{n}W - \frac{2}{n}\frac{1}{n-1}\left[\sum_{i,j=1}^{n} a_{i\Pi(j)} - \sum_{i}^{n} a_{i\Pi(i)}\right] \\
&= \frac{2}{n}W + \frac{2}{n}\frac{1}{n-1}W - \frac{2}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i\Pi(j)}\right] \\
&= \frac{2}{n}W\left(1 + \frac{1}{n-1}\right) - 0 \\
&= \frac{2}{n-1}W
\end{aligned}$$

This satisfies our contraction property with

$$\lambda = \frac{2}{n-1}. \tag{B.4}$$

To bound the variance component, compute

$$
\begin{aligned}
\mathbb{E}[(W-W')^2|\Pi] &= \mathbb{E}[(a_{I\Pi(I)} - a_{I\Pi(J)} + a_{J\Pi(J)} - a_{J\Pi(I)})^2|\Pi] \\
&= \mathbb{E}[a_{I\Pi(I)}^2 + a_{J\Pi(J)}^2 + a_{I\Pi(J)}^2 + a_{J\Pi(I)}^2 \\
&\quad - 2a_{I\Pi(I)}a_{I\Pi(J)} - 2a_{J\Pi(J)}a_{J\Pi(I)} - 2a_{I\Pi(I)}a_{J\Pi(I)} - 2a_{J\Pi(J)}a_{I\Pi(J)} \\
&\quad + 2a_{I\Pi(I)}a_{J\Pi(J)} + 2a_{I\Pi(J)}a_{J\Pi(I)}|\Pi] \\
&= \frac{2}{n}\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)}^2 \\
&\quad - \frac{4}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{i\Pi(j)} - \frac{4}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{j\Pi(i)} \\
&\quad + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{j\Pi(j)} + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)}a_{j\Pi(i)} \\
&= \frac{2}{n}\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\left(\sum_{i,j=1}^{n} a_{i\Pi(j)}^2 - \sum_{i=1}^{n} a_{i\Pi(i)}^2\right) \\
&\quad - \frac{4}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(a_{i\Pi(i)}\sum_{j=1}^{n}\left(a_{i\Pi(j)} + a_{j\Pi(i)}\right) - 2a_{i\Pi(i)}^2\right) \\
&\quad + \frac{2}{n}\frac{1}{n-1}\left(\sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right) \\
&= \frac{2}{n}\left(1 - \frac{1}{n-1}\right)\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n} \\
&\quad + \frac{8}{n}\frac{1}{n-1}\sum_{i=1}^{n} a_{i\Pi(i)}^2 \\
&\quad + \frac{2}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right) - \frac{4}{n}\frac{1}{n-1}\sum_{i=1}^{n} a_{i\Pi(i)}^2 \\
&= \frac{2}{n} + \frac{2(n+2)}{n(n-1)}\sum_{i=1}^{n} a_{i\Pi(i)}^2 + \frac{2}{n(n-1)}\sum_{i,j=1,i\neq j}^{n} (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})
\end{aligned}
\tag{B.5}
$$

From (B.5) and corollary A.2,

$$\begin{aligned}
\mathbb{E}[(W-W')^2|\Pi] &= \operatorname{Var}\left(\frac{2(n+2)}{n(n-1)}\sum_{i=1}^n a_{i\Pi(i)}^2 \right.\\
&\quad \left. + \frac{2}{n(n-1)}\sum_{i,j=1,i\neq j}^n (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})\right)\\
&\leq 2\left(\frac{4(n+2)^2}{n^2(n-1)^2}\operatorname{Var}\left(\sum_{i=1}^n a_{i\Pi(i)}^2\right) + \right.\\
&\quad \left. \frac{4}{n^2(n-1)^2}\operatorname{Var}\left(\sum_{i,j=1,i\neq j}^n (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})\right)\right)\\
&\leq \frac{32}{n^2}\operatorname{Var}\left(\sum_{i=1}^n a_{i\Pi(i)}^2\right) + \frac{32}{n^4}\operatorname{Var}\left(\sum_{i,j=1,i\neq j}^n (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})\right)
\end{aligned} \tag{B.6}$$

for $n \geq 2$ since $n-1 \geq n/2$ $\quad \frac{1}{(n-1)^2} \leq \frac{4}{n^2}$ for $n \geq 2$.

First, we address the first term in (B.6):

$$\operatorname{Var}\left(\sum_{i=1}^n a_{i\Pi(i)}^2\right) = \sum_{i=1}^n \operatorname{Var}(a_{i\Pi(i)}^2) + \sum_{i,j=1,i\neq j}^n \operatorname{Cov}(a_{i\Pi(i)}^2, a_{j\Pi(j)}^2),$$

with

$$\begin{aligned}
\sum_{i,j=1,i\neq j}^n \operatorname{Cov}(a_{i\Pi(i)}^2, a_{j\Pi(j)}^2) &= \sum_{i,j=1,i\neq j}^n \left(\frac{1}{n(n-1)}\sum_{k,l=1,k\neq l}^n a_{ik}^2 a_{jl}^2 - \left(\frac{1}{n}\sum_k a_{ik}^2\right)\left(\frac{1}{n}\sum_l a_{jl}^2\right)\right)\\
&= \sum_{i,j=1,i\neq j}^n \left(\frac{1}{n(n-1)}\sum_{k,l=1}^n a_{ik}^2 a_{jl}^2 - \frac{1}{n^2}\sum_k\sum_l a_{ik}^2 a_{jl}^2 - \frac{1}{n(n-1)}\sum_k a_{ik}^2 a_{jk}^2\right)\\
&= \frac{1}{n^2(n-1)}\sum_{i,j=1,i\neq j}^n \sum_{k,l=1}^n a_{ik}^2 a_{jl}^2 - \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^n \sum_k a_{ik}^2 a_{jk}^2\\
&\leq \frac{(n-1)^2}{n^2(n-1)}\\
&\leq \frac{1}{n}
\end{aligned}$$

It will be convenient to express our bound as a multiple of $\sum_{i,j=1}^{n} a_{i,j}^4$, so we establish a lower bound on that quantity. Our scaling is such that $\sum_{i,j=1}^{n} a_{i,j}^2 = n-1$, so if we write $a := [a_{11}^2\ a_{12}^2 \dots a_{nn}^2]^T$ out as a vector, $a^T 1 = n-1$. By Cauchy-Schwarz,

$$\begin{aligned}(n-1)^2 &= (a^T 1)^2 \\ &\leq ||a||_2^2 ||1||_2^2 \\ &= n^2 \sum_{i,j=1}^{n} a_{i,j}^4.\end{aligned}$$

Therefore, $\sum_{i,j=1}^{n} a_{i,j}^4 \geq 1$, so

$$\sum_{i,j=1,i\neq j}^{n} \mathrm{Cov}(a_{i\Pi(i)}^2, a_{j\Pi(j)}^2) \leq \frac{1}{n} \sum_{i,j=1}^{n} a_{i,j}^4. \tag{B.7}$$

For the second term in (B.6) we again apply corollary A.2:

$$\mathrm{Var}\left(\sum_{i,j=1,i\neq j}^{n} (a_{i\Pi(i)} a_{j\Pi(j)} + a_{i\Pi(j)} a_{j\Pi(i)})\right) < 2\,\mathrm{Var}\,(X) + 2\,\mathrm{Var}\,(Y)\,,$$

where $X = \sum_{i,j=1,i\neq j}^{n} a_{i\Pi(i)} a_{j\Pi(j)}$ and $Y = \sum_{i,j=1,i\neq j}^{n} a_{i\Pi(j)} a_{j\Pi(i)}$. We note that

$$X = \sum_{i=1}^{n} a_{i\Pi(i)} \sum_{j=1,j\neq i}^{n} a_{j\Pi(j)} = W^2 - \sum_{i=1}^{n} a_{i\Pi(i)}^2. \tag{B.8}$$

TODO: Finish including proof of Hoeffing's Combinatorial Central Limit Theorem. There's still that one bound that I cannot rederive. Well, I can just cite Stein. $\square$

B.0.2 Proof of Lemma 1.5

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.0.3 Proof of Theorem 1.6

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

$$
f'(w) - wf(w) = h_{z,\alpha}(w) - \Phi(z) \tag{B.9}
$$

for the smoothed indicator

$$h_{z,\alpha}(w) = \begin{cases} 1 & w \leq z \\ 1 + \frac{z-w}{\alpha} & z < w \leq z + \alpha \\ 0 & w > z + \alpha. \end{cases} \tag{B.10}$$

Therefore,

$$\begin{aligned} |P(W \leq z) - \Phi(z)| &= |\mathbb{E}[(f'(W) - Wf(W))]| \\ &= \left| \mathbb{E}\left[f'(W) - \frac{(W' - W)(f(W') - f(W))}{2\lambda} + f(W)R \right] \right| \\ &= \left| \mathbb{E}\left[f'(W)\left(1 - \frac{(W'-W)^2}{2\lambda}\right) \right. \right. \\ &\quad \left. \left. + \frac{f'(W)(W'-W)^2 - (f(W') - f(W))(W' - W)}{2\lambda} + f(W)R \right] \right| \\ &:= |\mathbb{E}[J_1 + J_2 + J_3]| \\ &\leq |\mathbb{E}J_1| + |\mathbb{E}J_2| + |\mathbb{E}J_3|. \end{aligned} \tag{B.11}$$

It is known from Chen and Shao (2004) that for all $w \in \mathbb{R}, 0 \leq f(w) \leq 1$ and $|f'(w)| \leq 1$. Then

$$|\mathbb{E}J_3| \leq \mathbb{E}|J_3| = \mathbb{E}|f(W)R| \leq \mathbb{E}|R| \tag{B.12}$$

and

$$
\begin{aligned}
|\mathbb{E}J_1| &= \left|\mathbb{E}\left[f'(W)\left(1-\frac{(W'-W)^2}{2\lambda}\right)\right]\right| \\
&\leq \mathbb{E}\left[\left|f'(W)\left(1-\frac{(W'-W)^2}{2\lambda}\right)\right|\right] \\
&\leq \mathbb{E}\left[\left|\left(1-\frac{(W'-W)^2}{2\lambda}\right)\right|\right] \\
&= \frac{1}{2\lambda}\mathbb{E}[|2\lambda-\mathbb{E}[(W'-W)^2|W]|] \\
&= \frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}W^2-\mathbb{E}WR)-\mathbb{E}[(W'-W)^2|W]+2\lambda(1-\mathbb{E}W^2+\mathbb{E}WR)|] \\
&\leq \frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}W^2-\mathbb{E}WR)-\mathbb{E}[(W'-W)^2|W]|]+\mathbb{E}|1-\mathbb{E}W^2+\mathbb{E}WR|
\end{aligned}
\tag{B.13}
$$

Note that

$$
\mathbb{E}[\mathbb{E}[(W'-W)^2|W]]=\mathbb{E}\Delta^2=2\lambda(\mathbb{E}W^2-\mathbb{E}WR), \tag{B.14}
$$

so

$$
\frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}W^2-\mathbb{E}WR)-\mathbb{E}[(W'-W)^2|W]|]\leq\frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])}. \tag{B.15}
$$

Combining with (B.13),

$$
\begin{aligned}
|\mathbb{E}J_1| &\leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])}+\mathbb{E}|1-\mathbb{E}W^2+\mathbb{E}WR| \\
&\leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])}+\mathbb{E}|1-\mathbb{E}W^2|+\mathbb{E}|WR|
\end{aligned}
\tag{B.16}
$$

Lastly, we bound the second term,

$$\begin{aligned}
J_2 &= \frac{1}{2\lambda}(W' - W)\int_W^{W'} (f'(W) - f'(t))dt \\
&= \frac{1}{2\lambda}(W' - W)\int_W^{W'}\int_t^W f''(u)dudt \\
&= \frac{1}{2\lambda}(W' - W)\int_W^{W'} (W' - u)f''(u)du.
\end{aligned} \tag{B.17}$$

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,

$$
\begin{aligned}
|\mathbb{E}J_2| &= \left|\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)\int_W^{W'}(W'-u)f''(u)du\right]\right| \\
&= \left|\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)\int_W^{W'}\left(\frac{W+W'}{2}-u\right)f''(u)du\right]\right| \\
&\leq \left|\mathbb{E}\left[||f''||\frac{1}{2\lambda}|W'-W|\int_{\min(W,W')}^{\max(W,W')}\left|\frac{W+W'}{2}-u\right|du\right]\right| \qquad \text{(B.18)} \\
&= \left|\mathbb{E}\left[||f''||\frac{1}{2\lambda}\frac{|W'-W|^3}{4}\right]\right| \\
&\leq \frac{\mathbb{E}|W'-W|^3}{4\alpha\lambda},
\end{aligned}
$$

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

$$
\begin{aligned}
P(W \leq z) &\leq \mathbb{E}h_{z,\alpha}(W) \\
&\leq Nh_{z,\alpha} + \frac{\mathbb{E}|W'-W|^3}{4\alpha\lambda} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} \\
&\quad + |1-\mathbb{E}W^2| + \mathbb{E}|WR| + \mathbb{E}|R| \qquad \text{(B.19)} \\
&\leq \Phi(z) + \frac{\alpha}{\sqrt{2\pi}} + \frac{\mathbb{E}|W'-W|^3}{4\alpha\lambda} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} \\
&\quad + |\mathbb{E}W^2-1| + \mathbb{E}|WR| + \mathbb{E}|R|
\end{aligned}
$$

The minimizer of the expression is

$$
\alpha = \frac{(2\pi)^{1/4}}{2}\sqrt{\frac{\mathbb{E}|W'-W|^3}{\lambda}}. \qquad \text{(B.20)}
$$

Plugging this in, we get the upper bound

$$P(W \leq z) - \Phi(z) \leq (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|W' - W|^3}{\lambda}} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} \\ + |\mathbb{E}W^2 - 1| + \mathbb{E}|WR| + \mathbb{E}|R| \tag{B.21}$$

Proving the corresponding lower bound in a similar manner completes the proof of the theorem. □

B.0.4 Proof of Theorem 1.7

Proof. Now we bound $|\mathbb{E}J_2|$ with $\delta \geq 0$. From (B.11),

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

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

$$\begin{aligned}
\mathbb{E}J_{22}\mathbf{1}_{\{\delta\leq W'-W\leq 0\}} &= \mathbb{E}\left[(W' - W)\mathbf{1}_{\{\delta\leq W'-W\leq 0\}} \int_0^{W'-W} (\mathbf{1}_{\{W\leq z\}} - \mathbf{1}_{\{W+t\leq z\}})dt\right] \\
&= \mathbb{E}\left[(W - W')\mathbf{1}_{\{\delta\leq W'-W\leq 0\}} \int_{W'-W}^{0} (\mathbf{1}_{\{z\leq W\leq z-t\}})dt\right] \\
&\leq \mathbb{E}\left[(W - W')^2\mathbf{1}_{\{\delta\leq W'-W\leq 0\}}\mathbf{1}_{\{z-\delta\leq W\leq z\}}\right] \\
&\leq 3\delta\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) \quad \text{by (1.5)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}J_{22}\mathbf{1}_{\{0<W'-W\leq\delta\}} &= \mathbb{E}\left[(W' - W)\mathbf{1}_{\{0<W'-W\leq\delta\}} \int_0^{W'-W} (\mathbf{1}_{\{W\leq z\}} - \mathbf{1}_{\{W+t\leq z\}})dt\right] \\
&= \mathbb{E}\left[(W' - W)\mathbf{1}_{\{0<W'-W\leq\delta\}} \int_0^{W'-W} \mathbf{1}_{\{z-t<W\leq z\}}dt\right] \\
&\leq \mathbb{E}\left[(W' - W)^2\mathbf{1}_{\{0<W'-W\leq\delta\}}\mathbf{1}_{\{z-\delta\leq W\leq z\}}\right] \\
&\leq 3\delta\lambda(\sqrt{\mathbb{E}W^2} + \mathbb{E}|R|) \quad \text{by (1.5)}
\end{aligned}$$

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.11), noting that J_1 and J_3 stay the same. $\square$

Appendix C

Rate of Convergence Bounds

C.0.5 Proof of Proposition 2.4

Proof.

$$
\begin{aligned}
\mathbb{E}T_\Pi^2 &= \frac{N-1}{N}\mathbb{E}\left[\left(\frac{q_\Pi}{d_\Pi}\right)^2\right] && \text{(C.1)}\\
&= \frac{N-1}{N}\mathbb{E}\left[\frac{4N^2\bar{u}_{2,\Pi}^2}{2N-2N\bar{u}_{2,\Pi}^2}\right] \qquad \text{from (2.8)}\\
&= 2(N-1)\mathbb{E}\left[\frac{\bar{u}_{2,\Pi}^2}{1-\bar{u}_{2,\Pi}^2}\right]\\
&= 2(N-1)\mathbb{E}g(\bar{u}_{2,\Pi}), && \text{(C.2)}
\end{aligned}
$$

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

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

$$
\mathbb{E}\bar{u}_{2,\Pi} = \frac{1}{N}\mathbb{E}\left[\sum_{i=N+1}^{2N} u_{\Pi(i)}\right] = \frac{1}{N}\sum_{i=N+1}^{2N}\mathbb{E}u_{\Pi(i)} = \frac{1}{N}\sum_{i=N+1}^{2N}\frac{1}{2N}\sum_{j=1}^{2N}u_j = 0
$$

Under independence, $\mathrm{Var}(\bar{u}_{2,\Pi})$ 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 $\mathrm{Var}(u_{\Pi(i)}) = 1$. Under independence and with $i \neq j$,

$\mathrm{Var}(u_{\Pi(i)} + u_{\Pi(j)}) = 2$. Summing only 2 (out of $2N$) values under permutation dependence, $\mathrm{Var}(u_{\Pi(i)} + u_{\Pi(j)}) = 2 - \frac{2}{2N-1}$.

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

$$\begin{aligned}
\mathrm{Var}(\bar{u}_{2,\Pi}) &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=N+1}^{2N} u_{\Pi(i)}\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=N+1}^{2N} u_{\Pi(i)}^2 + \sum_{i=N+1}^{2N}\sum_{j=N+1,j\neq i}^{2N} u_{\Pi(i)}u_{\Pi(j)}\right] \\
&= \frac{1}{N^2}\sum_{i=N+1}^{2N}\frac{1}{2N}\sum_{j=1}^{2N} u_j^2 + \frac{1}{N^2}\sum_{i=N+1}^{2N}\sum_{j=N+1,j\neq i}^{2N}\mathbb{E}[u_{\Pi(i)}u_{\Pi(j)}] \\
&= \frac{1}{N} + \frac{1}{N^2}\sum_{i=N+1}^{2N}\sum_{j=N+1,j\neq i}^{2N}\frac{1}{2N}\frac{1}{2N-1}\sum_{k=1}^{2N}\sum_{l=1,l\neq k}^{2N} u_k u_l \\
&= \frac{1}{N} + \frac{1}{N^2}\sum_{i=N+1}^{2N}\sum_{j=N+1,j\neq i}^{2N}\frac{1}{2N}\frac{1}{2N-1}\left(\left(\sum_{k=1}^{2N} u_k\right)^2 - \sum_{k=1}^{2N} u_k^2\right) \\
&= \frac{1}{N} + \frac{1}{N^2}\sum_{i=N+1}^{2N}\sum_{j=N+1,j\neq i}^{2N}\frac{1}{2N}\frac{1}{2N-1}\left(0^2 - 2N\right) \\
&= \frac{1}{N} + \frac{1}{N}(N^2 - N)\left(-\frac{1}{2N-1}\right) \\
&= \frac{2N-1}{N(2N-1)} + \frac{1-N}{N(2N-1)} \\
&= \frac{1}{2N-1}
\end{aligned}$$

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 $g(\bar{u}_{2,\Pi}) = \frac{\bar{u}_{2,\Pi}^2}{1-\bar{u}_{2,\Pi}^2}$ around $\mathbb{E}[\bar{u}_{2,\Pi}] = 0$:

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

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

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

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

Therefore,

$$\begin{aligned}
\left|\frac{\mathbb{E}T_\Pi^2}{2(N-1)} - \mathbb{E}\bar{u}_{2,\Pi}^2\right| &= \left|\frac{\mathbb{E}T_\Pi^2}{2(N-1)} - \frac{1}{2N-1}\right| \\
&\leq \mathbb{E}|R_3(\bar{u}_{2,\Pi})| \\
&= \mathbb{E}\left|\frac{24(5\xi_L^4 + 10\xi_L^2 + 1)}{4!\,(\xi_L - 1)^5}\bar{u}_{2,\Pi}^4\right| \\
&\leq \mathbb{E}\left|\frac{24(5\bar{u}_{2,\Pi}^4 + 10\bar{u}_{2,\Pi}^2 + 1)}{4!\,(\bar{u}_{2,\Pi} - 1)^5}\bar{u}_{2,\Pi}^4\right| \\
&\leq \frac{5B^4 + 10B^2 + 1}{|B-1|^5}\mathbb{E}\bar{u}_{2,\Pi}^4 \\
&\leq \frac{5B^4 + 10B^2 + 1}{|B-1|^5} f_{c_1}(4)N^{-2} \qquad \text{by (2.4)} \\
&:= c_1 N^{-2}
\end{aligned}$$

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

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.0.6 Proof of Proposition 2.3

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

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

From (2.11),

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

where

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

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

The $\mathbb{E}[u_{\Pi(J)} - u_{\Pi(I)}|\Pi=\pi]$ term is common to applications of Stein's method of exchangeable pairs. However, there is a complication in the d_Π 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

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

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

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

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

where

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

By Taylor's theorem,

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

with

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

Re-arranging, we get

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

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

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

with

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

Squaring, applying the bound, and re-arranging yields

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

Now we combine bounds to get

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{u}_{2,\Pi}^2}\right)^2\right]-\left(\mathbb{E}\left[\frac{1}{1-\bar{u}_{2,\Pi}^2}\right]\right)^2\right| \\
&=\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{u}_{2,\Pi}^2}\right)^2\right]-\left(\mathbb{E}\left[\frac{1}{1-\bar{u}_{2,\Pi}^2}\right]\right)^2+\frac{1}{(2N-1)^2}-\frac{1}{(2N-1)^2}\right| \\
&\leq\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{u}_{2,\Pi}^2}\right)^2\right]-\left(\mathbb{E}\left[\frac{1}{1-\bar{u}_{2,\Pi}^2}\right]\right)^2+\frac{1}{(2N-1)^2}\right|+\left|\frac{1}{(2N-1)^2}\right| \\
&\leq\left|\mathbb{E}\left[\left(\frac{1}{1-\bar{u}_{2,\Pi}^2}\right)^2\right]-1-\frac{2}{2N-1}-\left(\left(\mathbb{E}\left[\frac{1}{1-\bar{u}_{2,\Pi}^2}\right]\right)^2-\left(1+\frac{1}{2N-1}\right)^2\right)\right|+\left|\frac{1}{(2N-1)^2}\right| \\
&\leq c_4N^{-2}+2\left(1+\frac{1}{2N-1}\right)c_5N^{-2}+c_5^2N^{-4}+\left|\frac{1}{(2N-1)^2}\right| \\
&\leq (c_4+3c_5+c_5^2+\frac{1}{4})N^{-2} \\
&:= c_6N^{-2}
\end{aligned}
$$

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

$$
\operatorname{Var}(X)\leq 8(f_{c_2}(2))^2N^{-2}\operatorname{Var}(V)+8u_\Delta^4\frac{c_6}{4}N^{-4}
$$

with

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

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

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

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

$$\mathbb{E}[u_{\pi(I)}u_{\pi(J)}] = \mathbb{E}[u_{\pi(K)}u_{\pi(L)}] = \frac{1}{2N}\frac{1}{2N-1}\sum_{I=1}^{2N} u_I \sum_{J=1,J\neq I}^{2N} u_J = -\frac{1}{2N-1}$$

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

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

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

$$\begin{aligned}
&\mathbb{E}[u_{\pi(I)}u_{\pi(J)}u_{\pi(K)}u_{\pi(L)}] \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\sum_{I=1}^{2N}\sum_{J=1,J\neq I}^{2N}\sum_{K=1,K\neq I,J}^{2N}\sum_{L=1,L\neq I,J,K}^{2N} u_I u_J u_K u_L \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\sum_{I=1}^{2N} u_I \sum_{J=1,J\neq I}^{2N} u_J \sum_{K=1,K\neq I,J}^{2N} u_K(-u_I - u_J - u_K) \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\sum_{I=1}^{2N} u_I \sum_{J=1,J\neq I}^{2N} u_J((-u_I - u_J)(-u_I - u_J) + (u_I^2 + u_J^2 - 2N)) \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\sum_{I=1}^{2N} u_I \sum_{J=1,J\neq I}^{2N} u_J(2u_I^2 - 2N + 2u_J^2 + 2u_I u_J) \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\sum_{I=1}^{2N} u_I \left((2u_I^2 - 2N)(-u_I) + 2\sum_{J=1,J\neq I}^{2N} u_J^3 + 2u_I(2N - u_I^2)\right) \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\sum_{I=1}^{2N} u_I \left(-4u_I^3 + 6Nu_I + 2\left(\sum_{J=1}^{2N} u_J^3 - u_I^3\right)\right) \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\frac{1}{2N-3}\left(-6\sum_{I=1}^{2N} u_I^4 + 12N^2\right)
\end{aligned}$$

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

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

$$\begin{aligned}
\mathbb{E}[u_{\pi(I)}^2 u_{\pi(J)}^2] &= \frac{1}{2N}\frac{1}{2N-1}\sum_{I=1}^{2N}\sum_{J=1,J\neq I}^{2N} u_I^2 u_J^2 \\
&= \frac{1}{2N}\frac{1}{2N-1}\sum_{I=1}^{2N} u_I^2(2N-u_I^2) \\
&= \frac{2N}{2N-1} - \frac{1}{2N}\frac{1}{2N-1}\sum_{I=1}^{2N} u_I^4
\end{aligned}$$

for N^2 terms in the sum.

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

$$\begin{aligned}
\mathbb{E}[u_{\pi(I)}^2 u_{\pi(J)} u_{\pi(K)}] &= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\sum_{I=1}^{2N}\sum_{J=1,J\neq I}^{2N}\sum_{K=1,K\neq I,J}^{2N} u_I^2 u_J u_K \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\sum_{I=1}^{2N}\sum_{J=1,J\neq I}^{2N} u_I^2 u_J(0-u_I-u_J) \\
&= -\frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\left(\sum_{I=1}^{2N} u_I^3 \sum_{J=1,J\neq I}^{2N} u_J + \sum_{I=1}^{2N} u_I^2 \sum_{J=1,J\neq I}^{2N} u_J^2\right) \\
&= -\frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\left(\sum_{I=1}^{2N} -u_I^4 + \sum_{I=1}^{2N} u_I^2(2N-u_I^2)\right) \\
&= \frac{1}{2N}\frac{1}{2N-1}\frac{1}{2N-2}\left(2\sum_{I=1}^{2N} u_I^4 - 4N^2\right)
\end{aligned}$$

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

Putting it all together, we have

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

Therefore,

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

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,

$$
\begin{aligned}
\operatorname{Var}(Y) &= \operatorname{Var}\left(\mathbb{E}\left[\left(T'_\Pi \frac{d_\Pi - d'_\Pi}{d_\Pi}\right)^2 \middle| \Pi = \pi\right]\right) \\
&\leq \mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right)^4\right] \\
&\leq \sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^8\right]\mathbb{E}[(d_\Pi - d'_\Pi)^8]} \\
&\leq \sqrt{f_{c_6}(8)N^{-8/2} f_{c_4}(8)N^{-8}} \text{ from (2.10), (2.7)} \\
&= \sqrt{f_{c_6}(8) f_{c_4}(8)}N^{-6} \\
&:= c_7 N^{-6}
\end{aligned}
$$

And to the third,

$$
\begin{aligned}
\operatorname{Var}(Z) &= 4\operatorname{Var}\left(\mathbb{E}\left[\left(\frac{2u_{\Pi(J)} - 2u_{\Pi(I)}}{d_\Pi} T'_\Pi \frac{d_\Pi - d'_\Pi}{d_\Pi}\right) \middle| \Pi = \pi\right]\right) \\
&\leq 16u_\Delta^2 \mathbb{E}\left[\left(\frac{1}{d_\Pi}\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right)^2\right] \\
&\leq 16u_\Delta^2 f_{c_2}(2)N^{-2/2}\sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^4\right]\mathbb{E}[(d_\Pi - d'_\Pi)^4]} \text{ from (2.5)} \\
&\leq 16u_\Delta^2 f_{c_2}(2)N^{-1}\sqrt{f_{c_6}(4)N^{-4/2} f_{c_4}(4)N^{-4}} \text{ from (2.10), (2.7)} \\
&\leq 16u_\Delta^2 f_{c_2}(2)(f_{c_6}(4))^{-1/2}(f_{c_4}(4))^{-1/2}N^{-4} \\
&:= c_8 N^{-4}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi - T_\Pi)^2|T_\Pi])} \\
&= N\sqrt{(\mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z))} \\
&\le N\sqrt{8(f_{c_2}(2))^2\left(20 + 16\left(\sum_{i=1}^{2N} u_i^4\right)N^{-2}\right)N^{-4} + 8u_\Delta^4\frac{c_6}{4}N^{-4} + c_7N^{-6} + c_8N^{-4}} \\
&:= N^{-1}c_3\sqrt{20 + 16\frac{\sum_{i=1}^{2N} u_i^4}{N^2}}
\end{aligned}
$$

□

C.0.7 Proof of Proposition 2.2

Proof. The strategy is to break apart the remainder term from the main piece. From (2.11),

$$
\begin{aligned}
\mathbb{E}|T'_\Pi - T_\Pi|^3 &= \left(\frac{N-1}{N}\right)^{3/2}\mathbb{E}\left[d_\Pi^{-3}\left|2u_{\Pi(J)} - 2u_{\Pi(I)} + q'_\Pi\frac{d_\Pi - d'_\Pi}{d'_\Pi}\right|^3\right] \\
&\le 8\left(8u_\Delta^3\mathbb{E}[d_\Pi^{-3}] + \sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^6\right]\mathbb{E}[(d_\Pi - d'_\Pi)^6]}\right) \\
&\le 64u_\Delta^3 f_{c_2}(3)N^{-3/2} + 8\sqrt{f_{c_6}(6)N^{-6/2}f_{c_4}(6)N^{-6}} \text{ from } (2.5), (2.10), (2.7) \\
&\le \frac{c_9^2}{2}N^{-3/2}
\end{aligned}
$$

Therefore,

$$
(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi - T_\Pi|^3}{\lambda}} \le (2\pi)^{-1/4}c_9N^{-1/4}.
$$

□

C.0.8 Proof of Proposition 2.6

Proof.

$$
\begin{aligned}
\mathbb{E}|R| &= \mathbb{E}\left|\binom{N}{2}\sqrt{\frac{N-1}{N}}\frac{1}{d_\Pi}\mathbb{E}\left[q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\middle|T_\Pi\right]\right| \\
&\leq \frac{N}{2}\mathbb{E}\left|\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right| \\
&\leq \frac{N}{2}\sqrt{\mathbb{E}\left|\frac{q'_\Pi}{d_\Pi d'_\Pi}\right|^2\mathbb{E}[d_\Pi - d'_\Pi]^2} \\
&\leq \frac{N}{2}\sqrt{f_{c_6}(2)N^{-2/2}f_{c_4}(2)N^{-2}} \text{ from } (2.10), (2.7) \\
&= \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}
\end{aligned}
$$

□

C.0.9 Proof of Proposition 2.5

Proof.

$$
\begin{aligned}
\mathbb{E}|T_\Pi R| &= \mathbb{E}\left|T_\Pi\binom{N}{2}\sqrt{\frac{N-1}{N}}\frac{1}{d_\Pi}\mathbb{E}\left[q'_\Pi\frac{(d_\Pi - d'_\Pi)}{d'_\Pi}\middle|T_\Pi\right]\right| \\
&\leq \frac{N}{2}\mathbb{E}\left|T_\Pi\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi - d'_\Pi)\right| \\
&\leq \frac{N}{2}\sqrt{\mathbb{E}T_\Pi^2\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^2(d_\Pi - d'_\Pi)^2\right]} \\
&\leq \frac{N}{2}\sqrt{\mathbb{E}T_\Pi^2\sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^4\right]\mathbb{E}[(d_\Pi - d'_\Pi)^4]}} \\
&\leq \frac{N}{2}\sqrt{\mathbb{E}T_\Pi^2\sqrt{f_{c_6}(4)N^{-4/2}f_{c_4}(4)N^{-4}}} \text{ from } (2.10), (2.7) \\
&= \frac{N^{-1/2}}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{\mathbb{E}T_\Pi^2} \\
&\leq \frac{1}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{2+2c_1}N^{-1/2}
\end{aligned}
$$

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

References

[1] A Aizerman, Emmanuel M Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[2] V. Bentkus and F. Götze. The berry-esseen bound for student's statistic. *The Annals of Probability*, 24(1):491–503, 1996.

[3] Erwin Bolthausen. An estimate of the remainder in a combinatorial central limit theorem. *Probability Theory and Related Fields*, 66(3):379–386, 1984.

[4] 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.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] L.H.Y. Chen, L. Goldstein, and Q.M. Shao. *Normal Approximation by Stein's Method*. Springer Verlag, 2010.

[7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[8] H.A. David. The beginnings of randomization tests. *The American Statistician*, 62(1):70–72, 2008.

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

[10] P. Diaconis and E. Lehmann. Comment. *Journal of the American Statistical Association*, 103(481):16–19, 2008.

[11] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

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

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

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

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

[16] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[17] 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.

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

[19] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[20] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 999999:953–997, 2011.

[21] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[22] E.L. Lehmann. *Elements of large-sample theory*. Springer Verlag, 1999.

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] John C Platt. Fast training of support vector machines using sequential minimal optimization. 1999.

[30] 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.

[31] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. the MIT Press, 2002.

[32] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel methods in computational biology*. MIT press, 2004.

[33] R.J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

[34] Q.M. Shao. An explicit berry-esseen bound for student's t-statistic via stein's method. *Stein's Method and Applications (AD Barbour and LHY Chen eds). Lecture Notes Series, Institute for Mathematical Sciences, NUS*, 5:143–155, 2005.

[35] Sören Sonnenburg, Gunnar Rätsch, Sebastian Henschel, Christian Widmer, Jonas Behr, Alexander Zien, Fabio de Bona, Alexander Binder, Christian Gehl, and Vojtěch Franc. The shogun machine learning toolbox. *The Journal of Machine Learning Research*, 11:1799–1802, 2010.

[36] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

[37] C. Stein. Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7, 1986.

[38] G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In *SANTA FE INSTITUTE STUDIES IN THE SCIENCES OF COMPLEXITY-PROCEEDINGS VOLUME-*, volume 12, pages 95–95. ADDISON-WESLEY PUBLISHING CO, 1992.

[39] 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.

[40] SL Zabell. On Student's 1908 Article The Probable Error of a Mean. *Journal of the American Statistical Association*, 103(481):1–7, 2008.