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 for the randomization distribution.

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. We reproduce Stein's proof of the Hoeffding combinatorial central limit theorem (HCCLT) with explicit calculation of various constants. 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 an r.v. in the exchangeable pair.

1.2 Hoeffding combinatorial CLT

Theorem 1.1. *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 $n-1$:*

$$\sum_{j=1}^{n} a_{ij} = 0 \tag{1.1}$$

$$\sum_{i=1}^{n} a_{ij} = 0 \tag{1.2}$$

$$\sum_{i=1,j=1}^{n} a_{ij}^2 = n-1 \tag{1.3}$$

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

$$P(\Pi = \pi) = \frac{1}{n!}. \tag{1.4}$$

Define

$$W = \sum_{i=1}^{n} a_{i\Pi(i)} \tag{1.5}$$

to be the sum of a random diagonal. Then

$$|P(W \le w) - \Phi(w)| \le \frac{C}{\sqrt{n}} \left[\sqrt{\sum_{i,j=1}^{n} a_{ij}^4} + \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^3} \right]. \tag{1.6}$$

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:

$$P(I = i, J = j) = \frac{1}{n(n-1)} \quad i, j \in \{1,\ldots,n\}, i \neq j. \tag{1.7}$$

Define the random permutation Π' by

$$\Pi'(i) = \Pi \circ (I, J) = \begin{cases} \Pi(J) & i = I \\ \Pi(I) & i = J \\ \Pi(i) & \text{else.} \end{cases} \tag{1.8}$$

We construct our exchangeable pair by defining

$$W' = \sum_{i=1}^{n} a_{i\Pi'(i)} = W - a_{I\Pi(I)} + a_{I\Pi(J)} - a_{J\Pi(J)} + a_{J\Pi(I)}. \tag{1.9}$$

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

$$\lambda = \frac{2}{n-1}. \tag{1.10}$$

To bound the variance component, compute

$$\begin{aligned}
\mathbb{E}[(W-W')^2|\Pi] &= \mathbb{E}[(a_{I\Pi(I)} - a_{I\Pi(J)} + a_{J\Pi(J)} - a_{J\Pi(I)})^2|\Pi] \\
&= \mathbb{E}[a_{I\Pi(I)}^2 + a_{J\Pi(J)}^2 + a_{I\Pi(J)}^2 + a_{J\Pi(I)}^2 \\
&\quad - 2a_{I\Pi(I)}a_{I\Pi(J)} - 2a_{J\Pi(J)}a_{J\Pi(I)} - 2a_{I\Pi(I)}a_{J\Pi(I)} - 2a_{J\Pi(J)}a_{I\Pi(J)} \\
&\quad + 2a_{I\Pi(I)}a_{J\Pi(J)} + 2a_{I\Pi(J)}a_{J\Pi(I)}|\Pi] \\
&= \frac{2}{n}\sum_{i=1}^n a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^n a_{i\Pi(j)}^2 \\
&\quad - \frac{4}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^n a_{i\Pi(i)}a_{i\Pi(j)} - \frac{4}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^n a_{i\Pi(i)}a_{j\Pi(i)} \\
&\quad + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^n a_{i\Pi(i)}a_{j\Pi(j)} + \frac{2}{n}\frac{1}{n-1}\sum_{i,j=1,i\neq j}^n a_{i\Pi(j)}a_{j\Pi(i)} \\
&= \frac{2}{n}\sum_{i=1}^n a_{i\Pi(i)}^2 + \frac{2}{n}\frac{1}{n-1}\left(\sum_{i,j=1}^n a_{i\Pi(j)}^2 - \sum_{i=1}^n a_{i\Pi(i)}^2\right) \\
&\quad - \frac{4}{n}\frac{1}{n-1}\sum_{i=1}^n\left(a_{i\Pi(i)}\sum_{j=1}^n\left(a_{i\Pi(j)} + a_{j\Pi(i)}\right) - 2a_{i\Pi(i)}^2\right) \\
&\quad + \frac{2}{n}\frac{1}{n-1}\left(\sum_{i,j=1,i\neq j}^n a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right) \\
&= \frac{2}{n}\left(1 - \frac{1}{n-1}\right)\sum_{i=1}^n a_{i\Pi(i)}^2 + \frac{2}{n} \\
&\quad + \frac{8}{n}\frac{1}{n-1}\sum_{i=1}^n a_{i\Pi(i)}^2 \\
&\quad + \frac{2}{n}\frac{1}{n-1}\sum_{i=1}^n\sum_{j=1}^n\left(a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right) - \frac{4}{n}\frac{1}{n-1}\sum_{i=1}^n a_{i\Pi(i)}^2 \\
&= \frac{2}{n} + \frac{2(n+2)}{n(n-1)}\sum_{i=1}^n a_{i\Pi(i)}^2 + \frac{2}{n(n-1)}\sum_{i,j=1,i\neq j}^n\left(a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)}\right)
\end{aligned} \tag{1.11}$$

Theorem 1.2 (The c_r-inequality)**.** *Let $r > 0$. Suppose that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^r <$*

∞. *Then*

$$\mathbb{E}|X+Y|^r < c_r(\mathbb{E}|X|^r + \mathbb{E}|Y|^r), \tag{1.12}$$

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

Corollary 1.3. *Suppose that* $\mathrm{Var}(X) < \infty$ *and* $\mathrm{Var}(Y) < \infty$. *Then*

$$\mathrm{Var}(X+Y) < 2(\mathrm{Var}(X) + \mathrm{Var}(Y)). \tag{1.13}$$

Proof. This follows immediately by applying Theorem 1.2 to the centered random variables $X' = X - \mathbb{E}[X]$ and $Y' = Y - \mathbb{E}[Y]$. □

From (1.11) and corollary 1.3,

$$\begin{aligned}
\mathbb{E}[(W-W')^2|\Pi] &= \mathrm{Var}\left(\frac{2(n+2)}{n(n-1)}\sum_{i=1}^n a_{i\Pi(i)}^2\right. \\
&\quad \left. + \frac{2}{n(n-1)}\sum_{i,j=1,i\neq j}^n (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})\right) \\
&\leq 2\left(\frac{4(n+2)^2}{n^2(n-1)^2}\,\mathrm{Var}\left(\sum_{i=1}^n a_{i\Pi(i)}^2\right) + \right. \\
&\quad \left. \frac{4}{n^2(n-1)^2}\,\mathrm{Var}\left(\sum_{i,j=1,i\neq j}^n (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})\right)\right) \\
&\leq \frac{32}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^n a_{i\Pi(i)}^2\right) + \frac{32}{n^4}\,\mathrm{Var}\left(\sum_{i,j=1,i\neq j}^n (a_{i\Pi(i)}a_{j\Pi(j)} + a_{i\Pi(j)}a_{j\Pi(i)})\right)
\end{aligned} \tag{1.14}$$

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

First, we address the first term in (1.14):

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

with

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

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 $\mathbf{a} := [a_{11}^2 \; a_{12}^2 \ldots a_{nn}^2]^T$ out as a vector, $\mathbf{a}^T\mathbf{1} = n-1$. By Cauchy-Schwarz,

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

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

$$\sum_{i,j=1,i\neq j}^{n} \mathrm{Cov}(a_{i\Pi(i)}^2, a_{j\Pi(j)}^2) \leq \frac{1}{n} \sum_{i,j=1}^{n} a_{i,j}^4. \tag{1.15}$$

For the second term in (1.14) we again apply corollary 1.3:

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

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

$$X = \sum_{i=1}^{n} a_{i\Pi(i)} \sum_{j=1,j\neq i}^{n} a_{j\Pi(j)} = W^2 - \sum_{i=1}^{n} a_{i\Pi(i)}^2. \tag{1.16}$$

TODO: ... Maybe finish this up later? □

1.3 Exchangeable Pairs

TODO: Add a lot of development for exchangeable pairs. For now, focusing on generalizing the theorems in "Normal Approximation by Stein's Method."

Theorem 5.5 in "Normal Approximation by Stein's Method" concerns variance 1 exchangeable random variables. Our setting has the variance tending to 1, so we first prove a slight generalization of the theorem. Large parts of the proof are copied verbatim from the book.

1.4 Preliminaries

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

$$\mathbb{E}[W - W'|W] = \lambda(W - R) \tag{1.17}$$

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

Lemma 1.5. *If (W, W') is an exchangeable pair, then $\mathbb{E}[g(W, W')] = 0$ for all antisymmetric measurable functions such that the expected value exists.*

Here is a slight generalization of Lemma 2.7:

Lemma 1.6. *Let (W, W') be an approximate Stein pair and $\Delta = W - W'$. Then*

$$\mathbb{E}[W] = \mathbb{E}[R] \quad \text{and} \quad \mathbb{E}[\Delta^2] = 2\lambda\mathbb{E}[W^2] - 2\lambda\mathbb{E}[WR] \quad \text{if } \mathbb{E}[W^2] < \infty. \tag{1.18}$$

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

$$\mathbb{E}[Wf(W)] = \frac{1}{2\lambda} = \mathbb{E}[(W - W')(f(W) - f(W'))] + \mathbb{E}[f(W)R]. \tag{1.19}$$

Proof. From (1.17) 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,

$$\begin{aligned} \mathbb{E}[W'W] &= \mathbb{E}[\mathbb{E}[W'W|W]] \\ &= \mathbb{E}[W\mathbb{E}[W'|W]] \\ &= \mathbb{E}[W((1-\lambda)W + \lambda R)] \quad \text{from (1.17)} \\ &= (1-\lambda)\mathbb{E}[W^2] + \lambda\mathbb{E}[WR]. \end{aligned} \tag{1.20}$$

Then

$$\begin{aligned} \mathbb{E}[\Delta^2] &= \mathbb{E}[(W - W')^2] \\ &= \mathbb{E}[W^2] + \mathbb{E}[W'^2] - 2\mathbb{E}[W'W] \\ &= 2\mathbb{E}[W^2] - 2((1-\lambda)\mathbb{E}[W^2] + \lambda\mathbb{E}[WR]) \quad \text{from (1.20)} \\ &= 2\lambda\mathbb{E}[W^2] - 2\lambda\mathbb{E}[WR]. \end{aligned} \tag{1.21}$$

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 1.5,

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

Rearranging the expression yields

$$\mathbb{E}[Wf(W)] = \frac{1}{2\lambda}\mathbb{E}[(W-W')(f(W)-f(W'))] + \mathbb{E}[f(W)R]. \tag{1.22}$$

□

This is just a small part of Lemma 2.4:

Lemma 1.7. *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*

$$||f_h|| \leq 2||h'||. \tag{1.23}$$

1.5 Main Theorem

Generalization of Theorem 5.5:

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

$$\mathbb{E}[T'-T|T] = -\lambda(T-R)$$

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

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

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

$$f'(w) - wf(w) = h_{z,\alpha}(w) - \Phi(z) \tag{1.24}$$

for the smoothed indicator

$$h_{z,\alpha}(w) = \begin{cases} 1 & w \leq z \\ 1 + \frac{z-w}{\alpha} & z < w \leq z + \alpha \\ 0 & w > z + \alpha. \end{cases} \tag{1.25}$$

Therefore,

$$\begin{aligned} |P(W \leq z) - \Phi(z)| &= |\mathbb{E}[(f'(W) - Wf(W))]| \\ &= \left| \mathbb{E}\left[f'(W) - \frac{(W' - W)(f(W') - f(W))}{2\lambda} + f(W)R \right] \right| \\ &= \left| \mathbb{E}\left[f'(W)\left(1 - \frac{(W' - W)^2}{2\lambda}\right) \right.\right. \\ &\quad \left.\left. + \frac{f'(W)(W' - W)^2 - (f(W') - f(W))(W' - W)}{2\lambda} + f(W)R \right] \right| \\ &:= |\mathbb{E}[J_1 + J_2 + J_3]| \\ &\leq |\mathbb{E}[J_1]| + |\mathbb{E}[J_2]| + |\mathbb{E}[J_3]|. \end{aligned} \tag{1.26}$$

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

$$|\mathbb{E}[J_3]| \leq \mathbb{E}[|J_3|] = \mathbb{E}[|f(W)R|] \leq \mathbb{E}[|R|] \tag{1.27}$$

and

$$\begin{aligned}
|\mathbb{E}[J_1]| &= \left|\mathbb{E}\left[f'(W)\left(1-\frac{(W'-W)^2}{2\lambda}\right)\right]\right| \\
&\leq \mathbb{E}\left[\left|f'(W)\left(1-\frac{(W'-W)^2}{2\lambda}\right)\right|\right] \\
&\leq \mathbb{E}\left[\left|\left(1-\frac{(W'-W)^2}{2\lambda}\right)\right|\right] \\
&= \frac{1}{2\lambda}\mathbb{E}[|2\lambda - \mathbb{E}[(W'-W)^2|W]|] \\
&= \frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}[W^2]-\mathbb{E}[WR]) - \mathbb{E}[(W'-W)^2|W] + 2\lambda(1-\mathbb{E}[W^2]+\mathbb{E}[WR])|] \\
&\leq \frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}[W^2]-\mathbb{E}[WR]) - \mathbb{E}[(W'-W)^2|W]|] + \mathbb{E}[|(1-\mathbb{E}[W^2]+\mathbb{E}[WR])|]
\end{aligned} \tag{1.28}$$

Note that

$$\mathbb{E}[\mathbb{E}[(W'-W)^2|W]] = \mathbb{E}[\Delta^2] = 2\lambda(\mathbb{E}[W^2]-\mathbb{E}[WR]), \tag{1.29}$$

so

$$\frac{1}{2\lambda}\mathbb{E}[|2\lambda(\mathbb{E}[W^2]-\mathbb{E}[WR]) - \mathbb{E}[(W'-W)^2|W]|] \leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])}. \tag{1.30}$$

Combining with (1.28),

$$\begin{aligned}
|\mathbb{E}[J_1]| &\leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} + \mathbb{E}[|1-\mathbb{E}[W^2]+\mathbb{E}[WR]|] \\
&\leq \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2|W])} + \mathbb{E}[|1-\mathbb{E}W^2|] + \mathbb{E}|WR|
\end{aligned} \tag{1.31}$$

Lastly, we bound the second term,

$$\begin{aligned}
J_2 &= \frac{1}{2\lambda}(W'-W)\int_W^{W'}(f'(W)-f'(t))dt \\
&= \frac{1}{2\lambda}(W'-W)\int_W^{W'}\int_t^W f''(u)dudt \\
&= \frac{1}{2\lambda}(W'-W)\int_W^{W'}(W'-u)f''(u)du.
\end{aligned} \tag{1.32}$$

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

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

For the latter,

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

Since W and W' are exchangeable,

$$\begin{aligned}|\mathbb{E}[J_2]| &= \left|\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)\int_W^{W'}(W'-u)f''(u)du\right]\right| \\ &= \left|\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)\int_W^{W'}\left(\frac{W+W'}{2}-u\right)f''(u)du\right]\right| \\ &\leq \left|\mathbb{E}\left[||f''||\frac{1}{2\lambda}|W'-W|\int_{\min(W,W')}^{\max(W,W')}\left|\frac{W+W'}{2}-u\right|du\right]\right| \\ &= \left|\mathbb{E}\left[||f''||\frac{1}{2\lambda}\frac{|W'-W|^3}{4}\right]\right| \\ &\leq \frac{\mathbb{E}[|W'-W|^3]}{4\alpha\lambda},\end{aligned} \tag{1.33}$$

where the final inequality follows from the fact that $|h'_{z,\alpha}(x)| \leq 1/\alpha$ for all $x \in \mathbb{R}$ and Lemma 1.7.

Collecting the bounds, we obtain

$$\begin{aligned} P(W \leq z) &\leq \mathbb{E}[h_{z,\alpha}(W)] \\ &\leq Nh_{z,\alpha} + \frac{\mathbb{E}[|W' - W|^3]}{4\alpha\lambda} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} \\ &\quad + |1 - \mathbb{E}[W^2]| + \mathbb{E}|WR| + \mathbb{E}|R| \\ &\leq \Phi(z) + \frac{\alpha}{\sqrt{2\pi}} + \frac{\mathbb{E}[|W' - W|^3]}{4\alpha\lambda} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} \\ &\quad + |\mathbb{E}[W^2] - 1| + \mathbb{E}|WR| + \mathbb{E}|R| \end{aligned} \tag{1.34}$$

The minimizer of the expression is

$$\alpha = \frac{(2\pi)^{1/4}}{2}\sqrt{\frac{\mathbb{E}[|W' - W|^3]}{\lambda}}. \tag{1.35}$$

Plugging this in, we get the upper bound

$$\begin{aligned} P(W \leq z) - \Phi(z) \leq (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}[|W' - W|^3]}{\lambda}} + \frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(W' - W)^2|W])} \\ + |\mathbb{E}[W^2] - 1| + \mathbb{E}|WR| + \mathbb{E}|R| \end{aligned} \tag{1.36}$$

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

Chapter 2

Main Proof

In this chapter, we prove the core theoretical result of this thesis: a rate of convergence bound for the randomization distribution of the t-statistic, using theorem 1.8 of chapter 1.

2.1 Motivation

Motivated by concerns regarding normality assumptions in the hypothesis being tested, Fisher [**?**] 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 [**?**] 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 [**?**] or David [**?**]. Diaconis and Lehmann [**?**] in their comment on Zabell's paper further expanded on some properties of these randomization tests.

Ludbrook and Dudley [**?**] 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 [**?**], and Fisher's randomization model [**?**]. Add some more on these two models...

Under the randomization model and using the language of triangular arrays,

Lehmann [**?**] 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 [**?**], 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 [**?**] first obtained a Berry-Esseen bound for Student's statistic in the independent but non-identically distributed setting with additional work by Shao [**?**].

We use Stein's method of exchangeable pairs to prove a conservative bound of $O(N^{-1/4})$ on the rate of convergence of the randomization t-distribution to the standard normal distribution.

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 $S_2 = \{u_{\pi_0(i)}\}_{i=N+1}^{2N}$. 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

$$q_\Pi = \left(\sum_{i=1, i\neq I}^{N} u_{\Pi(i)} + u_{\Pi(I)} - \sum_{i=N+1, i\neq J}^{2N} u_{\Pi(i)} - u_{\Pi(J)} \right)$$

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

$$\bar{u}_{1,\Pi} = \frac{1}{N} \sum_{i=1}^{N} u_{\Pi(i)} \text{ and } \bar{u}_{2,\Pi} = \frac{1}{N} \sum_{i=N+1}^{2N} u_{\Pi(i)}$$

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 $I, J \in \{1, \ldots, N\}$ or $I, J \in \{N+1, \ldots, 2N\}$ 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 linear 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

$$z_i = \sqrt{\frac{2N}{\sum (u_i - \bar{u})^2}}(u_i - \bar{u}), \tag{2.1}$$

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:

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

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.

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

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

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

By assumption (2.3),

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

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,

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

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,

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

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

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

Therefore,

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

The important thing to do is to isolate $|h_\Pi|$, which is small in expectation, but not absolutely. By the c_r-inequality,

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

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,

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

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

2.5 Proof

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, $\pi^* = (2N, \ldots, 1) \circ \pi$. Since indices 1 to N correspond to the

first group and $N+1$ to $2N$ to the second, π^* flips the groups after π, so $T_{\pi^*} = -T_\pi$.

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

For convenience, we restate theorem 1.8 of chapter 1:

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

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

The difference of our exchangeable pair is given by

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

Note that

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

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

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

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

so

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

Proposition 2.2. $|\mathbb{E}T_\Pi^2-1|\le c_2N^{-1}$

Proof.

$$\mathbb{E}T_\Pi^2=\frac{N-1}{N}\mathbb{E}\left[\left(\frac{q_\Pi}{d_\Pi}\right)^2\right] \tag{2.13}$$
$$=\frac{N-1}{N}\mathbb{E}\left[\frac{4N^2\bar{u}_{2,\Pi}^2}{2N-2N\bar{u}_{2,\Pi}^2}\right]\quad\text{from (2.8)}$$
$$=2(N-1)\mathbb{E}\left[\frac{\bar{u}_{2,\Pi}^2}{1-\bar{u}_{2,\Pi}^2}\right]$$
$$=2(N-1)\mathbb{E}[g(\bar{u}_{2,\Pi})], \tag{2.14}$$

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 (2.14) 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,

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

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

This implies that

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

□

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

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

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

From (2.11),

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

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

$$\begin{aligned}\mathrm{Var}(WV) &= \mathrm{Var}(W(V-\mathbb{E}V) + W\mathbb{E}V) \\ &\le 2\,\mathrm{Var}(W(V-\mathbb{E}V)) + 2\,\mathrm{Var}(W\mathbb{E}V) \\ &\le 2\mathbb{E}[W^2(V-\mathbb{E}V)^2] + 2(\mathbb{E}V)^2\,\mathrm{Var}(W) \\ &\le 2(f_{c_2}(2))^2 N^{-2}\,\mathrm{Var}(V) + 2u_\Delta^4\,\mathrm{Var}(W). \end{aligned} \tag{2.15}$$

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

where

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

By Taylor's theorem,

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

with

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

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

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

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

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

with

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

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

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

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 $I \neq K, J = L$:

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

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,

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

And to the third,

$$\begin{aligned}
\mathrm{Var}(Z) &= 4\,\mathrm{Var}\left(\mathbb{E}\left[\left(\frac{2u_{\Pi(J)}-2u_{\Pi(I)}}{d_\Pi}T'_\Pi\frac{d_\Pi-d'_\Pi}{d_\Pi}\right)\middle|\Pi=\pi\right]\right)\\
&\le 16u_\Delta^2\mathbb{E}\left[\left(\frac{1}{d_\Pi}\frac{q'_\Pi}{d_\Pi d'_\Pi}(d_\Pi-d'_\Pi)\right)^2\right]\\
&\le 16u_\Delta^2 f_{c_2}(2)N^{-2/2}\sqrt{\mathbb{E}\left[\left(\frac{q'_\Pi}{d_\Pi d'_\Pi}\right)^4\right]\mathbb{E}[(d_\Pi-d'_\Pi)^4]}\text{ from (2.5)}\\
&\le 16u_\Delta^2 f_{c_2}(2)N^{-1}\sqrt{f_{c_6}(4)N^{-4/2}f_{c_4}(4)N^{-4}}\text{ from (2.10), (2.7)}\\
&\le 16u_\Delta^2 f_{c_2}(2)(f_{c_6}(4))^{-1/2}(f_{c_4}(4))^{-1/2}N^{-4}\\
&:= c_8N^{-4}
\end{aligned}$$

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

□

Proposition 2.4. $(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi-T_\Pi|^3}{\lambda}} < (2\pi)^{-1/4}c_9N^{-1/4}$.

Proof. The strategy is to break apart the remainder term from the main piece. From

(2.11),

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

Therefore,

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

□

Proposition 2.5. $\mathbb{E}|R| \leq \frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}$.

Proof.

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

□

Proposition 2.6. $\mathbb{E}|T_\Pi R| \leq \frac{1}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{2+2c_1}N^{-1/2}$.

Proof.

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

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

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

$$\begin{aligned}
\sup_{t\in\mathbb{R}}|P(T_\Pi\le t)-\Phi(t)| &\le (2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T'_\Pi-T_\Pi|^3}{\lambda}}+\frac{1}{2\lambda}\sqrt{\mathrm{Var}(\mathbb{E}[(T'_\Pi-T_\Pi)^2|T_\Pi])} \\
&\quad+|\mathbb{E}T_\Pi^2-1|+\mathbb{E}|T_\Pi R_\Pi|+\mathbb{E}|R_\Pi| \\
&\le N^{-1}c_3\sqrt{20+16\frac{\sum_{i=1}^{2N}u_i^4}{N^2}}+(2\pi)^{-1/4}c_9N^{-1/4}+c_2N^{-1} \\
&\quad+\frac{1}{2}(f_{c_6}(4)f_{c_4}(4))^{1/4}\sqrt{2+2c_1}N^{-1/2}+\frac{1}{2}\sqrt{f_{c_6}(2)f_{c_4}(2)}N^{-1/2}
\end{aligned}$$

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 $N \in \{\text{floor}(10^{.5+.5i})\}_{i=1}^{7}$. 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.

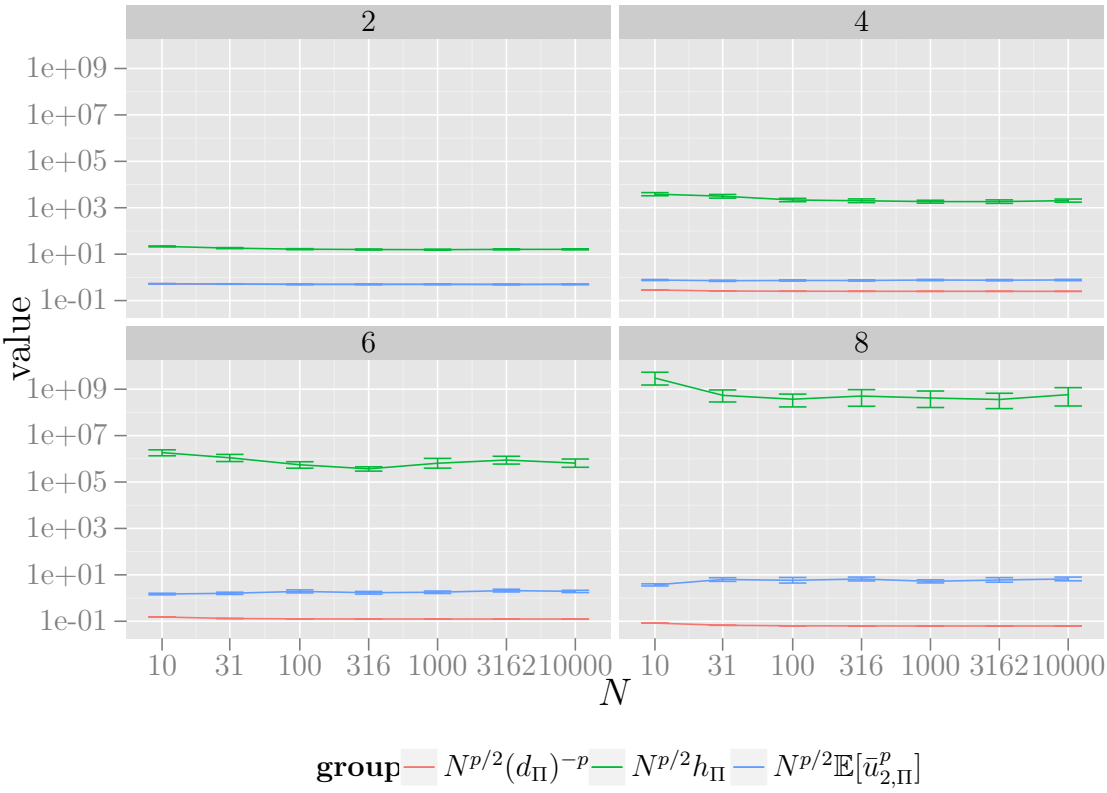


Figure 3.1: Log-log plots of values scaled by proven upper bounds of rates, faceted on p.

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

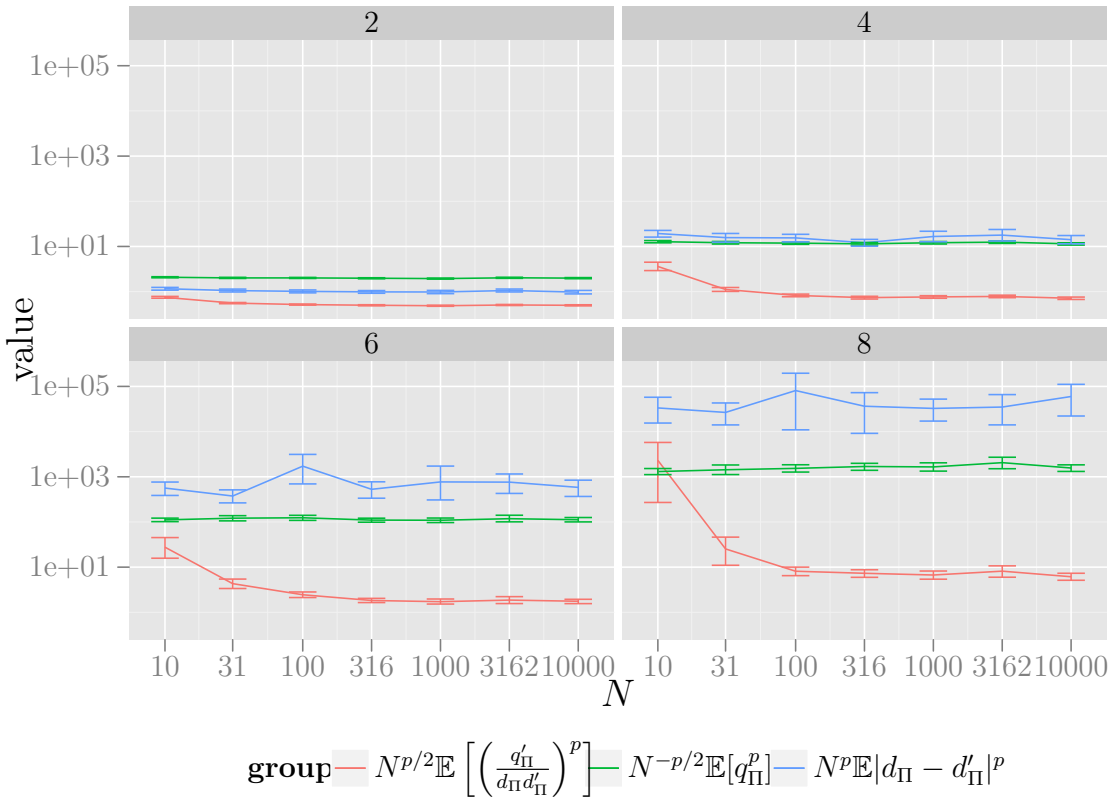


Figure 3.2: Log-log plots of values scaled by proven upper bounds of rates, faceted on p.

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 with 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 100 times.

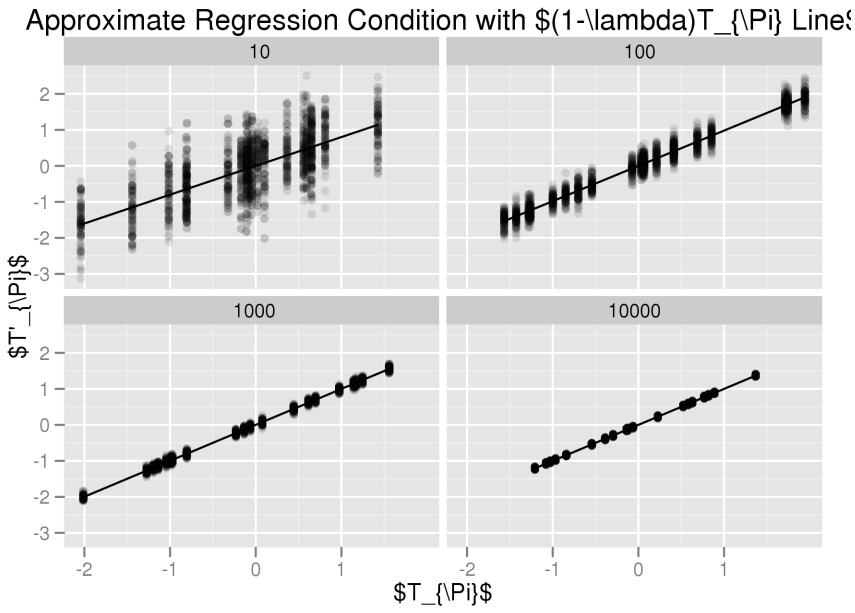


Figure 3.3: Faceted on per-group sample size, N.

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).

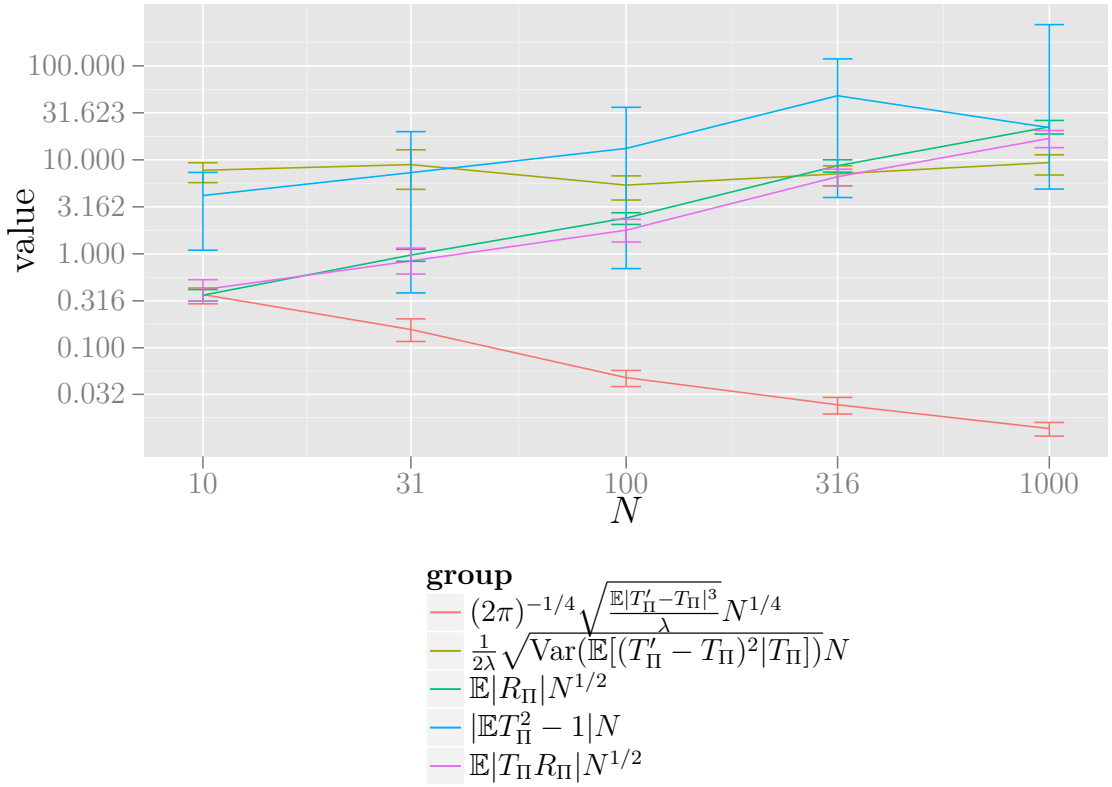


Figure 3.4: Log-log plot of values for each term in the bound, simulating the conditional expectation by Monte Carlo (1,000 MC draws, 100 permutations each).

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

3.3.2 Exact Conditional Expectation Calculations

Here we T' for all N^2 group-switching transpositions (I, J):

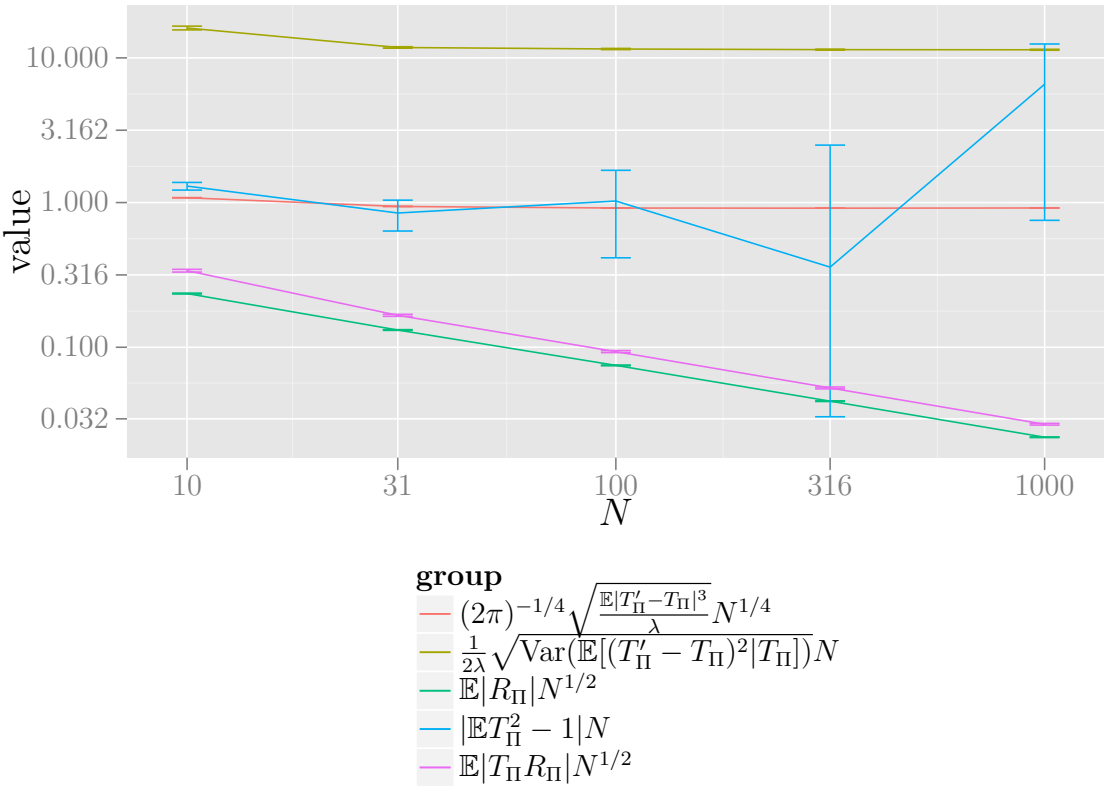
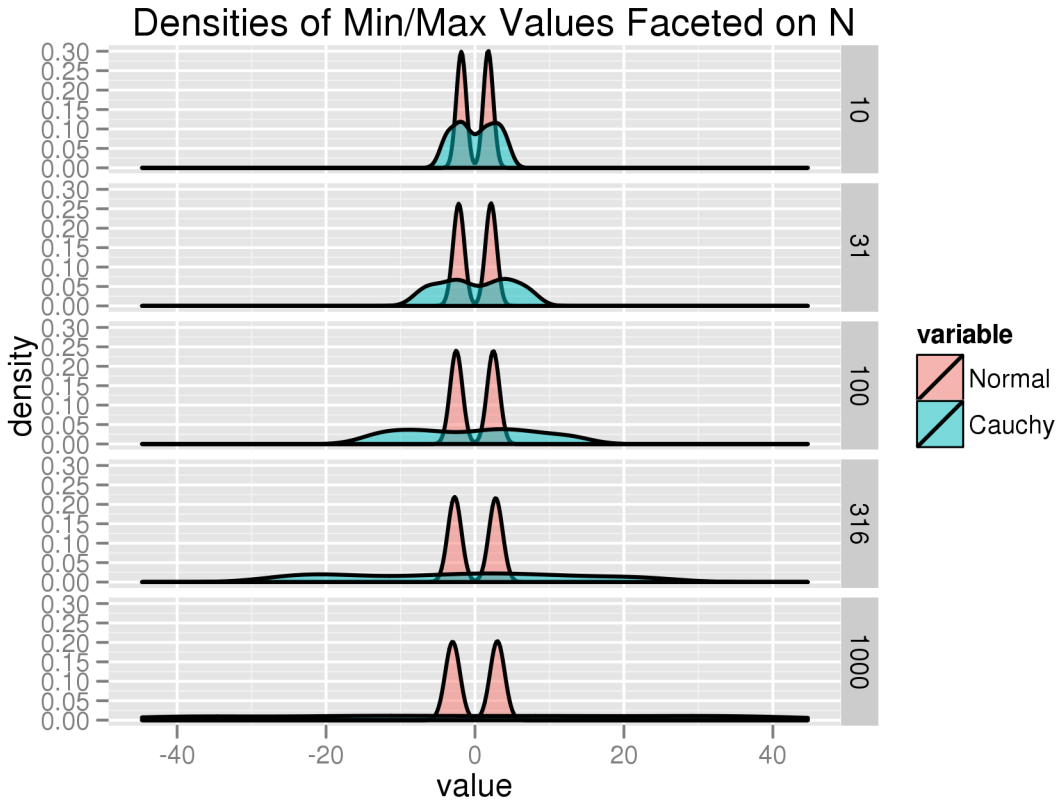


Figure 3.5: Log-log plot of values for each term in the bound, calculating the conditional expectation exactly (200,000 permutations each).

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.

3.4 True Rate

To assess the true rate of convergence, we consider two settings: the earlier Normal setting and group draws from a Cauchy distribution with location parameters -1 and 1 depending on the group. Our bounds include a dependence on $u_\Delta = \max_i u_i - \min_i u_i$. To better understand the differences between these two models, we simulate the minimum and maximum scaled (mean 0 and sum of squares $2N$) values:



For $N = 1000$, the Normal model typically has u_Δ values around 6. In contrast, the Cauchy model has u_Δ values closer to 40.

Here, we plot the empirical Kolmogorov-Smirnov test statistic in the following three settings:

1. a standard Normal draw of size N (repeated N times to get the empirical distribution)

2. the permutation t-statistic under Cauchy sampling (N permutations)

3. the permutation t-statistic under Normal sampling (N permutations)

We also add the sum of the five unscaled, simulated bound terms ($200,000$ permutations) from the previous section.

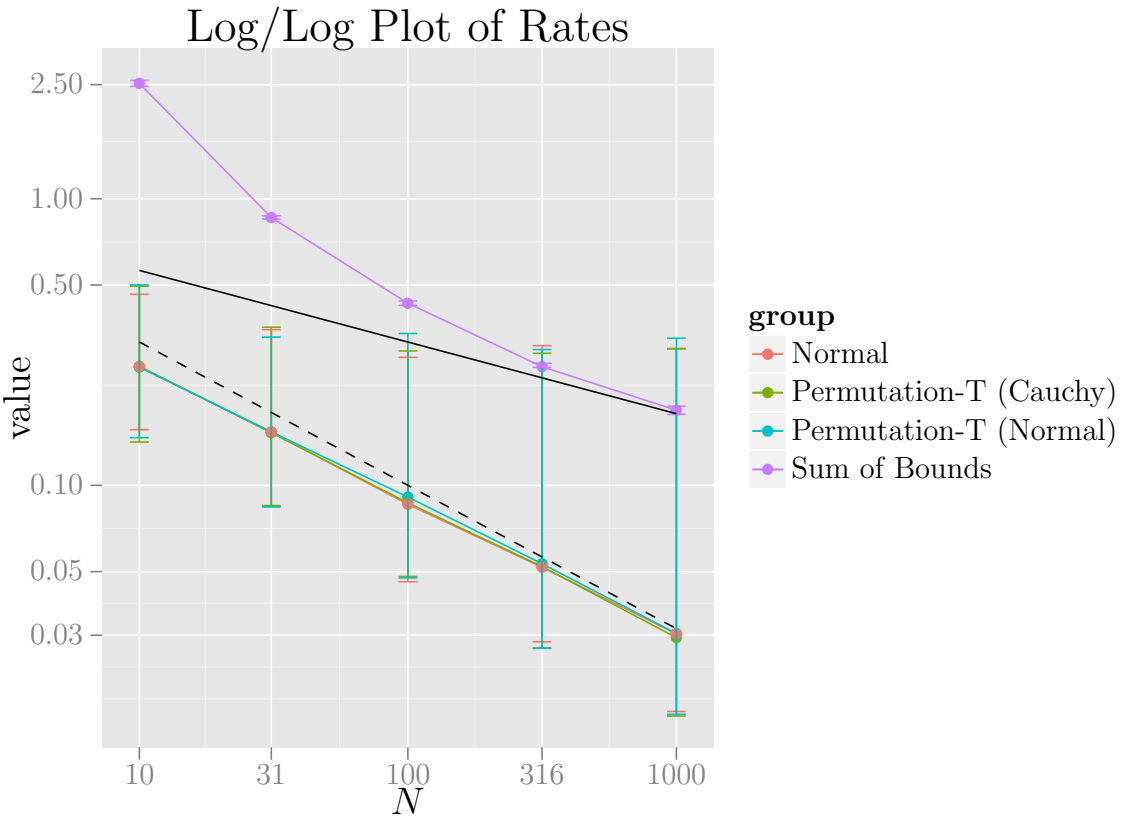


Figure 3.6: Solid black line: $N^{-1/4}$; dashed black line: $N^{-1/2}$

It's not a fair comparison to place the sum of the bounds on the same plot because that was computed over $200,000$ separate permutations instead of the 500 shared by the other three. Still, we can draw some general conclusions. The normal and two permutation-t K-S statistics decay perfectly at a rate of $N^{-1/2}$, and our bound follows a rate of $N^{-1/4}$, suggesting that the true rate of convergence is the former. Also, the error-bars seem to be increasing in size but are actually roughly constant due to the log-log scale.

Chen et al. [**?**] provide a simple example (pp.154-155) in which the sum of i.i.d. random variables yields

$$E|W' - W|^3 = \frac{4}{N^{3/2}}$$

with $\lambda = N^{-1}$. This leads to an $O(N^{-1/4})$ bound, which is suboptimal and apparently not uncommon when applying this kind of theorem.

3.5 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.

Consider a naïve implementation based on a double for-loop and recomputing each t-statistic anew versus a vectorized approach using the update formula:

```
computeAllCond2 <- function(T, N, u, l, x, y){
    minus <- which(l == -1)
    plus <- which(l == 1)
    Tprime <- 1:(N^2)
    for(j in 1:N){
      for(k in 1:N){
        swap <- c(minus[j], plus[k])
```

```
        l[swap] <- l[rev(swap)]
        Tprime[N*(j-1)+k] <- t.test(u[l==1], u[l==-1], var.equal=TRUE)$statistic
        l[swap] <- l[rev(swap)]
      }
    }
    data.frame("T" = T, "Tprime" = Tprime, "N" = N, "lambda" = 2 / N)
}

computeAllCond <- function(T, N, u, l, x, y){
  del <- rep(y, length(x)) - rep(x, each = length(y))
  xbar <- mean(x)
  ybar <- mean(y)
  Tprime <- -(xbar - ybar + 2/N*del) /
    (sqrt(2/N)*sqrt(sum(u^2)/(2*(N-1)) - 1/2*N/(N-1)*(xbar^2 + ybar^2 + 2*del/N*(x
  data.frame("T" = T, "Tprime" = Tprime, "N" = N, "lambda" = 2 / N)
}
```

We observe roughly a 2,000 times increase in speed on a problem instance of size $N = 100$. With byte-compilation and additional tuning, a four order of magnitude increase is possible.

```
> system.time(computeAllCond2(T, N, u, l, x, y))
   user  system elapsed
  7.333   0.000   7.334
> system.time(computeAllCond(T, N, u, l, x, y))
   user  system elapsed
  0.005   0.000   0.004
> sum((sort(computeAllCond(T, N, u, l, x, y)$Tprime) - sort(computeAllCond2(T, N,
[1] 3.137579e-27
dat <- ldply(rep(floor(10^(seq(1, 3.5, by=.5))), each = 8),
simulateBounds, .parallel = TRUE, .progress = "text")
> print(object.size(dat), units = "Gb")
```

```
2.6 Gb
```

3.6 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.

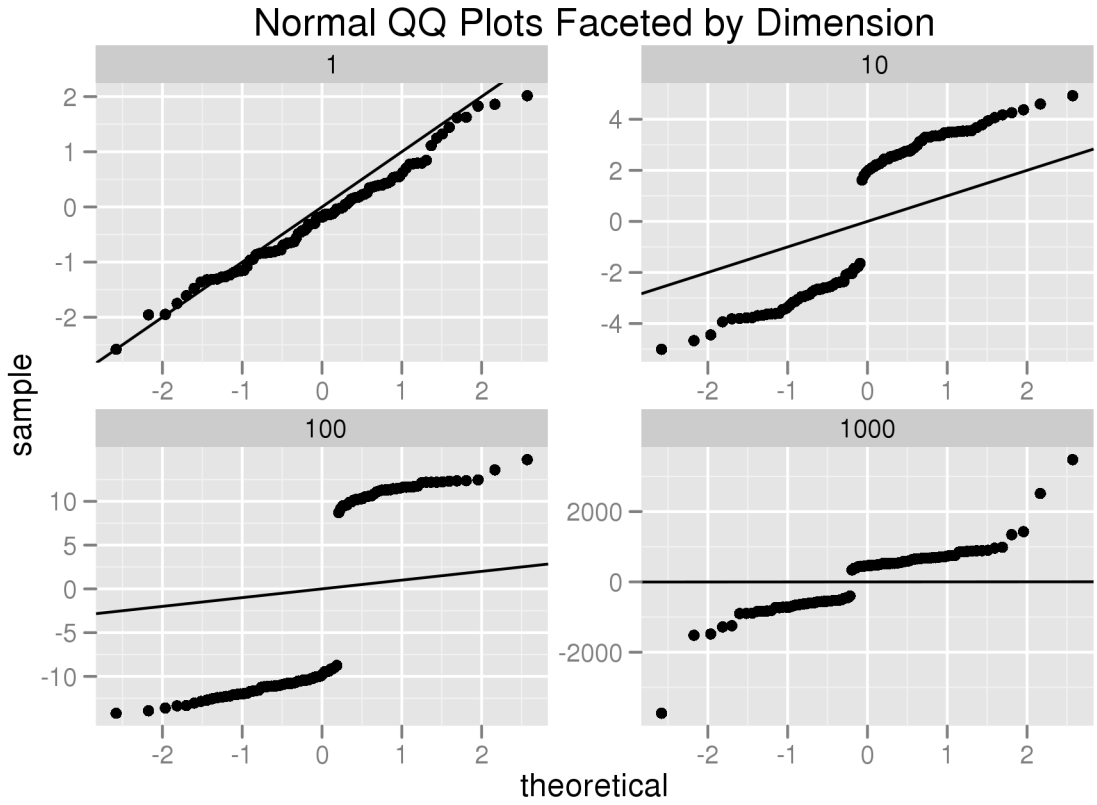
It's nice that the calculations are invariant to change in the exchangeable pair. Whether that holds true for more drastic changes (e.g. swapping more than 2 elements) is not known.

3.7 Generalizations (Null Distribution)

It is natural to consider generalizations from the univariate data, linear kernel setting. We explore whether the randomization distribution is still Normal with multivariate data and/or a non-linear kernel. Should the null distribution be non-Normal, we further attempt to determine whether the approximate regression condition holds.

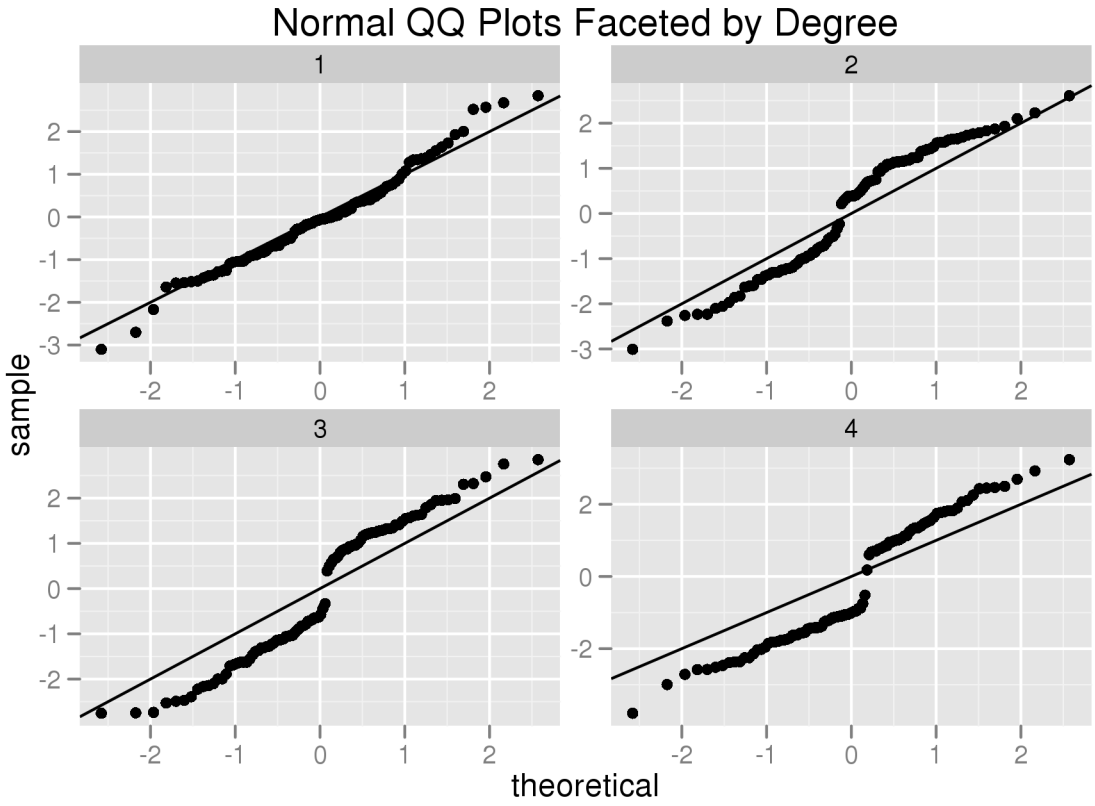
We generate 100 observations with dimensionality 1, 10, 100, and 1000. For each set of data, we permute 100 times and plot the Friedman statistic (t-statistic on SVM fitted values) with a linear kernel against the standard Normal quantiles. Note

that we take the sign of the Friedman statistic to be positive or negative with equal probability for ease of comparing distributions.



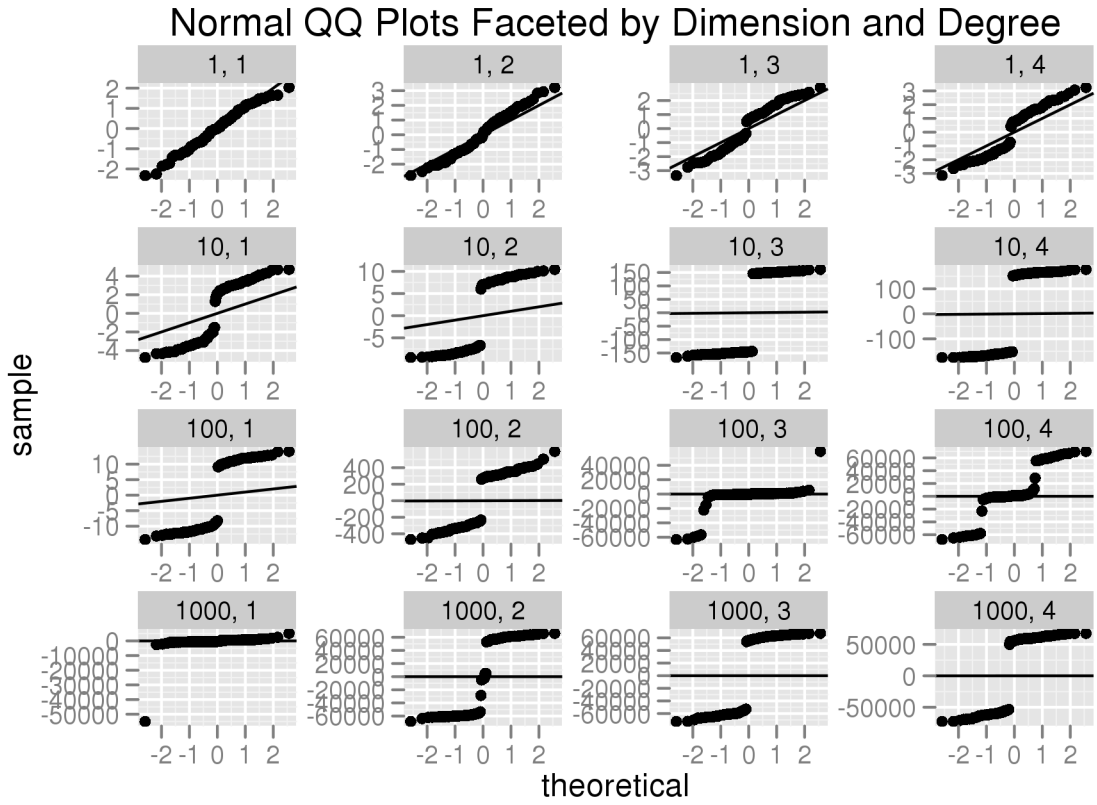
At a sample size of 100 and with univariate data, the standard Normal is a close fit. With increasing dimensionality, the Friedman statistic gets more and more extreme. One possible explanation is that it becomes easier to separate two sets of points as the dimensionality increases.

Here we look at univariate data but with an inhomogeneous kernel ($k(x, x') = (< x, x' > +1)^d$) of degree d:



The polynomial kernel for degrees greater than 1 yields null distributions with fatter tails than the standard Normal.

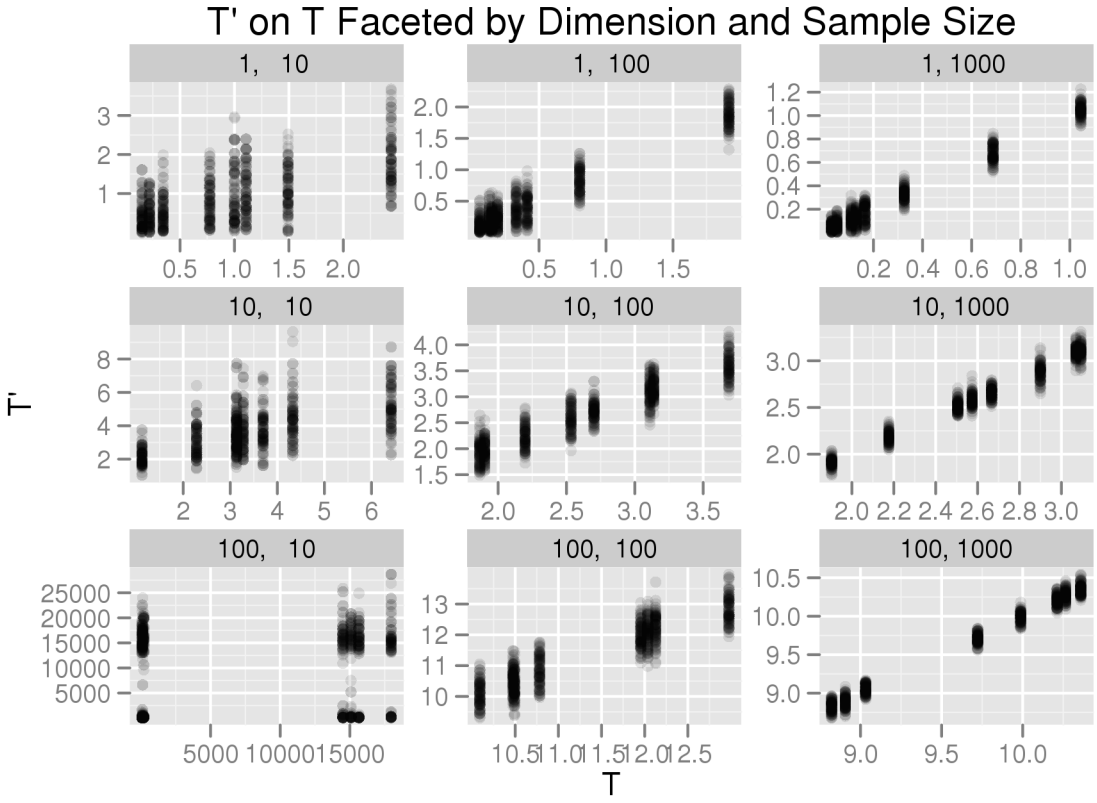
Here we look at the effects of both dimension of the underlying data and degree of the inhomogeneous kernel:



3.8 Generalizations (Approximate Regression Condition)

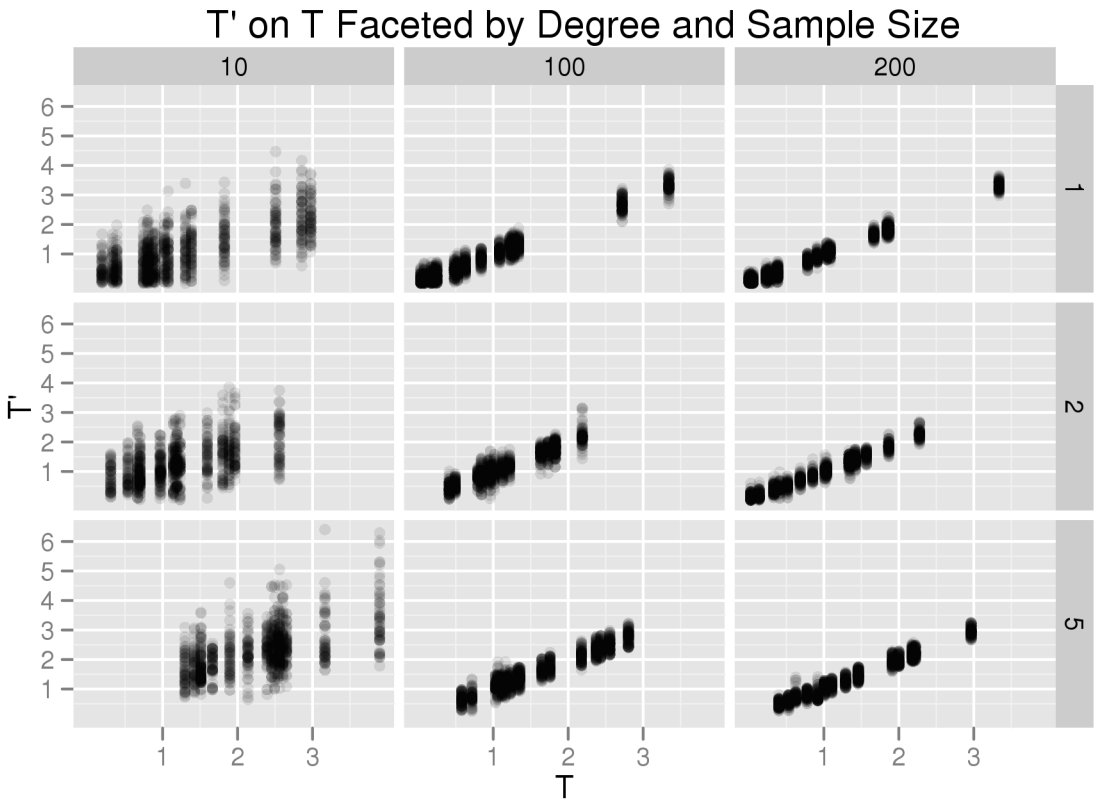
Here we plot T' on T, where T' results from swapping labels and refitting the SVM. Note that here we don't assign the statistics random signs because there is no clear way to maintain the coupling between T and T'.

The dimensions (rows) are 1, 10, and 100, and the sample sizes (columns) are 10, 100, and 1000.



Across a row, it is clear that a larger sample size decreases the variability of T' about T. And down a column, it is clear that increasing dimensionality results in a greater departure from (folded) Normality. It is not clear whether the reduction in variability is of order $1/N$.

Here we look at sample sizes (columns) of 10, 100, and 200 and inhomogeneous polynomial kernel degrees (rows) of 1, 2, and 5.



We again observe an approximately linear plus noise relationship between T' and T with the noise decreasing in sample size.

Chapter 4

Friedman's Test

In this chapter we describe Friedman's approach to the two-sample problem, give 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?

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.

4.2 The Friedman Two-Sample Test

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

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 $p < \alpha$.

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 $F(\mathbf{u})$.

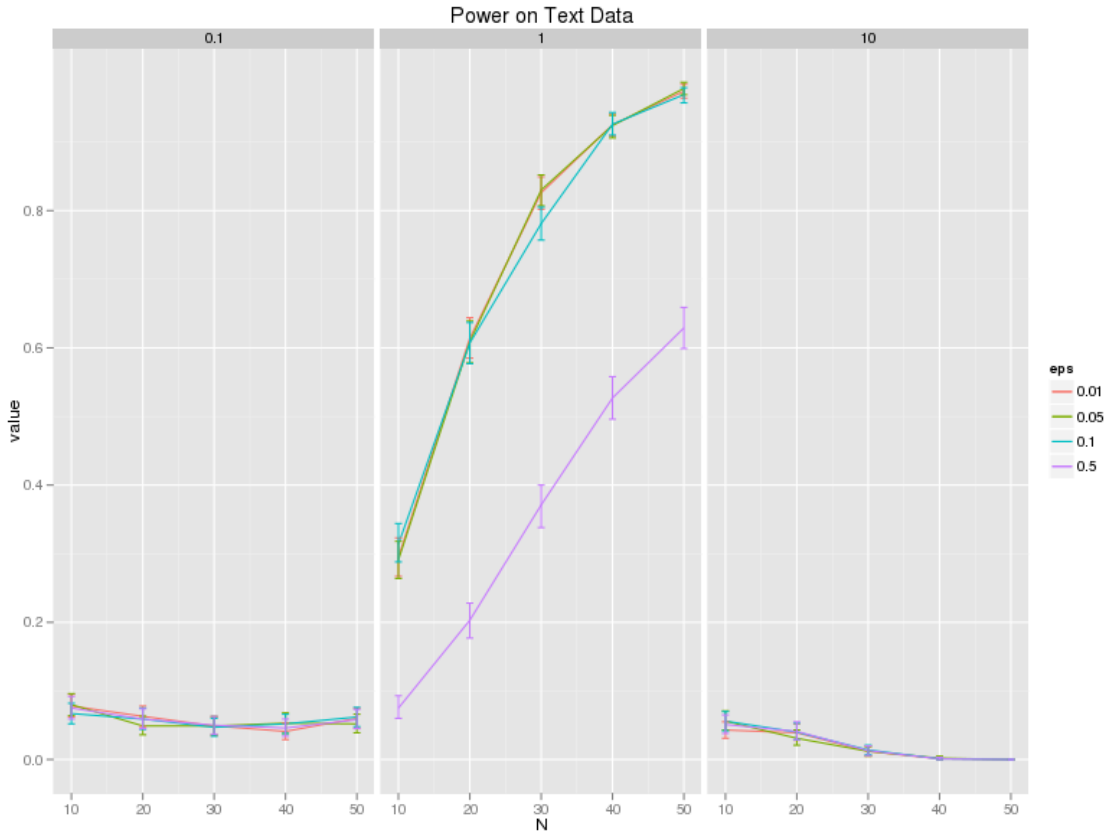


Figure 4.1: Friedman Test (SVM with 3-spectrum kernel) for Twitter data demonstrating power for columns $C \in \{.1, 1, 10\}$ and colors $\varepsilon \in \{.01, .05, .1, .5\}$. Error bars indicate 95% bootstrap confidence intervals. Tuning parameter choice is *critical*. We fix $C = 1$ and $\varepsilon = .1$ for computational considerations, but cross-validation is recommended.

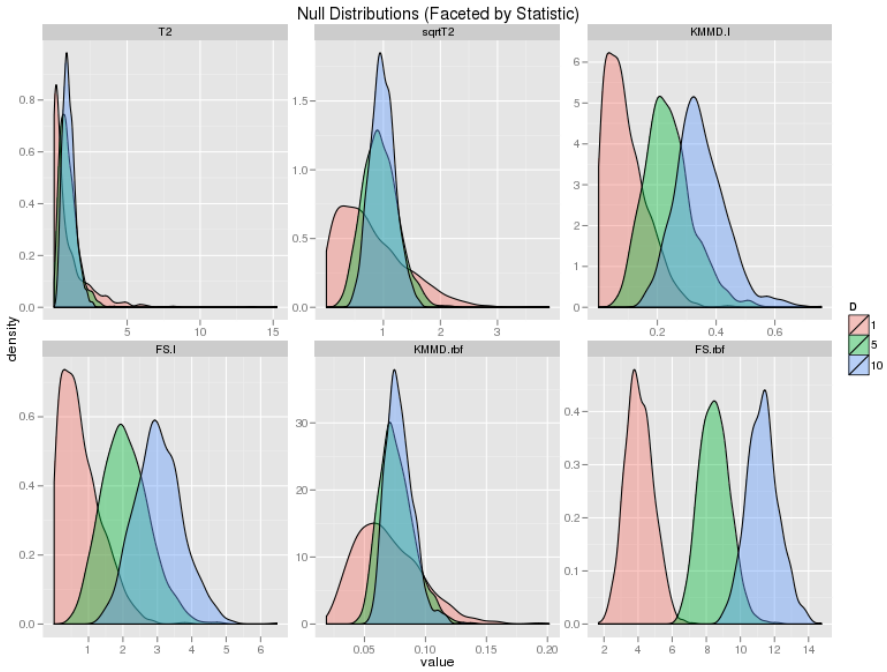


Figure 4.2: T2: Hotelling's T^2-statistic; sqrtT2: $|T|$; KMMD.l: kernel MMD with a linear kernel; FS.l: FS with a linear kernel; KMMD.rbf: kernel MMD with a radial basis function (RBF) kernel; FS.rbf: FS with RBF kernel

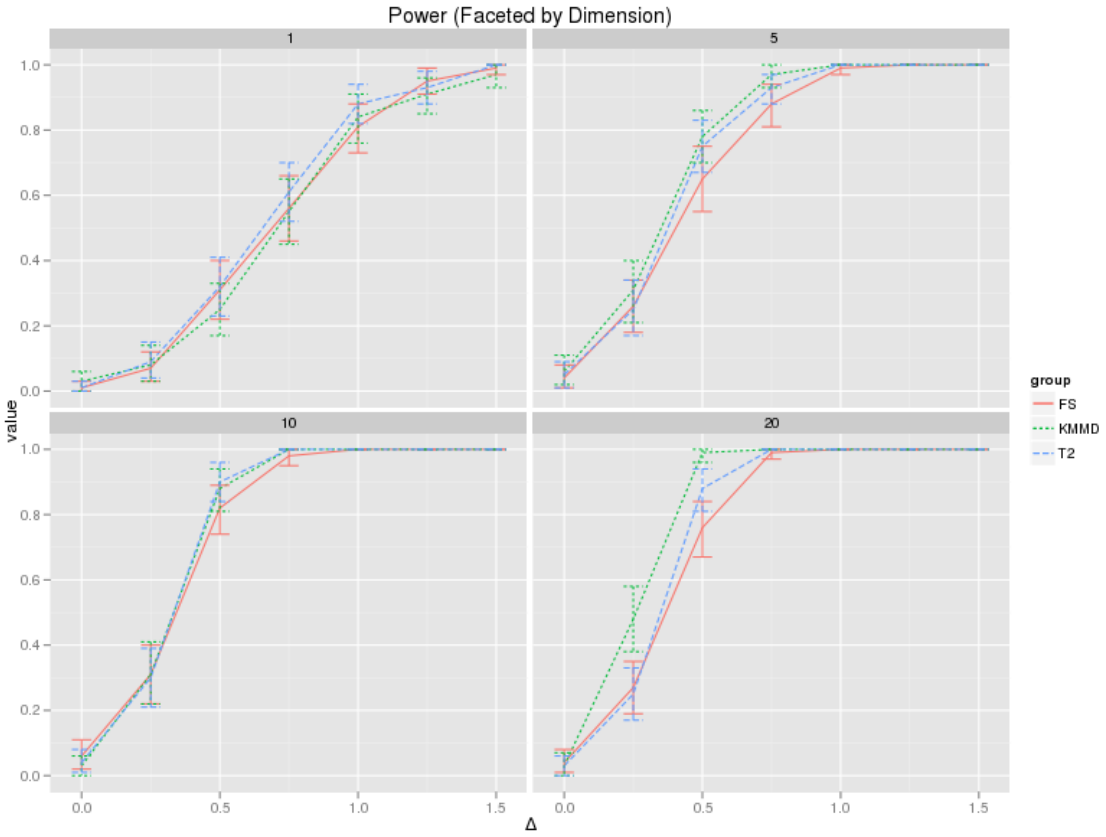


Figure 4.3: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; T2: Hotelling's T^2-statistic; Error bars indicate 95% bootstrap confidence intervals. The tests perform similarly, and the kernel-based tests use a linear kernel.

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 [**?**]. 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.

4.3 SVM

We experience better computational results with Support Vector Machine (SVM) regression rather than classification as implemented in the **ksvm** function of the **R** [**?**] package **kernlab** [**?**].

Recall that SVM regression solves the following problem [**?**]:

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

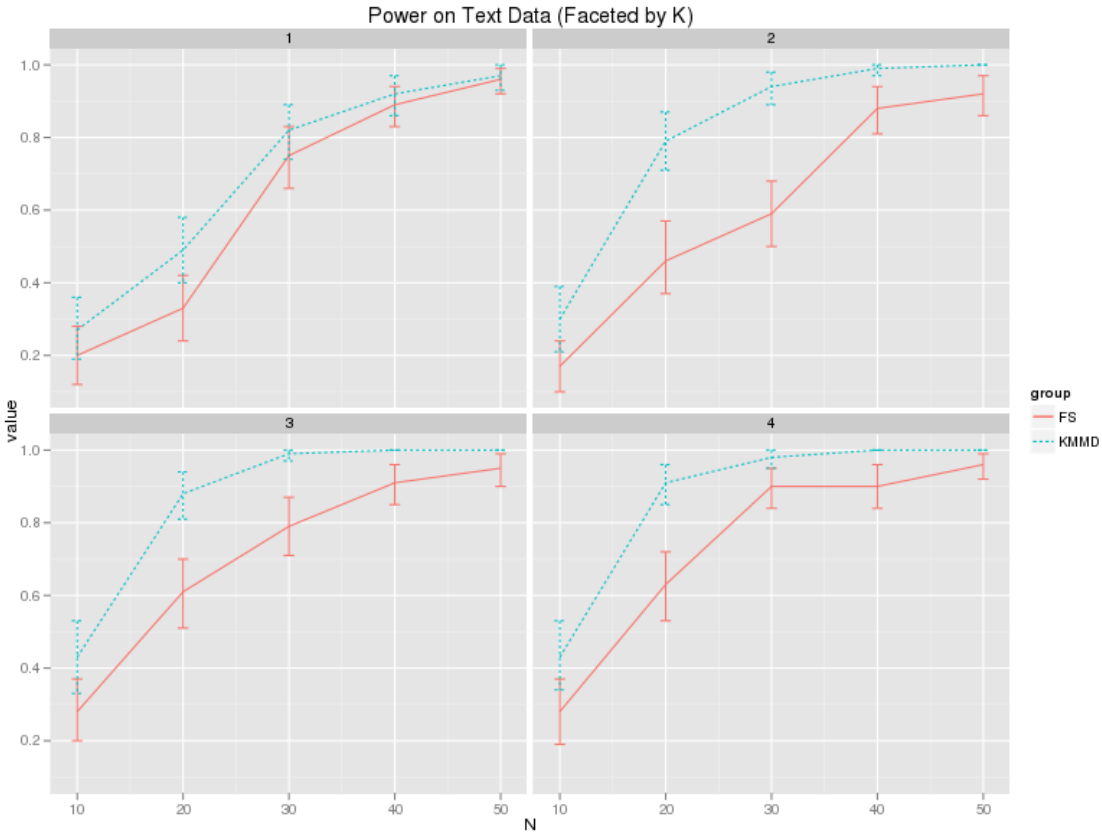


Figure 4.4: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; Error bars indicate 95% bootstrap confidence intervals. The MMD test is more powerful.

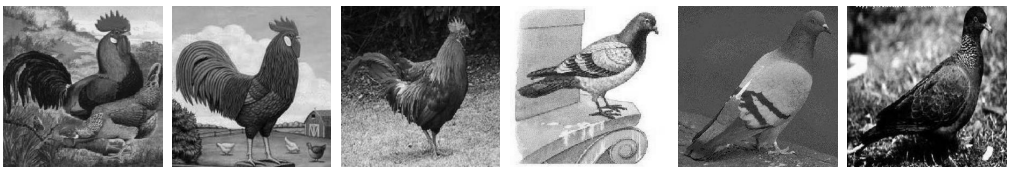


Figure 4.5: Images of roosters and pigeons for use in discrimination test.

with solution is given by

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

4.3.1 Tuning Parameters

The cost parameter C controls the complexity of the prediction function, and ε controls the leniency of the loss function. These parameters are typically chosen via cross-validation over a grid of choices. However, due to computational considerations, we mostly fix these values at $C = 1$ and $\varepsilon = .1$. In subsection 4.6.2, we describe a sample of string data from Twitter. In figure 4.1 we demonstrate the statistical power of the test for the Twitter data over a grid of SVM parameters. It is clear that these parameters play a *crucial* role in the operating characteristics of the resultant test.

We emphasize that the proper strategy is to conduct the search anew for each statistic calculation in each permutation. That is, use cross-validation to find the best performing pair (C_0, ε_0) in terms of the Friedman Statistic. For each permutation i, use cross-validation over the same grid to find the ith pair (C_i, ε_i). This ensures symmetry of protocol and enforces that the test have level α. The grid search likely maximizes the power over the set of tuning parameters: it is hoped that the search benefits the actual labeling of values by at least as much as it does permuted labels.

4.3.2 Equivalence to Permutation t-test

Theorem 4.1. *The Friedman Test paired with support vector regression generalizes the two-sample permutation t-test. Namely, the two procedures are equivalent with univariate data and a linear kernel.*

Proof.

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

since we have univariate data and a linear kernel. Therefore, the SVM score is simply a linear 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,

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

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. □

Despite the slight dependence, the randomization distribution of the t-statistic converges weakly to the normal distribution [**?**]. Anonymous and Anonymous [**?**] use Stein's Method of exchangeable pairs [**?**, **?**] to prove a conservative $\mathcal{O}(N^{-1/4})$ rate of convergence in Kolmogorov-Smirnov distance between the two distributions. The problem is not as straightforward as in the i.i.d. case because the permutation

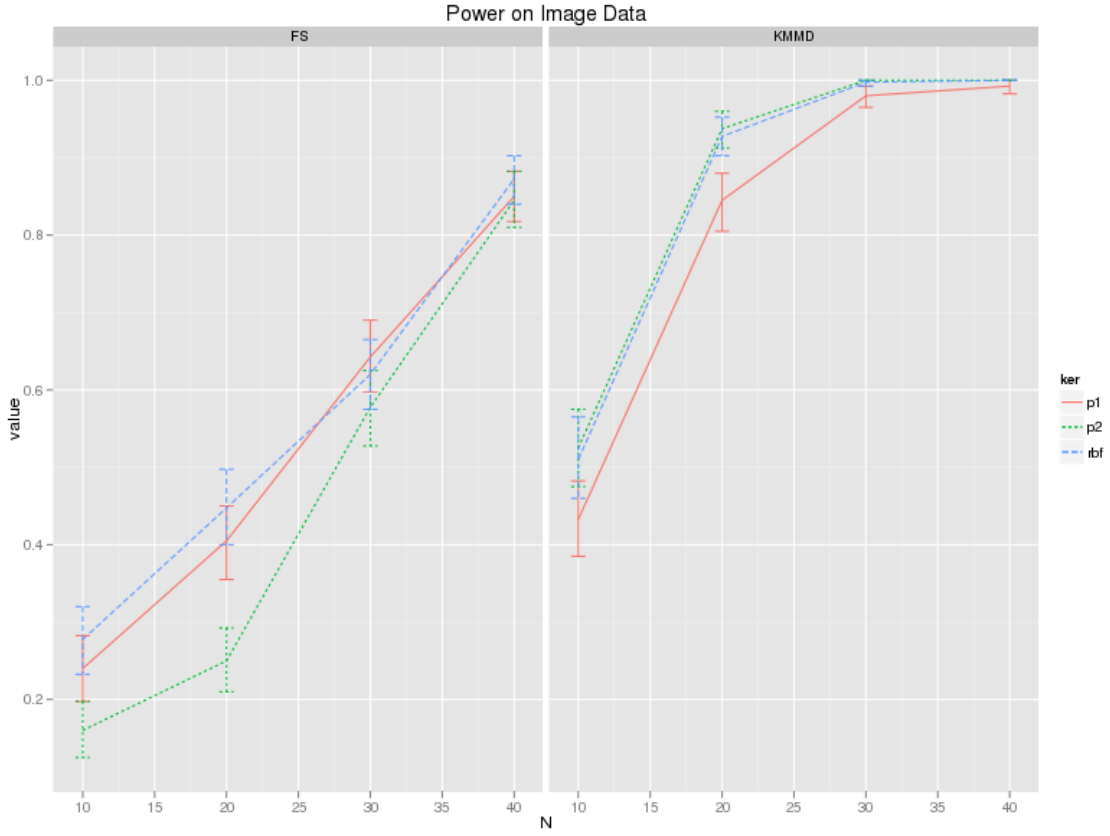


Figure 4.6: p1: linear kernel; p2: inhomogeneous degree 2 polynomial kernel; rbf: radial basis function kernel; Error bars indicate 95% bootstrap confidence intervals.

structure induces a global—though mild and diminishing in sample size—negative dependence in the data. This dependence thwarts traditional Fourier-analytic techniques yet can be managed via Stein's eponymous method of proof.

4.4 Maximum Mean Discrepancy

Gretton et al. [?, ?, ?, ?] 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 4.2. *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 its 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.

4.5 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 4.1, 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 $D \in \{1, 5, 10\}$.

For the FS, the SVM cost parameter C is fixed at 1, with $\varepsilon = .1$. We choose the RBF kernel hyperparameter via estimation techniques such as those implemented in

the **sigest** function of **kernlab** [**?**]. Due to the different scales, it is not easy to see that FS and $|T|$ in fact have the same distribution. 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.

4.6 Experiments

4.6.1 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 $\Delta \in \{0, .25, \ldots, 1.5\}$ in figure 4.3.

For FS and MMD, we used the the RBF kernel with the same method of hyperparameter estimation. In this simple setting, all three methods perform similarly with perhaps a small edge to MMD.

4.6.2 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** [**?**]. 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 [**?**] with $k = 4$ as our kernel for both the FT and MMD. Thus, each string is mapped to a 27^k dimensional feature vector of counts of the number of k letter and space combinations. 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. 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.

4.6.3 Image Data

We consider the task of discriminating between images of roosters and pigeons from the Caltech 101 Object Categories dataset [**?**]. 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 monochrome bitmap 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.

Again, MMD performs better. However, it appears that the linear kernel performs significantly worse for the MMD than for the FS. This could reflect a difference in the function classes over which each technique operates.

4.7 Extensions

4.7.1 Heterogeneous Data

This procedure extends naturally to the heterogeneous data setting via multiple kernel learning (MKL) [**?**, **?**]. Qiu et al. [**?**] develop MKL for support vector regression. Given j different data modalities, it suffices to match a kernel K_i to each—or perhaps more than one kernel for each data source, so as to better target specific features. The semidefinite programming approach (SDP) to MKL finds the best linear combination $K = \sum_{i=1}^{j} \mu_i K_i$ for some relevant objective function. For computational reasons, the best non-negative linear combination is frequently sought, as this yields a simpler quadratically constrained quadratic program (QCQP).

4.7.2 Missing Data

If we further consider entire missing modalities (e.g. one sample is missing some biometric reading), Poh et al. [**?**] develop the *neutral point substitution* technique to allow substitution of the missing modality with a new kernel that is *unbiased* with

regard to the classification at hand. This allows for full use of both modalities that are present for all samples as well as those that are present only for a subset of the samples and effective utilization of all the data in the training set. Panov et al. [**?**] modify the NPS method to allow for missing modalities in the test set.

4.7.3 Theoretical Guarantees

Having proved a bound in the univariate data, linear kernel case by constructing an exchangeable pair, Anonymous and Anonymous [**?**] use simulations to suggest that the same pair is likely to yield success in more general settings: the key *approximate regression condition* holds more universally for multivariate data, a non-linear kernel, and a combination of the two settings. Further simulations demonstrate that the $\mathcal{O}(N^{-1/4})$ rate of convergence does not appear to be tight and a more typical $\mathcal{O}(N^{-1/2})$ is within reach.

A rate of convergence result with known constant allows for a single calculation of the Friedman statistic—rather than the N_{perm} required for randomization-based inference. Theoretical inference could be done on the limiting distribution, with error characterized by the proven bound. This large savings in computation comes only at the known cost of the limiting distribution approximation, which falls rapidly in sample size.

4.8 Discussion

We have tested a two-sample testing method of Friedman's [**?**] with a particular choice of learning algorithm—support vector regression. This Friedman Test can be seen as a generalization of the celebrated permutation t-test, or randomization test. Without tuning, performance is competitive in some settings with the MMD test. Simulations suggest that more powerful tests may be achieved with the added complexity of tuning—at some computational cost. Further work is required to determine a good set of heuristic choices for the SVM tuning parameters.

Modern data sources often consist of different modalities. Wireless sensor networks (including cellular phones) are deployed to collect large quantities of *diverse* data. These networks may be heterogeneous, with newer and upgraded hardware logging novel sources of data. Because Friedman's idea leverages *any* learning algorithm, we can at present easily incorporate extensions such as both the treatment of heterogeneous data *and* an allowance for missing data modalities. Future developments in regression and classification can be incorporated to advance the state-of-the-art in two-sample testing.

Chapter 5

Multiple Kernels

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

5.1 Introduction

Introduction to MKL, theory of MKL, and optimization problems/tuning parameters.

5.2 Simulations

5.2.1 Vectorial Data Mixture Distribution

Let's look at mixtures of MVN. Let

$$\Sigma_1 = \begin{bmatrix} 1^2 & 0 \\ 0 & 1^2 \end{bmatrix}$$

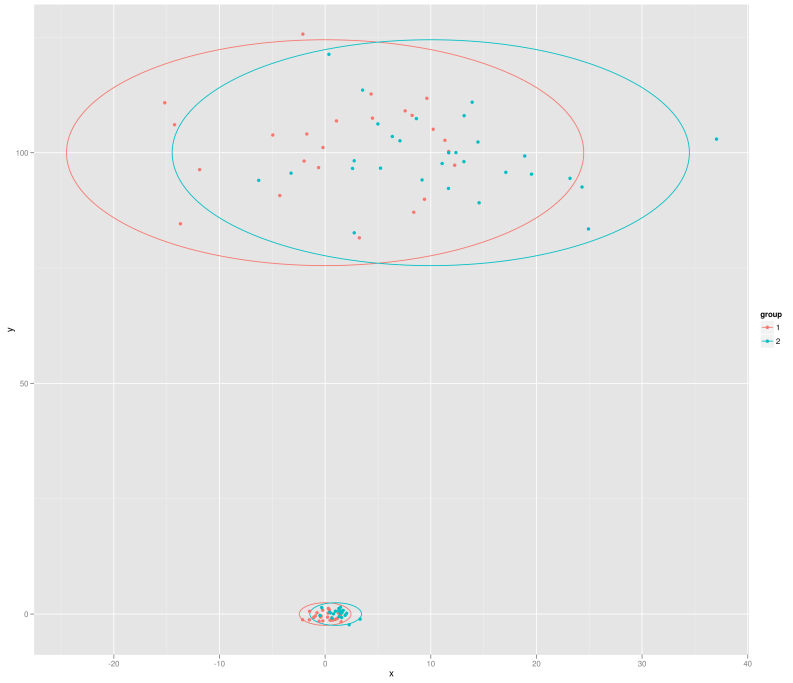
$$\Sigma_2 = \begin{bmatrix} 10^2 & 0 \\ 0 & 10^2 \end{bmatrix}$$

$$\mu_1(\delta_1) = [\delta_1, 0]^T$$

$$\mu_2(\delta_2) = [\delta_2, 100]^T.$$

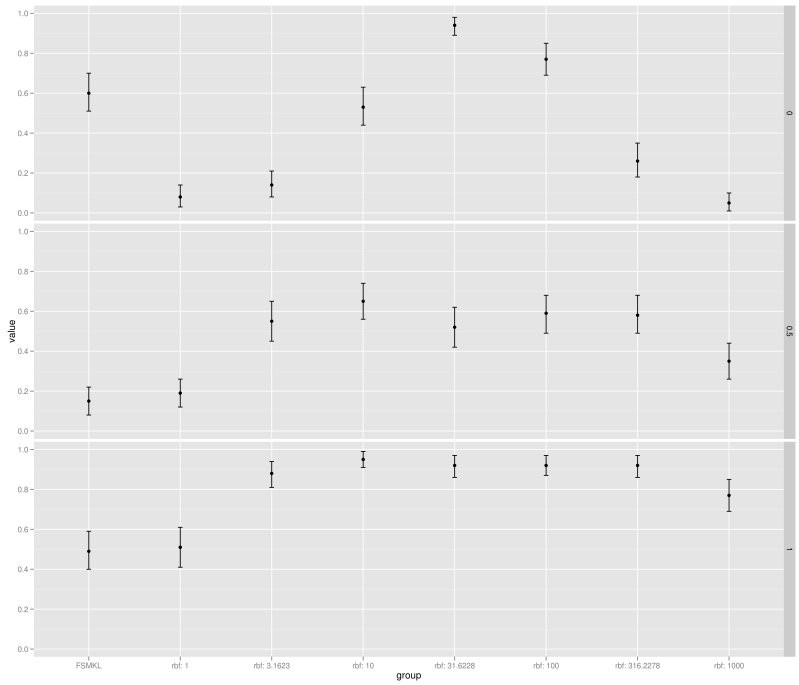
Let d_1 be a mixture distribution of $\mathcal{N}_2([0,0]^T, \Sigma_1)$ with probability p and $\mathcal{N}_2([0,100]^T, \Sigma_2)$ with probability $1-p$. Let d_2 be a mixture distribution of $\mathcal{N}_2([1,0]^T, \Sigma_1)$ with probability p and $\mathcal{N}_2([10,100]^T, \Sigma_2)$ with probability $1-p$. Note that $\delta_1 = 1$ and $\delta_2 = 10$ were chosen to be one standard deviation away (on the x-axis, see $(\Sigma_1)_{1,1}$ and $(\Sigma_2)_{1,1}$). In all the simulations, we draw $n = 50$ samples from each mixture distribution, d_1 and d_2. We take the mixture probability $p = .5$ unless otherwise specified.

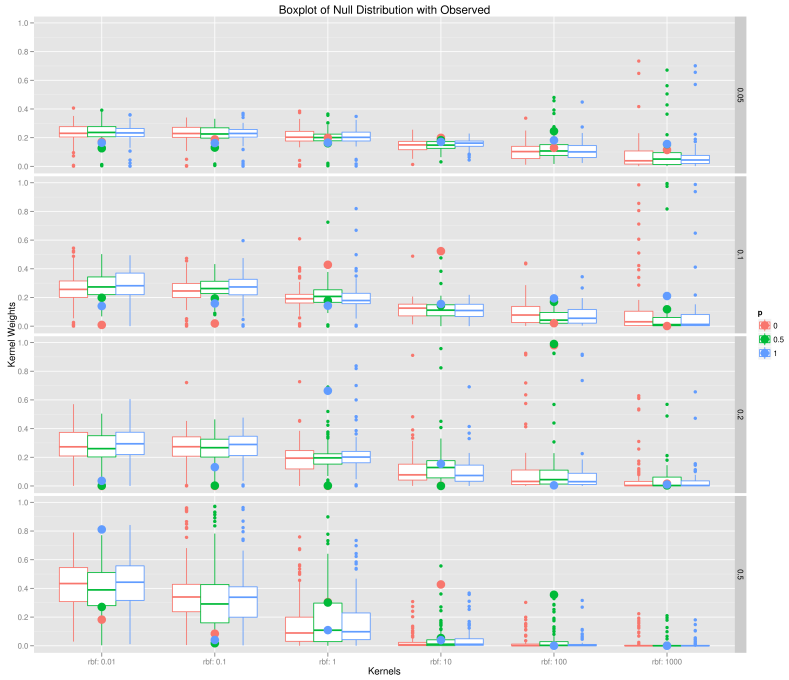
Here is a plot of the 95% confidence ellipses of the mixture distributions:



Here we plot the average power and bootstrap 95% confidence intervals (100 simulations) for each RBF kernel individually and MKL on all of them, faceted on the mixture probability. So for $p = 0$, we have all the weight on $\mathcal{N}_2(\mu_2, \Sigma_2)$, and for $p = 1$, we have all the weight on $\mathcal{N}_2(\mu_1, \Sigma_1)$. Since the latter is on a smaller scale, we expect the smaller width RBF kernels to do better. We take $C = 1$, and the widths to be from the middle run of the last section: The smaller distribution ($p = 1$) has higher power for smaller kernels, but the MKL power is about the same as compared with the $p = 0$ case. Both outperform the mixed setting.

Here are the null distributions of the weights and the observed weights:





5.2.2 Heterogeneous Data

I created a test heterogeneous dataset that has two components to it: DNA string and univariate. I created the DNA data via a Markov chain because I wanted the joint

distribution of 2-grams to be different from the product of two 1-grams. I wanted this dependence so that I could later pick out differences between two groups with a 2-spectrum kernel instead of a 1-spectrum kernel. For the first group, I randomly picked a starting string according to the stationary distribution and then proceeded via the transition probabilities. For the second group, I had independent draws from the stationary distribution.

The univariate data is simply $\mathcal{N}(\{-\mu, \mu\}, 1)$, and there are 20 samples in each group. I fixed the kernel training parameter values and trained a convex combination of 5 kernels on the data via MKL: Gaussian RBF kernels with parameter .1, .2, .5, and 1, and a 2-spectrum kernel (later I do want to test that the 2-spectrum is required in this case over the 1-spectrum because of the Markov chain construction of the data, but I had to add pre-processors and I'm still not too familiar with shogun yet).

I looked at the kernel weights (so no Friedman test yet) over $\mu = .5, .6, \ldots, 2.9, 3$. The weights on, say, the 2-spectrum kernel aren't monotonically decreasing because of sampling variance: I only ran each of these once. I have attached MKL1.png to show that the signal in the DNA data outweighs (in the sense of yielding a higher MKL weight) the signal in the univariate part of the data until $\mu = 1.5$ or so. At that point, the RBF1 kernel takes over.

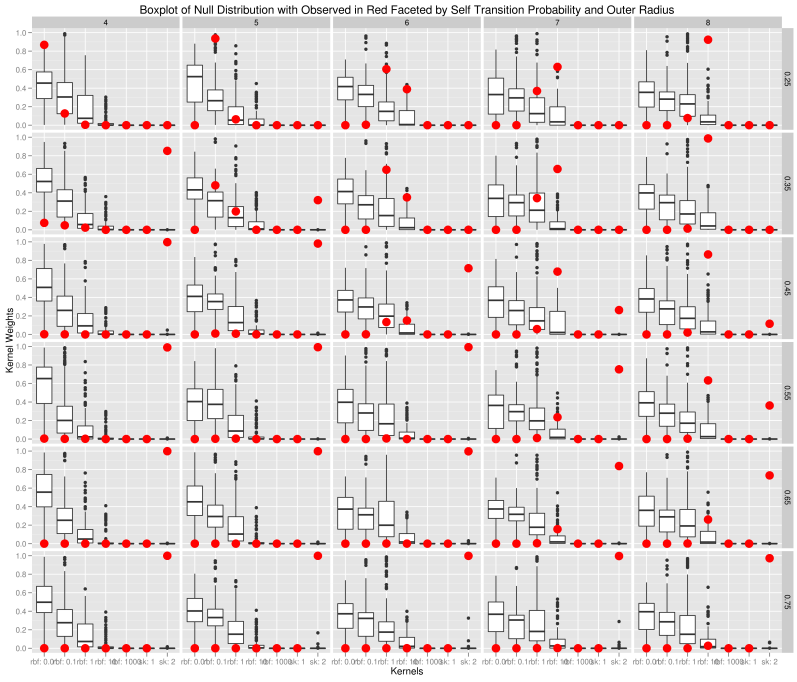
The pictures from our meeting weren't that compelling, so I've been looking for better examples. I decided a reasonable one was the Christmas star example on page 1548 of http://eprints.pascal-network.org/archive/00002269/01/sonnenburg06a.pdf

Imagine an outer star (radius 4, 5, 6, 7, 8) with an inner star (radius 4) inside. The different stars correspond to different labelings in the classification problem. I looked at 5 kernels individually (RBF with width .01, .1, 1, 10, 1000) and the MKL combination of all of them).

Here's a heterogeneous data example. I added string (DNA) data to the Christmas star example from last week. So each point is (l_i, x_i, y_i, s_i) = (label, Christmas star x-coordinate, Christmas star y-coordinate, DNA sequence). I generated the DNA by first picking a random (Poisson) length, sampling the starting point from the stationary distribution (all 1/4) of the Markov chain, and then picking according to the transition matrix M, where $M_{i,i} = s$ (for self transition probability) and

$M_{i,j} = \frac{(1-s)}{3}$ for $i \neq j$.

Here are the MKL weights:



And the power:

The effect of C:

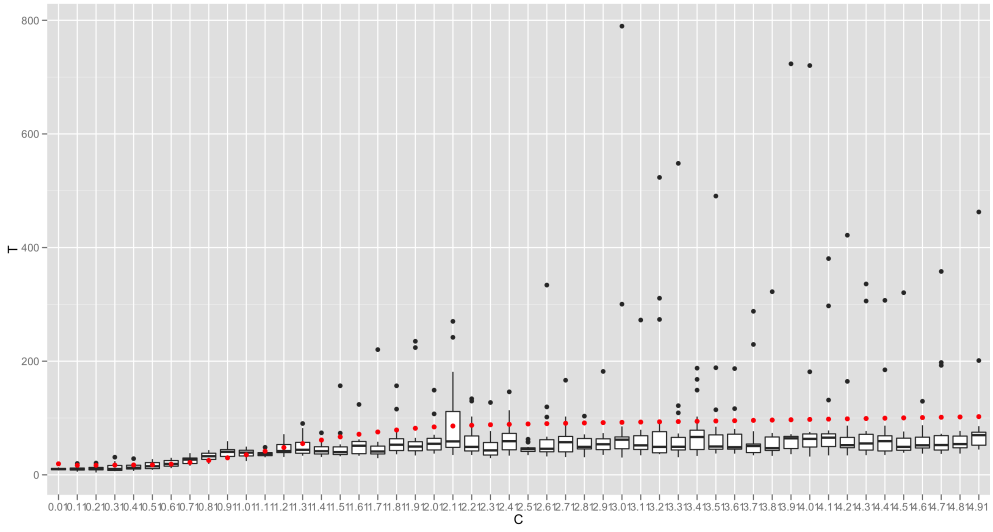
C clearly has an effect, especially if you get it very wrong (¿.8). I'm a little disappointed in the performance of MKL but pleased that it does pick out the right structure in the data. I'd like to say that if you know the structure of the data a priori and use that in the kernel, you will obviously get the best performance. It seems like you give up a lot of performance using MKL (maybe too much to justify its convenience), only doing better than the worst choices of kernels given.

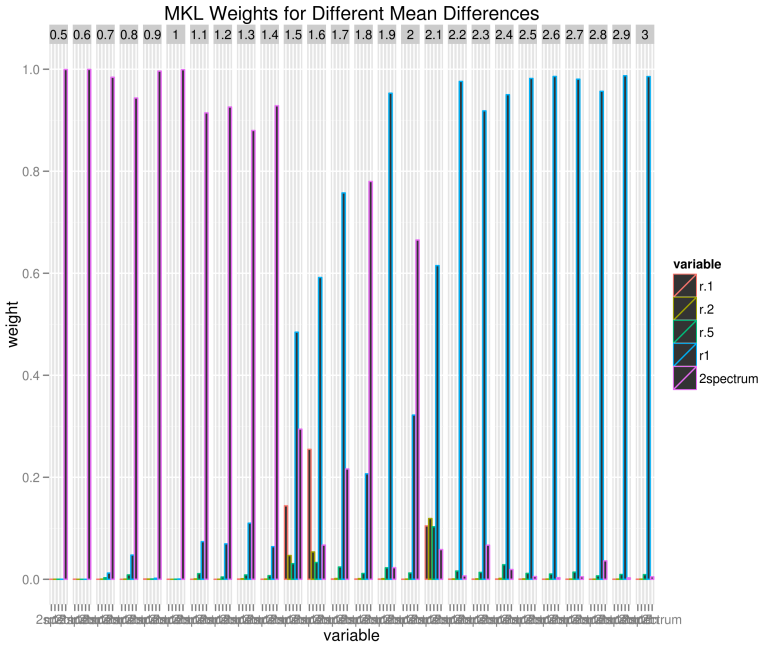
MKL can pick out the structure of the data:

5.3 Wine Example

TODO


Power (Christmas Star + DNA Example, C = .1, n = 50)
4
5
6
0.25
0.4
0.55
0.7
Power
Kernel
FSMKL
rbf: 1
rbf: 10
sk: 1
sk: 2
group


T
C


MKL Weights for Different Mean Differences
0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
1.0
0.8
0.6
0.4
0.2
0.0
weight
variable
variable
r.1
r.2
r.5
r1
2spectrum

References

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

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

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

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

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

[6] 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.

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] 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.

[19] 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.

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

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

[22] Q.M. Shao. An explicit berry-esseen bound for students t-statistic via steins method. *Steins Method and Applications (AD Barbour and LHY Chen eds). Lecture Notes Series, Institute for Mathematical Sciences, NUS*, 5:143–155, 2005.

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

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