CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

AN INVESTIGATION OF DISTANCE-BASED STATISTICAL METHODS IN HIGH DIMENSIONS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Mathematics

by

Arman Terzyan

August 2020

The thesis of Arman Terzyan is approved:

Alexander Alekseenko, Ph.D. Date

Ali Al-Sharadqah, Ph.D. Date

Mark Schilling, Ph.D., Chair Date

California State University, Northridge

Table of Contents

List of Tables

List of Figures

ABSTRACT

AN INVESTIGATION OF DISTANCE-BASED STATISTICAL METHODS IN HIGH DIMENSIONS

By

Arman Terzyan

Master of Science in Mathematics

When dealing with high dimensional data, common statistical tools often become ineffective. This thesis examines causes of these issues through an exploration of high dimensional spaces. We present some counterintuitive mathematical and statistical results from a geometric view. These results are exploited to modify an existing classification method and also introduce a new test statistic for the general distribution free two-sample problem. Other applications are briefly considered.

Chapter 1

Introduction

1.1 Motivation

With the abundance of data in the world today, the need for analysis is pressing. Classical tasks in statistics such as classification and the two-sample testing problem are commonplace in many different areas of research.

The goal of classification is to place an incoming observation into its respective category based off several variables pertaining to that observation. Existing methods include k-nearest neighbors, support vector machines, random forests, and more. A popular use of classification is seen in banks deciding whether or not to approve a loan through the use of numerous factors including the customer's age, gender, salary, credit history, etc. Another example is classifying an incoming email as spam or not spam by observing the frequency of a particular word.

In the same vein, the two-sample problem tests whether two independently drawn samples came from the same underlying distribution. Applications of the two-sample problem are plenty. In academia, one could test whether there is a significant difference in performance on a standardized test between the classes of two instructors. A more complex example would be testing for differences in medical profiles in patients with or without a newly introduced medication. The standard methods for performing these tasks often struggle in situations with a large number of variables. This is known as the "curse of dimensionality".

1.2 Structure of Thesis

In this thesis, following a look at the geometric properties of high dimensional spaces, we propose a new procedure for distance-based classification as well as a new test statistic for the two-sample problem in these scenarios. Chapter 2 provides a geometric study of high dimensions where we highlight the bizarre behavior that is exhibited both mathematically and statistically. In Chapter 3, we describe the task of classification in detail, the proposed new procedure, and the primary research paper that motivated it. The two-sample problem is formally introduced in Chapter 4. As this is a very popular area of study, many valuable and motivating articles are summarized. We present our own test statistic and evaluate its performance against several existing tests. In addition, we consider the use of these principles in the unsupervised problem of clustering as well as goodness of fit testing.

Chapter 2

Characteristics of High Dimensions

2.1 Prelude

To illustrate a preliminary paradoxical result, consider fitting a unit square inside the unit circle with both centered at the origin as in Figure 2.1.

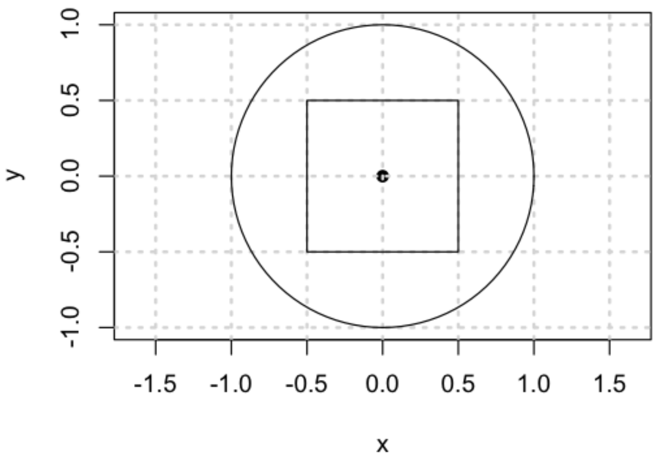


Figure 2.1: Unit square inscribed in unit circle.

The distance from any vertex of the square to the origin is

$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2} = \sqrt{2\left(\frac{1}{2}\right)^2} = \frac{\sqrt{2}}{2} < 1.$$

confirming what is shown in Figure 2.1 with the entire square lying inside the circle.

If we move into three dimensions, the square becomes a cube while the circle become a sphere. The entire unit cube still fits inside the unit sphere as the distance from any vertex of the cube to the origin is

$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2} = \sqrt{3\left(\frac{1}{2}\right)^2} = \frac{\sqrt{3}}{2} < 1.$$

Now consider the analogous situation as the dimension d grows. The general distance from the origin to any vertex is $\frac{\sqrt{d}}{2}$. Hence when $d = 4$ the edges of the hypercube touch the hypersphere and when $d \geq 5$ the edges begin poking out of the hypersphere. This behavior is illustrated as a projection into two dimensions in Figure 2.2.

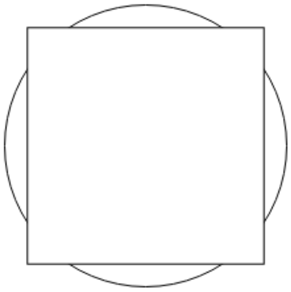


Figure 2.2: High dimensional projection of hypercube peeking out of hypersphere.

In addition to its edges poking out, the hypercube has the bulk of its volume outside of the hypersphere as the dimension grows. We can show this by calculating the probability that a random vector from the unit hypercube falls outside the unit hypersphere. We work with a d-dimensional uniform distribution with support $[-0.5, 0.5]^d$. For a vector $\mathbf{U} = (U_1, U_2, ..., U_d)$ from this distribution to lie outside the unit hypersphere, its distance from the origin must be greater than one. Thus we want to evaluate

$$P\left(\sqrt{\sum_{i=1}^{d} U_i^2} > 1\right) = P(\sum_{i=1}^{d} U_i^2 > 1).$$

It is useful to standardize the above expression thus the mean and variance of the sum of d squared uniform random variables is required. With the given support of [-0.5,0.5], $E(U_i) = 0$. To find $E(\sum_{i=1}^{d} U_i^2)$, we use the fact that $\text{Var}(U_i) = E(U_i^2) - E(U_i)^2 = E(U_i^2)$.

Hence

$$\begin{aligned}E(\sum_{i=1}^{d} U_i^2) &= \mathrm{Var}(\sum_{i=1}^{d} U_i)\\ &= \sum_{i=1}^{d} \mathrm{Var}(U_i)\\ &= \sum_{i=1}^{d} \frac{1}{12}\\ &= \frac{d}{12}.\end{aligned}$$

Calculating $\mathrm{Var}(\sum_{i=1}^{d} U_i^2)$ follows similarly below:

$$\begin{aligned}\mathrm{Var}(\sum_{i=1}^{d} U_i^2) &= \sum_{i=1}^{d} \mathrm{Var}(U_i^2)\\ &= \sum_{i=1}^{d} (E(U_i^4) - E(U_i^2)^2)\\ &= dE(U_i^4) - \frac{d}{144}.\end{aligned}$$

We have $E(U_i^4) = \int_{-0.5}^{0.5} u^4 \cdot 1du = \frac{1}{80}$. Thus $\mathrm{Var}(\sum_{i=1}^{d} U_i^2) = d(\frac{1}{80} - \frac{1}{144}) = \frac{d}{180}$. After standardizing we have

$$P\left(\frac{\sum_{i=1}^{d} U_i^2 - \frac{d}{12}}{\sqrt{\frac{d}{180}}} > \frac{1 - \frac{d}{12}}{\sqrt{\frac{d}{180}}}\right).$$

Applying the Central Limit Theorem, we obtain the approximation

$$P\left(Z > \frac{1 - \frac{d}{12}}{\sqrt{\frac{d}{180}}}\right).$$

Results for various dimensions are shown in Table 2.1.

Note that at 12 dimensions, the amount of volume in the hypercube that lies inside the

d	$P(\sum_{i=1}^{d} U_i^2 > 1)$
10	0.23975
12	0.5
15	0.8068
20	0.9773
25	0.9982

Table 2.1: Probability content outside hypersphere in growing dimensions.

hypersphere is approximately equal to the amount outside the hypersphere. After passing 20 dimensions, almost all probability content falls outside the hypersphere. This counterintuitive behavior is indicative of the sort of phenomena that can create challenges for statistical and probabilistic analysis in high dimensions.

2.2 Behavior of the Standard Uniform Distribution in High Dimensions

Suppose a sample of size five is taken from a ten dimensional uniform distribution, with support $[0, 1]^{10}$. The vastness of space in high dimensions pushes the data out into the fringes of the distribution. This can be seen using the *parallel coordinate* plot shown in Figure 2.3.

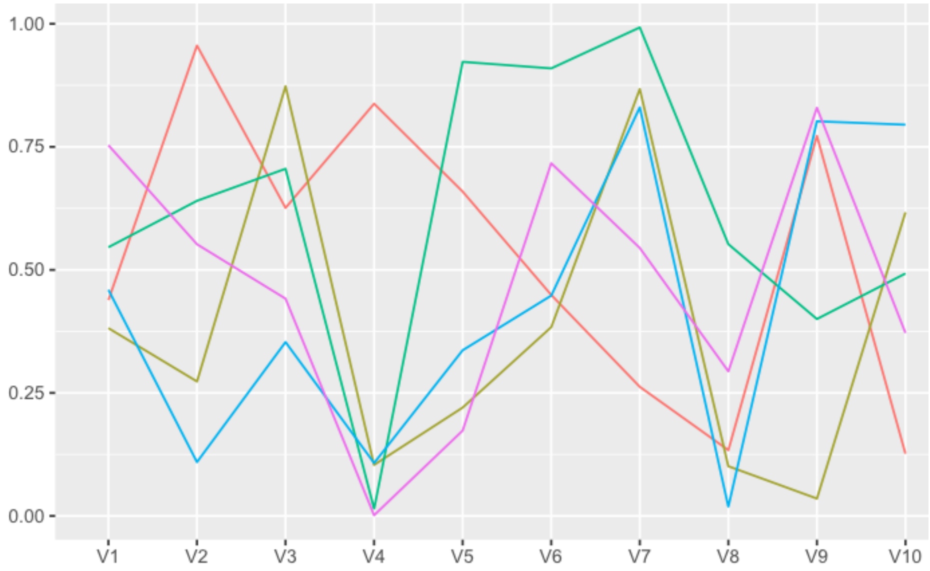
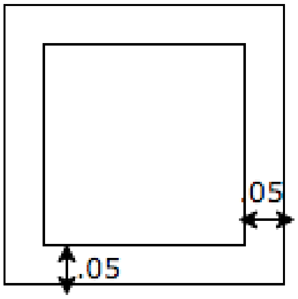


Figure 2.3: Parallel coordinates with five observations from ten dimensional uniform.

Each observation, represented by a single color, is a random point sampled from the above distribution. Notice that all the observations have at least one coordinate close to either zero or one, implying that the point is close to the boundary of the hypercube $[0,1]^{10}$. In fact we can show that the probability of being near the boundary becomes very high as the dimension grows:

Let's start in two dimensions. Consider the probability of a randomly selected point from a unit square falling within 0.05 of the boundary. Thus we simply want the area of the frame in Figure 2.4, $1-(1-2(.05))^2 = 0.01$. In ten dimensions, the volume of the corresponding frame is $1-(1-2(.05))^{10} = 0.65$. When we reach fifty dimensions, the volume of the frame is $1-(1-2(.05))^{50} = 0.995$. Thus the probability of a 50-dimensional uniform random vector falling within 0.05 of the boundary is a near certainty. A number of statistical procedures that work well in lower dimensional spaces are much less effective in high dimensions due to "boundary effects" of the sort that are exhibited here.

Figure 2.4: Frame in two dimensions.



2.3 The Normal Distribution in High Dimensions

Now let's explore the the most important distribution in statistics–the normal. The density for the d-dimensional standard normal is:

$$f(\mathbf{Z}) = f(z_1, \cdots, z_d) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{\sum_{i=1}^{d} {z_i}^2}{2}).$$

Since the univariate normal distribution is bell-shaped, intuition would suggest that most of the data sampled from this distribution will fall near the center where $\mathbf{Z} = \mathbf{0}$ and the density is maximized with value $f(\mathbf{0}) = \frac{1}{(2\pi)^{\frac{d}{2}}}$. However, this is not the case as d grows. Surprisingly, the bulk of the data will collect in areas where the density is low. This can be seen by looking at the Euclidean distance, $||\mathbf{Z}||$, between $\mathbf{Z}$ and the origin. To quantify a value as "low", we consider the ratio of the density value at $\mathbf{Z}$, $f(\mathbf{Z})$, to the maximum density value at $\mathbf{0}$, $f(\mathbf{0})$. We show that as the dimension grows, this ratio shrinks toward zero for most $\mathbf{Z}$.

Since $||\mathbf{Z}||^2$ is the sum of squared standard normals, it is distributed as a chi-squared random variable with d degrees of freedom. Let $||\mathbf{Z}||_{.05}$ represent the square root of $\chi^2_{.05}$ (the 0.05 quantile of the chi-squared distribution) with d degrees of freedom. The region where $||\mathbf{Z}||$ exceeds $||\mathbf{Z}||_{.05}$ holds 0.95 of the probability content for $||\mathbf{Z}||$. It is shown in Table 2.2 that the ratio $\frac{f(||\mathbf{Z}||_{.05})}{f(\mathbf{0})}$ shrinks dramatically as d increases. Values of $\mathbf{Z}$ for which

d	$\|\|\mathbf{Z}\|\|_{.05}$	Density Height Ratio
1	0.063	0.99803586
2	0.320	0.95000000
5	1.070	0.56397908
10	1.985	0.13943600
25	3.822	0.00067170
50	5.896	0.00000003

Table 2.2: Density ratio $\frac{f(||\mathbf{Z}||_{.05})}{f(\mathbf{0})}$ as dimension increases.

$||\mathbf{Z}||$ exceeds $||\mathbf{Z}||_{.05}$ have even lower ratios since the density decreases moving outward.

From Table 2.2, we see that in 25 dimensions 95% of the probability mass falls in the region where the height of the density is less than 0.067% of its height at the origin. After 50 dimensions, almost all **Z** will have a value where the density is imperceptibly small.

2.3.1 Two Interesting Geometric Results

Similar to what was discussed earlier with regards to the uniform distribution, we show two remarkable geometric results for multivariate normal samples in high dimensions.

2.3.1.1 Concentration in an Annulus

Consider again a d-dimensional Gaussian vector, $\mathbf{Z} = \mathbf{Z}_d$, with bounds to ensure that the probability that the distance to the origin lies in the interval (a, b) is equal to $1 - \alpha$:

$$P(a \leq ||\mathbf{Z}_d|| \leq b) = 1 - \alpha,$$

where α is small. That is, we wish to determine an annulus that has a high probability of containing **Z**. We work with the square of this interval:

$$P(a^2 \leq ||\mathbf{Z}_d||^2 \leq b^2) = 1 - \alpha.$$

Since $||\mathbf{Z}_d||^2$ is distributed as χ^2_d we can normalize:

$$P\left(\frac{a^2 - d}{\sqrt{2d}} \leq \frac{||\mathbf{Z}_d||^2 - d}{\sqrt{2d}} \leq \frac{b^2 - d}{\sqrt{2d}}\right) = 1 - \alpha.$$

By the Central Limit Theorem, we can set

$$\frac{a^2 - d}{\sqrt{2d}} \approx Z_{-\frac{\alpha}{2}} \qquad \frac{b^2 - d}{\sqrt{2d}} \approx Z_{\frac{\alpha}{2}}$$

for large d. Solving for the bounds a and b we get:

$$a \approx \sqrt{d + Z_{-\frac{\alpha}{2}}\sqrt{2d}} \qquad b \approx \sqrt{d + Z_{\frac{\alpha}{2}}\sqrt{2d}}.$$

We see that for large d, both a and b are close in a relative sense. Therefore, surprisingly, the bulk of the probability mass lies in a thin annulus. Let's inspect visually what happens as the dimension grows with $\alpha = 0.05$, so that 95% of the probabilty mass lies within the annulus. Figure 2.5 shows cross-sections of the annulus for various d where the radius of the red circle represents a while the radius of the blue circle represents b. Figures have been rescaled in order to show the relative magnitude of the annuli.

2.3.1.2 Concentration in an Equatorial Slice

We now also show a result seemingly contradictory to the above. Not only will volume concentrate in an annulus but it will also fall into a thin slice around the equator. In fact, every equatorial slice will work! Using the term "equatorial slice" in high dimensions describes the region formed when intersecting the density with a plane that crosses through

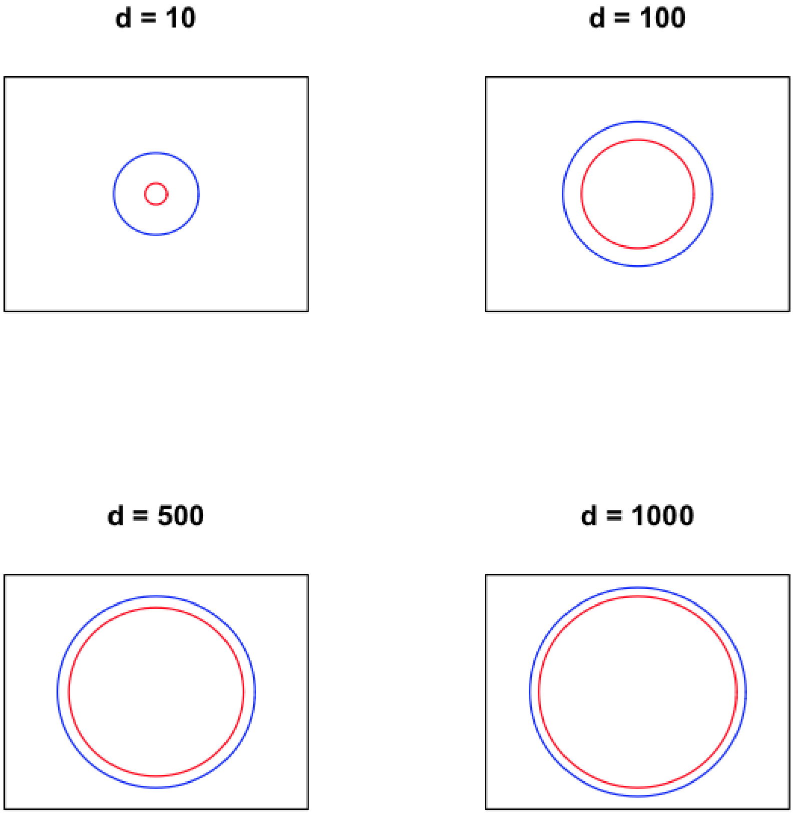


Figure 2.5: As dimension increases, 95% of the probability mass lies within a thin annulus.

the origin.

Let $\mathbf{Z}$ be a d-dimensional random normal vector. Let $\mathbf{U} = \frac{\mathbf{Z}}{||\mathbf{Z}||}$, to transform $\mathbf{Z}$ onto the unit d-sphere. We show for any thin margin $\delta > 0$, $P(|U_1| \leq \delta) \rightarrow 1$ as $d \rightarrow \infty$. We have

$$\begin{aligned} P(|U_1| \leq \delta) &= P(U_1^2 \leq \delta^2) \\ &= P(\frac{Z_1^2}{||\mathbf{Z}||^2} \leq \delta^2) \\ &= P\left(\frac{Z_1^2}{Z_1^2 + \displaystyle\sum_{i=2}^{d} Z_i^2} \leq \delta^2\right) \\ &= P\left(\sum_{i=2}^{d} Z_i^2 \geq \frac{Z_1^2(1-\delta^2)}{\delta^2}\right) \\ &= P\left(\frac{\sum_{i=2}^{d} Z_i^2}{Z_1^2} \geq \frac{1-\delta^2}{\delta^2}\right) \\ &= P\left(\frac{\frac{\sum_{i=2}^{d} Z_i^2}{d-1}}{\frac{Z_1^2}{1}} \cdot \frac{d-1}{1} \geq \frac{1-\delta^2}{\delta^2}\right). \end{aligned}$$

Note $Z_1^2 \sim \chi_1^2$ and $\displaystyle\sum_{i=2}^{d} Z_i^2 \sim \chi_{d-1}^2$. Thus the ratio of $\frac{\sum_{i=2}^{d} Z_i^2}{d-1}$ to $\frac{Z_1^2}{1}$ follows an $F_{d-1,1}$ distribution. We use the fact that this is equivalent to the distribution of $\frac{1}{t_{d-1}^2}$ where t_{d-1} is a Student's t with $d-1$ degrees of freedom. Thus we proceed as follows:

$$\begin{aligned} P\left(\frac{\frac{\sum_{i=2}^{d} Z_i^2}{d-1}}{\frac{Z_1^2}{1}} \cdot \frac{d-1}{1} \geq \frac{1-\delta^2}{\delta^2}\right) &= P\left(F_{d-1,1} \geq \frac{1-\delta^2}{(d-1)\delta^2}\right) \\ &= P\left(\frac{1}{t_{d-1}^2} \geq \frac{1-\delta^2}{(d-1)\delta^2}\right) \\ &= P\left(t_{d-1}^2 \leq \frac{(d-1)\delta^2}{1-\delta^2}\right) \\ &= 1 - P\left(t_{d-1}^2 > \frac{(d-1)\delta^2}{1-\delta^2}\right). \end{aligned}$$

By Markov's inequality we have:

$$1 - P\left(t_{d-1}^2 > \frac{(d-1)\delta^2}{1-\delta^2}\right) \leq 1 - \frac{E(t_{d-1}^2)}{\frac{(d-1)\delta^2}{1-\delta^2}}.$$

Since $E(t_{d-1}^2)$ is equivalent to $\mathrm{Var}(t_{d-1}) = \frac{d-1}{d-3}$. We have

$$\begin{aligned} 1 - \frac{E(t_{d-1}^2)}{\frac{(d-1)\delta^2}{1-\delta^2}} &= 1 - \frac{\frac{d-1}{d-3}}{\frac{(d-1)\delta^2}{1-\delta^2}} \\ &= 1 - \frac{1-\delta^2}{\delta^2(d-3)}, \end{aligned}$$

from which the desired result $P(|U_1| \leq \delta) \to 1$ follows as $d \to \infty$. Note that U_1 was chosen without loss of generality as any direction would work to reach this peculiar result. Observe the paradoxical nature of this result as it appears that the bulk of the probability mass is in many places at once. Even more paradoxically to what was shown in the previous section, the probability mass gathers in an annulus while also collecting near any thin slice around the "equator". It can be shown that this holds for any spherically symmetric distribution. A more exhaustive approach to these results and others similar is presented by Blum et al. (2020).

Finally, we present an additional result that, while equally as strange as those presented above, is quite useful in constructing statistical procedures for high dimensional situations, as we will show in the subsequent chapters of this thesis.

2.3.2 Vanishing Variability

An unexpected property of high dimensions is that the usual variability one expects in a random sample from a probability distribution disappears. When working with multivariate standard normal vectors, the distribution of the Euclidean distance to the origin as well as the distance between two such vectors will be shown through simple delta method calculations to approach constants.

Theorem 2.3.1 (Delta Method)**.** *Let T_d be a sequence of random variables such that*

$$\sqrt{d}\left(T_d - \theta\right) \Rightarrow N(0, \sigma^2).$$

Then

$$\sqrt{d}\Bigg(g(T_d) - g(\theta)\Bigg) \Rightarrow N(0, (g'(\theta))^2\sigma^2).$$

We apply Theorem 2.3.1 to derive two fundamental results for large dimensional statistical samples:

Theorem 2.3.2. *Let* **Z** *be a d-dimensional standard normal vector. As the dimension d of the space increases,* $||\mathbf{Z}|| = \sqrt{d} + O_p(1)$.

Proof. We apply the delta method. Let $T_d = \frac{1}{d}||\mathbf{Z}||^2$ and $g(x) = \sqrt{x}$. To find the mean $\theta = E(T_d)$, we use the fact that the mean of a χ^2_d random variable is d. Hence

$$E(\frac{1}{d}||\mathbf{Z}||^2) = \frac{1}{d} \cdot d = 1.$$

We also have

$$\begin{aligned}
\sigma^2 &= \mathrm{Var}(\sqrt{d}(T_d - \theta)) \\
&= d \cdot \mathrm{Var}(T_d - \theta) \\
&= d \cdot \mathrm{Var}\Bigg(\frac{1}{d}||\mathbf{Z}|||^2 - 1\Bigg) \\
&= \frac{d}{d^2}\mathrm{Var}\Bigg(||\mathbf{Z}||^2\Bigg) \\
&= \frac{d}{d^2}(2d) \\
&= 2.
\end{aligned}$$

The second to last line follows from the formula for the variance of a χ^2_d random variable. Finally $g'(x) = \frac{1}{2\sqrt{x}}$, so $g'(\theta) = \frac{1}{2}$. By Theorem 2.3.1 we have

$$\sqrt{d}(g(T_d) - g(\theta))$$

$$= \sqrt{d}\bigg(\frac{1}{\sqrt{d}}||\mathbf{Z}|| - 1\bigg) \to N\bigg(0, \bigg(\frac{1}{2}\bigg)^2 \cdot 2\bigg),$$

or

$$\bigg(||\mathbf{Z}|| - \sqrt{d}\bigg) \to N\bigg(0, \frac{1}{2}\bigg).$$

Therefore,

$$||\mathbf{Z}|| = \sqrt{d} + O_p(1).$$

□

Theorem 2.3.3. *Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be independent d-dimensional standard normal vectors. As d increases, $||\mathbf{Z}_1 - \mathbf{Z}_2|| = \sqrt{2d} + O_p(1)$.*

Proof. We apply the delta method once again. Let $T_d = \frac{1}{2d}||\mathbf{Z}_1 - \mathbf{Z}_2||^2$ and $g(x) = \sqrt{x}$. We proceed by calculating the mean and variance of $||\mathbf{Z}_1 - \mathbf{Z}_2||^2$. Note that $\frac{\mathbf{Z}_1 - \mathbf{Z}_2}{\sqrt{2}} \sim N(0,1)$, thus $\frac{(\mathbf{Z}_1 - \mathbf{Z}_2)^2}{2} \sim \chi_1^2$. Therefore $E(||\mathbf{Z}_1 - \mathbf{Z}_2||^2) = 2d$ and thus $\theta = E(T_d) = 1$. Also $\mathrm{Var}(\frac{\mathbf{Z}_1 - \mathbf{Z}_2)^2}{2}) = 2d$ leading to $\mathrm{Var}((\mathbf{Z}_1 - \mathbf{Z}_2)^2) = 8d$. Moving ahead we evaluate

$$\begin{aligned}
\sigma^2 &= \mathrm{Var}(\sqrt{d}(T_d - \theta)) \\
&= d \cdot \mathrm{Var}(T_d - \theta) \\
&= d \cdot \mathrm{Var}\bigg(\frac{1}{2d}||\mathbf{Z}_1 - \mathbf{Z}_2|||^2 - 1\bigg) \\
&= \frac{d}{(2d)^2}\mathrm{Var}\bigg(||\mathbf{Z}_1 - \mathbf{Z}_2||^2\bigg) \\
&= \frac{d}{4d^2}(8d) \\
&= 2.
\end{aligned}$$

Using Theorem 2.3.1 yields

$$\sqrt{d}(g(T_d) - g(\theta))$$

$$= \sqrt{d}\left(\frac{1}{\sqrt{2d}}||\mathbf{Z}_1 - \mathbf{Z}_2|| - 1\right) \rightarrow N\left(0, \left(\frac{1}{2}\right)^2 \cdot 2\right),$$

or

$$\left(\frac{1}{\sqrt{2}}||\mathbf{Z}_1 - \mathbf{Z}_2|| - \sqrt{d}\right) \rightarrow N\left(0, \frac{1}{2}\right).$$

Therefore,

$$||\mathbf{Z}_1 - \mathbf{Z}_2|| = \sqrt{2d} + O_p(1).$$

□

From these theorems we conclude that since the relative distances to the origin of d-dimensional standard normal observations are nearly constant and the pairwise distances between such observations are nearly constant, a multivariate normal sample in high dimensions will be arranged approximately as a simplex centered at the origin where the data values converge to the vertices and the edges are approximately equal in length. This is shown below in Figure 2.6, for the case $n = 4$.

Hall et al. (2005) additionally prove that under certain weak assumptions, this result holds for any high dimensional distribution. The variability from sample to sample is captured only by the rotation of the simplex. This “concentration of distance” phenomena will be exploited to develop statistical procedures effective in high dimensions in subsequent sections.

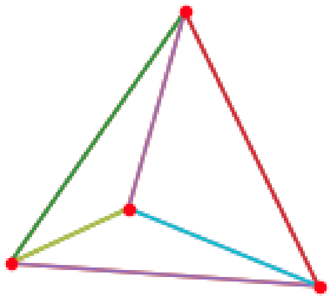


Figure 2.6: Multivariate normal sample converges into a simplex as dimension increases.

Chapter 3

Classification

Given sets of observations from different underlying populations, a classification task looks to assign a new observation to one of those populations. This is seen in Figure 3.1 where the goal would be to classify the unknown black point into either the blue or red group.

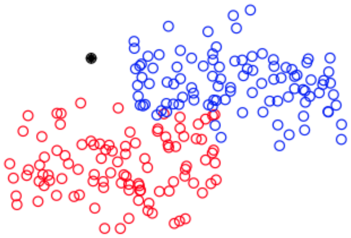


Figure 3.1: Basic classification task.

One of the well-known methods used in classification is k-Nearest Neighbors. k-Nearest Neighbors or k-NN is a nonparametric method that classifies a new observation based off the majority class from its k nearest neighbors. The distance metric often used is Euclidean distance. Figure 3.2 demonstrates how the value chosen for k could alter the results. Suppose we are trying to classify the center black point. If $k = 3$ is chosen, the unknown black point will be classified as a green triangle since it is a majority of the nearest neighbors. However, if $k = 5$ is chosen, the unknown black point will be classified as a blue circle since that is now the majority of the nearest neighbors.

k-NN unfortunately suffers from the "curse of dimensionality." The method functions under the assumption that closer points are similar to one another. But as has been demonstrated in Chapter 2, as dimension increases distance becomes essentially irrelevant since the observations become nearly equidistant from one another. An ideal algorithm would minimize misclassification even in large dimensional situations.

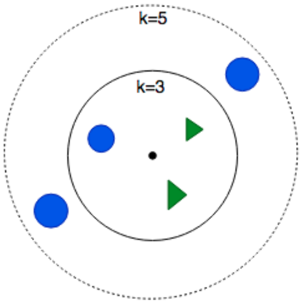
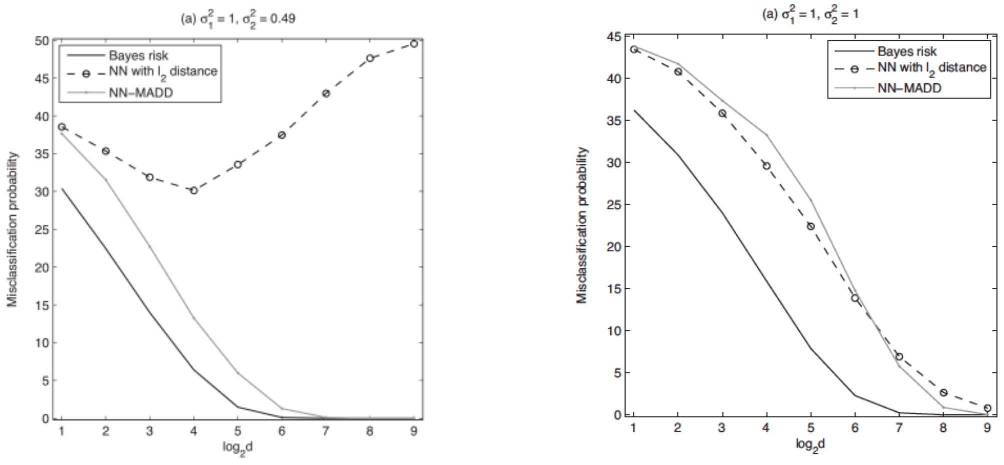


Figure 3.2: k-Nearest Neighbors

Pal et al. (2016) introduced the *Nearest Neighbor Mean Absolute Difference of Distance* (NN-MADD) method. This method exploits the fact that all pairwise distances among observations in one particular class concentrate around a single value. The steps of the algorithm to classify a single objservation $\mathbf{z}$ are as follows:

1. Take a vector $\mathbf{w} \in S_1 \bigcup S_2$ where $S_1 \bigcup S_2$ represents the training set with known underlying classes for each element. Calculate $\sum\limits_{\mathbf{t} \in (S_1 \bigcup S_2) - \mathbf{w}} ||\mathbf{z} - \mathbf{t}||_p - ||\mathbf{w} - \mathbf{t}||_p$
2. Calculate this sum for all $\mathbf{w}$ and find the minimum.
3. Classify $\mathbf{z}$ as the underlying class, S_1 or S_2, of the vector $\mathbf{w}$ that returned the minimum.

In their paper, NN-MADD was tested on the task to classify between two d-dimensional normal distributions with respective mean vectors $\boldsymbol{\mu}_{1,d} = (0, \ldots, 0)$ and $\boldsymbol{\mu}_{2,d} = (0.5, \ldots, 0.5)$ with dispersion matrices $\sum_{1,d} = \sigma_1^2 \boldsymbol{I}_d$ and $\sum_{2,d} = \sigma_2^2 \boldsymbol{I}_d$. Tests were conducted for cases of both equal and unequal dispersion matrices where $\sigma_1^2 = \sigma_2^2 = 1$ or $\sigma_1^2 = 1$ and $\sigma_2^2 = .49$. The training set was formed with 10 observations from each class while 100 observations from each class were used to form the test set. This process was repeated 250 times to get the average misclassification rate. The results, reproduced from Pal et al. (2016), are shown in Figure 3.3 for the cases of both equal and unequal dispersion matrices. Also shown are the Bayes risk to display the minimum possible error, as well as misclassifica-



(a) Different variance

(b) Same variance

Figure 3.3: Misclassification rates using MADD and traditional k-NN.

tion rates using the traditional nearest neighbors approach. In the case of unequal variance, misclassification rates for the nearest neighbors approach pivot at around 16 dimensions and begin to climb dramatically henceforth.

We present a modification of the NN-MADD by taking the “concentration of distance” fact a step further. As all pairwise distances among observations in one particular class concentrate around a single value along with the distances between observations in opposing classes, why not just use the averages of those distances? This would after all best represent the expected values for the populations. The steps of our algorithm are as follows:

1. Calculate the average of the pairwise distances within each sample S_1 and S_2 along with the average pairwise distance between the samples. These values are labeled as D_{S_1}, D_{S_2}, and D_{S_1,S_2}.

2. For each new test point, **z**, calculate its average distance to both S_1 and S_2. These values are labeled as D_{zS_1} and D_{zS_2}.

3. To classify **z** , define T_1 and T_2 as below and find which returns the minimum.

$$T_1 = |D_{zS_1} - D_{S_1}| + |D_{zS_2} - D_{S_1,S_2}|$$

$$T_2 = |D_{zS_2} - D_{S_2}| + |D_{zS_1} - D_{S_1,S_2}|$$

- If T_1 is smaller, then **z** behaves similarly to the members S_1 and hence **z** will be classified as such. The opposite holds if T_2 is smaller.

The misclassification rates of our method alongside the MADD are shown in Figures 3.4-3.6 using the same format as that of Pal et al. (2016) where the x-axis again shows the value $\log_2 d$ to simplify viewing. In addition to the cases investigated by Pal et al. (2016), we evaluate the results for equal location and larger location differences with larger differences in variance as well.

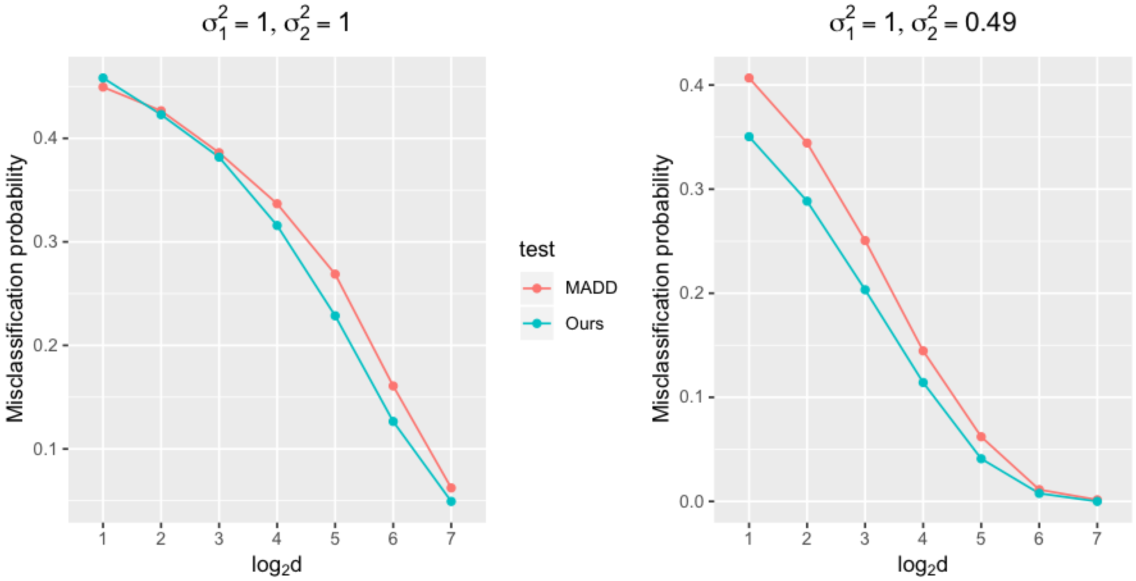


Figure 3.4: Misclassification rates where $\mu_1 = 0, \mu_2 = 0.5$.

The cases from Pal et. al (2016) are shown in Figure 3.4. A sample interpretation of the right figure is that in 8 dimensions, the MADD misclassifies around 25% of the time while our method misclassifies only 18% of the time. Though both the MADD and our method reach near 0% misclassification as the dimension surpasses 128, our method offers a fair boost up to that point.

Figure 3.5 displays the case where only a scale difference exists. The MADD appears to supersede our method at 64 and 128 dimensions but is inferior before that. In the case shown in Figure 3.6, where both a location and scale difference exists, our method again offers better results than the MADD until a high enough dimension is reached and both converge to 0% misclassification.

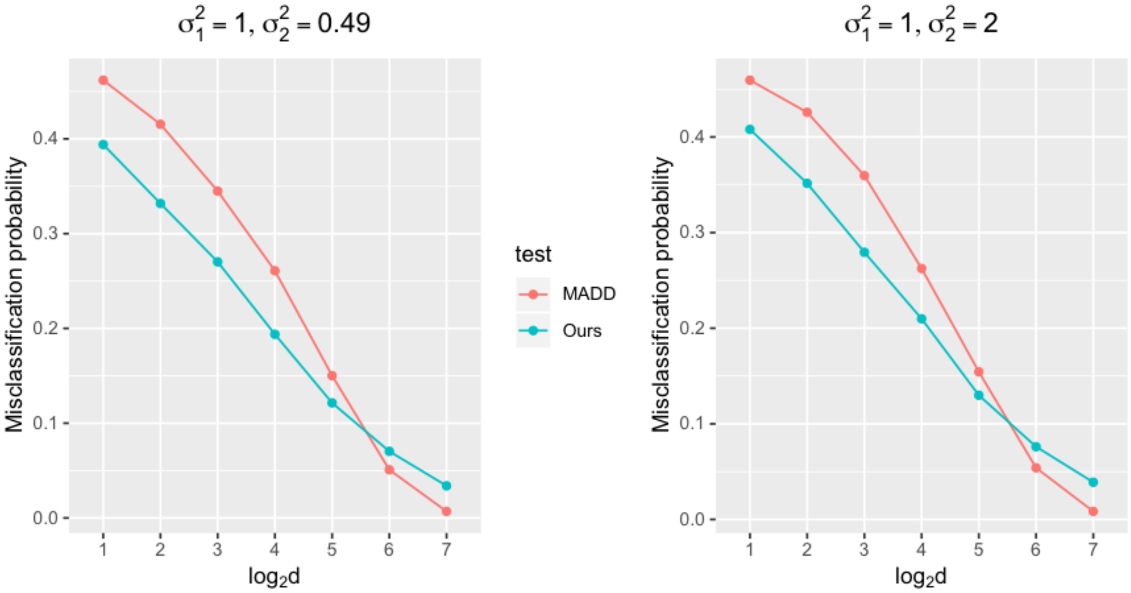


Figure 3.5: Misclassification rates where $\mu_1 = \mu_2 = 0$.

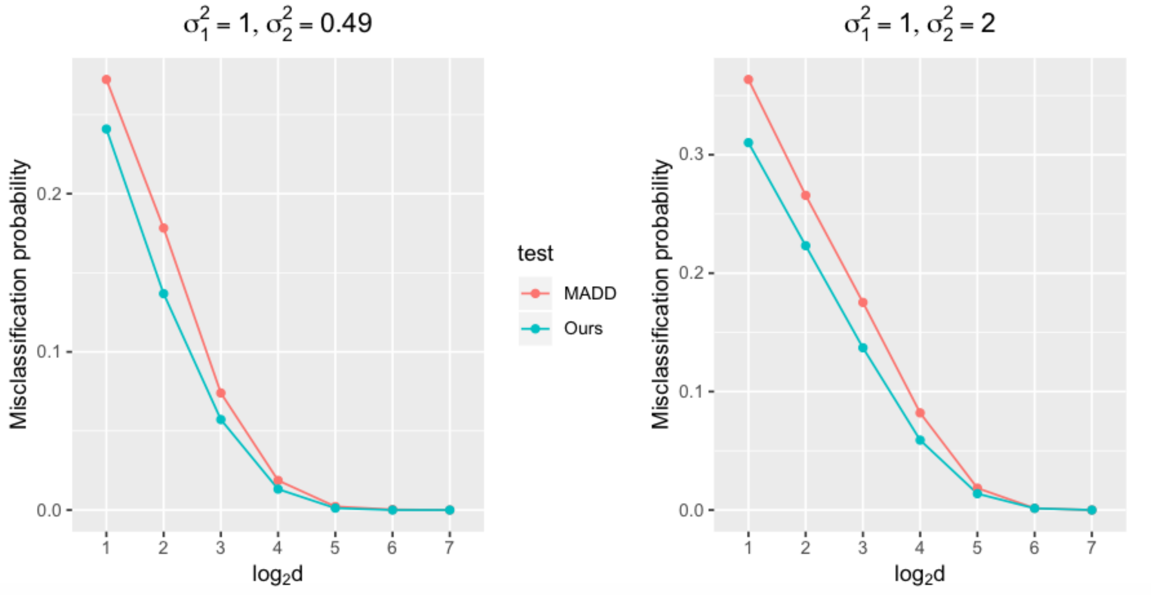


Figure 3.6: Misclassification rates where $\mu_1 = 0, \mu_2 = 1$.

Chapter 4

Two-Sample Problem

The general two-sample problem is to test the hypothesis that two samples were drawn from the same population,

$$H_0 : F = G$$

$$H_a : F \neq G$$

where F and G are absolutely continuous and completely unknown. A common approach is to utilize pairwise distances through the use of either graphs or some composition of these distances. The goal of these methods is to maximize power–the probability of rejecting the null hypothesis when it is false.

4.1 Graph-Based Approach

A graph is simply a collection of vertices connected by lines or edges. These edges are often assigned a weight which coincides with the context of the underlying problem. A path is a sequence of vertices that must be passed through to reach an ending vertex. A spanning tree is a subgraph that has the same set of vertices of the prinicpal graph wherein any two vertices are connected by one path. Friedman and Rafsky (1979) introduced a method for the two sample problem through the use of a minimum spanning tree (MST). The procedure constructs a graph with the sample points acting as vertices along with edges linking all pairs. Edge weights are assigned by the Euclidean distance between samples. The MST is a subset of the edges that minimizes the total edge weight while connecting all the vertices through a path. The test statistic T_{MST} counts the number of edges in the MST that connect vertices from different samples. T_{MST} is shown below:

$$T_{MST} = 1 + \sum_{i=1}^{n-1} \lambda_i,$$

where λ_i acts as an indicator that returns 1 if the i^{th} edge connects vertices from different samples and 0 otherwise. High values of T_{MST} indicate mixing of the distributions thus a failure to reject the null hypothesis.

Schilling (1986) considers a similar approach incorporating the k-nearest neighbors for some k for each point $\mathbf{Z}_i$. The Euclidean norm is chosen as the metric for dissimilarity though many others could be used. The introduced test statistic $T_{k,n}$ is shown below:

$$T_{k,n} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{r=1}^{k} I_i(r).$$

$I_i(r)$ represents an indicator that takes the value 1 when the r^{th} nearest neighbor belongs to same sample as the point $\mathbf{Z}_i$ being considered and 0 otherwise. Hence $T_{k,n}$ sums over the k-nearest neighbors that are members of the same sample of the particular test point while iterating over all points. This value is divided by the total number of samples and the value chosen for k. Thus the test statistic is a proportion expressing the probability of any chosen point being closest to points from the same underlying population. The null hypothesis is rejected for high values as this would indicate a lack of mixing between the two samples. Essentially, this test employs a subgraph for each point in the graph of the merged samples where the subgraph consists of the primary point connected with its k-nearest neighbors.

Unfortunately, the tests above struggle in high dimensions due to the previously discussed concentration of distance. Points that are nearest neighbors may be no more representative of the relative local densities than points that are not since the former will not be much closer to the primary point than the latter. Recall in Theorem 2.3.2 and Theorem 2.3.3 we showed the distance from a d-dimensional standard normal vector $\mathbf{Z}_1$ to the origin is approximately $\sqrt{d}$ and the distance from $\mathbf{Z}_1$ to another d-dimensional standard normal vector $\mathbf{Z}_2$ is approximately $\sqrt{2d}$ for large d. This constant distance result holds for cases outside the standard normal distribution, as mentioned earlier. If $\mathbf{X}$ and $\mathbf{Y}$ are independent samples from d-dimensional distributions F and G with variance σ_1^2 and σ_2^2 respectively, then as $d \to \infty$ $\mathbf{X}$ and $\mathbf{Y}$ form simplexes with edge length $2\sigma_1^{\frac{1}{2}}$ and $2\sigma_2^{\frac{1}{2}}$ respectively after rescaling by $d^{-\frac{1}{2}}$ and in addition the distance between any $\mathbf{X}_i$ and $\mathbf{Y}_j$ converges in probability to $(\sigma_1^2 + \sigma_2^2 + \mu^2)^{\frac{1}{2}}$ where μ^2 indicates the average sum of squared distances between each respective dimension of $\mathbf{X}_i$ and $\mathbf{Y}_j$ (Hall et al. 2005).

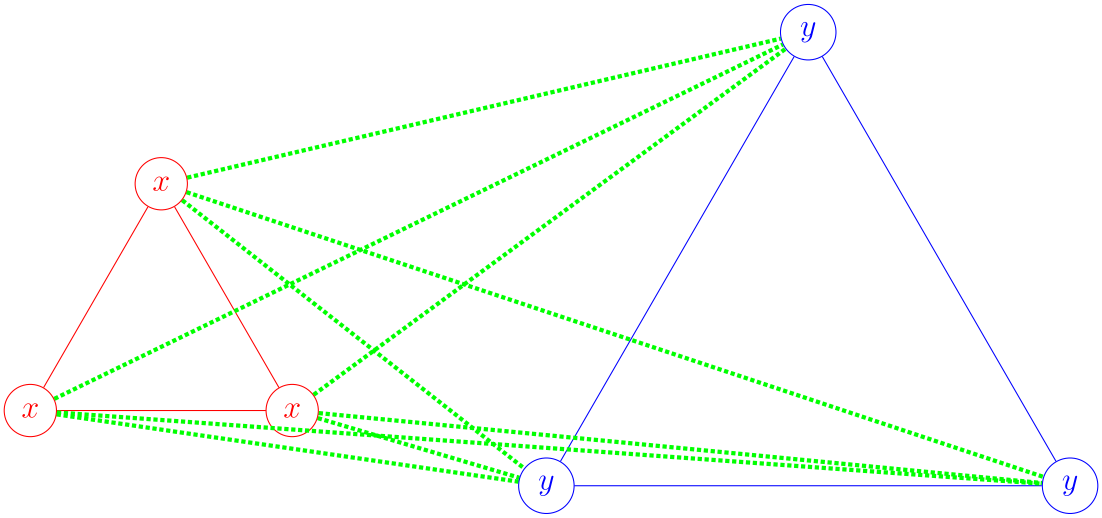


Figure 4.1: Distribution of within and between distances for two samples from unequal distributions. Within sample distances each converge to a constant, as do between sample distances.

Figure 4.1 illustrates the underlying situation for H_a where $F \neq G$. As we are assuming H_a, Figure 4.1 exhibits the concentration of distance phenomena where within sample distances represented by the red and blue lines, as well as between class distances represented by the green lines, each converge to single values as $d \to \infty$.

A numeric illustration of Figure 4.1 is a distance matrix showing within and between sample distances for $\mathbf{X}$ and $\mathbf{Y}$. For example, drawing samples of size three from joint

$N(0,1)$ and $N(0.5,2)$ distributions respectively produced the distance matrices shown in Figure 4.2 and 4.3 for $d = 16$ and $d = 1024$.

$$\begin{pmatrix} 0 & 5.39 & 3.99 & 10.82 & 8.22 & 10.33 \\ 5.39 & 0 & 6.04 & 10.95 & 9.10 & 9.59 \\ 3.99 & 6.04 & 0 & 8.89 & 8.19 & 9.07 \\ 10.82 & 10.95 & 8.89 & 0 & 13.40 & 11.27 \\ 8.22 & 9.10 & 8.19 & 13.40 & 0 & 10.16 \\ 10.33 & 9.58 & 9.07 & 11.27 & 10.16 & 0 \end{pmatrix}$$

Figure 4.2: $d = 16$

$$\begin{pmatrix} 0 & 47.16 & 44.52 & 74.09 & 74.17 & 73.53 \\ 47.16 & 0 & 44.89 & 73.34 & 73.69 & 73.38 \\ 44.52 & 44.89 & 0 & 74.27 & 73.17 & 72.19 \\ 74.09 & 73.34 & 74.27 & 0 & 92.49 & 91.69 \\ 74.17 & 73.69 & 73.17 & 92.49 & 0 & 88.91 \\ 73.53 & 73.38 & 72.19 & 91.69 & 88.91 & 0 \end{pmatrix}$$

Figure 4.3: $d = 1024$

The 3×3 blocks on the top left and bottom right of each matrix represent the within sample distances for **X** and **Y** respectively while the 3×3 blocks on the top right and bottom left symmetrically represent the between sample distances. The values produced in each submatrix align with results from Hall et al. (2005), as the distances in each 3×3 block cluster around a particular value. This trend becomes more clear as d increases in

Figure 4.3.

Biswas and Ghosh (2014) and Chen and Friedman (2017) noted that in the case of $\sigma_2^2 > \sigma_1^2 + \nu$, an observation from F will have its nearest neighbors from F, while an observation from G will have its nearest neighbors come from F as well. Figures 4.2 and 4.3 exhibit this property where the between sample distances are collectively less than the within sample distances for $\mathbf{Y}$. Thus tests like the MST have low power when a scale difference exists. Chen and Friedman (2017) introduced a modification to tests that utilize the MST to compensate for this issue. The test finds the number of the edges that connect observations from the **X** sample and the number of edges that connect observations from **Y** sample. One of these sums is scaled by a factor of two. These values are referred to as R_1 and R_2. The proposed test statistic is

$$S = (R_1 - \mu_1, R_2 - \mu_2)\Sigma^{-1}(R_1 - \mu_1, R_2 - \mu_2)^T,$$

where μ_1 and μ_2 are the expected value of R_1 and R_2 respectively. In addition, Sarkar et al. (2019) proposed modifications of MST and nearest neighbor based tests by using the MADD, discussed in Chapter 3, as a dissimilarity index.

4.2 Non-Graph Based Approach

A conjecture from Deuber (1998) stated that for equal numbers of black and white points in Euclidean space the sum of the distances between points of equal color is less than or equal to the sum of the pairwise distances between points of different color. This conjecture was eventually proven by Morgenstern (2001). Several test statistics have been introduced in the nature of Deuber's conjecture above.

Baringhaus and Franz (2004) along with Szekely and Rizzo (2004) introduced nearly identical test statistics compounding the sum of interpoint distances between the samples with the sum of distances within each sample as shown in the two formulas below:

$$T_{BFSR}(X,Y) = \frac{mn}{m+n}\Bigg[\frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}||X_j - Y_k|| - \frac{1}{2m^2}\sum_{j=1}^{m}\sum_{k=1}^{n}||X_j - X_k|| - \frac{1}{2n^2}\sum_{j=1}^{m}\sum_{k=1}^{n}||Y_j - Y_k||\Bigg]$$

or equivalently

$$T_{BFSR}(X,Y) = D_{XY} - \frac{D_{XX} + D_{YY}}{2},$$

where D_{XY} represents the average distance between the two samples indicated by the first term in brackets while D_{XX} and D_{YY} represent the average distance within the respective samples indicated by the second and third term in the brackets. Note that when calculating D_{XX} and D_{YY}, the Euclidean distance from an observation to itself is also being taken into account for the averaging process. Larger values of T_{BFSR} suggest a rejection of the null hypothesis. Using this statistic, the researchers were able to achieve power similar to the t and T^2-tests for normal location alternatives.

Biswas and Ghosh (2014) reformatted the testing by considering the pairs of interpoint distances as bivariate distributions with their own respective means. From here instead of testing whether $F = G$, they equivalently test the equality of these newly found means. The authors calculate

$$\hat{\boldsymbol{\mu}}_{D_F} = \Bigg[\hat{\mu}_{FF} = \binom{m}{2}^{-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}||\mathbf{x}_i - \mathbf{x}_j||, \hat{\mu}_{FG} = (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathbf{x}_i - \mathbf{y}_j||\Bigg],$$

$$\hat{\boldsymbol{\mu}}_{D_G} = \left[\hat{\mu}_{FG} = (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathbf{x}_i - \mathbf{y}_j||, \hat{\mu}_{GG} = \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}||\mathbf{y}_i - \mathbf{y}_j||\right],$$

with H_0 testing $\boldsymbol{\mu}_{D_F} = \boldsymbol{\mu}_{D_G}$ where higher values of the test statistic $||\hat{\boldsymbol{\mu}}_{D_F} - \hat{\boldsymbol{\mu}}_{D_G}||^2$ lead to rejection.

Recently, Li (2018) also approached the two-sample problem with a focus on interpoint distances. Li (2018) shows that in a high dimensional setting, the asymptotic distribution of the interpoint distances is normal as the dimension grows to infinity. Subsequently, two separate test statistics are introduced for location differences and scale differences. Li's tests are superior in cases he examined where both a location and scale difference exist. We however focus on the proposals of both Baringhaus and Franz and Szekely and Rizzo.

T_{BFSR} performs well when tasked with normal location alternatives but struggles with scale alternatives. To account for this we build upon T_{BFSR} and introduce

$$T_{new} = |D_{XX} - D_{YY}|.$$

These two test statistics are combined to form T_{merge} which simply appends a rejection of H_0 if either T_{BFSR} or T_{new} rejects.

To see how T_{merge} performs, we utilize a permutation test. Most statistical methods require lots of assumptions but permutation tests are nonparametric, flexible and easily applicable. The procedures are as follows:

1. Calculate the test statistic from the initial data.
2. Permute or shuffle the observations between the data samples X and Y and recalculate the test statistic from this new permuted data set.

3. Repeat step 2 a large amount of times.

4. Build the sampling distribution of the test statistic when H_0 is true from the results of step 2 and 3.

5. Calculate the *p-value:* probability of observing a value as or more extreme than the actual test statistic under H_0 by counting how many of the permuted test statistics meet or exceed the actual test statistic and dividing by the number of permutations performed.

6. Reject H_0 if the p-value is less than a pretermined α. We use $\alpha = 0.05$ in our studies.

7. To calculate the power of the test, repeat the whole permutation process many times and find the proportion of times in which H_0 was rejected.

4.3 Power studies

We conduct our first power analysis on the proposed test statistic using the permutation procedure approach described above and compare it to the results attained by Baringhaus and Franz (2004) which aligned closely to those attained by Szekely and Rizzo (2004). We show results for T_{merge} and T_{BFSR} using simulations consisting of 200 repetitions with 400 permutations each. For the simulations, we consider samples from two d-dimensional multivariate normal distributions of differing location and/or scale. We fix $\mathbf{X}$ to have a $N_d(\mathbf{0}, \mathbf{I}_d)$ distribution while $\mathbf{Y}$ varies between $N_d(\boldsymbol{\mu}_{\boldsymbol{Y}}, \mathbf{I}_d)$ for location alternatives and $N_d(\mathbf{0}, \delta\mathbf{I}_d)$ for scale alternatives. For simulations in lower dimensions, $\boldsymbol{\mu}_{\boldsymbol{Y}}$ is of the form $(\Delta, 0, \ldots, 0)$ to align with the results from Baringhaus and Franz (2004) and Szekely and Rizzo (2004). Neither paper tested their respective method for spaces over six dimensions but if the same pattern of only varying $\boldsymbol{\mu}_{\boldsymbol{Y}}$ in one dimension is applied in higher dimensional spaces, the resulting power tends to 0 as $d \to \infty$, whereas if $\boldsymbol{\mu}_{\boldsymbol{Y}}$ varies in every dimension power will tend to 1. To account for this in our simulations, we sometimes allowed $\boldsymbol{\mu}_{\boldsymbol{Y}}$ to vary in $\lceil \sqrt{d} \rceil$ of the dimensions. Our reasoning follows from the procedures of design for a typical power study where stability is reached with alternatives differing from the null by $O(\frac{1}{\sqrt{n}})$ where n represents the sample size.

A significance level $\alpha = 0.05$ was used along with different cases for the sample sizes, n_1 from **X** and n_2 from **Y**. However, as T_{merge} is a combination of T_{BFSR} and T_{new}, we apply a Bonferroni correction by lowering the significance level for each component test from 0.05 to 0.025. This enables the size of the test to not exceed 0.05.

4.3.1 Location Alternatives

For each dimension we use combinations of sample size $n_1 = n_2 = 20$ and $n_1 = 20, n_2 = 50$. Table 4.1 shows power results for different dimensional spaces where $\boldsymbol{\mu_Y}$ varies in only a single dimension. As expected, T_{BFSR} is slightly superior to T_{merge} in nearly every category in detecting a location difference. In addition, the clear decrease in power seen as $d \to \infty$ supports our earlier reasoning for allowing $\boldsymbol{\mu_Y}$ to vary in $\lceil\sqrt{d}\rceil$ dimensions. As such we reproduce the results from Table 4.1 in this fashion for $d = 16$ and above. The new results are shown in Table 4.2. The updated values of both tests when $\Delta = 0.5$ show roughly stable power. Though T_{merge} exhibits substantial improvements, it is still somewhat deficient in comparison to T_{BFSR}.

	Sample Size		$\Delta = 0.5$		$\Delta = 1$	
d	n_1	n_2	T_{BFSR}	T_{merge}	T_{BFSR}	T_{merge}
2	20	20	26	21	78	69
	20	50	37	31	92	85
6	20	20	17	13	59	56
	20	50	24	19	81	75
16	20	20	15	12	36	30
	20	50	16	15	67	60
32	20	20	10	7	30	24
	20	50	16	11	52	43
64	20	20	7	6	16	13
	20	50	9	6	31	23
128	20	20	11	9	17	12
	20	50	10	9	18	16

Table 4.1: Significant tests as percentage for multidimensional normal location alternatives with $\boldsymbol{\mu_Y}$ varying in a single dimension.

	Sample Size		$\Delta = 0.5$		$\Delta = 1$	
d	n_1	n_2	T_{BFSR}	T_{merge}	T_{BFSR}	T_{merge}
16	20	20	49	33	99	99
	20	50	68	53	100	99
32	20	20	49	33	98	99
	20	50	74	62	99	99
64	20	20	50	40	100	99
	20	50	64	66	100	100
128	20	20	47	37	100	99
	20	50	70	69	100	100

Table 4.2: Significant tests as percentage for multidimensional normal location alternatives with $\boldsymbol{\mu_Y}$ varying in $\lceil\sqrt{d}\rceil$ dimensions.

4.3.2 Scale Alternatives

As discussed previously, for normal scale alternatives we consider a $\mathbf{Y}$ distribution of the form $N_d(0, \delta \mathbf{I}_d)$. To align with results from Baringhaus and Franz (2004), we select $n_1 = 20, n_2 = 50$ and test at $d = 6, 32$ and 128. Recall in T_{merge} we incorporate $T_{new} = |D_{XX} - D_{YY}|$ where D_{XX} and D_{YY} represents the average of pairwise distances within the $\mathbf{X}$ and $\mathbf{Y}$ sample respectively. Our conjecture that the addition of this term could expose a scale difference and boost power scores is confirmed in Figures 4.4-4.6. T_{merge} clearly outperforms T_{BFSR} by a large margin in detecting scale alternatives.

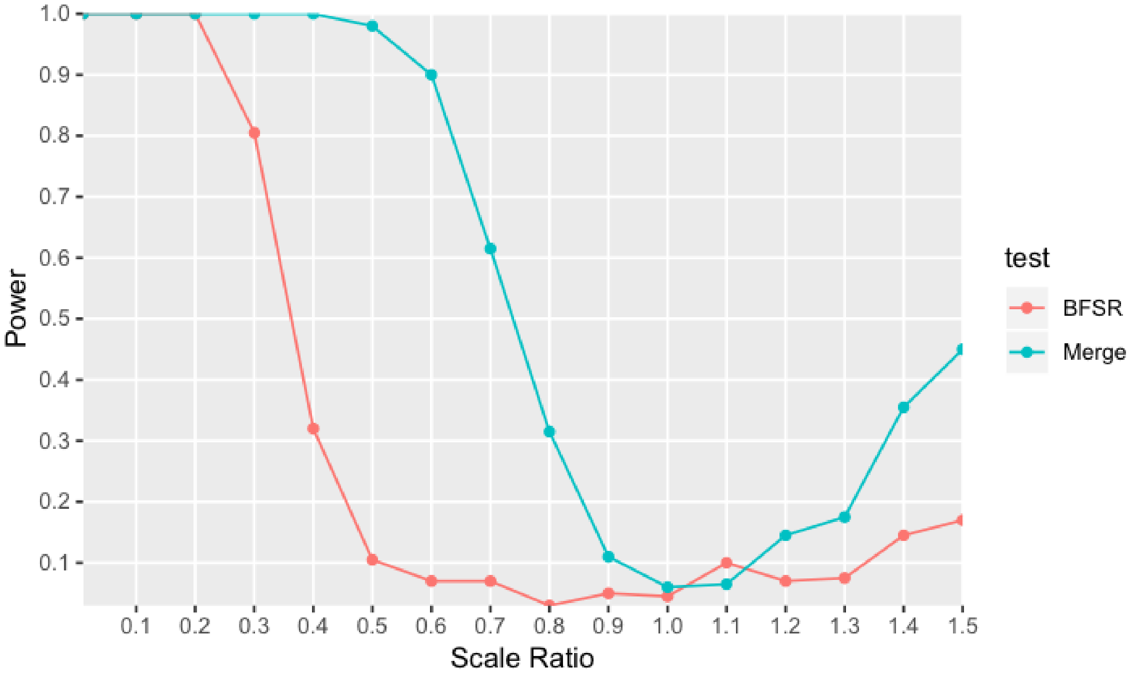


Figure 4.4: Approximate power at $d = 6$.

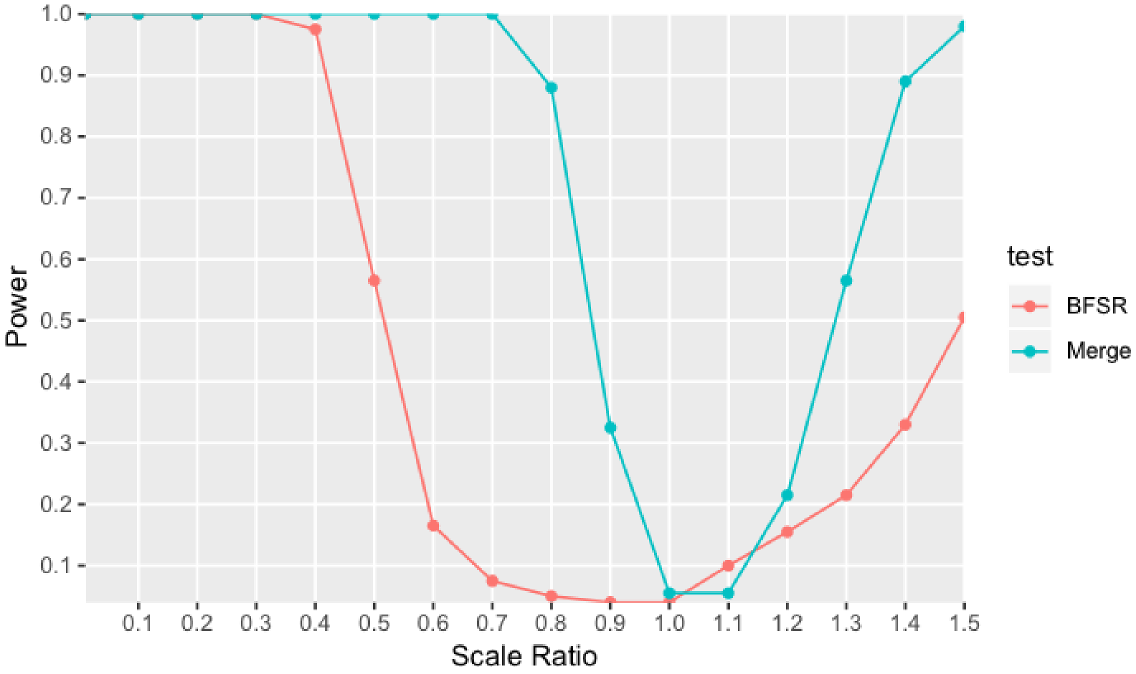


Figure 4.5: Approximate power at $d = 32$.

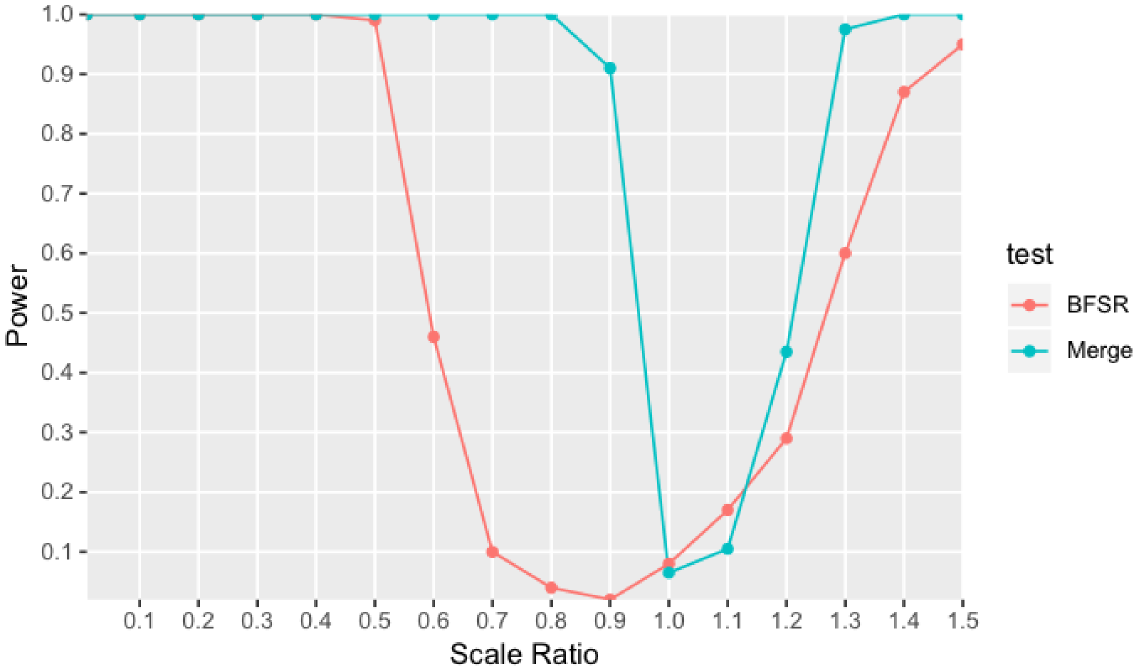


Figure 4.6: Approximate power at $d = 128$.

4.3.3 Additional power studies

We investigate the power of other tests for further comparison to gain an ample understanding of how our own test statistic performs. Hall and Tajvidi (2002) proposed a test that finds the number of nearest neighbors of a particular observation that are of an opposing class and compares that count with the expected value of the underlying hypergeometric distribution. Rosenbaum (2005) suggested using a test statistic based on an optimal non-bipartite matching. The method follows by joining pooled sample points into disjoint pairs with the goal of minimizing the total sum of distances within pairs. After finding the optimal matching, the test statistic is the number of pairs with elements from different samples. We will also include the previously discussed tests of Baringhaus and Franz (2002), Szekeley and Rizzo (2002), Friedman and Rafsky (1979), and also Biswas and Ghosh (2014) from which we use the chosen parameters along with the conducted power studies' results of these aforementioned methods.

The parameters are $m = n = 20$ and $d = 500$ with a significance level of 0.05. We test for differences in location and scale individually in addition to a combination of the two. Though the mean vector for F is the origin, when a location difference exists, unlike the previous power studies, the mean vector for G varies in every dimension, i.e, $\mu_F = (0, 0, ..., 0)$, $\mu_G = (\Delta, \Delta, ..., \Delta)$. The dispersion matrix for F is $\Sigma_1 = \mathbf{S}$ where **S** is a $d \times d$ matrix of the form $s_{i,j} = (0.5)^{|i-j|}$ with $\Sigma_2 = \sigma^2\mathbf{S}$ as the dispersion matrix for G. We first consider cases where F and G are both multivariate normal distributions. Observed powers are shown below in Table 4.3. Some of the test names are abbreviated with BG as Biswas and Ghosh (2014), FR as Friedman and Rafsky (1979), and HT as Hall and Tajvidi (2002).

$Test$	$\Delta = 0.25, \sigma^2 = 1$	$\Delta = 0, \sigma^2 = 1.25$	$\Delta = 0.1, \sigma^2 = 1.1$
BFSR	100.0	20.0	37.0
BG	82.0	100.0	94.0
FR	77.0	0.0	7.5
HT	83.5	100.0	92.5
Rosenbaum	67.5	9.0	11.0
T_{merge}	100.0	100.0	92.0

Table 4.3: Significant tests as a percentage when comparing multivariate normal distributions at $d = 500$.

When comparing two multivariate normal distributions, our test reached maximal power in cases of a sole location or scale difference and was only slightly outperformed by Hall and Tajvidi (2002) and Biswas and Ghosh (2014) when both scale and location are different.

We now consider the case where F and G are Laplace distributions. The Laplace distribution is important to study as a contrast to the normal distribution due to its sharp central peak and heavier tails. Results are shown in Table 4.4.

$Test$	$\Delta = 0.25, \sigma^2 = 1$	$\Delta = 0, \sigma^2 = 1.25$	$\Delta = 0.1, \sigma^2 = 1.1$
BFSR	100.0	28.0	48.5
BG	69.5	100.0	85.5
FR	91.0	0.0	6.0
HT	70.0	100.0	79.0
Rosenbaum	86.5	10.0	11.5
T_{merge}	100.0	81.0	80.5

Table 4.4: Significant tests as a percentage when comparing multivariate Laplace distributions at $d = 500$.

When comparing two multivariate Laplace distributions, our test reached maximal power when a location difference exists and ranked in the top three with a scale or mixed alternative.

We also examine the case where F is a standard normal distribution while G is a standard Laplace distribution. Results are shown in Table 4.5.

BFSR	100.0
BG	100.0
FR	0.0
HT	100.0
Rosenbaum	88.5
T_{merge}	100

Table 4.5: Significant tests as a percentage when testing between standard normal distribution and standard Laplace distribution at $d = 500$.

Our test along with several others reach maximal power in this scenario. In line with previous discussion, Friedman and Rafsky struggles tremendously due to the large disparity in the scales of the two distributions.

More power studies are appropriated for comparison from Mondal et al. (2005). The parameters are near identical as before but **S** is a $d \times d$ matrix of the form $s_{i,j} = (-0.5)^{|i-j|}$. Mondal et al. (2005) modified the test proposed by Schilling (1986) by comparing the proportions of neighbors of each observation that come from it's underlying class to the expected values of these proportions . Results from Mondal et al. (2005) will be abbreviated as MBG in the results shown in Table 4.6 and 4.7.

d	BFSR	FR	HT	MBG	T_{merge}
25	75.6	30.4	9.6	23.0	74.2
50	97.8	44.6	17.8	27.2	88.2
100	100.0	67.2	36.8	47.8	99.0

Table 4.6: Significant tests as percentage for multivariate normal location alternatives where $\Delta = 0.3$.

d	BFSR	FR	HT	MBG	T_{merge}
25	7.0	3.8	47.8	44.8	64.0
50	10.6	3.0	85.8	76.8	91.8
100	13.6	2.4	99.2	95.8	97.8

Table 4.7: Significant tests as percentage for multivariate normal scale alternatives where $\sigma^2 = 1.3$.

For the above location alternative, our test along with that of Baringhaus and Franz (2002) resulted in the far higher power for all tested dimensions than the other tests. For a scale alternative, our test had the highest power in all dimensions cases followed by the test of Hall and Tajvidi (2002).

Several other tests exist for the general two-sample problem. Srivastava et al. (2016) test solely for equality of means by projecting the data matrix onto lower dimensional subspaces. Good results are achieved with a high dimensional gene expression data set. Tsukada (2019) offers a combination of the tests from Baringhaus and Franz (2004) and Biswas and Ghosh (2014). Other approaches have been taken by Aslan and Zech (2005), who employed an energy function, Chen and Qin (2010), who modified Hotelling's T^2 test, Aupetit and Catz (2005) who implemented a Delauney graph, and Biswas et al. (2015) who used Hamiltonian paths. Ruth (2014) and Petrie (2016) used similar approaches to Rosenbaum (2005). Further tests were introduced by Klebanov et al. (2006) and Zhou et al. (2017).

4.3.4 Applications to real data

We applied some of the previously discussed tests on two actual collected data sets. The first consists of various voice measurements of participants from whom a portion have Parkinson's disease. There are 22 variables for evaluating the voices with 24 out of 32 participants actually having Parkinson's with the other eight serving as controls. We tested to see if there is a significant difference between the groups and see how our test compares to those of Schilling (1986) and Baringhaus and Franz (2004). Recall our test statistic T_{merge} is a combination of T_{BF} from Baringhaus and Franz (2004) as well as $|D_{XX} - D_{YY}|$ checking for scale differences, wherein T_{merge} rejects at significance level α when either of the comprising tests rejects at $\frac{\alpha}{2}$. Resulting p-values were as follows: 0.719 for Schilling's test using $k = 3$ neighbors; 0.020 for T_{BF}, and 0.296 using $|D_{XX} - D_{YY}|$. As T_{BF} rejects at $\alpha = 0.025$, T_{merge} rejects at $\alpha = 0.05$. Closer inspection of the data reveals strong location differences between the samples which explains the superior performance of T_{BF}; however T_{merge} did nearly as well. The ineffectiveness of Schilling's test is unsurprising due to the high dimensional low sample size data situation here.

The second data set consists of various medical measurements obtained from colposcopy, a procedure which is used for cervical cancer screening, where observations were obtained through two modalities. Resulting p-values were as follows: 0.468 for Schilling's test using $k = 5$ neighbors (other values of k gave similar results); 0.354 for T_{BF}, and 0.048 using $|D_{XX} - D_{YY}|$. As $|D_{XX} - D_{YY}|$ is significant at $\alpha = 0.05$, T_{merge} rejects at $\alpha = 0.10$, whereas T_{BF} and Schilling's test give no evidence whatsoever against the null hypothesis that the two modalities produce the same distribution of measurements. The subpar results from Baringhaus and Franz's test and Schilling's test in contrast to the significant results from $|D_{XX} - D_{YY}|$ is due to the appreciable scale differences in the measurements between the two modalities, while location differences are almost nonexistent. Thus these two applications demonstrate that the T_{merge} test we have introduced has the ability to detect both differences in location and differences in scale.

Chapter 5

Further Statistical Applications

5.1 Clustering

In Chapter 3 we discussed classification in high dimensions. To test our classification method, we first trained it on a data set which had known target values. In other words, our analysis was an example of supervised learning. But what if the underlying class of each observation was unknown to us? Or what if the number of classes is not even known? One approach to gain some insight into the data is clustering.

Clustering is an unsupervised method which aims to merge individual observations in the sample into select groups populated by other observations they are most similar to. Ideally this will offer some insight into the occupants of each of these clusters. Again the issue of similarity comes up. How can we quantify the similarity of two observations? A very common and simple approach is to consider the Euclidean distance between them. Though several methods exist, we focus on k-means clustering. The procedures for k-means are as follows:

1. Select the number of clusters k. Choosing an optimal value for k is a heavily researched topic though we do not focus on that.
2. Randomly select k of the data points. These represent the initial cluster centers.
3. Measure Euclidean distance (or some other chosen metric) from every observation to the cluster centers.
4. Assign each observation to its closest cluster.
5. With these k initialized clusters, calculate the centroid for each to be the new cluster centers and repeat steps 3 and 4.

6. The algorithm converges when no observation switches its assigned cluster.

Clearly as the dimension increases, the "concentration of distance" phenomena we have discussed arises again. Steps 3 and 4 in the k-means algorithm above will be essentially useless as the calculated distances all converge to a single value. Several approaches exist to mediate this difficulty. Perhaps it is preferrable to utilize a different metric altogether. Aggarwal et al. (2001) found that the Manhattan metric (L_1 norm) is preferable in high dimensions. In addition fractional norms are introduced which offer several advantages in k-means.

Though it doesn't play into the core ideas of this work on problems associated with distance-based methods, principal component analysis (PCA) is still a viable method in many situations by possibly reducing the number of dimensions greatly. Principal component analysis performs a singular value decomposition on the training set to factorize it into three separate matrices that are used to extract the principal components. These components spot the axes that explain the largest amount of variance. In an ideal situation, a significant portion of the variance can be explained by a number of components that is far less than the original dimensions of the data. The original d-dimensional data set can be projected into a p-dimensional hyperplane where p is the number of principal components. From here standard clustering techniques like k-means can be employed. However, like many other dimensionality reduction techniques, PCA struggles to offer fruitful results when there is a lack of strong correlations among the variables in the original data. There is potential for a high amount of information loss.

An approach proposed by Sarkar and Ghosh (2019) circumvents dimensionality reduction and offers an effective approach to conduct clustering that exploits the constancy of distance phenomenon introduced in Section 2.3.2. Recall the use of the *Nearest Neighbor Mean Absolute Difference of Distance* (NN-MADD) introduced by Pal et al. (2016) and applied in chapter 3 for the purpose of classification. The formula for the MADD, again, is as follows:

$$\rho(z,w) = (n-2)^{-1} \sum_{t \in \chi - \{z,w\}} ||\mathbf{z} - \mathbf{t}||_p - ||\mathbf{w} - \mathbf{t}||_p.$$

The above formulation assesses the dissimilarity between two d-dimensional observations z and w within the training set χ by finding the sum of their respective distances to every other observation t in the training set. If z and w belong to the same underlying population/cluster, $d^{-\frac{1}{2}}\rho(z,w) \to 0$ as $d \to \infty$. Results from numerous simulations from Sarkar and Ghosh (2019) showed that clustering based on the MADD led to almost perfect clustering on high dimensional data with low sample size demonstrating a stark contrast to methods based solely on Euclidean distance.

5.2 Goodness of Fit

Testing whether a collected sample follows a particular known distribution is another fundamental task in statistics. Several methods exist for this purpose. Most rely on a ranking of the observations followed by a distance-based comparison between the sample and predicted distribution. A common method used in univariate situations is the Kolmogorov-Smirnov Test (to be referred to as KST). The KST proceeds by measuring the largest distance between the cumulative distribution functions of both the sample and specified distribution. Unfortunately the simplicity of the KST and similar procedures such as the Cramer von Mises and Anderson-Darling tests does not extend easily for multidimensional problems. χ^2 tests are an option though an appropriate binning of the data must be chosen. Two-sample testing can also be used to perform goodness of fit. Existing methods generally again function by inspecting an observation's nearest neighbors (see for example Schilling (1983a,b)). This diminishes power as the dimension increases. Tests designed in the frame of those discussed that exploit the underlying geometry in high dimensional spaces, could offer benefit in goodness of fit testing. We leave this for future work.

Chapter 6

Conclusion

In this work, we offered several illustrations of the bizarre behavior exhibited in high dimensions. Particularly important results obtained revealed the phenomena of concentration of distance in high dimensions. Exposure of this phenomena led to a new binary classification method and a new test statistic for the general two-sample problem. For classification, we provided evidence of reduction in misclassification rates with our method compared to the method introduced in the motivating literature through the use of simulations with a growing number of dimensions. Similarly our proposed test statistic for the two-sample problem demonstrated performance superior to other tests suggested in previous works for cases of normal scale alternatives and managed comparable results for normal location alternatives. Permutation procedures were detailed and employed for this analysis. The efficacy of our proposed test statistic was reaffirmed through the use of real data sets. In addition, we discussed of some of the approaches that can be taken when dealing with the unsupervised task of clustering as well as how the constancy of distance phenomenon might be used both there and for multidimensional goodness of fit testing.

References

[1] Hall, Peter, James Stephen Marron, and Amnon Neeman. “Geometric representation of high dimension, low sample size data.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.3 (2005): 427-444.

[2] Pal, Arnab K., Pronoy K. Mondal, and Anil K. Ghosh. “High dimensional nearest neighbor classification based on mean absolute differences of inter-point distances.” Pattern Recognition Letters 74 (2016): 1-8.

[3] Baringhaus, Ludwig, and Carsten Franz. “On a new multivariate two-sample test.” Journal of Multivariate Analysis 88.1 (2004): 190-206.

[4] Schilling, Mark F. “Multivariate two-sample tests based on nearest neighbors.” Journal of the American Statistical Association 81.395 (1986): 799-806.

[5] Li, Jun. “Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem.” Biometrika 105.3 (2018): 529-546.

[6] Sarkar, Soham, Rahul Biswas, and Anil K. Ghosh. “On high-dimensional modifications of some graph-based two-sample tests.” arXiv preprint arXiv:1806.02138 (2018).

[7] Friedman, Jerome H., and Lawrence C. Rafsky. “Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests.” The Annals of Statistics (1979): 697-717.

[8] Blum, Avrim, John Hopcroft, and Ravi Kannan. Foundations of data science. Cambridge University Press, 2020.

[9] Szkely, Gbor J., and Maria L. Rizzo. “Testing for equal distributions in high dimension.” InterStat 5.16.10 (2004): 1249-1272.

[10] Sarkar, Soham, and Anil K. Ghosh. “On perfect clustering of high dimension, low sample size data.” IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[11] Biswas, Munmun, and Anil K. Ghosh. “A nonparametric two-sample test applicable to high dimensional data.” Journal of Multivariate Analysis 123 (2014): 160-171.

[12] Hall, Peter, and Nader Tajvidi. “Permutation tests for equality of distributions in high-dimensional settings.” Biometrika 89.2 (2002): 359-374.

[13] Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. “On the surprising behavior of distance metrics in high dimensional space.” International Conference on Database Theory. Springer, Berlin, Heidelberg, 2001.

[14] Srivastava, Muni S., Shota Katayama, and Yutaka Kano. “A two sample test in high dimensional data.” Journal of Multivariate Analysis 114 (2013): 349-358.

[15] Chen, Hao, and Jerome H. Friedman. “A new graph-based two-sample test for multivariate and object data.” Journal of the American Statistical Association 112.517 (2017): 397-409.

[16] Rosenbaum, Paul R. “An exact distribution-free test comparing two multivariate distributions based on adjacency.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.4 (2005): 515-530.

[17] Ruth, David M. “A new multivariate two-sample test using regular minimum-weight spanning subgraphs.” Journal of Statistical Distributions and Applications 1.1 (2014): 22.

[18] Petrie, Adam. “Graph-theoretic multisample tests of equality in distribution for high dimensional data.” Computational Statistics & Data Analysis 96 (2016): 145-158.

[19] Aslan, B., and G. Zech. “A new test for the multivariate two-sample problem based on the concept of minimum energy.” Journal of Statistical Computation and Simulation 75.2 (2005): 109-119.

[20] Chen, Song Xi, and Ying-Li Qin. “A two-sample test for high-dimensional data with applications to gene-set testing.” The Annals of Statistics 38.2 (2010): 808-835.

[21] Morgenstern, Dietrich. “Proof of a conjecture by Walter Deuber concerning the distances between points of two types in R^d.” Discrete Mathematics 226.1-3 (2001): 347-349.

[22] Aupetit, Michal, and Thibaud Catz. “High-dimensional labeled data analysis with topology representing graphs.” Neurocomputing 63 (2005): 139-169.

[23] Biswas, Munmun, Minerva Mukhopadhyay, and Anil K. Ghosh. “On some exact distribution-free one-sample tests for high dimension low sample size data.” Statistica Sinica (2015): 1421-1435.

[24] Tsukada, Shin-ichi. “High dimensional two-sample test based on the inter-point distance.” Computational Statistics 34.2 (2019): 599-615.

[25] Mondal, Pronoy K., Munmun Biswas, and Anil K. Ghosh. “On high dimensional two-sample tests based on nearest neighbors.” Journal of Multivariate Analysis 141 (2015): 168-178.

[26] Klebanov, Lev, et al. “A permutation test motivated by microarray data analysis.” Computational Statistics & Data Analysis 50.12 (2006): 3619-3628.

[27] Zhou, Wen-Xin, Chao Zheng, and Zhen Zhang. “Two-sample smooth tests for the equality of distributions.” Bernoulli 23.2 (2017): 951-989.

[28] W. Deuber, 2. Problem 303, Discrete Math. 192 (1998) 348.

[29] Schilling, Mark F. “An infinite-dimensional approximation for nearest neighbor goodness of fit tests.” The Annals of Statistics (1983a): 13-24.

[30] Schilling, Mark F. “Goodness of Fit Testing in $\mathbb{R}^m$ Based on the Weighted Empirical Distribution of Certain Nearest Neighbor Statistics.” The Annals of Statistics 11.1 (1983b): 1-12.

[31] Little, Max A., et al. “Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection.” Biomedical engineering online 6.1 (2007): 23.

[32] Fernandes, Kelwin, Jaime S. Cardoso, and Jessica Fernandes. “Transfer learning with partial observability applied to cervical cancer screening.” Iberian conference on pattern recognition and image analysis. Springer, Cham, 2017.