ON MANIFOLD LEARNING FOR SUBSEQUENT INFERENCE

by

Mingyue Gao

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2019

© 2019 Mingyue Gao

All rights reserved
Abstract

Manifold learning algorithms are successfully used in machine learning and statistical pattern recognition to learn a meaningful low-dimensional representation of high-dimensional data. Manifold learning is often applied as the first step in addressing a certain ultimate inference task, such as classification and hypothesis testing, in order to obtain better performance on the inference task. Fundamental questions arise about the utilities of manifold learning in the subsequent inference: (1) If the true low-dimensional manifold is known, is the subsequent inference in that manifold superior to that in the high-dimensional ambient space? (2) Does the subsequent inference in the learnt low-dimensional manifold recover that in the true manifold?

In this work, we explore answers to these two questions in several inference tasks. We start by considering the power of Likelihood Ratio Tests (LRTs). In multinomial models, we demonstrate that the power in the true manifold is not uniformly superior to that in the high-dimensional ambient space when there are finite samples. We then consider the expected error of classification in multinomial models and observe that classification in the true manifold fails to yield a uniformly smaller expected error.
ABSTRACT

than that in the ambient space. Lastly, for network models, we consider the adjacency spectral embedded space as a manifold, where a lower-dimensional submanifold potentially exists. Under appropriate conditions, after applying Isomap in the estimated adjacency spectral embedded space to learn the submanifold, we establish that the hypothesis testing in the learnt submanifold is asymptotically equal to that in the true submanifold.

Primary Reader: Carey E. Priebe
Secondary Reader: Minh Tang
Acknowledgments

For the past four and a half years, I am grateful for the company of all the lovely
people around me. First I would like to express my deep gratitude to my advisor,
Dr. Carey E. Priebe, for his unconditional support and valuable advice at all levels.
I would also like to thank Dr. Minh Tang and Dr. Michael Trosset, who provided
varied, but consistently invaluable, guidance. I thank my co-authors and collaborators
that all offered their help with patience in discussions over the past few years. Special
thanks go to Dr. Randal Burns, Dr. Jason Eisner and Dr. Joshua Vogelstein for being
on my GBO committee. I also want express my appreciation to the faculty and staff
of the Applied Maths and Statistics Department at Johns Hopkins University. My
thanks go to my fellow students and friends at Hopkins for all the time we have spent
together.
Dedication

This thesis is dedicated to my parents, for their endless love.
Contents

Abstract ii

Acknowledgments iv

List of Tables ix

List of Figures x

1 Introduction 1

2 On the Power of Likelihood Ratio Tests (LRT's) in Dimension-Restricted Submodels 10

2.1 Restricted LRT Power Superiority Literature Review . . . . . . . . . . 11

2.2 Asymptotic Theory for LRT's . . . . . . . . . . . . . . . . . . . . . 16

2.3 One-Sample LRT's in Multinomial Models . . . . . . . . . . . . . . . 17

2.3.1 Hardy-Weinberg Equilibrium . . . . . . . . . . . . . . . . . . . . 18

2.3.2 Linear Restricted Submodels . . . . . . . . . . . . . . . . . . . . 30
### CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>$\pi_M$ vs. $\pi_{\tilde{M}}$</td>
<td>30</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Multinomial Models</td>
<td>32</td>
</tr>
<tr>
<td>2.5</td>
<td>Two-Sample LRT’s in Multinomial Models</td>
<td>41</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Hardy-Weinberg Equilibrium</td>
<td>41</td>
</tr>
<tr>
<td>2.6</td>
<td>Appendix: Proofs</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>On the Expected Error of Classification in Dimension-Restricted Submodels</td>
<td>56</td>
</tr>
<tr>
<td>3.1</td>
<td>Classification in Continuous Models (Gaussians)</td>
<td>59</td>
</tr>
<tr>
<td>3.2</td>
<td>Classification in Discrete Models (Multinomials)</td>
<td>65</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Linear Restricted Submodels</td>
<td>66</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Hardy-Weinberg Equilibrium</td>
<td>73</td>
</tr>
<tr>
<td>4</td>
<td>Manifold Learning in Network Inference</td>
<td>77</td>
</tr>
<tr>
<td>4.1</td>
<td>Random Graph Models</td>
<td>79</td>
</tr>
<tr>
<td>4.2</td>
<td>Manifold Learning in ASE Space</td>
<td>84</td>
</tr>
<tr>
<td>4.3</td>
<td>One-Sample Testing</td>
<td>92</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Power Comparison: $\pi_{\tilde{H}}$ vs. $\pi_H$</td>
<td>102</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Power Analysis: $\pi_{R^a}$ vs. $\pi_H$</td>
<td>107</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Power Comparison: $\pi_H$ vs. $\pi_{\tilde{H}}$</td>
<td>114</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Simulations</td>
<td>116</td>
</tr>
<tr>
<td>4.4</td>
<td>Two-Sample Testing</td>
<td>119</td>
</tr>
</tbody>
</table>
List of Tables

4.1 Estimated powers of the two-sample test with $\alpha = 0.05$, $n_3 = 15$ and $n_4 = 16$ ................................................................ 129
4.2 Estimated powers $\pi_{d_1,d_2}$ ................................................................ 140
List of Figures

2.1 The curve shows the power difference \( \pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) - \pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) \) at alternatives \( \tau_A \in (0, 1) \), when \( n = 10, \tau_0 = 0.3, \alpha = 0.05 \). The red portion shows where the difference is positive. Both powers are computed exactly in R to default significant digits.

2.2 From a 3D visualization of the Hardy-Weinberg example, we see the \( \theta_i \) (red points) on the HW curve \( H \) and the \( \hat{\theta}_i \) (green points) in the simplex \( \Delta^2 \) (blue surface) but off the curve.

2.3 The parameters are \( \tau_0 = 0.3, \alpha = 0.05, m = 25, n_{rand} = 20, s = 10, k = 10 \) (nearest neighbor size in Isomap). By visual inspection, it is clear that under this setting \( T_{\hat{H}} \approx T_H \), there is no power dominance between \( T_H \) and \( T_{\Delta^2} \). The unrestricted and restricted test powers are computed exactly in R. The learnt restricted power curve is estimated via a Monte Carlo simulation with 100 iterations. In each iteration, we first sample \( \tau_1, \ldots, \tau_m \sim U(0, 1) \); then we sample \( \{x_{ij}\}_{j=1}^{n_{rand}} \sim \text{Tri}(\psi(\tau_i)) \) for \( i = 1, \ldots, m \) and \( \{x_{i}^{*}\}_{i=1}^{s} \sim \text{Tri}(\psi(\tau^*)) \) (\( \tau^* = \tau_0 \) under \( H_0 \) and \( \tau^* = \tau_A \) at alternative \( \tau_A \)); and next we apply Isomap to \( \{\theta_i\}_{i=1}^{m} \cup \theta \cup \psi(\tau_0) \) and we obtain \( \hat{Z}^*, \hat{Z}_0 \).

3.1 Assuming \( a = 2 \) in \( \psi_L \), the curve shows the difference \( E[L_{\psi_L}] - E[L_{\Delta^2}] \) when \( N_0 = N_1 = 5 \) (\( p_0 = 0.5 \)), \( \tau_0 = 0.1 \), and ranging \( a\tau_1 \in (0, 1) \).

3.2 The curve shows the difference \( E[L_{\hat{H}}] - E[L_{\Delta^2}] \) when \( n = 10, \tau_0 = 0.3, p_0 = 0.2 \) and ranging \( \tau_1 \in (0, 1) \).

4.1 Model Selection of embedding dimension of \( ASE(A) \), embedding dimension \( D = 3 \), i.e., the top 3 singular values and their associated singular vectors are chosen by singular value thresholding.

4.2 The green dots are \( X_i \), forming \( \mathcal{H} \); the red dots are \( \hat{X}_i \), estimated latent positions before the proper rotation; and the black dots are \( X_i \) by rotating \( X_i \) with the rotation matrix \( W_0 \).
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>$X_i$ are the red dots in $\mathbb{R}^3$ around the true latent positions $X_i \in \mathcal{H} \subset \Delta^2$ ($\mathcal{H}$ in black and simplex in purple).</td>
<td>95</td>
</tr>
<tr>
<td>4.4</td>
<td>The scree plot of the embedding dimension of $\mathcal{X}$ via Isomap [2]. The large gap between the first two dimensions indicates that the optimal dimensionality of the submanifold is 1.</td>
<td>96</td>
</tr>
<tr>
<td>4.5</td>
<td>Map of the four tests that will be discussed in the following subsections.</td>
<td>102</td>
</tr>
<tr>
<td>4.6</td>
<td>The power comparison between $\pi_{R^3}$ and $\pi_{H}$.</td>
<td>113</td>
</tr>
<tr>
<td>4.7</td>
<td>ROC curves comparison between $\pi_{R^3}$ and $\pi_{H}$.</td>
<td>113</td>
</tr>
<tr>
<td>4.8</td>
<td>Estimated test densities of $T_{iso}$ and $T_{H}^c$. The vertical lines mark the estimated critical values, respectively.</td>
<td>117</td>
</tr>
<tr>
<td>4.9</td>
<td>Estimated ROC curves illustrate relations between $\alpha$ (false positive) and $\pi_{R^3}, \hat{\pi}<em>{H,MDE}, \hat{\pi}</em>{H,MLE}, \hat{\pi}<em>{H,MLE}, \hat{\pi}</em>{iso}$ (estimated true positive).</td>
<td>118</td>
</tr>
<tr>
<td>4.10</td>
<td>Estimated ROC curves of the unrestricted test (black), the restricted test (red), and the learnt restricted test (blue).</td>
<td>130</td>
</tr>
<tr>
<td>4.11</td>
<td>A 2D visualization of 3D data. The green dots represent rows of $X$ (Hardy-Weinberg curve), the red dots are rows of $X$, the black dots are rows of $X$, and the blue dots are rows of $X$. The black dots and the blue dots largely overlap.</td>
<td>133</td>
</tr>
<tr>
<td>4.12</td>
<td>On proving $T_{iso} \rightarrow T_{H}^c$. Dashed lines denote indirect relations.</td>
<td>144</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Real-world data, such as fMRI scans, speech signals and networks, are usually measured or presented in high dimensions. In practice, the high-dimensional representation is often redundant with noisy dimensions. To achieve a better understanding of the data, the model and more successful methodologies, meaningful dimensionality reduction needs to be applied. The first of its kind, Principal Component Analysis (PCA) is a powerful linear dimensionality reduction technique that maps high-dimensional data onto a low-dimensional hyperplane. Other widely used linear dimensionality reduction techniques include Classical Multidimensional Scaling (CMDS), which finds a low-dimensional representation to best preserve pairwise distances. CMDS as a generalization of PCA is applicable to all sorts of distance measures, but it is still not adequate in handling complex data that are not concentrated around a hyperplane and that have more complicated forms.
CHAPTER 1. INTRODUCTION

Various manifold learning, or nonlinear dimensionality reduction, algorithms have been proposed to address this problem. These methods aim to recover the intrinsic geometric structure of smooth nonlinear manifolds. For instance, Isomap \[2\] as a global approach provides a low-dimensional representation that preserves pairwise geodesic distances on quasi-isometric manifolds. Local approaches, such as Locally Linear Embedding (LLE) \[3\] and Laplacian Eigenmaps \[4\], attempt to preserve the local geometry of the data and map nearby points on the manifold to nearby points on the learnt low-dimensional manifold. Maaten \[5\] gave a comparative review of some major manifold learning algorithms.

A common question that arises is how well a manifold learning algorithm learns the high-dimensional data, or in other words whether the algorithm reveals the underlying low-dimensional structure of the manifold. Unlike supervised learning tasks such as classification, where the performance of a classifier can be naturally evaluated through classification accuracy, manifold learning is unsupervised and does not have such default criteria. Some embedding quality assessment methods, as summarized in \[6\], have been proposed to measure the performance of the learnt embedding. For a manifold learning algorithm, these assessment methods usually return a score for the learnt low-dimensional representation as a measure of how well the algorithm recovers the original domain structure, up to some scaling and rotation.

However, learnt low-dimensional representation itself is rarely the goal, except when the ultimate task is data compression or data visualization. In most cases
researchers utilize manifold learning as the first step in a particular exploitation task. Thus, a proper assessment of a manifold learning algorithm can be derived from the performance of the subsequent inference in the learnt low-dimensional manifold.

A fundamental setting in statistics and machine learning contains $\mathcal{X}$, the space of data points and $\Theta$, the space of statistical generative models $\{p(x; \theta) : \theta \in \Theta\}$ or discriminative models $\{p(y; x, \theta) : \theta \in \Theta\}$. In the field of information geometry, $\Theta$ is treated as a manifold, known as the statistical manifold. For instance, univariate Gaussians can be parameterized as $\{\mathcal{N}(x; \mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$, and $\Theta = \{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$ is the corresponding statistical manifold.

Clearly, there are two possible manifold learnings, in $\mathcal{X}$ and in $\Theta$, respectively. When considering manifold learnings in $\mathcal{X}$, we say $\mathcal{X}$ is a data manifold. Consider a data set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, where each $x_i$ is represented in $D'$-dimension. Manifold learning constructs an embedding of $\mathcal{X}$ to a space of dimension $d'$ with $d' < D'$ and produces a $d'$-dimensional coordinate $y_i$ for each $x_i$. In data manifolds, manifold learning is widely studied and has been successfully applied to image data and other forms of high-dimensional data.

On the other hand, manifold learning in statistical manifolds has not been explored to the same extent. Unlike data manifolds whose points $x_i$ can be sampled directly from manifolds, we do not have access to many samples from statistical manifolds in practice. For instance, consider a hypothesis testing of a parametric distribution $p(x; \theta)$ for $\theta \in \Theta$, $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$, where $\theta_0 \in \Theta$. If no further information
CHAPTER 1. INTRODUCTION

is available, the Neyman-Pearson framework is typically used. Suppose $\Theta$ is not the optimally restricted space, but there exists an embedded submanifold $M \subset \Theta$, such that $\theta, \theta_0 \in M$. For example, if we consider the distribution of fourth grade boys’ allowances in a certain county using univariate Gaussians, then the distribution can be parameterized as $\mathcal{N}(x; \mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R}_0^+ \times \mathbb{R}^+$ with a correct restriction that $\mu$ is nonnegative.

If we can correctly restrict $\Theta$ to $M$, then one would expect that the power, or more generally, the inference, should be at least not worse, or perhaps even better by doing so. Al-Rawwash’s 1986 PhD dissertation [7] stated a general Likelihood Ratio Tests (LRTs) conjecture that “the more restrictions that are put on the alternative space, the higher the power of the LRT,” attributing it to a 1982 NSF proposal submitted by his advisor, J. Marden. However, Al-Rawwash [7] only studied Gaussians with conic submodels. In 1992, Tsai [8] stated that “the longtime conjecture of the power superiority of the restricted LRT to its unrestricted version in the entire parameter space of alternatives for the general setting is of considerably analytic difficulty and lacks definitive results.”

If the correctly restricted statistical submanifold $M \subset \Theta$ has a lower intrinsic dimension $d$ than $D$ (the dimension of $\Theta$), we say $M$ is a dimension-restricted submanifold of $\Theta$. Though literature such as [9, 7, 8, 10] has considered the power comparison between LRTs in $\Theta$ vs. LRTs in a restricted area of $\Theta$, it is unexplored when it comes to dimension-restricted $M$, which will be discussed later in Chapter [2]
CHAPTER 1. INTRODUCTION

Manifold learning on statistical manifolds to learn a more optimally dimension-restricted submanifold is possible if we have access to more points from $\mathcal{M}$. For instance, we consider an online learning task and assume samples from each time window $i$ are drawn from the distribution $p(x; \theta_i)$, where $\theta_i \in \mathbb{R}^D$ is from an unknown embedded submanifold $\mathcal{M}$ of intrinsic dimension $d$ of a known statistical manifold $\Theta$ ($D$-dim), so that $\theta_i \in \mathcal{M} \subset \Theta \subset \mathbb{R}^D$, for $i = 1, \ldots, m$. Let $\psi$ be the unknown smooth manifold mapping from $\mathbb{R}^d$ to $\mathbb{R}^D$.

We are interested in testing a hypothesis $H_0 : \theta^* = \theta_0$, conditioned upon $\theta^*, \theta_0 \in \mathcal{M}$, where $p(x; \theta^*)$ is the distribution from which samples from time window $m + 1$ are drawn. Then there exists unknown $u^*, u_0 \in \mathbb{R}^d$ such that $\theta^* = \psi(u^*)$ and $\theta_0 = \psi(u_0)$. The simple testing procedure is to consider the hypothesis testing task in $\mathbb{R}^D$. During time window $m + 1$, we observe samples $\{x_t\}_{t=1}^s$, which are i.i.d from $p(x; \theta^*)$, the distribution of interest. From $\{x_t\}_{t=1}^s$ we get an estimator $\hat{\theta}^* \in \mathbb{R}^D$ and reject $H_0$ if $\hat{\theta}^*$ is far from $\theta_0$. If $\mathcal{M}$ or $\psi$ is known, then we can correctly restrict the hypothesis and return to the dimension-restricted test comparison scheme above. On the other hand, when $\mathcal{M}$ is unknown but we can utilize some information to learn it, then we can conduct the testing on the learnt submanifold. If the manifold learning is good enough, then the testing on the learnt submanifold is expected to be close enough to that on $\mathcal{M}$.

During time window $i$, though $\theta_i$ is unknown, we have access to samples $\{x_{ij}\}_{j=1}^{n_{rand}}$ that are i.i.d from $p(x; \theta_i)$, with which we can easily get an estimator $\hat{\theta}_i \in \mathbb{R}^D$ for
CHAPTER 1. INTRODUCTION

\( \theta_i \). After having the collection \( \{\hat{\theta}_i\}_{i=1}^m \), which can be interpreted as noisy samples from \( \mathcal{M} \), if there is a uniformly small enough bound on \( \|\theta_i - \hat{\theta}_i\| \), then applying some manifold learning algorithm to \( \{\hat{\theta}_i\}_{i=1}^m \) provides a mapping estimate \( \psi^{-1} \) to \( \Psi^{-1} \). By considering the distance between \( \psi^{-1}(\hat{\theta}^*) \) and \( \psi^{-1}(\theta_0) \), the null hypothesis is rejected if the distance is great.

As one utilizes manifold learning in \( \mathcal{X} \) or in \( \Theta \), there are two fundamental questions to pose before drawing any inference conclusion: If the true embedded low-dimensional manifold is known, is the subsequent inference in the manifold superior to that in the high-dimensional ambient space; more specifically, is more power yielded in hypothesis testing or a smaller expected error in classification? If so, since in practice we always try to learn the low-dimensional manifold, does the subsequent inference in the learnt low-dimensional manifold recover that in the true manifold, or at least asymptotically so?

The answer to the first question seems to be yes in general. In 2003, however, Abu-Dayyeh, Al-Jararha, and Madan [9] constructed a counterexample to the above statement using Gaussian random variables with nonconic submodels of the form \( \mu \in [-k,k]^2 \) and an identity covariance matrix. Surprisingly, smaller values of \( k \) give more restricted alternatives, but not uniformly greater power by numerical calculations. This example, as a first counterexample to the longtime conjecture [8], considered the restricted model in the same dimension but in a smaller area, compared to the unrestricted model. In general, manifold learning algorithms learn a
CHAPTER 1. INTRODUCTION

lower-dimensional representation than that in which data are represented. Therefore, if we would like to answer the first fundamental question, it is necessary to dig into restricted models of smaller dimensionality compared to unrestricted models. We will discuss the counterexamples for LRTs and classification, concluding that, counterintuitively, the subsequent inference in the true manifold $\mathcal{M}$ has no guarantee of being superior to that in the ambient space $\Theta$.

The second question is surprisingly unexplored. Literature has been focusing on using subsequent inference, namely classification error rate, as criteria to compare different manifold learning algorithms [5]. Due to the lack of theoretical analysis on whether $\mathcal{I}_{\mathcal{M}} \rightarrow \mathcal{I}_{\mathcal{M}}$, in this work we will explore this matter in detail, mainly in Section 2.4 and Chapter 4.

We start by noting some basics of manifolds and submanifolds. Let $\mathcal{A}$ be a topological space and $U \subset \mathcal{A}$, $V \subset \mathbb{R}^D$ denote open sets. If $\psi : U \rightarrow V$ is a homeomorphism, then $\psi(u) = (\psi_1(u), \ldots, \psi_D(u))$ defines a coordinate system on $U$, and the functions $\psi_1, \ldots, \psi_D$ are the coordinate functions. The pair $(U, \psi)$ is called a chart on $\mathcal{A}$. The inverse map $\psi^{-1}$ is a parameterization of $U$.

An atlas on $\mathcal{A}$ is a collection of charts $\{U_\alpha, \psi_\alpha\}$ such that $U_\alpha$ cover $\mathcal{A}$. $\mathcal{A}$ is a $D$-dimensional topological manifold if it admits an atlas $\{U_\alpha, \psi_\alpha\}$. We say $\mathcal{A}$ is smooth if all partial derivatives of $\psi_\alpha$ exist and are continuous. A subset $\mathcal{M} \subset \mathcal{A}$ is a $d$-dimensional embedded submanifold if for every $x \in \mathcal{M}$, there exists a chart $(U, \psi)$ such that $x \in U$ and $\psi(U \cap \mathcal{M})$ is the intersection of a $d$-dimensional hyperplane with
CHAPTER 1. INTRODUCTION

ψ(U).

For linear dimensionality deduction methods, in general, the main goal is: for given \( n \) data points \( \{x_i\}_{i=1}^{n} \) drawn approximately from a \( d \)-dimensional affine subspace \( \mathcal{P} = \{x : x = \mu + Ry, y \in \mathbb{R}^d\} \), where \( \mu \in \mathbb{R}^D \) is an arbitrary point in \( \mathcal{P} \) and \( R \in \mathbb{R}^{D \times d} \) is a basis for \( \mathcal{P} \), to find \( \mu, R, \{y_i\}_{i=1}^{n} \) such that \( \sum_{i=1}^{n} \|x_i - \mu - Ry_i\| \) is minimized. We can say the collection \( \{y_i\}_{i=1}^{n} \) is the best linear representation of \( \{x_i\}_{i=1}^{n} \) in \( d \)-dimension.

The term “manifold learning”, sometimes referred to as “nonlinear dimensionality reduction” in machine learning, can be viewed as a generalization of linear dimensionality reduction. Manifold learning studies subspace not only linear or affine but in more complex forms. Analogous to linear dimensionality reduction, the main goal in manifold learning is: given \( \{x_i\}_{i=1}^{n} \) drawn approximately from a manifold of intrinsic dimension \( d \), manifold learning does not seek to learn the shape of \( \mathcal{M} \), but rather the low-dimensional representation or manifold coordinates \( \{u_i\}_{i=1}^{n} \), together with the parameterization \( \psi^{-1} \). Throughout this thesis, when we say the learnt manifold \( \hat{\mathcal{M}} \), we mean the estimated low-dimensional representation.

This study is organized as follows:

- In Chapter 1 we introduce manifold learning and provide an overview of related work on subsequence inference analysis on it. We ask two fundamental questions about manifold learning subsequent inference:

**Q1:** If the true low-dimensional manifold is known, is the subsequent inference in the manifold superior to that in the high-dimensional ambient space?
Q2: Does the subsequent inference in the learnt manifold recover that in the true manifold?

- In Chapter 2 we discuss the LRT power comparison of restricted models on the true manifold and unrestricted models in ambient space.

- In Chapter 3 we study the classification expected error comparison of restricted models on the true manifold and unrestricted models in ambient space.

- In Chapter 4 we focus on network models and Random Dot Product Graphs (RDPGs), in particular, the latent positions. Manifold learning and its subsequent inference in the estimated latent spaces of RDPGs are discussed.

- In Chapter 5 the results and their extensions are discussed from different perspectives.
Chapter 2

On the Power of Likelihood Ratio Tests (LRTs) in Dimension-Restricted Submodels

Manifold learning, as the first step in dealing with high-dimensional data in order to get a meaningful low-dimensional representation, bears the expectation that the subsequent inference could be superior. However, the fundamental question of whether this expectation is true is unexplored. In this chapter we first study $Q1$, which was raised in Chapter [1] and which asks whether the power of LRTs restricted on the true manifold is uniformly superior to that in ambient space. First, we assume the true submanifold where the parameter of interest lies is given and compare the LRT powers of restricted submodels and of unrestricted models. Here, we use the
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

Hardy-Weinberg equilibrium as an example for a nonlinear submanifold embedded in a two-dimensional simplex. We observe a phenomenon whereby the power of restricted LRTs is not uniformly superior to that of unrestricted LRTs. Second, we investigate $Q_2$, which asks whether the inference on the learnt submanifold yields equivalent power to that on the true submanifold. Finally, a power comparison of the two-sample LRTs is studied, and it is shown that the aforementioned phenomenon exists for linear submanifolds.

In Section 2.1, we review the literature on the LRT power comparison between restricted models and unrestricted models. Section 2.2 introduces asymptotic results of LRT power superiority of restricted models over unrestricted ones. The main theoretical results with respect to $Q_1$ are presented in Section 2.3.1 and simulation results with respect to $Q_2$ are presented in Section 2.4. Finally, Section 2.5 shows the theoretical results of the power comparison of two-sample LRTs.

2.1 Restricted LRT Power Superiority Literature Review

We start by reviewing the existing literature on the power superiority comparison between restricted LRTs and unrestricted LRTs. Al-Rawwash’s 1986 Ph.D. dissertation [7] offered a general LRT conjecture that “the more restrictions that are put on the alternative space, the higher the power of the LRT,” attributing it to a 1982 NSF
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

proposal submitted by his advisor, J. Marden. However, Al-Rawwash \[7\] only studied Gaussians with conic submodels. In 1992, Tsai \[8\] stated, “The longtime conjecture of the power superiority of the restricted LRT to its unrestricted version in the entire parameter space of alternatives for the general setting is of considerable analytic difficulty and lacks definitive results.” In Tsai’s work, a multivariate Gaussian model with known covariance matrix was considered, and the hypothesis was the mean vector. It has been shown that by restricting the alternative to the positive orthant space $\mathbb{R}^+$, the LRT yields more power than the unrestricted LRT.

The literature on the power superiority comparison between restricted LRTs and unrestricted LRTs has focused on multivariate Gaussian models. The most general result for Gaussian random variables is formulated by Praestgaard \[10\]:

Let $X \in \mathbb{R}^D$ and $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mathcal{N}$ denotes the Gaussian distribution and covariance $\Sigma$ is known. Let $\mathcal{C} \subset \mathbb{R}^D$ be a closed convex cone that contains a linear space $\mathcal{L} \subset \mathcal{C}$. Consider the restricted LRT for null and alternative hypotheses,

$$H_0: \mu \in \mathcal{L}, H_A: \mu \in \mathcal{C}/\mathcal{L}.$$ 

It has been proved that the restricted LRT is uniformly more powerful over $\mu \in \mathcal{C}$ than the unrestricted test,

$$H_0: \mu \in \mathcal{L}, H_A: \mu \in \mathbb{R}^D/\mathcal{L}.$$
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

However, Abu-Dayyeh et al. [9] constructed a counterexample to falsify the general LRT conjecture using Gaussian random variables with nonconic submodels of the form \( \mu \in [-k,k]^2 \) and identity covariance matrix. Surprisingly, smaller values of \( k \) give more restricted alternatives, but not uniformly greater power by numerical calculations. This counterexample restricts the alternative to a smaller region while maintaining its dimension. It is inviting to modify the general LRT conjecture and speculate that it holds if the restricted alternative has lower dimensions than the unrestricted. Despite considerable interest in Marden’s general LRT conjecture, we are not aware of any previous statements of the more plausible dimension-restricted LRT conjecture other than [11].

Initially, we formulate the dimension-restricted LRT problem under the popular setting with Gaussian random variables as follows:

Let \( X \in \mathbb{R}^D \) and \( X \sim \mathcal{N}(\mu, \Sigma) \), where \( \mu \in \mathcal{M} \subset \mathcal{A} \subset \mathbb{R}^D \), where \( \mathcal{A} \) is a manifold of dimension \( D \), and \( \mathcal{M} \) is an embedded submanifold of dimension \( d \) with \( D > d \). Consider the unrestricted LRT for null and alternative hypotheses,

\[
H_0 : \mu = \mu_0 \ vs. \ H_A : \mu \neq \mu_0, \mu \in \mathcal{A},
\]

and the restricted LRT,

\[
H_0 : \mu = \mu_0 \ vs. \ H_A : \mu \neq \mu_0, \mu \in \mathcal{M},
\]
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

where \( \mu_0 \in \mathcal{M} \).

We first consider a simple case where \( \mathcal{M} \) is a linear submanifold and \( \mathcal{A} \) is the \( D \)-dim Euclidean space. Let \( X \in \mathbb{R}^D \) and \( X \sim \mathcal{N}(\mu, \Sigma) \), where \( \mu \in \mathcal{P} \subset \mathbb{R}^D \), and where \( \mathcal{P} \) is a hyperplane of dimension \( d \) lying in \( D \) Euclidean space where \( D > d \). There exists a matrix \( R \in \mathbb{R}^{D \times d} \) such that \( \mu = Rw \) for some \( w \in \mathbb{R}^d \). The unrestricted LRT and the restricted LRT can be written as \( H_0 : \mu = \mu_0, H_A : \mu \neq \mu_0, \mu \in \mathbb{R}^D \) and \( H_0 : \mu = \mu_0, H_A : \mu \neq \mu_0, \mu \in \mathcal{P} \), respectively. Without loss of generality, we assume \( \mu_0 = 0 \). We prove that under this setting, the dimension-restricted LRT is superior to the unrestricted LRT.

**Theorem 1.** Let \( X \in \mathbb{R}^D \) and \( X \sim \mathcal{N}(\mu, \Sigma) \) with \( \Sigma \) known, where \( \mu \in \mathcal{P} \subset \mathbb{R}^D \) and \( \mu = Rw \) for some known matrix \( R \in \mathbb{R}^{D \times d} \). When testing \( H_0 : \mu = \mu_0 = Rw_0, H_A : \mu \neq \mu_0 \) with fixed type I error \( \alpha \), the power of the dimension-restricted LRT on \( \mathcal{P} \) is uniformly more than that of the unrestricted LRT at alternatives of form \( \mu_A = Rw_A \).

**Proof.** The unrestricted MLE of \( \mu \) is \( \hat{\mu}_{MLE} = X \). The unrestricted LRT formula is

\[
T_{RD}(X) = -2 \log \frac{L(x; \mu_0)}{L(x; \hat{\mu}_{MLE})} = X^T \Sigma^{-1} X,
\]

where \( L(x; \mu) \) is the likelihood function of a \( D \)-dim multivariate Gaussian random variable with known \( \Sigma \). Under the null hypothesis, the random variable \( T_{RD}(X) \) follows a \( \chi^2_D \) distribution. Then the critical value \( C_{RD} \) satisfies \( \mathbb{P}[\chi^2_D \geq C_{RD}] = \alpha \).
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

Suppose that we know that $\mu \in \mathcal{P}$, with which we can rewrite the hypothesis as $H_0 : \mu = \mu_0, H_A : \mathcal{P}/\mu_0$. The MLE of $w$ is $\hat{w}_{MLE} = (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} X$. Thus, the MLE of $\mu$ on $d$-dim $\mathcal{P}$ is $\hat{\mu}_{MLE,\mathcal{P}} = R \hat{w}_{MLE}$. The restricted LRT formula is

$$T_\mathcal{P}(X) = -2 \log \frac{L(x; \mu_0)}{L(x; \hat{\mu}_{MLE,\mathcal{P}})} = X^T \Sigma^{-1} X - (X - \hat{\mu}_{MLE,\mathcal{P}})^T \Sigma^{-1} (X - \hat{\mu}_{MLE,\mathcal{P}})$$

$$= (R^T \Sigma^{-1} X)^T (R^T \Sigma^{-1} R)^{-1} (R^T \Sigma^{-1} X).$$

Under the null hypothesis, the random variable $T_\mathcal{P}(X) \sim \chi_d^2$. The critical value $C_\mathcal{P}$ satisfies $\mathbb{P}[\chi_d^2 \geq C_\mathcal{P}] = \alpha$.

Now we consider the random variables $T_{\mathcal{D}}(X)$ and $T_\mathcal{P}(X)$ at alternatives of the form $\mu_A = Rw_A$ as

$$T_{\mathcal{D}}(X) \sim \chi_D^2(\delta_D^2), \text{ where } \delta_D^2 = (Rw_A)^T \Sigma^{-1} (Rw_A);$$

$$T_\mathcal{P}(X) \sim \chi_d^2(\delta_d^2), \text{ where } \delta_d^2 = (R^T \Sigma^{-1} Rw_A)^T (R^T \Sigma^{-1} R)^{-1} (R^T \Sigma^{-1} Rw_A).$$

Since $\delta_D^2 = \delta_d^2$, by Lemma 1 in the following page, the powers $\pi_{\mathcal{D}}(Rw_A; \alpha, Rw_0)$ and $\pi_\mathcal{P}(Rw_A; \alpha, Rw_0)$ have the following relationship:

$$\pi_{\mathcal{D}}(Rw_A; \alpha, Rw_0) = \mathbb{P}[\chi_D^2(\delta_D^2) > C_{\mathcal{D}}]$$

$$< \mathbb{P}[\chi_d^2(\delta_d^2) > C_\mathcal{P}] = \pi_\mathcal{P}(Rw_A; \alpha, Rw_0).$$
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

Lemma 1. \( \chi^2 \) (Das Gupta and Perlman [12]): Fix \( \alpha \in (0, 1) \) and \( \lambda > 0 \). For \( m \geq 1 \), define \( c_{m, \alpha} \) by \( \mathbb{P}(\chi_m^2 \geq c_{m, \alpha}) = \alpha \). Then

\[
\mathbb{P}(\chi_m^2(\lambda) \geq c_{m, \alpha})
\]

is strictly decreasing in \( m \).

Theorem 1 shows the power superiority of linearly restricted LRTs over unrestricted LRTs for Gaussian random variables. When the restricted alternative space is not linear but a nonlinear submanifold, whether Theorem 1 still holds is of considerable analytical difficulty and yet lacks definitive results. If we allow the testing sample size to go to infinity, then the well-known chi-square approximation of the LRT [13] indicates the power superiority of dimension-restricted LRTs [11]. To complete the theory, we state the asymptotic results from [11] in Section 2.2.

Another unexplored yet practical extension to current literature is to not only consider Gaussians but also other multivariate random variables. In Section 2.3 of this study we focus on multinomials.

2.2 Asymptotic Theory for LRTs

Let \( X_1, \ldots, X_n \) be independent and identically distributed as \( P_\theta \), and the statistical model \( \{P_\theta : \theta \in \Theta\} \) is indexed by a \( D \)-dimensional manifold \( \Theta \) while the statistical submodel \( \{P_\theta : \theta \in \Psi\} \) is indexed by a \( d \)-dimensional submanifold \( \Psi \subset \Theta \).
Suppose that $\Theta_0 \subset \Psi$ is an $d_0$-dimensional submanifold. For testing $H_0 : \theta \in \Theta_0$ vs. $H_A : \theta \notin \Theta_0$, the following theorem from [11] shows the power superiority of dimension-restricted LRTs over unrestricted LRTs:

**Theorem 2.** For $\theta_0 \in \Theta_0$ and $h$ such that $\theta_A = \theta_0 + \frac{h}{\sqrt{n}} \in \Psi$ for $n \geq N$. If $\Psi$ is an embedded submanifold of $\Theta$, then under alternative $\theta_A$, the restricted LRT has more power than the unrestricted LRT for a sufficiently large $n$.

The proof of this theorem is provided in [11]. The following section compares dimension-restricted LRTs and unrestricted LRTs for finite samples under the setting of multinomial models.

### 2.3 One-Sample LRTs in Multinomial Models

In this section, we use bold fonts $\mathbf{\theta}$ to represent a vector and $\theta_i$ to denote its entries.

Consider the general multinomial model, $\text{Multi}(\theta)$, where the outcomes occur with probability $\theta = (\theta_1, \ldots, \theta_D)$, which is parameterized by the unit simplex $\Delta^{D-1} \subset \mathbb{R}^D$, where $\Delta^{D-1} = \{ \theta \in [0,1]^D : \sum_{i=1}^D \theta_i = 1 \}$. Suppose the outcome of a drawing $n$ i.i.d. observations from $\text{Multi}(\theta)$ is $\mathbf{x} = (x_1, \ldots, x_D)$, where $x_i$ records the number of
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

occurrences of category \(i\). The unrestricted likelihood function of \(\theta\) is

\[
L(x; \theta) = \frac{n!}{\prod_{i=1}^{D} x_i!} \prod_{i=1}^{D} \theta_i^{x_i},
\]

and the unrestricted MLE of \(\theta\) is \(\hat{\theta} = (\frac{x_1}{n}, \ldots, \frac{x_D}{n})\). The unrestricted LRT rejects \(H_0 : \theta = \hat{\theta}\) if and only if

\[
\Lambda_{\Delta D - 1}(x) = -2 \log \frac{L(x; \hat{\theta})}{L(x; \theta)} = -2 \log \frac{\prod_{i=1}^{D} \hat{\theta}_i^{x_i}}{\prod_{i=1}^{D} (\frac{x_i}{n})^{x_i}}
\]

is sufficiently large.

2.3.1 Hardy-Weinberg Equilibrium

Consider a trinomial model which is a multinomial model with \(D = 3\). Define \(\psi : [0, 1] \to \Theta\) by \(\psi(\tau) = \{(\tau^2, 2\tau(1 - \tau), (1 - \tau)^2), \tau \in [0, 1]\} \subset \mathbb{R}^3\). The Hardy-Weinberg (HW) subfamily of trinomial distributions is parameterized by the embedded submanifold \(\mathcal{H} = \{\psi(\tau), \tau \in [0, 1]\}\).

The HW as a trinomial distribution is parameterized by \(\tau\) and its likelihood function is

\[
L(x; \psi(\tau)) = \frac{n!}{\prod_{i=1}^{3} x_i!} 2^{x_2} \tau^m (1 - \tau)^{2n - m},
\]

where \(m = 2x_1 + x_2\).
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

The MLE of $\tau$ is $\hat{\tau} = \frac{m}{2n}$ and the restricted MLE of $\theta$ in HW is $\tilde{\theta} = \psi(\hat{\tau})$. The restricted LRT rejects $H_0 : \theta = \tilde{\theta} = \psi(\tau_0)$ if and only if

$$\Lambda_{HW}(x) = -2 \log \frac{L(x; \hat{\theta})}{L(x; \tilde{\theta})} = -2 \log \frac{L(x; \psi(\tau_0))}{\hat{L}(x; \psi(\hat{\tau}))} = -2 \log \frac{\tau_0^m (1 - \tau_0)^{2n-m}}{\hat{\tau}^m (1 - \hat{\tau})^{2n-m}}$$

is sufficiently large. The unrestricted LRT $\Lambda_{\Delta^2}(x)$ is defined in (2.1) with $D = 3$.

We consider the powers $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ and $\pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ with respect to the unrestricted LRT and the restricted LRT. The asymptotic theory for LRTs stated in Section 2.2 says $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) < \pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ is true when considering $\tau_A$ in the neighborhood of $\tau_0$ when $n \to \infty$. We now investigate the relationship between two powers when $n < \infty$. Figure 2.1 shows a finite $n$ example where $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) < \pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ is no longer uniformly true.

![Figure 2.1](image)

Figure 2.1: The curve shows the power difference $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) - \pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ at alternatives $\tau_A \in (0, 1)$, when $n = 10, \tau_0 = 0.3, \alpha = 0.05$. The red portion shows where the difference is positive. Both powers are computed exactly in R to default significant digits.
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

Moreover, the phenomenon that $\pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ is not uniformly superior to $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ is far more general than the simple example in Figure 2.1.

**Theorem 3.** Let $\psi$ parameterize the Hardy-Weinberg submodel of the trinomial experiment with $n \geq 2$ and consider the null hypothesis $H_0 : \theta = \psi(\tau_0)$. For almost every $\tau_0 \in (0.2, 0.8)$, there exist $\alpha \in (0, 1)$ and $\tau_A \in (0, 1)$, such that the restricted LRT of size $\alpha$ is less powerful at alternative $\psi(\tau_A)$ than the unrestricted LRT of size $\alpha$.

The complete proof is provided in Section 2.6. To prove the theorem, the following conditions need to be satisfied:

**C1:** Suitable $\tau_0$ under which $\Lambda_{\Delta^2}(x)$ and $\Lambda_{HW}(x)$ order the possible outcomes differently;

**C2:** Type I error $\alpha$ is large enough such that $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ and $\pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ are not equal;

**C3:** The choice of $\tau_A$ under which $\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$ exceeds $\pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0))$.

In what follows, it will be convenient to work with the inverse likelihood ratios.

We define

$$T_{\Delta^2}(x) = \frac{L(x; \hat{\theta})}{L(x, \psi(\tau_0))} = \frac{(\frac{x_1}{n})^{x_1}(\frac{x_2}{n})^{x_2}(\frac{x_3}{n})^{x_3}}{2^{x_2} \tau_0 (1 - \tau_0)^{2n-m}}$$ (2.2)
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

and

\[
T_{HW}(\mathbf{x}) = \frac{L(\mathbf{x}; \psi(\hat{\tau}))}{L(\mathbf{x}; \psi(\tau_0))} = \frac{(\frac{m}{2n})^m (1 - \frac{m}{2n})^{2n-m}}{\tau_0^m (1 - \tau_0)^{2n-m}} \tag{2.3}
\]

with respect to \( \Lambda_{\Delta^2}(\mathbf{x}) \) and \( \Lambda_{HW}(\mathbf{x}) \).

We first verify C1. Consider ordering all unique \( T_{\Delta^2}(\mathbf{x}) \) from the largest to the smallest of all possible outcomes \( \mathbf{x} \). Let \( T^{(i)}_{\Delta^2} \) denote the \( i^{th} \) element in the ordered sequence of all unique \( T_{\Delta^2}(\mathbf{x}) \) values. The ordering of their \( T_{\Delta^2}(\mathbf{x}) \) values defines the order of all possible outcomes \( \mathbf{x} \). We call such an ordering of all \( \mathbf{x} \) “the order of all outcomes in the unrestricted LRT,” and we say \( \mathbf{x}_1 \succ \mathbf{x}_2 \) in the unrestricted LRT if \( T_{\Delta^2}(\mathbf{x}_1) > T_{\Delta^2}(\mathbf{x}_2) \). The ordered sequence of outcomes can be written as \( \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \).

It is worth pointing out that there might be two (or more) different possible outcomes \( \mathbf{x}_1, \mathbf{x}_2 \) such that \( T_{\Delta^2}(\mathbf{x}_1) = T_{\Delta^2}(\mathbf{x}_2) \), in which case we say their orders are interchangable. For example, consider \( \mathbf{x}_1, \mathbf{x}_2 \) such that \( T_{\Delta^2}(\mathbf{x}_1) = T_{\Delta^2}(\mathbf{x}_2) \), then both \( \mathbf{x}_1 \succ \mathbf{x}_2 \) and \( \mathbf{x}_1 \prec \mathbf{x}_2 \) are valid and will both be evaluated when referring to “the order of all outcomes in the unrestricted LRT.”

Let \( O^i_{\Delta^2} \) denote the \( i^{th} \) level set of all outcomes \( \mathbf{x} \) that yield the same \( T^{(i)}_{\Delta^2} \). We define \( \alpha^J_{\Delta^2} \), “the type I error that contains outcomes till level \( J \) for the unrestricted LRT in \( \Delta^2 \),” as
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

\[ \alpha_{\Delta^2}^J = \mathbb{P}_{\tau_0} \left( \bigcup_{i=1}^{J} O_{\Delta^2}^i \right) = \sum_{i=1}^{J} \mathbb{P}_{\tau_0} (O_{\Delta^2}^i). \]  

Then for testing \( H_0 : \theta = \psi(\tau_0) \), the critical region of the unrestricted LRT of size \( \alpha_{\Delta^2}^J \) is

\[ C_{\Delta^2}(\alpha_{\Delta^2}^J) = \bigcup_{i=1}^{J} O_{\Delta^2}^i; \]  

i.e., the unrestricted LRT rejects \( H_0 \) if and only if \( x \in C_{\Delta^2}(\alpha_{\Delta^2}^J) \).

For a specific \( \alpha \in (0,1) \), we define \( J^* = \max_j \{ \alpha \geq \alpha_{\Delta^2}^j \} \). The rule for the unrestricted LRT to reject \( H_0 \) of size \( \alpha \) can be written as:

\[
\mathbb{P}_{\tau_0}[\text{rejects } H_0 \text{ given } x] = \begin{cases} 
1 & \text{if } x \in C_{\Delta^2}(\alpha_{\Delta^2}^{J^*}) \\
\frac{\alpha - \alpha_{\Delta^2}^{J^*}}{\alpha_{\Delta^2}^{J^*+1} - \alpha_{\Delta^2}^{J^*}} & \text{if } x \in O_{\Delta^2}^{J^*+1} \\
0 & \text{otherwise.}
\end{cases}
\]

Similarly for the restricted LRT, we define the ordered test statistics sequence \( T_{HW}^{(i)} \) of all unique \( T_{HW}(x) \) values in decreasing order and “the order of all outcomes in the restricted LRT” with respect to \( x \). The ordered sequence of outcomes can be written as \( x_{HW}^{(1)}, x_{HW}^{(2)}, \ldots \).

Let \( O_{HW}^i \) denote the \( i \)th level set of all outcomes \( x \) that yield \( T_{HW}^{(i)} \). We define
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

\(\alpha^J_{HW}\), “the type I error that contains outcomes till level \(J\) for the restricted LRT in HW,” as

\[
\alpha^J_{HW} = \mathbb{P}_{\tau_0} \left( \bigcup_{i=1}^{J} O^i_{HW} \right) = \sum_{i=1}^{J} \mathbb{P}_{\tau_0}(O^i_{HW}). \tag{2.6}
\]

Then for testing \(H_0 : \theta = \psi(\tau_0)\), the critical region of the restricted LRT of size \(\alpha^J_{HW}\) is

\[
C_{HW}(\alpha^J_{HW}) = \bigcup_{i=1}^{J} O^i_{HW}; \tag{2.7}
\]

i.e., the restricted LRT rejects \(H_0\) if and only if \(x \in C_{HW}(\alpha^J_{HW})\).

For a specific \(\alpha \in (0, 1)\), we define \(J^{**} = \max_j \{ \alpha \geq \alpha^j_{HW} \}\). The rule for the restricted LRT to reject \(H_0\) of size \(\alpha\) can be written as:

\[
\mathbb{P}_{\tau_0}[\text{rejects } H_0 \text{ given } x] = \begin{cases} 
1 & \text{if } x \in C_{HW}(\alpha^{J^{**}}_{HW}) \\
\frac{\alpha - \alpha^{J^{**}}_{HW}}{\alpha^{J^{**}+1}_{HW} - \alpha^{J^{**}}_{HW}} & \text{if } x \in O^{J^{**}+1}_{HW} \\
0 & \text{otherwise.}
\end{cases} \tag{2.8}
\]

Now we investigate Condition I for Theorem 3. From Equation (2.3), \(T_{HW}(x)\) as a function of \(x\) depends only on \(m\), or in other words a level set \(O^i_{HW}\) contains all outcomes \(x\) with the same \(m\) value. Suppose there exists some \(\tau_0 \in (0, 1)\), \(m_1 \neq m_2\) and \(m_1, m_2 \in [0, 2n] \cap \mathbb{Z}\), such that some level set \(O^i_{HW}\) contains all outcomes \(x\) of
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

value \( m_1 \) or \( m_2 \); i.e.,

\[
\left( \frac{m_1}{2n} \right)^{m_1} \frac{(1 - \frac{m_1}{2n})^{2n-m_1}}{\tau_0^{m_1}(1 - \tau_0)^{2n-m_1}} = \left( \frac{m_2}{2n} \right)^{m_2} \frac{(1 - \frac{m_2}{2n})^{2n-m_2}}{\tau_0^{m_2}(1 - \tau_0)^{2n-m_2}},
\]

which can be written as

\[
\left( \frac{\tau_0}{1 - \tau_0} \right)^{m_1-m_2} = \frac{m_1^{m_1}(2n - m_1)^{2n-m_1}}{m_2^{m_2}(2n - m_2)^{2n-m_2}}.
\]

(2.9)

We observe that the number of all ways to choose the pair \((m_1, m_2)\) is \(n(2n + 1)\) and that there is at most one choice of \(\tau_0\) associated with each \((m_1, m_2)\) pair. Thus, there are at most \(n(2n + 1)\) choices of \(\tau_0\), under which there exists some level set \(O_{iHW}^i\) that contains all outcomes \(x\) with non-unique \(m\) values. Thus, the following lemma is true:

**Lemma 2.** Let \( E = \{\tau_0 \in (0, 1) : O_{iHW}^i \text{ is associated with a single } m, \text{ for all } i\} \).

Then the set \( E \) has measure 1.

On the other hand, consider \( x = (x_1, x_2, x_3) \) and \( y = (y_1, y_2, y_3) \) such that \( T_{\Delta^2}(x) = T_{\Delta^2}(y) \), and \( m_1 = 2x_1 + x_2 \) and \( m_2 = 2y_1 + y_2 \); i.e.,

\[
\frac{(\frac{x_1}{2n})^{x_1}(\frac{x_2}{2n})^{x_2}(\frac{x_3}{2n})^{x_3}}{2^{x_2} \tau_0^{m_1}(1 - \tau_0)^{2n-m_1}} = \frac{(\frac{y_1}{2n})^{y_1}(\frac{y_2}{2n})^{y_2}(\frac{y_3}{2n})^{y_3}}{2^{y_2} \tau_0^{m_2}(1 - \tau_0)^{2n-m_2}},
\]
which can be rewritten as
\[
\frac{(2x_1)^{x_1}(x_2)^{x_2}(2x_3)^{x_3}}{(2y_1)^{y_1}(y_2)^{y_2}(2y_3)^{y_3}} = \left( \frac{\tau}{1 - \tau} \right)^{m_1 - m_2}.
\] (2.10)

Similar to the argument in Lemma 2, let \( E' = \{ \tau_0 \in (0, 1) : O_{D_2}^i \text{ is associated with a single } m, \text{ for all } i \} \). Then the set \( E' \) has measure 1 as well. The proof of Theorem 3 needs \( \tau_0 \in E \) and argues that the occurrence of \( O_{D_2}^i \) associated with multiple \( m \) does not jeopardize the proof.

For C1, we establish the following conditions under which the unrestricted LRT and the restricted LRT order the possible outcomes differently; i.e., where there exist \( x \) and \( y \), such that \( T_{HW}(x) > T_{D_2}(x) \) and \( T_{HW}(y) < T_{D_2}(y) \):

**Lemma 3.** The unrestricted LRT and the restricted LRT order the possible outcomes differently under any of the following conditions:

1. \( \tau_0 \in (0.2, \frac{1}{3}) \cup (\frac{2}{3}, 0.8) \);
2. \( \tau_0 \in (\frac{1}{3}, \frac{2}{3}) \) and \( n \geq 2 \);
3. \( \tau_0 \in \{ \frac{1}{3}, \frac{2}{3} \} \) and \( n \geq 3 \);

**Proof.** For \( \tau_0 < \frac{1}{3} \), the unrestricted LRT orders \((0, n, 0) \succ (0, 0, n) \) in that

\[
T_{D_2}((0, n, 0)) = \frac{n^n}{2^n (\frac{\tau_0}{1 - \tau_0})^n} > \frac{n^n}{2^n (\frac{\tau_0}{1 - \tau_0})^0} = T_{D_2}((0, 0, n))
\]

\[
\iff 2 \left( \frac{\tau_0}{1 - \tau_0} \right) < 1,
\]

25
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

and for $\frac{1}{5} < \tau_0 < \frac{1}{2}$, the restricted LRT orders $(0, n, 0) \prec (0, 0, n)$ as

$$T_{HW}((0, n, 0)) = \frac{n^n n^n}{(\frac{\tau_0}{1-\tau_0})^n} < \frac{(2n)^{2n}}{(\frac{\tau_0}{1-\tau_0})^0} = T_{HW}((0, 0, n))$$

$$\iff 2^2 \left( \frac{\tau_0}{1-\tau_0} \right) > 1.$$

Thus, for $\tau_0 \in (0.2, \frac{1}{3})$, the unrestricted and restricted LRTs order the possible outcomes differently for $n \geq 1$. By symmetry, the unrestricted and restricted orders also differ if $\tau_0 \in (\frac{2}{3}, 0.8)$. The first condition is justified.

Second, assume that $n \geq 2$. For $\frac{1}{3} < \tau_0 < \frac{1}{2}$, the unrestricted LRT orders $(n-1, 1, 0) \prec (n-1, 0, 1)$ because

$$T_{\Delta^2}((n-1, 1, 0)) = \frac{(n-1)^{n-1}}{2(\frac{\tau_0}{1-\tau_0})^{2n-1}} < \frac{(n-1)^{n-1}}{2^0(\frac{\tau_0}{1-\tau_0})^{2n-2}} = T_{\Delta^2}((n-1, 0, 1))$$

$$\iff 2^2 \left( \frac{\tau_0}{1-\tau_0} \right) > 1.$$

On the other hand, the restricted LRT orders $(n-1, 1, 0) \succ (n-1, 0, 1)$ if $\tau_0 \leq 0.5$, in that

$$T_{HW}((n-1, 1, 0)) = \frac{(2n-1)^{2n-1}/(2n)^{2n}}{(\frac{\tau_0}{1-\tau_0})^{2n-1}} > \frac{(2n-2)^{2n-2}/(2n)^{2n}}{(\frac{\tau_0}{1-\tau_0})^{2n-2}} = T_{HW}((n-1, 0, 1))$$

$$\iff \frac{(2n-1)^{2n-1}}{(2n-2)^{2n-2}} < \frac{\tau_0}{1-\tau_0},$$

and the left side of the above inequality as a function of $n$ achieves its minimal when
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

\( n = 2; \text{ i.e.,} \)

\[
\min_{n \geq 2} \frac{(2n - 1)^{2n-1}}{(2n - 2)^{2n-2}2^2} = \frac{3^3}{2^22^2}, \\
\implies \frac{27}{16} > \frac{\tau_0}{1 - \tau_0}.
\]

It follows that the unrestricted and restricted LRTs order all possible outcomes differently if \( \tau_0 \in (\frac{1}{3}, 0.5] \) for \( n \geq 2 \), and by symmetry also if \( \tau_0 \in [0.5, \frac{2}{3}] \). Thus, the second condition is justified.

Third, let \( \tau_0 = \frac{1}{3} \) and \( n \geq 3 \). The unrestricted LRT orders \((1, 0, n - 1) \succ (0, 1, n - 1)\), for

\[
T_{\Delta^2}((1, 0, n - 1)) = \frac{(n - 1)^{n-1}/n^n}{(\frac{1}{3})^2(\frac{2}{3})^{2n-2}} > \frac{(n - 1)^{n-1}/n^n}{(2\frac{1}{3})^1(\frac{2}{3})^{2n-1}} = T_{\Delta^2}((0, 1, n - 1))
\]

\( \iff \frac{1}{3}^2(\frac{2}{3})^{2n-2} < 2(\frac{1}{3})^1(\frac{2}{3})^{2n-1}. \)

In contrast, the restricted LRT orders \((1, 0, n - 1) \prec (0, 1, n - 1)\) since

\[
T_{HW}((1, 0, n - 1)) = \frac{2^2(2n - 2)^{2n-2}/(2n)^{2n}}{(\frac{1}{3})^2(\frac{2}{3})^{2n-2}} > \frac{(2n - 1)^{2n-1}/(2n)^{2n}}{(2\frac{1}{3})^1(\frac{2}{3})^{2n-1}} = T_{HW}((0, 1, n - 1))
\]

\( \iff \frac{8(2n - 2)^{2n-2}}{(2n - 1)^{2n-1}} > 1, \text{ for } n \geq 3. \)

By symmetry, the unrestricted and restricted LRTs order all possible outcomes dif-
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

ferently when \( \tau_0 = \frac{2}{3} \) and \( n \geq 3 \). The last condition is verified.

The above lemma provides the region of \( \tau_0 \) satisfying \( C_1 \) for a very small \( n \). On the other hand, we observe that as \( n \) grows, the region of \( \tau_0 \) satisfying \( C_1 \) grows.

**Lemma 4.** For any \( \tau_0 \in (0, 1) \), there exists \( N_0 \), such that for all \( n \geq N_0 \), the unrestricted LRT and the restricted LRT order the possible outcomes differently.

We define \( L = (2, 1, 0)^T \) for the following:

**Proof.** When \( \tau_0 \in (0.2, 0.8) \), Lemma 3 defines \( N_0 \) explicitly.

When \( \tau_0 \leq 0.2 \), let \( x = (0, n, 0) \) and \( y = (1, n - 1, 0) \). Then \( 2n\tau_0 \leq 0.4n < L^Tx = n < n+1 = L^Ty \). Since \( T_{HW} \) as a function of \( m \) achieves its minimum when \( m \) is equal to \( 2n\tau_0 \), while decreasing when \( m \in [0, 2n\tau_0] \) and increasing when \( m \in [2n\tau_0, 2n] \), it can be easily concluded that \( T_{HW}(x) < T_{HW}(y) \).

On the other hand, for the unrestricted LRT, if \( T_{\Delta^2}(x) > T_{\Delta^2}(y) \), then we have

\[
T_{\Delta^2}(x) > T_{\Delta^2}(y)
\iff \left( \frac{n}{2} \right)^n \left( \frac{1 - \tau_0}{\tau_0} \right)^n > \left( \frac{n-1}{2} \right)^{n-1} \left( \frac{1 - \tau_0}{\tau_0} \right)^{n-1}
\iff \left( \frac{n}{n-1} \right)^n > \frac{1 - \tau_0}{\tau_0} \frac{2}{n-1}.
\]

For \( \tau_0 \leq 0.2 \), there exists \( N_0 \), such that for all \( n \geq N_0 \), the inequality (2.11) holds.

By symmetry, this is true for \( \tau_0 \geq 0.8 \) as well.

\( \square \)
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

After discussing \( C_1 \), we now shift our focus to \( C_2 \). If \( \alpha \) is chosen such that two tests have the same critical region; i.e., \( C_{HW}(\alpha) = C_\Delta(\alpha) \), then \( \pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) = \pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) \). Lemma 3 indicates that if any of the conditions are met, then there exists \( \alpha \), such that \( C_{HW}(\alpha) \neq C_\Delta(\alpha) \) and potentially, \( \pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) \neq \pi_{HW}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) \). \( C_3 \) is discussed in the proof of Theorem 3.

The phenomenon in Theorem 3 is more general than trinomial. We recall that in biology, the Hardy-Weinberg equilibrium describes allele and genotype frequencies in a population and can be generalized for more than two alleles and polyploidy. For instance, consider a three-allele case, which can be parameterized by \( \psi_A : [0, 1]^2 \to \Theta \) by \( \Psi_A = \{ (\tau^2, \nu^2, \eta^2, 2\tau\nu, 2\tau\eta, 2\eta\nu) : \tau, \nu, \eta \in [0, 1], \tau + \nu + \eta = 1 \} \subseteq \Delta^5 \subseteq \mathbb{R}^6 \). Then the proof for Theorem 3 can be easily generalized to submodels \( \Psi_A \) of \( \Delta^5 \) and to the Hardy-Weinberg equilibrium of more than two alleles and polyploidy. Thus, the phenomenon that the dimension-restricted LRT is not uniformly superior to the unrestricted LRT stated in Theorem 3 is not unique to the trinomial.

As a submodel of the trinomial that is a two-parameter exponential family distribution \[14\], the Hardy-Weinberg model itself parameterizes a regular one-parameter exponential family distribution after reparameterization \( \eta = \frac{\tau}{1-\tau} \). It is obviously a curve in the sense of geometric curvature, while its statistical curvature \[15\] is zero.

One may suspect that the phenomenon is a result of the geometric curvature. For this reason, next we consider a submanifold of zero geometric curvature – a linear manifold, instead of Hardy-Weinberg, as the submodel of the trinomial and
demonstrate that the phenomenon still exists. It is also worth pointing out that the phenomenon in Theorem 3 is not unique to Hardy-Weinberg models.

2.3.2 Linear Restricted Submodels

We define $L : [0, 1] \to \Theta$ by $L = \{(\tau, a\tau, 1 - (a + 1)\tau) : \tau \in [0, \frac{1}{a+1}]\} \subset \Delta^2 \subset \mathbb{R}^3$, for some constant $a > 0$. Consider a linear subfamily of trinomial distributions parameterized by the embedded submanifold $\Psi_L = \{\psi_L(\tau), \tau \in [0, 1, \frac{1}{a+1}]\}$.

It is clear that $\Psi_L$ has zero statistical curvature and zero geometric curvature.

**Theorem 4.** Let $\psi_L$ parameterize a linear submodel of the trinomial experiment with $n \geq 2$ and consider the null hypothesis $H_0 : \theta = \psi_L(\tau_0)$. For almost every $\tau_0 \in (0, 1)$, there exist $\alpha \in (0, 1)$ and $\tau_A \in (0, 1)$, such that the restricted LRT of size $\alpha$ is less powerful at alternatives $\psi_L(\tau_A)$ than the unrestricted LRT of size $\alpha$.

The proof of Theorem 4 is analogous to that of Theorem 3. Theorem 4 can be generalized to any 1-dim linear submodel of multinomial models where $D \geq 3$.

2.4 $\pi_M$ vs. $\hat{\pi_M}$

To answer Q1 raised in Chapter 1 Sections 2.1 - 2.3 compare the power of restricted LRTs and that of unrestricted LRTs. Now we shift our focus to Q2. In this section, we study Q2 in hypothesis testing, in particular in regards to the one-sample testing problem. We start by defining the problem properly:
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

Let \( \mathcal{F} = \{ f(x; \theta), x \in \mathbb{R}^D, \theta \in \Theta \} \) be a parametric model, with \( \Theta \subset \mathbb{R}^D \) being the statistical manifold. Let \( \mathcal{M} \subset \Theta \) be a \( d \)-dimensional smooth, closed and compact submanifold. So there exists a corresponding manifold mapping \( \psi : U \rightarrow \Theta \), where \( U \subset \mathbb{R}^d \). We consider the one-sample hypothesis testing: \( H_0 : \theta^* = \hat{\theta} \) vs. \( H_A : \theta^* \neq \hat{\theta} \).

If \( \mathcal{M} \) is known, then we conduct restricted tests, although they are not necessarily more powerful than unrestricted tests as we discussed in Section 2.3.1, where we considered the trinomial model with \( D' = 3, D = 2, d = 1 \). When \( \mathcal{M} \) is unknown and we only have samples \( \{ x_t \}_{t=1}^s \) i.i.d. drawn from \( f(x; \theta^*) \), then the only option is to perform unrestricted tests as we have no other information about \( \mathcal{M} \) to help restrict alternatives. We denote \( \hat{\theta}^* \) as the unrestricted estimator of \( \theta^* \).

We assume that the existence and \( d \) of \( \mathcal{M} \) is known, but given \( \{ x_{ij} \}_{j=1}^{n_{rand}} \) that are uniformly drawn from \( f(x; \theta_i) \) for \( i = 1, \ldots, m \), where \( \{ \theta_i \}_{i=1}^m \) are uniformly chosen on \( \mathcal{M} \) and \( \{ x_t \}_{t=1}^s \) i.i.d. drawn from \( f(x; \theta^*) \), where \( \theta^* \) is the parameter of testing interest. \( \theta_i \) are assumed unknown; otherwise, we would not be concerned with \( \{ x_{ij} \}_{j=1}^{n_{rand}} \). An example is discussed in Chapter 1. For each \( \theta_i \), there is an unrestricted estimator \( \hat{\theta}_i \) for \( i = 1, \ldots, m \).

If \( \mathcal{M} \) is unknown, we can learn it efficiently with \( \hat{\theta}_i \) via manifold learning algorithms and test it with the learnt low-dimensional representation. We call such a test a learnt restricted test. Thus, Q2 asks whether \( \pi_{\mathcal{M}_1} \), the power of the learnt restricted test, converges asymptotically to \( \pi_{\mathcal{M}_1} \), the power of the corresponding restricted test. If the error between \( \hat{\theta}_i \) and \( \theta_i \) is sufficiently small or more strongly \( \hat{\theta}_i \overset{d}{\rightarrow} \theta_i \), then we
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

would expect $\pi_{\mathcal{M}} \to \pi_M$.

In the following lines, we elaborate on the procedure of learnt restricted tests. Suppose the $\theta_i$, which are precisely on $\mathcal{M}$, are unknown. Given samples $\{x_{ij}\}_{j=1}^{n_{\text{rand}}}$ from $f(x; \theta_i)$, we are easily able to get an estimator $\hat{\theta}_i$ for $\theta_i$ for each $i = 1, \ldots, m$.

From testing samples $\{x_{\ell}\}_{\ell=1}^s$, an unrestricted estimator $\hat{\theta}^*$ can be obtained easily. The collection $\{\hat{\theta}_i\}_{i=1}^m$ and $\hat{\theta}^*$ can be interpreted as noisy samples from $\mathcal{M}$. If there exists a uniformly sufficient small bound on $||\theta_i - \hat{\theta}_i||$, a corresponding reasonable approximation $\hat{\psi}^{-1}$ to $\psi^{-1}$ is expected. By considering the distance between $\hat{\psi}^{-1}(\hat{\theta}^*)$ and $\hat{\psi}^{-1}(\theta_0)$, the null hypothesis is rejected if the distance is great.

We focus on multinomial distributions in the following subsection to demonstrate some results of the power comparison between restricted tests and learnt restricted tests.

2.4.1 Multinomial Models

Consider when $\mathcal{F}$ is the multinomial family $\text{Multi}(\theta)$ indexed by $\theta \in \Theta$, and a one-sample testing problem $H_0 : \theta^* = \hat{\theta}$ vs. $H_A : \theta^* \neq \hat{\theta}$, where $\hat{\theta} \in \mathcal{M} \subset \Theta$.

Assume $\mathcal{M}$ is an unknown submanifold of intrinsic dimension $d < D$ embedded in $\Theta$, a known statistical manifold of dimension $D$, such as a simplex $\Delta^D$. Let $\psi : U \to \Theta$ parameterize the embedded submanifold $\mathcal{M}$, where $U \subset \mathbb{R}^d$. The one-sample hypothesis can be rewritten as $H_0 : \theta^* = \psi(u_0)$ vs. $H_0 : \theta^* \neq \psi(u_0)$, where $u_0 \in U$ satisfies $\psi(u_0) = \hat{\theta}$.
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED
SUBMODELS

For any $\theta_i \in \mathcal{M}$, its learnt embedding $\psi^{-1}(\hat{\theta}_i)$ is no longer in $U$. Thus, instead of
LRTs, we consider a more general test, the information test, to measure the distance
between two points on a statistical manifold, which is also valid for $\hat{u}_i = \psi^{-1}(\hat{\theta}_i) \notin U$.

We start by recalling that the Hellinger distance between two probability densities
$f$ and $g$ is given by

$$D_H(f, g) = \left[ \int (\sqrt{f} - \sqrt{g})^2 \right]^{\frac{1}{2}} = (2 - 2 \int (fg)^{\frac{1}{2}})^{\frac{1}{2}},$$

and the integral is taken with respect to the domain measure. Let $\rho(f, g) = \int (fg)^{\frac{1}{2}}$.

The minimum Hellinger distance estimator (MHDE) of $\theta$, which parameterizes a
density $f_\theta$, is defined to be $\hat{\theta}_{MHDE} = \arg\min_{\theta \in \Theta} D_H^2(f_\theta, \hat{f}_n)$, where $\hat{f}_n$ is the empirical
density for a discrete distribution $f_\theta$.

The MHDE is known to be asymptotically equivalent to the maximum likelihood
estimator for a parametric multinomial distribution [16]. Simpson [17] constructed a
Hellinger deviance test that is asymptotically equivalent to the LRT $\Lambda_{D-1}$ in

$$H_n = 8n[\rho(f_{\hat{\theta}_{MHDE}}, \hat{f}_n)] \xrightarrow{p} -2\frac{\log L(x; \hat{\theta}_{MLE})}{\log L(x; \theta)} = \Lambda_{D-1}(x),$$

where $n$ is the sample size.

Kass [18] showed the information distance $D_I$ between $\theta$ and $\hat{\theta}$ is equivalent to the
minimum geodesic distance $D_{G,S}$ on the spherical reparameterization system where
$\theta = (\theta_1, \ldots, \theta_D)$ is reparameterized as $z(\theta) = (2\sqrt{\theta_1}, \ldots, 2\sqrt{\theta_D})$ and $D_I$ is a transfor-
mation of Hellinger distance $D_H$ via

$$D_H(\theta, \hat{\theta}) = 2 \sin(D_I(\theta, \hat{\theta})/4).$$

It can be shown via geometry that $D_I(\theta, \hat{\theta}) = D_{G,S}(\theta, \hat{\theta}) = 2 \arccos \sum_{i=1}^{D} \sqrt{\theta_i \hat{\theta}_i}$.

We define the test statistics of information tests in $\Theta$ and on $\mathcal{M}$ to be the information distance with the unrestricted MHDE and the restricted MHDE, respectively. For $\theta^*$, let $\hat{\theta}^{\ast}_{MHDE, \Theta}$ denote its MHDE in $\Theta$ and $\hat{\theta}^{\ast}_{MHDE, \mathcal{M}}$ denote its MHDE restricted in $\mathcal{M}$. The restricted MHDE can be written as $\hat{\theta}^{\ast}_{MHDE, \mathcal{M}} = \psi(\hat{u}^{\ast}_{MHDE})$, where $\hat{u}^{\ast}_{MHDE}$ is the minimum Hellinger distance estimator of $u^*$ while assuming $\theta^* = \psi(u^*) \in \mathcal{M}$.

It is obvious that $\hat{\theta}^{\ast}_{MHDE, \Delta^0-1}$ equals the unrestricted MLE in simplex, defined in Section 2.3. Thus, the new unrestricted and restricted tests with respect to Hellinger distance, respectively, are defined as

$$T_{\Theta} = D_I(\hat{\theta}^{\ast}_{MHDE, \Theta}, \hat{\theta}) \quad (2.12)$$

$$T_{\mathcal{M}} = D_I(\hat{\theta}^{\ast}_{MHDE, \mathcal{M}}, \hat{\theta}) = D_I(\hat{u}^{\ast}_{MHDE}, u_0). \quad (2.13)$$

Now we consider the learnt restricted test, which is a test defined on $\hat{\mathcal{M}}$. We expect learnt restricted tests to be asymptotically equivalent to a transformation of restricted tests. In this work, we consider learnt restricted tests defined by Isomap since Isomap provides bounded errors on the difference between learnt pairwise geodesic distances and the corresponding true geodesic distances. Learnt restricted tests have
the following procedure:

1. We are given samples \( \{x_{ij}\}_{j=1}^{nr\text{and}} \) from \( f(x; \theta_i) \), with which it is easy to obtain an estimator \( \hat{\theta}_i \) for \( \theta_i \), for \( i = 1, \ldots, m \). For instance, we consider \( \hat{\theta}_i \) being the MLE. Such unrestricted estimators \( \hat{\theta}_i \) are in \( \Theta \).

2. We then apply Isomap to \( \{\hat{\theta}_i\}_{i=1}^m \) which can be interpreted as noisy samples from \( \mathcal{M} \), to produce an approximation \( \tilde{\psi}^{-1} \) to \( \psi^{-1} \). Let \( \hat{Z}_i = \tilde{\psi}^{-1}(\hat{\theta}_i) \).

3. Now we apply the out-of-sample embedding extension \[19\] to the estimate \( \hat{\theta}^* \in \Theta \) from \( \{x_{i\ell}\}_{\ell=1}^s \) (e.g. the MLE) to obtain a \( d \)-dim estimate \( \hat{Z}^* = \tilde{\psi}^{-1}(\hat{\theta}^*) \) and a \( d \)-dim embedding \( \hat{Z}_0 = \tilde{\psi}^{-1}(\hat{\theta}) \). We reject the null hypothesis if the distance between \( \hat{Z}^* \) and \( \hat{Z}_0 \) is sufficiently large.

Alternatively, for simplicity, we combine Steps 2 and 3 into one:

2&3: We apply Isomap to \( \{\hat{\theta}_i\}_{i=1}^m, \hat{\theta}^* \) and \( \hat{\theta} \) to produce an approximation \( \tilde{\psi}^{-1} \). Then we have the \( d \)-dim embedding \( \hat{Z}^* = \tilde{\psi}^{-1}(\hat{\theta}^*) \) and \( \hat{Z}_0 = \tilde{\psi}^{-1}(\hat{\theta}) \) for \( \hat{\theta} \). We reject the null hypothesis if the distance between \( \hat{Z}^* \) and \( \hat{Z}_0 \) is sufficiently large.

In what follows, we consider 2&3 to avoid out-of-sample embedding analysis.

The test statistic for the learnt restricted test is defined as

\[
T_{\mathcal{M}} = \|\hat{Z}^* - \hat{Z}_0\|. \tag{2.14}
\]
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

The learnt restricted test formula \( (2.14) \) is approximately the learnt geodesic distance by Isomap. The goal of learnt restricted tests is to approach restricted tests asymptotically. We consider \( d_{\mathcal{M}}(\hat{\theta}^*, \hat{\theta}) \), which can be approximately represented by \( \| \hat{Z}^* - \hat{Z}_0 \| \).

Figure 2.2: From a 3D visualization of the Hardy-Weinberg example, we see the \( \theta_i \) (red points) on the HW curve \( \mathcal{H} \) and the \( \hat{\theta}_i \) (green points) in the simplex \( \Delta^2 \) (blue surface) but off the curve.

In learnt restricted tests, asymptotic inference arises in several places, namely in parameters \( s, m, \) and \( nrand \). Large values of \( s \) and \( nrand \) mean tighter bounds on the unrestricted estimators \( \hat{\theta}_i, \hat{\theta}^* \) in \( \Theta \), and the large value of \( m \) provides more sample points for better manifold learning. When \( m \) is small, even if we have \( s, nrand \to \infty \), it is hardly true that \( \| \hat{Z}^* - \hat{Z}_0 \| \to d_{\mathcal{M}}(\hat{\theta}^*, \hat{\theta}) \to d_{\mathcal{M}}(\hat{\theta}^*_M, \hat{\theta}) \). This is because the recovery error of the geodesic distance \( d_{\mathcal{M}} \) via Isomap depends on \( m \) (see Theorem \[11\]). On the other hand, when \( m \to \infty \), then the recovery error of the geodesic
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

distance comes from the off-manifold noise; i.e., the unrestricted estimation error 
\[ \| \hat{\theta}_i - \theta_i \| \text{ and } \| \hat{\theta}^* - \theta^* \|. \] Thus, the error between \( \| \hat{Z}^* - \hat{Z}_0 \| \) and \( d_M(\hat{\theta}^*_M, \hat{\theta}) \) relies on \( m, nr \text{rand}, \) and \( s \). As there is a single \( \theta^* \) but many more \( \theta_i \), the impact of \( \hat{\theta}^* \) on manifold learning is negligible when \( m \) is sufficiently large. A natural conjecture that arises, therefore, is as \( m, nr \text{rand} \to \infty \), we have \( \| \hat{Z}^* - \hat{Z}_0 \| \xrightarrow{d} d_M(\hat{\theta}^*_M, \hat{\theta}) \) and thus \( \pi_{M^*} \to \pi_M \).

We construct an example containing a finite sample of a trinomial model. Let \( \Theta = \Delta^2 \) and \( M = \mathcal{H} \), the Hardy-Weinberg submanifold defined in Section 2.3.1. Figure 2.3a shows the powers of \( T_H \) and \( T_H^* \) defined in Equations (2.13) and (2.14), respectively. Figure 2.3b shows the two power differences, \( \pi_{\Delta^2} - \pi_H \) and \( \pi_{\Delta^2} - \pi_{\mathcal{H}} \).
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

(a) Power functions of the restricted test and the learnt restricted test.

(b) Power function difference between the restricted test/the learnt restricted test and the unrestricted test.

Figure 2.3: The parameters are $\tau_0 = 0.3$, $\alpha = 0.05$, $m = 25$, $nrand = 20$, $s = 10$, $k = 10$ (nearest neighbor size in Isomap). By visual inspection, it is clear that under this setting $T_H \approx T_{\Delta^2}$, there is no power dominance between $T_H$ and $T_{\Delta^2}$. The unrestricted and restricted test powers are computed exactly in R. The learnt restricted power curve is estimated via a Monte Carlo simulation with 100 iterations. In each iteration, we first sample $\tau_1, \ldots, \tau_m \sim U(0, 1)$; then we sample $\{x_{ij}\}_{j=1}^{nrand}$ and $\{x^*_i\}_{i=1}^m$ and $\{x^*_{ij}\}_{j=1}^{nrand}$; finally, we apply Isomap to $\{\theta_i\}_{i=1}^m \cup \hat{\theta}^* \cup \psi(\tau_0)$ and we obtain $Z^*, Z_0$.

Since the $m$ chosen in the above example is relatively small, although the set $\{\theta_i\}_{i=1}^m$ is randomly sampled, it can be rather unbalanced on $\mathcal{H}$ due to the inadequacy of $m$. Thus, we consider the power function of the learnt restricted test $\pi_{\mathcal{M}}(\tau_A; \tau_0, \alpha, m, nrand, s, k)$ being the expectation $E_{\{\theta_i\}_{i=1}^m}[\pi_{\mathcal{M}}(\tau_A; \tau_0, \alpha, m, nrand, s, k|\{\theta_i\}_{i=1}^m)]$ over all possible configurations of $\{\theta_i\}_{i=1}^m \subset (0, 1)$. Similarly, $\pi_{\mathcal{M}}(\tau_A; \tau_0, \alpha, m, nrand, s, k)$ should be an expectation $E_{\{\theta_i\}_{i=1}^m}\{E_{\{x_{ij}\}_{j=1}^{nrand}}[\pi_{\mathcal{M}}(\tau_A; \tau_0, \alpha, m, nrand, s, k|x_{ij}\}_{j=1}^{nrand}|\{\theta_i\}_{i=1}^m]}}$.
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

over all possible configurations of \( \{x_{ij}\}_{j=1}^{nrand} \) given \( \{\theta_i\}_{i=1}^{m} \), where \( \theta_i = \psi(\tau_i) \).

There is a bound on \( ||\hat{\theta}_i - \theta_i|| \) similar to that in Equation (4.9) in Section 4.3.1, where \( \hat{\theta}_i \) is the MLE from samples \( \{x_{ij}\}_{j=1}^{nrand} \) for \( \theta_i \). We view \( \hat{\theta}_i \) as a noisy sample from \( \mathcal{M} \). The bound \( ||\theta_i - \hat{\theta}_i|| \) is derived based on large sample theory.

**Lemma 5.** For a fixed \( \theta_i \), if \( nrand \) is large enough compared to \( D \), then

\[
||\theta_i - \hat{\theta}_i|| < h \frac{\log(nrand)}{nrand} \quad \text{with probability} \quad 1 - O\left( \frac{(\log(nrand))^{D/2-1}}{(nrand)^{D/2}} \right),
\]

for some \( h = \Theta(1) \) and \( h > 0 \).

The assumption that \( nrand \) is large enough relative to \( D \) is rather essential and reasonable because of the curse of dimensionality.

**Proof.** Let \( x_i = (x_{i1}, \ldots, x_{iD}) \) be the outcome of \( nrand \) i.i.d. observations from \( \text{Multi}(\theta_i) \). By the multivariate Central Limit Theorem, for the unrestricted MLE \( \hat{\theta}_i = \left( \frac{x_{i1}}{nrand}, \ldots, \frac{x_{iD}}{nrand} \right) \), we have

\[
\hat{\theta}_i - \theta_i \sim \mathcal{N}(0, \frac{\Sigma_{\theta}}{nrand}),
\]

where \( \Sigma_{\theta} = \text{diag}(\theta_i) - \theta_i \theta_i^T \). Thus, the squared \( L_2 \) norm \( ||\theta_i - \hat{\theta}_i|| \) is a linear combination of chi-square distributions \( ||\theta_i - \hat{\theta}_i||^2 \sim \frac{1}{nrand} \sum_{i=1}^{D-1} \lambda_i \chi_i^2 \) (Lemma [1]), where \( \lambda_i \)'s are the nonzero eigenvalues of \( \Sigma_{\theta} \), and 0 is the smallest eigenvalue of \( \Sigma_{\theta} \).

As \( \sum_{i=1}^{D-1} \lambda_i = 1 - ||\theta_i||^2 \) and \( \lambda_i \geq 0 \), there is a loose bound \( 1 > \lambda_{\text{max}} > \frac{1}{D} \). Denote
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED
SUBMODELS

\[ Z = \| \theta_i - \hat{\theta}_i \|^2, \]

\[ P[Z \geq z] = P\left[ \frac{1}{n_{\text{rand}}} \sum_{i=1}^{D-1} \lambda_i \chi_i^2 \geq z \right] \]
\[ < P\left[ \frac{1}{n_{\text{rand}}} \sum_{i=1}^{D-1} \lambda_{\text{max}} \chi_i^2 \geq z \right] \]
\[ = P[\chi_{D-1}^2 \geq \frac{z}{\lambda_{\text{max}}} * n_{\text{rand}}] \]
\[ < P[\chi_{D-1}^2 \geq z * n_{\text{rand}}]. \]

If \( z = h \log(n_{\text{rand}}) \), with \( h = \Theta(1) \) and \( h > 0 \), then by the upper bound of chi-square

CDF by Inglot [20], we have

\[ P[\chi_{D-1}^2 \geq z * n_{\text{rand}}] = P[\chi_{D-1}^2 \geq h \log(n_{\text{rand}})] \]
\[ \leq \frac{1}{\pi} \frac{h \log(n_{\text{rand}})}{h \log(n_{\text{rand}}) - D + 3} \exp\left\{ -\frac{1}{2} (h \log(n_{\text{rand}}) - (D - 1) - (D - 3) \log(h \log(n_{\text{rand}})/(D - 1)) \right\} \]
\[ + \log(D - 1)) \}
\[ = \frac{1}{\pi} \frac{h \log(n_{\text{rand}})}{h \log(n_{\text{rand}}) - D + 3} (D - 1) e^{(D-1)/2} (\log(h/(D - 1)))^{(D-3)/2} \frac{\log(n_{\text{rand}})^{(D-3)/2}}{(n_{\text{rand}})^{h/2}}, \]

where \( D \geq 3 \), which is true for all multinomial models except binomials.

Isomap, as a manifold learning algorithm, is known to preserve geodesic distance asymptotically. Consider a collection of randomly drawn samples \( \{ \theta_i \} \) from \( \mathcal{M} \); i.e.,
there exists a sequence \( \{ u_i \} \subset U \) (for Hardy-Weinberg, \( U = [0, 1] \)) randomly selected
and \( \theta_i = \psi(u_i) \). The proof of Isomap preserving geodesic distance asymptotically
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED
SUBMODELS

restated in Section 4.6 is completed by proving that the shortest path distance on the
constructed graph $G$ is approximately the geodesic distance on $M$.

Theorem 11 is given based on exact samples $\{\theta_i\}_{i=1}^m$ on $M$. However, we consider
$\{\hat{\theta}_i\}_{i=1}^m$, which are noisy versions of $\{\theta_i\}_{i=1}^m$. To show $\|\hat{Z}^* - \hat{Z}_0\| \to d_M(\theta_M^*, \hat{\theta})$ as
$nrand \to \infty$ and $m \to \infty$, Theorem 14 in Section 4.3.1 provides a similar proof
of the asymptotic equivalence of the restricted test and the learnt restricted test in
networks.

If $\sup_{1 \leq i \leq m} \|\theta_i - \hat{\theta}_i\|$ converges to 0 fast enough as $nrand \to \infty$ and $m$ is sufficiently
large, then $T_{M^*}$ is closely equivalent to $T_M$, and thus the learnt restricted test converges
to the restricted test. Let $n = m + 2$ and $\{\theta_i\}_{i=1}^n = \{\theta_i\}_{i=1}^m \cup \{\theta^*, \hat{\theta}\}$ and $\{\hat{\theta}_i\}_{i=1}^n = \{\hat{\theta}_i\}_{i=1}^m \cup \{\theta^*, \hat{\theta}\}$.

2.5 Two-Sample LRTs in Multinomial Models

2.5.1 Hardy-Weinberg Equilibrium

Consider a trinomial model that is a multinomial model with $D = 3$. We define $\psi : [0, 1] \to \Theta$ by $\psi(\tau) = \{(\tau^2, 2\tau(1-\tau), (1-\tau)^2), \tau \in [0, 1]\} \subset \mathbb{R}^3$. The
Hardy-Weinberg (HW) subfamily of trinomial distributions is parameterized by the
embedded submanifold, $\mathcal{H} = \{\psi(\tau), \tau \in [0, 1]\}$.
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

Suppose that one draws \( n_1 \) i.i.d observations from Trinomial(\( \theta_1 \)) and counts \( x = (x_1, x_2, x_3) \), where \( x_i \) records the number of occurrences of outcome \( i \). One also independently draws \( n_2 \) i.i.d observations from Trinomial(\( \theta_2 \)) and counts \( z = (z_1, z_2, z_3) \), where \( z_i \) records the number of occurrences of outcome \( i \). Let \( \theta_1 = \psi(\tau_1) \) and \( \theta_2 = \psi(\tau_2) \) for some unknown \( \tau_1, \tau_2 \in [0, 1] \). The two-sample restricted LRT can be formulated as

\[
H_0: \theta_1 = \theta_2 \text{ vs. } H_{HA}^{\Delta^2}: \theta_1 \neq \theta_2, \text{ with } \theta_1, \theta_2 \in \mathcal{H},
\]

and that of the unrestricted LRT as

\[
H_0: \theta_1 = \theta_2 \text{ vs. } H_{AA}^{\Delta^2}: \theta_1 \neq \theta_2, \text{ with } \theta_1, \theta_2 \in \Delta^2.
\]

The unrestricted joint likelihood function of two samples in \( \Delta^2 \) is

\[
L(x, z; \theta_1, \theta_2) = \frac{n_1! n_2!}{\prod_{i=1}^{3} x_i! \prod_{j=1}^{3} z_j!} \prod_{i=1}^{3} \theta_{x_i}^{x_i} \prod_{j=1}^{3} \theta_{z_j}^{z_j}.
\]

The joint likelihood function of two samples in HW parametrized by \( \tau_1, \tau_2 \) is

\[
L(x, z; \psi(\tau_1), \psi(\tau_2)) = \frac{n_1! n_2!}{\prod_{i=1}^{3} x_i! \prod_{j=1}^{3} z_j!} 2^{x_2+z_2} \tau_1^{m_1} (1 - \tau_1)^{2n_1-m_1} \tau_2^{m_2} (1 - \tau_2)^{2n_2-m_2}.
\]

Let \( m_1 = 2x_1 + x_2 \) and \( m_2 = 2z_1 + z_2 \). The MLE of \( \tau_1, \tau_2 \) can be written as
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

\( \hat{\tau}_{1,MLE} = \frac{m_1}{2n_1}, \hat{\tau}_{2,MLE} = \frac{m_2}{2n_2} \), respectively, under \( H_0 \) and \( \hat{\tau}_{1,MLE} = \hat{\tau}_{2,MLE} = \frac{m_1+m_2}{2n_1+2n_2} \) under \( H^A_{HW} \). The restricted LRT rejects \( H_0 : \psi(\tau_1) = \psi(\tau_2) \) if

\[
\Lambda_{HW}(x, z) = -2 \log \frac{L(x, z; \tau_1 = \tau_2 = \frac{m_1+m_2}{2n_1+2n_2})}{L(x, z; \tau_1 = \frac{m_1}{2n_1}, \tau_2 = \frac{m_2}{2n_2})}
\]

\[
= -2 \log \frac{(\frac{m_1}{2n_1})^{m_1}(1 - \frac{m_1+m_2}{2n_1+2n_2})^{2n_1-m_1+2n_2-m_2}}{(\frac{m_1}{2n_1})^{m_1}(1 - \frac{m_1+m_2}{2n_1+2n_2})^{2n_1-m_1+2n_2-m_2}}
\]

is sufficiently large. Moreover, the unrestricted LRT rejects \( H_0 : \psi(\tau_1) = \psi(\tau_2) \) if

\[
\Lambda_{\Delta^2}(x, z) = -2 \log \frac{L(x, z; \theta_1, \theta_2)}{L(x, z; \theta_1, \theta_2)}
\]

\[
= -2 \log \frac{2^{x_2+2z}(\frac{m_1}{2n_1})^{m_1+m_2}(1 - \frac{m_1+m_2}{2n_1+2n_2})^{2n_1-m_1+2n_2-m_2}}{(\frac{x_1}{n_1})^{x_1}(\frac{x_2}{n_1})^{x_2}(\frac{x_3}{n_2})^{x_3}(\frac{z_1}{n_1})^{z_1}(\frac{z_2}{n_2})^{z_2}(\frac{z_3}{n_2})^{z_3}}
\]

is sufficiently large.

In what follows, it is equivalent and convenient to consider the inverse likelihood ratios,

\[
T_{HW}(x, z) = \frac{(\frac{m_1}{2n_1})^{m_1}(1 - \frac{m_1}{2n_1})^{2n_1-m_1}(\frac{m_2}{2n_2})^{m_2}(1 - \frac{m_2}{2n_2})^{2n_2-m_2}}{(\frac{m_1}{2n_1})^{m_1+m_2}(1 - \frac{m_1+m_2}{2n_1+2n_2})^{2n_1-m_1+2n_2-m_2}}
\]

\[
T_{\Delta^2}(x, z) = \frac{(\frac{x_1}{n_1})^{x_1}(\frac{x_2}{n_1})^{x_2}(\frac{x_3}{n_2})^{x_3}(\frac{z_1}{n_1})^{z_1}(\frac{z_2}{n_2})^{z_2}(\frac{z_3}{n_2})^{z_3}}{2^{x_2+2z}(\frac{m_1}{2n_1})^{m_1+m_2}(1 - \frac{m_1+m_2}{2n_1+2n_2})^{2n_1-m_1+2n_2-m_2}}
\]

instead of \( \Lambda_{HW}(x, z) \) and \( \Lambda_{\Delta^2}(x, z) \).

We consider the powers \( \pi_{\Delta^2}(\psi(\tau_1), \psi(\tau_2); \alpha, n_1, n_2) \) and \( \pi_{HW}\pi_{\Delta^2}(\psi(\tau_1), \psi(\tau_2); \alpha, n_1, n_2) \) when \( n < \infty \), corresponding to the unrestricted LRT and the restricted LRT respec-
We expect the power superiority loss of dimension-restricted LRTs over unrestricted LRTs similar to that in one-sample tests, as shown in Theorem 3, to exist in two-sample tests as well. The setting of the two-sample is more complex since we need to consider sample sizes and alternatives in pairs. Because the two-sample testing problem that we introduce here has a composite null hypothesis, the type I error calculation becomes more complex and makes it hard to track the critical region.

The following theorem establishes that the phenomenon in Theorem 3 holds for two-sample LRTs as well, and thus completes the theory of LRTs; this result is the first one with respect to two-sample LRTs of which we are aware:

**Theorem 5.** Let $\psi$ parameterize the Hardy-Weinberg submodel of the trinomial experiment and consider the null hypothesis $H_0 : \psi(\tau_1) = \psi(\tau_2)$, and the experiment is of size $n_1 = n_2 = n$ for both populations. For any $n > 3$, there exists $\alpha \in (0, 1)$ and a pair of alternative $(\tau_1, \tau_2) \in (0, 1) \times (0, 1)$, such that the restricted LRT of size $\alpha$ is less powerful at $(\tau_1, \tau_2)$ than the unrestricted LRT of size $\alpha$.

The proof of Theorem 5 is not identical to that of Theorem 3. As both theorems consider LRTs in the Hardy-Weinberg submodel, both proofs are similar. However, the critical difference in defining $\alpha$ between one-sample tests and two-sample tests makes Theorem 5 trickier and less generalizable than Theorem 3.

We assume $n_1 = n_2 = n$ hence forth unless specified otherwise. Then $T_{HW}(x, z)$
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

and $T_\Delta^2(x, z)$ become

$$T_{HW}(x, z) = \frac{\left(\frac{m_1}{2n}\right)^{m_1} \left(1 - \frac{m_1}{2n}\right)^{2n_1 - m_1} \left(\frac{2n_2}{m_2}\right)^{m_2} \left(1 - \frac{m_2}{2n}\right)^{2n_2 - m_2}}{\left(\frac{m_1 + m_2}{4n}\right)^{m_1 + m_2} \left(1 - \frac{m_1 + m_2}{4n}\right)^{2n_1 + m_1 + 2n_2 - m_2}}.$$  

$$T_\Delta^2(x, z) = \frac{2^{2m_2 + n_2} \left(\frac{m_1 + m_2}{4n}\right)^{m_1 + m_2} \left(1 - \frac{m_1 + m_2}{4n}\right)^{2n_1 + m_1 + 2n_2 - m_2}}{\left(\frac{m_1}{n}\right)^{x_1} \left(\frac{m_2}{n}\right)^{x_2} \left(\frac{2n}{n}\right)^{z_1} \left(\frac{2n}{n}\right)^{z_2} \left(\frac{2n}{n}\right)^{z_3}}.$$  

We consider ordering all unique $T_\Delta^2(x, z)$ values from the largest to the smallest by varying the possible outcome pairs $[x, z]$, with $T_\Delta^{(i)}$ denoting the $i^{th}$ element in the ordered $T_\Delta^2(x, z)$ sequence. Similar to the ordering definition of $T_\Delta^2$ for one-sample LRTs in Section 2.3.1, we say $[x_1, z_1] \succ [x_2, z_2]$ under the unrestricted LRT if $T_\Delta^2(x_1, z_1) \geq T_\Delta^2(x_2, z_2)$. Let $O_\Delta^i$ denote the set of all outcome pairs $[x, z]$ that yield the same $T_\Delta^{(i)}$.

Analogously, the ordered sequence of all unique $T_{HW}(x, z)$ values from the largest to the smallest is defined, with $T_{HW}^{(i)}$ denoting the $i^{th}$ element and $[x_1, z_1] \succ [x_2, z_2]$ under the restricted LRT if $T_{HW}(x_1, z_1) \geq T_{HW}(x_2, z_2)$. Let $O_{HW}^i$ denote the set of all outcome pairs $[x, z]$ that yield the same $T_{HW}^{(i)}$.

As both groups have the same experiment size and our interest lies in whether their underlying distributions are the same under $H_0$, there is a perfect symmetry with respect to $[x, z]$ when considering $T_{HW}$ or $T_\Delta^2$. Under $H_0$, it is sufficient to consider any pair $[x, z]$ without specifying group labels; i.e., $T_{HW}(x, z) = T_{HW}(z, x)$ and $T_\Delta^2(x, z) = T_\Delta^2(z, x)$.

**Lemma 6.** When $n > 3$, the unrestricted LRT and the restricted LRT order all
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

possible outcomes differently.

Proof. When \( n > 3 \), the order of \( T_{HW}(x, z) \) is:

1. \( T_{HW}^{(1)} = T_{HW}((n, 0, 0), (0, 0, n)) \) and \( O_{HW}^{1} = \{(n, 0, 0), (0, 0, n)\} \),

2. \( T_{HW}^{(2)} = T_{HW}((n, 0, 0), (0, 1, n - 1)) = T_{HW}((0, 0, n), (n - 1, 1, 0)) \) and \( O_{HW}^{2} = \{(n, 0, 0), (0, 1, n - 1), [(0, 0, n), (n - 1, 1, 0)]\} \);

and the order of \( T_{\Delta^2}(x, z) \) is:

1. \( T_{\Delta^2}^{(1)} = T_{\Delta^2}((n, 0, 0), (0, 0, n)) \) and \( O_{\Delta}^{1} = \{(n, 0, 0), (0, 0, n)\} \),

2. \( T_{\Delta^2}^{(2)} = T_{\Delta^2}((n, 0, 0), (1, 0, n - 1)) = T_{\Delta^2}((0, 0, n), (n - 1, 0, 1)) \) and \( O_{\Delta^2}^{2} = \{(n, 0, 0), (1, 0, n - 1), [(0, 0, n), (n - 1, 0, 1)]\} \);

thus, the smallest \( I \) such that \( O_{\Delta}^{I} \neq O_{HW}^{I} \) is \( I = 2 \).

We borrow the definitions of \( \alpha_{HW}^{I}, \alpha_{\Delta^2}^{I} \) and \( C_{HW}(\alpha), C_{\Delta^2}(\alpha) \) from Section 2.3.1. When assigning \( \alpha = \alpha_{HW}^{1} = \alpha_{\Delta}^{1} \), as \( C_{HW}(\alpha) = C_{\Delta^2}(\alpha) \), then \( \pi_{HW} = \pi_{\Delta^2} \) under any alternative.

Proof of Theorem 5:

Proof. For simplicity, let \( L_{\tau_1, \tau_2}[x, z] \) denote the likelihood function of \( L[x, z; \tau_1, \tau_2] = L[x, z; \psi(\tau_1), \psi(\tau_2)] \).

Picking \( \alpha \in (\alpha_{HW}^{1}, \min\{\alpha_{HW}^{2}, \alpha_{\Delta}^{2}\}) \) can possibly lead to \( \pi_{HW}(\psi(\tau_1), \psi(\tau_2); \alpha, n, n) \neq \pi_{\Delta^2}(\psi(\tau_1), \psi(\tau_2); \alpha, n, n) \) as \( C_{HW}(\alpha) \neq C_{\Delta^2}(\alpha) \) here, containing the existence of ordering differences between the restricted LRT and the unrestricted LRT. We pick \( \alpha_0 \)
as

\[\alpha_0 = P_{\tau_1 = \tau_2}[O_{\Delta^2}^1] + r_{\Delta^2}P_{\tau_1 = \tau_2}[O_{\Delta^2}^2] \]
\[= 2 \sup_{\tau \in (0,1)} \{ L_{\tau,\tau}[(n,0,0),(0,0,n)] + r_{\Delta^2}(L_{\tau,\tau}[(n,0,0),(1,0,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,0,1)])\} \]
\[= P_{\tau_1 = \tau_2}[O_{HW}^1] + r_{HW}P_{\tau_1 = \tau_2}[O_{HW}^2] \]
\[= 2 \sup_{\tau \in (0,1)} \{ L_{\tau,\tau}[(n,0,0),(0,0,n)] + r_{HW}(L_{\tau,\tau}[(n,0,0),(0,1,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,1,0)])\}, \]

for some \(r_{\Delta^2}, r_{HW} \in [0,1]\).

If we analyze each part from above separately, then we observe

\[\arg \max_\tau L_{\tau,\tau}[(n,0,0),(0,0,n)] = \frac{1}{2} \]
\[\arg \max_\tau \{ L_{\tau,\tau}[(n,0,0),(1,0,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,0,1)]\} = \frac{1}{2} \]
\[\arg \max_\tau \{ L_{\tau,\tau}[(n,0,0),(0,1,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,1,0)]\} = \frac{1}{2}. \]

When \(\frac{1}{3} < \tau < \frac{2}{3}\), we have \(L_{\tau,\tau}[(n,0,0),(1,0,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,0,1)] < L_{\tau,\tau}[(n,0,0),(0,1,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,1,0)]\), i.e. \(L_{\tau,\tau}[O_{HW}^2] > L_{\tau,\tau}[O_{\Delta^2}^2]\). As a result, \(r_{\Delta^2} < r_{HW}\) should hold. Without loss of generality, we set \(r_{\Delta^2} = 1\), then \(\alpha_0 = 2\{(L_{\tau,\tau}[(n,0,0),(0,0,n)] + (L_{\tau,\tau}[(n,0,0),(1,0,n-1)] + L_{\tau,\tau}[(0,0,n),(n-1,0,1)])\});

therefore, the value of \(r_{HW}\) can be computed, i.e., \(r_{HW} = 1/2\).
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

We rewrite the power difference between the unrestricted LRT and the restricted LRT under alternative \((\tau_1, \tau_2)\) as a function \(f\) of \((\tau_1, \tau_2)\),

\[
f(\tau_1, \tau_2) = \pi_{\Delta^2} - \pi_{HW} = r_{\Delta^2}L_{\tau_1, \tau_2}[O_{\Delta^2}^2] - r_{HW}L_{\tau_1, \tau_2}[O_{HW}^2]
\]

\[
= \{L_{\tau_1, \tau_2}[(n, 0, 0), (1, 0, n - 1)] + L_{\tau_1, \tau_2}[(1, 0, n - 1), (n, 0, 0)]
\]
\[
+ L_{\tau_1, \tau_2}[(0, 0, n), (n - 1, 0, 1)] + L_{\tau_1, \tau_2}[(n - 1, 0, 1), (0, 0, n)]\}
\]
\[
- \frac{1}{2} \{L_{\tau_1, \tau_2}[(n, 0, 0), (0, 1, n - 1)] + L_{\tau_1, \tau_2}[(0, 1, n - 1), (n, 0, 0)]
\]
\[
+ L_{\tau_1, \tau_2}[(0, 0, n), (n - 1, 1, 0)] + L_{\tau_1, \tau_2}[(n - 1, 1, 0), (0, 0, n)]\}.
\]

(2.15)

Some basic characteristics of the function \(f(a, b)\) are

I. \(f(a, b) = f(b, a)\),

II. \(f(a, b) = f(1 - a, 1 - b)\),

III. \(\frac{f(a, b)}{n} = [a^{2n}(1 - b)^{2n-2}b - (1 - a)^{2n}b^{2n-2}(1 - b)][2b - 1]
\]
\[
+ [b^{2n}(1 - a)^{2n-2}a - (1 - b)^{2n}a^{2n-2}(1 - a)][2a - 1].
\]

I and II are straightforward due to the symmetry of two-sample LRTs and the Hardy-Weinberg model. III is derived after simplifying Equation (2.15).

Based on II, it is sufficient to consider \(a = \tau_1 < 1/2, b = \tau_2 < 1/2\) in what follows.

Based on III, if \(a^{2n}(1 - b)^{2n-2}b - (1 - a)^{2n}b^{2n-2}(1 - b) < 0\) and \(b^{2n}(1 - a)^{2n-2}a -
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

$(1 - b)^{2n}a^{2n-2}(1 - a) < 0$ are both true for some pair of $[a, b]$, then $f(a, b) > 0$ holds for that $[a, b]$ pair.

We define $p_1 = \frac{\tau_1}{1 - \tau_1}$ and $p_2 = \frac{\tau_2}{1 - \tau_2}$. Then Equation (2.15) can be rewritten as

$$\frac{f(\tau_1, \tau_2)}{n} = (1 - \tau_1)^{2n}(1 - \tau_2)^{2n-2}\tau_2[p_1^{2n} - p_2^{2n-3}](2\tau_2 - 1)$$

$$+ (1 - \tau_2)^{2n}(1 - \tau_1)^{2n-2}\tau_1[p_2^{2n} - p_1^{2n-3}](2\tau_1 - 1).$$

$p_1^{2n} < p_2^{2n-3}$ and $p_2^{2n} < p_1^{2n-3}$ both being true is a sufficient condition to guarantee $\frac{f(\tau_1, \tau_2)}{n} > 0$ and the positivity of Equation (2.15).

For any $n < \infty$, when choosing $\tau_1, \tau_2$ such that $p_1^{2n} = p_2^{2n-4}$, we have $p_1^{2n} < p_2^{2n-3}$ and $p_2^{2n} < p_1^{2n-3}$ both hold. Thus, there exists at least one pair of alternative $(\tau_1, \tau_2)$ under which $\pi_\Delta^2(\psi(\tau_1), \psi(\tau_2); \alpha_0, n, n) > \pi_{HW}(\psi(\tau_1), \psi(\tau_2); \alpha_0, n, n)$ holds. It is obvious that the set of such alternatives $(\tau_1, \tau_2)$ has a positive measure.

Theorem 5 considers two-sample LRTs when $n_1 = n_2$. From numerical calculations via R, we conjecture that the phenomenon described in Theorem 5 exists with unequal experiment sizes as well.

With both Theorem 3 and Theorem 5 we have established that the dimensionality conjecture of LRTs, which states that “the dimension-restricted LRT is more powerful than the unrestricted LRT,” does not hold for one-sample or two-sample tests. This phenomenon is hardly unique for the Hardy-Weinberg submodel or linear submodels, for we have constructed other submodels of multinomial models, including submodels...
of dimension and/or co-dimension greater than one, that also falsify the conjecture. Our current efforts are focused on identifying conditions under which the dimension-restricted LRT conjecture does, or does not, hold.

In the next chapter we will continue the discussion under another inference problem: classification.

2.6 Appendix: Proofs

Observation 1. The sum of (in)dependent squared univariate Gaussians is a linear combination of independent chi-square variables, each with one degree of freedom.

Proof. Let $Y = (Y_1, \ldots, Y_k) \sim \mathcal{N}(\mu, \Sigma)$. The quadratic form of $Y$ can be rewritten as

$$Q(Y) = Y^TLY$$

$$= \sum_{i=1}^{k} \lambda_i (U_i + b_i)^2,$$

where $\Sigma^{\frac{1}{2}}L\Sigma^{\frac{1}{2}} = V^T\Lambda V$, for some orthogonal matrix $V$ and $\Lambda$ is a diagonal matrix with $\lambda_1, \ldots, \lambda_k$ as diagonal entries. $U_i$ are columns of the matrix $U = VL$ and $b = V\Sigma^{-\frac{1}{2}}\mu$. Each $U_i$ represents a standard normal variable. Columns of $V^T$ are the eigenvectors of $\Sigma^{\frac{1}{2}}L\Sigma^{\frac{1}{2}}$. $\square$

Proof of Theorem 3

Proof. Without loss of generality, we assume $\tau_0 < 0.5$ and $\tau_0 \in E$ defined in Lemma
We consider “the order of all outcomes in the restricted LRT” and “the order of all outcomes in the unrestricted LRT.” There are possible ties in either ordering; for example, if \( x \neq y \) but \( L^T x = L^T y \), then \( T_{HW}(x) = T_{HW}(y) \). Then \( x \succ y \) and \( y \succ x \) are both valid orderings. Thus, there are multiple valid ways to order all outcomes in the restricted LRT as well in the unrestricted LRT. Of the valid options presented, we pick the one in which the two orderings agree the most.

We define \( I_1 = \min \{ x_{HW}^{(i)} \neq x_{\Delta_2}^{(i)} \} \), \( J_1 = \{ j : x_{HW}^{(j')} \in O_{HW}^j \} \) and \( J_2 = \{ j : x_{\Delta_2}^{(j')} \in O_{\Delta}^j \} \). The definitions capture the first pair of outcomes that differ in the two orderings, implying that \( \bigcup_{i=1}^{J_1-1} O_{HW}^i = \bigcup_{i=1}^{J_2-1} O_{\Delta}^i \). Lemma 2 suggests that there are three cases describing \( O_{HW}^{I_1} \) and \( O_{\Delta_2}^{I_2} \):

I. \( O_{HW}^{I_1} \) and \( O_{\Delta_2}^{I_2} \) are associated with different values of \( 2X_1 + X_2 \), and \( O_{\Delta_2}^{I_2} \) is associated with one value of \( 2X_1 + X_2 \);

II. \( O_{HW}^{I_1} \) and \( O_{\Delta_2}^{I_2} \) are associated with different values of \( 2X_1 + X_2 \), and \( O_{\Delta_2}^{I_2} \) is associated with more than one value of \( 2X_1 + X_2 \);

III. \( O_{HW}^{I_1} \) and \( O_{\Delta_2}^{I_2} \) have one common value of \( 2X_1 + X_2 \), and \( O_{\Delta_2}^{I_2} \) is associated with a second value of \( 2X_1 + X_2 \).

For Case I, let \( m_1, m_2 \) denote the values of \( 2X_1 + X_2 \) associated with \( O_{HW}^{I_1} \) and \( O_{\Delta_2}^{I_2} \), respectively. Let \( \alpha = \min(\alpha_{HW}^{I_1}, \alpha_{\Delta_2}^{I_2}) \), so that the restricted and the unrestricted
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

LRTs of size $\alpha$ have critical regions that are identical except for outcomes in $O_{HW}^{J_1}$ and $O_{\Delta_2}^{J_2}$. Any power difference between the two LRTs will accrue from these outcomes.

Then with respect to the restricted LRT, $\alpha$ can be written as

$$\alpha = P_{\tau_0} \left[ \bigcup_{i=1}^{J_1-1} O_{HW}^i \right] + r_1 P_{\tau_0}[O_{HW}^{J_1}];$$

and with respect to the unrestricted LRT, $\alpha$ can be written as

$$\alpha = P_{\tau_0} \left[ \bigcup_{i=1}^{J_2-1} O_{\Delta_2}^i \right] + r_2 P_{\tau_0}[O_{\Delta_2}^{J_2}];$$

where $r_1, r_2$ represent the randomization probability defined in Equation (2.8) with respect to $\alpha_{HW}^{J_1}$ and $\alpha_{\Delta_2}^{J_2}$, respectively. As a result, we have the relationship between $r_1$ and $r_2$ as

$$r_1 P_{\tau_0}[O_{HW}^{J_1}] = r_2 P_{\tau_0}[O_{\Delta_2}^{J_2}]$$

$$\iff r_1 \tau_0^{m_1} (1 - \tau_0)^{2n-m_1} \sum_{x \in O_{HW}^{J_1}} \frac{2^{x_2} n!}{x_1! x_2! x_3!} = r_2 \tau_0^{m_2} (1 - \tau_0)^{2n-m_2} \sum_{x \in O_{\Delta_2}^{J_2}} \frac{2^{x_2} n!}{x_1! x_2! x_3!}.$$

(2.16)

Thus, the difference between the power of the unrestricted LRT and that of the
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

restricted LRT under alternative $\psi(\tau_A)$ is

$$\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) - \pi_{\text{HW}}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) = r_2\mathbb{P}_{\tau_A}[O_{\Delta^2}^2] - r_1\mathbb{P}_{\tau_A}[O_{\text{HW}}^2].$$

(2.17)

Then from the relationship between $r_1$ and $r_2$ illustrated in Equation (2.16), the power difference (2.17) is positive if

$$\frac{\mathbb{P}_{\tau_A}[O_{\Delta^2}^2]}{\mathbb{P}_{\tau_A}[O_{\text{HW}}^2]} > \frac{r_1}{r_2} = \frac{\mathbb{P}_{\tau_0}[O_{\Delta^2}^2]}{\mathbb{P}_{\tau_0}[O_{\text{HW}}^2]},$$

$$\iff \left(\frac{\tau_A}{1 - \tau_A}\right)^{m_2 - m_1} > \left(\frac{\tau_0}{1 - \tau_0}\right)^{m_1 - m_2},$$

which is possible for some choices of $\tau_A$ since $m_1 \neq m_2$.

For Case II, let $m_1$ denote the values of $2X_1 + X_2$ associated with $O_{\text{HW}}^{J_2}$ and let $m_2, m'_2$ denote those of $O_{\Delta^2}^{J_2}$. Let $\alpha = \min(\alpha_{\text{HW}}^{J_1}, \alpha_{\Delta^2}^{J_2})$. Then we have the same argument as in Case I:

$$r_1\mathbb{P}_{\tau_0}[O_{\text{HW}}^{J_2}] = r_2\mathbb{P}_{\tau_0}[O_{\Delta^2}^{J_2}],$$

$$\pi_{\Delta^2}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) - \pi_{\text{HW}}(\psi(\tau_A); \alpha, n, \psi(\tau_0)) = r_2\mathbb{P}_{\tau_A}[O_{\Delta^2}^2] - r_1\mathbb{P}_{\tau_A}[O_{\text{HW}}^2].$$

Thus, the power difference is positive if and only if

$$\frac{\mathbb{P}_{\tau_A}[O_{\Delta^2}^2]}{\mathbb{P}_{\tau_A}[O_{\text{HW}}^2]} > \frac{r_1}{r_2} = \frac{\mathbb{P}_{\tau_0}[O_{\Delta^2}^2]}{\mathbb{P}_{\tau_0}[O_{\text{HW}}^2]},$$
CHAPTER 2. ON THE POWER OF LRTS IN DIMENSION-RESTRICTED SUBMODELS

which can be rewritten as

\[
\begin{align*}
\frac{\tau_A^{m_2} (1 - \tau_A)^{2n-m_2} \sum_{x \in O^I_{m_2}} \frac{2^n 2n!}{x_1!x_2!x_3!} + \tau_A^{m'_2} (1 - \tau_A)^{2n-m'_2} \sum_{y \in O^I_{m'_2}} \frac{2^n 2n!}{y_1!y_2!y_3!}}{\tau_A^{m_1} (1 - \tau_A)^{2n-m_1} \sum_{x \in O^I_{m_1}} \frac{2^n 2n!}{x_1!x_2!x_3!} + \tau_A^{m'_1} (1 - \tau_A)^{2n-m'_1} \sum_{y \in O^I_{m'_1}} \frac{2^n 2n!}{y_1!y_2!y_3!}} > 1,
\end{align*}
\]

(2.18)

In what follows, we will argue that \((m_2 - m_1)(m'_2 - m_1) > 0\), based on which we can conclude that there exist some choices of \(\tau_A\) such that Equation (2.18) holds.

Suppose \(m_2 < m_1\) and \(m_1 > 2n \tau_0\), then one can conclude \(m'_2 < m_1\). If we otherwise assume \(m'_2 > m_1\) is true, we have \(T_{HW}(m'_2) > T_{HW}(m_1)\) based on two facts: \(T_{HW}(m)\) as a function of \(m\) increases on \([2n \tau_0, 2n]\) and \(m_1 = L^T \bar{x}_{HW}\), where \(I^*\) is the smallest index such that “the order for all outcomes in the restricted LRT” and “the order for all outcomes in the unrestricted LRT” do not agree. In other words, if \(m'_2 > m_1\), then the order for the restricted LRT and that for the unrestricted LRT would respectively be

\[
T_{HW}(m'_2) > T_{HW}(m_1) > T_{HW}(m_2)
\]

\[
T_{\Delta^2}(m'_2) = T_{\Delta^2}(m_2) > T_{\Delta^2}(m_1),
\]

54
which contradicts the definition of $I^*$. Thus, if $m_2 < m_1$ and $m_1 > 2n\tau_0$, then $m'_2 < m_1$. By symmetry, because $T_{HW}(m)$ as a function of $m$ decreases on $[0,2n\tau_0]$, if $m_2 > m_1$ and $m_1 < 2n\tau_0$, then $m'_2 > m_1$.

Similarly, suppose $m_2 > m_1$ and $m_1 > 2n\tau_0$. If $m'_2 < m_1$, then the order for the restricted LRT and that for the unrestricted would respectively be

$$T_{HW}(m_2) > T_{HW}(m_1) > T_{HW}(m'_2) \text{ OR } T_{HW}(m_2) > T_{HW}(m'_2) > T_{HW}(m_1);$$

$$T_{\Delta^2}(m'_2) = T_{\Delta^2}(m_2) > T_{\Delta^2}(m_1),$$

with either case contradicting the definition of $I^*$. Thus, if $m_2 > m_1$ and $m_1 > 2n\tau_0$, then $m'_2 > m_1$. By symmetry, if $m_2 < m_1$ and $m_1 < 2n\tau_0$, then $m'_2 < m_1$.

Thus, in Equation (2.18), $(m_2 - m_1)(m'_2 - m_1) > 0$ holds and there exist some choices of $\tau_A$ such that Equation (2.18) holds.

Case III is similar to Case II. \[\square\]
Chapter 3

On the Expected Error of Classification in Dimension-Restricted Submodels

The last chapter discusses the power comparison between unrestricted and dimension-restricted models for LRTs. In this chapter, we will shift our focus to classification problems to discuss the expected error comparison between dimension-restricted and unrestricted models. We start by recalling some basics of statistical classification problems.

Consider \((X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} f_\theta(x, y)\), where each \(X_i \in \mathbb{R}^D\) is an observed feature vector and \(Y_i \in \{0, 1\}\) is the corresponding observed class label. There is another sequence of observed feature vectors \(X_1^*, \ldots, X_N^*\) all from an unobserved class
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

$Y^* \in \{0, 1\}$.

Let $f_0(x) = f_{\theta_0}(x) = f_{\theta}(x|Y = 0)$ and $f_1(x) = f_{\theta_1}(x) = f_{\theta}(x|Y = 1)$ denote the class-conditional probability densities. Let priori class probabilities be $p_0 = \mathbb{P}[Y = 0] > 0$ and $p_1 = \mathbb{P}[Y = 1] = 1 - p_0 > 0$, so that we have $f_{\theta}(x) = p_0 f_0(x) + p_1 f_1(x)$.

In pertinent literature, there are two major settings for classification problems, one with query sample size $N = 1$ (Setting I) and one with $N = \infty$ (Setting II).

We first recall some basics of classification under Setting I. A function $g : \mathbb{R}^D \to \{0, 1\}$ defines a classifier. The Bayes decision function,

$$
g^*(x) = \begin{cases} 
1 & \text{if } \mathbb{P}[Y = 1|X = x] > 1/2 \\
0 & \text{otherwise},
\end{cases}
$$

minimizes the error probability $\mathbb{P}[g(X) \neq Y]$. Here, $g^*$ as the optimal decision among all possible decision rules is called the Bayes decision. The Bayes error is defined as $L^* = \mathbb{P}[g^*(X) \neq Y]$.

Under Setting II, it is equivalent to consider $X_1^* \overset{iid}{\sim} f_{\theta^*}$ with $\theta^* \in \{\theta_0, \theta_1\}$ and upon testing the hypothesis,

$$
H_0 : \theta^* = \theta_0, \\
H_1 : \theta^* = \theta_1.
$$
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

For any test of $H_0$ vs. $H_1$, we define the expected probability of error as

$$P_e(N) = \alpha(N)p_0 + \beta(N)p_1,$$

where $\alpha(N)$ and $\beta(N)$ are Type I and Type II errors of the test, respectively. The minimum attainable probability of error can be related to the Chernoff information asymptotically [21],

$$\lim_{N \to \infty} \frac{1}{N} \log P_e(N) = - \sup_{\alpha \in [0,1]} D_{\alpha}(f_1||f_0),$$

where $D_{\alpha}(f_1||f_0) = -\log \int f_0^{\alpha}(z)f_1^{1-\alpha}(z)dz$ defines the unnormalized $\alpha$-divergence between $f_0$ and $f_1$ of fractional order $\alpha$. As a generalization of Kullback-Leibler divergence (where $\alpha \to 1$), the unnormalized $\alpha$-divergence measures the similarity relating to the overlapping of $f_0$ and $f_1$. The value $\sup_{\alpha \in [0,1]} D_{\alpha}(f_1||f_0)$ is also called the Chernoff information.

Suppose $\theta_0, \theta_1 \in \mathcal{M} \subset \Theta \subset \mathbb{R}^D$, where $\mathcal{M}$ is a smooth, compact, and low-dimensional submanifold of intrinsic dimension $d$ embedded in the manifold $\Theta$ of intrinsic dimension $D$. When $\theta_0$ and $\theta_1$ are unknown, instead of using the Bayes decision function, a natural substitution is the Bayes plug-in decision function. We define unrestricted classifiers to be Bayes plug-in decision functions with the unrestricted estimation of $\theta_0, \theta_1$ on $\mathcal{A}$ and restricted classifiers to be Bayes plug-in decision functions with the restricted estimation of $\theta_0, \theta_1$ on $\mathcal{M}$. Section 3.1 discusses the ex-
expected classification error comparison between unrestricted classifiers and restricted classifiers in Gaussian models under Setting II. Section 3.2 reveals that the expected classification error of the restricted classifier is not uniformly smaller than that of the unrestricted classifier in multinomial models under Setting I.

3.1 Classification in Continuous Models

(Gaussians)

Let \( f_k(x) \) be in the Gaussian distribution family. Let \( f_k(x) = \mathcal{N}(x; \mu_k, \Sigma) \) for \( k = 0, 1 \), with \( \Sigma \) known and \( \mu_k \in \mathcal{P} \subset \mathbb{R}^D \), where \( \mathcal{P} \) is a hyperplane of dimension \( d \) that goes through the origin, lying in \( D \)-dimensional Euclidean space where \( D > d \). There exists a matrix \( R \in \mathbb{R}^{D \times d} \), such that \( \mu_k = Rw_k \) for some \( w_k \in \mathbb{R}^d \). There is a sequence \( X_1^*, \ldots, X_N^* \overset{iid}{\sim} \mathcal{N}(x; \mu^*, \Sigma) \), and the goal is to decide \( k^* \in \{0, 1\} \) such that \( \mu_k^* = \mu^* \).

If \( \mu_k \) are known, then the classification problem is equivalent to the Neyman-Pearson LRTs with simple null and simple alternative hypotheses: \( H_0 : \mu^* = \mu_0 \) vs. \( H_1 : \mu^* = \mu_1 \). Thus, knowing the statistical submanifold \( \mathcal{P} \) has no impact on the inference.

It is more interesting to ponder when \( \mu_k \) are unknown and the classification task will depend on the estimation of \( \mu_k \), i.e., the MLE of \( \mu_k \). The procedure of classifying the sequence \( X_1^*, \ldots, X_N^* \) begins with us obtaining unrestricted estimators \( \hat{\mu}_{k, \Theta} \) on the
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

statistical manifold $\Theta$ and restricted estimators $\hat{\mu}_{k,M}$ on the submanifold $\mathcal{M}$. Then

the Linear Discriminant Analysis (LDA) transforms $X_i^*$ linearly by $w = \Sigma^{-1}(\hat{\mu}_0. - \hat{\mu}_1.)$ and $c = \frac{1}{2}(\hat{\mu}_{1,0}^T \Sigma^{-1} \hat{\mu}_1. - \hat{\mu}_{0,0}^T \Sigma^{-1} \hat{\mu}_0.)$. The classification rule becomes $I_{w^TX_i^* > c}$ where

$w = \Sigma^{-1}(\hat{\mu}_{0,\Theta} - \hat{\mu}_{1,\Theta})$, $c = \frac{1}{2}(\hat{\mu}_{1,\Theta}^T \Sigma^{-1} \hat{\mu}_{1,\Theta} - \hat{\mu}_{0,\Theta}^T \Sigma^{-1} \hat{\mu}_{0,\Theta})$ for unrestricted classifiers and

$w = \Sigma^{-1}(\hat{\mu}_{0,M} - \hat{\mu}_{1,M})$, $c = \frac{1}{2}(\hat{\mu}_{1,M}^T \Sigma^{-1} \hat{\mu}_{1,M} - \hat{\mu}_{0,M}^T \Sigma^{-1} \hat{\mu}_{0,M})$ for restricted classifiers.

We emphasize that in this section $\mathcal{M} = \mathcal{P}$.

The unrestricted MLE of $\mu_k$ in $\Theta$ is given by $\hat{\mu}_k^D = \bar{X}_k = \frac{\sum_{j=1}^n x_j I_{Y_j=k}}{\sum_{j=1}^n I_{Y_j=k}}$ and the restricted MLE of $\mu_k$ in $\mathcal{P}$ is given by $\hat{\mu}_k^d = (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} \bar{X}_k$ for $k = 0, 1$. As

the classification task under Setting II is equivalent to testing the hypothesis,

$$H_0 : \mu^* = \mu_0$$
$$H_1 : \mu^* = \mu_1,$$  \hspace{1cm} (3.2)

the unrestricted classification task is equivalent to testing the hypothesis,

$$H_0 : (X_1 - X_0)^T \Sigma^{-1} \mu^* = (X_1 - X_0)^T \Sigma^{-1} \mu_0$$
$$H_1 : (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu^* = (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu_1,$$

and similarly, the restricted classification task can be rewritten as testing the hypoth-
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

\[ H_0 : (R^T\Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu^* = (R^T\Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu_0 \]

\[ H_1 : (R^T\Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu^* = (R^T\Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu_1. \]

We establish the following theorem on classification error comparison analogous to the LRT power comparison stated in Theorem 1:

**Theorem 6.** Assume access to a sequence of labeled instances \((X_1, Y_1), ..., (X_n, Y_n) \overset{iid}{\sim} p_0\mathcal{N}(x; \mu_0, \Sigma) + p_1\mathcal{N}(x; \mu_1, \Sigma)\), where \(X_i \in \mathbb{R}^D\) represents an observed feature vector and \(Y_i \in \{0, 1\}\) is the corresponding observed class label for \(i = 1, \ldots, n\). Suppose \(\mu_k \in \mathcal{P} \subset \mathbb{R}^D\) and \(\mu_k = Rw_k\) for some known matrix \(R \in \mathbb{R}^{D \times d}\) and \(k = 0, 1\).

Assume that \(p_k\) and \(\mu_k\) are unobserved. Now we observe a sequence \(\{(X^*_i, Y^*_i)\}_{i=1}^N\) from an unobserved class label \(y^* \in \{0, 1\}\). When predicting \(y^*\), as \(N \to \infty\), the expected error of the restricted classifier in \(\mathcal{P}\) is uniformly smaller than that of the unrestricted classifier in \(\mathbb{R}^D\).

**Proof.** For multivariate Gaussians, it can be easily retrieved that the \(\alpha\)-divergence between \(f_0(x) = \mathcal{N}(x; \mu_0, \Sigma_0)\) and \(f_1(x) = \mathcal{N}(x; \mu_1, \Sigma_1)\) is

\[
D_\alpha(f_1||f_0) = \frac{1}{2} \log \frac{\left|\Sigma_0\right|^\alpha \left|\Sigma_1\right|^{1-\alpha}}{\left|\alpha \Sigma_0 + (1-\alpha) \Sigma_1\right|} + \frac{\alpha(1-\alpha)}{2} (\mu_0 - \mu_1)^T (\alpha \Sigma_0 + (1-\alpha) \Sigma_1)^{-1} (\mu_0 - \mu_1).
\]

When \(\Sigma_0 = \Sigma_1 = \Sigma\), the Chernoff information which is the optimal \(\alpha\)-divergence
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

can be retrieved when \( \alpha = \frac{1}{2} \) and can be written as

\[
\sup_{\alpha \in [0, 1]} D_\alpha(f_1||f_0) = \frac{1}{8}(\mu_0 - \mu_1)^T\Sigma^{-1}(\mu_0 - \mu_1). 
\] (3.3)

When predicting the class label of the sequence \( \{X_i^*\}_{i=1}^N \), the unrestricted classifier considers testing the hypothesis,

\[
H_0^D : (\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}\mu^* = (\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}\mu_0 \\
H_1^D : (\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}\mu^* = (\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}\mu_1,
\]

where \( \bar{X}_k = \frac{\sum_{j=1}^n X_j I_{Y_j = k}}{\sum_{j=1}^n I_{Y_j = k}} \); i.e., the mean of all samples in class \( k \) is the unrestricted MLE for \( k = 0, 1 \). Based on the fact that \( (\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}X_i^* \sim N((\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}\mu^*, (\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}(\bar{X}_1 - \bar{X}_0)) \), the Chernoff information defined in Equation (3.3) between \( H_0^D \) and \( H_1^D \) becomes

\[
CI_D = \frac{1}{8}[(\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}(\mu_1 - \mu_0)]^T((\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}(\bar{X}_1 - \bar{X}_0))^{-1}[(\bar{X}_1 - \bar{X}_0)^T\Sigma^{-1}(\mu_1 - \mu_0)] \\
= \frac{1}{8}\frac{[(\mu'_1 - \mu'_0)^T(U_1 - U_0)]^2}{\|U_1 - U_0\|^2}, 
\] (3.4)

where we define \( \mu'_k = \Sigma^{-1/2}\mu_k \) and \( U_k = \Sigma^{-1/2}\bar{X}_k \) for \( k = 0, 1 \).

As \( \bar{X}_k \sim N(\mu_k, \Sigma \frac{I}{N_k}) \), where \( N_k = \sum_{j=1}^n I_{Y_j = k} \), we have \( U_k \sim N(\Sigma^{-1/2}\mu_k, \frac{I}{N_k}) \) with \( I_D \) being a \( D \times D \) identity matrix. Then, geometrically Equation (3.4) is the square of a transformation of the correlation between a Gaussian variable \( (U_1 - U_0) \) and its
mean vector. Based on the classical result that the square of the correlation between Gaussian variables is Beta distributed \([22]\), we have

\[
\frac{[(\mu'_1 - \mu'_0)^T(U_1 - U_0)]^2}{\|U_1 - U_0\|^2\|\mu'_1 - \mu'_0\|^2} \sim \text{Beta}(\frac{1}{2}, \frac{D - 1}{2}).
\]

Thus, the expected error of the unrestricted classifier satisfies the following equation:

\[
E_D[\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}_e(N)] = -E_{X_0, X_1} \left[ \frac{1}{8} \frac{[(\mu'_1 - \mu'_0)^T(U_1 - U_0)]^2}{\|U_1 - U_0\|^2} \right]
\]

\[
\propto - \frac{1}{D} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1),
\]

where the proportion sign accounts for a function depending only on \(N_0, N_1\).

On the other hand, the restricted classifier considers testing the hypothesis,

\[
H^d_0 : (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu^* = (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu_0,
\]

\[
H^d_1 : (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu^* = (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu_1.
\]

Similar to the expected error analysis for the unrestricted classifier, based on the fact that \((R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} X_i^* \sim \mathcal{N}((R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} \mu^*, (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} R(R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0))\) for \(i = 1, 2, \ldots, D\)
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

\[ CI_d = \frac{1}{8} \frac{[(\mu''_1 - \mu''_0)^T (U'_1 - U'_0)]^2}{\|U'_1 - U'_0\|^2}, \quad (3.5) \]

where we define \( \mu''_k = (R^T \Sigma^{-1} R)^{-1/2} R^T \Sigma^{-1} \mu_k \) and \( U'_k = (R^T \Sigma^{-1} R)^{-1/2} R^T \Sigma^{-1} \tilde{X}_k \) for \( k = 0, 1 \). In that \( U'_k \sim \mathcal{N}(\mu''_k, \frac{l_k}{N_k}) \), we have similar results as in the unrestricted case,

\[
\frac{[(\mu''_1 - \mu''_0)^T (U'_1 - U'_0)]^2}{\|U'_1 - U'_0\|^2} \sim \text{Beta}(\frac{1}{2}, \frac{d - 1}{2}).
\]

Thus, the expected error of the restricted classifier satisfies the following equation:

\[
E_d[\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}_e(N)] = -E_{\tilde{X}_0, \tilde{X}_1} \left[ \frac{1}{8} \frac{[(\mu''_1 - \mu''_0)^T (U'_1 - U'_0)]^2}{\|U'_1 - U'_0\|^2} \right] \\
\propto -\frac{1}{d} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1),
\]

where the proportion sign accounts for the same function of \( N_0, N_1 \) as in the unrestricted case. The inequality,

\[
E_D[\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}_e(N)] > E_d[\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}_e(N)],
\]

concludes the theorem. \( \square \)

Theorem 6 is a result of the finite training sample size \( n \) and the infinite testing sample size \( N \).
3.2 Classification in Discrete Models (Multinomials)

In this section, we consider the classification expected error comparison of unrestricted classifiers and restricted classifiers in multinomial models under Setting I.

Classification tasks where \( f_{\theta_k}(x) \) is a multinomial distribution are ubiquitous in practice when the feature is categorical. For instance, text classification is a task of classifying documents by contents, such as by which words appear in the document. A document can be represented by a feature vector \( w = (w_1, \ldots, w_D) \), where \( w_k \) are the frequencies or the existence of the \( k \)th specific word. Often the assumption of the text classification problem is that \( w|C \) follows a multinomial distribution, where \( C \) is the document class. A typical text classification procedure starts with defining a vocabulary \( V \) where \( D = |V| \) is the dimension of feature vectors. Next, given a training set of documents, one can estimate \( p_k \) (priori probability of class \( k \)) and \( f_{\theta_k}(x) \) (conditional density of class \( k \)). Then, for a new document we estimate its class membership based on its feature vector.

The above example of text classification describes one application of the problem discussed in this section. Now we provide a general description of the problem.

Consider a multinomial model, \( Multi(\theta) \), where the outcomes occur with probabilities \( \theta = (\theta_1, \ldots, \theta_D) \), which is parametrized by the \( D - 1 \) dimensional unit simplex
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

$\Delta^{D-1} \in \mathbb{R}^D$, where $\Delta^{D-1} = \{\theta \in [0,1]^D : \sum_{i=1}^D \theta_i = 1\}$.

Suppose the outcome of a drawing of $n$ i.i.d. observations from $\text{Multi}(\theta)$ is $x = (x_1, \ldots, x_D)$, where $x_i$ records the number of occurrences of category $i$.

The unrestricted likelihood function of $\theta$ is

$$L_{\theta}(x) = \frac{n!}{\prod_{i=1}^D x_i!} \prod_{i=1}^D \theta_i^{x_i},$$

and the unrestricted MLE of $\theta$ is $\hat{\theta} = (\frac{x_1}{n}, \ldots, \frac{x_D}{n})$.

Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid $f_\theta(x, y)$, where $f_\theta(x) = p_0 f_0(x) + p_1 f_1(x)$ with $f_0(x) = \text{Multi}(\theta_0)$ and $f_1(x) = \text{Multi}(\theta_1)$ denote the class conditional probability densities. If $\theta_k$ is unknown for $k = 0, 1$, then the unrestricted Bayes plug-in decision function is

$$g^*_{\Delta^{D-1}}(x) = \begin{cases} 1 & \text{if } p_0 \hat{f}_0^{\Delta^{D-1}}(x) < p_1 \hat{f}_1^{\Delta^{D-1}}(x) \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{f}_k^{\Delta^{D-1}} = \text{Multi}(\hat{\theta}_k)$ with $\hat{\theta}_0 = (\frac{x_1}{N_0}, \ldots, \frac{x_D}{N_0})$, $\hat{\theta}_1 = (\frac{z_1}{N_1}, \ldots, \frac{z_D}{N_1})$, $x = (x_1, \ldots, x_D)$ and $z = (z_1, \ldots, z_D)$ record numbers of occurrences of class 0 and 1 respectively, i.e., $X_i I_{Y_i=0}$ and $X_i I_{Y_i=1}$.

3.2.1 Linear Restricted Submodels

Consider a trinomial model that is a multinomial model with $D = 3$. Define $\psi_L : [0,1] \to \Theta$ by $\psi_L(\tau) = (\tau, a\tau, 1 - (a + 1)\tau)$ for $\tau \in [0, \frac{1}{a+1}]$ and some constant $a$.
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

$a > 0$. Consider a linear subfamily that is parametrized by $\Psi_L = \{\psi_L(\tau), \tau \in [0, \frac{1}{a+1})\} \subset \Delta^2 \subset \mathbb{R}^3$ for $a > 0$.

The joint distribution $f_\theta(X, Y)$ has a conditional density distribution as $f(X)|_{Y=k} \sim \text{multinomial}(1, \theta_k)$ where $\theta_k = \psi_L(\tau_k) \in \Psi_L \subset \Delta^2$, with some $\tau_k \in [0, \frac{1}{a+1}]$ for $k = 0, 1$. For $k = 0, 1$, the restricted MLE of $\tau_k$ is $\hat{\tau}_k = \frac{m_k}{N_k(1+a)}$, where $m_0 = x_1 + x_2$ and $m_1 = z_1 + z_2$.

If $\tau_k$ is unknown, then the restricted Bayes plug-in decision function is

$$g^*_L(x) = \begin{cases} 1 & \text{if } p_0 \tilde{f}^\Psi_L(x) < p_1 \tilde{f}^\Psi_L(x) \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{f}^\Psi_L = \text{Multi}(\psi_L(\hat{\tau}_k))$.

We consider the expected errors $E[L_{\Psi_L}]$ (the restricted classification rule $g^*_L$) and $E[L_{\Delta^2}]$ (the unrestricted classification rule $g^*_{\Delta^2}$). We now investigate the relationship between the two expected errors when $n < \infty$. Figure 3.1 shows an example for finite $n$ and the fact that there is no uniform dominance between $E[L_{\Psi_L}]$ and $E[L_{\Delta^2}]$, which is computed exactly via R up to a default precision level with seven significant digits.
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

Figure 3.1: Assuming \( a = 2 \) in \( \psi_L \), the curve shows the difference \( E[L_{\psi_L}] - E[L_{\Delta^2}] \) when \( N_0 = N_1 = 5 \) \((p_0 = 0.5), \tau_0 = 0.1\), and ranging \( \alpha \tau_1 \in (0, 1) \).

Moreover, the phenomenon that \( E[L_{\psi_L}] \) is not uniformly smaller than \( E[L_{\Delta^2}] \) is far more general than the simple example in Figure 3.1.

**Theorem 7.** Let \( \Psi_L \) parametrize a submanifold of the trinomial with \( n \geq 2 \), and consider the classification problem with Setting I. For any \( n \geq 2 \), there exists a pair \((\tau_0, \tau_1)\) and \( p_0, p_1 \in (0, 1)\) as the true but unknown parameter of class 0,1, such that \( E[L_{\Delta^2}] < E[L_{\psi_L}] \), where \( L_{\Delta^2} \) and \( L_{\psi_L} \) are the errors of the unrestricted and the restricted Bayes plug-in rules.

**Proof.** Let \( N_k = np_k \) be the number of training observations from class \( k = 0 \) or 1. Let \((x_1, x_2, x_3) = \sum_{i=1}^{n} X_i I_{Y_i=0}\) and \((z_1, z_2, z_3) = \sum_{i=1}^{n} X_i I_{Y_i=1}\). Note that \( N_0 = \sum_{i=1}^{3} x_i \) and \( N_1 = \sum_{i=1}^{3} z_i \). Let \( \mathcal{X} = \{X_i\}_{i=1}^{n} \) denote the set of training data. Let \((e_1, e_2, e_3)\) be the standard basis for the 3-dimensional space.
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

Without loss of generality, suppose \( \tau_0 > \tau_1 \). Then the expected Bayes plug-in error of a testing instance, \( \mathbf{X}^* = (X_1^*, X_2^*, X_3^*) \), can be written as

\[
E_X[L(\mathbf{X}^*)] = E_X[\mathbb{P}(g^*(\mathbf{X}^*) \neq Y)]
\]

\[
= E_X[\sum_{i=1}^{3} \mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_i) < p_1 \hat{f}_1(\mathbf{e}_i)) f_0(\mathbf{e}_i) + \sum_{i=1}^{3} \mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_i) \geq p_1 \hat{f}_1(\mathbf{e}_i)) f_1(\mathbf{e}_i)].
\]

(3.6)

The unrestricted estimated densities are \( \hat{f}_0^\Delta^2(\mathbf{X}^*) = \frac{N_0!}{\prod_{i=1}^{3} x_i!} \left( \frac{x_1}{N_0} \right)^{X_1^*} \left( \frac{x_2}{N_0} \right)^{X_2^*} \left( \frac{x_3}{N_0} \right)^{X_3^*} \) and \( \hat{f}_1^\Delta^2(\mathbf{X}^*) = \frac{N_0!}{\prod_{i=1}^{3} x_i!} \left( \frac{m_1}{N_0(1+a)} \right)^{X_1^*} \left( \frac{n_0}{N_0(1+a)} \right)^{X_2^*} \left( 1 - \frac{m_1}{N_0(1+a)} \right)^{X_3^*} \). The restricted estimated densities are \( \hat{f}_0^\Psi^L(\mathbf{X}^*) = \frac{N_0!}{\prod_{i=1}^{3} x_i!} \left( \frac{m_0}{N_0(1+a)} \right)^{X_1^*} \left( \frac{m_1}{N_0(1+a)} \right)^{X_2^*} \left( 1 - \frac{m_1}{N_0(1+a)} \right)^{X_3^*} \) \( (1 - m_0)^{X_3^*} \).

As \( f_0(\mathbf{e}_i) \) is constant when taking the expected value \( E_X[L(\mathbf{X}^*)] \) over \( \mathcal{X} \) for \( i = 1, 2, 3 \), Equation (3.6) can be expansively written as

\[
E_X[L(\mathbf{X}^*)] = p_0 \tau_0 E_X[\mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_1) < p_1 \hat{f}_1(\mathbf{e}_1))] + p_1 \tau_1 E_X[\mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_1) \geq p_1 \hat{f}_1(\mathbf{e}_1))] \\
+ p_0 a \tau_0 E_X[\mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_2) < p_1 \hat{f}_1(\mathbf{e}_2))] + p_1 a \tau_1 E_X[\mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_2) \geq p_1 \hat{f}_1(\mathbf{e}_2))] \\
+ p_0 [1 - (1 + a) \tau_0] E_X[\mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_3) < p_1 \hat{f}_1(\mathbf{e}_3))] \\
+ p_1 [1 - (1 + a) \tau_1] E_X[\mathbb{P}(p_0 \hat{f}_0(\mathbf{e}_3) \geq p_1 \hat{f}_1(\mathbf{e}_3))].
\]

Furthermore, \( p_0 \hat{f}_0^\Delta^2(\mathbf{e}_1) < p_1 \hat{f}_1^\Delta^2(\mathbf{e}_1) \) is equivalent to \( \frac{N_0}{n} \frac{x_1}{N_0} < \frac{N_1}{n} \frac{z_1}{N_1} \), i.e., \( x_1 < z_1 \).

Similarly, \( p_0 \hat{f}_0^\Psi^L(\mathbf{e}_1) < p_1 \hat{f}_1^\Psi^L(\mathbf{e}_1) \) is equivalent to \( \frac{N_0}{n} \frac{m_0}{N_0(1+a)} < \frac{N_1}{n} \frac{m_1}{N_1(1+a)} \), i.e., \( m_0 < m_1 \).
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

$m_1$. Similar arguments can be extended to terms with respect to $e_2$ and $e_3$.

We expansively write the difference between the expected Bayes plug-in error of
the unrestricted model and that of the restricted model as

$$E_X[L_{D^2}(X^*)] - E_X[L_{\Psi_L}(X^*)]$$

$$= p_0 \tau_0 [\mathbb{P}(x_1 < z_1) - \mathbb{P}(m_0 < m_1)] + p_1 \tau_1 [\mathbb{P}(x_1 \geq z_1) - \mathbb{P}(m_0 \geq m_1)]$$

$$+ p_0 a \tau_0 [\mathbb{P}(x_2 < z_2) - \mathbb{P}(m_0 < m_1)] + p_1 a \tau_1 [\mathbb{P}(x_1 \geq z_1) - \mathbb{P}(m_0 \geq m_1)]$$

$$+ p_0 [1 - (a + 1) \tau_0] [\mathbb{P}(x_3 < z_3) - \mathbb{P}(N_0 - m_0 < N_1 - m_1)]$$

$$+ p_1 [(1 - (a + 1) \tau_1)] [\mathbb{P}(x_3 \geq z_3) - \mathbb{P}(N_0 - m_0 \geq N_1 - m_1)].$$

As $x_3 = N_0 - m_0 \sim Binomial(N_0, 1 - (1 + a) \tau_0)$ and $z_3 = N_1 - m_1 \sim Binomial(N_1, 1 - (1 + a) \tau_1)$, the last two terms are 0. We are then left to consider:

$$E_X[L_{D^2}(X^*)] - E_X[L_{\Psi_L}(X^*)] = (p_0 \tau_0 - p_1 \tau_1) [\mathbb{P}(x_1 < z_1) - \mathbb{P}(m_0 < m_1)]$$

$$+ (p_0 a \tau_0 - p_1 a \tau_1) [\mathbb{P}(x_2 < z_2) - \mathbb{P}(m_0 < m_1)]. \quad (3.7)$$

For a $D$-categorical multinomial distribution $Multi(\theta), \theta \in \mathbb{R}^D$, we can view it as a $(D - \ell)$-categorical multinomial by combining $\ell + 1$ categories into one. For instance, for a multinomial model indexed by $\theta = (\theta_1, \ldots, \theta_D)$, we say there exists a reduced multinomial model indexed by $(\sum_{i=1}^{\ell+1} \theta_i, \theta_{\ell+2}, \ldots, \theta_D)$. Here we observe binomial models reduced from the trinomial model: $x_1 \sim Binomial(N_0, \tau_0), z_1 \sim$
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

Binomial($N_1, \tau_1$) and $x_2 \sim Binomial(N_0, a\tau_0)$, $z_2 \sim Binomial(N_1, a\tau_1$). Similarly, $m_0 \sim Binomial(N_0, (1 + a)\tau_0$) and $m_1 \sim Binomial(N_1, (1 + a)\tau_1$).

If assuming that $n$ tends to infinity, we use a univariate Gaussian distribution to approximate a binomial distribution $Binomial(n, p)$, then by the Berry-Esseen theorem [23], with $0.4097 \leq C < 0.4748$, the difference between the binomial distribution and the approximated Gaussian distribution is bounded by

$$|F_n(x) - \Phi(x)| \leq \frac{Cp(1-p)(p^2 + (1-p)^2)}{(\sqrt{p(1-p)})^3 \sqrt{n}} = \frac{C(p^2 + (1-p)^2)}{\sqrt{p(1-p)} \sqrt{n}},$$

where $F_n(x)$ denotes the binomial distribution CDF and $\Phi(x)$ denotes a univariate Gaussian CDF.

The probability $P(x_1 < z_1)$ can be approximated by $\Phi(-\frac{N_0\tau_0 - N_1\tau_1}{\sqrt{N_0\tau_0(1-\tau_0) + N_1\tau_1(1-\tau_1)}})$, and the approximation error is of order $O(\frac{1}{\sqrt{n}})$. Similarly, the probability $P(m_0 < m_1)$ can be approximated by $\Phi(-\frac{(N_0\tau_0 - N_1\tau_1)\sqrt{1+a}}{\sqrt{N_0\tau_0(1-(1+a)\tau_0) + N_1\tau_1(1-(1+a)\tau_1)}})$, and the approximation error is of order $O(\frac{1}{\sqrt{n}})$ as well.

If $p_0\tau_0 - p_1\tau_1 = \frac{h}{\sqrt{n}}$, with $h = \Theta(1)$ and $h > 0$, then:

$$P(x_1 < z_1) - P(m_0 < m_1) + O(\frac{1}{\sqrt{n}})$$

$$= \Phi(-\frac{N_0\tau_0 - N_1\tau_1}{\sqrt{N_0\tau_0(1-\tau_0) + N_1\tau_1(1-\tau_1)}}) - \Phi(-\frac{(N_0\tau_0 - N_1\tau_1)\sqrt{1+a}}{\sqrt{N_0\tau_0(1-(1+a)\tau_0) + N_1\tau_1(1-(1+a)\tau_1)}})$$

$$= \Phi(-\frac{h}{\sqrt{p_0\tau_0(1-\tau_0) + p_1\tau_1(1-\tau_1)}}) - \Phi(-\frac{h\sqrt{1+a}}{\sqrt{p_0\tau_0(1-(1+a)\tau_0) + p_1\tau_1(1-(1+a)\tau_1)}}) > 0.$$

(3.8)
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

The difference in Equation (3.8) is of order $\Omega(h)$; thus, the difference $\mathbb{P}(x_1 < z_1) - P(m_0 < m_1) > 0$. Similarly, $\mathbb{P}(x_2 < z_2) - \mathbb{P}(m_0 < m_1) > 0$ can be shown. The inequality in Equation (3.8) is due to the fact that $f(a) = \frac{h\sqrt{1+a}}{\sqrt{p_0\tau_0(1-(1+a)\tau_0)+p_1\tau_1(1-(1+a)\tau_1)}}$ as a function of $a$ increases as $a > 0$ increases and it achieves its minimum when $a = 0$.

It follows that if $p_0\tau_0 - p_1\tau_1 = \frac{h}{\sqrt{n}}$, with $h = \Theta(1)$, then asymptotically the expected error of the restricted classifier is no greater than that of the unrestricted classifier, i.e.,

$$\lim_{n \to \infty} \{E_X[L_{\Delta^2}(X^*)] - E_X[L_{\Psi_n}(X^*)]\} \geq 0.$$ 

On the other hand, we consider the matter when $n$ is finite. The goal is to check if Equation (3.7) has a positive probability of being negative.

If we choose $\tau_0 < \min\{\frac{w-1}{w(a+1)-a}, \frac{v-1}{v(a+1)-1}\}$, where $w = (\frac{1+a}{a})^{\frac{1}{2n-1}}$ and $v = (1 + a)^{\frac{1}{2n-1}}$, then the following statements are true simultaneously:

$$\mathbb{P}(x_1 < z_1) - \mathbb{P}(m_0 < m_1) < 0,$$

$$\mathbb{P}(x_2 < z_2) - \mathbb{P}(m_0 < m_1) < 0.$$ 

We emphasize that the choice of $\tau_0, \tau_1$ specified here is not a necessary condition, but a sufficient one.
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

Then for $\tau_1$ such that $\tau_1 < \min\{\frac{N_0}{N_1}, \tau_0\}$, we observe

$$p_0\tau_0 - p_1\tau_1 > 0.$$ 

Together, we conclude that by choosing $\tau_0 < \min\{\frac{w-1}{w(a+1)-a}, \frac{v-1}{v(a+1)-1}\}$ and $\tau_1 < \min\{\frac{N_0}{N_1}, \tau_0\}$, where $w = \left(\frac{1-a}{a}\right)^{\frac{1}{2n-1}}$ and $v = (1+a)^{\frac{1}{2n-1}}$, the unrestricted classifier yields smaller expected error than the restricted classifier, i.e., $E[L_{\Delta^2}] - E[L_{\Psi_L}] < 0.$

The above theorem reveals that the phenomenon that existed in testing, in which restricted models do not have uniformly better inference than unrestricted models, also appears in classification.

### 3.2.2 Hardy-Weinberg Equilibrium

Chapters 2 and 3 both compare inference in unrestricted models and in restricted models. The difference is that Chapter 2 considers LRTs as the inference task, while Chapter 3 considers classification. To tie the two chapters more closely together, we consider the Hardy-Weinberg equilibrium as the embedded submanifold of $\Delta^2$ in classification in hopes of obtaining analogous results to Theorem 3 in Chapter 2.

Consider $\Psi$ that parametrizes the Hardy-Weinberg submodel of $\Delta^2$ defined in Section 2.3.1. The joint distribution $f_0(X, Y)$ has a conditional density distribution as $f(X)|_{Y=k} \sim \text{multinomial}(1, \theta_k)$ where $\theta_k = \psi(\tau_k) \in \mathcal{H} \subset \Delta^2$, with some $\tau_k \in [0, 1]$.
for $k = 0, 1$. The restricted MLE of $\tau_k$ is $\hat{\tau}_k = \frac{m_k}{N_k}$ where $m_0 = x_1 + \frac{1}{2} x_2$ and $m_1 = z_1 + \frac{1}{2} z_2$.

If $\tau_k$ are unknown, then the restricted Bayes plug-in decision function is

$$g_H^*(x) = \begin{cases} 1 & \text{if } p_0 \hat{f}_0^H(x) < p_1 \hat{f}_1^H(x) \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{f}_k^H = Multi(\psi(\hat{\tau}_k))$ for $k = 0, 1$.

We consider the expected errors $E[L_{\Delta^2}]$ and $E[L_H]$ corresponding to the unrestricted classification rule $g_{\Delta^2}^*$ and the restricted classification rule $g_H^*$, respectively. We now investigate the relationship between the two expected errors when $n < \infty$.

Figure 3.2 shows an example for finite $n$ and the fact that there is no uniform dominance between $E[L_H]$ and $E[L_{\Delta^2}]$, which is computed exactly via R by Equation (3.9) up to a default precision level with seven significant digits.
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

Figure 3.2: The curve shows the difference $E[L_H] - E[L_{\Delta^2}]$ when $n = 10$, $\tau_0 = 0.3$, $p_0 = 0.2$ and ranging $\tau_1 \in (0, 1)$.

The restricted estimated densities are $\hat{f}_0^H(X^*) = \frac{N_0!}{\prod_{i=1}^{n_0} x_i!} 2^{X_1^2} (\frac{m_0}{N_0})^{2X_1^2 + X_2^2} (1 - \frac{m_0}{N_0})^{2X_3^2 + X_2^2}$ and $\hat{f}_1^H(X^*) = \frac{N_1!}{\prod_{i=1}^{n_1} z_i!} 2^{X_1^2} (\frac{m_1}{N_1})^{2X_1^2 + X_2^2} (1 - \frac{m_1}{N_1})^{2X_3^2 + X_2^2}$. We expansively write the difference between the expected Bayes plug-in error of the unrestricted model and that of the restricted model as

$$E_X[L_{\Delta^2}] - E_X[L_H] = p_0\tau_0^2[\mathbb{P}(x_1 < z_1) - \mathbb{P}(m_0 < m_1)] + p_1\tau_1^2[\mathbb{P}(x_1 \geq z_1) - \mathbb{P}(m_0 \geq m_1)]$$

$$+ p_02\tau_0(1 - \tau_0)[\mathbb{P}(x_2 < z_2) - \mathbb{P}(m_0(N_0 - m_0) < m_1(N_1 - m_1))]$$

$$+ p_12\tau_1(1 - \tau_1)[\mathbb{P}(x_1 \geq z_1) - \mathbb{P}(m_0(N_0 - m_0) \geq m_1(N_1 - m_1))]$$

$$+ p_0(1 - \tau_0)^2[\mathbb{P}(x_3 < z_3) - \mathbb{P}(N_0 - m_0 < N_1 - m_1)]$$

$$+ p_1(1 - \tau_1)^2[\mathbb{P}(x_3 \geq z_3) - \mathbb{P}(N_0 - m_0 \geq N_1 - m_1)].$$

(3.9)
CHAPTER 3. ON CLASSIFICATION IN DIMENSION-RESTRICTED SUBMODELS

Here we observe binomial models reduced from the trinomial model: 

\[ x_1 \sim \text{Binomial}(N_0, \tau_0^2), \]
\[ z_1 \sim \text{Binomial}(N_1, \tau_1^2), \]
\[ x_2 \sim \text{Binomial}(N_0, 2\tau_0(1 - \tau_0)), \]
\[ z_2 \sim \text{Binomial}(N_1, 2\tau_1(1 - \tau_1)) \]
\[ x_3 \sim \text{Binomial}(N_0, (1 - \tau_0)^2), \]
\[ z_3 \sim \text{Binomial}(N_1, (1 - \tau_1)^2). \]

Similarly, we observe 

\[ m_0 \sim \text{Binomial}(N_0, \tau_0) \]
\[ m_1 \sim \text{Binomial}(N_1, \tau_1). \]

The proof for Theorem 7 cannot be easily generalized here. However, from vast experiments with different choices of \( n \), we conjecture that the following is true:

**Conjecture 1.** Consider the Hardy-Weinberg as submodel of \( \Delta^2 \). For any \( 2 < n < \infty \), there exists a pair \( (\tau_0, \tau_1) \) as the true parameter of class 0/1, respectively, and a prior probability \( p_0 \), such that the restricted classifier has a greater expected error than the unrestricted classifier.
Chapter 4

Manifold Learning in Network Inference

One form of high-dimensional data emerges in network analysis. To gain insight into a network, researchers typically start with the identification of its latent structure. Given an adjacency matrix representing a network, depending on how much information about the network is available, different methodologies can be introduced to more adequately fulfill the structure discovery task. In this chapter, we shift our focus to graphs that by nature consist of high-dimensional data. The dimensionality of graph data is twofold: one dimension is in the adjacency matrix and the other is in its latent position space.

Initially, we consider an adjacency matrix representing a network of $n$ nodes, either unweighted or weighted. It is obvious that each node $i$ has a feature vector in $\mathbb{R}^n$ with
its dimension \( j \) providing the connectivity information between nodes \( i \) and \( j \). Thus, we have a data set of size \( n \) in \( n \)-dim; the dimensionality increases when more nodes come into the network. To resolve the exploding dimensionality issue and to gain more generalizable information, one often utilizes a Latent Position Model (LPM) \([24]\), which assumes that the adjacency between any pair of nodes depends on their intrinsic features via a link function. The intrinsic features of each node form a latent position vector, and all the latent position vectors form the latent position space.

A Random Dot Product Graph (RDPG) \([25]\), as a widely used class of LPM, defines the link function to be the inner product of two latent position vectors. Given an adjacency matrix as an instance of a probabilistic model such as the RDPG, the latent position space is usually unobserved. Adjacency Spectral Embedding (ASE) \([26]\) or Laplacian Spectral Embedding (LSE) \([27]\) provide a consistent and asymptotically normal estimate of the latent position space. Since ASE as a linear embedding method is not able to decode nonlinear submanifolds embedded in the latent position space, we propose applying manifold learning algorithms to the latent position space to achieve a better understanding of the nonlinear structure for the sake of subsequent inference. Such a nonlinear structure can exist only in parts of nodes within the network \([28]\).

This chapter begins with an overview of some basic concepts of random graphs for network inference, in particular RDPGs. Next, we consider a framework for network structure discovery via manifold learning applied to the ASE space for an RDPG.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Given a hypothesis testing problem, the unrestricted test is performed in the ASE space $\mathcal{A}$, while the restricted test takes place in the true embedded manifold space $\mathcal{M}$ where the latent positions of RDPGs belong. We define the learnt restricted test in the learnt embedded manifold space $\hat{\mathcal{M}}$. To answer $\textbf{Q1}$ and $\textbf{Q2}$ raised in Chapter 1, we compare the three powers of $\pi_{\mathcal{A}}, \pi_{\mathcal{M}},$ and $\pi_{\hat{\mathcal{M}}}$ corresponding to the unrestricted test, the restricted test, and the learnt restricted test respectively. We show that the RDPG network inference procedure developed here yields higher power if suitable manifold learning has been applied before the inference task.

In this chapter, we use boldface $\mathbf{X}$ to denote a matrix and $X_i$ to represent the $i$th row of the matrix.

4.1 Random Graph Models

We start by recalling some basic concepts of random graphs and random graph inference. A graph $G$ is an ordered pair of $(V, E)$, where $V$ is the vertex or node set and $E$ is the edge set. Let $|V|$ denote the number of vertices. There is an edge between vertex $i$ and $j$ if $(i, j) \in E$. Each graph $G = (V, E)$ with $|V| = n$, can be represented by an adjacency matrix $A \in \{0, 1\}^{n \times n}$, such that $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise.

For a random graph, the adjacency between a pair of vertices depends on their unobserved properties. To elucidate, we consider a social network, where whether
any two individuals are friends or not depends on their characters, hobbies, etc. We may represent all relevant features of an individual by a vector. Another example of this can be found in a brain connectome, in which connectivity patterns may be relevant to brain regions and neuron types. Latent Position Models (LPM) proposed by Hoff [24] captures such network structures, where the adjacency between any pair of nodes depends on a link function of the unobserved intrinsic properties of the node pair of interest.

One type of LPM of broad interest is the the Random Dot Product Graph (RDPG), which defines the inner product as the link function.

**Definition 1** (Directed RDPGs). Let \( X_i = (X_{i,L}, X_{i,R}) \) denote the latent position of vertex \( i \) and the latent position matrix \( X = [X_1, \ldots, X_n]^T = [X_L, X_R] \), where \( X_{i,L} \in \mathbb{R}^{d_L}, X_{i,R} \in \mathbb{R}^{d_R} \). Let \( F \) be a distribution on a set \( \mathcal{X} = \mathcal{X}^L \times \mathcal{X}^R \subseteq \mathbb{R}^{d_L} \times \mathbb{R}^{d_R} \) such that the inner product of any \( x_L \in \mathbb{R}^{d_L} \) and any \( x_R \in \mathbb{R}^{d_R} \) is a probability. Suppose \( X_i \overset{iid}{\sim} F \) and \( P = \rho_n X_L X_R^T \), with sparsity factor \( \rho_n \in (0, 1) \) known. We observe a hollow matrix \( A \in \{0, 1\}^{n \times n} \) and \( A \) is defined as

\[
P(A | X) = \prod_{i \neq j} (\rho_n X_{i,L}^T X_{j,R})^{A_{ij}} (1 - \rho_n X_{i,L}^T X_{j,R})^{1-A_{ij}},
\]

then \((A, X) \sim \text{RDPG}(F)\) is an instance of a directed RDPG.

For an *undirected RDPG*, we have \( X_{i,L} = X_{i,R} = X_i \in \mathcal{X} \subseteq \mathbb{R}^D \) as the latent position of vertex \( i \) and we consider \( X = [X_1, \ldots, X_n]^T \). Now \( F \) is a distribution on a
set $\mathcal{X} \times \mathcal{X} \subset \mathbb{R}^D \times \mathbb{R}^D$ such that the inner product of any $x, y \in \mathbb{R}^D$ is a probability. Suppose $X_i \overset{iid}{\sim} F$ and $P = \rho_n \mathbf{X}\mathbf{X}^T$, with sparsity factor $\rho_n \in (0, 1]$ known. Thus, $A$ is sampled according to $\mathbb{P}(A|X) = \prod_{i \neq j} (\rho_n X_i^T X_j)^{A_{ij}}(1 - \rho_n X_i^T X_j)^{1 - A_{ij}}$ in an undirected RDPG model.

It is worth mentioning that RDPGs are non-identifiable, or in other words for any orthogonal matrix $W \in \mathbb{R}^{D \times D}$, $XW$ as the latent position matrix after rotation $W$ yields the same probability matrix $P$ as $X$ does, for $P = XX^T = (XW)(XW)^T$.

For RDPGs, the latent position vectors $X_i$ form meaningful low-dimensional representations. Furthermore, the latent position space as a subset of $\mathbb{R}^D$ may form a manifold of lower intrinsic dimension $d$. We call such a model the Latent Structure Model. The one-dimensional latent structure model has been introduced in [29].

**Definition 2** (1-Dimensional Latent Structure Model). Let $C$ be a smooth curve of intrinsic dimension $d = 1$ and $\psi(t) : [0, 1] \rightarrow C$ denote the parametrization of $C$. Let $G$ be a family of distributions on $[0, 1]$ and $F$ be the induced distributions on $C$: $\mu_F(B) = \mu_G(\psi^{-1}(B))$ for any set $B \subset C$, where $\mu_F$ and $\mu_G$ represent the distribution measures associated with $F$ and $G$. An RDPG with i.i.d latent position matrix $X$ is a parametric latent structure random graph with known univariate support $C$ and underlying distribution $G$ if the latent position vectors $X_i$ are distributed according to $F = G(\psi^{-1})$ where $G$ belongs to some regular parametric family $G_\theta = \{G_\theta : \theta \in$
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

\[ \Theta \in \mathbb{R}^D \] on \([0, 1]\) with \(\psi\) and \(C\) known,

\[ X_i \sim F = G_{\psi}(\psi^{-1}). \]

An RDPG with i.i.d latent position matrix \(X\) is a nonparametric latent structure random graph with known univariate support \(C\) if \(\psi\) and \(C\) are both known, and \(F = G(\psi^{-1})\), where \(G \in \mathcal{G}\) with \(\mathcal{G}\) being a family of distributions on \([0, 1]\) that is not a subset of any regular parametric family of distributions on \([0, 1]\).

Given a network and its adjacency matrix, we are interested in the inference under the assumption that the adjacency matrix is an RDPG. To estimate the latent positions of an RDPG, Adjacency Spectral Embedding has been shown to yield consistent estimates [26].

**Definition 3** (Adjacency Spectral Embedding). Given a positive integer \(D\), the Adjacency Spectral Embedding (ASE) of an adjacency matrix \(A\) is defined as \(\bar{X} = U_A \mathcal{S}^{\frac{1}{2}}_A\), where

\[
A = U_A \mathcal{S}_A U_A^T + U_A^\perp \mathcal{S}_A^\perp U_A^\perp T
\]

is the spectral decomposition of \(A\), \(\mathcal{S}_A\) is the diagonal matrix of the \(D\) largest eigenvalues of \(A\) in non-increasing order and \(U_A\) is the \(n \times D\) matrix whose columns are the corresponding eigenvectors. Similarly, \(\mathcal{S}_A^\perp\) is the diagonal matrix of the remaining
n – D eigenvalues of $A$ in non-increasing order and $U_A^+$ is the $n \times (n - D)$ matrix whose columns are the corresponding eigenvectors.

Consider the ASE of a probability matrix $P = XX^T$ in an undirected RDPG setting, where the latent position matrix is $X \in \mathbb{R}^{n \times D}$, making the $D$-dimensional ASE of $P$ recover $X$ up to an orthogonal rotation. Given an adjacency matrix $A$ that is an instance of the undirected RDPG with $P$ as the probability matrix, then $A$ can be viewed as $P$ plus a matrix of noise. Thus, the $D$-dimensional ASE of $A$ is a reasonable estimation of $X$ up to an orthogonal rotation.

The theorem in [30] shows that the latent position estimates via ASE are asymptotically Gaussian and that they are distributed around the true latent positions.

**Theorem 8** (Central Limit Theorem for ASE in RDPGs). Let $(A_n, X_n) \sim RDPG(F)$ be a sequence of instances of a $d$-dimensional random dot product graph according to an inner product distribution $F$. Let $\Phi(x, \Sigma)$ denote the cdf of a (multivariate) Gaussian distribution with mean zero and covariance matrix $\Sigma$, evaluated at $x \in \mathbb{R}^D$. Then there exists a sequence of orthogonal $D \times D$ matrices $\{W_n\}$ such that for all $z \in \mathbb{R}^D$ and for any fixed index $i$,

$$
\lim_{n \to \infty} P[\sqrt{n}(\hat{X}_iW_n - X_i) \leq z] = \int \Phi(z, \Sigma(x))dF(x),
$$
where

\[
\Sigma(x) = \Delta^{-1} E[(x^T X_1 - (x^T X_1)^2)X_1X_1^T] \Delta^{-1}
\] (4.1)

with \( \Delta = E[X_1X_1^T] \) and \( X_1 \sim F \).

In the following sections we focus on undirected RDPGs, and the default setting of the sparse factor is \( \rho_n = 1 \) unless specified otherwise.

### 4.2 Manifold Learning in ASE Space

Graphs form an important class of high-dimensional data. The dimensionality of graphs is twofold: one dimension lies in the adjacency matrix and the other dimension lies in the latent position space. We focus on the second dimension in this study. When we estimate the latent position space to reduce the dimensions of the adjacency matrix, Adjacency spectral embedding (ASE) and Laplacian spectral embedding (LSE) provide consistent solutions [26, 27]. These methods seek to find the linearly embedded space of RDPGs. When we encounter a \( D \)-dimensional RDPG that is also a 1-dimensional LSM, these methods themselves will detect the \( D \)-dimensional spectral embedded space while failing to shed light on the 1-dimensional nonlinear structure.

For a 1-dimensional LSM, which is the focus of our inference task, we propose to first treat it as an RDPG, using spectral embedding methods to obtain its \( D \)-
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

dimensional latent positions; second, we apply manifold learning algorithms to the $D$-dimensional latent positions to obtain the learnt 1-dimensional representations.

Manifold learning algorithms aim to map data from high-dimensions to low-dimensions such that their intrinsic geometry is preserved. This more concise data representation is practical for data visualization and subsequent inference such as classification, among others. Various manifold learning algorithms have been proposed in the last two decades, among which practitioners typically select the ones to suit their task needs. For instance, t-SNE [31] is widely used when it comes to embedding high-dimensional data for visualization purpose. While we are concerned with an efficient low-dimensional representation of each node in the network, often the ultimate goal of such a network structure discovery process is to answer a certain inference query. Suppose we have sufficient computational capacity. Then, global manifold learning approaches may be favored since our interest is to preserve geometry at all scales for the whole latent position space.

Moreover, when the ultimate inference task involves measuring distances between pairs of latent position vectors, we often consider Euclidean and geodesic distances. While the Euclidean distance is an intuitive solution when we consider the latent space of RDPGs, the geodesic distance is usually a better choice when there exists an embedded submanifold in the latent space. Only a few manifold learning algorithms provide the guarantee of preserving any sort of distance. Normalized-output algorithms, including Locally Linear Embedding (LLE) [3], Laplacian Eigenmaps [32],
Local Tangent Space Alignment (LTSA) [33], Hessian Eigenmaps (HLLE) [34], and Diffusion maps [35] are not expected to recover geodesic distances [36]. Verma [37] recently proposed two low-dimensional embedding methods to approximately preserve geodesic distances for general manifolds. As a more widely-used method, Isomap asymptotically preserves pairwise geodesic distances as one increases the amount of data sampled from the underlying manifold that satisfies certain assumptions [2]. In this study, we focus on Isomap.

Despite the desired properties of Isomap, little is known about how Isomap functions when there is noise in the data. [38] reported that Isomap suffers from topological instabilities when the input data from the Swiss roll are noised with a small yet designed noise magnitude. Tenenbaum replied that it can be solved by finding a suitable neighborhood size to yield topologically stable embeddings. Pertinent literature has focused on developing variations of Isomap to improve the robustness when confronting noisy input data for all kinds of applications.

In our study, we investigate the application of Isomap in the latent position space of RDPGs and provide the approximate equivalence between the estimated geodesic distances via Isomap and the corresponding true geodesic distances. This result is the first one regarding the asymptotic convergence of Isomap on noisy data of which we are aware.

To bring this to light, we consider an inference task on the graph of the right hemisphere mushroom body (MB) connectome of the larval Drosophila brain [28].
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

The connectome consists of four distinct types of neurons - Kenyon Cells (KC), Input Neurons (MBIN), Output Neurons (MBON), Projection Neurons (PN) – with directed activity that can be represented by a stochastic block matrix $B$, where a 0 entry indicates no connection at all between the corresponding two types of neurons. There are a total of 213 neurons, among which 100 are KC neurons and 7536 directed edges. While $B$ defaults to 4 clusters – each of one type of neurons – Priebe [28] instead observed 6 clusters among the 213 neurons. The $GMM \circ ASE$ method that composes Gaussian mixture modeling with adjacency spectral embedding proposed in [28] estimates the neurons’ latent positions $\hat{X}_1, \ldots, \hat{X}_{213}$, which are embedded in $\mathbb{R}^6$, and suggests that all KC neurons’ estimated latent positions $\hat{X}_{KC,1}, \ldots, \hat{X}_{KC,100}$ form a 1-dimensional quadratic curve $C_{KC}$ (1-dimensional LSM).

Li [39] described so-called “claws” associated with each KC neuron and suggests a relationship between the neuron age and its number of claws. A question raised is whether 3-claw neurons and 4-claw neurons have the same generating latent position vector. If we formulate the question as a two-sample testing problem that treats 3-claw neurons and 4-claw neurons as two independent populations, then the null hypothesis becomes: the two samples of estimated latent positions suggest that the two populations have the same generating point. To perform the test in $\mathbb{R}^6$ with $\hat{X}_{KC,3\text{-claw},1}, \ldots, \hat{X}_{KC,3\text{-claw},n_3}$ and $\hat{X}_{KC,4\text{-claw},1}, \ldots, \hat{X}_{KC,4\text{-claw},n_4}$ is straightforward, yet hardly optimal as $n_3 = 15$ (number of 3-claw neurons) and $n_4 = 16$ (number of 4-claw neurons) are inadequate for testing 6-dimensional vectors. Due to the curve
structure of all KC neurons, instead of testing 6-dimensional vectors, we propose performing the test on the 1-dimensional representations of the two samples. All other KC neurons except 3-claw and 4-claw ones may be viewed as auxiliaries to help estimate the curve $C_{KC}$. This procedure is similar to that described in Section 2.4.

In the following sections, we show via theory and simulations that testing with the $d$-dimensional ($d = 1$) manifold representations is superior to that with the $D$-dimensional ASE representations. We start by outlining the problem considered for our theorems and simulations.

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} F$ on $M \subset A \subset \mathbb{R}^D$, where $M$ is a smooth, compact low-dimensional submanifold of intrinsic dimension $d$ embedded in a high-dimensional manifold $A$ of intrinsic dimension $D$. There exists a mapping $\rho : U \to M$ from a low-dimensional Euclidean feature space $U \subset \mathbb{R}^d$ to $M \subset \mathbb{R}^D$.

Consider $G = G(X)$ to be the RDPG latent position model, where $X$ denotes the $n \times D$ matrix with rows given by $X_i$. Let $\hat{X} = ASE(G)$ be the $n \times D$ matrix with rows given by the adjacency spectral embedding (ASE) latent position estimates $\hat{X}_i$. The Central Limit Theorem (CLT) of RDPGs says that $\hat{X}_i W$ will be approximately Gaussian around $X_i$ in $\mathbb{R}^{\text{rank}(XX^T)=D}$ for some orthogonal (rotation) matrix $W$.

If $M$ is a linear submanifold of $A$, in other words there exists $R \in \mathbb{R}^{d \times D}$, such that $X = wR$ for some $w \in \mathbb{R}^{d \times n}$, then the optimal dimension of latent position space chosen should be $d$ instead of $D$, because there exists a certain $w_t \in \mathbb{R}^{n \times d}$ such that $P = XX^T = wRR^T w^T = w_tw_t^T$. 

88
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

On the other hand, if $\mathcal{M}$ is a nonlinear submanifold embedded in $\mathcal{A}$, then ASE is not able to observe its presence. In general, the ASE should only be able to detect $D$, though there is a lower-dimensional representation for the latent positions. We propose applying manifold learning algorithms such as Isomap on the estimated ASE space to obtain a mapping estimate $\rho^{-1}$ of $\rho^{-1} : \mathcal{M} \rightarrow U$ with the collection $\{\hat{X}_i, W\}_{i=1}^n$. For convenience, let $\hat{\mathcal{M}}$ denote the collection of learnt low-dimensional representations in $d$ or the learnt mapping $\rho^{-1}$. In manifold learning, often it is the lower-dimensional representation of the mapping $\rho^{-1}$ that a manifold learning algorithm learns.

For network models – and we consider RDPG models in particular – we compare the efficiency of inference on $\mathcal{A}$, on $\mathcal{M}$, and on $\hat{\mathcal{M}}$ in order to pursue answers for $Q1$ and $Q2$ raised in Chapter 1.

Given a subsequent inference, we consider the hypothesis testing specified by $(H_0, H_A, \alpha)$ and a test statistic $T(\hat{X})$ measuring the distance between the null and the sample estimate. We have three power functions to consider, $\pi_A$ vs. $\pi_M$ vs. $\pi_{\hat{M}}$, which correspond to the unrestricted test, the restricted test and the learnt restricted test, respectively. We will compare $\pi_A$ vs. $\pi_M$ for $Q1$ and $\pi_M$ vs. $\pi_{\hat{M}}$ for $Q2$. In the following sections, we will consider both one-sample testing and two-sample testing.

In one-sample testing, we consider a graph of $n$ vertices and assume $s \ll n$ of them are restricted to or around a particular latent position $X^* \in \mathcal{A}$, i.e., $X_1, \ldots, X_s \overset{iid}{\sim} h_{X^*}$ for a function $h_{X^*}$ of small or zero variance, where $\mathcal{A}$ represents the latent position.
space. The one-sample testing considers whether $X^*$ is the same as a specific latent position $p_0$, where $p_0 \in \mathcal{M} \subset \mathcal{A}$. The restricted test assumes $X^* \in \mathcal{M}$, while the unrestricted test works with $X^* \in \mathcal{A}$. We define the learnt restricted test as testing with the estimated low-dimensional representation, and we consider the following procedure. We first learn $\hat{\rho}^{-1}$ via a manifold learning algorithm with the collection of ASE latent position estimates after the appropriate rotation, i.e., $\{\hat{X}_iW\}_{i=1}^n$. Second, we perform the test with the $d$-dimensional embeddings of $\hat{X}_iW$ and $p_0$, i.e., $\{\rho^{-1}(\hat{X}_iW)\}_{i=1}^n$ and $\hat{\rho}^{-1}(p_0)$. This learnt restricted test procedure is similar to that in Section 2.4. Both problems learn a $d$-dimensional representation via estimated manifold mapping with auxiliary data that are noisy samples from $\mathcal{M}$ and employ the $d$-dimensional representation to perform the test.

In two-sample testing, there are two groups of vertices with group $k$ of size $s_k \ll n$, and they are restricted to or around one specified latent position $X_k^*$, for $k = 1, 2$, i.e., $X_1, \ldots, X_{s_1} \sim h_{X_1^*}$ and $X_{s_1+1}, \ldots, X_{s_1+s_2} \sim h_{X_2^*}$. The two-sample test probes $H_0 : X_1^* = X_2^*$ vs. $H_A : X_1^* \neq X_2^*$. The restricted test assumes $X_k^* \in \mathcal{M}$, while the unrestricted test works with $X_k^* \in \mathcal{A}$ only. The learnt restricted test is defined as analogous to that in one-sample testing.

There are mainly two types of RDPGs in the network research community: those whose latent positions $X_i$’s are fixed ($h_{X^*} \propto \delta(X^*)$) and those whose latent positions are themselves random ($h_{X^*} \propto \delta(X^*)$). For a graph inference task with the former, the randomness of an actual adjacency matrix $\mathcal{A}$ has its source from the sampling
randomness conditioned upon the given latent positions. For a graph inference task with the latter, another resource of randomness of $A$ is from the randomness in the latent positions. For convenience, in the following, we consider two functions are equal if their ratio is constant, i.e., $h_{X^*} = h'_{X^*}$ if $h_{X^*} \propto h'_{X^*}$.

An alternative approach to ASE in learning the latent positions $\hat{X}_i$ is the Laplacian Spectral Embedding (LSE). Unlike ASE, LSE provides a transformation of $\hat{X}_i$, which results in the inference being different. An analogous CLT for estimated latent positions obtained by LSE is provided by Tang [27].

In this study, we will focus on the RDPGs that are parametric 1-dim LSMs. For LSMs of two or more dimensions, literature has not provided a clear definition. On the other hand, it is more straightforward to define restricted tests for parametric LSMs. In particular, we focus on 1-dim curves that can be written in the form of Bézier curves.

In the following sections, we will use a simple example, in which the Hardy-Weinberg curve is used as the embedded submanifold $\mathcal{M}$ to show some results. We focus on undirected RDPG models. In Section 4.3, we describe one-sample tests in 1-dimensional LSMs and define the three tests explicitly with our main theoretical results presented in its subsections. Section 4.4 discusses the setting and simulations of two-sample tests in 1-dimensional LSMs. As an extension to the main problem aforementioned, Section 4.5 investigates the general network structure discovery framework and uses simulations to demonstrate the trade-off in dimensionality selection.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

4.3 One-Sample Testing

We consider $\mathcal{H} \subset \mathbb{R}^3$, where $\mathcal{H}$ is the Hardy-Weinberg curve given by $\tau \in [0, 1]$ and $\psi : [0, 1] \rightarrow \mathcal{H}$ with $\psi(\tau) = [\tau^2, 2\tau(1-\tau), (1-\tau)^2]^T$. Let $\tau_1, \ldots, \tau_m \overset{iid}{\sim} U[0, 1]$ and $p_i = \psi(\tau_i)$. Here, the manifold $\mathcal{M} = \mathcal{H}$ and the ambient space $\mathcal{A} = \mathbb{R}^3$. Let the ss $p_i$ make up the rows of $X_m \in \mathbb{R}^{m \times 3}$. Suppose there is a group of $s$ latent positions, $X_1^*, \ldots, X_s^* \overset{iid}{\sim} h_{p^*}$ with $p^* = \psi(\tau^*)$ for some unknown $\tau^* \in [0, 1]$ and $h_{p^*}$ is a function indexed by the parameter $p^*$. For example, $h_{p^*} = \delta(p^*)$ where $\delta(x)$ is the Dirac delta function; or $h_{p^*} = \psi(\mathcal{N}(\tau; \tau^*, \sigma^2))$, where $\sigma^2$ is very small. Let $X_{m+i} = X_i^*$ for $i = 1, \ldots, s$ and $n = m + s$ denote the graph cardinality.

Now the $p_i$ are interpreted as the latent positions $X_i$ for an RDPG. We consider the one-sample testing question of whether $p^*$ comes from the same latent position as the null on $\mathcal{H}$,

$$H_0 : p^* = p_0$$

$$H_A : p^* \neq p_0,$$ (4.2)

where $p_0 = \psi(\tau_0)$ for some known $\tau_0 \in (0, 1)$. The goal is to compare the testing powers of the unrestricted test, the restricted test, and the learnt restricted test. We will first use $h_{p^*} = \delta(p^*)$ as an example to explain the testing procedure for the three tests.

The RDPG considered in one-sample testing has $n = m + s$ nodes. Under $H_0$,
the latent position matrix $X_0 = (X_1, \ldots, X_n)'$ is the $n \times 3$ matrix where the $i$th row $X_i = p_i$ for the first $m$ rows and $X_i = p_0$ for the last $s$ rows. Let an adjacency matrix $A$ be sampled according to the probability matrix $P^0 = X_0X_0^T$, i.e., $A_{ij} \sim \text{Bern}(P^0_{ij})$ for $i < j$, for we consider undirected graphs here. Given an adjacency matrix instance $A$, we apply ASE to $A$ and the resulting scree plot of the eigenvalues in descending order is shown in Figure 4.1 where $D = 3$ is chosen by the ‘elbow rule’ \[1\] and is indeed the same as that of the true latent position space. $ASE(A)$ provides the estimated latent position matrix $\hat{X}_0 \in \mathbb{R}^3$, whose rows are $\hat{X}_i$ for $i = 1, \ldots, n$.

By the CLT for RDPGs stated in Theorem 8, $\hat{X}_iW_0 | X_i = x_i$ is approximately Gaussian around $x_i$ in $\mathbb{R}^{\text{rank}(X_0X_0^T)=3}$, for $i = m + 1, \ldots, n$, for some rotation matrix $W_0$. Let $\hat{X}_0 = \hat{X}_0W_0$ and rows of $\hat{X}_0$ be denoted by $\hat{X}_i$. Figure 4.2 is a 2D visualization of $\{X_i\}_{i=1}^n$, $\{\hat{X}_i\}_{i=1}^n$ and $\{\hat{X}_i\}_{i=1}^n$. Figure 4.3 shows the collection $\{X_i\}_{i=1}^n$ and $\{\hat{X}_i\}_{i=1}^n$ in a 3D plot.
Figure 4.1: Model Selection of embedding dimension of $ASE(A)$, embedding dimension $D = 3$, i.e., the top 3 singular values and their associated singular vectors are chosen by singular value thresholding.

Figure 4.2: The green dots are $X_i$, forming $\mathcal{H}$; the red dots are $\hat{X}_i$, estimated latent positions before the proper rotation; and the black dots are $\hat{X}_i$ by rotating $\hat{X}_i$ with the rotation matrix $W_0$. 
Figure 4.3: $\tilde{X}_i$ are the red dots in $\mathbb{R}^3$ around the true latent positions $X_i \in \mathcal{H} \subset \Delta^2$ ($\mathcal{H}$ in black and simplex in purple).

As ASE is not able to detect $\mathcal{M}$, here the ASE space may be treated as $\mathcal{A}$, from which we hope to learn the embedded submanifold $\mathcal{M}$. We focus on Isomap throughout this work. We apply Isomap on $\{\tilde{X}_i\}_{i=1}^n$ to detect the embedded submanifold. Following the steps of Isomap, once we obtain the pairwise shortest path distance matrix, whose eigenvalues are shown in Figure 4.4 in descending order, the scree plot successfully detects the dimensionality of the embedded submanifold in the estimated ASE space $\tilde{X}$. We emphasize that the submanifold dimensionality detection in $\{\tilde{X}_i\}_{i=1}^n$ and that in $\{\hat{X}_i\}_{i=1}^n$ are equivalent. This is because the crucial step in Isomap, constructing the pairwise shortest path distance matrix, only depends on pairwise Euclidean distances in $\mathbb{R}^3$, which are invariant under orthogonal transformation.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Figure 4.4: The scree plot of the embedding dimension of $\tilde{X}$ via Isomap [2]. The large gap between the first two dimensions indicates that the optimal dimensionality of the submanifold is 1.

Now we define the three tests. We begin with defining the unrestricted test. Let $\bar{X} = \frac{1}{s} \sum_{i=m+1}^{m+s} \tilde{X}_i$ be the sample average in $\mathbb{R}^3$. Under $H_0$, for $i = m + 1, \ldots, m + s$, 
\[ \sqrt{n}(\tilde{X}_i - p_0)|X_0 \rightarrow \mathcal{N}(x; 0, \Sigma_0), \]
where $\Sigma_0$ is a covariance matrix defined in Theorem 8.

Then, by Corollary 3 in [30], which states that any fixed and finite collection of residuals $\tilde{X}_i - X_i$ is asymptotically independent, we have 
\[ \sqrt{n}(\bar{X} - p_0)|X_0 \rightarrow \mathcal{N}(x; 0, \Sigma_0/s). \]

For each $\bar{X}_i$, its projection on the curve may be defined as the point that is closest to $\bar{X}_i$ in Euclidean distance, denoted by $\psi(\hat{\tau}_i)$, where $\hat{\tau}_i = \arg\min_{r \in [0,1]} \|\psi(r) - \bar{X}_i\|_2$ for $i = m + 1, \ldots, m + s$. We say $\bar{X}_i$ is indexed by $\hat{\tau}_i$ on the curve or $\hat{\tau}_i$ may be viewed as the 1-dim coordinate on $\mathcal{H}$. 

96
In that $\hat{X}_i$ is not exactly on $H$, $\hat{\tau}_i$ as a 1-dim representation on $H$ is an estimation of the curve representation of $\hat{X}_i$ under the rule of minimum distance to the curve, and for this reason we call this representation the “Minimum Distance Estimator” (MDE). There is an alternative way to define the curve representation of $\hat{X}_i$. As $X_i = p_i = \psi(\tau_i)$, then $\tau_i$ and $X_i$ have the exact linear relationship as $\tau_i = L^T X_i$, with $L = [1, \frac{1}{2}, 0]^T$. This linear transformation is inspired by the MLE of the Hardy-Weinberg described in Section 2.3.1; therefore, we call this representation “Maximum Likelihood Estimator” (MLE). Note that such a linear transformation from $X_i$ to $\tau_i$ is unique unless we consider a shift $v \in \mathbb{R}$ s.t. $\tau_i = L^T X_i + v$. As a result, $\hat{\tau}_i = L^T \hat{X}_i$ is a well-defined manifold representation of $\hat{X}_i$. Let $\bar{\tau} = \frac{1}{s} \sum_{i=m+1}^{m+s} \hat{\tau}_i$ and $\bar{\tau} = \frac{1}{s} \sum_{i=m+1}^{m+s} \hat{\tau}_i$ be the average for the MDE and MLE representations, respectively.

When considering the hypothesis (4.2), we first define two types of tests: the unrestricted test as the distance between $p_0$ and $\bar{X}$ in $\mathbb{R}^3$, and the restricted test as the distance on $H$ between $\tau_0$ and $\bar{\tau}$ (or $\bar{\tau}$). An intuitive choice for the former distance is the Euclidean distance. For the latter distance, the two natural ones to consider are the Euclidean distance in $\mathbb{R}$ and the geodesic distance on $H$. Taking $\bar{\tau}$ for instance, the Euclidean distance in $\mathbb{R}$ refers to $|\bar{\tau} - \tau_0|$, and the geodesic distance on $H$ is equivalent to the curve distance $|\int_{\bar{\tau}}^{\tau_0} \|\psi'(u)\|du|$. For the restricted test, we consider both.

Let $T_{\mathbb{R}^3}$ denote the unrestricted test statistics. Under consideration there are four restricted tests on $H$, each of which is defined with one of two representations, $\bar{\tau}$.
(MLE) or $\tilde{\tau}$ (MDE), and one of two distances, Euclidean or geodesic, such that each test is defined with a unique combination of one representation and one distance. We define the unrestricted test $T_{R^3}(A)$ and the four types of restricted tests, $T_{H,MDE}(A)$, $T_{c,H,MDE}(A)$, and $T_{c,H,MLE}(A)$ as

$$T_{R^3}(A) = \|\tilde{X} - p_0\|_2$$  \hspace{1cm} (4.3)

$$T_{H,MDE}(A) = |\tilde{\tau} - \tau_0|$$  \hspace{1cm} (4.4)

$$T_{c,H,MDE}(A) = |\int_{\tilde{\tau}}^{\tau_0} \|\psi'(u)\|du|$$  \hspace{1cm} (4.5)

$$T_{H,MLE}(A) = |\tilde{\tau} - \tau_0|$$  \hspace{1cm} (4.6)

$$T_{c,H,MLE}(A) = |\int_{\tilde{\tau}}^{\tau_0} \|\psi'(u)\|du|$$  \hspace{1cm} (4.7)

where $A$ is a sampled adjacency matrix instance under the null hypothesis or the alternative.

For the learnt restricted test, we propose to learn the manifold mapping inverse $\psi^{-1}$ based on $\{\tilde{X}_i\}_{i=1}^n$ and $p_0$. Let $\psi_{iso}^{-1}: \tilde{X}_i \rightarrow \hat{Z}_i$ be the mapping learnt from dimension $D = 3$ to dimension $d = 1$. In Hardy-Weinberg models, we have the estimated 1-dim representation of $\tilde{X}_i$ as $\hat{Z}_i = \psi_{iso}^{-1}(\tilde{X}_i)$, for $i = 1, \ldots, n$. Let $\hat{Z} = \frac{1}{s} \sum_{i=m+1}^{m+s} \hat{Z}_i$ be the sample average. Let $\hat{Z}_0 = \psi_{iso}^{-1}(p_0)$ be the learnt 1-dim representation of $p_0$. We emphasize that $\hat{Z}_0$ may also be learnt through out-of-sample embedding (Section 2.4 for more details).

Let $T_{iso}$ denote the learnt restricted test statistics. $T_{iso}$ is defined to be the Eu-
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

clidean distance between the sample mean and \( p_0 \) in the learnt 1-dim coordinate representation, and it may be viewed as the distance on the learnt manifold,

\[
T_{iso}(A) = |\tilde{Z} - \hat{Z}_0|.
\] (4.8)

When fixing the size of tests to be \( \alpha \), the critical value of the unrestricted test, denoted by \( C_{R^3} \), is defined as \( C_{R^3} = \{ C : \alpha = \mathbb{P}_0[T_{R^3}(A) \geq C] \} \). We define \( C_{H,MDE}, C^c_{H,MDE}, C_{H,MLE}, C^c_{H,MLE} \), and \( C_{iso} \) accordingly.

To calculate the powers under alternative \( p_A = \psi(\tau_A) \), let \( X_A = (X_1, \ldots, X_n)' \) be the \( n \times 3 \) matrix where the \( i \)th row \( X_i = p_i \) for the first \( m \) rows and \( X_i = p_A \) for the last \( s \) rows. Note that \( X_0 \) and \( X_A \) differ merely on the last \( s \) rows. Let an adjacency matrix \( A \) be sampled according to the probability matrix \( P^A = X_A X_A' \).

The estimated latent position matrix \( \hat{X}_A \) is obtained from \( A \) via ASE. Applying the rotation matrix \( W_A \) to \( \hat{X}_A \) provides the correctly aligned estimated latent positions for the purpose of the CLT in Theorem 8. Let \( \hat{X}_A = \hat{X}_A W_A \) and denote rows of \( \hat{X}_A \) as \( \hat{X}_i \).

Once we have \( \hat{X}_i \), the power of the unrestricted test is defined as \( \pi_{R^3}(p_A; p_0, \alpha, m, s) = \mathbb{P}_{\psi(\tau_A)}[T_{R^3}(A) \geq C_{R^3}] \). We define \( \pi_{H,MDE}(p_A; p_0, \alpha, m, s), \pi^c_{H,MDE}(p_A; p_0, \alpha, m, s), \pi_{H,MLE}(p_A; p_0, \alpha, m, s), \pi^c_{H,MLE}(p_A; p_0, \alpha, m, s) \) and \( \pi_{iso}(p_A; p_0, \alpha, m, s) \) accordingly.

We construct an experiment using Monte-Carlo (MC) \([14]\) to explore relationships among the powers of the unrestricted test, the restricted tests and the learnt restricted
Example 1. Let $\alpha = 0.05, m = 1000, s = 5, \tau_0 = 0.3, \tau_A = 0.35$, and $K = 100$ (the number of nearest neighbors chosen in Isomap). Define $h_{p^*} = \delta(p^*)$.

In the first place, we approximate critical values via MC simulation with $B_0$ iterations. In iteration $b$, where $b = 1, \ldots, B_0$, we sample a graph $A^b$ from the probability matrix $P^b = X_0X_0^T$, and $\hat{X}_0^b = ASE(A^b)$ is the estimated latent position matrix. After rotating $\hat{X}_0^b$ to $X_0^b$, we calculate the unrestricted test statistics $T_{\mathbb{R}^3}(A^b)$ in $\mathbb{R}^3$ and the four restricted test statistics $T_{H,MDE}(A^b)$, $T_{c,H,MDE}(A^b)$, $T_{H,MLE}(A^b)$ and $T_{c,H,MLE}(A^b)$ on $H$. We then apply Isomap to the rows of $X_0^b$ and $p_0$, and we obtain learnt manifold representations $\hat{Z}_i^b$ for $i = 0, \ldots, n$ and the learnt restricted test statistics $T_{iso}(A^b)$.

The ordered collection of $\{T_{\mathbb{R}^3}(A^b)\}_{b=1}^{B_0}$ is denoted by $\{T_{\mathbb{R}^3}^{(i)}\}_{i=1}^{B_0}$ from the smallest to the largest. Then the approximated critical value $\hat{C}_{\mathbb{R}^3} \approx T_{\mathbb{R}^3}^{(1-\alpha)(B_0+1)}$, which satisfies $P_0[T_{\mathbb{R}^3}(A^b) > \hat{C}_{\mathbb{R}^3}] = \alpha$. Similarly, we have $\hat{C}_{H,MDE}$, $\hat{C}_{c,H,MDE}$, $\hat{C}_{H,MLE}$, $\hat{C}_{c,H,MLE}$, and $\hat{C}_{iso}$, each corresponding to its ordered test statistics series, respectively.

The powers are approximated with $\hat{C}_{\mathbb{R}^3}, \hat{C}_{H,MDE}, \hat{C}_{c,H,MDE}, \hat{C}_{H,MLE}, \hat{C}_{c,H,MLE}$, and $\hat{C}_{iso}$ via MC simulation with $B_A = 1000$ iterations at $p_A$. In iteration $b$, where $b = 1, \ldots, B_A$, we sample a graph $A^b$ from the probability matrix $P^A = X_AX_A^T$, and $\hat{X}_A^b = ASE(A^b)$ denotes the estimated latent position matrix. Thus, with the new collection $\{T_{\mathbb{R}^3}(A^b)\}_{b=1}^{B_A}$, we estimate the power by $\hat{P}_{\mathbb{R}^3}(p_A; p_0, \alpha, m, s) = P_{\psi(\tau_A)}[T_{\mathbb{R}^3}(A^b) \geq \hat{C}_{\mathbb{R}^3}]$. Similarly, we obtain $\hat{P}_{H,MDE}(p_A; p_0, \alpha, m, s)$, $\hat{P}_{c,H,MDE}(p_A; p_0, \alpha, m, s)$, $\hat{P}_{H,MLE}(p_A; p_0, \alpha, m, s)$,
\[
\hat{\pi}^c_{\mathcal{H},MLE}(p_A; p_0, \alpha, m, s), \quad \text{and} \quad \hat{\pi}_{iso}(p_A; p_0, \alpha, m, s, K), \quad \text{each corresponding to its test statistics series, respectively. The estimated powers are,}
\]

\[
\hat{\pi}^c_{\mathbb{R}^2}(p_A; p_0, \alpha, m, s) = 0.771 \\
\hat{\pi}^c_{\mathcal{H},MDE}(p_A; p_0, \alpha, m, s) = 0.807 \\
\hat{\pi}^c_{\mathcal{H},MDE}(p_A; p_0, \alpha, m, s) = 0.797 \\
\hat{\pi}_{\mathcal{H},MLE}(p_A; p_0, \alpha, m, s) = 0.966 \\
\hat{\pi}^c_{\mathcal{H},MLE}(p_A; p_0, \alpha, m, s) = 0.964 \\
\hat{\pi}_{iso}(p_A; p_0, \alpha, m, s, K) = 0.984.
\]

Now we consider using McNemar’s test to check whether the difference between each pair of estimated powers is significant with \( H_0 : \pi_k = \pi_\ell \), and with \( k, \ell \) each referring to one of the six tests we have been considering. For each McNemar test, we set \( \alpha = 0.05 \).

Results indicate that the power of the unrestricted test is significantly different from the others. For the restricted tests, the power of Euclidean distance tests is not different from that of geodesic distance tests under the same MLE, or MDE, representation; however, the two tests with different representations are significantly different from each other under the same distance. Another interesting observation is that \( \pi^c_{\mathcal{H},MLE} \) and \( \pi_{iso} \) are not significantly different.

We expect that the power relations discovered in the above example stand in gen-
eral with \( s \) finite and \( m \to \infty \). We consider \( \pi_{\mathcal{H},\text{MLE}}(p_A; p_0, \alpha, m, s) \approx \pi_{\mathcal{H},\text{MLE}}^c(p_A; p_0, \alpha, m, s) \approx \pi_{\text{iso}}(p_A; p_0, \alpha, m, s) \) in Section 4.3.3 and Section 4.3.1.

In what follows, unless specified otherwise, we abbreviate \( T_{\mathcal{H},\text{MLE}} \) as \( T_{\mathcal{H}} \) and \( T_{\mathcal{H},\text{MLE}}^c \) as \( T_{\mathcal{H}}^c \) and their powers, for we will mainly discuss the MLE representation for restricted tests. Figure 4.5 shows a map of the tests that will be discussed in the following subsections.

\[
\begin{align*}
\pi_{\text{iso}} & \quad \text{Section 4.3.1} \\
\pi_{\mathcal{H}}^c & \quad \text{Section 4.3.3} \\
\pi_{\mathcal{R}^3} & \quad \text{Section 4.3.2} \\
\pi_{\mathcal{H}} & \quad \text{Section 4.3.2}
\end{align*}
\]

Figure 4.5: Map of the four tests that will be discussed in the following subsections.

In Sections 4.3.2 and 4.3.3 we study \( h_{p^*} = \delta(p^*) \). In Section 4.3.1 we assume \( h_{p^*} = \psi(\mathcal{N}(\tau; \tau^*, \sigma^2)) \).

### 4.3.1 Power Comparison: \( \pi_{\mathcal{H}}^c \) vs. \( \pi_{\mathcal{H}} \)

Assume \( h_{p^*} = \psi(\mathcal{N}(\tau; \tau^*, \sigma^2)) \), with \( \sigma^2 \) being very small. To study the conjecture that the power of the learnt restricted test is asymptotically equal to that of the restricted test, it is essential to show that \( T_{\text{iso}} \to T_{\mathcal{H}}^c \) as \( m \) goes to infinity, where \( T_{\text{iso}} \) and \( T_{\mathcal{H}}^c \) are defined in Equations (4.8) and (4.7), respectively.

We start by reviewing Isomap and its major assumptions. For a set of \( D- \)
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INference

dimensional data that are sampled from a quasi-isometric, geodesically convex and $d$-dimensional manifold $\mathcal{M}$, Isomap \cite{15} learns a low-dimensional representation that recovers pairwise geodesic distances on $\mathcal{M}$. More specifically, Isomap takes as input all pairwise distances in the $D$-dimensional space, measured in some metric $d_X$, and outputs the $d$-dimensional representation $\{Z_i\}_{i=1}^n$ with $D > d$. Given data $\{X_i\}_{i=1}^n$, the steps of Isomap are:

1. Construct a graph $G = (V, E)$, where $V = \{X_i\}_{i=1}^n$ and $e(X_i, X_j) \in E$ if $d_X(X_i, X_j) \leq \epsilon$ ($\epsilon$-rule) or if point $i$ is one of the $K$ nearest neighbors of point $j$ ($K$-NN rule). The edge weight equals $d_X(X_i, X_j)$.

2. Compute the shortest path distance matrix $[d_G]$, where the entry $ij$ is $d_G(X_i, X_j)$, the shortest graph distance from point $i$ to $j$.

3. Apply Classical Multidimensional Scaling (CMDS) to the dissimilar matrix $[d_G]$ and obtain the $d$-dimensional representation $\{Z_i\}_{i=1}^n$.

$[d_G]$ as an approximation to the geodesic distance matrix $[d_M]$, is taken by Isomap as the dissimilar matrix into CMDS to obtain a $d$-dimensional embedding. Isomap thus may be viewed as a nonlinear generalization of CMDS.

We now recall some basics of Isomap assumptions, namely, quasi-isometric and geodesically convex manifolds.

**Definition 4** (Quasi-Isometric). Let $\mathcal{M}$ be a manifold and assume there exists a mapping $\rho : U \rightarrow \mathcal{M}$, where $U \subset \mathbb{R}^d$, $\mathcal{M} \subset \mathbb{R}^D$. We say that $\mathcal{M}$ is quasi-isometric if

103
there exist constants $C, K$ such that

$$\forall x, y \in U : C^{-1}d(x, y) - K \leq d_M(\rho(x), \rho(y)) \leq Cd(x, y) + K$$

, where $d(\cdot, \cdot)$ is the Euclidean metric and $d_M(\cdot, \cdot)$ is the manifold geodesic metric.

**Definition 5** (Geodesically Convex). A manifold $\mathcal{M}$ is said to be geodesically convex if any two points $x, y \in \mathcal{M}$ are connected by a geodesic of length $d_M(x, y)$.

**Example 2.** The Hardy-Weinberg curve $\mathcal{H}$ is a quasi-isometric manifold because

$$\forall x, y \in (0, 1) : 2^{-1}d(x, y) \leq d_H(\psi(x), \psi(y)) \leq 2d(x, y).$$

$\mathcal{H}$ is obviously geodesically convex.

Suppose the probability matrix of an RDPG is $P = \rho_n XX^T$, where the sparsity factor $0 < \rho_n \leq 1$ is known. Given latent positions $\{X_i\}_{i=1}^n$ and their properly rotated estimates $\{\tilde{X}_i\}_{i=1}^n$, where $X_i \in \mathcal{M} \subset \mathbb{R}^D$ and $\tilde{X}_i \in \mathbb{R}^D$, the approximated equality between $T_{iso}$ and $T_{\mathcal{H}}^c$ is established based on a series of necessary intermediate steps:

I. Consider an Isomap learning on $\{X_i\}_{i=1}^n \cup \{p_0\}$. Let $G$ denote the corresponding shortest path distance graph in Isomap, which is generated based on pairwise Euclidean distances between $X_p$ and $X_q$ for $p \neq q$ and $p, q = 1, \ldots, n$. An $\epsilon_n$-ball (or $K$-NN) rule is chosen such that $G$ contains all edges $e(X_p, X_q)$ of length $\|X_p - X_q\| \leq \epsilon_n = \frac{1}{r(n)}$, where $r(n) = o(\sqrt{n\rho_n}/\log^2 n)$. For simplicity,
we assume \( r(n) = n^a \sqrt{\rho_n} / \log^2 n \) for some \( a \in (0, 1/2) \). Note that if \( a = 0 \) and \( \rho_n = 1 \), then Isomap learning will make all the points neighbors, in which case Isomap is equivalent to PCA/CMDS.

If the assumptions of Theorem 11 are met, then \( d_M(X_p, X_q) \approx d_G(X_p, X_q) \), for any \( X_p, X_q \in \mathcal{M} \), which will be restated in Corollary 1.

II. Consider an Isomap learning on \( \{\tilde{X}_i\}_{i=1}^n \cup \{p_0\} \). Let \( \hat{G} \) denote the corresponding shortest path distance graph in Isomap, which is generated based on pairwise Euclidean distances between \( \tilde{X}_p \) and \( \tilde{X}_q \) for \( p \neq q \) and \( p, q = 1, \ldots, n \). An appropriate \( \hat{e}_n \) (or \( \hat{K}\text{-NN} \)) rule shall be defined and a corresponding graph \( \hat{G} \) is generated such that if \( X_p \) and \( X_q \) are neighbors in \( G \), then \( \tilde{X}_p \) and \( \tilde{X}_q \) are neighbors in \( \hat{G} \), while \( \hat{G} \) contains all edges \( e(\tilde{X}_p, \tilde{X}_q) \) of length \( ||\tilde{X}_p - \tilde{X}_q|| \leq \hat{e}_n \) as well.

A natural selection of \( \hat{e}_n \) is \( e_n \). By the uniform control of deviations between the estimated and true latent positions of RDPGs [29], with high probability we have

\[
\sup_{1 \leq k \leq n} ||\tilde{X}_k / \sqrt{\rho_n} - X_k|| \leq h(n),
\]  \hspace{1cm} (4.9)

where \( h(n) = O(\frac{\log^2 n}{\sqrt{n \rho_n}}) \).

III. Lemma 12 proves \( d_G(\tilde{X}_p, \tilde{X}_q) \to d_G(X_p, X_q) \) as \( m \) tends toward infinity. Thus, with \( d_M(X_p, X_q) \approx d_G(X_p, X_q) \), Lemma 2 ties \( d_G(\tilde{X}_p, \tilde{X}_q) \) and \( d_M(X_p, X_q) \) to-
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

gether.

IV. Both $T^c_H$ and $T_{iso}$ are estimates of $d_M(p_0, X^*)$. While $T^c_H = d_M(p_0, \psi(\tilde{\tau}))$ and $d_M(p_0, X^*)$ are both exact geodesic distances on $\mathcal{H}$, $T^c_H$ approximates $d_M(p_0, X^*)$ via the fact that $\psi(\tilde{\tau}) \approx \tilde{X} \approx X^*$. On the other hand, $T_{iso} = d_M(p_0, \tilde{X})$ is an approximate to $d_M(p_0, X^*)$ since $d_M(p_0, \tilde{X}) \approx d_G(p_0, \tilde{X}) \approx d_G(p_0, X^*)$.

When we investigate the aforementioned approximate equality between the restricted test $T^c_H$ and the learnt restricted test $T_{iso}$ via Isomap, the above steps outline the proof of the following lemma:

**Lemma 7.** When $\tau_0, \tau_A, \alpha$ and $s$ are fixed and finite, under both null and alternative hypotheses $T_{iso} \xrightarrow{d} T^c_H$ as $m \to \infty$.

The details of the proof is stated in Section 4.6.

The conclusion is that since $T_H^c \xrightarrow{d} T^c_H$ as $m \to \infty$, their corresponding powers are equivalent in the limit.

**Theorem 9.** Suppose $X_1, \ldots, X_m \overset{iid}{\sim} \text{Uniform}(\mathcal{H})$ and $X_{m+1}, \ldots, X_{m+s} \overset{iid}{\sim} \mathcal{N}(\tau; \tau^*, \sigma^2)$, where $p^* = \psi(\tau^*) \in \mathcal{H}$ and $\sigma^2$ is small. Let $n = m + s$ and $X = [X_1, \ldots, X_n]^T$. In the RDPG setting, with $P = \rho_n XX^T$ and the sparsity factor $0 < \rho_n \leq 1$ known, consider the hypothesis testing problem, $H_0 : p^* = p_0$ vs. $H_0 : p^* \neq p_0$, observing just $A \sim P$. At significance level $\alpha$, with $s$ fixed and finite, at alternative $p_A \in \mathcal{H}$, $\lim_{m \to \infty} \pi_{iso} - \pi_H^c = 0$. 

106
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

There are two possible generalizations to this theorem. First, we observe that the function $h_{p^*}$ may technically be in any form that allows $X_{m+1}, \ldots, X_{m+s}$, not all from one point but rather around some point on $\mathcal{H}$. This condition is to ensure $\lim_{m \to \infty} \pi_{iso} < 1$ and $\lim_{m \to \infty} \pi_{h}^c < 1$ when $s$ is finite. For a network of size $n$, the goal is to test whether or not the latent position vectors of the group $X_{m+1}, \ldots, X_{m+s}$ are as specified in the null hypothesis. While this group of nodes is fixed, if we may enlarge the whole network by introducing more auxiliary nodes, then Theorem 9 suggests that performing the test after the manifold learning step will yield the same power asymptotically as conducting the test while knowing the embedded submanifold in the ASE space.

In addition, if one wishes to generalize $\mathcal{H}$ to Bézier curves, then instead of $\pi_{h,MLE}^c$, associated with the MLE representation, one may consider $\pi_{h,MDE}^c$ since it is clear that $\max_{m+1 \leq i \leq m+s} |\hat{\tau}_i - \tau_i| = O(h(n))$. Thus, the above theorem is expect to hold for both $\pi_{h}^c = \pi_{h,MLE}^c$ and $\pi_{h}^c = \pi_{h,MDE}^c$. This observation equips us to generalize $\mathcal{H}$ to any Bézier curve $B(t)$ that is quasi-isometric, concluding $\lim_{m \to \infty} \pi_{iso}^c - \pi_{h,MDE}^c = 0$.

4.3.2 Power Analysis: $\pi_{\mathbb{R}^3}$ vs. $\pi_{\mathcal{H}}$

In this section we consider $h(p^*) = \delta(p^*)$. We compute the powers $\pi_{\mathbb{R}^3}$ and $\pi_{\mathcal{H}}$ by looking at the theory in large samples, or more precisely, as $m \to \infty$ while $s$ is finite. The idea is that, at least for models estimated using large finite samples, these asymptotic results will provide a useful approximation.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

In Theorem 8 for $i = m + 1, \ldots, m + s$, the CLT says that $\sqrt{n}(\bar{X}_i - X_i)|X_i = p^* \to \mathcal{N}(x; 0, \Sigma(p^*))$, where $\Sigma(x) = \Delta^{-1}E[X_jX_j^T(x^T X_j - (x^T X_j)^2)]\Delta^{-1}$ with $\Delta = E[X_n X_n^T]$.

The power $\pi_{\mathbb{R}^3}(p_A; p_0, \alpha, m, s)$ can be well approximated by considering $\sqrt{n}(\bar{X}_i - X_i)|X_i = p^*$ as a Gaussian distribution under both the null and the alternative hypotheses when $m$ is large enough. We compute the power $\pi_{\mathbb{R}^3}(p_A; p_0, \alpha, m, s)$ by considering that under $H_0$, $\bar{X} \to \mathcal{N}(x; p_0, \Sigma(p_0)_{ns})$ and at the alternative $p_A = \psi(\tau_A)$, $\bar{X} \to \mathcal{N}(x; p_A, \Sigma(p_A)_{ns})$. Thus, the test statistic $T_{\mathbb{R}^3}(A)$ defined in Equation (4.3) is the $L_2$ norm of a Gaussian random variable, which is equivalent to a linear combination of independent noncentral chi-square variables. Observation provides the necessary tool to compute $\pi_{\mathbb{R}^3}(p_A; p_0, \alpha, m, s)$.

From Observation under $H_0$ we have

$$T_{\mathbb{R}^3}^2(A)|_{H_0} = \|\bar{X} - p_0\|_2^2|_{H_0} \to \sum_{i=1}^{D} \frac{\lambda_i}{n s} \chi_1^2,$$

where $\lambda_i$ are the eigenvalues of $\Sigma(p_0)$ in descending order. Define $f_0$ and $F_0$ to be the density function and the distribution function of $\sum_{i=1}^{D} \lambda_i \chi_1^2$ respectively. Let $Q_F(p)$ denote the quantile function of a CDF $F$ at probability $p$. Then $C_{\mathbb{R}^3} = \{C : \alpha =$
\[ \mathbb{P}_0[\text{T}_{R^3}(A) \geq C] \] is of order \( \Theta(\frac{1}{m^2}) \),

\[
\alpha = \mathbb{P}_0[\sum_{i=1}^{D} \frac{\lambda_i}{n^s} \chi_1^2 > C_{R^3}]
\]

\[
= \mathbb{P}_0[\sum_{i=1}^{D} \lambda_i^2 > n^s C_{R^3}]
\]

\[ \implies n^s C_{R^3} = Q_{F_0}(1 - \alpha). \]

Here \( \lambda_i, Q_{F_0}(1 - \alpha) \) are of \( O(1) \) with respect to \( n \) and \( s \).

Under the alternative \( p_A \), we have

\[
\text{T}'_{R^3}(A)|_{H_A} = \|X - p_0\|_2^2 |_{H_A} \rightarrow \sum_{i=1}^{D} \frac{\lambda_i'}{n^s} \chi_1^2(b_i^2),
\]

where \( \lambda_i' \) are the eigenvalues of \( \Sigma(p_A) \) in descending order. We define \( f_A \) and \( F_A \) to be the density function and the distribution function of \( \sum_{i=1}^{D} \lambda_i' \chi_1^2(b_i^2) \), respectively.

Then \( \pi_{R^3}(p_A; p_0, \alpha, m, s) \) is

\[
\pi_{R^3}(p_A; p_0, \alpha, m, s) = \mathbb{P}_A[\sum_{i=1}^{D} \frac{\lambda_i'}{n^s} \chi_1^2(b_i^2) > C_{R^3}]
\]

\[
= \mathbb{P}_A[\sum_{i=1}^{D} \lambda_i' \chi_1^2(b_i^2) > Q_{F_0}(1 - \alpha)] = 1 - F_A(Q_{F_0}(1 - \alpha)),
\]

where \( b = (b_1, b_2, b_3) = \sqrt{n^s} \text{diag}(\frac{1}{\sqrt{\lambda_1'}}, \frac{1}{\sqrt{\lambda_2'}}, \frac{1}{\sqrt{\lambda_3'}}) \text{V}(p_A - p_0) \), and \( \Sigma(p_A) = \text{V}^T \text{diag}(\lambda_1', \lambda_2', \lambda_3') \text{V}. \)

If \( \tau_A - \tau_0 = \frac{h}{\sqrt{n}} \), with \( h = \Theta(1) \), then \( p_A - p_0 = \frac{h}{\sqrt{n}}[(2\tau_0 + \frac{h}{\sqrt{n}}), 2(1 - 2\tau_0 - \frac{h}{\sqrt{n}}), -(2 - 2\tau_0 - \frac{h}{\sqrt{n}})]' \). When \( \tau_A \) is in the neighbor of \( \frac{h}{\sqrt{n}} \) of \( \tau_0 \), the noncentral parameter vector
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

\[ b = \sqrt{\hat{a}} \text{diag}(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \frac{1}{\sqrt{\lambda_3}}) V[(2\tau_0 + \frac{h}{\sqrt{n}}), 2(1 - 2\tau_0 - \frac{h}{\sqrt{n}}), -(2 - 2\tau_0 - \frac{h}{\sqrt{n}})]' \] does not decay when \( n \) grows.

Since there is no closed form of the quantile or CDF of linear combinations of (non)central chi-square distributions, we will approximate \( C_{\mathbb{R}^3} \) and \( \pi_{\mathbb{R}^3} \) to the desired accuracy via the R package “CompQuadForm”, which contains various state-of-the-art approximation methods to calculate the CDF of linear combinations of (non)central chi-square distributions, including [40], [41] and [42]. For other approximation algorithms not included in the package, see [43] and [44].

Example 3. Continue Example 1. The covariance matrix \( \Sigma(p_0) \) of \( \tilde{X} \) under \( H_0 \) is

\[
\Sigma(p_0) = \begin{bmatrix}
1.21 & -1.47 & 0.531 \\
-1.47 & 4.31 & -1.96 \\
0.531 & -1.96 & 2.15
\end{bmatrix}.
\]

We calculate the critical value \( C_{\mathbb{R}^3} \) satisfying \( P_{\pi_0}[T_{\mathbb{R}^3} = \|Y\|^2 \geq C_{\mathbb{R}^3}] = \alpha \), where \( Y|_{H_0} \sim \mathcal{N}(x; 0, \Sigma_{p_0}^{(n)}) \), by Observation 1 with \( U_i \overset{iid}{\sim} \mathcal{N}(0, 1) \),

\[
T_{\mathbb{R}^3}|_{H_0} = \|Y\|^2 = 0.001186U_1^2 + 0.000231U_2^2 + 0.000109U_3^2.
\]

Thus, \( C_{\mathbb{R}^3} \approx 0.00494 \). Under the alternative \( p_A = \psi(0.35) \), we have \( \pi_{\mathbb{R}^3}(p_A; p_0, \alpha, m, s) = P_{\pi_A}[T_{\mathbb{R}^3} = \|Y\|^2 \geq C_{\mathbb{R}^3}] \), with \( Y|_{H_A} \sim \mathcal{N}(x; p_A - p_0, \Sigma_{p_A}^{(n)}) \), so \( b = \mathbf{V} \mathbf{V}^T \mathbf{A}^{-1/2} \mathbf{V}(p_A - \)
\[ p_0 = \Lambda^{-1/2}V(p_A - p_0) = [-1.40, -4.18, 1.51]^T, \] which gives

\[ T_{R^3|H_A} = \|Y\|^2 = 0.001221(U_1 - 1.40)^2 + 0.000239(U_2 - 4.18)^2 + 0.000119(U_3 + 1.51)^2. \]

In the package “CompQuadForm”, we use the “davis” function to obtain

\[ \pi_{R^3}(p_A; p_0, 0.05, 1000, 5) = 0.791. \]

Now we consider \( \pi_H \). Since \( \tilde{r}_i \) is a linear transformation of \( \tilde{X}_i \) and \( \tilde{X}_i \) converges to a Gaussian distribution, then \( \tilde{r}_i \rightarrow \mathcal{N}(x; \mu_{\tau_i}, \sigma_{\tau_i}^2) \), where \( \mu_{\tau_i} = [1, \frac{1}{2}, 0]^T X_i = \tau_i \) and \( \sigma_{\tau_i}^2 = \frac{1}{n_s} [1, \frac{1}{2}, 0]^T \Sigma(X_i) [1, \frac{1}{2}, 0] \) for \( i = 0, A \). Thus, under \( H_0 \), \( \tilde{r} \rightarrow \mathcal{N}(x; \mu_{\tau_0}, \frac{\sigma_{\tau_0}^2}{s}) \) and at \( \tau_A \), \( \tilde{r} \rightarrow \mathcal{N}(x; \mu_{\tau_A}, \frac{\sigma_{\tau_A}^2}{s}) \), we may calculate \( \pi_H \) exactly.

Suppose \( \tau_A = \tau_0 + \frac{h}{\sqrt{n}} \), with \( \mathcal{FN} \) representing the Folded Normal distribution. Under \( H_0 \), the test statistic is

\[ T_H(A)|_{H_0} = |\tilde{r} - \tau_0|_{H_0} \sim \mathcal{FN}(x; 0, \frac{c_0}{n_s}), \] (4.10)

where \( c_0 = [1, \frac{1}{2}, 0]^T \Sigma(p_0) [1, \frac{1}{2}, 0] \). The critical value \( C_H \) is then of order \( \Theta\left(\frac{1}{\sqrt{n_s}}\right)\),

\[ \alpha = \mathbb{P}_0[T_H(A) > C_H] \rightarrow C_H = \sqrt{\frac{c_0}{n_s}} \sqrt{2} \text{erf}^{-1}(1 - \alpha), \]

where \( \text{erf}(x) = 2\Phi(x\sqrt{2}) - 1 \) is the “error function” encountered in integrating the standard Gaussian distribution.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Under the alternative $A$, the test statistic is

$$T_H(A)|_{H_A} = |\bar{r} - \tau_0|_{H_A} \sim \mathcal{F}\mathcal{N}(x; \tau_A - \tau_0, \frac{c_A}{nS})$$ \hspace{1cm} (4.11)$$

where $c_A = [1, \frac{1}{2}, 0]^T \Sigma(p_A)[1, \frac{1}{2}, 0]$. When $\tau_A - \tau_0 = \frac{h}{\sqrt{n}}$, with $h = O(1)$, the power $\pi_H(p_A; p_0, \alpha, m, s)$ may be written as

$$\pi_H(p_A; p_0, \alpha, m, s) = \mathbb{P}_A[T_H > C_H] = \mathbb{P}_A[T_H > \sqrt{\frac{c_0}{nS}} \sqrt{2} \text{erf}^{-1}(1 - \alpha)]$$

$$= 1 - \frac{1}{2} \left[ \text{erf}\left(\frac{\frac{1}{\sqrt{nS}} \sqrt{2} \text{erf}^{-1}(1 - \alpha) + \frac{h}{\sqrt{n}}}{\sqrt{\frac{c_A}{nS}}}\right) + \text{erf}\left(\frac{\frac{1}{\sqrt{nS}} \sqrt{2} \text{erf}^{-1}(1 - \alpha) - \frac{h}{\sqrt{n}}}{\sqrt{\frac{c_A}{nS}}}\right) \right]$$

$$= 1 - \left[ \Phi\left(\frac{\sqrt{c_0} \sqrt{2} \text{erf}^{-1}(1 - \alpha) + h}{\sqrt{\frac{c_A}{s}}}\right) - \Phi\left(\frac{h - \sqrt{c_0} \sqrt{2} \text{erf}^{-1}(1 - \alpha)}{\sqrt{\frac{c_A}{s}}}\right) \right].$$

**Example 4.** Under the settings in Example 1, the power of the restricted test calculated by “CompQuadForm” is $\pi_{H,MLE}(p_A; p_0, 0.05, 1000, 5) = 0.987$. Now we fix $\alpha = 0.05, n = 1000, s = 5$ and vary the alternative $\tau_A \in (0, 1)$ to generate the power curves of $T_{\mathbb{R}^3}$ and $T_H$.

Figure 4.6a implies that one should be concerned only when $\tau_A$ is in the neighborhood of $\tau_0$ because both $\pi_{\mathbb{R}^3}$ and $\pi_H$ converge to 1 as $\tau_A$ moves away from $\tau_0$. As both powers converge to 1 as $\tau_A - \tau_0$ becomes larger, $\pi_H$ converges at a faster rate than $\pi_{\mathbb{R}^3}$ does.
(a) Power curves of $\pi_{R^3}$ and $\pi_H$ at null $\tau_0 = 0.3$ with $\alpha = 0.05$.

(b) The power difference curve of $\pi_H - \pi_{R^3}$ at null $\tau_0 = 0.3$ with $\alpha = 0.05$.

Figure 4.6: The power comparison between $\pi_{R^3}$ and $\pi_H$.

(a) ROC curves of $\pi_{R^3}$ and $\pi_H$ at null $\tau_0 = 0.3$ and alternative $\tau_A = 0.35$ varying different $\alpha$.

(b) The ROC difference curve of $\pi_H - \pi_{R^3}$ at null $\tau_0 = 0.3$ and alternative $\tau_A = 0.35$ varying different $\alpha$.

Figure 4.7: ROC curves comparison between $\pi_{R^3}$ and $\pi_H$. 

113
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

4.3.3 Power Comparison: $\pi_\mathcal{H}$ vs. $\pi^c_\mathcal{H}$

For a Bézier curve \( \{B(t), t \in [0, 1]\} \), the geodesic distance between any two points \( x = B(t_1) \) and \( y = B(t_2) \) on the curve equals the curve distance between \( x \) and \( y \), i.e., \(| \int_{t_1}^{t_2} \|B'(u)\|du |\). Since the Hardy-Weinberg curve is a 3-D Bézier curve and the norm of its derivative is \( \|\psi'(t)\| = \sqrt{24}\sqrt{t^2 - \frac{1}{3}} \), the test statistic is \( T^c_\mathcal{H} = \left| \int_{\tau}^{\tau_0} \|\psi'(u)\|du \right| \).

Without loss of generality, we assume \( \tau_0 \leq 0.5 \).

**Lemma 8.** Consider the powers of the restricted tests defined in Equations (4.6) and (4.7) at the alternative \( p_A = \psi(\tau_A) \), then \( |\pi^c_\mathcal{H}(p_A; p_0, \alpha, m, s) - \pi_\mathcal{H}(p_A; p_0, \alpha, m, s)| \leq \mathcal{O}(\frac{1}{n}) \).

**Proof.** From calculus, as \( \psi'(x) \) is a concave function, if \( 0 < t_1 < t_2 \) and \( \tau_0 \leq 0.5 \), then the following inequality is true:

\[
\int_{\tau_0 - t_2}^{\tau_0} \|\psi'(u)\|du > \int_{\tau_0 - t_1}^{\tau_0} \|\psi'(u)\|du \geq \int_{\tau_0}^{\tau_0 + t_1} \|\psi'(u)\|du. \tag{4.12}
\]

The second inequality is strict, except when \( \tau_0 = 0.5 \).

By Equations (4.6) and (4.10), because \( \mathbb{P}_0[|\bar{\tau} - \tau_0| \geq C_\mathcal{H}] = \alpha \), the rejection region of \( T_\mathcal{H} \) is when \( \bar{\tau} \in (-\infty, \tau_0 - C_\mathcal{H}) \cup [\tau_0 + C_\mathcal{H}, \infty) \), which is symmetric about \( \tau_0 \). We denote the rejection region of \( T^c_\mathcal{H} \) as \( \bar{\tau} \in (-\infty, \tau_0 - a_1] \cup [\tau_0 + a_2, \infty) \), where \( a_1, a_2 > 0 \). Due to the fact that the first inequality in Equation (4.12), \( a_1 \) and \( a_2 \) satisfy \( 0 < a_1 < C_\mathcal{H} < a_2 \) and \( C_\mathcal{H} - a_1 < a_2 - C_\mathcal{H} \).
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

For $t_1 = O\left(\frac{1}{\sqrt{n}}\right)$, from calculus, we have

$$
\int_{\tau_0 - t_1}^{\tau_0} \|\psi'(u)\| du - \int_{\tau_0}^{\tau_0 + t_1} \|\psi'(u)\| du < t_1 [Q(\tau_0 - t_1) - Q(\tau_0 + t_1)]
$$

$$= O\left(\frac{1}{\sqrt{n}}\right)O\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{n}\right),$$

where $\frac{\partial Q(x)}{\partial x} = q(x) := \|\psi'(x)\|$. Thus, it follows that $|a_1 - C_H| < O\left(\frac{1}{n}\right)$ and $|a_2 - C_H| < O\left(\frac{1}{n}\right)$.

On the other hand, with Equation (4.11) the powers $\pi_H(p_A; p_0, \alpha, m, s)$ and $\pi^c_H(p_A; p_0, \alpha, m, s)$ may be written as:

$$
\pi_H(p_A; p_0, \alpha, m, s) = \mathbb{P}_{\tau_A}[\bar{\tau} - \tau_0 \leq - C_H] + \mathbb{P}_{\tau_A}[\bar{\tau} - \tau_0 \geq C_H]
$$

$$= \Phi\left(\frac{\mu_{\tau_A} - C_H}{\sigma_{\tau_A}}\right) + 1 - \Phi\left(\frac{\mu_{\tau_A} + C_H}{\sigma_{\tau_A}}\right),$$

$$\pi^c_H(p_A; p_0, \alpha, m, s) = \mathbb{P}_{\tau_A}[\bar{\tau} - \tau_0 \leq - a_1] + \mathbb{P}_{\tau_A}[\bar{\tau} - \tau_0 \geq a_2]
$$

$$= \Phi\left(\frac{\mu_{\tau_A} - a_1}{\sigma_{\tau_A}}\right) + 1 - \Phi\left(\frac{\mu_{\tau_A} + a_2}{\sigma_{\tau_A}}\right),$$

where $\mu_{\tau_A} = \tau_A - \tau_0$. Then the difference between $\pi_H(p_A; p_0, \alpha, m, s)$ and $\pi^c_H(p_A; p_0, \alpha, m, s)$ is

$$
\pi^c_H(p_A; p_0, \alpha, m, s) - \pi_H(p_A; p_0, \alpha, m, s) =
$$

$$\left[\Phi\left(\frac{\mu_{\tau_A} - a_1}{\sigma_{\tau_A}}\right) - \Phi\left(\frac{\mu_{\tau_A} - C_H}{\sigma_{\tau_A}}\right)\right] - \left[\Phi\left(\frac{\mu_{\tau_A} + a_2}{\sigma_{\tau_A}}\right) - \Phi\left(\frac{\mu_{\tau_A} + C_H}{\sigma_{\tau_A}}\right)\right].$$

115
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Since \(|a_1 - C_H| < O(\frac{1}{n})\) and \(|a_2 - C_H| < O(\frac{1}{n})\), we may bound the difference between the two powers \(\pi_H(p_A; p_0, \alpha, m, s)\), \(\pi^c_H(p_A; p_0, \alpha, m, s)\) as:

\[
|\pi^c_H(p_A; p_0, \alpha, m, s) - \pi_H(p_A; p_0, \alpha, m, s)| < |a_1 - C_H| \phi\left(\frac{\mu_A - a_1}{\sigma_A}\right) + |a_2 - C_H| \phi\left(\frac{\mu_A + C_H}{\sigma_A}\right) = O\left(\frac{1}{n}\right),
\]

where \(\phi(x)\) denotes the probability density of a standard normal. We have \(\phi\left(\frac{\mu_A - a_1}{\sigma_A}\right) = O(1)\) and \(\phi\left(\frac{\mu_A + C_H}{\sigma_A}\right) = O(1)\) since \(\sigma_A = O(\frac{1}{\sqrt{n}})\), \(C_H = O(\frac{1}{\sqrt{n}})\).

The above lemma shows that the difference between \(\pi_H\) and \(\pi^c_H\) is negligible when \(m\) is large and the difference \(\tau_A - \tau_0 = \Omega\left(\frac{1}{n}\right) = \Omega\left(\frac{1}{m}\right)\).

4.3.4 Simulations

For the Hardy-Weinberg submanifold, the RDPG CLT covariance matrix defined in Equation (4.1) is a function of \(\tau \in (0, 1)\) and therefore, is not uniform over \(\tau \in (0, 1)\). In this section, we discuss some observations about the six powers in addition to those in Example 1.

Example 5. Let \(m = 1000\), \(s = 5\), \(\tau_0 = 0.3\), \(\tau_A = 0.35\), and \(K = 100\) (the number of nearest neighbors chosen in Isomap). Define \(h_{p^*} = \delta(p^*)\). We start with plotting the densities of \(T^c_{H}\) and \(T_{iso}\) under the choices of the null and the alternative.

Under \(H_0\), for \(i = m + 1, \ldots, m + s\), since \(\tilde{X}_i - p_0 \sim \mathcal{N}(x; 0, \frac{\Sigma(p_0)}{ns})\) and considering the fact that \(K\) is large enough, while constructing \(\hat{G} = (\hat{V}, \hat{E})\), we have \(\hat{V} = \{\tilde{X}_i\}_{i=1}^n \cup \)
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

(a) Estimated test densities at null $\tau_0 = 0.3$.

(b) Estimated test densities at alternative $\tau_A = 0.35$ with $\tau_0 = 0.3$.

Figure 4.8: Estimated test densities of $T_{iso}$ and $T_{C_H}$. The vertical lines mark the estimated critical values, respectively.

Next we plot six estimated ROC curves with $\alpha$ (false positive) being the $x$-axis and $\pi$ (true positive) being the $y$-axis.

$\hat{\pi}_{H,MDE} \approx \hat{\pi}_{H,MDE}^C$ and $\hat{\pi}_{H,MLE} \approx \hat{\pi}_{H,MLE}^C$ imply that the restricted test defined with low-dimensional Euclidean distances is approximately the same as that defined with geodesic distances for the Hardy-Weinberg submanifold. This observation is expected to hold for a broader class of smooth quasi-isometric curves.

When computing powers analytically in Section 4.3.2, Figures 4.6b and 4.7b show $\pi_{\mathbb{R}^3} < \pi_H$. This experiment demonstrates a consistent result as the power superiority
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Figure 4.9: Estimated ROC curves illustrate relations between $\alpha$ (false positive) and $\hat{\pi}_{R^3}^c$, $\hat{\pi}_{H,MLE}^c$, $\hat{\pi}_{H,MDE}^c$, $\hat{\pi}_{H,MLE}$, $\hat{\pi}_{H,MDE}$, $\hat{\pi}_{iso}$ (estimated true positive).

The fact that $\hat{\pi}_{iso}$ converges to 1 faster than $\hat{\pi}_{H,MLE}^c$ seems surprising at first since the “truth” is generally assumed to be preferred over the “estimated”. However, for the fixed collection $\{\tilde{X}_i\}_{i=1}^n$ the true Hardy-Weinberg curve $\mathcal{H}$ is not necessarily the best fit; rather, some other curve $\tilde{\mathcal{H}}$ is. For instance, if we fit $\{\tilde{X}_i\}_{i=1}^n$ parametrically via a Bézier curve as described in Section 4.5, then Figure 4.11 shows that the best fitted Bézier curve is not $\mathcal{H}$. We denote the learnt manifold via Isomap by $\tilde{\mathcal{H}}$. The robustness of Isomap suggests that $\tilde{\mathcal{H}}$ explains and fits $\{\tilde{X}_i\}_{i=1}^n$ better than $\mathcal{H}$ does. It is obvious that $\tilde{\mathcal{H}}$, $\tilde{\mathcal{H}}$, and $\mathcal{H}$ are not necessarily equal, nor do they follow some deterministic relationship. Thus, projecting $\{\tilde{X}_i\}_{i=1}^n$ onto $\mathcal{H}$ does not necessarily provide the “optimal” representation, not thereby yielding more superior inference than $\tilde{\mathcal{H}}$. 

118
The result that $\hat{\pi}_{H,MLE}$ converges faster than $\hat{\pi}_{H,MDE}$ is unexpected since there seems to be no assertion regarding whether one manifold representation is preferred to the other. Example 6 in Section 4.4 provides a more detailed discussion on the comparison of $\hat{\pi}_{H,MLE}$ vs. $\hat{\pi}_{H,MDE}$. In either case, the restricted test $\hat{\pi}_{H,*}$ converges faster than the unrestricted test $\hat{\pi}_{R^3}$.

4.4 Two-Sample Testing

We now shift our focus to two-sample testing. Consider $H \subset \mathbb{R}^3$, where $H$ is the Hardy-Weinberg curve given by $\tau \in [0,1]$ and $\psi : [0,1] \rightarrow H$ with $\psi(\tau) = [\tau^2, 2\tau(1-\tau), (1-\tau)^2]'$. Let $\tau_1, \ldots, \tau_m \overset{iid}{\sim} U[0,1]$ and $p_i = \psi(\tau_i)$. Here, the manifold $\mathcal{M} = H$ and the ambient space $\mathcal{A} = \mathbb{R}^3$. Let the auxiliaries $p_1, \ldots, p_m$ make up the rows of $X_m \in \mathbb{R}^{m \times 3}$.

Consider two groups of latent positions of size $s_1, s_2$ respectively. Let $X_{s_1+1}^* \sim \overset{iid}{h}_{p_1^*}$ in group one and $X_{s_1+s_2}^* \sim \overset{iid}{h}_{p_2^*}$ in group two, with $p_k^* = \psi(\tau_k^*)$, where $\tau_k^* \in [0,1]$ for $k = 1, 2$. Let $X_{m+i} = X_i^*$ for $i = 1, \ldots, s_1+s_2$. Denote $N = m + s_1 + s_2$.

We consider the following hypothesis for two-sample testing:

\[ H_0 : p_1^* = p_2^* \]
\[ H_A : p_1^* \neq p_2^* \quad (4.13) \]

In the following, we take $h_{p^*} = \delta(p^*)$ as an example.
Let $\mathbf{X}_0 = (X_1, \ldots, X_{m+s_1+s_2})'$ be a $(m + s_1 + s_2) \times 3$ matrix where the $i$th row $X_i = p_i$ except the last $s_1 + s_2$ rows, which are $p_0^*$. The null hypothesis assumes that the observed adjacency matrix $\mathbf{A}$ is sampled from the probability matrix $\mathbf{P}^0 = \mathbf{X}_0 \mathbf{X}_0^T$, i.e., $A_{ij} \sim \text{Bern}(P_{ij}^0)$ for $i < j$.

By the RDPG CLT stated in Theorem 8, $\hat{\mathbf{X}}_i \mathbf{W}_0 | X_i = x_i$ is approximately Gaussian around $x_i$ in $\mathbb{R}^{\text{rank}(\mathbf{X}_0 \mathbf{X}_0^T) = 3}$, for $i = m + 1, \ldots, m + s_1 + s_2$, for some rotation matrix $\mathbf{W}_0$. Denote $\hat{\mathbf{X}}_0 = \hat{\mathbf{X}}_0 \mathbf{W}_0$ and rows of $\hat{\mathbf{X}}_0$ as $\hat{X}_i$.

Under $H_0$, assume $p_1^* = p_2^* = p_0^*$. Since for $i = m + 1, \ldots, m + s_1 + s_2$, $\hat{X}_i | \mathbf{X}_0 \rightarrow \mathcal{N}(x; p_0^* \mathbf{W}_0^T, \mathbf{W}_0 \Sigma(p_0^*) \mathbf{W}_0^T / N)$, we observe the unaligned ASE estimates $\hat{X}_i | \mathbf{X}_0 \rightarrow \mathcal{N}(x; p_0^* \mathbf{W}_0^T, \mathbf{W}_0 \Sigma(p_0^*) \mathbf{W}_0^T / N)$.

Let $\tilde{X}^1 = \frac{1}{s_1} \sum_{i=m+1}^{m+s_1} \hat{X}_i$ and $\tilde{X}^2 = \frac{1}{s_2} \sum_{i=m+s_1+s_2}^{m+s_1+s_2} \hat{X}_i$ be the sample averages of group one and two in $\mathbb{R}^3$, respectively. We observe $\tilde{X}^k | \mathbf{X}_0 \rightarrow \mathcal{N}(x; p_0^* \mathbf{W}_0^T, \frac{\mathbf{W}_0 \Sigma(p_0^*) \mathbf{W}_0^T / N}{s_k})$ for $k = 1, 2$. We define the test statistic in $\mathbb{R}^3$ to be the $L_2$ norm, denoted by $T_{\mathbb{R}^3}$,

$$T_{\mathbb{R}^3}(\mathbf{A}) = \| \tilde{X}^1 - \tilde{X}^2 \|_2 = \| \tilde{X}^1 - \tilde{X}^2 \|_2,$$  

(4.14)

where $\mathbf{A}$ is a sampled adjacency matrix.

We use the rotated $\tilde{X}_i$ to obtain the MDE representation $\hat{\tau}_i$ and the MLE representation $\hat{\tau}_i$ defined in Section 4.3. Under consideration there are four restricted tests on $\mathcal{H}$, each of which is defined with one of two representations, (MLE) or (MDE), and one of two distances, Euclidean or geodesic, such that each test is defined with
a unique combination of one representation and one distance. We define the unrestricted test and the four types of restricted tests in two-sample testing as:

$$T_{H,MDE}(A) = |\tilde{r}_1 - \tilde{r}_2|,$$  \hspace{1cm} (4.15)

$$T_{c,H,MDE}^c(A) = |\int_{\tilde{r}_1}^{\tilde{s}_2} \|\psi'(u)\| du|,$$  \hspace{1cm} (4.16)

$$T_{H,MLE}(A) = |\tilde{r}_1 - \tilde{r}_2| = |L^T \tilde{X}_1 - L^T \tilde{X}_2| = |L^T (\tilde{X}_1 W_0^T - \tilde{X}_2 W_0^T)|,$$  \hspace{1cm} (4.17)

$$T_{c,H,MLE}^c(A) = |\int_{\tilde{r}_1}^{\tilde{s}_2} \|\psi'(u)\| du|.$$  \hspace{1cm} (4.18)

One subtle simplification of two-sample testing compared to one-sample testing is that the nonidentifiability of RDPGs is not an issue for the unrestricted test $T_{\mathbb{R}^3}$, the restricted tests $T_{H,MDE}, T_{H,MLE}$, or the learnt restricted test.

In defining Equation (4.14), we argue that using $\tilde{X}_i$ is equivalent to using $\hat{X}_i$ for two-sample testing as an orthogonal rotation or shift does not change the distance $\|\tilde{X}_i - \tilde{X}_j\|$ between any pair of operating points $\tilde{X}_i$ and $\tilde{X}_j$; while for one-sample testing the test statistic measures the distance between a group of rotated points and a fixed point $p_0$ on $H$. We note that defining the two test statistics on $H$ with geodesic distances requires the rotated $\tilde{X}_i$ because the definition of $\tilde{r}_i$ and geodesic distance utilize the form of $\psi$.

Now we learn the manifold mapping inverse $\psi^{-1}$ based on the collection $\{\tilde{X}_i\}_{i=1}^n$, or equivalently, $\{\hat{X}_i\}_{i=1}^n$. The learnt low-dimensional representation via Isomap is invariant to orthogonal rotations, since computing pairwise Euclidean distances in
\( \mathbb{R}^D \) as the first step in Isomap is invariant to orthogonal rotations. Let \( \psi_{iso}^{-1} : \hat{X}_i \rightarrow \hat{Z}_i \) be the mapping learnt from dimension \( D = 3 \) to \( d = 1 \). We have the estimated 1-dim representation of \( \hat{X}_i \) as \( \hat{Z}_i = \psi_{iso}^{-1}(\hat{X}_i) \), for \( i = 1, \ldots, m + s_1 + s_2 \). We define \( \hat{Z}_1 = \frac{1}{s_1} \sum_{i=m+1}^{m+s_1} \hat{Z}_i \) and \( \hat{Z}_2 = \frac{1}{s_2} \sum_{i=m+s_1+1}^{m+s_1+s_2} \hat{Z}_i \) as the sample averages.

Let \( T_{iso} \) denote the learnt test statistics via Isomap. \( T_{iso} \) for hypothesis (4.13) is defined as the “estimated manifold distance” between two samples,

\[
T_{iso}(A) = |\hat{Z}_1 - \hat{Z}_2|. \tag{4.19}
\]

Given the test size \( \alpha \), the critical value of the unrestricted test denoted by \( C_{\mathbb{R}^3} \) is defined to be \( C_{\mathbb{R}^3} = \{ C : \alpha = \sup_{p_0 \in [0,1]} P(p_0)p_0^3 \geq C \} \). Similarly, we define \( C_{H,MDE}, C_{H,MDE}^c, C_{H,MLE}, C_{H,MLE}^c \), and \( C_{iso} \), respectively.

To calculate the powers, under the alternative pair \( (p_1^* = \psi(\tau^*_1), p_2^* = \psi(\tau^*_2)) \) for some \( \tau^*_1 \neq \tau^*_2 \in [0,1] \), let \( X_A = (X_1, \ldots, X_{m+s_1+s_2})' \) be a \( (m + s_1 + s_2) \times 3 \) matrix where the \( i \)th row \( X_i = p_i \) for the first \( m \) rows, \( X_i = p_1^* \) for rows \( m+1, \ldots, m+s_1 \) and \( X_i = p_2^* \) for rows \( m+s_1+1, \ldots, m+s_2 \). Note that \( X_0 \) and \( X_A \) differ solely on the last \( s_1 + s_2 \) rows. Let an adjacency matrix \( A \) be sampled from the probability matrix \( P^A = X_A X_A^T \). The estimated latent position matrix \( \hat{X}_A \) is obtained from \( A \) via ASE.

Applying the rotation matrix \( W_A \) to \( \hat{X}_A \) to obtain the correctly aligned estimated latent positions for the purpose of CLT in Theorem 8, denote \( \hat{X}_A = \hat{X}_A W_A \) and rows of \( \hat{X}_A \) as \( \hat{X}_i \).
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

The power of the unrestricted test is $\pi_{\mathbb{R}^3}(p_1^*, p_2^*; \alpha, m, s) = \mathbb{P}(p_1^*, p_2^*; T_{\mathbb{R}^3} \geq C_{\mathbb{R}^3})$.

Likewise, we define $\pi_{\mathcal{H}, MDE}(p_1^*, p_2^*; \alpha, m, s)$, $\pi_{\mathcal{H}, MDE}(p_1^*, p_2^*; \alpha, m, s)$, $\pi_{\mathcal{H}, MLE}(p_1^*, p_2^*; \alpha, m, s)$, $\pi_{\mathcal{H}, MLE}(p_1^*, p_2^*; \alpha, m, s)$, and $\pi_{iso}(p_1^*, p_2^*; \alpha, m, s)$, respectively.

Borrowing the inference in Section 4.3.2 on $\mathbb{R}^3$, since $\tilde{X}_1 - \tilde{X}_2|p_0^* \to N(x; 0, \frac{W_0\Sigma(p_0^*\Sigma)^T}{N}(\frac{1}{s_1} + \frac{1}{s_2}))$, the critical value $C_{\mathbb{R}^3}$ may be calculated by

$$T_{\mathbb{R}^3}(A)|_{p_0^*} = \|\tilde{X}_1 - \tilde{X}_2\|_2^2_{p_0^*} \to \sum_{i=1}^{D} \frac{\lambda_i}{N}(\frac{1}{s_1} + \frac{1}{s_2}) \chi_1^2,$$

where $\lambda_i$ are the eigenvalues of $\Sigma(p_0^*)$ in descending order. Define $f_0^*$ and $F_0^*$ to be the density function and the distribution function of $\sum_{i=1}^{D} \lambda_i \chi_1^2$ respectively. Then we have $C_{\mathbb{R}^3}$ as a function of $\alpha$,

$$N \frac{s_1 s_2}{s_1 + s_2} C_{\mathbb{R}^3} = Q_{F_0}(1 - \alpha).$$

Since under alternatives of the form $(p_1^*, p_2^*) = (\psi(\tau_1^*), \psi(\tau_2^*))$, we have $\tilde{X}_1 - \tilde{X}_2|p_1^* - p_2^*, \frac{W_0\Sigma(p_1^*)\Sigma^T}{N s_1} + \frac{W_0\Sigma(p_2^*)\Sigma^T}{N s_2}) \to N(x; 0, \frac{W_0\Sigma(p_1^*)\Sigma^T}{N s_1} + \frac{W_0\Sigma(p_2^*)\Sigma^T}{N s_2})$, and

$$T_{\mathbb{R}^3}(A)|_{(p_1^*, p_2^*)} = \|\tilde{X}_1 - \tilde{X}_2\|_2^2_{(p_1^*, p_2^*)} \to \sum_{i=1}^{D} \frac{\lambda'_i}{N}(\frac{1}{s_1} + \frac{1}{s_2}) \chi_1^2.$$

We define $f_A$ and $F_A$ to be the density function and the distribution function of $\sum_{i=1}^{D} \lambda'_i \chi_1^2(b^2_i)$ respectively, where $\lambda'_i$ are the eigenvalues of $\Sigma(p_A) = \frac{W_0\Sigma(p_1^*)\Sigma^T}{s_1} + \frac{W_0\Sigma(p_2^*)\Sigma^T}{s_2}$.
in descending order. The power for the unrestricted test is

\[ \pi_{R^3}(p_1^*, p_2^*; \alpha, m, s) = \mathbb{P}_{p_1^*, p_2^*}\left[ \sum_{i=1}^{D} \chi_i^1 \chi_i^2 (b_i^2) \geq Q_{F_0}(1 - \alpha) \right] = 1 - F_A(Q_{F_0}(1 - \alpha)), \]

where \((b_1, b_2, b_3) = b = \sqrt{ns} \text{diag}(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \frac{1}{\sqrt{\lambda_3}})V (p_A - p_0), \) and \(\Sigma(p_A) = V^T \text{diag}(\lambda_1', \lambda_2', \lambda_3')V.\)

### 4.4.1 Simulation: Hardy-Weinberg Example Continued

Continuing the example in one-sample testing, we construct an experiment using Monte-Carlo (MC) to explore the relationship among the powers of the unrestricted test, the four restricted tests, and the learnt restricted test in two-sample testing.

**Example 6.** Let \(\alpha = 0.05, m = 1000, s_1 = s_2 = 5, \tau_1^* = 0.4, \tau_2^* = 0.45,\) and \(K = 100\) (the number of neighbors chosen in Isomap). Define \(h_{p^*} = \delta(p^*).\)

Initially, we approximate critical values via an MC simulation with \(B_0 = 999\) iterations. In each iteration \(b,\) we sample a graph \(A^b\) from the probability matrix \(P_0 = X_0X_0^T,\) where \(p_0^* = \psi(\tau_0^*),\) with \(\tau_0^* = (\tau_1^* + \tau_2^*)/2.\) The choice of \(\tau_0^*\) only affects the covariance matrix, as \(\tilde{X}_1 - \tilde{X}_2| \sigma_0^* \to \mathcal{N}(x; 0, \frac{W_0 \Sigma(p_0^*) W_0^T}{N} (\frac{1}{s_1} + \frac{1}{s_2})).\) We know \(\tau_0^* \in [\tau_1^*, \tau_2^*];\) thus, allowing \(\tau_0^*\) to be the middle point of \(\tau_1^*, \tau_2^*\) is a reasonable assumption when \(s_1 = s_2.\) The remainder of the simulation procedure to estimate critical values is similar to that in Example 1.
The powers are approximated with the estimated critical values and with $B_A = 1000$ iterations at $(p_1^*, p_2^*)$,

\[
\hat{\pi}_{R^3}(p_1^*, p_2^*; \alpha, m, s) = 0.220
\]
\[
\hat{\pi}_{H,MDE}(p_1^*, p_2^*; \alpha, m, s) = 0.765
\]
\[
\hat{\pi}_{c,H,MDE}(p_1^*, p_2^*; \alpha, m, s) = 0.769
\]
\[
\hat{\pi}_{H,MLE}(p_1^*, p_2^*; \alpha, m, s) = 0.692
\]
\[
\hat{\pi}_{c,H,MLE}(p_1^*, p_2^*; \alpha, m, s) = 0.691
\]
\[
\hat{\pi}_{iso}(p_1^*, p_2^*; \alpha, m, s) = 0.885.
\]

Now we consider McNemar’s test to check whether or not the difference between each pair of powers is significant, with $H_0 : \pi_k = \pi_\ell$, where $k, \ell$ each refers to one of the six tests we have been considering. For each McNemar test, we set $\alpha = 0.05$.

As in one-sample testing, some expected results are observed. The power of the unrestricted test is significantly different from the others; for restricted tests, the power of Euclidean distance tests is not different from that of geodesic distance tests under the MLE (or MDE) representation, and the two tests with different manifold representations under the same distance are significantly different from each other. Unlike the mirror example in one-sample testing, $\hat{\pi}_{H,MLE}$ is smaller than $\hat{\pi}_{H,MDE}$. This difference indicates that the MDE representation is more robust than the MLE.
one, since as \( \tau \) gets nearer to 0.5 the magnitude of the variance of \( \hat{X}_i \) increases, as Figure 4.3 suggests. To unify the current two-sample testing example with that of one-sample testing, we change the true positions to \( \tau_1^* = 0.3, \tau_2^* = 0.35 \) with \( \alpha, m, s_1, s_2, K \) unchanged, and the powers become

\[
\hat{\pi}_{R3}(p_1^*, p_2^*; \alpha, m, s) = 0.281
\]
\[
\hat{\pi}_{H,MDE}(p_1^*, p_2^*; \alpha, m, s) = 0.571
\]
\[
\hat{\pi}_{c,H,MDE}(p_1^*, p_2^*; \alpha, m, s) = 0.570
\]
\[
\hat{\pi}_{H,MLE}(p_1^*, p_2^*; \alpha, m, s) = 0.733
\]
\[
\hat{\pi}_{c,H,MLE}(p_1^*, p_2^*; \alpha, m, s) = 0.744
\]
\[
\hat{\pi}_{iso}(p_1^*, p_2^*; \alpha, m, s) = 0.895.
\]

The above result is consistent with that in Example 4.1 on the relationship between MLE and MDE representations, and thus verifies the assumption that MDE representations are more robust than MLE ones.

\[\square\]

### 4.4.2 Real Data: the Connectome Experiment

We consider the graph of the right hemisphere mushroom body (MB) connectome of the larval Drosophila brain [28]. The connectome consists of four distinct types of neurons: Kenyon Cells (KC), Input Neurons (MBIN), Output Neurons (MBON),
and Projection Neurons (PN). A directed adjacency matrix $A_{213}$ represents the connectivity among all 213 nodes in the connectome, among which 100 are KC neurons. The $GMM \circ ASE$ method, which composes Gaussian mixture modeling with adjacency spectral embedding and which was proposed in [28] estimates the neurons’ latent positions $\hat{X}_1, \ldots, \hat{X}_{213}$ that are embedded in $\mathbb{R}^6$ and suggests that all KC neurons’ estimated latent positions $\hat{X}_{KC,1}, \ldots, \hat{X}_{KC,100}$ form a 1-dimensional quadratic curve $C_{KC}$ (1-dimensional LSM). Let $\hat{C}_{KC}$ denote the semiparametric GMM estimation of $C_{KC}$ in [28], where $\hat{C}_{KC}$ is indexed as $\hat{C}_{KC} = \{(\mu(t), C(t)) : \mu(t) = m_0 t^2 + 2m_1 t (1 - t) + m_2 t^2, C(t) = (\sigma_0 (1 - t)^2 + \sigma_1 (1 - t)^2)I_6, t \in [0,1]\}$, where $m_0, m_1, m_2 \in \mathbb{R}^6$ and $\sigma_0, \sigma_1 \in \mathbb{R}$.

Li [39] described so-called “claws” associated with each KC neuron and suggests a relationship between the neuron age and its number of claws. A question raised is whether 3-claw and 4-claw neurons have the same generating latent position vector. If we formulate the question as a two-sample testing problem that treats both 3-claw and 4-claw neurons as two separate populations, then the hypothesis test is as follows:

$H_0 :$ 3-claw and 4-claw neurons both have the same generating point

$H_A :$ the generating point of 3-claw neurons is different from that of 4-claw neurons

Let $\hat{X}_{KC,k\text{-claw},i}$ denote the $i$th estimated latent position vector of $k$-claw neurons, where $k = 0, 1, \ldots, 6$ and $i = 1, 2, \ldots, n_k$ with $n_k$ being the total number of $k$-claw
neurons. Note that $\sum_{k=0}^{6} n_k = 100$. We model all KC neurons by assuming every $k$-claw neuron has the same generating point $X_{KC,k}$-claw and the 1-dim LSM indicates $X_{KC,k}$-claw $\in \mathcal{C}_{KC} \subset \mathbb{R}^6$. Note that this assumption makes this 1-dim LSM a stochastic block model (SBM), which is a special case of RDPGs. As $\mathcal{C}_{KC}$ is not known and $\hat{\mathcal{C}}_{KC}$ is the "optimal" estimation available for it, we treat $\hat{\mathcal{C}}_{KC}$ as the truth while conducting the following two-sample testing experiment.

Under $H_0$, we then have an SBM of 6 blocks with one block containing 3-claw neurons and 4-claw neurons while under $H_A$, we have another SBM of 7 blocks. To be more specific, let $X_{KC,0}$ and $X_{KC,A}$ denote the latent position matrix of all KC neurons under $H_0$ and $H_A$, respectively. Under $H_0$, rows $1 + \sum_{i=0}^{k-1} n_i$ to $\sum_{i=0}^{k} n_i$ of $X_{KC,0}$ are equal to $X_{KC,k}$-claw, for $k = 0, 1, 2, 5, 6$, and rows $1 + \sum_{i=0}^{2} n_i$ to $\sum_{i=0}^{4} n_i$ of $X$ are equal to $X_{KC,3/4}$-claw, which is the weighted average of $X_{KC,3}$-claw and $X_{KC,4}$-claw; while under $H_A$, rows $1 + \sum_{i=0}^{k-1} n_i$ to $\sum_{i=0}^{k} n_i$ of $X_{KC,A}$ equal to $X_{KC,k}$-claw, for $k = 0, 1, 2, 3, 4, 5, 6$. The hypothesis test may be rephrased as which of $X_{KC,0}$ and $X_{KC,A}$ is the correct latent position matrix for the KC neurons.

We perform the unrestricted two-sample test in $\mathbb{R}^6$, the restricted test in $\hat{\mathcal{C}}_{KC}$ and the learnt restricted test with the 1-dim representation obtained via Isomap. The unrestricted test statistic is defined as $T_{\mathbb{R}^6} = \|\hat{X}_{KC,3}$-claw $- \hat{X}_{KC,4}$-claw$. The restricted test in $\hat{\mathcal{C}}_{KC}$ rejects the null hypothesis if $T_{\hat{\mathcal{C}}_{KC}} = |\hat{t}_{KC,3}$-claw $- \hat{t}_{KC,4}$-claw$|$ is sufficiently large, where $\hat{t}_{KC,k}$-claw, $i = \text{arg min}_{t \in [0,1]} \| \mu(t) - \hat{X}_{KC,k}$-claw, $i \|_1$ is the MDE representation that utilizes the aligned latent position estimate $\hat{X}_{KC,k}$-claw, $i = \hat{X}_{KC,k}$-claw, $i W$. As
Section 4.3.3 suggests, the difference between the two powers $\pi_{KC}$ and $\pi_{C}$ defined by 1-dim Euclidean distance and geodesic distance, respectively is negligible. The learnt restricted test statistic is $T_{iso} = |\tilde{Z}_{KC,3}\text{-claw} - \tilde{Z}_{KC,4}\text{-claw}|$ with $\tilde{Z}_{KC,k}\text{-claw, i}$ representing the 1-dim Isomap embedding of $\hat{X}_{KC,k}\text{-claw, i}$. Note that only the restricted test statistic $T_{KC}$ requires the rotated latent position vectors $\hat{X}_{KC,k}\text{-claw, i}$ due to the definition of 1-dim representations $\hat{X}_{KC,k}\text{-claw, i}$.

We run 999 iterations and 1000 iterations, respectively to estimate the critical values and powers of the three tests. The estimated powers are as follows:

<table>
<thead>
<tr>
<th>$\hat{\pi}_R^6$</th>
<th>$\hat{\pi}_{KC}$</th>
<th>$\hat{\pi}_{iso}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.133</td>
<td>0.812</td>
<td>0.655</td>
</tr>
</tbody>
</table>

Table 4.1: Estimated powers of the two-sample test with $\alpha = 0.05$, $n_3 = 15$ and $n_4 = 16$.

By varying the size of tests $\alpha$, we have the following ROC curves.
Figure 4.10: Estimated ROC curves of the unrestricted test (black), the restricted test (red), and the learnt restricted test (blue).

In light of the curve structure of all KC neurons, we expect that the restricted test and the learnt restricted test are superior to the unrestricted one. For the learnt restricted test, all other KC neurons except 3-claw and 4-claw ones may be viewed as auxiliaries to help estimate the curve $C_{KC}$. As expected, Figure 4.10 suggests both the restricted test and the learnt restricted test converge faster than the unrestricted test. Instead of the aforementioned Hardy-Weinberg example, the restricted test has better performance than the learnt restricted test. As this real-data example contains a network of much smaller size ($n = 100$ vs. $n = 1010$ in Example 6), the embedding result of Isomap is less trustworthy when there are fewer samples (smaller $n$) with greater errors ($\sup_{1 \leq i \leq n} \| \tilde{X}_i - X_i \|$). Thus, we are not surprised to see that $\hat{\pi}_{C_{KC}}$ is superior to $\hat{\pi}_{iso}$ here. In either case, learnt restricted tests are superior to, thus preferred over, unrestricted tests.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

4.5 Learning Hierarchy

When comparing the efficiencies of inference in $\mathcal{A}$, $\mathcal{M}$ and $\hat{\mathcal{M}}$, Section 4.3 and Section 4.4 are based on the assumption that $\hat{X}_i$ is perfectly recovered from $\hat{X}_i$ to align $X_i$ for $i = 1, \ldots, n$. This assumption requires knowledge of $X_i$, while in practice the ASE algorithm serves the function of recovering $X_i$.

Given an adjacency matrix representing a network, the structure discovery task of the network is significant in many applications. Depending on how much information is provided about the network, such as application fields and other prior information, different assumptions about the network model may be made to better fulfill the structure discovery task. Here is an outline of the network learning hierarchy in terms of the information known for the network of interest and the network’s embedded space, in the context of RDPGs:

(I) the network is an instance of an RDPG;

(II) the dimension of the latent position space is $D$;

(III) there exists a submanifold $\mathcal{M}$ of intrinsic dimension $d$ embedded in the latent position space;

(IV) the “form” or “skeleton” of $\mathcal{M}$ is known, for instance, $\mathcal{M}$ is a quadratic curve;

(V) the manifold mapping function $\rho$ of $\mathcal{M}$ is known, including the boundary of $\mathcal{M}$;

(VI) the true latent position matrix $X$ is known.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Progressing in this learning hierarchy means that more information is available for the network. For example, knowing level (IV) means knowing level (I-III) as well.

Consider the most ideal senario in level (VI), where the rotation matrix $W$ in Theorem 8 can be well approximated via procrustes, i.e., $\hat{W} = \arg\min_{\hat{W}} \|X - \hat{X}W\|_F$ and $\hat{W}$ is a good approximation of $W$. If $Q = X^T X \in \mathbb{R}^{D \times D}$ has distinct eigenvalues, then the columns of $W$ are formed by eigenvectors of $Q$ corresponding to eigenvalues in descending order. Once we have $\hat{X} = \hat{X}W$, which is the estimated latent position matrix after aligning to the true manifold $M$, the RDPG CLT follows. An example has been demonstrated in Figure 4.2 when $M = H$.

We now shift our focus to (V). Knowing $X$ helps us derive theoretical results, but this is usually not practical. If we know the “form” of $M$, then we model $\hat{X}$ according to the structure of $M$, which will give us an estimated mapping $\rho^{-1}$ from $X \in \mathbb{R}^{n \times D}$ to $\hat{Z} \in \mathbb{R}^{n \times d}$. $\rho^{-1}$ is an orthogonal rotation to the mapping $\rho^{-1}'$ that is estimated from $\hat{X} \in \mathbb{R}^{n \times D}$ to $\hat{Z} \in \mathbb{R}^{n \times d}$.

We take $M = H$ as an example. The Hardy-Weinberg curve can be written in the form of a quadratic Bézier curve, $\psi(t) = e_1 t^2 + e_2 2t(1 - t) + e_3 (1 - t)^2$, where $e_i$ represents a 3-dim vector with 0 entries except the $i$th entry being 1, for $i = 1, 2, 3$. The optimal fitted quadratic Bézier curve with respect to $\{\hat{X}_i\}_{i=1}^n$ is denoted by $\hat{\psi}(t) = M_1 t^2 + M_2 2t(1 - t) + M_3 (1 - t)^2$, where $M_1, M_2, M_3 \in \mathbb{R}^3$. Let $\hat{X}$ denote the alignment of $\hat{X}$ in level (V) via Bézier curve fitting.

We observe that the rotation from $\hat{X}$ to $\hat{X}$ is the same as the rotation from
(M₁, M₂, M₃) to (e₁, e₂, e₃) with the boundary of M being the two endpoints of X. After finding the optimal Bézier curve of X with control points (M₁, M₂, M₃) and with the endpoints constraint satisfied, the estimated rotation matrix W then satisfies (M₁, M₂, M₃)T W = (e₁, e₂, e₃)T. Thus, X = WX. Figure 4.11 provides a 2D visualization of rows of X, X, X and X, showing that X and X are mostly identical.

If we consider the hypothesis problem defined in Equation (4.2) at level (V), then with X learnt from W, we define test statistics Tₚ₃, Tₕ, Tₕ c and T iso with respect to X in the way we define test statistics Tₚ₃, Tₕ, Tₕ c and T iso with respect to X. If X → X as m → ∞, then we expect π(·) ≈ π(·) for the same type of test (·), where π(·) is the power corresponding to the test statistics T(·).

Figure 4.11: A 2D visualization of 3D data. The green dots represent rows of X (Hardy-Weinberg curve), the red dots are rows of X, the black dots are rows of X, and the blue dots are rows of X. The black dots and the blue dots largely overlap.

We generalize the procedure above for learning W to any M that can be written
as a Bézier curve \( B(t) = \sum_{k=0}^{K-1} h_k t^k \) of \( K \) control points, where each \( h_k \in \mathbb{R}^D \).

Assuming \( h_k \) given \( k = 0, \ldots, K - 1 \), we estimate the rotation matrix through the following steps:

1. First we find the optimal fitted Bézier curve of order \( K - 1 \) with respect to \( \hat{X} \in \mathbb{R}^{n \times D} \), denoted by \( \hat{B}(t) = \sum_{k=0}^{K-1} \hat{h}_k t^k \).

2. An estimation \( \hat{W} \) of \( W \) is obtained by \( [\hat{h}_0, \ldots, \hat{h}_{K-1}]^T \hat{W} = [h_0, \ldots, h_{K-1}]^T \).

Remark 1. The learning of the optimal Bézier curve is typically based on minimizing the residual standard error between \( \hat{X} \) and the Bézier curve iteratively. To fit a \( K \) order Bézier curve on \( \hat{X} \), we can formulate the following:

\[
(\hat{h}_0, \ldots, \hat{h}_{K-1}) = \arg \min_{h_0, \ldots, h_{K-1}, t_1, \ldots, t_n} \sum_{i=1}^{n} (\hat{X}_i - \sum_{k=0}^{K-1} h_k t_i^k)^2.
\]

We consider a generalized hypothesis of that in Equation (4.2) to \( p^*, p_0 \in \mathcal{M} \), where \( \mathcal{M} \) can be written as a Bézier curve \( B(t) = \sum_{k=0}^{K-1} h_k t^k \) of \( K \) control points and \( p_0 = B(\tau_0) \) for some \( \tau_0 \in [0, 1] \). The unrestricted test \( T_{\mathcal{R},0} \), the restricted tests \( T_{\mathcal{M}} \), \( T_{\mathcal{M}^c} \), and the learnt restricted test \( T_{iso} \) are defined according to Equations (4.3), (4.4), (4.5) and (4.8). For a generalized two-sample testing with the hypothesis defined in Equation (4.13), we define the unrestricted test \( T_{\mathcal{R},D} \), the restricted tests \( T_{\mathcal{M}}, T_{\mathcal{M}^c} \), and the learnt restricted test \( T_{iso} \) according to Equations (4.14), (4.15), (4.16), and (4.19).

When estimating \( W \), in addition to the estimator \( \hat{W} \) by the Bézier curve fitting,
there is an easier method. As $X_i$ are i.i.d. sampled from $\mathcal{M}$, if we uniformly randomly sample another $n$ points, say $Y_i$ from $\mathcal{M}$, then $X^T X$ and $Y^T Y$ are expected to be very close, and so are their eigenvectors. As an extension to Theorem 8, we show in Section 4.5.1 that under certain assumptions of the sparsity factor $\rho_n$, there exists an estimated rotation matrix $\hat{W}_1$ independent of $X$, such that $\sqrt{n\rho_n}(\hat{X}_i \hat{W}_1 / \sqrt{\rho_n} - X_i) |X$ is asymptotically normally distributed with mean 0.

For level (IV), for example, the order of the Bézier curve is known to be $K$ and the control points $h_k$ for $k = 0, \ldots, K - 1$ are unknown, then semiparametric methods can be useful in estimating $h_k$ [28].

For level (III), where we assume the intrinsic dimensionality of $\mathcal{M}$ is known but not its form, under certain assumptions on $\mathcal{M}$ such as smoothness and quasi-isometric property, we can apply manifold learning algorithms such as Isomap to learn $\mathcal{M}$. If $\mathcal{M}$ is a smooth 1-dim curve, Bézier curve fitting algorithms can be applied to estimate the number of control points and their exact nature.

For level (II), with $\hat{X}$, an estimate for the intrinsic dimension of $\mathcal{M}$ from scree plots is selected, for example, by Isomap. Section 4.5.2 studies an example with respect to level (II) and (III) in two-sample testing.

### 4.5.1 CLT with Estimated Rotation Matrix

Since the probability matrix $P$ is symmetric and positive semidefinite with rank $D$, it can also be written as $P = USU^T$, where $U \in \mathbb{R}^{n \times D}$ has orthonormal columns
and \( S \) is diagonal with positive decreasing entries along the diagonal. There exists some random orthogonal matrix \( W \in \mathbb{R}^{D \times D} \) such that \( US^{1/2} = XW \), where the columns of \( W = XU \) are the eigenvectors of \( X^TX \) corresponding to the eigenvalues in descending order. Assume \( X^TX \) has distinct eigenvalues. For an RDPG instance \( A \), let \( \hat{X} = ASE(A) \), by Theorem 8, \( \sqrt{n}(\hat{X}W - X_i) |X \sim N(x; 0, \Sigma(X_i)) \). If \( X^TX \) has repeated eigenvalues, then \( W \) is more complicated.

Now if \( X \) is unknown, \( W \) can be estimated via \( \hat{X} \) and the fact that \( X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{H} \) (the Hardy-Weinberg curve). Suppose we independently generate another set of random points \( Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{H} \), and the \( i \)th row of the matrix \( Y \in \mathbb{R}^{n \times D} \) is given by \( Y_i \) for \( i = 1, \ldots, n \). Let the columns of \( \hat{W}_1 \) be the eigenvectors of \( Y^TY \) corresponding to the eigenvalues in decreasing order, then we expect \( \hat{W}_1 \approx W \). Furthermore, we consider \( Y^TY = E[X^TX] \).

Suppose \( P = \rho_nXX^T \) and the sparsity factor \( 0 < \rho_n \leq 1 \) is known, then Theorem 8 implies \( \sqrt{n\rho_n}(\hat{X}W/\sqrt{\rho_n} - X_i) |X \sim N(x; 0, \Sigma(X_i)) \). Let \( W \) and \( \hat{W}_1 \) be defined as stated above. In such a case, their difference is bounded of order \( \frac{1}{\sqrt{n}} \):

**Lemma 9.** \( \|\hat{W}_1 - W\| \leq \frac{c}{\sqrt{n}} \) for some constant \( c \).

**Proof.** Let \( Y_D = \rho_nY^TY \) and \( X_D = \rho_nX^TX \). There exists a matrix \( E \in \mathbb{R}^{D \times D} \) such that \( Y_D = X_D + E \). By Theorem 8, \( E_{ij} \sim N(x; 0, n\rho_n^2\sigma_{ij}^2) \), where \( \sigma_{ij} \) is constant with respect to \( n, \rho_n \). Thus, \( \|E\| = O(\sqrt{n}\rho_n) \). Let \( \lambda_i \) denote the \( i \)th eigenvalue of \( X_D \) in decreasing order and \( \hat{\lambda}_i \) denote the \( i \)th eigenvalue of \( Y_D \) in decreasing order. Let \( X_D = V\Lambda V^T \), where \( V \) is the non-singular eigenvector matrix and \( \Lambda \) is the diagonal
matrix with \( \lambda_i \) as diagonal entries in descending order.

By the Bauer-Fike Theorem \([46]\), with the conditional number \( \kappa(V) = \|V\|_2\|V^{-1}\|_2 \) of \( V \) being constant, we have an upper bound on the difference between the eigenvalues of \( X_D \) and those of \( Y_D \),

\[
|\lambda_i - \hat{\lambda}_i| \lesssim \kappa(V)O(\sqrt{n\rho_n})
\]

\[
\lesssim O(\sqrt{n\rho_n}).
\]

Therefore, we can write \( \hat{\lambda}_i = \lambda_i + \epsilon \lambda_i \), where \( \epsilon = O(1/\sqrt{n}) \) since \( \lambda_i = \Theta(n\rho_n) \).

Thus, the difference between the eigenvectors \( v_i \) of \( X_D \) and \( \hat{v}_i \) of \( Y_D \) is bounded by

\[
\|\hat{v}_i - v_i\| = \|(X_D - \lambda_i I_D)^{-1}(\epsilon \lambda_i E)\hat{v}_i\|
\]

\[
\leq \|(X_D - \lambda_i I)^{-1}\| \cdot \|\epsilon \lambda_i E\| \cdot \|\hat{v}_i\|
\]

\[
\lesssim \frac{1}{n\rho_n} \cdot \sqrt{n\rho_n}
\]

\[
= O\left(\frac{1}{\sqrt{n}}\right).
\]

As the columns of \( W \) are \( v_i \) and the columns of \( \hat{W}_1 \) are \( \hat{v}_i \), we conclude that \( \|\hat{W}_1 - W\| \leq \frac{c}{\sqrt{n}} \) for some \( c = O(1) \). \qed
Suppose we have the estimate \( \hat{W}_1 \) instead of \( W \), then

\[
\sqrt{n\rho_n(\hat{X}_i\hat{W}_1/\sqrt{\rho_n} - X_i)}|\mathbf{X} = \sqrt{n\rho_n(\hat{X}_iW/\sqrt{\rho_n} - X_i)}|\mathbf{X} + \sqrt{n\rho_n}\hat{X}_i(\hat{W}_1 - W)/\sqrt{\rho_n}|\mathbf{X}
\]

\[
= \sqrt{n}(\hat{X}_iW - \sqrt{\rho_n}X_i)|\mathbf{X} + \sqrt{n}\hat{X}_i(\hat{W}_1 - W)|\mathbf{X}
\]

\[
\rightarrow \mathcal{N}(x; 0, \Sigma(\sqrt{\rho_n}X_i)) + O(\|\hat{X}_i\|)
\]

\[
\rightarrow \mathcal{N}(x; 0, \Sigma(\sqrt{\rho_n}X_i)) + O(\sqrt{\rho_n}).
\]

If the sparse factor \( \rho_n = o(1) \), then for an estimated \( \hat{W}_1 \) that does not depend on \( X \) but rather the fact that rows of \( X \) are i.i.d. from the Hardy-Weinberg curve, we have the Central Limit Theorem for \( \hat{X}_i \) after rotating with the estimated rotation matrix \( \hat{W}_1 \),

\[
\sqrt{n\rho_n(\hat{X}_i\hat{W}_1/\sqrt{\rho_n} - X_i)}|\mathbf{X} \rightarrow \mathcal{N}(x; 0, \Sigma(\sqrt{\rho_n}X_i)).
\]

### 4.5.2 Superfluous Embedding Dimensions

In Section 4.3 since \( X_i \in \mathbb{R}^3 \), the optimal dimension of \( \hat{X}_i \) should be three for the sake of CLT. Figure 4.1 shows that \( D = 3 \) is chosen by the ‘elbow rule’ for the embedding dimension of \( ASE(A) \). While the ‘elbow rule’ is ubiquitous in model selection, this sort of visual analysis is subjective. For instance, according to preference one may argue that \( D = 2 \) or \( D = 4 \) should be chosen in Figure 4.1.

In this section we construct an experiment to explore the effect of ASE dimen-
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Dimensionality selection on the network inference, in particular in the two-sample testing defined in Equation (4.13). We consider choosing \( D = 5 \) for the estimated latent position space, denoted by \( \hat{Y} = ASE(A) \) whose rows are denoted by \( \hat{Y}_i \), where \( \hat{Y}_i \in \mathbb{R}^5 \).

We emphasize that \( \hat{X}_i \) is equal to the first 3-dim of \( \hat{Y}_i \), i.e., \( \hat{X}_i = \hat{Y}_i[1 : 3] \).

Since there are two dimensionality selections when defining \( T_{iso} \), one for ASE space and one for \( M \), we now more generally define this as \( T_{3,1} \), with 3 being the dimension of the ASE space and 1 being that of the manifold. To generalize this definition, we define test statistics \( T_{d_1,d_2} \) with \( d_1 \geq d_2 \), where \( d_1 \) is the dimension selected for the latent position space by \( ASE(A) \) and \( d_2 \) is the dimension selected for the intrinsic dimension of \( M \). Let \( \pi_{d_1,d_2} \) denote the corresponding power.

For instance, the test statistic \( T_{3,1}(\tilde{t}_{3,1}^1, \tilde{t}_{3,1}^2) \) equals the Euclidean distance between \( \tilde{t}_{3,1}^1 \) and \( \tilde{t}_{3,1}^2 \), where \( \tilde{t}_{3,1}^1 = \sum_{i=m+1}^{m+s_1+t_3^1} \tilde{t}_{3,1}^i \) and \( \tilde{t}_{3,1}^2 = \sum_{i=m+s_1+s_2}^{m+s_1+s_2} \tilde{t}_{3,1}^i \) and each \( \tilde{t}_{3,1}^i \in \mathbb{R} \) is the estimated 1-dim representation of \( \hat{X}_i \) via Isomap. If \( d_1 = d_2 \), then \( T_{d_1,d_1} \) denotes an unrestricted test statistic with the ASE space dimension being \( d_1 \). For instance, the test statistic \( T_{5,5}(\tilde{Y}_1, \tilde{Y}_2) \) equals the Euclidean distance between \( \tilde{Y}_1 \) and \( \tilde{Y}_2 \), where \( \tilde{Y}_1 = \sum_{i=m+1}^{m+s_1} \tilde{Y}_i \) and \( \tilde{Y}_2 = \sum_{i=m+s_1+s_2}^{m+s_1+s_2} \tilde{Y}_i \).

We consider the aforementioned unrestricted tests \( T_{d_1,d_1} \) and the learnt restricted tests \( T_{d_1,d_2} \). One reasonable expectation is that \( \pi_{d_1,d_2} \approx \pi_{d'_1,d_2} \), where \( d_1 > d'_1 \geq d_2 \) and \( \pi_{d_1,d_2} > \pi_{d_1,d'_2} \) if \( d_2 < d'_2 \leq d_1 \). We note that the rotation matrix \( W \) does not require our attention as we consider two-sample unrestricted tests and learnt restricted tests here.
Example 7. We show some power results via Monte-Carlo with iterations $B_0 = 999$ to estimate critical values and $B_A = 1000$ to estimate powers $\pi_{d_1,d_2}$ for various choices of $d_1, d_2$, with $\alpha = 0.05, m = 1000, s_1 = s_2 = 5, \tau_1 = 0.4, \tau_2 = 0.45, K = 100$ (the number of neighbors in Isomap):

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>5</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.052</td>
<td>0.055</td>
<td>0.079</td>
<td>0.875</td>
</tr>
<tr>
<td>3</td>
<td>0.210</td>
<td>0.118</td>
<td>0.885</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.890</td>
<td>0.885</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.059</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Estimated powers $\hat{\pi}_{d_1,d_2}$

The results of the experiment indicate some expected results, such as $\hat{\pi}_{5,1} \approx \hat{\pi}_{3,1} \approx \hat{\pi}_{2,1}$. However, we also observe some compelling relations: $\hat{\pi}_{5,5} \approx \hat{\pi}_{5,3} < \hat{\pi}_{3,3}$.

In practice, given a data set of high dimensions, PCA or some other linear dimensionality reduction method is typically applied first to reduce the dimensions before the manifold learning step [47, 5, 31]. One practical reason for this is that most manifold learning algorithms are computationally more expensive than linear dimensionality reduction methods. From the subsequent inference perspective, one should prefer linear to nonlinear dimensionality reductions when the underlying submanifold is approximately linear. For instance, Figure 4.1 shows that the last two dimensions
of $\hat{Y}_i \in \mathbb{R}^5$ are redundant. Linear dimensionality reduction on $\hat{Y}_i$ gives $\hat{X}_i$ and the corresponding power $\hat{\pi}_{3,3}$ is greater than $\hat{\pi}_{5,3}$, which in turn corresponds to a nonlinear reduction on $\hat{Y}_i$ and to the same optimal number of dimensions in the linear reduction scree plot.

4.6 Appendix: Proofs

Given a data set $\{x_i\}_{i=1}^n \subset \mathcal{M}$, we define three different distances between two points $x_0$ and $x_p$ in the set. These distances are manifold metric $d_{\mathcal{M}}$ (i.e. geodesic distance), shortest path distance $d_S$ and graph distance $d_G$ respectively, and they are defined as

$$d_{\mathcal{M}}(x_0, x_p) = \inf_{\gamma} \{ \text{length}(\gamma) \}$$

$$d_G(x_0, x_p) = \min_P (\|x_0 - x_1\| + \cdots + \|x_{p-1} - x_p\|)$$

$$d_S(x_0, x_p) = \min_P (d_{\mathcal{M}}(x_0, x_1) + \cdots + d_{\mathcal{M}}(x_{p-1}, x_p))$$

where $\gamma$ is set of all piecewise smooth arcs connecting $x_0, x_p$ in $\mathcal{M}$ and $x_i \in \mathcal{M}, x_i \in \mathbb{R}^D$ for $i = 0, \ldots, p$.

The following theorem proves $d_S \approx d_{\mathcal{M}}$:

Theorem 10. Let $\epsilon$ and $\delta$ be positive, with $4\delta < \epsilon$. Suppose:

1. The graph $G$ contains all edges $xy$ for which $d_{\mathcal{M}}(x, y) \leq \epsilon$; and
2. For every point \( m \in M \), there exists a data point \( x_i \) for which \( d_M(m, x_i) \leq \delta \).

Then for all pairs of data points \( x, y \) we have

\[
d_M(x, y) \leq d_S(x, y) \leq (1 + 4\delta/\epsilon) d_M(x, y)
\]

The second condition is referred as the \( \delta – \text{sampling condition} \).

The following lemma shows \( d_G \approx d_S \):

**Lemma 10.** Let \( \lambda > 0 \) be given. Suppose the points \( x_i, x_{i+1} \in M \) satisfy the condition

\[
\|x_i - x_{i+1}\| \leq \min(s_0, (2/\pi)r_0\sqrt{24\lambda}).
\]

Suppose that there is a geodesic arc of length \( d_M(x_i, x_{i+1}) \) connecting \( x_i \) to \( x_{i+1} \). Then

\[
(1 - \lambda)d_M(x_i, x_{i+1}) \leq \|x_i - x_{i+1}\| \leq d_M(x_i, x_{i+1})
\]

where minimum branch separation \( s_0 = s_0(M) \) is the largest positive number for which \( \|x - y\| < s_0 \) implies \( d_M(x, y) \leq \pi r_0 \) for \( x, y \in M \) and the minimum radius of curvature \( r_0 = r_0(M) = 1/\max_t \{\|\dot{\gamma}(t)\|\} \).

The main theorem that establishes \( d_M \approx d_G \) is as follows:

**Theorem 11.** Let \( \{x_i\} \) be a finite set of data points in \( M \), given a graph \( G \) on \( \{x_i\} \)
and positive real numbers $\lambda_1, \lambda_2 < 1$. We also refer to positive real numbers $\epsilon_{\min}, \epsilon_{\max}$ and $\delta$. Suppose

1. $G$ contains all edges $xy$ of length $\|x - y\| \leq \epsilon_{\min}$;

2. All edges of $G$ have length $\|x - y\| \leq \epsilon_{\max}$;

3. $\{x_i\}$ satisfies the $\delta$–sampling condition in $\mathcal{M}$;

4. $M$ is geodesically convex, i.e. any two points $x, y \in \mathcal{M}$ are connected by a geodesic of length $d_M(x, y)$;

5. $\epsilon_{\max} \leq \min(s_0(\mathcal{M}), (2/\pi)r_0(\mathcal{M})\sqrt{24\lambda_1})$; and

6. $\delta \leq \lambda_2\epsilon_{\min}/4$;

then

$$(1 - \lambda_1)d_M(x, y) \leq d_G(x, y) \leq (1 + \lambda_2)d_M(x, y).$$

In what follows, we present the proofs of the results in Section 4.3.1. In summary, we use the diagram in Figure 4.12 to show the transition relationship between $T_{iso}$, $T^\gamma_{iso}$ and the intermediate distance objects defined.
We start by checking the assumptions in Theorem 11 for Isomap learning on \{X_i\}_{i=1}^n. Since \cal H is geodesically convex, we only need to check the \(\delta\)-sampling condition defined in Theorem 10. We will use the fact that \(X_i = \psi(\tau_i)\) for \(i = 1, \ldots, m\) with \(\tau_i\) uniform on \([0, 1]\).

Let \(\tau_1, \ldots, \tau_m \sim_i U(0, 1)\) and \(\tau(1), \ldots, \tau(m)\) be the corresponding order statistics such that \(\tau(s) > \tau(r)\) if \(s > r\). Then by the fact that \(\cal H\) is quasi-isometric, to satisfy the \(\delta\)-sampling condition it is sufficient to satisfy \(\max_{1 \leq r \leq m-1} \tau(r+1) - \tau(r) \leq u_0\), where \(u_0 = C\delta\) for some constant \(C > 0\). For the purpose of this work, we are interested when \(\delta\) is of order \(O(\log^2 m/\sqrt{m}p_m)\). The following lemma shows that when \(m \to \infty\), the maximum gap between any two adjacent points that are uniformly sampled on \([0, 1]\) is bounded by \(\log^2 m/\sqrt{m}p_m\) with high probability.

**Lemma 11.** Suppose \(\tau_1, \ldots, \tau_m \sim_i U(0, 1)\). If \(m \to \infty\), then the \(\delta\)-sampling condition is satisfied with probability \(\exp(-m/\sqrt{m}p_m})\).
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

Proof. Let $\tau_1, \ldots, \tau_m$ be the order statistics of $\tau_1, \ldots, \tau_m$. We define the gap statistics as $U_r = \tau_{(r+1)} - \tau_{(r)}$, for $1 \leq r \leq m - 1$. Using the fact that for $1 \leq r < s \leq m$,

$$\tau_s - \tau_r \sim \text{Beta}(s - r, m - s + r + 1),$$

we may write down the distribution on $U_r$, i.e. the distance between any two adjacent points that are uniformly sampled on $[0, 1]$ as $U_r \sim \text{Beta}(1, m)$.

Let $U_1, \ldots, U_{m-1}$ denote the order statistics of $U_1, \ldots, U_{m-1}$. Holst [48] showed that $P[U_{(m-1)} \leq u] = P[(m-1)U_{(m-1)} - \log(m-1) \leq (m-1)u - \log(m-1)] \to \exp(-e^{-(m-1)u + \log(m-1)}) = \exp(-\frac{m-1}{e^{(m-1)u}})$ if $m \to \infty$ and $u \to 0$. Letting $u = \log^2 m/\sqrt{mp_m}$ yields $\sup_{1 \leq i \leq m-1} P[|\tau_{(i+1)} - \tau_{(i)}| \leq \log^2 m/\sqrt{mp_m}] = \exp(-\frac{m}{e^{\sqrt{(m \log^2 m)\rho_m}}})$ for large $m$.

Since $\mathcal{H}$ is quasi-isometric, for every $i, j \in (0, 1)$, we observe that $d_{\mathcal{M}}(\psi(\tau_i), \psi(\tau_j)) \leq 2|\tau_i - \tau_j|$. Thus, $\sup_{1 \leq i \leq m-1} P[d_{\mathcal{M}}(\psi(\tau_{(i+1)}), \psi(\tau_{(i)})) \leq 2/\sqrt{mp_m}] \geq \exp(-\frac{m}{e^{\sqrt{(m \log^2 m)\rho_m}}})$.

Therefore, with probability $\exp(-\frac{m}{e^{\sqrt{(m \log^2 m)\rho_m}}})$, the collection $\psi(\tau_1), \ldots, \psi(\tau_m)$ satisfies the $\delta$-sampling condition defined in Lemma [10] which states that for every point $x \in \mathcal{H}$ there exists a point $\psi(\tau_i)$ such that $d_{\mathcal{M}}(\psi(\tau_i), x) \leq \delta$.

Lemma [11] considers the probability that the $\delta$-sampling condition is met with $\delta = \log^2 m/\sqrt{mp_m}$ when samples are chosen uniformly on $\mathcal{H}$. For another random sample selection process, [2] provides a probability for the $\delta$-sampling condition when samples are chosen from a Poisson process.
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

We emphasize that since \( s \) is finite and \( m \to \infty \), the asymptotic analysis of \( n = m + s \) is equivalent to that of \( m \).

Theorem 11 shows \( d_M(X_i, X_j) \approx d_G(X_i, X_j) \), for any \( X_i, X_j \in \mathcal{M} \). Suppose the \( \delta \)-condition of the geodesically convex manifold is satisfied with \( \delta = \log^2 n / \sqrt{n \rho_n} \). We restate Theorem 11 with \( \delta = 1 / \sqrt{n \rho_n} \) in the following corollary:

**Corollary 1.** Let \( \{X_i\}_{i=1}^n \) be the set of data points in \( \mathcal{H} \). If the parameters are chosen as

\[
\epsilon_{\text{min}} = \epsilon_{\text{max}} = \epsilon = O(1/r(n)) = O(\log^2 n/n^{a}\sqrt{\rho_n}),
\]

\[
\delta = \log^2 n / \sqrt{n \rho_n},
\]

then for \( \lambda_1 = O(\log^4 n/n^{2a} \rho_n) \) and \( \lambda_2 = O(\log^2 n/n^{\frac{1}{2}-a} \sqrt{\rho_n}) \), where \( a \in (0, 1/2) \), the relation between \( d_G \) and \( d_M \) may be written as

\[
(1 - \lambda_1)d_M(x, y) \leq d_G(x, y) \leq (1 + \lambda_2)d_M(x, y).
\]

After discussing Isomap on \( \{X_i\}_{i=1}^n \), we now shift our focus to Step 2, where we consider Isomap on \( \{\bar{X}_i\}_{i=1}^n \) and first establish the relation \( d_G(\bar{X}_p, \bar{X}_q) \rightarrow d_G(X_p, X_q) \).

**Lemma 12.** \( |d_G(\bar{X}_p, \bar{X}_q) - d_G(X_p, X_q)| \leq O(r(n)h(n)) \) for \( p, q = 1, \ldots, n \).

**Proof.** Let \( G = (V, E) \) be the graph generated in Isomap on \( \{X_i\}_{i=1}^n \) with \( \epsilon_n \)-ball rule,
where \( V = \{ X_i \}_{i=1}^n \) and each edge \( e(X_i, X_j) \in E(G) \) satisfies \( \| X_i - X_j \| \leq \epsilon_n \). Let 
\( \hat{G} = (\hat{V}, \hat{E}) \) be the graph generated in Isomap on \( \{ \hat{X}_i \}_{i=1}^n \) with \( \hat{\epsilon}_n \)-ball rule, where 
\( \hat{V} = \{ \hat{X}_i \}_{i=1}^n \cup \{ p_0 \} \) and each edge \( e(\hat{X}_i, \hat{X}_j) \in E(\hat{G}) \) satisfies \( \| \hat{X}_i - \hat{X}_j \| \leq \hat{\epsilon}_n \).

For a pair of fixed \( p, q \in \{1, \ldots, n\} \), let \( \gamma \) be the geodesic shortest path connecting \( X_p \) to \( X_q \). The path is denoted by \( X_p = y_0, y_1, \ldots, y_{L-1}, y_L = X_q \), where \( L = O(r(n)) \) since \( G \) contains all edges \( e(X_i, X_j) \) of length \( \| X_i - X_j \| \leq \epsilon_n = \frac{1}{r(n)} \). Each \( y_t \) represents some point \( \in \{ X_i \}_{i=1}^n \) for \( t = 1, \ldots, L-1 \). Then we have 
\[
\begin{align*}
\| \hat{X}_p - X_p \| + \sum_{i=1}^L \| \hat{y}_i - \hat{y}_{i-1} \| + \| \hat{X}_q - X_q \| \\
\leq \| \hat{X}_p - X_p \| + \sum_{i=1}^L \| \hat{y}_i - y_t \| + \sum_{t=1}^L \| \hat{y}_t - \hat{y}_{t-1} \| + \| \hat{X}_q - X_q \| \\
= O(r(n)h(n)) + d_G(X_p, X_q)
\end{align*}
\]

where for each \( y_t \), the \( \hat{y}_t \) represents the corresponding noisy version and is in \( \{ \hat{X}_i \}_{i=1}^n \), for \( t = 1, \ldots, L \).

On the other hand, consider \( \tilde{\gamma} \), the geodesic shortest path connecting \( \tilde{X}_p \) and \( \tilde{X}_q \) in \( \tilde{G} \). The path may be represented by \( \tilde{X}_p = \hat{v}_0, \hat{v}_1, \ldots, \hat{v}_L = \tilde{X}_q \), where \( \hat{v}_t \) represents some point \( \in \{ \hat{X}_i \}_{i=1}^n \). Similarly, \( \tilde{L} = O(1/\hat{\epsilon}_n) = O(r(n)) \). Then we have 
\[
\begin{align*}
d_G(\tilde{X}_p, \tilde{X}_q) = \sum_{i=1}^{\tilde{L}} \| \tilde{v}_i - \tilde{v}_{i-1} \| \\
\end{align*}
\]

Consider \( d_G(X_p, X_q) \), which is the shortest path from \( X_p \) to \( X_q \) on \( G \) and is bounded above by the path distance of any other path,
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

\[ d_G(X_p, X_q) \leq \|\bar{X}_p - X_p\| + \sum_{i=1}^i \|v_i - v_{i-1}\| + \|\bar{X}_q - X_q\| \]
\[ = \|\bar{X}_p - X_p\| + \sum_{t=1}^i \|v_t - v_{t-1}\| + \sum_{t=1}^i \|v_t - \hat{v}_t\| + \|\bar{X}_q - X_q\| \]
\[ = O(r(n)h(n)) + d_G(\bar{X}_p, \bar{X}_q) \]

where for each \( \hat{v}_t \), the \( v_t \) represents the corresponding denoised version and is in \( \{X_i\}_{i=1}^n \), for \( t = 1, \ldots, \hat{L} \).

The difference between \(|d_G(\bar{X}_p, \bar{X}_q) - d_G(X_p, X_q)|\) is then bounded by

\[ |d_G(\bar{X}_p, \bar{X}_q) - d_G(X_p, X_q)| \leq O(r(n)h(n)). \]

\[ \square \]

The above lemma establishes the approximate equality between \( d_G(\bar{X}_p, \bar{X}_q) \) and \( d_G(X_p, X_q) \) as \( n \) tends to infinity. Together with the corollary \( d_G(X_p, X_q) \approx d_M(X_p, X_q) \), we establish that \( d_G(\bar{X}_p, \bar{X}_q) \approx d_M(X_p, X_q) \).

**Corollary 2.** If \( n \to \infty \), \( |d_G(\bar{X}_p, \bar{X}_q) - d_M(X_p, X_q)| \leq O(r(n)h(n)) \) for \( p, q = 1, \ldots, n \).

By applying the triangle inequality on \( |d_G(\bar{X}_p, \bar{X}_q) - d_M(X_p, X_q)| \leq |d_G(\bar{X}_p, \bar{X}_q) - d_G(X_p, X_q)| + |d_M(X_p, X_q) - d_G(X_p, X_q)| \), we achieve the bound in the lemma. Lemma
states that the estimated shortest path distance via noisy samples \( \{\hat{X}_i\}_{i=1}^n \) on \( \hat{G} \) is a good approximation to the true geodesic distance on \( \mathcal{H} \). After tying \( d_G \) to \( d_M \), now we tie \( T_{iso} \) to \( d_M \) through \( d_G \).

In Isomap, after one obtains the shortest path distance matrix \([d_G]\), whose entry \( ij \) is \( d_G(X_i, X_j) \), the final step is to find a low-dimensional representation that best preserves the dissimilar matrix \([d_G]\). The classical method is to apply CMDS to \([d_G]\), and this recovers a low-dimensional embedding such that pairwise Euclidean distances in this new embedding space match the corresponding geodesic distances. The following lemma argues that the low-dimensional representation \( \hat{Z} \) approximately preserves the estimated shortest path distance \([d_G]\), whose entry \( ij \) is \( d_G(\hat{X}_i, \hat{X}_j) \):

**Lemma 13.** If \( n \to \infty \), \( |d_G(\hat{X}_p, \hat{X}_q) - |\hat{Z}_p - \hat{Z}_q| | \leq O(r(n)h(n)) \), for any \( p, q = 1, \ldots, n \).

**Proof.** Let \([d_M]\) denote the matrix whose entry \( ij \) is the exact manifold distance between \( X_i \) and \( X_j \). Suppose for the collection \( \{X_i\}_{i=1}^n \subset \mathbb{R}^D \) there exists an exact low dimensional representation \( \{\tau_i\}_{i=1}^n \subset \mathbb{R}^d \), denoted by \( Z^E \in \mathbb{R}^{n \times d} \), which preserves pairwise shortest path distances in \([d_M]\).

Now, instead of \([d_M]\), we consider the estimated graph distance matrix \([d_G]\). \([d_G]\) may be viewed as a noisy version of \([d_M]\), i.e. \([d_G] = [d_M] + \delta_n F\), where \( \delta_n = O(r(n)h(n)) \) and \( F \) is a symmetric matrix with zero diagonals and entries of order \( \Theta(1) \).

Let \( \hat{Z} \) be the \( d \)-dimensional embedding matrix of \( \hat{X} \) via the application of CMDS to
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFECTION

the distance matrix $[d_G]$. From section 2.3 in Cox [49], we have $\max_{1 \leq i \leq n} |Z_i^E - \hat{Z}_i| \leq O(\delta_n)$, where $Z_i^E$, $\hat{Z}_i$ represent the $i^{th}$ row of $Z^E$ and $\hat{Z}$, respectively.

We now have the main components to show $T_{iso} \rightarrow T_{\mathcal{H}}^c$.

**Lemma 14.** $|T_{\mathcal{H}}^c - T_{iso}| \leq O(r(n)h(n))$ as $n \rightarrow \infty$.

**Proof.** The geodesic distance on the Bézier curve $\mathcal{H}$ may be written as $d_M(X_i, X_j) = \int_{\tau_i}^{\tau_j} \|\psi'(u)\| du$, where $\tau_i = [1, \frac{1}{2}, 0]^T X_i$. Based on the uniform upper bound between $X_i$ and $\tilde{X}_i$, a bound between $\tau_i$ and $\hat{\tau}_i$ may be established as

$$
\max_{1 \leq i \leq m} |\tau_i - \hat{\tau}_i| = \max_{1 \leq i \leq m} |(1, \frac{1}{2}, 0)^T (X_i - \tilde{X}_i)| = O(\|X_i - \tilde{X}_i\|) \leq O(h(n)).
$$

To bound the difference between $T_{\mathcal{H}}^c(\tilde{\tau}, \tau_0)$ and $d_M(X^*, p_0)$, we have

$$
|T_{\mathcal{H}}^c(\tilde{\tau}, \tau_0) - d_M(X^*, p_0)| = \left| \int_{\tilde{\tau}}^{\tau_0} \|\psi'(u)\| du - \int_{\tau^*}^{\tau_0} \|\psi'(u)\| du \right| 
\leq \left| \int_{\tilde{\tau}}^{\tau^*} \|\psi'(u)\| du \right|
\leq 2|\tilde{\tau} - \tau^*| = O(h(n)),
$$

and the last inequality is due to the fact that $\mathcal{H}$ is quasi-isometric with constant $C \leq 2$.

**Lemma 13** establishes $|d_G(X_p, X_q) - |\hat{Z}_p - \hat{Z}_q|| \leq O(r(n)h(n))$, for any $p, q =$
CHAPTER 4. MANIFOLD LEARNING IN NETWORK INFERENCE

$1, \ldots, n$, from which we have $\left| d_G(p_0, \tilde{X}) - |\tilde{Z}_0 - \tilde{Z}| \right| \leq O(r(n)h(n))$. The relation $| \int_{\tilde{h}} \| \psi'(u) \| du | = |\tilde{Z}_i - \tilde{Z}_0| + O(r(n)h(n))$ is built on a series of intermediate steps:

$$T_{\mathcal{H}}^c = | \int_{\tilde{h}} \| \psi'(u) \| du | \approx d_M(p_0, X^*) \approx d_G(p_0, X^*) \approx d_G(p_0, \hat{X}) \approx |\tilde{Z}_0 - \tilde{Z}| = T_{iso}. $$

$\square$
Chapter 5

Discussion

When considering an inference task on high-dimensional data, one common scheme starts with learning a meaningful lower-dimensional representation via an applicable manifold learning algorithm. The bias-variance tradeoff seems to suggest that low-dimensional inference may be more powerful than that performed in the high-dimensional ambient space. However, even this simple “manifold learning for subsequent inference” task is surprisingly unexplored. In this work, we investigate this task in two steps which are formalized by the two questions Q1 and Q2. In Chapters 2 and 3 we consider the inference task in the statistical manifold Θ while in Chapter 4 we consider the inference task in the data manifold X.

For Q1, we study LRTs in Chapter 2 and classification in Chapter 3. The longtime conjecture that restricting the set of alternatives increases the power of an LRT was proved false by a counterexample that restricts the alternative to a smaller square
CHAPTER 5. CONCLUSIONS

area than the unrestricted alternative [9]. As manifold learning typically involves
the reduction of dimensionality, it is inviting to focus on the conjecture when the
submodel has lower dimensions. Chapter 2 provides the Hardy-Weinberg submodel
of the trinomial experiment as a counterexample with a finite sample size. The
phenomenon that the power of the restricted LRT is not uniformly superior to that
of the unrestricted LRT is more common than the Hardy-Weinberg submodel, and
it even exists in linear submodels as presented in Section 2.3.2. After establishing
the LRT power comparison with respect to $Q_1$, we then shift our focus to when the
subsequent inference task is classification and we compare the expected error with
respect to $Q_1$.

Both inference tasks, LRTs and classification, suggest that inference on the true
manifold is not uniformly more powerful than that in the ambient space, at least
for a finite sample size. The fact that for any sample size $n < \infty$ there exist settings
($\alpha, \tau_0, \tau_A$ in LRTs, $p_0, \tau_0, \tau_A$ in classification) for which the restricted model is less
powerful does not imply that we should prefer the unrestricted model. Our examples
displayed in Figure 2.1, 3.1 and 3.2 suggest that the restricted model outperforms the
unrestricted model in more cases and by wider margins than the inverse. However,
one should be more cautious about this compelling phenomenon when conducting an
inference task on a high-dimensional data set.

To answer $Q_2$, we study hypothesis tests that use manifold learning as the first
step, focusing on those in $\Theta$ in Section 2.4 and those in $\mathcal{X}$ in Chapter 4. Both
CHAPTER 5. CONCLUSIONS

sections utilize $\mathcal{H}$, the Hardy-Weinberg curve, as an example. Section 2.4 considers manifold learning with $\{\hat{\theta}_i\}_{i=1}^m$ that are noisy samples of $\{\theta_i\}_{i=1}^m \subset \Theta = \Delta^{D-1}$, and each $\hat{\theta}_i$ as the distribution parameter estimation is in $\Delta^{D-1}$ as shown in Figure 2.2. Chapter 4 considers manifold learning with $\{\hat{X}_i\}_{i=1}^n$ that are noisy samples of $\{X_i\}_{i=1}^n \subset \mathcal{X} = \mathbb{R}^D$, and each $\hat{X}_i$ as the sample estimation is in $\mathbb{R}^D$ as shown in Figure 4.3.

In addition to answering Q2, Chapter 4 also formulates the network structure discovery as a manifold learning problem in the ASE space and builds a practical scheme for identifying meaningful structures in networks. Finally, Section 4.5 studies the tradeoff of dimensionality selection of both ASE space and embedded submanifold, and it will be a good starting point for future work on network structure inference tasks.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


Vita

Mingyue Gao received her B.S. degree in mathematics and applied mathematics from Dalian University of Technology, China in 2012. After that, she moved to the US and received the M.A. degree in mathematics from Wake Forest University, NC in 2014 and enrolled in the Applied Mathematics and Statistics Ph.D. program at Johns Hopkins University in 2014. Her research interests include subsequent inference in manifold learning, network analysis and deep learning applications in time series.