Distance-Based Methods for Detecting Associations in Structured Data with Applications in Bioinformatics

A thesis presented for the degree of

Doctor of Philosophy of Imperial College London

by

Christopher Minas

Department of Mathematics

Imperial College

180 Queen's Gate

London SW7 2BZ

United Kingdom

March 12, 2013

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Christopher Minas

Copyright

To all those who have not had the opportunities I have been blessed with.

Abstract

In bioinformatics applications samples of biological variables of interest can take a variety of structures. For instance, in this thesis we consider vector-valued observations of multiple gene expression and genetic markers, curve-valued gene expression time courses, and graph-valued functional connectivity networks within the brain. This thesis considers three problems routinely encountered when dealing with such variables: detecting differences between populations, detecting predictive relationships between variables, and detecting association between variables.

Distance-based approaches to these problems are considered, offering great flexibility over alternative approaches, such as traditional multivariate approaches which may be inappropriate. The notion of distance has been widely adopted in recent years to quantify the dissimilarity between samples, and suitable distance measures can be applied depending on the nature of the data and on the specific objectives of the study. For instance, for gene expression time courses modeled as time-dependent curves, distance measures can be specified to capture biologically meaningful aspects of these curves which may differ. On obtaining a distance matrix containing all pairwise distances between the samples of a given variable, many distance-based testing procedures can then be applied. The main inhibitor of their effective use in bioinformatics is that p-values are typically estimated by using Monte Carlo permutations. Thousands or even millions of tests need to be performed simultaneously, and time/computational constraints lead to a low number of permutations being enumerated for each test.

The contributions of this thesis include the proposal of two new distance-based statistics, the DBF statistic for the problem of detecting differences between populations, and the GRV coefficient for the problem of detecting association between variables. In each case approximate null distributions are derived, allowing estimation of p-values with reduced computational cost, and through simulation these are shown

to work well for a range of distances and data types. The tests are also demonstrated to be competitive with existing approaches. For the problem of detecting predictive relationships between variables, the approximate null distribution is derived for the routinely used distance-based pseudo F test, and through simulation this is shown to work well for a range of distances and data types. All tests are applied to real datasets, including a longitudinal human immune cell *M. tuberculosis* dataset, an Alzheimer's disease dataset, and an ovarian cancer dataset.

Acknowledgements

First and foremost I would like to thank my supervisor, Dr Giovanni Montana, without whose guidance this thesis would not have come to fruition. I also wish to thank the Engineering and Physical Sciences Research Council for their generous funding allowing me to study towards the PhD, and the IT support staff of the Mathematics department for never letting me be without my tools. Thanks also to Dr Simon Waddell and Dr Edward Curry for supplying us with interesting datasets and biological insights.

Many thanks go to my fellow office-dwellers and friends for their jokes, tricks, conversations and support. In no particular order, I would like to thank Orlando Doehring, Matt Silver, James Martin, Anna Fowler, Todd Kuffner, Swati Chandna, Dean Bodenham, Georg Hahn, Paul Ginzberg, Elena Ehrlich, Din-Houn Lau, Da Ruan, Aidan O'Sullivan, Eva Janousova, Maria Vounou, Ioannis Phinikettos, Edward Cohen, Brian McWilliams, Maurice Berk, Badr Missaoui, Simon Smith, Mohammed Morshedi and Alexis Dritsas.

Special thanks are due to my family for their tireless love and support. My parents, Maria and Andreas, grandparents, Stavrolilia and Krinos, and aunt, Nona, have always impressed upon me the importance of studying and given me the emotional support needed on many occasions. My godfather, 'uncle Andreas', has always guided me and provided extra support when things have not gone so well. My brothers, Anthony and George, and sister, Marina, have also always been there for extra support in there own special ways.

Finally, I wish to thank my lovely fiancée, Michelle Lakeridou, for her never-ending encouragement and inspiration, and her family for their support and always positive outlook.

Christopher Minas

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The work presented in this thesis is motivated by statistical problems commonly arising in bioinformatics, which describes the study of data obtained from biological experiments. The overall aim of such experiments is to understand the complex mechanisms governing particular biological processes, such as susceptibility to disease.

Three problems commonly encountered in the bioinformatics literature can be stated as follows:

(i) Detecting differences between populations: The interest is in detecting if the behaviour of random variable $\mathcal{Y}$ alters in different populations. $\mathcal{Y}$ is observed on sampling units representing the different populations, and the membership of sampling units to a population can be captured by discrete-valued variable $\mathcal{X}$.

(ii) Detecting predictive relationships between variables: The interest is in detecting if the behaviour of random variable $\mathcal{Y}$ can be predicted by the behaviour of random variable $\mathcal{X}$.

(iii) Detecting association between variables: The interest is in detecting if the behaviour of random variables $\mathcal{Y}$ and $\mathcal{X}$ is associated.

The biological variables of interest in these problems can range from scalar-valued gene expression to curve-valued gene expression time courses and graph-valued functional connectivity networks (comprised of nodes connected by edges); see Table 1.1 for more examples. For scalar- and vector-valued variables (note that a vector-valued random variable is also referred to as a random vector, i.e., a vector whose elements

are scalar-valued random variables), traditional multivariate approaches are applied to model the variation exhibited by these variables. For curve- and graph-valued variables, specialized methods are needed to capture and model the complex variational properties (see for example, Liu and Yang (2009) and Lord *et al.* (2011)).

In this thesis we embrace the notion of distance to model the variation of a given variable, regardless of its nature; it can be scalar-, vector-, curve- or graph-valued, that is, the observations may be single values, vectors, functions (curves) or graphs, respectively. For N observations of the given variable, the idea is to compute the $N \times N$ distance matrix harbouring the distances between all pairwise combinations of observations, and use this within the statistical testing framework instead of the original observations themselves. This offers several advantages over existing methods. For instance, for observations of any type, many distances are available, and each can be associated with a different biological meaning. We have demonstrated this for curve-valued observations in our published article Minas *et al.* (2011), the material of which is included in this thesis. Furthermore, distance-based testing procedures are typically routed in well-known traditional methods, as demonstrated in the coming chapters. They therefore generalize traditional approaches, which fosters their understanding, making them more accessible to biologists who may not fully understand the specialized approaches. Lastly, on computing the distance matrix, many complementary methods are immediately available, such as clustering and visualization (Pekalska and Duin, 2005). These can be used to supplement any statistical analyses performed.

Distance-based approaches have been proposed in the bioinformatics literature and elsewhere for the problems of interest in this thesis. However, they all suffer from limitations which inhibit their effective use in bioinformatics. The biggest impediment is that the distribution of distance-based statistics under the null hypotheses of interest are unknown (Mantel, 1967; McArdle and Anderson, 2001). In real applications, the null distribution is commonly approximated by a discrete distribution generated by using Monte Carlo permutations to shuffle the observations and recompute the statistic, i.e., a permutation distribution. P-values are estimated from this distribution as the proportion of permuted statistic values as extreme or more extreme than the observed statistic value (see, for instance, Beckmann *et al.* (2005), Wessel and Schork (2006) and Salem *et al.* (2010)).

Table 1.1: Biological variables of interest which have been considered in the literature under each statistical problem, with the nature of these variables stated in brackets, and example articles in which they have been considered. The following terminology/notation has been used: time-dependent gene expression, gene expression at different time-points over a given time-range; dose-dependent gene expression, gene expression at different dosage levels of a given treatment; SNPs, single nucleotide polymorphisms.

Statistical Problem	$\mathcal{Y}$		$\mathcal{X}$		Example Study
Detecting differences between populations	time-dependent gene expression	(curve-valued)	population-membership	(scalar-valued)	Storey *et al.* (2005b)
	multiple SNPs	(vector-valued)	population-membership	(scalar-valued)	Wu *et al.* (2010)
	functional connectivity network	(graph-valued)	population-membership	(scalar-valued)	Lord *et al.* (2011)
Detecting predictive relationship of $\mathcal{X}$ on $\mathcal{Y}$	gene expression	(scalar-valued)	multiple SNPs	(vector-valued)	Storey *et al.* (2005a)
	expression of multiple genes	(vector-valued)	age and sex	(vector-valued)	Zapala and Schork (2006)
	expression of multiple genes	(vector-valued)	multiple SNPs	(vector-valued)	Wessel *et al.* (2007)
	dose-dependent gene expression	(curve-valued)	multiple SNPs	(vector-valued)	Salem *et al.* (2010)
Detecting association between $\mathcal{Y}$ and $\mathcal{X}$	expression of multiple genes	(vector-valued)	multiple SNPs	(vector-valued)	Wessel and Schork (2006)
	voxels of brain	(vector-valued)	multiple SNPs	(vector-valued)	Vounou *et al.* (2010)

Permutation approaches, however, are computationally intensive and introduce sampling errors (Berry and Mielke, 1983). Moreover, whereas large p-values can be well-approximated by a Monte Carlo approach, smaller ones will be estimated less accurately (Mielke and Berry, 2007; Knijnenburg *et al.*, 2009). In particular, it has been shown that in order to obtain a permutation p-value within 10^{-5} of the true p-value, $O(10^7)$ permutations are required. We have observed that in real applications a much smaller number of permutations are performed, giving rise to concerns about the accuracy of the resulting p-values.

This is a major issue when many tests are simultaneously performed, as is typically the case in bioinformatics; hundreds of thousands of genes are tested in gene expression experiments (see, for instance, Storey *et al.* (2005b)) and millions of genetic markers such as single nucleotide polymorphisms (SNPs) are tested in genome-wide association (GWA) studies (see, for instance, Vounou *et al.* (2010)). In such experiments biological variables identified as significant are pushed forward for further analysis, which of course depletes resources of time and money. Type I errors therefore need to be minimized in the testing phase, and much work has been done using multiple-testing corrections on observed p-values to achieve this (see, for instance, Benjamini and Hochberg (1995), Reiner *et al.* (2003) and Storey and Tibshirani (2003)).

When using Monte Carlo permutations to estimate p-values in such cases, the total number of permutations per test may be vastly limited due to time/computational constraints. This has been shown to increase familywise type I error rates (Phipson and Smyth, 2010). This is because for N_π permutations, p-values of zero are obtained with probability $1/(N_\pi + 1)$ under the null hypothesis (Phipson and Smyth, 2010). Thus even if no tests should be significant, a larger number of tests will yield permutation p-values of zero and hence be deemed significant regardless of multiple-testing corrections and significance levels applied. For example, suppose $N_\pi = 10^3$ Monte Carlo permutations are performed for a test of the null hypothesis of equality between populations for each of 10^5 genes. The expected number of genes with corresponding permutation p-values of zero will be $10^5/(1001) \approx 100$, even when no genes are responsible for differential actions between populations.

Therefore, for distance-based approaches, minimizing type I error rates equates to estimating small p-values more accurately, i.e., without using Monte Carlo permuta-

tions, before applying multiple-testing corrections. For the distance-based approaches considered for each of the three problems of interest in this thesis, computationally cheap estimation of small p-values is a recurring theme.

1.1 Summary of Contributions

The contributions of this thesis by problem are:

(i) Detecting differences between populations: On applying distances to the observations of $\mathcal{Y}$, a distance-based variance decomposition which generalizes the multivariate analysis of variance (MANOVA) decomposition is derived. A distance-based F (DBF) statistic is derived from this, and its null distribution is approximated by a continuous distribution using moment matching, allowing p-values to be estimated without expensive permutations for any distance. The exact moments of the permutation distribution which would be generated using all possible permutations are used for this result, and are obtained analytically by applying the results of Kazi-Aoual *et al.* (1995). The DBF test is shown to generalize some MANOVA testing procedures, and simulations are provided to support this claim. This test is applied to a case-control GWA study of Alzheimer's disease, and to our knowledge provides the first distance-based case-control study of this disease.

(ii) Detecting predictive relationships between variables: The pseudo F statistic of McArdle and Anderson (2001) generalizes a statistic proposed for testing a null hypothesis of zero-valued regression coefficients within a multivariate multiple linear regression framework. It is routinely used with permutations in bioinformatics applications, where distances are applied to the observations of $\mathcal{Y}$ and $\mathcal{X}$ is vector-valued. We approximate its null distribution by a continuous distribution using moment matching, allowing p-values to be estimated without permutations for any distance. For this result expressions for the exact moments of the permutation distribution generated by using all permutations are derived as no suitable results exist in the literature. The pseudo F test using the approximate null distribution is applied to a candidate-phenotype GWA study of Alzheimer's disease,

which to our knowledge provides the first distance-based regression analysis of such studies.

(iii) Detecting association between variables: We propose a new test statistic, the generalized RV (GRV) statistic, where distances are applied to both the observations of $\mathcal{Y}$ and $\mathcal{X}$, and derive its approximate null distribution by moment matching; the moments are obtained analytically by applying the results of Kazi-Aoual *et al.* (1995). Simulation experiments are performed to demonstrate competitiveness with the well-known distance-based standardized Mantel test, and better performance for specific experimental setups. We also show theoretically and through simulation that the GRV test generalizes two well-established multivariate approaches to the problem; the RV test of correlation between random vectors of Escoufier (1973), and the distance correlation (dCor) test of dependence between random vectors of Székely *et al.* (2007). The GRV test is applied to a candidate-phenotype GWA study of Alzheimer's disease and a gene expression quantitative trait loci (eQTL) mapping study of ovarian cancer, providing, to our knowledge, the first fully distance-based analyses of such studies.

1.2 Thesis Structure

The thesis is partitioned into three parts; Part I: Background Literature is comprised of Chapters 2, 3 and 4, Part II: Methodology is comprised of Chapters 5, 6 and 7, and Part III: Applications is comprised of Chapters 8 and 9. See Figure 1.1 for a schematic of the thesis structure.

Part I: Background Literature contains separate reviews for each of the three problems. Each is comprised of the problem statement and existing approaches in the multivariate and distance-based settings, after which the limitations of existing distance-based approaches are highlighted. Chapter 2 considers the problem of detecting differences between populations. Chapter 3 considers the problem of detecting predictive relationships between two random vectors. In the multivariate setting the variables are vector-valued but in the distance-based setting only the predictor variable is strictly vector-valued. Chapter 4 considers the problem of detecting association between two random vectors. In the multivariate setting the variables are vector-valued

but in the distance-based setting the variables can be of any type.

Part II: Methodology contains the main contributions for each problem. In Chapter 5 we derive the DBF statistic to test a null hypothesis of equality between populations, and describe the permutation approach to estimate p-values. This was the initial approach used to assess significance, and has been published in Minas *et al.* (2011). The approximate null distribution of the statistic detailed in this chapter was developed after our publication. Theoretical connections are made with traditional multivariate methods for vector-valued variables, and simulations are also provided. For a range of data types and distance measures we demonstrate the applicability of the DBF test on simulated data, in addition to real data. Finally, two power studies are performed, one demonstrating the competitive performance of the DBF test with existing approaches suitable for curve-valued variables, and one demonstrating the computational advantage of using the approximate distribution over permutations for several distances and data types.

In Chapter 6 we derive the approximate null distribution of the pseudo F statistic used to test a null hypothesis of no predictive relationship between two variables. This distribution is shown to be applicable for a range of distance measures and data types, on simulated and real data.

In Chapter 7 we derive the GRV statistic to test a null hypothesis of no association between two variables. The approximate null distribution is derived and through simulation and real data examples it is shown to be applicable for a range of data types and distance measures. For vector-valued variables theoretical connections with the RV and dCor tests are discussed, and simulations are performed to demonstrate these. Competitiveness with other distance-based tests is demonstrated through power studies, including the standardized Mantel test.

Part III: Applications contains two chapters in which the different distance-based tests proposed are applied to real datasets. In Chapter 8 we provide a brief review of existing approaches used in GWA studies of Alzheimer's disease, emphasizing the distance-based approaches (or lack of). We present the findings from using the DBF test to perform a case-control study, and the pseudo F and GRV tests to perform candidate-phenotype studies.

In Chapter 9 we describe microarray gene expression studies, and pay particular

attention to two variants; longitudinal microarray time course studies and eQTL mapping studies. For longitudinal microarray time course studies we provide a brief review of existing methods, and highlight some key limitations. These limitations have been presented in our article Minas *et al.* (2011), and justify the use of different distances for longitudinal microarray time course analysis. The findings of a differential analysis of *M.tuberculosis* performed using the DBF test with several distance measures and using the permutation approach to perform inference are then presented (these have also been published in Minas *et al.* (2011)). For eQTL mapping studies we provide a brief review of existing methods and emphasize distance-based approaches which have been proposed. Using the GRV test we perform an eQTL pathway analysis mapping of ovarian cancer and present the findings.

Conclusions and directions for further work are presented in Chapter 10.

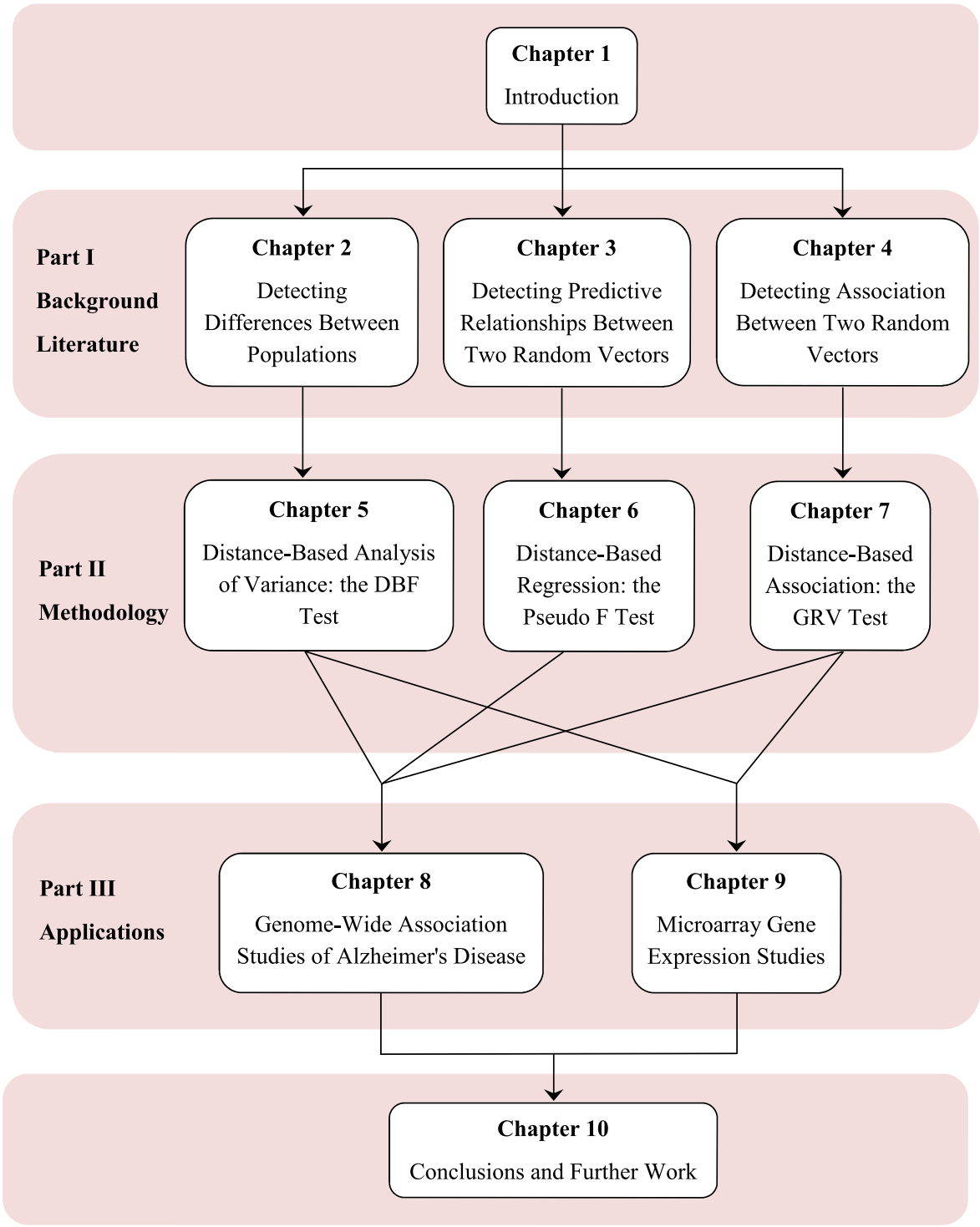

Figure 1.1: Schematic of the thesis structure. The problem statements and corresponding reviews for the problems of interest are contained in Part I: Background Literature. New methodology for each problem is contained in Part II: Methodology, with the arrows from the review chapters indicating which methodological chapters are related to which problem. Arrows from the methodological chapters indicate the study in Part III: Applications where the methodology has been applied. Conclusions and directions for further work are presented in Chapter 10.

Part I

Background Literature

Chapter 2

Detecting Differences Between Populations

In this chapter we introduce the problem of detecting differences between populations. We begin by reviewing the problem in the classical multivariate framework where observations are assumed to be real and vector-valued, and describe traditional multivariate analysis of variance methods. For observations which do not conform to the required assumptions of these methods, we review distance-based approaches which can be applied. The chapter concludes with a summary of the limitations of the existing distance-based methods.

2.1 Multivariate Approaches

2.1.1 Problem Statement

Consider N independent observations of Q-dimensional real-valued random vector $\mathcal{Y} = (\mathcal{Y}_1, \ldots, \mathcal{Y}_Q)^T$, given by $\{\boldsymbol{y}_i\}_{i=1}^N$. Assume that these observations belong to one of G populations with means $\{\boldsymbol{\mu}_g\}_{g=1}^G$, each of size N_g such that $N = \sum_{g=1}^G N_g$. The null hypothesis of interest is typically stated as

$$H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_G, \tag{2.1}$$

(Mardia *et al.*, 1979). Under this hypothesis there is equality between the means of the populations from which the observations are drawn. The alternative hypothesis is

that the means are not equal for at least one group.

2.1.2 Traditional Multivariate Analysis of Variance

Traditional approaches to testing (2.1) are based on the multivariate analysis of variance (MANOVA). This is the approach whereby the total variance exhibited by the N observations is partitioned into within- and between-group variance. This is achieved as follows.

Define the overall sample mean by $\bar{\boldsymbol{y}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i$ and each within-group sample mean by $\bar{\boldsymbol{y}}_g = \frac{1}{N_g}\sum_{i=1}^{N} \boldsymbol{y}_i I_{gi}$ for $g = 1, \ldots, G$, where I_{gi} is an indicator variable taking the value 1 if observation $\boldsymbol{y}_i$ is in group g and 0 otherwise. The $Q \times Q$ covariance matrix of $\mathcal{Y}$ is estimated by

$$\boldsymbol{S} = \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T .$$

The total variance of the observations can be quantified by $\mathrm{tr}(\boldsymbol{S})$, that is, by the summation of the variance of each of the Q variables comprising $\mathcal{Y}$. Typically in multivariate analysis, the $Q \times Q$ total sum of squares matrix given by

$$\boldsymbol{T} = \sum_{i=1}^{N} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T ,$$

is used to represent the information in $\boldsymbol{S}$, since it is within a multiplicative factor (Mardia *et al.*, 1979). The quantity $\mathrm{tr}(\boldsymbol{T})$ is defined as the total sum of squares, and is a multivariate analogue of the sum of squares used for centered univariate observations (Anderson, 2001). Since this is related to the total variance, we refer to $\mathrm{tr}(\boldsymbol{T})$ as the variability, to distinguish from the variance, as they both provide information about scatter from the mean. The total sum of squares matrix can be expressed as the sum of between- and within-group sum of squares matrices,

$$\boldsymbol{B} = \sum_{g=1}^{G} N_g (\bar{\boldsymbol{y}}_g - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_g - \bar{\boldsymbol{y}})^T \quad \text{and} \quad \boldsymbol{W} = \sum_{g=1}^{G}\sum_{i=1}^{N} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_g)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_g)^T I_{gi},$$

respectively. That is, $\boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W}$.

MANOVA test statistics make use of different elements of this total sum of squares decomposition to quantify differences in the amount of variability explained by the G groups. For $G = 2$, the well-known Hotelling's T^2 statistic is given by

$$T^2 = \frac{N_1 N_2 (N-2)(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^T \boldsymbol{W}^{-1} (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)}{N}.$$

Larger values of this statistic provide evidence against the null. For $G > 2$, MANOVA statistics include Wilks' Λ,

$$\Lambda = \det(\boldsymbol{W}) / \det(\boldsymbol{T}),$$

the Lawley-Hotelling trace,

$$\text{LH} = \text{tr}\left(\boldsymbol{W}^{-1}\boldsymbol{B}\right),$$

and the Pillai trace,

$$\text{PT} = \text{tr}\left(\boldsymbol{T}^{-1}\boldsymbol{B}\right)$$

(Rencher, 2002; Krzanowski, 2000). Wilks' Λ uses the ratio of the determinants of the within-group and total sum of squares matrices to yield a measure of the proportion of variability in the given dataset explained by the within-group sum of squares. A small statistic value indicates that the within-group variability accounts for a small proportion of the total variability, meaning that the between-group variability accounts for the remaining large proportion. Thus evidence against the null is provided. The Lawley-Hotelling trace considers the matrix generalization of the fraction of two scalar values, by multiplying the inverse of $\boldsymbol{W}$ by $\boldsymbol{B}$. The trace of this quantifies how much greater the effect of $\boldsymbol{B}$ is than $\boldsymbol{W}$, such that larger values provide evidence against the null. The Pillai trace similarly uses the trace operator, but compares the between-group to the total sum of squares. Again, larger values of this statistic provide evidence against the null.

Under the assumption that the observations are independent and identically distributed from a Multivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, some distributional results are available. For instance, for $G = 2$, Hotelling's T^2 statistic multiplied by a constant depending on N and Q has an exact F distribution under the null;

$$\frac{N-Q-1}{(N-2)Q} T^2 \sim F_{Q,N-Q-1},$$

where $F_{Q,N-Q-1}$ denotes the F distribution with degrees of freedom Q and $N-Q-1$. For $G > 2$, Wilks' Λ, the Pillai trace and the Lawley-Hotelling trace statistics can all be similarly transformed to statistics which are well-approximated by the F distribution with degrees of freedom dependent on N, G and Q (see, for example, Rencher (2002)).

When $N < Q$, as is typically the case with genomic datasets, the classical MANOVA tests cannot be applied directly. This is because the $\boldsymbol{T}$ and $\boldsymbol{W}$ matrices are singular, and so have at least one zero-valued eigenvalue and cannot be inverted. Several high-dimensional MANOVA settings with $N < Q$ have been considered in the literature, and tests of equality between groups have been proposed, some using traditional MANOVA statistics with generalized inverses (see Srivastava (2007) and Schott (2007) for good reviews, and Tsai and Chen (2009) for an application to gene expression data).

One of the first tests was proposed by Dempster for $G = 2$ (Dempster, 1960), where an F-type statistic defined as the ratio of within- to between-group variability was proposed. This statistic was generalized for $G > 2$ and named the Dempster trace criterion four decades later by Fujikoshi *et al.* (2004). They noticed that the trace operator could be applied to the $\boldsymbol{B}$ and $\boldsymbol{W}$ sum of squares matrices to yield equivalent expressions to those proposed by Dempster. Although not stated explicitly, they use the sum of squares of each of the Q variables, i.e., $\text{tr}(\boldsymbol{T})$, to represent the total variability of the N observations. They then partition this into between- and within-group components via

$$\text{tr}(\boldsymbol{T}) = \text{tr}(\boldsymbol{B}) + \text{tr}(\boldsymbol{W}), \tag{2.2}$$

which follows by applying the trace operator to the decomposition $\boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W}$. The Dempster trace criterion is then defined as

$$\frac{\text{tr}(\boldsymbol{B})}{\text{tr}(\boldsymbol{W})}, \tag{2.3}$$

and a transformation of this statistic is shown to be asymptotically Gaussian.

While MANOVA approaches are applicable in a wide range of scientific areas, including genomics (Szabo *et al.*, 2003; Tsai and Chen, 2009; Shen *et al.*, 2011), they may be inappropriate for at least two main reasons. Firstly, when the observations are multivariate, the multivariate normality assumption may not necessarily hold, e.g.,

if the observations are heavily skewed or discrete-valued, such as the genetic data typically encountered in GWA studies (see, for instance, Wu *et al.* (2010) and Chapter 8). In such studies minor allele counts are observed for hundreds of thousands of genetic markers across the genome. Secondly, observations may not be represented by vectorial data structures. In an increasingly large number of applications they are functional (i.e., curves), as in longitudinal microarray time course studies (see Chapter 9). In such studies gene expression measurements are observed over time, and are modeled as smooth functions of time (Berk and Montana, 2009; Berk *et al.*, 2012). Observations can also be graph-valued, such as trees and networks; they are used in neuroimaging studies to model functional connectivity between regions of the brain (Rubinov and Sporns, 2010). In such cases we can turn to distance-based methods to test for equality between populations, where only distances between observations are required.

2.2 Distance-Based Approaches

2.2.1 Problem Statement

Consider N independent observations $\{y_i\}_{i=1}^N$ of random variable $\mathcal{Y}$ belonging to G groups with means $\{\mu_g\}_{g=1}^G$. We place no restriction on the nature of these observations; they can be of any form, for example, scalar-, vector-, curve- or graph-valued. The fundamental assumption is that we are able to define a distance measure $d_{\mathcal{Y}}(\cdot,\cdot)$ which quantifies the dissimilarity between any pair of observations in the random sample (note the terms 'distance' and 'dissimilarity' are used interchangeably). It is further assumed that $d_{\mathcal{Y}}$ is either semi-metric or metric. It is semi-metric if it satisfies the properties of identity ($\{d_{\mathcal{Y}}(y_i,y_i)=0\}_{i=1}^N$), non-negativity ($\{d_{\mathcal{Y}}(y_i,y_j)\geq 0\}_{i,j=1}^N$), and symmetry ($\{d_{\mathcal{Y}}(y_i,y_j)=d_{\mathcal{Y}}(y_j,y_i)\}_{i>j}$) (see, for example, Mardia *et al.* (1979)). If it additionally satisfies the triangle inequality, that is, $d_{\mathcal{Y}}(y_i,y_j)\leq d_{\mathcal{Y}}(y_i,y_k)+d_{\mathcal{Y}}(y_k,y_j)$ for $i,j,k=1,\ldots,N$, then $d_{\mathcal{Y}}$ is metric.

The choice of distance depends on the type of data and scientific problem at hand. Having chosen a suitable distance, arrange all pairwise distances in the $N\times N$ distance

matrix $\mathbf{\Delta}_{\mathcal{Y}} = \{d_{\mathcal{Y}}(y_i, y_j)\}_{i,j=1}^{N}$. We are then interested in testing the null hypothesis

$$H_0 : d_{\mathcal{Y}}(\mu_i, \mu_j) = 0, \tag{2.4}$$

for all $i \neq j \in \{1, \ldots, G\}$. Here, there is equality between the group means with respect to the chosen distance measure $d_{\mathcal{Y}}$. The alternative hypothesis is that a difference exists between any two groups.

In the literature there are two main distance-based methods suitable for testing (2.4) which have been applied in bioinformatics applications. These are the multi-response permutation procedure (MRPP) of Mielke *et al.* (1976) and the Mantel test of Mantel (1967), which are described in detail below. In ecology methods also include ANOSIM which is based on ranks (Legendre and Legendre, 1998), and a non-parametric analysis of variance approach (Anderson, 2001). These are not described here.

2.2.2 The Multi-Response Permutation Procedure Test

The MRPP statistic (Mielke *et al.*, 1976; Mielke and Berry, 2007) is formulated as the weighted sum of within-group distances. In particular, it is defined as

$$\delta(\mathbf{\Delta}) = \sum_{g=1}^{G} \frac{c_g}{N_g(N_g - 1)} \sum_{k<j} d_{\mathcal{Y}}^2(y_k, y_j) I_{gk} I_{gj},$$

where $\{c_g\}_{g=1}^{G}$ are positive weights such that $\sum_{g=1}^{G} c_g = 1$. The weights can be chosen to reflect the type of averaging required by the practitioner. Examples include $\{c_g = 1/N_g\}_{g=1}^{G}$ reflecting the view that each group contributes equally to the overall statistic, $\{c_g = N_g/N\}_{g=1}^{G}$, indicating that larger groups contribute more to the overall statistic, and $\{c_g = (N_g - 1)/(N - 2)\}_{g=1}^{G}$, magnifying the contribution provided by larger groups and damping the contribution provided by smaller groups. The idea of this non-negative statistic is to provide a measure of the within-group variability, so that small values are indicative of groupings containing similar observations.

Inference is typically performed by using a Monte Carlo permutation procedure where the observed statistic, denoted $\hat{\delta}(\mathbf{\Delta})$, is compared against a permutation sampling distribution generated under the null. This is achieved by defining a set of

N_π Monte Carlo permutations $\pi \in \Pi$, where each π is a one-to-one mapping of the set $\{1, \ldots, N\}$ to itself. For each permutation, the observed permuted statistic is computed as $\hat{\delta}(\boldsymbol{\Delta}_\pi)$, where $\boldsymbol{\Delta}_\pi$ denotes the distance matrix with rows and columns simultaneously permuted by π. The set $\{\hat{\delta}(\boldsymbol{\Delta}_\pi)\}_{\pi \in \Pi}$ then defines the sampling distribution under the null. The p-value of the observed $\hat{\delta}(\boldsymbol{\Delta})$ can then be approximated by

$$\frac{\#(\hat{\delta}(\boldsymbol{\Delta}_\pi) \leq \hat{\delta}(\boldsymbol{\Delta}))}{N_\pi}.$$

Note that this is a left-tailed test since smaller values of $\hat{\delta}(\boldsymbol{\Delta})$ indicate smaller within-group variability.

An alternative approach has been proposed where the exact permutation distribution which would be obtained by using all $N!$ permutations is approximated by the Pearson type III distribution (Mielke and Berry, 2007). This is a skewed distribution which includes the Normal and Chi-squared distributions as special cases, and is thus able to capture skewness which may be observed in the sampling distribution.

2.2.3 The Mantel Test

The Mantel statistic was originally proposed by Mantel (1967) to test a null hypothesis regarding association between two distance matrices. This testing paradigm is discussed in Chapter 4, where details of the Mantel statistic and corresponding testing procedure are also provided.

In the context of testing (2.4), a form of the Mantel statistic can be used which requires specification of a model matrix encoding group membership (Legendre and Legendre, 1998). This matrix, which we denote $\boldsymbol{M}$, is a symmetric $N \times N$ matrix with binary-valued elements $\{m_{ij}\}_{i,j=1}^N$ such that $m_{ij} = 1$ if y_i and y_j are in different groups and $m_{ij} = 0$ if they are in the same group. This matrix thus corresponds to the alternative hypothesis that the group means are dissimilar, as the within-group portions of the matrix are set to zero while the between-group portions are non-zero. The Mantel statistic is then given by

$$M(\boldsymbol{\Delta}) = \sum_{i<j} d_{\mathcal{Y}}(y_i, y_j) m_{ij}.$$

Since the elements m_{ij} take the value of 1 only for between-group distances, $M(\boldsymbol{\Delta})$ is a weighted sum of the between-group distances. It therefore provides a measure of between-group variability. It is non-negative, and large values provide evidence against the null hypothesis.

Inference can be performed by using Monte Carlo permutations as with the MRPP test. For N_π permutations $\pi \in \Pi$, the p-value of an observed $\hat{M}(\boldsymbol{\Delta})$ is estimated by

$$\frac{\#(\hat{M}(\boldsymbol{\Delta}_\pi) \geq \hat{M}(\boldsymbol{\Delta}))}{N_\pi},$$

which is a right-tailed test. For large N, Mantel (1967) has provided a Normal approximation for this sampling distribution using the exact mean and variance that would be obtained by using all possible permutations. However, its use has been cautioned where the sampling distribution exhibits non-normal, i.e., skewed, tendencies (Mantel, 1967).

2.3 Summary

We have reviewed the traditional MANOVA approaches used to test null hypothesis (2.1) of equality between populations. For many applications of interest in bioinformatics, the data is of the form of discrete-, curve-, or graph-valued objects, for which MANOVA approaches are inappropriate. Distance-based approaches can be used instead, requiring only a suitably defined distance measure between observations of any type. For the corresponding distance-based null hypothesis (2.4) of equality between populations, we have reviewed the MRPP and Mantel tests.

Although these tests have been applied in bioinformatics applications, they suffer from a few limitations. Firstly, the MRPP and Mantel statistics are not suitably interpretable. The MRPP statistic, for example, may yield small values when between-group distances are small. That is, no consideration is made for the between-group distances, and hence true clustering effects of the distances across the groups cannot be highlighted by the statistic alone. It is only through permutations of the sampling units across groups that any clustering effects will become apparent, since the within-group distances will change each time. Similarly, the Mantel statistic does not consider within-group distances, so while between-group distances may be large, the within-

group distances may also be large in comparison. Thus, these statistics alone do not provide a direct overview of all the distance information provided. A test statistic for null hypothesis (2.4) should be interpretable and give a summary of all available distance information.

Secondly, in drawing inferences in real applications, Monte Carlo permutations seem to be applied rather than using the distributional approximations which exist (see, for instance, Reiss *et al.* (2009) and Beckmann *et al.* (2005)). In addition, a small number of permutations are used. For instance, the MRPP test has been applied with $O(10^4)$ permutations to vector-valued neuroimaging data with $N = 38$ samples (Reiss *et al.*, 2009), and the Mantel test has been applied with $O(10^3)$ permutations to discrete-valued genetic polymorphism data with $N = 500$ samples (Beckmann *et al.*, 2005). Using approximations of the null sampling distribution instead of permutations will yield more accurate p-value estimates, in addition to reducing the computational cost of performing inferences.

In Chapter 5 we propose a distance-based F (DBF) test for testing null hypothesis (2.4) with an approximate null distribution. The DBF statistic is more interpretable than the MRPP and Mantel statistics, and the approximate null distribution allows p-values to be estimated without permutations and hence with a dramatically reduced computational cost.

Chapter 3

Detecting Predictive Relationships Between Two Random Vectors

In this chapter we introduce the problem of detecting predictive relationships between two variables, a response variable and a predictor variable, in a linear regression setting. The problem is first reviewed from the multivariate perspective, where both variables are vector-valued, i.e., random vectors. For response observations of other types, we review a distance-based regression approach which can be applied. A summary of limitations of the distance-based method concludes the chapter.

3.1 Multivariate Approach

3.1.1 Problem Statement

Consider explaining the Q-dimensional random vector $\mathcal{Y} = (\mathcal{Y}_1, \ldots, \mathcal{Y}_Q)^T \in \mathbb{R}^Q$ comprised of scalar-valued variables $\{\mathcal{Y}_q\}_{q=1}^Q$ in terms of the M-dimensional random vector $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_M)^T \in \mathbb{R}^M$ comprised of scalar-valued variables $\{\mathcal{X}_m\}_{m=1}^M$. Each of the scalar-valued variables comprising $\mathcal{X}$ are referred to as predictor variables and are independent to each other. The variables comprising $\mathcal{Y}$ are referred to as the response or dependent variables (see, for example, Rencher (2002)).

The problem entails modeling the response variables as a linear function of the predictor variables, i.e., setting up a multivariate multiple linear regression (MMLR) model. The random vector $\mathcal{Y}$ is observed on N sampling units yielding the $N \times Q$

response matrix $\boldsymbol{Y}$, which is typically then column-centered, and the random vector $\mathcal{X}$ is observed on the same N sampling units yielding the $N \times M$ predictor matrix $\boldsymbol{X}$. The MMLR regression model can then be stated as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \tag{3.1}$$

where $\boldsymbol{B}$ is the $M \times Q$ matrix of regression coefficients, and $\boldsymbol{E}$ is the $N \times Q$ matrix containing errors in the model. Column j of $\boldsymbol{B}$ contains the unknown coefficients which model the j^{th} response variable as a linear combination of the M predictor variables.

The null hypothesis of interest is that the response variables cannot be modeled as a linear combination of the predictor variables, i.e.,

$$H_0 : \boldsymbol{B} = \boldsymbol{0}, \tag{3.2}$$

(Rencher, 2002). The alternative is $\boldsymbol{B} \neq \boldsymbol{0}$, i.e., at least one coefficient is non-zero. In this case the response variables can be modeled as a linear combination of the predictor variables. Note here that the emphasis is on detecting a relationship given the observed data, rather than one from which the response observations associated with new observations of the predictor variables can be predicted.

3.1.2 Multivariate Multiple Linear Regression

The traditional approach of testing null hypothesis (3.2) requires using a decomposition of the total sum of squares matrix $\boldsymbol{Y}^T\boldsymbol{Y}$ (which equals $\boldsymbol{T}$ defined in Section 2.1.2 as $\boldsymbol{Y}$ is centered), into elements explained and unexplained by the regression model. This exact expression is deferred for the moment, as it requires first estimating the optimal $\boldsymbol{B}$.

The optimal $\boldsymbol{B}$ is found by minimizing the errors in $\boldsymbol{E}$, typically approached by minimizing the quantity

$$\operatorname{tr}\left(\boldsymbol{E}^T\boldsymbol{E}\right) = \operatorname{tr}\left(\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\right)\right).$$

This least squares approach results in a least squares estimator of $\boldsymbol{B}$, denoted $\hat{\boldsymbol{B}}$, given by $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. On defining the idempotent and symmetric projection matrix

$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$, referred to as the hat matrix, the fitted values of the regression model are given by $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{H}\boldsymbol{Y}$, yielding residuals $\boldsymbol{R} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y}$ where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix.

The above exposition assumes that $\boldsymbol{H}$ is well-defined, which in turn depends on $\boldsymbol{X}^T\boldsymbol{X}$ being non-singular. For this to be the case, $\boldsymbol{X}$ must be of full rank, i.e., the M columns of $\boldsymbol{X}$ must be linearly independent. This will not be so if $N < M$, since the rank of $\boldsymbol{X}$ is bounded from above by the minimum of N and M, and hence the rank cannot equal M. If $N = M$, the rank of $\boldsymbol{X}$ may equal M and hence $\boldsymbol{X}^T\boldsymbol{X}$ may be singular. However, in this case $\boldsymbol{H}$ will equal the $N \times N$ identity matrix $\boldsymbol{I}_N$ (see Appendix A for proof), and so will not be of any use in the regression model. It is typically assumed that $N >> M$ and that $\boldsymbol{X}$ is of full rank (see, for instance, Mardia *et al.* (1979) and Bingham and Fry (2010)).

The sum of squares decomposition is then given by

$$\boldsymbol{Y}^T\boldsymbol{Y} = \hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}} + \boldsymbol{R}^T\boldsymbol{R}, \tag{3.3}$$

where $\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}$ is the sum of squares matrix predicted by the regression model (referred to as the predicted sum of squares matrix) and $\boldsymbol{R}^T\boldsymbol{R}$ is the sum of squares matrix of the residual errors in the predicted model (referred to as the residual sum of squares matrix). The predicted sum of squares matrix provides the sum of squares components explained by the fitted regression model, while the residual sum of squares matrix provides the sum of squares components which are unexplained by the model. The decomposition is analogous to the sum of squares decomposition described in Section 2.1.2 for MANOVA.

The traditional statistics used for MANOVA can also be applied to test (3.2). For instance, Wilks' Λ can be defined as

$$\Lambda = \frac{\det\left(\boldsymbol{R}^T\boldsymbol{R}\right)}{\det\left(\boldsymbol{Y}^T\boldsymbol{Y}\right)},$$

ranging between 0 and 1, with smaller values providing evidence against the null. This is because $\det\left(\boldsymbol{R}^T\boldsymbol{R}\right)$ provides a scalar-valued quantification of the size of the values in $\boldsymbol{R}^T\boldsymbol{R}$, and as this decreases, a greater proportion of the total sum of squares is explained by the regression model (i.e., the values in $\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}$ are larger). Similarly, the

Lawley-Hotelling and Pillai trace statistics can be defined as

$$\text{LH} = \text{tr}\left(\left(\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right) \quad \text{and} \quad \text{PT} = \text{tr}\left(\left(\boldsymbol{Y}^T\boldsymbol{Y}\right)^{-1}\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right).$$

Significance can be assessed via distributional results which are available under the assumption that the errors in the regression model are distributed Multivariate Normal (see, for instance, Rencher (2002)).

In bioinformatics applications, it is often the case that the number of response variables exceeds the number of sampling units, i.e., $N < Q$ (see, for instance, Schork and Zapala (2012), where random vector $\mathcal{Y}$ is taken to be the expression levels of multiple genes). This causes problems for the Lawley-Hotelling and Pillai trace statistics, since the total and residual sum of squares matrices are singular and cannot be inverted.

An alternative statistic which can be used in this case is the pseudo F statistic proposed by McArdle and Anderson (2001). Analogously to the classical F statistic applied in linear regression where $\mathcal{Y}$ is scalar-valued, it is defined as the ratio of the total variability of $\mathcal{Y}$ explained by the fitted regression model to that which is unexplained by the fitted model. The required variability terms are obtained by applying the trace operator to decomposition (3.3). In particular,

$$\text{tr}\left(\boldsymbol{Y}^T\boldsymbol{Y}\right) = \text{tr}\left(\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right) + \text{tr}\left(\boldsymbol{R}^T\boldsymbol{R}\right) \tag{3.4}$$

where $\text{tr}\left(\boldsymbol{Y}^T\boldsymbol{Y}\right) = \text{tr}\left(\boldsymbol{T}\right)$ is the observed total variability of $\mathcal{Y}$, $\text{tr}\left(\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right)$ is the variability explained by the fitted regression model, and $\text{tr}\left(\boldsymbol{R}^T\boldsymbol{R}\right)$ is the variability unexplained by the fitted regression model. The pseudo F statistic is then defined as

$$F = \frac{\text{tr}\left(\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right)}{\text{tr}\left(\boldsymbol{R}^T\boldsymbol{R}\right)}, \tag{3.5}$$

taking non-negative values. Larger values provide evidence against the null hypothesis, since larger values of $\text{tr}\left(\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right)$ and smaller values of $\text{tr}\left(\boldsymbol{R}^T\boldsymbol{R}\right)$ indicate that much of the observed variability is explained by the fitted regression model. If $Q = 1$, that is, there is only one response variable comprising $\mathcal{Y}$, this statistic reduces to the classical F statistic applied in standard linear regression, ignoring degrees of freedom divisors. For $Q > 1$, the null sampling distribution of F is unknown, so significance is typically

assessed non-parametrically via permutations of the N sampling units.

3.2 Distance-Based Approach

3.2.1 Problem Statement

Consider now that random variable $\mathcal{Y}$ may not necessarily be vector-valued, i.e., it may not be a random vector, but may curve- or graph-valued, for instance. Denote the N observations of $\mathcal{Y}$ by $\{y_i\}_{i=1}^N$, and assume that a suitable semi-metric or metric distance function, $d_{\mathcal{Y}}(\cdot,\cdot)$, is defined yielding the $N \times N$ distance matrix $\boldsymbol{\Delta}_{\mathcal{Y}} = \{d_{\mathcal{Y}}(y_i, y_j)\}_{i,j=1}^N$.

From distance matrix $\boldsymbol{\Delta}_{\mathcal{Y}}$ it is possible to obtain a scalar-valued measure of the spread of the N observations of $\mathcal{Y}$, also referred to as the variability of $\mathcal{Y}$. This is achieved through principal coordinate analysis, which is described in Section 3.2.2. The problem of interest then entails modeling this variability in terms of the variability exhibited by the predictor variables comprising random vector $\mathcal{X}$. Under the null hypothesis, the variability in $\boldsymbol{\Delta}_{\mathcal{Y}}$ is not explained by $\mathcal{X}$. A mathematical expression of this null hypothesis is deferred until Section 3.2.3. Under the alternative hypothesis the predictor variables do explain the observed variability in $\boldsymbol{\Delta}_{\mathcal{Y}}$.

3.2.2 Principal Coordinate Analysis

Given $N \times N$ distance matrix $\boldsymbol{\Delta}_{\mathcal{Y}}$, principal coordinate analysis, also known as classical multidimensional scaling (MDS), seeks to represent each observation as an N-dimensional vector in Euclidean space. In particular, these vectors are sought such that their pairwise Euclidean distances equal the corresponding pairwise distances in $\boldsymbol{\Delta}_{\mathcal{Y}}$ (Torgerson, 1952; Gower, 1966). Thus, $\mathcal{Y}$ with observations $\{y_i\}_{i=1}^N$ is represented by N-dimensional random vector $\tilde{\mathcal{Y}}$ with centered observations $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^N$ such that $d^2_{\mathcal{Y}}(y_i, y_j) = (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_j)^T (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_j)$.

In MDS the centered $N \times N$ matrix $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_N)^T$ is sought and is referred to as the principal coordinate matrix. It can be found via a three-step procedure which derives and solves an equation containing the known distances $\boldsymbol{\Delta}_{\mathcal{Y}}$ and the unknown $\tilde{\boldsymbol{Y}}$ (Borg and Groenen, 2005). Begin by storing the squared pairwise Euclidean distances between $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^N$ in the matrix $\boldsymbol{\Delta}^2$ with elements $\left\{d^2_{\tilde{\mathcal{Y}}}(\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_j) = (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_j)^T (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_j)\right\}_{i,j=1}^N$.

In the first step, $\boldsymbol{\Delta}^2$ is expressed in terms of the $N \times N$ outer product matrix $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$. This matrix contains the inner products between all N vectors, that is, the $(i,j)^{\text{th}}$ element is given by $\tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_j$ for $i,j = 1,\ldots,N$, so that the elements of $\boldsymbol{\Delta}^2$ are given by

$$d^2_{\mathcal{Y}}(\tilde{\boldsymbol{y}}_i,\tilde{\boldsymbol{y}}_j) = \tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_i + \tilde{\boldsymbol{y}}_j^T\tilde{\boldsymbol{y}}_j - 2\tilde{\boldsymbol{y}}_i^T\tilde{\boldsymbol{y}}_j,$$

for all i,j. The matrix version of this relationship is given by

$$\boldsymbol{\Delta}^2 = \boldsymbol{d}\mathbf{1}_N^T + \mathbf{1}_N\boldsymbol{d}^T - 2\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T, \tag{3.6}$$

where $\boldsymbol{d} = \left(\tilde{\boldsymbol{y}}_1^T\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_N^T\tilde{\boldsymbol{y}}_N\right)^T$ is the column vector containing the diagonal elements of $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$.

The second step consists of replacing the unknown $\boldsymbol{\Delta}^2$ in (3.6) with the known $\boldsymbol{\Delta}^2_{\mathcal{Y}}$, and deriving an equation in terms of $\boldsymbol{\Delta}^2_{\mathcal{Y}}$ and $\tilde{\boldsymbol{Y}}$. This replacement yields

$$\boldsymbol{\Delta}^2_{\mathcal{Y}} = \boldsymbol{d}\mathbf{1}_N^T + \mathbf{1}_N\boldsymbol{d}^T - 2\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T, \tag{3.7}$$

which can be simplified so that only $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$ remains on the right-hand side. This is achieved by using the symmetric $N \times N$ centering matrix $\boldsymbol{C} = (\boldsymbol{I}_N - \boldsymbol{J}_N/N)$, where $\boldsymbol{J}_N$ is the square matrix of ones, to remove the terms containing $\boldsymbol{d}$. In particular, $\mathbf{1}_N^T\boldsymbol{C} = \mathbf{0}_N^T$ and $\boldsymbol{C}\mathbf{1}_N = \mathbf{0}_N$ where $\mathbf{0}_N$ is the N-dimensional column vector of zeros, so that the elements of $\boldsymbol{d}$ are weighted by zeros. These zero vectors arise by performing a double-centering, i.e., multiplying on both sides by the centering matrix, as follows:

$$\begin{aligned}
\boldsymbol{C}\boldsymbol{\Delta}^2_{\mathcal{Y}}\boldsymbol{C} &= \boldsymbol{C}\boldsymbol{d}\left(\mathbf{1}_N^T\boldsymbol{C}\right) + \left(\boldsymbol{C}\mathbf{1}_N\right)\boldsymbol{d}^T\boldsymbol{C} - 2\boldsymbol{C}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\boldsymbol{C} \\
&= \boldsymbol{C}\boldsymbol{d}\mathbf{0}_N^T + \mathbf{0}_N\boldsymbol{d}^T\boldsymbol{C} - 2\boldsymbol{C}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\boldsymbol{C} \\
&= -2\boldsymbol{C}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\boldsymbol{C}.
\end{aligned}$$

It follows that

$$-\frac{1}{2}\boldsymbol{C}\boldsymbol{\Delta}^2_{\mathcal{Y}}\boldsymbol{C} = \boldsymbol{C}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\boldsymbol{C},$$

and since $\tilde{\boldsymbol{Y}}$ is assumed to be centered, $\boldsymbol{C}\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Y}}$, so that

$$-\frac{1}{2}\boldsymbol{C}\boldsymbol{\Delta}^2_{\mathcal{Y}}\boldsymbol{C} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T.$$

The double-centered matrix $-\frac{1}{2}\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{Y}}^2\boldsymbol{C}$ is referred to as the centered inner product matrix since it contains the inner products of the centered vectors $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^N$, and we denote it by $\boldsymbol{G}_{\mathcal{Y}}$. In terms of the known distances $\{d_{\mathcal{Y}}(y_i, y_j)\}_{i,j=1}^N$, the elements of $\boldsymbol{G}_{\mathcal{Y}}$ are given by

$$\begin{aligned} g_{\mathcal{Y}}(y_i, y_j) &= -\frac{1}{2}\left(\frac{1}{N}d_{\mathcal{Y}}^2(y_i, y_j) - \frac{1}{N}\sum_{k=1}^{N} d_{\mathcal{Y}}^2(y_i, y_k) - \frac{1}{N}\sum_{l=1}^{N} d_{\mathcal{Y}}^2(y_l, y_j)\right. \\ &\quad \left. + \frac{1}{N^2}\sum_{l=1}^{N}\sum_{k=1}^{N} d_{\mathcal{Y}}^2(y_l, y_k)\right), \end{aligned}$$

for $i, j = 1, \ldots, N$, where $\sum_{k=1}^N d_{\mathcal{Y}}^2(y_i, y_k)/N$ is the mean of the i^{th} row of squared distances, $\sum_{l=1}^N d_{\mathcal{Y}}^2(y_l, y_j)/N$ is the mean of the j^{th} column of squared distances, and $\sum_{k=1}^N d_{\mathcal{Y}}^2(y_l, y_k)/N^2$ is the total mean of all squared distances. This symmetric and real-valued matrix depends solely on known quantities, so we obtain the equation

$$\boldsymbol{G}_{\mathcal{Y}} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T. \tag{3.8}$$

For the final step of the MDS procedure, notice that $\tilde{\boldsymbol{Y}}$ can be found via spectral decomposition of $\boldsymbol{G}_{\mathcal{Y}}$. This is given by

$$\boldsymbol{G}_{\mathcal{Y}} = \boldsymbol{U}_{\mathcal{Y}}\boldsymbol{\Lambda}_{\mathcal{Y}}\boldsymbol{U}_{\mathcal{Y}}^T,$$

where $\boldsymbol{U}_{\mathcal{Y}}$ contains the N eigenvectors of $\boldsymbol{G}_{\mathcal{Y}}$ with corresponding ordered eigenvalues $\{\lambda_{\mathcal{Y},i}\}_{i=1}^N$ on the diagonal of $\boldsymbol{\Lambda}_{\mathcal{Y}}$. $\boldsymbol{U}_{\mathcal{Y}}$ is referred to as the standard coordinate matrix and its columns represent the N orthogonal dimensions comprising the Euclidean space, i.e., $\boldsymbol{U}_{\mathcal{Y}}^T\boldsymbol{U}_{\mathcal{Y}} = \boldsymbol{U}_{\mathcal{Y}}\boldsymbol{U}_{\mathcal{Y}}^T = \boldsymbol{I}_N$. The ordered eigenvalues represent the importance of each associated dimension. $\tilde{\boldsymbol{Y}}$ is given by $\boldsymbol{U}_{\mathcal{Y}}\boldsymbol{\Lambda}_{\mathcal{Y}}^{\frac{1}{2}}$, where $\boldsymbol{\Lambda}_{\mathcal{Y}}^{\frac{1}{2}}$ is the diagonal matrix containing the square-rooted eigenvalues of $\boldsymbol{\Lambda}_{\mathcal{Y}}$.

If $\mathcal{Y}$ is a real-valued random vector with observations centered and $d_{\mathcal{Y}}$ is the Euclidean distance measure, then $\boldsymbol{G}_{\mathcal{Y}} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T = \boldsymbol{Y}\boldsymbol{Y}^T$, but $\tilde{\boldsymbol{Y}}$ is not necessarily equal to $\boldsymbol{Y}$ (indeed, the dimensions of the matrices will not even match if $N \neq Q$). Furthermore, MDS is equivalent to classical principal component analysis (PCA) in this case (Krzanowski, 2000), which is commonly used as a dimensionality reduction and visualization tool. PCA represents the N observations in $\boldsymbol{Y}$ by an orthogonal config-

uration in Euclidean space, $\tilde{\boldsymbol{Y}}$, such that the variance of the N values in each column (where each column represents an orthogonal direction) is maximized. The configuration $\tilde{\boldsymbol{Y}}$ is found via an eigenanalysis of the covariance matrix of $\boldsymbol{Y}$, i.e., $\boldsymbol{Y}^T\boldsymbol{Y}/(N-1)$. The eigenvectors represent the orthogonal directions and the eigenvalues represent the variance in the corresponding directions. The sum of the eigenvalues equals the total sample variance of $\mathcal{Y}$, found as the sum of the sample variance of each variable comprising $\mathcal{Y}$. The proportion of variance explained by each direction can then be found by comparing the eigenvalue of that dimension with the sum of all the eigenvalues.

MDS is also often used for dimensionality reduction and visualization. For example, the inter-point relationships between the original observations $\{y_i\}_{i=1}^N$ can be viewed by plotting the first 2 or 3 elements of the vectors $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^N$, which due to the ordering of the eigenvalues, are deemed the most important. However, here MDS is described as a means of quantifying the spread of the N observations of $\mathcal{Y}$ given $d_{\mathcal{Y}}$, where $\mathcal{Y}$ may not necessarily be a random vector. Since $\tilde{\mathcal{Y}}$ represents $\mathcal{Y}$, we can consider the sample total sum of squares of $\tilde{\mathcal{Y}}$ as an appropriate measure, given by $\operatorname{tr}\left(\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}\right)$ (since $\tilde{\boldsymbol{Y}}$ is centered). Since $\operatorname{tr}\left(\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}\right) = \operatorname{tr}\left(\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\right) = \operatorname{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right)$, the total variability of $\tilde{\mathcal{Y}}$ equals the quantity

$$\operatorname{tr}(\boldsymbol{G}_{\mathcal{Y}}) = \frac{1}{N}\sum_{i>j} d_{\mathcal{Y}}^2(\boldsymbol{y}_i, \boldsymbol{y}_j).$$

Therefore we define $\operatorname{tr}(\boldsymbol{G}_{\mathcal{Y}})$ as the total variability of $\mathcal{Y}$ with respect to $d_{\mathcal{Y}}$. Analogously to PCA, the standard coordinates $\boldsymbol{U}_{\mathcal{Y}}$ can be thought of as directions of variability which explain the total variability. Each corresponding eigenvalue can be compared with $\operatorname{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right) = \sum_{i=1}^N \lambda_{\mathcal{Y},i}$ to yield the proportion of the sample total variability explained by the given direction of variability. For example, the first direction accounts for $\left(\lambda_{\mathcal{Y},1}/\operatorname{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right)\right) \times 100\%$ of the total variability.

So far no mention has been made of the nature of the eigenvalues $\{\lambda_{\mathcal{Y},i}\}_{i=1}^N$, and in particular, when they are non-negative. It has been shown that they will be non-negative if the distance function $d_{\mathcal{Y}}$ is metric (Krzanowski, 2000). Consequently, some eigenvalues will be negative for semi-metric distance functions, such as those encountered in genetics, for example. This yields coordinate matrix $\tilde{\boldsymbol{Y}}$ with complex-valued components, which hinders the configuration being presented well in Euclidean space. Since the eigenvalues are ordered, the non-negative eigenvalues associated with the

first few directions of variability will be positive, and hence these more important directions can be viewed in Euclidean space. The final few directions, however, will be associated with negative eigenvalues, and therefore complex-valued axes are required.

There exist several approaches to dealing with negative eigenvalues (see, for example, Pekalska and Duin (2005)). If the negative eigenvalues are small in comparison to the positive eigenvalues, the directions of variability they represent can be disregarded as noise (Pekalska and Duin, 2005). For negative eigenvalues of larger magnitude, a correction can be applied to the off-diagonal elements of the original distance matrix. This is achieved by adding a suitable constant such that the resulting distances satisfy the triangle inequality (Legendre and Anderson, 1999). However, there is no conclusive proof that altering distance matrices in this manner is beneficial in providing a configuration in Euclidean space (Pekalska and Duin, 2005), as the structure of the observed distances must be altered.

Regardless of the distance $d_{\mathcal{Y}}$ being metric or semi-metric, the total variability of $\mathcal{Y}$ with respect to $d_{\mathcal{Y}}$ can still be described by $\operatorname{tr}(\boldsymbol{G}_{\mathcal{Y}})$, since it is non-negative and real-valued (as it equals a sum of squared distances).

3.2.3 The Pseudo F Test

McArdle and Anderson (2001) define the distance-based version of pseudo F statistic (3.5), also named the pseudo F statistic, by writing it in terms of Euclidean distances between the rows of $\boldsymbol{Y}$ and then generalizing to any suitable distance.

In order to achieve this, sum of squares decomposition (3.4) is written in terms of the Euclidean distances between the rows of $\boldsymbol{Y}$. This is done by first applying the property of the trace operator that $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of suitable dimensions, as follows:

$$\begin{aligned}
\operatorname{tr}\left(\boldsymbol{Y}^T\boldsymbol{Y}\right) &= \operatorname{tr}\left(\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\right) + \operatorname{tr}\left(\boldsymbol{R}^T\boldsymbol{R}\right) \\
\Rightarrow \operatorname{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right) &= \operatorname{tr}\left(\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}^T\right) + \operatorname{tr}\left(\boldsymbol{R}\boldsymbol{R}^T\right) \\
\Rightarrow \operatorname{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right) &= \operatorname{tr}\left(\boldsymbol{H}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{H}\right) + \operatorname{tr}\left(\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\boldsymbol{Y}\boldsymbol{Y}^T\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\right) \\
\Rightarrow \operatorname{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right) &= \operatorname{tr}\left(\boldsymbol{H}\boldsymbol{Y}\boldsymbol{Y}^T\right) + \operatorname{tr}\left(\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\boldsymbol{Y}\boldsymbol{Y}^T\right). \qquad (3.9)
\end{aligned}$$

Expressing the sum of squares decomposition in this form presents the dependence

on the $N \times N$ matrix product $\boldsymbol{Y}\boldsymbol{Y}^T$. This matrix is related to the $N \times N$ distance matrix $\boldsymbol{\Delta}_{\mathcal{Y}}$ containing the Euclidean distances between the rows of $\boldsymbol{Y}$ via $\boldsymbol{G}_{\mathcal{Y}}$ (defined in Section 3.2.2). In particular, $\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{G}_{\mathcal{Y}}$ since $\boldsymbol{Y}$ is column-centered, so that (3.9) can be written as

$$\operatorname{tr}(\boldsymbol{G}_{\mathcal{Y}}) = \operatorname{tr}(\boldsymbol{H}\boldsymbol{G}_{\mathcal{Y}}) + \operatorname{tr}((\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{G}_{\mathcal{Y}}). \tag{3.10}$$

In this decomposition, $\operatorname{tr}(\boldsymbol{G}_{\mathcal{Y}})$ quantifies the total variability exhibited by the samples of $\mathcal{Y}$ with respect to $d_{\mathcal{Y}}$, $\operatorname{tr}(\boldsymbol{H}\boldsymbol{G}_{\mathcal{Y}})$ quantifies the variability explained by the predictor variables, and $\operatorname{tr}((\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{G}_{\mathcal{Y}})$ quantifies the remaining variability.

In the computation of $\boldsymbol{G}_{\mathcal{Y}}$ the distances stored in $\boldsymbol{\Delta}_{\mathcal{Y}}$ can be generalized to be of any suitable type. It then follows that the response observations can be of any type, not just vector-valued, provided a suitable distance measure is also defined. In this case null hypothesis (3.2) is not strictly valid since it pertains to the original regression model in which the response observations are vector-valued. This is not described explicitly in McArdle and Anderson (2001), so we describe the reasoning below.

The regression model being considered can be expressed in terms of the $N \times N$ principal coordinate matrix $\tilde{\boldsymbol{Y}}$ arising from $\boldsymbol{G}_{\mathcal{Y}}$. By substituting $\tilde{\boldsymbol{Y}}$ into (3.9) and reverse engineering the decomposition, we obtain

$$\operatorname{tr}\left(\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}\right) = \operatorname{tr}\left(\hat{\tilde{\boldsymbol{Y}}}^T\hat{\tilde{\boldsymbol{Y}}}\right) + \operatorname{tr}\left(\boldsymbol{R}_1^T\boldsymbol{R}_1\right), \tag{3.11}$$

where $\hat{\tilde{\boldsymbol{Y}}} = \boldsymbol{H}\tilde{\boldsymbol{Y}}$ and $\boldsymbol{R}_1 = \tilde{\boldsymbol{Y}} - \hat{\tilde{\boldsymbol{Y}}}$ is the $N \times N$ residual sum of squares matrix. We observe that this can come from applying the trace operator to the sum of squares matrix decomposition

$$\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} = \hat{\tilde{\boldsymbol{Y}}}^T\hat{\tilde{\boldsymbol{Y}}} + \boldsymbol{R}_1^T\boldsymbol{R}_1.$$

This in turn can be seen to come from the MMLR model

$$\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B}_1 + \boldsymbol{E}_1, \tag{3.12}$$

where $\boldsymbol{B}_1$ is the $M \times N$ matrix of regression coefficients and $\boldsymbol{E}_1$ is the $N \times N$ matrix of errors. This is clearly different from the original MMLR model given by (3.1); the

dimensions of the regression coefficient and error matrices are different.

The null hypothesis of no predictive relationship based on (3.12) can be stated as

$$H_0 : \boldsymbol{B}_1 = \boldsymbol{0}. \tag{3.13}$$

This is equivalent to (3.2) in the case of $d_{\mathcal{Y}}$ being the Euclidean distance applied to centered observations stored in $\boldsymbol{Y}$. This is because $\boldsymbol{G}_{\mathcal{Y}} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T = \boldsymbol{Y}\boldsymbol{Y}^T$, so that (3.11) is equivalent to (3.9), which in turn was derived from the original MMLR model.

The distance-based pseudo F statistic used to test (3.13) is expressed as

$$F = \frac{\operatorname{tr}(\boldsymbol{H}\boldsymbol{G}_{\mathcal{Y}})}{\operatorname{tr}((\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{G}_{\mathcal{Y}})}, \tag{3.14}$$

and quantifies the ratio of variability in $\boldsymbol{\Delta}_{\mathcal{Y}}$ explained and unexplained by the predictor variables. Larger values of this statistic provide evidence against the null.

Given an observed value of the test statistic, $\hat{F}$, inference is performed using permutations. Given N_π Monte Carlo permutations $\pi \in \Pi$, the set $\{\hat{F}_\pi\}_{\pi\in\Pi}$ is generated where $\hat{F}_\pi$ is the pseudo F statistic evaluated with $\boldsymbol{G}_{\mathcal{Y},\pi}$ instead of $\boldsymbol{G}_{\mathcal{Y}}$, where $\boldsymbol{G}_{\mathcal{Y},\pi}$ denotes $\boldsymbol{G}_{\mathcal{Y}}$ with rows and columns simultaneously permuted by π. The p-value is then computed as the proportion of the N_π permuted statistics greater than or equal to the observed $\hat{F}$, i.e.,

$$\frac{\#(\hat{F}_\pi \geq \hat{F})}{N_\pi}.$$

Clearly, this is a one-sided test, since only larger values of F provide evidence against the null.

3.3 Summary

In bioinformatics, regression models are routinely deployed to relate variables such as genes to predictor variables such as genetic polymorphisms and environmental variables (see, for instance, Salem *et al.* (2010)). Where the response observations are real and vector-valued, they are typically high-dimensional with $N < Q$, causing problems for traditional multivariate approaches. Multivariate distances can be applied to the responses in such cases, allowing the use of the distance-based pseudo F statistic to test for no predictive relationship.

The pseudo F test has been applied to many problems in bioinformatics, for which the response observations are not necessarily real and vector-valued. In particular, distances have been applied to response observations including real and vector-valued gene expression and imaging data, discrete and vector-valued SNP data and curve-valued dose-dependent gene expression data (Zapala and Schork, 2006; Wessel and Schork, 2006; Salem *et al.*, 2010). This exemplifies the broad utility of the pseudo F test.

A limitation in its use in these applications, however, is the relatively low number of Monte Carlo permutations enumerated when performing inference. For instance Salem *et al.* (2010) use $O(10^4)$ permutations for $N = 49$ samples, and Wessel and Schork (2006) use $O(10^5)$ permutations for $N = 57$ samples. Thus, the permutation procedure used in conjunction with the pseudo F statistic is subject to the problems discussed in the introduction. This limitation has also been highlighted recently by Schork and Zapala (2012), especially inlight of the increasing need to perform repeated tests, such as in GWA studies.

It has been unanimously agreed that the null distribution of the pseudo F statistic cannot be derived exactly (Zapala and Schork, 2006; McArdle and Anderson, 2001), since it is dependent on the particular distance measure being applied. No attempts have been made in the literature to approximate this distribution. As a result, the full potential of applying the pseudo F test in studies requiring tens of thousands of tests has yet to be examined.

In Chapter 6 we provide an approximation for the null distribution of the pseudo F statistic. We also demonstrate its applicability in imaging genetics, where the interest is in modeling SNPs as predictor variables of observed imaging data.

Chapter 4

Detecting Association Between Two Random Vectors

In this chapter we introduce the problem of testing for no association between two random vectors. This problem has received much interest in the multivariate literature, where the random vectors are comprised of real and scalar-valued random variables. While many approaches exist, we review a selected few as representative approaches for the two separate multivariate null hypotheses which are typically tested. This is followed by a review of distance-based approaches which are applicable when we have random variables of different structures instead of two vector-valued random variables. A summary concludes the chapter with limitations of the existing distance-based methods.

4.1 Multivariate Approaches

4.1.1 Problem Statement

Consider the random vectors $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_P)^T \in \mathbb{R}^P$ and $\mathcal{Y} = (\mathcal{Y}_1, \ldots, \mathcal{Y}_Q)^T \in \mathbb{R}^Q$. Having observed these on the same N sampling units, the aim is to infer if an association exists between $\mathcal{X}$ and $\mathcal{Y}$. In particular, the term association is used to mean 'some relationship' which is symmetric, that is, the variables comprising either $\mathcal{X}$ or $\mathcal{Y}$ are not deemed predictor variables of the other, as in the regression approach.

Classical approaches such as Canonical Correlation Analysis (Hotelling, 1936) and

the RV test (Escoufier, 1973) consider linear relationships between the variables, as captured by the $(P+Q)\times(P+Q)$ covariance matrix of $(\mathcal{X},\mathcal{Y})$ given by

$$\begin{pmatrix} \mathbf{\Sigma}_{\mathcal{X}\mathcal{X}} & \mathbf{\Sigma}_{\mathcal{X}\mathcal{Y}} \\ \mathbf{\Sigma}_{\mathcal{Y}\mathcal{X}} & \mathbf{\Sigma}_{\mathcal{Y}\mathcal{Y}} \end{pmatrix},$$

where $\mathbf{\Sigma}_{\mathcal{X}\mathcal{X}} = \{\text{cov}(\mathcal{X}_i,\mathcal{X}_j)\}_{i,j=1}^{P}$, $\mathbf{\Sigma}_{\mathcal{Y}\mathcal{Y}} = \{\text{cov}(\mathcal{Y}_i,\mathcal{Y}_j)\}_{i,j=1}^{Q}$, $\mathbf{\Sigma}_{\mathcal{X}\mathcal{Y}} = \{\text{cov}(\mathcal{X}_i,\mathcal{Y}_j)\}$ for $i = 1,\ldots,P$ and $j = 1,\ldots,Q$, and $\mathbf{\Sigma}_{\mathcal{Y}\mathcal{X}} = \mathbf{\Sigma}_{\mathcal{X}\mathcal{Y}}^T$, where $\text{cov}(\cdot,\cdot)$ is the classical covariance function.

The null hypothesis of interest is typically stated as

$$H_0 : \mathbf{\Sigma}_{\mathcal{X}\mathcal{Y}} = \mathbf{0}, \tag{4.1}$$

thats is, the variables comprising $\mathcal{X}$ are uncorrelated with those comprising $\mathcal{Y}$. The alternative is $\mathbf{\Sigma}_{\mathcal{X}\mathcal{Y}} \neq \mathbf{0}$, i.e., that they are correlated. It is clear that this is a generalization of the classical Pearson's correlation test of no correlation between two random variables.

On assuming that $\mathcal{X}$ and $\mathcal{Y}$ have density functions $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$, respectively, and joint density function $f_{\mathcal{X}\mathcal{Y}}$, a null hypothesis of independence can be stated as

$$H_0 : f_{\mathcal{X}\mathcal{Y}} = f_{\mathcal{X}} f_{\mathcal{Y}} \tag{4.2}$$

(Székely *et al.*, 2007). The alternative hypothesis is $f_{\mathcal{X}\mathcal{Y}} \neq f_{\mathcal{X}} f_{\mathcal{Y}}$, and hence that a nonlinear relationship exists between $\mathcal{X}$ and $\mathcal{Y}$.

We review the RV test of Escoufier (1973) and the distance correlation (dCor) test of Székely *et al.* (2007) which can be used to test (4.1) and (4.2), respectively.

4.1.2 The RV Test

The RV test of Escoufier (1973) was proposed as a generalization of Pearson's correlation test to real-valued random vectors. It uses generalizations of the classical (univariate) notions of covariance, variance and correlation to define scalar-valued expressions for multivariate covariance, variance and correlation. We describe these below.

The covariance between $\mathcal{X}$ and $\mathcal{Y}$, denoted COVV$(\mathcal{X}, \mathcal{Y})$, is defined as the sum of the squared covariances between every random variable comprising $\mathcal{X}$ with every random variable comprising $\mathcal{Y}$. That is,

$$\text{COVV}\left(\mathcal{X}, \mathcal{Y}\right) = \sum_{i=1}^{P} \sum_{j=1}^{Q} \text{cov}^2(\mathcal{X}_i, \mathcal{Y}_j),$$

which can be written in matrix form as $\text{COVV}\left(\mathcal{X}, \mathcal{Y}\right) = \text{tr}\left(\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{Y}}\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{X}}\right)$. This definition serves two purposes. Firstly, it permits an intuitive partitioning of COVV when $\mathcal{X}$ is partitioned into two separate random vectors of reduced length. To see this, define $\mathcal{X}^1$ and $\mathcal{X}^2$ such that $\mathcal{X} = (\mathcal{X}^1, \mathcal{X}^2)$ where $\mathcal{X}^1 = (\mathcal{X}_1, \ldots, \mathcal{X}_K)^T$ and $\mathcal{X}^2 = (\mathcal{X}_{K+1}, \ldots, \mathcal{X}_P)^T$ for some $1 < K < P$. Then,

$$\begin{aligned}
\text{COVV}\left(\mathcal{X}, \mathcal{Y}\right) &= \text{COVV}\left((\mathcal{X}^1; \mathcal{X}^2), \mathcal{Y}\right) \\
&= \sum_{i=1}^{P} \sum_{j=1}^{Q} \text{cov}^2(\mathcal{X}_i, \mathcal{Y}_j) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{Q} \text{cov}^2(\mathcal{X}_i, \mathcal{Y}_j) + \sum_{i=K+1}^{P} \sum_{j=1}^{Q} \text{cov}^2(\mathcal{X}_i, \mathcal{Y}_j) \\
&= \text{COVV}\left(\mathcal{X}^1, \mathcal{Y}\right) + \text{COVV}\left(\mathcal{X}^2, \mathcal{Y}\right).
\end{aligned}$$

The second purpose of the definition is to ensure that the scalar-valued variance of $\mathcal{X}$, denoted VAV$\left(\mathcal{X}\right)$ and defined by

$$\begin{aligned}
\text{VAV}\left(\mathcal{X}\right) &= \text{COVV}\left(\mathcal{X}, \mathcal{X}\right) \\
&= \text{tr}\left(\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{X}}\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{X}}\right),
\end{aligned}$$

is non-negative, inline with the fact that the variance of a real and scalar-valued random variable is non-negative.

The correlation between $\mathcal{X}$ and $\mathcal{Y}$, denoted RV$(\mathcal{X}, \mathcal{Y})$, is then defined by substituting the multivariate covariance and variance definitions into the classical definition

of correlation to yield

$$\begin{aligned} \mathrm{RV}(\mathcal{X},\mathcal{Y}) &= \frac{\mathrm{COVV}(\mathcal{X},\mathcal{Y})}{\sqrt{\mathrm{VAV}(\mathcal{X})\,\mathrm{VAV}(\mathcal{Y})}} \\ &= \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{Y}}\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{X}}\right)}{\sqrt{\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{X}}\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{X}}\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{Y}}\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{Y}}\right)}}. \end{aligned} \tag{4.3}$$

It ranges between 0 and 1, with 0 indicating no association when $\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{Y}} = \mathbf{0}$ (i.e., $\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{Y}}\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{X}}\right) = \mathbf{0}$), and 1 indicating perfect association when $\mathcal{X} = a\mathcal{Y}$ for some real-valued constant a (i.e., $\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{Y}} = a\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{Y}}$ and $\boldsymbol{\Sigma}_{\mathcal{X}\mathcal{X}} = a^2\boldsymbol{\Sigma}_{\mathcal{Y}\mathcal{Y}}$). It generalizes Pearson's correlation coefficient, denoted by the correlation function $\mathrm{cor}(\cdot,\cdot)$, in the following way: if $P = Q = 1$, $\mathrm{RV}(\mathcal{X},\mathcal{Y}) = \mathrm{cor}^2(\mathcal{X},\mathcal{Y})$.

Given centered observations $\boldsymbol{X}$ and $\boldsymbol{Y}$ of $\mathcal{X}$ and $\mathcal{Y}$, respectively, empirical values of $\mathrm{COVV}(\mathcal{X},\mathcal{Y})$, $\mathrm{VAV}(\mathcal{X})$ and $\mathrm{VAV}(\mathcal{Y})$ can be directly substituted into (4.3) to yield an empirical RV coefficient. Define the $(P+Q)\times(P+Q)$ sample covariance matrix of $(\mathcal{X},\mathcal{Y})$ by

$$\frac{1}{N-1}\begin{pmatrix} \boldsymbol{T}_{\mathcal{X}\mathcal{X}} & \boldsymbol{T}_{\mathcal{X}\mathcal{Y}} \\ \boldsymbol{T}_{\mathcal{Y}\mathcal{X}} & \boldsymbol{T}_{\mathcal{Y}\mathcal{Y}} \end{pmatrix},$$

where $\boldsymbol{T}_{\mathcal{X}\mathcal{X}} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{T}_{\mathcal{Y}\mathcal{Y}} = \boldsymbol{Y}^T\boldsymbol{Y}$, $\boldsymbol{T}_{\mathcal{X}\mathcal{Y}} = \boldsymbol{X}^T\boldsymbol{Y}$ and $\boldsymbol{T}_{\mathcal{Y}\mathcal{X}} = \boldsymbol{T}_{\mathcal{X}\mathcal{Y}}^T$. Then the sample variance of $\mathcal{X}$ is given by $\mathrm{tr}(\boldsymbol{T}_{\mathcal{X}\mathcal{X}}\boldsymbol{T}_{\mathcal{X}\mathcal{X}})/(N-1)$, and similarly for $\mathcal{Y}$, and the sample covariance between $\mathcal{X}$ and $\mathcal{Y}$ is given by $\mathrm{tr}(\boldsymbol{T}_{\mathcal{X}\mathcal{Y}}\boldsymbol{T}_{\mathcal{Y}\mathcal{X}})/(N-1)$. The empirical RV coefficient is then obtained as

$$\begin{aligned} \mathrm{RV}(\boldsymbol{X},\boldsymbol{Y}) &= \frac{\mathrm{tr}(\boldsymbol{T}_{\mathcal{X}\mathcal{Y}}\boldsymbol{T}_{\mathcal{Y}\mathcal{X}})}{\sqrt{\mathrm{tr}(\boldsymbol{T}_{\mathcal{X}\mathcal{X}}\boldsymbol{T}_{\mathcal{X}\mathcal{X}})\mathrm{tr}(\boldsymbol{T}_{\mathcal{Y}\mathcal{Y}}\boldsymbol{T}_{\mathcal{Y}\mathcal{Y}})}} \\ &= \frac{\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X})}{\sqrt{\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X})\mathrm{tr}(\boldsymbol{Y}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{Y})}}, \end{aligned} \tag{4.4}$$

with larger values providing evidence against null hypothesis (4.1). No association exists when $\boldsymbol{X}^T\boldsymbol{Y} = \mathbf{0}$, and perfect association exists when $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}$ for some mapping matrix $\boldsymbol{B} \in \mathbb{R}^{P\times Q}$ such that $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{I}_P$. That is, when there exists a linear mapping which relates every P-dimensional observation in $\boldsymbol{X}$ to every Q-dimensional observation in $\boldsymbol{Y}$.

Robert and Escoufier (1976) have shown that the RV coefficient can be interpreted in terms of the Euclidean distances arising from $\boldsymbol{X}$ and $\boldsymbol{Y}$. Due to the properties of

the trace operator, the RV coefficient can be written as

$$\begin{aligned} \mathrm{RV}(\boldsymbol{X},\boldsymbol{Y}) &= \frac{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T)}{\sqrt{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T)\mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{Y}\boldsymbol{Y}^T)}} \\ &= \frac{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T)}{||\boldsymbol{X}\boldsymbol{X}^T||||\boldsymbol{Y}\boldsymbol{Y}^T||}, \end{aligned} \tag{4.5}$$

where $||\cdot||$ denotes the Frobenius norm defined by $||\boldsymbol{A}|| = \sqrt{\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})}$ for matrix $\boldsymbol{A}$. (4.5) differs from (4.4) in that emphasis is placed on the two symmetric $N \times N$ matrices $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{Y}\boldsymbol{Y}^T$, instead of the four covariance matrices $\boldsymbol{X}^T\boldsymbol{Y} \in \mathbb{R}^{P\times Q}$, $\boldsymbol{Y}^T\boldsymbol{X} \in \mathbb{R}^{Q\times P}$, $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{P\times P}$ and $\boldsymbol{Y}^T\boldsymbol{Y} \in \mathbb{R}^{Q\times Q}$.

The matrices $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{Y}\boldsymbol{Y}^T$ contain information on the Euclidean distances between the N observations in the P- and Q-dimensional spaces of $\mathcal{X}$ and $\mathcal{Y}$, respectively, since

$$\boldsymbol{X}\boldsymbol{X}^T = -\frac{1}{2}\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}^2\boldsymbol{C} \quad \text{and} \quad \boldsymbol{Y}\boldsymbol{Y}^T = -\frac{1}{2}\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{Y}}^2\boldsymbol{C}, \tag{4.6}$$

with Euclidean distance matrices $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$ and centering matrix $\boldsymbol{C}$ (as described in Section 3.2.2). These matrices are invariant to rotations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and can be made invariant to scale by dividing by their respective Frobenius norms. Thus $\boldsymbol{X}\boldsymbol{X}^T/||\boldsymbol{X}\boldsymbol{X}^T||$ and $\boldsymbol{Y}\boldsymbol{Y}^T/||\boldsymbol{Y}\boldsymbol{Y}^T||$ are comparable, and differences in the pairwise Euclidean distances between the N observations in each space can be detected by considering the Frobenius distance between them. This distance is given by

$$\begin{aligned} d_F\left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{||\boldsymbol{X}\boldsymbol{X}^T||}, \frac{\boldsymbol{Y}\boldsymbol{Y}^T}{||\boldsymbol{Y}\boldsymbol{Y}^T||}\right) &= \left|\left|\frac{\boldsymbol{X}\boldsymbol{X}^T}{||\boldsymbol{X}\boldsymbol{X}^T||} - \frac{\boldsymbol{Y}\boldsymbol{Y}^T}{||\boldsymbol{Y}\boldsymbol{Y}^T||}\right|\right| \\ &= \sqrt{2\left(1 - \frac{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T)}{||\boldsymbol{X}\boldsymbol{X}^T||||\boldsymbol{Y}\boldsymbol{Y}^T||}\right)}, \end{aligned} \tag{4.7}$$

which on substitution of (4.5) yields

$$d_F\left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{||\boldsymbol{X}\boldsymbol{X}^T||}, \frac{\boldsymbol{Y}\boldsymbol{Y}^T}{||\boldsymbol{Y}\boldsymbol{Y}^T||}\right) = \sqrt{2\left(1 - \mathrm{RV}(\boldsymbol{X},\boldsymbol{Y})\right)}. \tag{4.8}$$

Thus, an RV coefficient value of 1 is equivalent to a Frobenius distance of 0 between the rotation and scale invariant configurations arising from the Euclidean distances. This distance representation of the RV coefficient can therefore be used to measure the dissimilarity between $\mathcal{X}$ and $\mathcal{Y}$ given their respective observations of possibly different

dimensions.

Inference of an observed RV coefficient can be performed by using many permutations of the rows of one of the data matrices, and each time recomputing the RV coefficient to generate a null sampling distribution, to which the observed RV coefficient is compared. Since permutations are computationally expensive, alternative approaches consisting of approximating the exact permutation distribution which would be obtained by using all possible permutations have been proposed. For instance, the Normal, Lognormal and Pearson type III distributions have all been proposed, such that the p-value can be obtained by comparing the observed RV value against the given distributional approximation (Josse *et al.*, 2008).

4.1.3 The Distance Correlation Test

The dCor test of Székely *et al.* (2007) uses the same idea as the RV coefficient, namely that a generalization of Pearson's correlation coefficient can be obtained by substituting the classical covariance and variance definitions for multivariate versions. For dCor, the scalar-valued multivariate versions of covariance, variance and correlation are defined with respect the notion of independence. These quantities are called distance covariance (dCov), distance variance (dVar) and distance correlation (dCor), and are defined below.

The distance covariance between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\mathrm{dCov}(\mathcal{X},\mathcal{Y}) = \sqrt{\frac{1}{c_P c_Q}\int_{\mathbb{R}^{P+Q}} \frac{||f_{\mathcal{X},\mathcal{Y}}(\boldsymbol{t},\boldsymbol{s}) - f_{\mathcal{X}}(\boldsymbol{t})f_{\mathcal{Y}}(\boldsymbol{s})||^2}{||\boldsymbol{t}||^{1+P}||\boldsymbol{s}||^{1+Q}} d\boldsymbol{t}d\boldsymbol{s}},$$

where $||\cdot||^2$ denotes the squared Euclidean norm, and constants c_P and c_Q are defined as $\pi^{(1+P)/2}/\Gamma((1+P)/2)$ and $\pi^{(1+Q)/2}/\Gamma((1+Q)/2)$, respectively, with $\Gamma(\cdot)$ the Gamma function, and π is the standard mathematical constant. This is a weighted L_2 norm, and is defined in such a way that the resulting distance correlation defined below is invariant to scale transformations $(\mathcal{X},\mathcal{Y}) \to \epsilon(\mathcal{X},\mathcal{Y})$ for positive ϵ. Furthermore, this definition ensures that $\mathrm{dCov}(\mathcal{X},\mathcal{Y}) = 0$ only if $\mathcal{X}$ and $\mathcal{Y}$ are independent, due to the inclusion of $||f_{\mathcal{X},\mathcal{Y}}(\boldsymbol{t},\boldsymbol{s}) - f_{\mathcal{X}}(\boldsymbol{t})f_{\mathcal{Y}}(\boldsymbol{s})||^2$ (Székely *et al.*, 2007).

The distance variance of $\mathcal{X}$ is defined as $\mathrm{dVar}(\mathcal{X}) = \mathrm{dCov}(\mathcal{X},\mathcal{X})$, and similarly for $\mathcal{Y}$. Distance correlation is then defined by substituting the expressions for dCov and

dVar into the classical definition of correlation to yield

$$\mathrm{dCor}(\mathcal{X},\mathcal{Y}) = \frac{\mathrm{dCov}(\mathcal{X},\mathcal{Y})}{\sqrt{\mathrm{dVar}(\mathcal{X})\mathrm{dVar}(\mathcal{Y})}}.$$

This ranges between 0 and 1, with 0 characterizing independence of $\mathcal{X}$ and $\mathcal{Y}$, and hence no association. The value of 1 indicates maximum association, and hence larger values provide evidence against the null hypothesis. If $P = Q = 1$, then $\mathrm{dCor}(\mathcal{X},\mathcal{Y}) \leq |\mathrm{cor}(\mathcal{X},\mathcal{Y})|$ with equality when $\mathrm{cor}(\mathcal{X},\mathcal{Y}) = \pm 1$.

Given centered observations $\boldsymbol{X}$ and $\boldsymbol{Y}$, empirical values of dCov, dVar and hence dCor can be obtained using Euclidean distances between the observations (Székely *et al.*, 2007). In particular, define the Euclidean distance matrices $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$, and apply a double-centering to these to yield the centered matrices $\boldsymbol{D}_{\mathcal{X}} = \boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}\boldsymbol{C}$ and $\boldsymbol{D}_{\mathcal{Y}} = \boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{Y}}\boldsymbol{C}$. The empirical value of dCov is then shown to be given by

$$\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{N}\sqrt{\mathrm{tr}\left(\boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{Y}}\right)},$$

so that the empirical value of dVar is given by

$$\mathrm{dVar}(\boldsymbol{X}) = \frac{1}{N}\sqrt{\mathrm{tr}\left(\boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{X}}\right)},$$

and similarly for $\boldsymbol{Y}$. The empirical value of dCor is then given by

$$\mathrm{dCor}(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\frac{\mathrm{tr}\left(\boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{Y}}\right)}{\sqrt{\mathrm{tr}\left(\boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{X}}\right)\mathrm{tr}\left(\boldsymbol{D}_{\mathcal{Y}}\boldsymbol{D}_{\mathcal{Y}}\right)}}}.$$

The $\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y})$ quantity is key in determining the empirical association between $\boldsymbol{X}$ and $\boldsymbol{Y}$. No association exists when $\mathrm{dCor}(\boldsymbol{X},\boldsymbol{Y}) = 0$ which occurs when $\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}) = 0$. Székely *et al.* (2007) show that $\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}) = 0$ equates to the empirical marginal and joint density functions of $\mathcal{X}$ and $\mathcal{Y}$ satisfying the definition of independence. Maximum association occurs when $\mathrm{dCor}(\boldsymbol{X},\boldsymbol{Y}) = 1$, which occurs if the double-centered Euclidean distance matrices are related via a scaling factor; $\boldsymbol{D}_{\mathcal{X}} = a\boldsymbol{D}_{\mathcal{Y}}$ for some non-zero constant a (Székely *et al.*, 2007). It has been shown that this equates to $\boldsymbol{Y}$ and $\boldsymbol{X}$ being equal up to a translation, rotation and scaling with factor a.

A permutation test based on dCov is used to test null hypothesis (4.2), rather than dCor. The permutation p-values of an observed dCor and dCov are identical, but dCov is used due to being computationally less expensive. The empirical p-value of an observed dCov statistic is found by comparing against the null sampling distribution obtained by permutations of the rows of one of the data matrices.

Many theoretical properties of this approach have been discussed in the literature; key papers include Székely *et al.* (2007) and Székely and Rizzo (2009). Two properties of practical significance which have been highlighted in the review article of Newton (2009) are the consistency of the test against all types of dependent alternatives, and no assumption of normality being required for valid inferences. The dCor test provides a method of detecting nonlinear relationships, which until recently have been considered 'beyond the scope of ordinary applied statistics' (Newton, 2009).

4.2 Distance-Based Approaches

4.2.1 Problem Statement

Suppose that either $\mathcal{X}$, or $\mathcal{Y}$, or both are comprised of a single random variable which is not scalar-valued. So for instance, $\mathcal{X}$ could be a graph-valued random variable while $\mathcal{Y}$ is either a random vector or a curve-valued random variable. Denote the N observations of $\mathcal{X}$ and $\mathcal{Y}$ by $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, respectively, and assume that suitably defined semi-metric or metric distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are defined yielding $N \times N$ distance matrices $\mathbf{\Delta}_{\mathcal{X}} = \{d_{\mathcal{X}}(x_i, x_j)\}_{i,j=1}^N$ and $\mathbf{\Delta}_{\mathcal{Y}} = \{d_{\mathcal{Y}}(y_i, y_j)\}_{i,j=1}^N$.

The problem entails inferring if an association exists between $\mathcal{X}$ and $\mathcal{Y}$ given distance matrices $\mathbf{\Delta}_{\mathcal{X}}$ and $\mathbf{\Delta}_{\mathcal{Y}}$. Typically, the null hypothesis is expressed as

$$H_0 : \mathbf{\Delta}_{\mathcal{X}} \neq a\mathbf{\Delta}_{\mathcal{Y}}, \tag{4.9}$$

for some positive constant a. The constant represents possible scaling differences between the elements of each distance matrix, which may arise because of the chosen distance measures. For example, distances between the observations of $\mathcal{X}$ may lie in $[0, 1]$, while distances between the observations of $\mathcal{Y}$ may not be confined to the same range (the minimum will be 0, but the maximum may not necessarily be 1). Under

this null hypothesis, the pairwise distances in the $\mathcal{X}$ space are not linearly related to those of the $\mathcal{Y}$ space. The alternative hypothesis is that the distance matrices are equal up to a constant.

The most common approach to testing this hypothesis is via the standardized Mantel test (Mantel, 1967). We review this method below, in addition to the less well-known related MDS (RMDS) (Arenas and Cuadras, 2004), η^2 (Cuadras, 2008) and PROTEST (Jackson, 1995) approaches.

4.2.2 The Standardized Mantel Test

A classical approach which can be used to test (4.9) is the Mantel test (Mantel, 1967) and its standardized version. The statistics associated with these tests provide measures of agreement between the distance elements of each matrix. In particular, they seek to quantify the degree to which clustering effects are exhibited in both matrices.

In its original form, the Mantel test statistic is computed as the sum of the element-by-element product of the $A = N(N-1)/2$ upper-triangular values of the two distance matrices, that is,

$$M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}}) = \sum_{i>j} d_{\mathcal{X}}(x_i, x_j) d_{\mathcal{Y}}(y_i, y_j),$$

(Mantel, 1967). A more widely used version of this is its standardized version, which has been proposed as a more interpretable statistic as it is bounded while $M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}})$ has no upper limit to quantify perfect association (see, for instance, Legendre and Legendre (1998) and Schneider and Borlund (2007)).

The standardized Mantel statistic is defined by applying the original Mantel statistic with standardized distance elements. In particular, the A distances $\{d_{\mathcal{X}}(x_i, x_j)\}_{i>j}$ are standardized by subtracting their mean and dividing by their standard deviation, i.e.,

$$\frac{d_{\mathcal{X}}(x_i, x_j) - \bar{x}}{s_x},$$

where $\bar{x} = \sum_{i>j} d_{\mathcal{X}}(x_i, x_j)/A$, $s_x^2 = \sum_{i>j}(d_{\mathcal{X}}(x_i, x_j) - \bar{x})^2/(A-1)$, and similarly for the distances $\{d_{\mathcal{Y}}(y_i, y_j)\}_{i>j}$. Although some of the distances are correlated (as there exists some dependence between them; for example, $d_{\mathcal{X}}(x_1, x_2)$ and $d_{\mathcal{X}}(x_1, x_3)$ both contain observation x_1), standardizing the upper-triangular distances in this way essentially considers the distances as independent observations of a random variable

with sample mean 0 and sample variance 1. Applying this standardization to both distance matrices maps their elements to a space where they can be directly compared.

The standardized Mantel statistic is then given by

$$r_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}}) = \frac{1}{A-1} \sum_{i>j} \left(\frac{d_{\mathcal{X}}(x_i, x_j) - \bar{x}}{s_x} \right) \left(\frac{d_{\mathcal{Y}}(y_i, y_j) - \bar{y}}{s_y} \right),$$

which equals Pearson's correlation coefficient between the A-dimensional vectors $\boldsymbol{d}_{\mathcal{X}}$ and $\boldsymbol{d}_{\mathcal{Y}}$ containing the standardized distance elements $\{(d_{\mathcal{X}}(x_i, x_j) - \bar{x})/s_x\}_{i>j}$ and $\{(d_{\mathcal{Y}}(y_i, y_j) - \bar{y})/s_y\}_{i>j}$, respectively. That is,

$$r_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}}) = \operatorname{cor}(\boldsymbol{d}_{\mathcal{X}}, \boldsymbol{d}_{\mathcal{Y}}),$$

so that $r_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}})$ is bounded by ± 1 and quantifies the linear correlation between the distances in each distance matrix. As in the case of Pearson's correlation coefficient, a value of -1 indicates a perfect negative correlation (distances in $\boldsymbol{\Delta}_{\mathcal{X}}$ are large when distances in $\boldsymbol{\Delta}_{\mathcal{Y}}$ are small, and vice versa), whereas 1 indicates perfect positive correlation (distances in $\boldsymbol{\Delta}_{\mathcal{X}}$ are large when distances in $\boldsymbol{\Delta}_{\mathcal{Y}}$ are large, and similarly for small distances). A value of 0 indicates no correlation. Thus values tending towards ± 1 provide evidence against the null hypothesis.

Inference is typically performed by using a Monte Carlo permutation procedure where the observed statistic, denoted $\hat{r}_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}})$, is compared against a permutation sampling distribution generated under the null. For each permutation $\pi \in \Pi$, the observed permuted statistic is computed as $\hat{r}_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y},\pi})$, where $\boldsymbol{\Delta}_{\mathcal{Y},\pi}$ denotes $\boldsymbol{\Delta}_{\mathcal{Y}}$ with rows and columns simultaneously permuted by π. The set $\{\hat{r}_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y},\pi})\}_{\pi \in \Pi}$ then defines the sampling distribution under the null. The p-value of the observed $\hat{r}_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}})$ can then be approximated by

$$\frac{\#(|\hat{r}_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y},\pi})| \geq |\hat{r}_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}})|)}{N_\pi},$$

where $|\cdot|$ is the absolute operator. Note that this is a right-tailed test since larger values of $|r_M(\boldsymbol{\Delta}_{\mathcal{X}}, \boldsymbol{\Delta}_{\mathcal{Y}})|$ indicate greater association. An alternative approach has also been proposed for large N, where the exact permutation distribution which would be obtained by using all possible permutations is approximated by the Normal dis-

tribution (Mantel, 1967). However, its use has been cautioned where the sampling distribution appears skewed (Mantel, 1967).

The standardized Mantel test has been widely used in the literature. Some examples include Dow *et al.* (1987), Heywood (1991) and Legendre and Legendre (1998). However, the test suffers from a limitation when $\mathcal{X}$ and $\mathcal{Y}$ are P- and Q-dimensional real-valued random vectors, respectively. In particular, when the N centered observations stored in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are such that $\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{0}$, so that $\mathcal{X}$ and $\mathcal{Y}$ are not associated via the traditional CCA or RV tests, say. In this case the standardized Mantel test will detect a linear relationship between their respective Euclidean distance matrices (when one does not exist between the raw data matrices). This discrepancy has been shown empirically (see, for example, Peres-Neto and Jackson (2001) and Section 7.6.1). A mathematical explanation has been offered recently by Legendre and Fortin (2010), based on the connection between the sum of squares components arising from a linear regression analysis of $\boldsymbol{d}_\mathcal{Y}$ on $\boldsymbol{d}_\mathcal{X}$ and Pearson's correlation between them, $\text{cor}(\boldsymbol{d}_\mathcal{X}, \boldsymbol{d}_\mathcal{Y})$.

In particular, a linear regression analysis of $\boldsymbol{d}_\mathcal{Y}$ on $\boldsymbol{d}_\mathcal{X}$ yields the sum of squares decomposition

$$\boldsymbol{d}_\mathcal{Y}^T\boldsymbol{d}_\mathcal{Y} = \boldsymbol{d}_\mathcal{Y}^T\boldsymbol{H}_{d_\mathcal{Y}|d_\mathcal{X}}\boldsymbol{d}_\mathcal{Y} + \boldsymbol{d}_\mathcal{Y}^T\left(\boldsymbol{I}_A - \boldsymbol{H}_{d_\mathcal{Y}|d_\mathcal{X}}\right)\boldsymbol{d}_\mathcal{Y},$$

where $\boldsymbol{H}_{d_\mathcal{Y}|d_\mathcal{X}} = \boldsymbol{d}_\mathcal{X}\boldsymbol{d}_\mathcal{X}^T\boldsymbol{d}_\mathcal{Y}/(\boldsymbol{d}_\mathcal{X}^T\boldsymbol{d}_\mathcal{X})$ is the $A \times A$ hat matrix. The sum of squares of $\boldsymbol{d}_\mathcal{Y}$ is represented by $\boldsymbol{d}_\mathcal{Y}^T\boldsymbol{d}_\mathcal{Y}$, the sum of squares explained by $\boldsymbol{d}_\mathcal{X}$ is represented by $\boldsymbol{d}_\mathcal{Y}^T\boldsymbol{H}_{d_\mathcal{Y}|d_\mathcal{X}}\boldsymbol{d}_\mathcal{Y}$, and the residual sum of squares (unexplained by $\boldsymbol{d}_\mathcal{X}$) is represented by $\boldsymbol{d}_\mathcal{Y}^T\left(\boldsymbol{I}_A - \boldsymbol{H}_{d_\mathcal{Y}|d_\mathcal{X}}\right)\boldsymbol{d}_\mathcal{Y}$. It can then be shown that

$$\text{cor}^2\left(\boldsymbol{d}_\mathcal{X}, \boldsymbol{d}_\mathcal{Y}\right) = \frac{\boldsymbol{d}_\mathcal{Y}^T\boldsymbol{H}_{d\mathcal{Y}|d_\mathcal{X}}\boldsymbol{d}_\mathcal{Y}}{\boldsymbol{d}_\mathcal{Y}^T\boldsymbol{d}_\mathcal{Y}}, \tag{4.10}$$

where the term on the right is known as the coefficient of determination (Weisberg, 1985).

Now consider a linear regression analysis of $\boldsymbol{Y}$ on $\boldsymbol{X}$, yielding sum of squares decomposition

$$\text{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right) = \text{tr}\left(\boldsymbol{H}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{H}\right) + \text{tr}\left(\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\boldsymbol{Y}\boldsymbol{Y}^T\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\right),$$

where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. The multivariate analogue of the coefficient of deter-

mination is given by

$$\frac{\text{tr}\left(\boldsymbol{H}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{H}\right)}{\text{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)}, \tag{4.11}$$

(Legendre and Fortin, 2010). Note that $\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{0}$ (since $\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{0}$), indicating no effect of $\boldsymbol{X}$ in explaining the sum of squares of $\boldsymbol{Y}$, but we do not simplify the expressions accordingly in this exposition.

The key point highlighted by Legendre and Fortin (2010) is that the sum of squares expressions in the denominators of (4.10) and (4.11) are not equal, that is, the total sum of squares of $\mathcal{Y}$ ($\text{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)$) is not equal to the sum of squares of the Euclidean distances between the observations of $\mathcal{Y}$ ($\boldsymbol{d}_{\mathcal{Y}}^T\boldsymbol{d}_{\mathcal{Y}}$). This can be seen directly by representing $\text{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)$ in terms of Euclidean distances and comparing with $\boldsymbol{d}_{\mathcal{Y}}^T\boldsymbol{d}_{\mathcal{Y}}$. From Section 3.2.2 we have that

$$\text{tr}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right) = \frac{1}{N}\sum_{i>j} d_{\mathcal{Y}}^2(\boldsymbol{y}_i, \boldsymbol{y}_j), \tag{4.12}$$

where $d_{\mathcal{Y}}$ is the Euclidean distance (as $\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{G}_{\mathcal{Y}}$), whereas the expression $\boldsymbol{d}_{\mathcal{Y}}^T\boldsymbol{d}_{\mathcal{Y}}$ equals

$$\sum_{i>j}\left(d_{\mathcal{Y}}(\boldsymbol{y}_i, \boldsymbol{y}_j) - \bar{y}\right)^2 = \sum_{i>j} d_{\mathcal{Y}}^2(\boldsymbol{y}_i, \boldsymbol{y}_j) - A\bar{y}^2. \tag{4.13}$$

Since (4.12) and (4.13) are not equal, the coefficients of determination do not measure the same relationship. It is therefore argued that in the case where the original vector-valued observations are available, a test of no association that operates directly on these, and not on the derived distances, should be used.

4.2.3 The RMDS Coefficient

The RMDS coefficient of Arenas and Cuadras (2004) measures association by considering the notion of a distance matrix which combines information from the individual distance matrices. This distance matrix, denoted $\boldsymbol{\Delta}_{\mathcal{X}\mathcal{Y}}$ and termed the joint distance matrix, satisfies certain properties relating to the corresponding principal coordinates arising from MDS. These represent an average configuration of the N sampling units with respect to the distances in both distance matrices, and can be used to measure the level of redundancy between the separate coordinates arising from $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$.

The properties satisfied by $\boldsymbol{\Delta}_{\mathcal{XY}}$ are given as follows. On altering one distance matrix by a multiplicative constant so that $\text{tr}(\boldsymbol{G}_{\mathcal{X}}) = \text{tr}(\boldsymbol{G}_{\mathcal{Y}})$, i.e., the respective variabilities are equal, obtain the centered inner product matrices $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ and the corresponding principal coordinate matrices $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$. If $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{Y}}$ so that $\boldsymbol{\Delta}_{\mathcal{X}} = \boldsymbol{\Delta}_{\mathcal{Y}}$, then $\boldsymbol{\Delta}_{\mathcal{XY}} = \boldsymbol{\Delta}_{\mathcal{X}} = \boldsymbol{\Delta}_{\mathcal{Y}}$. Only the principal coordinates from one distance matrix are required to fully represent the information provided by both distance matrices, so there is maximum redundancy (one distance matrix can be ignored). If $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} = \boldsymbol{0}$, this means that $\boldsymbol{\Lambda}_{\mathcal{X}}^{\frac{1}{2}}\boldsymbol{U}_{\mathcal{X}}^T\boldsymbol{U}_{\mathcal{Y}}\boldsymbol{\Lambda}_{\mathcal{Y}}^{\frac{1}{2}} = \boldsymbol{0}$ so that the directions of variability, $\boldsymbol{U}_{\mathcal{X}}$ and $\boldsymbol{U}_{\mathcal{Y}}$, are orthogonal. Then $\boldsymbol{\Delta}_{\mathcal{XY}}^2 = \boldsymbol{\Delta}_{\mathcal{X}}^2 + \boldsymbol{\Delta}_{\mathcal{Y}}^2$, and there is minimum redundancy since both distance matrices are required to yield an overall view of the N sampling units with respect to $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$. For intermediate cases where $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are not equal and not orthogonal, $\boldsymbol{\Delta}_{\mathcal{XY}}$ contains some form of average of the individual distance matrices.

$\boldsymbol{\Delta}_{\mathcal{XY}}$ is not explicitly defined by Arenas and Cuadras (2004). Instead, its centered inner product matrix, denoted $\boldsymbol{G}_{\mathcal{XY}}$, is defined;

$$\boldsymbol{G}_{\mathcal{XY}} = \boldsymbol{G}_{\mathcal{X}} + \boldsymbol{G}_{\mathcal{Y}} - \frac{1}{2}\left(\boldsymbol{G}_{\mathcal{X}}^{\frac{1}{2}}\boldsymbol{G}_{\mathcal{Y}}^{\frac{1}{2}} + \boldsymbol{G}_{\mathcal{Y}}^{\frac{1}{2}}\boldsymbol{G}_{\mathcal{X}}^{\frac{1}{2}}\right),$$

where $\boldsymbol{G}_{\mathcal{X}}^{\frac{1}{2}} = \boldsymbol{U}_{\mathcal{X}}\boldsymbol{\Lambda}_{\mathcal{X}}^{\frac{1}{2}}\boldsymbol{U}_{\mathcal{X}}^T$ and $\boldsymbol{G}_{\mathcal{Y}}^{\frac{1}{2}} = \boldsymbol{U}_{\mathcal{Y}}\boldsymbol{\Lambda}_{\mathcal{Y}}^{\frac{1}{2}}\boldsymbol{U}_{\mathcal{Y}}^T$ (from (3.7), $\boldsymbol{\Delta}_{\mathcal{XY}}$ can be obtained from $\boldsymbol{G}_{\mathcal{XY}}$ as $\boldsymbol{\Delta}_{\mathcal{XY}} = \left(\boldsymbol{d}_G\boldsymbol{1}_N^T + \boldsymbol{1}_N\boldsymbol{d}_G^T - 2\boldsymbol{G}_{\mathcal{XY}}\right)^{\frac{1}{2}}$ where $\boldsymbol{d}_G$ is the column vector containing the diagonal elements of $\boldsymbol{G}_{\mathcal{XY}}$). Note that $\boldsymbol{G}_{\mathcal{X}} = \boldsymbol{G}_{\mathcal{Y}}$ is equivalent to $\boldsymbol{\Delta}_{\mathcal{X}} = \boldsymbol{\Delta}_{\mathcal{Y}}$, and in this case $\boldsymbol{G}_{\mathcal{XY}} = \boldsymbol{G}_{\mathcal{X}} = \boldsymbol{G}_{\mathcal{Y}}$, which is equivalent to $\boldsymbol{\Delta}_{\mathcal{XY}}^2 = \boldsymbol{\Delta}_{\mathcal{X}}^2 = \boldsymbol{\Delta}_{\mathcal{Y}}^2$. If $\boldsymbol{U}_{\mathcal{X}}^T\boldsymbol{U}_{\mathcal{Y}} = \boldsymbol{0}$, then $\boldsymbol{G}_{\mathcal{XY}} = \boldsymbol{G}_{\mathcal{X}} + \boldsymbol{G}_{\mathcal{Y}}$ which is equivalent to $\boldsymbol{\Delta}_{\mathcal{XY}}^2 = \boldsymbol{\Delta}_{\mathcal{X}}^2 + \boldsymbol{\Delta}_{\mathcal{Y}}^2$. The distances in $\boldsymbol{\Delta}_{\mathcal{XY}}$ are with respect to unknown distance function $d_{\mathcal{XY}}$, and the sample variability with respect to $d_{\mathcal{XY}}$ can be quantified by $\text{tr}\left(\boldsymbol{G}_{\mathcal{XY}}\right)$. Due to the properties defined above, $\text{tr}\left(\boldsymbol{G}_{\mathcal{XY}}\right) = \text{tr}\left(\boldsymbol{G}_{\mathcal{X}}\right) = \text{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right)$ if $\boldsymbol{\Delta}_{\mathcal{X}} = \boldsymbol{\Delta}_{\mathcal{Y}}$, and $\text{tr}\left(\boldsymbol{G}_{\mathcal{XY}}\right) = \text{tr}\left(\boldsymbol{G}_{\mathcal{X}} + \boldsymbol{G}_{\mathcal{Y}}\right)$ if $\boldsymbol{U}_{\mathcal{X}}^T\boldsymbol{U}_{\mathcal{Y}} = \boldsymbol{0}$.

The RMDS coefficient compares the sample variabilities with respect to $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ against the sample variability observed with respect to $d_{\mathcal{XY}}$. It is defined as

$$\text{RMDS}(\mathcal{X}, \mathcal{Y}) = 2\left(1 - \frac{\text{tr}\left(\boldsymbol{G}_{\mathcal{XY}}\right)}{\text{tr}\left(\boldsymbol{G}_{\mathcal{X}} + \boldsymbol{G}_{\mathcal{Y}}\right)}\right),$$

and ranges between 0 and 1. The minimum is attained when $\boldsymbol{U}_{\mathcal{X}}^T\boldsymbol{U}_{\mathcal{Y}} = \boldsymbol{0}$, that is, when there is minimum redundancy as information from both distance matrices is

required. The maximum is attained when $\boldsymbol{\Delta}_{\mathcal{X}} = \boldsymbol{\Delta}_{\mathcal{Y}}$, that is, when there is maximum redundancy as one distance matrix can be ignored. Larger values provide evidence against the null hypothesis.

Although no procedure is described to assess the significance of an observed association value, a Monte Carlo permutation procedure can be invoked to achieve this. A sampling distribution of RMDS can be generated under the null hypothesis by simultaneously permuting the rows and columns of $\boldsymbol{G}_{\mathcal{Y}}$ (or $\boldsymbol{G}_{\mathcal{X}}$) many times, and each time recomputing $\boldsymbol{G}_{\mathcal{XY}}$ and hence the RMDS value. Comparing the observed RMDS value against this sampling distribution yields an estimate of the p-value.

A limitation of the RMDS coefficient is that $\boldsymbol{G}_{\mathcal{XY}}$ may not be well-defined, as it may not be real-valued. This occurs when some eigenvalues in $\boldsymbol{\Lambda}_{\mathcal{X}}$ and $\boldsymbol{\Lambda}_{\mathcal{Y}}$ are negative, resulting in complex-valued elements of $\boldsymbol{G}_{\mathcal{X}}^{\frac{1}{2}}$ and $\boldsymbol{G}_{\mathcal{Y}}^{\frac{1}{2}}$, and hence of $\boldsymbol{G}_{\mathcal{XY}}$. This would imply that some squared distances with respect to $d_{\mathcal{XY}}$ are complex-valued, and hence the distances are negative. To overcome this a correction for negative eigenvalues can be applied. However, this alters the information provided in the original distance matrices.

4.2.4 The η^2 Coefficient

The η^2 coefficient of Cuadras (2008) considers a reduced dimension approach to measuring association based on the directions of variability arising from MDS of $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$. It considers the orthogonality between the first few directions of variability of each, and uses the determinant operator to yield a scalar-valued measurement of association.

On denoting the first S columns of $\boldsymbol{U}_{\mathcal{X}}$ by $\boldsymbol{U}_{\mathcal{X},S}$, and the first K columns of $\boldsymbol{U}_{\mathcal{Y}}$ by $\boldsymbol{U}_{\mathcal{Y},K}$, the coefficient is defined as

$$\eta^2(\mathcal{X},\mathcal{Y}) = det\left(\boldsymbol{U}_{\mathcal{X},S}^T \boldsymbol{U}_{\mathcal{Y},K} \boldsymbol{U}_{\mathcal{Y},K}^T \boldsymbol{U}_{\mathcal{X},S}\right).$$

Cuadras (2008) has shown that this coefficient ranges between 0 and 1. It equals 0 if the standard coordinates are orthogonal representing no overlap in the most important S and K directions of variability of $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$, respectively. Conversely, it equals 1 if the chosen directions of variability are the same, and hence the variabilities with

respect to $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are explained by the same important directions of variability. Larger values provide evidence against the null hypothesis.

As with RMDS, there is no described procedure for assessing the significance of an observed association value. However, a Monte Carlo permutation procedure can be applied to generate a sampling distribution of η^2 under the null hypothesis, to which the observed η^2 value can be compared. This requires permuting the rows of either $\boldsymbol{U}_{\mathcal{X},S}$ or $\boldsymbol{U}_{\mathcal{Y},K}$, and recomputing the coefficient value many times.

A difficulty in applying this test, however, lies in the computation of the coefficient. No guidance is provided in choosing the number of dimensions S and K. It can be argued that it is natural to set $S = K = N$, so that there is no possible loss of information induced by considering $S, K < N$. This yields a coefficient value of 1 because $\boldsymbol{U}_{\mathcal{Y},K}\boldsymbol{U}_{\mathcal{Y},K}^T = \boldsymbol{U}_{\mathcal{X},S}\boldsymbol{U}_{\mathcal{X},S}^T = \boldsymbol{I}_N$, and hence is not an appropriate choice as then $\mathcal{X}$ and $\mathcal{Y}$ would always be perfectly associated.

4.2.5 PROTEST

The PROTEST procedure of Jackson (1995) refers to the application of Monte Carlo permutations to the well-known Procrustes procedure comparing two data matrices of interest (see, for example, Mardia *et al.* (1979)). The Procrustes procedure translates, rotates and dilates one matrix optimally to match the other. The corresponding PROTEST statistic is derived from a measure of the goodness-of-fit of the two matrices after transformation.

The PROTEST statistic can be applied to the two principal coordinate matrices, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$, arising from $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$ in order to test (4.9) (Gower, 1971). They are initially scaled such that their respective total variabilities are equal to 1, that is, $\text{tr}(\boldsymbol{G}_{\mathcal{X}}) = \text{tr}(\boldsymbol{G}_{\mathcal{Y}}) = 1$. This ensures that the same result is obtained regardless of which matrix is kept fixed during these transformations (as will be shown later in this section).

Since $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are column-centered, their centroids are at the origin so no translation is required. Thus, consider applying only the rotation and dilation transformations to $\tilde{\boldsymbol{Y}}$, represented mathematically by letting $\tilde{\boldsymbol{Y}}$ equal $r\boldsymbol{A}\tilde{\boldsymbol{Y}}$, where r is the dilation parameter and $\boldsymbol{A}$ is an $N \times N$ orthogonal rotation matrix. The aim is to find the optimal r and $\boldsymbol{A}$ such that the goodness-of-fit criterion, defined as the resid-

ual sum of squares between all paired observations, is minimized. This is equivalent to minimizing the squared Frobenius distance between $\tilde{\boldsymbol{X}}$ and $r\boldsymbol{A}\tilde{\boldsymbol{Y}}$, given by $d_F^2(\tilde{\boldsymbol{X}}, r\boldsymbol{A}\tilde{\boldsymbol{Y}}) = \text{tr}\left((\tilde{\boldsymbol{X}} - r\boldsymbol{A}\tilde{\boldsymbol{Y}})^T(\tilde{\boldsymbol{X}} - r\boldsymbol{A}\tilde{\boldsymbol{Y}})\right)$. Thus the optimization problem of interest can be stated as

$$\min_{r,A}\left\{d_F^2(\tilde{\boldsymbol{X}}, r\boldsymbol{A}\tilde{\boldsymbol{Y}})\right\} = \min_{r,A}\left\{\text{tr}\left(\boldsymbol{G}_{\mathcal{X}}\right) + r^2\text{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right) - 2r\text{tr}\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}\boldsymbol{A}\right)\right\}.$$

subject to $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_N$. The optimal dilation, $\hat{r}$, can be found immediately by differentiation of the objective function $(\text{tr}(\boldsymbol{G}_{\mathcal{X}}) + r^2\text{tr}(\boldsymbol{G}_{\mathcal{Y}}) - 2r\text{tr}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}\boldsymbol{A}))$, as $\hat{r} = \text{tr}\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}\boldsymbol{A}\right)/\text{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right)$. Note the term $\text{tr}(\boldsymbol{G}_{\mathcal{Y}})$ in the denominator; this leads to a different solution if applying the transformation to the $\tilde{\boldsymbol{X}}$ matrix instead. However, since $\text{tr}(\boldsymbol{G}_{\mathcal{Y}}) = \text{tr}(\boldsymbol{G}_{\mathcal{X}}) = 1$, the solutions are the same.

Given $\hat{r}$, the optimal $\hat{\boldsymbol{A}}$ is then found as the maximizer of $\text{tr}\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}\boldsymbol{A}\right)$ subject to $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_N$. It can be shown that by using Lagrange multipliers and the singular value decomposition of $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$ ($\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} = \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{U}^T$), $\hat{\boldsymbol{A}} = \boldsymbol{V}\boldsymbol{U}^T$ (Mardia *et al.*, 1979). The PROTEST coefficient is then defined as the minimum squared Frobenius distance, $d_F^2(\tilde{\boldsymbol{X}}, \hat{r}\hat{\boldsymbol{A}}\tilde{\boldsymbol{Y}})$, given by

$$\text{PROTEST}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = 1 - \text{tr}\left(\boldsymbol{\Gamma}\right)^2.$$

This ranges between 0 and 1, with perfect association indicated by a value of 0 when $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are linearly related, and no association indicated by the value of 1 when $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} = \boldsymbol{0}$. Thus smaller values provide evidence against the null hypothesis since the configurations are deemed less dissimilar after optimal transformation. The p-value of an observed association value can be approximated by permuting the rows of one of the data matrices and recomputing the coefficient value for many permutations.

A limitation of this procedure is in requiring a correction for negative eigenvalues if distance functions are semi-metric, as advocated by Peres-Neto and Jackson (2001). On applying a correction, the principal coordinate configurations can be represented in real-valued Euclidean space (as eigenvalues are non-negative), but at the cost of altering the information provided in the distance matrices to force such a configuration.

4.3 Summary

When $\mathcal{X}$ and $\mathcal{Y}$ are real and vector-valued, we have reviewed the suitable multivariate RV and dCor tests of no association. While the RV test considers linear relationships between the variables comprising each vector, dCor considers more general, nonlinear relationships. When $\mathcal{X}$ and/or $\mathcal{Y}$ are comprised of a single random variable which is not scalar-valued, such as a curve- or graph-valued random variable, and multivariate approaches cannot be applied, we have reviewed a handful of suitable distance-based approaches.

Of the distance-based approaches reviewed, only the standardized Mantel test has been applied to genetics data, and in particular to relate discrete-valued genetic polymorphism data to environmental variables (see, for instance, Legendre and Fortin (2010)). Thus, while a few distance-based approaches exist which are clearly advantageous over multivariate approaches when one has non-vector-valued observations, their utility has yet to be fully investigated in the field of bioinformatics.

Based on the strengths and weaknesses of each distance-based approach, we form a list of requirements which should comprise a 'good' testing procedure for null hypothesis (4.9), such that it may be easily applied in bioinformatics applications. Firstly, given orthogonal real and vector-valued data which is centered, a distance-based approach applied with Euclidean distances should yield no association. That is, the distance-based approach should yield equivalent results to non-distance-based approaches. The main reason for this is to assist in the understanding of how these methods work, and hence help foster their use. Standardized Mantel does not maintain this equivalence, but it appears that methods with multivariate foundations, such as PROTEST, do maintain this equivalence. Thus a good distance-based approach should be a generalization of a well-understood multivariate approach.

Secondly, no dimensionality reduction should be applied to standard or principal coordinates arising from distance matrices as this could lead to a loss of information. The η^2 coefficient requires such a dimensionality reduction, but no guidance is offered on how best to obtain the resulting reduced dimension. Standardized Mantel, RMDS and PROTEST do not require dimensionality reduction, and as such retain all information provided in the original distance matrices (unless a correction is applied for semi-metric distances).

Thirdly, for semi-metric distance functions no alterations to the distance matrices should be applied. These distort the observed distances, again, possibly leading to a loss of information. Alterations are required for PROTEST and RMDS, but not standardized Mantel or η^2. Since they are not known to be beneficial for the problem at hand (Pekalska and Duin, 2005), such extra computation should be applied with caution. In fact, we show in Section 7.6.3 that applying such corrections can lead to a loss of power for a given method by applying it with and without a correction.

Finally, inference should be drawn without permutations. Standardized Mantel and PROTEST are typically used with permutations (Schneider and Borlund, 2007), regardless of the Normal approximation available for the permutation distribution of standardized Mantel. Furthermore, the Normal approximation will be inappropriate when the sampling distribution appears skewed, as is often the case in practice (Mantel, 1967).

In Chapter 7 we propose a distance-based statistic to test null hypothesis (4.9) which is generalized from the RV coefficient. It satisfies the above requirements, including having an approximate null distribution which can model skewed distributions. We provide evidence that it performs competitively with the standardized Mantel and PROTEST approaches, and since no permutations are required to assess significance, is particularly suited to applications where many tests need to be performed. We also show that for a specific distance measure the statistic equals the dCor statistic. Thus it encompasses both the linear (RV) and nonlinear (dCor) multivariate statistics, and is thus able to test null hypotheses (4.1) and (4.2), respectively, without permutations.

Part II

Methodology

Chapter 5

Distance-Based Analysis of Variance: the DBF Test

In this chapter we derive a distance-based generalization of the MANOVA decomposition used for high-dimensional vector-valued data. An interpretable distance-based statistic using both within- and between-group distance information is then defined to test null hypothesis (2.4). We derive an approximate null sampling distribution allowing inference to be performed without permutations, and demonstrate its applicability to a wide range of real applications. Several simulation studies are performed to highlight key advantages over competing methods.

5.1 The Distance-Based Variance Decomposition

In this section we generalize the MANOVA decomposition (2.2) by first showing that it can be written in terms of Euclidean distances. The distance-based variance decomposition then results by substituting any distance in place of the Euclidean distance.

Begin by considering the quantity $\text{tr}(\boldsymbol{T})$ associated with vector-valued observations $\{\boldsymbol{y}_i\}_{i=1}^N$. It is a measure of spread found by summing the squared Euclidean distance of each observation to the population mean vector. This quantity can be equivalently

written using only pairwise Euclidean distances between observations, as follows:

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{T}) &= \sum_{i=1}^{N}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^T(\boldsymbol{y}_i-\bar{\boldsymbol{y}}),\\
&= \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{y}_i^T\boldsymbol{y}_i+\frac{1}{2}\sum_{j=1}^{N}\boldsymbol{y}_j^T\boldsymbol{y}_j-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{y}_i^T\boldsymbol{y}_j\\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{y}_i-\boldsymbol{y}_j)^T(\boldsymbol{y}_i-\boldsymbol{y}_j)\\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}d_E^2(\boldsymbol{y}_i,\boldsymbol{y}_j),
\end{aligned}$$

where d_E denotes the Euclidean distance. Thus, $\mathrm{tr}(\boldsymbol{T})$ is proportional to the sum of squared inter-point Euclidean distances between all N observations. This well-known connection shows that the total variability of a given set of vectorial observations, traditionally found using the population mean, can be computed using only the inter-point Euclidean distances (Gower and Krzanowski, 1999; Anderson, 2001). In an analogous manner, the within- and between-group variability quantities $\mathrm{tr}(\boldsymbol{W})$ and $\mathrm{tr}(\boldsymbol{B})$ can also be written in terms of squared Euclidean distances, as follows:

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{W}) &= \sum_{g=1}^{G}\sum_{i=1}^{N}\left(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_g\right)^T\left(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_g\right)I_{gi}\\
&= \sum_{g=1}^{G}\left(\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{y}_i^T\boldsymbol{y}_iI_{gi}+\frac{1}{2}\sum_{j=1}^{N}\boldsymbol{y}_j^T\boldsymbol{y}_jI_{gj}-\frac{1}{N_g}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{y}_i^T\boldsymbol{y}_jI_{gi}I_{gj}\right)\\
&= \sum_{g=1}^{G}\left(\frac{1}{2N_g}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\boldsymbol{y}_i-\boldsymbol{y}_j\right)^T\left(\boldsymbol{y}_i-\boldsymbol{y}_j\right)I_{gi}I_{gj}\right)\\
&= \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}d_E^2(\boldsymbol{y}_i,\boldsymbol{y}_j)\frac{I_{gi}I_{gj}}{N_g},
\end{aligned}$$

and since $\mathrm{tr}(\boldsymbol{B})=\mathrm{tr}(\boldsymbol{T})-\mathrm{tr}(\boldsymbol{W})$, we obtain

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{B}) &= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}d_E^2(\boldsymbol{y}_i,\boldsymbol{y}_j)-\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}d_E^2(\boldsymbol{y}_i,\boldsymbol{y}_j)\frac{I_{gi}I_{gj}}{N_g}\\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}d_E^2(\boldsymbol{y}_i,\boldsymbol{y}_j)\left(1-\sum_{g=1}^{G}N\frac{I_{gi}I_{gj}}{N_g}\right).
\end{aligned}$$

Generalizations of these quantities can be defined by replacing the Euclidean distance, d_E, with any distance $d_{\mathcal{Y}}$. Thus we can define the total variability of a set of observations $\{y_i\}_{i=1}^N$ with respect to distance $d_{\mathcal{Y}}$ as

$$T_{\boldsymbol{\Delta}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{\mathcal{Y}}^2(y_i, y_j),$$

the within-group variability with respect to $d_{\mathcal{Y}}$ as

$$W_{\boldsymbol{\Delta}} = \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{\mathcal{Y}}^2(y_i, y_j)\frac{I_{gi}I_{gj}}{N_g},$$

and the between-group variability with respect to $d_{\mathcal{Y}}$ as

$$B_{\boldsymbol{\Delta}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{\mathcal{Y}}^2(y_i, y_j)\left(1 - \sum_{g=1}^{G} N\frac{I_{gi}I_{gj}}{N_g}\right).$$

The total variability in the data captured by $T_{\boldsymbol{\Delta}}$ can hence be decomposed into the sum of two components quantifying within- and between-group variability. That is, $T_{\boldsymbol{\Delta}} = W_{\boldsymbol{\Delta}} + B_{\boldsymbol{\Delta}}$, analogously to the decomposition $\mathrm{tr}(\boldsymbol{T}) = \mathrm{tr}(\boldsymbol{W}) + \mathrm{tr}(\boldsymbol{B})$ used for high-dimensional vectorial data. This distance-based variability decomposition holds for any distance.

We can write $T_{\boldsymbol{\Delta}}$, $B_{\boldsymbol{\Delta}}$ and $W_{\boldsymbol{\Delta}}$ more compactly in matrix form by using the centered inner product matrix $\boldsymbol{G}_{\mathcal{Y}}$. This matrix contains all the information on the inter-point distances between the N observations, and is such that its trace equals $T_{\boldsymbol{\Delta}}$;

$$\begin{aligned}
\mathrm{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right) &= \mathrm{tr}\left(\left(\boldsymbol{I}_N - \frac{1}{N}\boldsymbol{J}_N\right)\left(-\frac{1}{2}\boldsymbol{\Delta}_{\mathcal{Y}}^2\right)\right) \\
&= \mathrm{tr}\left(-\frac{1}{2N}\boldsymbol{\Delta}_{\mathcal{Y}}^2 + \frac{1}{2N}\boldsymbol{J}_N\boldsymbol{\Delta}_{\mathcal{Y}}^2\right) \\
&= \frac{1}{2N}\mathrm{tr}\left(\boldsymbol{J}_N\boldsymbol{\Delta}_{\mathcal{Y}}^2\right) \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{\mathcal{Y}}^2(y_i, y_j).
\end{aligned}$$

Therefore we rewrite $T_{\boldsymbol{\Delta}}$ more conveniently as $\mathrm{tr}(\boldsymbol{G}_{\mathcal{Y}})$. For $W_{\boldsymbol{\Delta}}$ and $B_{\boldsymbol{\Delta}}$ we define the centered $N \times N$ matrix of constants encoding group membership of each observation

to one of the G groups as

$$\boldsymbol{H}_c = \begin{pmatrix} \frac{1}{N_1}\boldsymbol{J}_{N_1} & & & 0 \\ & \frac{1}{N_2}\boldsymbol{J}_{N_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{N_G}\boldsymbol{J}_{N_G} \end{pmatrix} - \frac{1}{N}\boldsymbol{J}_N, \tag{5.1}$$

where $\boldsymbol{J}_a$ is the square matrix of ones of size a. Since this matrix is centered, we have that $\boldsymbol{C}\boldsymbol{H}_c\boldsymbol{C} = \boldsymbol{H}_c$ for centering matrix $\boldsymbol{C}$, and we use this fact in the evaluation of the quantity $\text{tr}(\boldsymbol{H}_c\boldsymbol{G}_\mathcal{Y})$ to derive expressions for $W_\boldsymbol{\Delta}$ and $B_\boldsymbol{\Delta}$ in terms of $\boldsymbol{G}_\mathcal{Y}$. We have

$$\begin{aligned} \text{tr}(\boldsymbol{H}_c\boldsymbol{G}_\mathcal{Y}) &= \text{tr}\left(\left(\left(\begin{pmatrix} \frac{1}{N_1}\boldsymbol{J}_{N_1} & & & 0 \\ & \frac{1}{N_2}\boldsymbol{J}_{N_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{N_G}\boldsymbol{J}_{N_G} \end{pmatrix} - \frac{1}{N}\boldsymbol{J}_N\right)\left(-\frac{1}{2}\boldsymbol{\Delta}_\mathcal{Y}^2\right)\right)\right) \\ &= \frac{1}{2N}\text{tr}\left(\boldsymbol{J}_N\boldsymbol{\Delta}_\mathcal{Y}^2\right) - \frac{1}{2}\text{tr}\left(\begin{pmatrix} \frac{1}{N_1}\boldsymbol{J}_{N_1} & & & 0 \\ & \frac{1}{N_2}\boldsymbol{J}_{N_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{N_G}\boldsymbol{J}_{N_G} \end{pmatrix}\boldsymbol{\Delta}_\mathcal{Y}^2\right) \\ &= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_\mathcal{Y}^2(y_i, y_j) - \frac{1}{2}\sum_{g=1}^{G}\frac{1}{N_g}\sum_{i=1}^{N}\sum_{j=1}^{N} d_\mathcal{Y}^2(y_i, y_j)I_{gi}I_{gj} \\ &= T_\boldsymbol{\Delta} - W_\boldsymbol{\Delta}, \end{aligned}$$

and since $B_\boldsymbol{\Delta} = T_\boldsymbol{\Delta} - W_\boldsymbol{\Delta}$, we have that $B_\boldsymbol{\Delta} = \text{tr}(\boldsymbol{H}_c\boldsymbol{G}_\mathcal{Y})$. Also, since $W_\boldsymbol{\Delta} = T_\boldsymbol{\Delta} - B_\boldsymbol{\Delta}$, we find that $W_\boldsymbol{\Delta} = \text{tr}\left((\boldsymbol{I}_N - \boldsymbol{H}_c)\boldsymbol{G}_\mathcal{Y}\right)$.

5.2 The Distance-Based F Statistic

Making use of the distance-based variance decomposition above, we can generalize the Dempster trace criterion (given by (2.3)) by replacing $\text{tr}(\boldsymbol{B})$ with $B_\boldsymbol{\Delta}$ and $\text{tr}(\boldsymbol{W})$ with $W_\boldsymbol{\Delta}$. That is, we define the distance-based F (DBF) statistic as

$$F_\boldsymbol{\Delta} = \frac{B_\boldsymbol{\Delta}}{W_\boldsymbol{\Delta}} = \frac{\text{tr}(\boldsymbol{H}_c\boldsymbol{G}_\mathcal{Y})}{\text{tr}\left((\boldsymbol{I}_N - \boldsymbol{H}_c)\boldsymbol{G}_\mathcal{Y}\right)}. \tag{5.2}$$

Analogously to the Dempster trace criterion and Lawley-Hotelling trace statistic, this statistic considers a ratio of between- to within-group variability. Larger values provide evidence against the null hypothesis, as larger between-group variability and smaller within-group variability suggest that observations in the same group are more similar than observations in different groups. A statistic of similar form was proposed by Anderson (2001) for application in ecology, but with degrees of freedom divisors $G-1$ and $N-G$ in the numerator and denominator, respectively.

5.3 Connection with MANOVA Statistics

It can be shown that $F_{\boldsymbol{\Delta}}$ is monotonically related to several MANOVA statistics when the observations are Q-dimensional vectors.

When $Q = 1$ and the Euclidean distance is applied, upon which we denote $F_{\boldsymbol{\Delta}}$ by $F_{\boldsymbol{\Delta}_E}$, $F_{\boldsymbol{\Delta}_E}$ is identical to the classical one-way ANOVA F statistic, ignoring the degrees of freedom divisors $G-1$ and $N-G$ in the numerator and denominator, respectively. Thus,

$$F_{\boldsymbol{\Delta}_E}\left(\frac{N-G}{G-1}\right) \sim F_{G-1,N-G}$$

(Anderson, 2001).

For $Q > 1$ and $G > 2$, we can show that $F_{\boldsymbol{\Delta}}$ is related to the Lawley-Hotelling and Pillai-trace statistics by using distance measures involving the within-group and between-group total sum of squares matrices. That is, although the DBF statistic is derived based on Euclidean distances which do not account for correlation amongst the variables comprising $\mathcal{Y}$, this information can be incorporated by an appropriate choice of distance measure (when $N > Q$). We show this in the following proposition.

Proposition 1 *On defining the distance matrices* $\boldsymbol{\Delta}_W = \{d_W(\boldsymbol{y}_i, \boldsymbol{y}_j)\}_{i,j=1}^N$ *and* $\boldsymbol{\Delta}_T = \{d_T(\boldsymbol{y}_i, \boldsymbol{y}_j)\}_{i,j=1}^N$ *with*

$$d_W^2(\boldsymbol{y}_i, \boldsymbol{y}_j) = (\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{W}^{-1} (\boldsymbol{y}_i - \boldsymbol{y}_j) \quad \textit{and} \quad d_T^2(\boldsymbol{y}_i, \boldsymbol{y}_j) = (\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{T}^{-1} (\boldsymbol{y}_i - \boldsymbol{y}_j),$$

we have that

$$F_{\boldsymbol{\Delta}_W} = \frac{LH}{Q} \quad \textit{and} \quad F_{\boldsymbol{\Delta}_T} = \frac{PT}{Q - PT}.$$

Proof. $F_{\boldsymbol{\Delta}}$ can be rearranged in terms of only $T_{\boldsymbol{\Delta}}$ and $W_{\boldsymbol{\Delta}}$ as

$$F_{\boldsymbol{\Delta}} = \frac{T_{\boldsymbol{\Delta}}}{W_{\boldsymbol{\Delta}}} - 1. \tag{5.3}$$

We show that LH and PT can be written in terms of $T_{\boldsymbol{\Delta}_W}$ and $W_{\boldsymbol{\Delta}_T}$, respectively, and these expressions can be substituted into (5.3) to obtain the required relationships. We begin by re-writing LH as

$$\begin{aligned}
\text{LH} &= \text{tr}(\boldsymbol{W}^{-1}\boldsymbol{B}) \\
&= \text{tr}(\boldsymbol{W}^{-1}(\boldsymbol{T} - \boldsymbol{W})) \\
&= \text{tr}(\boldsymbol{W}^{-1}\boldsymbol{T}) - Q \\
&= \frac{1}{2N}\text{tr}\left(\boldsymbol{W}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^T\right) - Q \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left((\boldsymbol{y}_i - \boldsymbol{y}_j)^T\boldsymbol{W}^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j)\right) - Q \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_W^2(\boldsymbol{y}_i, \boldsymbol{y}_j) - Q \\
&= T_{\boldsymbol{\Delta}_W} - Q.
\end{aligned}$$

From this we have that $T_{\boldsymbol{\Delta}_W} = \text{LH} + Q$, which we substitute into (5.3) to obtain

$$F_{\boldsymbol{\Delta}_W} = \frac{\text{LH} + Q}{W_{\boldsymbol{\Delta}_W}} - 1.$$

On expanding $W_{\boldsymbol{\Delta}_W}$ we find that it equals Q, yielding

$$F_{\boldsymbol{\Delta}_W} = \frac{\text{LH}}{Q},$$

as required.

On following a similar argument with PT we obtain

$$\begin{aligned}
\text{PT} &= \text{tr}(\boldsymbol{T}^{-1}\boldsymbol{B}) \\
&= Q - \text{tr}(\boldsymbol{T}^{-1}\boldsymbol{W}) \\
&= Q - \frac{1}{2}\text{tr}\left(\boldsymbol{T}^{-1}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^T\left(\frac{I_{gi}I_{gj}}{N_g}\right)\right) \\
&= Q - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{y}_i - \boldsymbol{y}_j)^T\boldsymbol{T}^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j)\sum_{g=1}^{G}\frac{I_{gi}I_{gj}}{N_g} \\
&= Q - W_{\boldsymbol{\Delta}_T}.
\end{aligned}$$

Since $T_{\boldsymbol{\Delta}_T} = Q$ (easily shown by expansion), we find that

$$F_{\boldsymbol{\Delta}_T} = \frac{\text{PT}}{Q - \text{PT}},$$

as required.

■

For $G = 2$, it follows from this proposition and the fact that $T^2 = (N-2)\text{PT}/(1-\text{PT})$ (Rencher, 2002) that $F_{\boldsymbol{\Delta}}$ with the Mahalanobis-like distance d_T, denoted $F_{\boldsymbol{\Delta}_T}$, is monotonically related to Hotelling's T^2 statistic via the equation

$$T^2 = \frac{(N-2)QF_{\boldsymbol{\Delta}_T}}{1 + (1-Q)F_{\boldsymbol{\Delta}_T}}. \tag{5.4}$$

Expanding on this relationship, since we know that $(N-Q-1)T^2/((N-2)Q)$ has an F distribution under the null, it follows that

$$\frac{(N-Q-1)F_{\boldsymbol{\Delta}_T}}{1 + (1-Q)F_{\boldsymbol{\Delta}_T}} \sim F_{Q,N-Q-1}.$$

That is, a transformation of $F_{\boldsymbol{\Delta}_T}$ follows the F distribution with degrees of freedom Q and $N-Q-1$ under the null.

5.4 Inference

Given an observed value of the test statistic, $\hat{F}_{\boldsymbol{\Delta}}$, computed for any suitably chosen distance measure $d_{\mathcal{Y}}$, inference can be performed using a non-parametric approach.

That is, the p-value can be estimated using permutations. Given N_π permutations $\pi \in \Pi$, the set $\{\hat{F}_{\boldsymbol{\Delta}_\pi}\}_{\pi\in\Pi}$ is generated by recalculating $B_{\boldsymbol{\Delta}}$ for each permutation, denoted $\hat{B}_{\boldsymbol{\Delta}_\pi}$, and using the monotonic relationship

$$\hat{F}_{\boldsymbol{\Delta}_\pi} = \frac{\hat{B}_{\boldsymbol{\Delta}_\pi}}{T_{\boldsymbol{\Delta}} - \hat{B}_{\boldsymbol{\Delta}_\pi}}. \tag{5.5}$$

$T_{\boldsymbol{\Delta}}$ is fixed for all permutations so that permuted values of $F_{\boldsymbol{\Delta}}$ are monotonically related to permuted values of $B_{\boldsymbol{\Delta}}$. The p-value is then estimated as the proportion of the N_π permuted statistics greater than or equal to the observed $\hat{F}_{\boldsymbol{\Delta}}$, i.e.,

$$\frac{\#(\hat{F}_{\boldsymbol{\Delta}_\pi} \geq \hat{F}_{\boldsymbol{\Delta}})}{N_\pi}.$$

Clearly, this is a one-sided test, since larger values of $F_{\boldsymbol{\Delta}}$ provide evidence against the null.

As an alternative to this expensive permutation-based testing approach, we consider an approximate distribution for the null sampling distribution of $F_{\boldsymbol{\Delta}}$, as this would allow p-values to be well-approximated without permutations. Since $F_{\boldsymbol{\Delta}}$ is related to $B_{\boldsymbol{\Delta}}$ via (5.5), we first consider approximating the null distribution of $B_{\boldsymbol{\Delta}}$, that is, the between-group variability.

5.4.1 The Approximate Null Distribution of the Between-Group Variability

For general data structures and distance measures, the null sampling distribution of the DBF test statistic (5.2) is unknown. This is because the between-group variability quantity, $B_{\boldsymbol{\Delta}}$, which features in the statistic will, in general, follow some unknown distribution which depends on the specific distance measure being used (Mantel, 1967). On denoting the $(i,j)^{\text{th}}$ element of $\boldsymbol{H}_c$ by h_{ij} and recalling that $\boldsymbol{H}_c$ is centered, $B_{\boldsymbol{\Delta}}$ can be expressed as the weighted sum of squared distances

$$B_{\boldsymbol{\Delta}} = -\frac{1}{2}\sum_{i\neq j} d_{\mathcal{Y}}^2(y_i, y_j)h_{ij}.$$

Thus even if each $d^2_{\mathcal{Y}}(y_i, y_j)$, for $i \neq j$, was assumed to be a random variable with known distribution, $B_{\mathbf{\Delta}}$ would be a weighted sum of correlated and uncorrelated random variables, whose distribution would be difficult to evaluate. For instance, the problem of evaluating the sum of correlated and uncorrelated Chi-squared and Gamma random variables has been considered extensively (see, for example, Solomon and Stephens (1977) and Kourouklis and Moschopoulos (1985)). Although it has been argued that a quantity of the form of $B_{\mathbf{\Delta}}$ has the appearance of a U-statistic which is asymptotically normal (Mantel, 1967; Hoeffding, 1948), in our experience with different data types, even for large sample sizes, $B_{\mathbf{\Delta}}$ often appears to be skewed to various degrees.

To demonstrate this, we explore the empirical permutation distribution of $B_{\mathbf{\Delta}}$ for four real datasets involving different data structures and distances:

(i) Vectorial and real-valued data: the data consists of $Q = 50$ gene expression measurements observed on $N = 103$ biological samples from the Novartis multi-tissue dataset described in Monti *et al.* (2003). In this case $G = 4$, corresponding to four different tissues. For this dataset, we consider the Euclidean, Mahalanobis and Manhattan distances (details provided in Appendix B.1).

(ii) Vectorial and discrete-valued data: the data consists of $Q = 5$ randomly selected SNPs observed on $N = 254$ samples from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (see Section 8.2 for further details). The observation of each sample at each SNP is the number of minor alleles, taking one value in $\{0, 1, 2\}$. In this case $G = 2$, corresponding to the two groups being compared, healthy controls and Alzheimer's disease patients. Here we use the identity-by-state (IBS), Rogers and Tanimoto I, and Sokal and Sneath genetic distances. The IBS distance compares the number of minor alleles at each SNP in the set of SNPs, while the Rogers and Tanimoto I and Sokal and Sneath distances use a function of the total number of matches of minor alleles across the whole set of SNPs (see Appendix B.3 for further details).

(iii) Functional data (curves): the data consists of $N = 18$ gene expression functional data replicates for a randomly selected gene in a dataset on *M.tuberculosis* analyzed by Tailleux *et al.* (2008). In this case $G = 2$, corresponding to two different types of cell, and replicate time courses were observed at 4 time-points. These

were smoothed via cubic smoothing splines to yield the 18 replicate curves (Minas *et al.*, 2011). Figure 5.1 shows observed time courses and their fitted curves for two randomly selected genes. The L_2, Visual L_2 and Curvature distance measures are applied to this dataset. The L_2 measure captures the difference in magnitude between curves, the Visual L_2 measure captures their scale-invariant differences in shape, and the Curvature measure captures their difference in rate of change regardless of direction (see Appendix B.2 for further details).

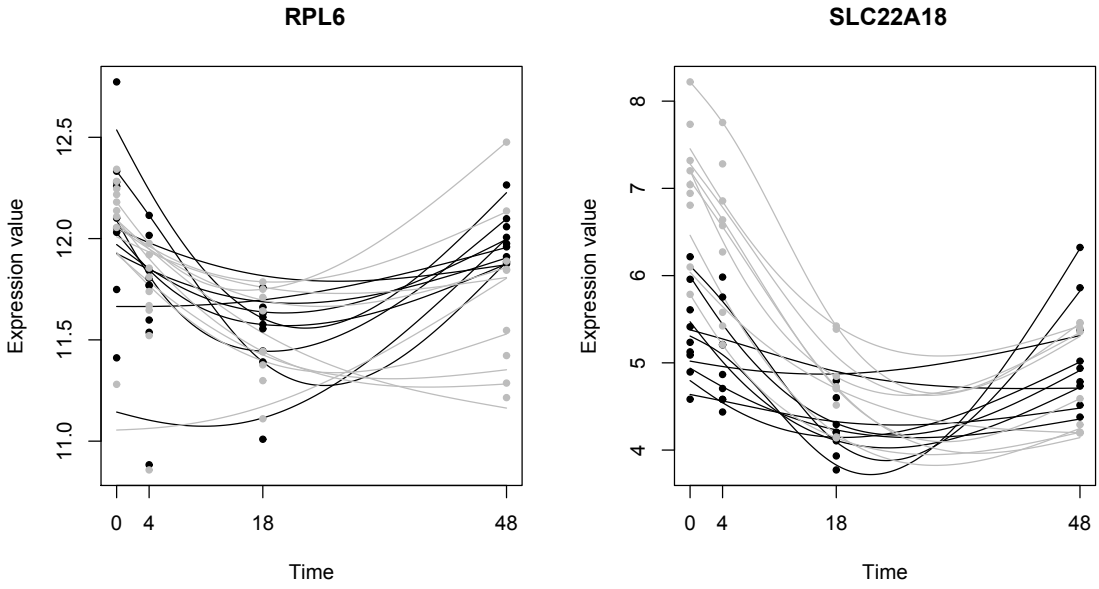

Figure 5.1: Replicate gene expression time courses modeled as time-dependent curves for genes RPL6 and SLC22A18 of the *M. tuberculosis* dataset; black for dendritic cells and gray for macrophages. The points represent the original gene expression time course measurements.

(iv) Graph data: the data consists of $N = 91$ graphs representing the functional connectivity networks from a functional MRI (fMRI) dataset on Schizophrenia described in Lord *et al.* (2011). In this case $G = 5$, corresponding to different levels of 'at-risk mental state' (ARMS) to which subjects can be diagnosed. Each graph is comprised of 19 vertices, with each representing a region of interest (ROI) across the brain. Figure 5.2 presents two graphs from this dataset; one observed on a control subject, and one observed on a subject with high ARMS, denoted ARMS-H. We apply the Hamming, Graph Edit, and Maximum Common Subgraph (MCS) distances. The Hamming distance captures the number of

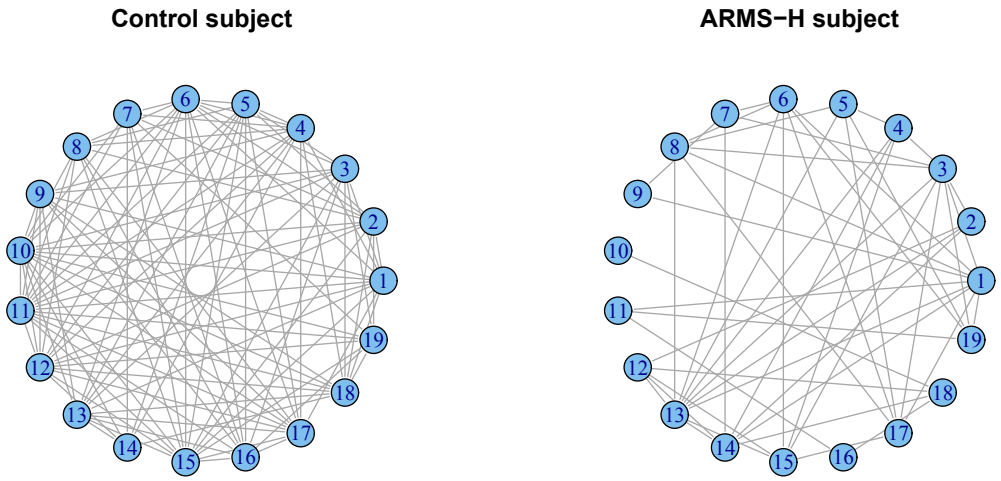

Figure 5.2: Sample control and ARMS-H brain connectivity graphs from the functional MRI dataset. Each circle represents a vertex, and the number within the circle denotes the corresponding ROI of the brain. The gray line connecting any pair of vertices represents an edge, and hence indicates some relationship between the two ROIs represented by the vertices. The control subject exhibits a much richer functional connectivity network between the ROIs than the ARMS-H subject, as indicated by the visibly larger number of edges.

common edges across any pair of graphs, the Graph Edit distance quantifies the number of edge deletions, insertions and substitutions required to transform one graph into another, and the MCS distance captures the proportional size of the maximum common subgraph between any pair of graphs (see Appendix B.4 for further details).

The exact permutation distribution of $B_{\mathbf{\Delta}}$ in each case would be given by the set $\{\hat{B}_{\mathbf{\Delta}_\pi}\}_{\pi\in\Pi}$ where Π contains all $N!$ permutations π of the elements of $\{1,\ldots,N\}$. Due to the computational effort required in enumerating all possible permutations, even for moderate size N, the exact distribution is generally unavailable. Figure 5.3 shows the approximate sampling distribution of $B_{\mathbf{\Delta}}$ obtained by using 10^6 Monte Carlo permutations for each of the data types and distance measures considered. The distributions exhibit varying degrees of skewness, even for large sample sizes.

Since the exact permutation distribution of $B_{\mathbf{\Delta}}$ is computationally and analytically

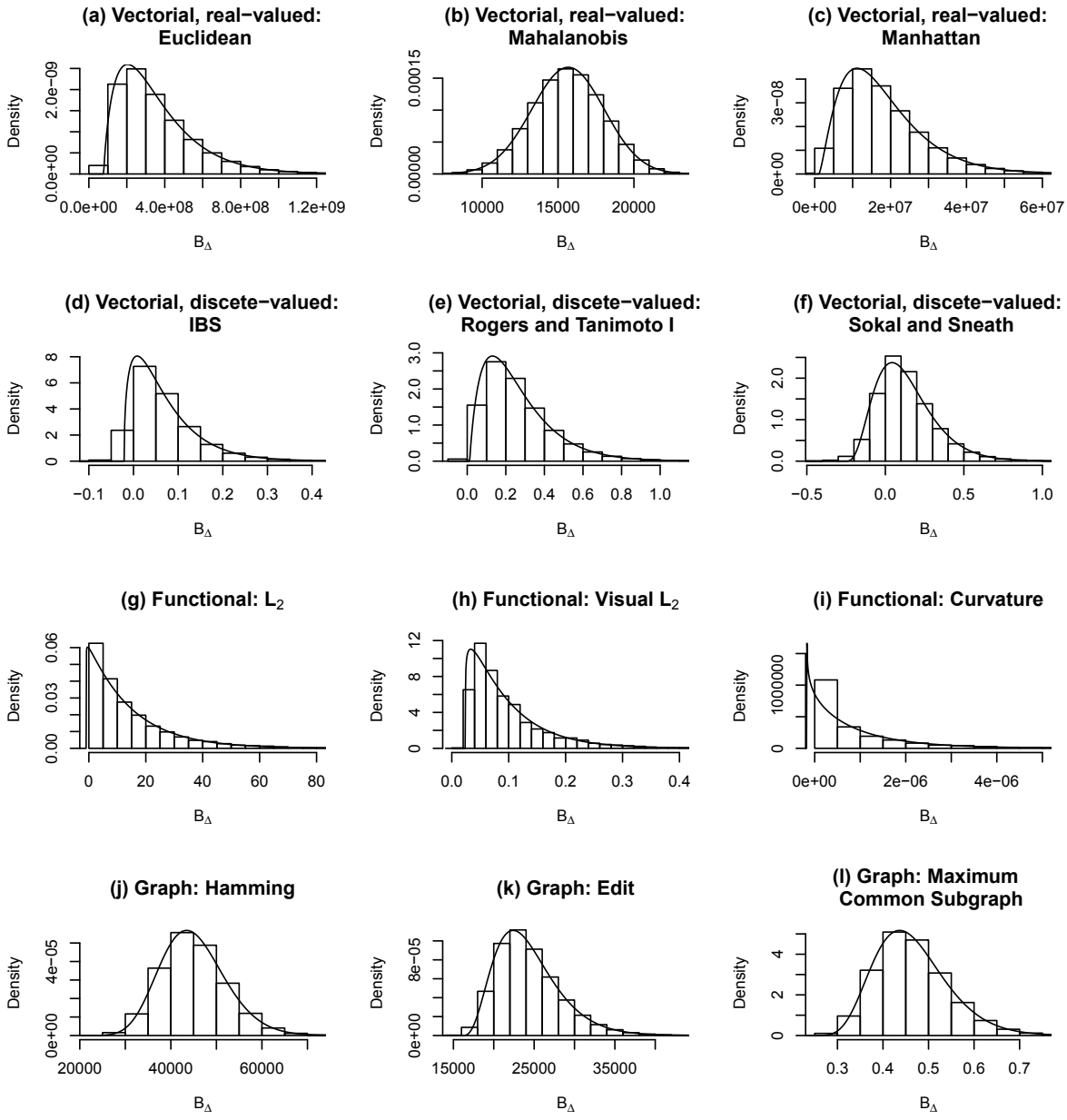

Figure 5.3: Sampling distributions of $B_{\mathbf{\Delta}}$ obtained by using 10^6 Monte Carlo permutations for four different data types and corresponding distances. (a)-(c) Vectorial and real-valued gene expression data with $N = 103$. (d)-(f) Vectorial and discrete-valued SNP data with $N = 254$. (g)-(i) Functional representation of longitudinal gene expression data with $N = 18$. (j)-(l) Graph representation of fMRI data with $N = 91$. Overlayed is the proposed approximate null probability density function described in Section 5.4.2.

intractable, we propose approximating it via moment matching. In moment matching the unknown distribution is approximated by a continuous distribution whose first few

moments 'match' those of the unknown distribution (Pearson, 1963; Johnson *et al.*, 1994).

The procedure is comprised of two steps. Firstly, the first three or four moments of the unknown distribution are either estimated or their exact values obtained. Exact values, which are obtained analytically, are preferred since they are not subject to sampling variability (Solomon and Stephens, 1978). Secondly, candidate distributions are considered for approximation. For example, Gamma and Lognormal distributions have both been applied to model the skewness observed in the sampling distributions of various multivariate and distance-based statistics (Berry and Mielke, 1983; Kazi-Aoual *et al.*, 1995; Josse *et al.*, 2008).

The choice of distribution is not limited to these, and systems of distributions have been proposed to 'provide approximations to as wide a variety of observed distributions as possible' (Johnson *et al.*, 1994). These systems are comprised of several distributions parameterized by the moments, often referred to as 'types'. For instance, the Pearson system (Pearson, 1895, 1901) proposed to capture different degrees of observed skewness is comprised of seven distributions. They encompass the Gamma, Beta, Exponential and Normal distributions by considering the first three or four moments. Another system often adopted is the Johnson system (Johnson, 1949), which is comprised of three types of distribution and encompasses the Lognormal distribution. This uses log-transformations of the first two moments of the variable of interest with the aim of removing skewness yielding a transformed distribution which appears normally distributed (see, for instance, Josse *et al.* (2008)). Other approaches include using polynomial expansions, such as the Gram-Charlier and Edgeworth expansions (see, for instance, Wallace (1958) and Johnson *et al.* (1994)). The expansion coefficients are given by the moments, and typically the first three moments are used (Josse *et al.*, 2008). However, these can yield negative densities over the support and can exhibit multimodal features. These are undesirable properties in application to observed data, prompting caution when used (see, for instance, Barton and Dennis (1952), Johnson *et al.* (1994) and Josse *et al.* (2008)).

The choice of distribution within a given system depends on practical considerations such as ease of implementation and theoretical arguments in their favour. We use the Pearson type III distribution, which encompasses the Gamma, Exponential

and Normal distributions, as it is flexible enough to capture the varying degrees of skewness often observed in real data (see Figure 5.3). While the distribution of $B_{\boldsymbol{\Delta}}$ is skewed for many distances, it also exhibits negligible skewness for some distances. For example, Figures 5.3 (b) and (j) demonstrate that $B_{\boldsymbol{\Delta}}$ appears normally distributed for the Mahalanobis and Hamming distances, and these can be captured by the Pearson type III distribution.

Furthermore, it retrieves the distributions expected in special cases of data and distance measure. To see this, recall from Section 5.2 that the ANOVA F and Hotelling's T^2 statistics follow F distributions under the null when the observed data is normally distributed. The DBF statistic is monotonically related to these (with the relevant distance measures), so that it follows the F distribution; indeed, this is supported by our simulation results in Section 5.5.1. Since the F distribution arises from the ratio of two Chi-squared and hence Gamma distributions, it follows that $B_{\boldsymbol{\Delta}}$ (and hence $W_{\boldsymbol{\Delta}}$) follows the Gamma distribution, and this is encompassed within the Pearson type III distribution.

Using the Pearson type III distribution to approximate the distribution of $B_{\boldsymbol{\Delta}}$ requires the mean, variance and skewness of the exact permutation distribution of $B_{\boldsymbol{\Delta}}$, which are given by

$$\mu_B = \frac{1}{N!}\sum_{\pi\in\Pi}\hat{B}_{\boldsymbol{\Delta}_\pi}, \quad \sigma_B^2 = \frac{1}{N!}\sum_{\pi\in\Pi}\hat{B}^2_{\boldsymbol{\Delta}_\pi} - \mu_B^2 \quad \text{and} \quad \gamma_B = \frac{\frac{1}{N!}\sum_{\pi\in\Pi}\hat{B}^3_{\boldsymbol{\Delta}_\pi} - 3\mu_B\sigma_B^2 - \mu_B^3}{\sigma_B^3},$$

respectively. Kazi-Aoual *et al.* (1995) have evaluated these expressions analytically and provide closed form manipulations allowing their efficient computation for $N > 6$ without the need for permutations. These closed form expressions require that $\boldsymbol{H}_c$ and $\boldsymbol{G}_{\mathcal{Y}}$ are square, symmetric and centered, which they are by definition. The expressions for μ_B, σ_B^2 and γ_B are provided in Appendix C.

On standardizing $B_{\boldsymbol{\Delta}}$ by subtracting μ_B and dividing by σ_B, the Pearson type III distribution is parameterized by the skewness γ_B. That is,

$$B^s_{\boldsymbol{\Delta}} = \frac{B_{\boldsymbol{\Delta}} - \mu_B}{\sigma_B} \sim PT_{III}\left(\gamma_B\right),$$

where PT_{III} denotes the Pearson type III distribution. By assumption of this dis-

tribution, the support of random variable $B^s_{\boldsymbol{\Delta}}$ is given by $[-2/\gamma_B, \infty)$ if $\gamma_B > 0$, $(-\infty, -2/\gamma_B]$ if $\gamma_B < 0$, and $(-\infty, \infty)$ if $\gamma_B = 0$. We denote the cumulative distribution function (CDF) of $B^s_{\boldsymbol{\Delta}}$ by $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(b; \gamma_B)$, and the probability density function (PDF) of $B^s_{\boldsymbol{\Delta}}$ by $f_{B^s_{\boldsymbol{\Delta}}}(b; \gamma_B)$. The PDF $f_{B^s_{\boldsymbol{\Delta}}}(b; \gamma_B)$ is defined by

$$\frac{(2/\gamma_B)^{4/\gamma_B^2}}{\Gamma\left(4/\gamma_B^2\right)} \left(\frac{2+\gamma_B b}{\gamma_B}\right)^{(4-\gamma_B^2)/\gamma_B^2} \exp\left(-\frac{2(2+\gamma_B b)}{\gamma_B^2}\right)$$

for $\gamma_B > 0$ and $-2/\gamma_B \leq b < \infty$, where $\Gamma(\cdot)$ denotes the usual Gamma function,

$$\frac{(-2/\gamma_B)^{4/\gamma_B^2}}{\Gamma\left(4/\gamma_B^2\right)} \left(\frac{-(2+\gamma_B b)}{\gamma_B}\right)^{(4-\gamma_B^2)/\gamma_B^2} \exp\left(-\frac{2(2+\gamma_B b)}{\gamma_B^2}\right)$$

for $\gamma_B < 0$ and $-\infty < b \leq -2/\gamma_B$, and

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{b^2}{2}\right)$$

for $\gamma_B = 0$, i.e., the standard Normal distribution (Mielke and Berry, 2007).

5.4.2 The Approximate Null Distribution of the DBF Statistic

We aim to approximate the null distribution of $F_{\boldsymbol{\Delta}}$ in terms of the distribution of $B^s_{\boldsymbol{\Delta}}$ by using the one-to-one function $h : B^s_{\boldsymbol{\Delta}} \mapsto F_{\boldsymbol{\Delta}}$ defined by

$$h\left(B^s_{\boldsymbol{\Delta}}\right) = \frac{\mu_B + \sigma_B B^s_{\boldsymbol{\Delta}}}{T_{\boldsymbol{\Delta}} - \mu_B - \sigma_B B^s_{\boldsymbol{\Delta}}}, \tag{5.6}$$

with inverse $h^{-1} : F_{\boldsymbol{\Delta}} \mapsto B^s_{\boldsymbol{\Delta}}$ defined by

$$h^{-1}\left(F_{\boldsymbol{\Delta}}\right) = \frac{\left(T_{\boldsymbol{\Delta}} - \mu_B\right) F_{\boldsymbol{\Delta}} - \mu_B}{\sigma_B\left(1 + F_{\boldsymbol{\Delta}}\right)}. \tag{5.7}$$

Transformation h must be continuous over the support of $B^s_{\boldsymbol{\Delta}}$. We have observed that for real datasets γ_B is not equal to 0 exactly (see Figure 5.3) so we only consider the cases where $\gamma_B > 0$ and $\gamma_B < 0$, and do not consider the case of $\gamma_B = 0$ in this exposition.

Transformation h is not continuous in the positive plane at $\beta = (T_{\boldsymbol{\Delta}} - \mu_B)/\sigma_B$ because $T_{\boldsymbol{\Delta}} = \operatorname{tr}(\boldsymbol{G}_y) > \mu_B$ due to $\operatorname{tr}(\boldsymbol{H}_c) = 1$. The boundary of the support of $B^s_{\boldsymbol{\Delta}}$

depends on the skewness, so the position of β may or may not affect the continuity of the distribution of $B^s_{\boldsymbol{\Delta}}$ over the support for the particular case of skewness. We thus consider dealing with the discontinuity separately for both cases of skewness.

First consider the positive skewness case, where the support of $B^s_{\boldsymbol{\Delta}}$ is $[-2/\gamma_B, \infty)$. Since $\gamma_B > 0$, $-2/\gamma_B$ is negative and the discontinuity can be observed as $B^s_{\boldsymbol{\Delta}}$ increases form $-2/\gamma_B$ to ∞. Figure 5.4 (a) shows how $F_{\boldsymbol{\Delta}}$ behaves as a function of $B^s_{\boldsymbol{\Delta}}$ over this support; $F_{\boldsymbol{\Delta}}$ is an increasing function of $B^s_{\boldsymbol{\Delta}}$ on both sides of the discontinuity at β.

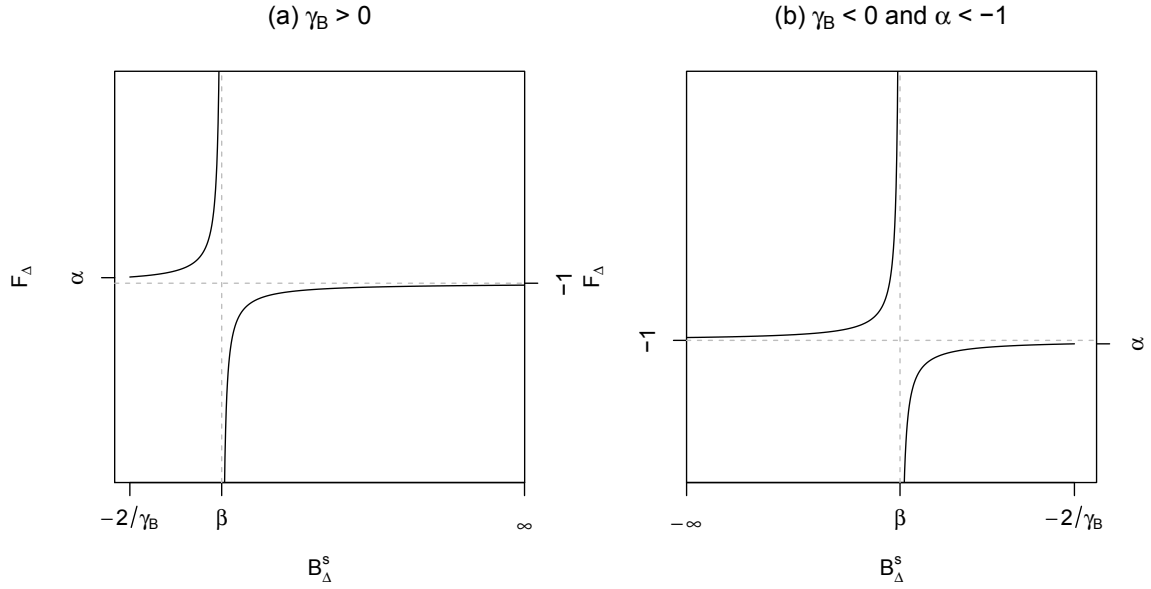

Figure 5.4: $F_{\boldsymbol{\Delta}}$ as a function of $B^s_{\boldsymbol{\Delta}}$. (a) $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ are monotonically related everywhere except at $B^s_{\boldsymbol{\Delta}} = \beta$ for $\gamma_B > 0$ over the support $[-2/\gamma_B, \infty)$. (b) $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ are monotonically related everywhere except at $B^s_{\boldsymbol{\Delta}} = \beta$ for $\gamma_B < 0$ over the support $(-\infty, -2/\gamma_B]$ when $\alpha < -1$.

We thus divide the support of $B^s_{\boldsymbol{\Delta}}$ into the two regions

$$\left[\frac{-2}{\gamma_B}, \beta\right) \quad \text{and} \quad (\beta, \infty)\,,$$

where the equivalent regions of support of $F_{\boldsymbol{\Delta}}$ are given by $[\alpha, \infty)$, where

$$\alpha = \frac{\gamma_B \mu_B - 2\sigma_B}{\gamma_B \left(T_{\boldsymbol{\Delta}} - \mu_B\right) + 2\sigma_B} \tag{5.8}$$

satisfies $h(\alpha) = -2/\gamma_B$, and $(-\infty, -1)$. We can show by contradiction that $\alpha > -1$,

since $\alpha \leq -1$ implies $T_{\boldsymbol{\Delta}}\gamma_B \leq 0$, and by definition both $T_{\boldsymbol{\Delta}}$ and γ_B are positive. In these regions we can apply the transformation since there are no discontinuities, and define the CDF of $F_{\boldsymbol{\Delta}}$ in terms of the CDF of $B^s_{\boldsymbol{\Delta}}$ by

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) = \begin{cases} \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) & -\infty < f < -1 \\ 1 + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) & \alpha \leq f < \infty \end{cases} \tag{5.9}$$

for $\gamma_B > 0$. The derivations are provided in Appendix D.1, including a proof that this is a valid CDF.

Now we turn our attention to the negative skewness case, where the support of $B^s_{\boldsymbol{\Delta}}$ is $(-\infty, -2/\gamma_B]$. In this case, $\gamma_B < 0$ so that $-2/\gamma_B$ is positive. The discontinuity at $B^s_{\boldsymbol{\Delta}} = \beta$ thus only needs to be considered if β is to the left of $-2/\gamma_B$, otherwise it can be ignored since it is not included in the support of $B^s_{\boldsymbol{\Delta}}$. We consider these two cases separately. First consider the case where $\beta < -2/\gamma_B$. We have that

$$\beta < \frac{-2}{\gamma_B} \Rightarrow \frac{-\mu_B}{\sigma_B} < \frac{T_{\boldsymbol{\Delta}} - \mu_B}{\sigma_B} < \frac{-2}{\gamma_B},$$

since $T_{\boldsymbol{\Delta}} > \mu_B$, from which we find

$$0 < \frac{\gamma_B T_{\boldsymbol{\Delta}}}{\gamma_B \mu_B - 2\sigma_B} < 1. \tag{5.10}$$

Applying this with the equation for α given by (5.8) yields $\alpha < -1$. Thus we define the occurrence of this first case when $\alpha < -1$. Figure 5.4 (b) shows how $F_{\boldsymbol{\Delta}}$ behaves as a function of $B^s_{\boldsymbol{\Delta}}$ over the support $(-\infty, -2/\gamma_B]$ when $\alpha < -1$. As with the positive skewness case, $F_{\boldsymbol{\Delta}}$ is an increasing function of $B^s_{\boldsymbol{\Delta}}$ on both sides of the discontinuity at β. We thus divide the support of $B^s_{\boldsymbol{\Delta}}$ into the two regions

$$(-\infty, \beta) \quad \text{and} \quad \left(\beta, \frac{-2}{\gamma_B}\right],$$

with equivalent supports of $F_{\boldsymbol{\Delta}}$ given by $(-1, \infty)$ and $(-\infty, \alpha]$, in which $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$

are monotonically related with no discontinuities. Thus we define the CDF of $F_{\boldsymbol{\Delta}}$ as

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) = \begin{cases} \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) & -\infty < f \leq \alpha \\ 1 + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) & -1 < f < \infty \end{cases} \tag{5.11}$$

for $\gamma_B < 0$ and $\alpha < -1$. The derivations are provided in Appendix D.2, including a proof that this is a valid CDF.

Now consider the case where $\beta > -2/\gamma_B$; this is equivalent to $\alpha > -1$. In this case there are no discontinuities in the support of $B^s_{\boldsymbol{\Delta}}$, so $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ are monotonically related everywhere. The support for $B^s_{\boldsymbol{\Delta}}$ given by $(-\infty, -2/\gamma_B]$ is equivalent to the support for $F_{\boldsymbol{\Delta}}$ of $(-1, \alpha]$. Thus the CDF of $F_{\boldsymbol{\Delta}}$ is defined as

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) = \begin{cases} 0 & -\infty < f \leq -1 \\ \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B) & -1 < f \leq \alpha \\ 1 & \alpha < f < \infty \end{cases} \tag{5.12}$$

for $\gamma_B < 0$ and $\alpha > -1$. This is a valid CDF as $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\cdot;\gamma_B)$ is a valid CDF.

Using these results, the approximate p-value of an observed $\hat{F}_{\boldsymbol{\Delta}}$ can be readily obtained without permutations. On computing the permutational mean μ_B, variance σ_B^2, and skewness γ_B, and additionally α if $\gamma_B < 0$, the p-value is given by $1 - \mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(\hat{F}_{\boldsymbol{\Delta}};\mu_B,\sigma_B,\gamma_B\right)$.

For the given case of skewness and α value, the PDF of $F_{\boldsymbol{\Delta}}$, denoted $f_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B)$, is given in terms of $f_{B^s_{\boldsymbol{\Delta}}}(\cdot;\gamma_B)$ by differentiating the CDF. Thus we have that

$$\begin{aligned} f_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) &= \left|\frac{d}{df}h^{-1}(f)\right| f_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B) \\ &= \frac{T_{\boldsymbol{\Delta}}}{\sigma_B(1+f)^2} f_{B^s_{\boldsymbol{\Delta}}}(h^{-1}(f);\gamma_B), \end{aligned}$$

where the range of f is given by the selected case of CDF.

5.5 Simulation Experiments

We provide a range of simulation results for the DBF test. In Section 5.5.1 we present empirical evidence in support of the results stated in Section 5.3, showing that the approximate null distribution matches those of the ANOVA F and Hotelling's T^2 tests. Section 5.5.2 details a power study demonstrating the competitiveness of the DBF test with the Mantel test and existing tests for the specific problem of testing a null hypothesis of equality between functions (curves). In Section 5.5.3 we illustrate how the approximate null distribution of the DBF statistic compares with the Monte Carlo permutation distribution for a number of data types and distances. In Section 5.5.4 we compare the permutation and approximation approaches of performing inference of an observed DBF statistic.

5.5.1 Comparison of DBF with ANOVA and MANOVA

Given that $F_{\mathbf{\Delta}}$ equals the ANOVA F statistic up to a constant as a special case for univariate data, we verify that the proposed approximate distribution of $F_{\mathbf{\Delta}}$ approximates that of the ANOVA F statistic well as N and G increase. Also, since $F_{\mathbf{\Delta}}$ is related to Hotelling's T^2 as a special case for multivariate data with $G = 2$, we verify that our proposed distribution, transformed via (5.4), approximates the distribution of T^2 well as N increases. That is, we aim to show that for the special cases the DBF test is approximately equivalent to the ANOVA F and Hotelling's T^2 tests, respectively.

For the univariate case, data is generated under the null and the DBF statistic with the Euclidean distance and the ANOVA F statistic are computed. P-values are found by comparing against their respective distributions. For $N = 40, 100, 500, 1000$ and $G = 2, 4, 5$, the k^{th} Monte Carlo run consists of simulating $y_1, \ldots, y_N \sim N(\mu_k, \sigma_k^2)$, where $\mu_k \sim U(-10, 10)$, $\sigma_k^2 \sim U(0, 10)$ (where $U(a, b)$ denotes the Uniform distribution over $[a, b]$). The mean and standard deviation of the absolute differences between the p-values obtained for $B = 200$ Monte Carlo simulations are reported in Table 5.1. It can be seen that as N and G increase, the absolute difference between the p-values decreases, thus showing that the approximate distribution of the DBF statistic behaves as expected in this case.

For the multivariate case, the DBF statistic using the Mahalanobis-like distance

Table 5.1: Mean (and standard deviation) of the absolute differences between p-values of the DBF statistic, and ANOVA F ($Q = 1$) and Hotelling's T^2 ($Q = 10$) statistics, under the null for 200 Monte Carlo runs.

	$Q = 1$				$Q = 10$	
	$G = 2$		$G = 4$	$G = 5$	$G = 2$	
N						
40	0.0252	(0.0342)	0.00670 (0.00535)	0.00565 (0.00464)	0.004702	(0.00272)
100	0.0137	(0.0225)	0.00306 (0.00267)	0.00254 (0.00196)	0.00200	(0.00121)
500	0.00474	(0.00951)	0.000594 (0.000525)	0.000541 (0.000383)	0.000392	(0.000260)
1000	0.00276	(0.00599)	0.000344 (0.000272)	0.000278 (0.000196)	0.000212	(0.000132)

measure d_T (defined in Proposition 1) and the Hotelling's T^2 statistic are computed. P-values are found by comparing against their respective distributions. For $N = 40, 100, 500, 1000$ and $Q = 10$, the k^{th} Monte Carlo run consists of simulating $\boldsymbol{y}_1, \ldots \boldsymbol{y}_N \sim N_Q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kP})^T$ with $\mu_{jk} \sim U(-6, 6)$ for $j = 1, \ldots, Q$, and $\boldsymbol{\Sigma}_k$ a random Wishart matrix of size $Q \times Q$. The mean and standard deviation of the absolute differences between the p-values obtained for $B = 200$ Monte Carlo runs are reported in Table 5.1. As N increases the difference between the p-values decreases, showing that the DBF and Hotelling's T^2 tests are approximately equivalent as N increases.

A further experiment is performed to show that, as N increases, the proposed approximate null distribution of the DBF statistic approximates the true ANOVA F and Hotelling's T^2 distributions, on applying the required transformations. In particular, we show that it yields a better approximation than a permutation-based CDF, especially when the number of permutations is low.

For $Q = 1$, $G = 2$, and each of $N = 50, 70$, one set of univariate observations is generated under the null from a Normal distribution as above. The DBF null CDF, suitably transformed, and the ANOVA F CDF are obtained, and the Kolmogorov-Smirnov (KS) statistic is used to compute the difference between these distributions. This statistic is computed as the maximum distance between two vectors representing the CDFs of interest; we use a vector of 1000 equally spaced points across the range of the approximate DBF distribution. For the given dataset for each N, and for each of $B = 200$ Monte Carlo runs, we use an increasing set of Monte Carlo permutations to compute the permutation CDF of the DBF statistic. We use $10^3, 10^4, 5 \times 10^4$ and 10^5 permutations, so that for each Monte Carlo run, 10^3 Monte Carlo permutations are enumerated, then 9×10^3 Monte Carlo permutations are added to yield the larger set of 10^4 permutations and so on. For each of these four sets of permutations the KS statistic depicting the difference between the DBF permutation CDF and the ANOVA F CDF is computed. This yields an empirical distribution of 200 KS statistic values for each set of permutations. The results of this experiment are shown in Figures 5.5 (a) and (b). We see that for $N = 50$, using more than 5×10^4 permutations yields a permutation distribution which is directly comparable with our approximate distribution. For $N = 70$, however, the approximate DBF distribution better approximates the true

underlying ANOVA F distribution than the permutation distributions typically used in practice; typically not more than 10^5 permutations are used for real data analyses.

For $Q = 10$, $G = 2$ and $N = 50$, one set of multivariate observations is generated under the null from a Multivariate Normal distribution as described above. The DBF null CDF, suitably transformed, and the Hotelling's T^2 CDF are obtained. Repeating as above, and using the KS statistic to quantify the difference between the transformed DBF permutation CDF and true Hotelling's T^2 CDF, the results are given in Figure 5.5 (c). We see that for $N = 50$ the approximate DBF distribution yields a better approximation of the true distribution than the permutation distributions.

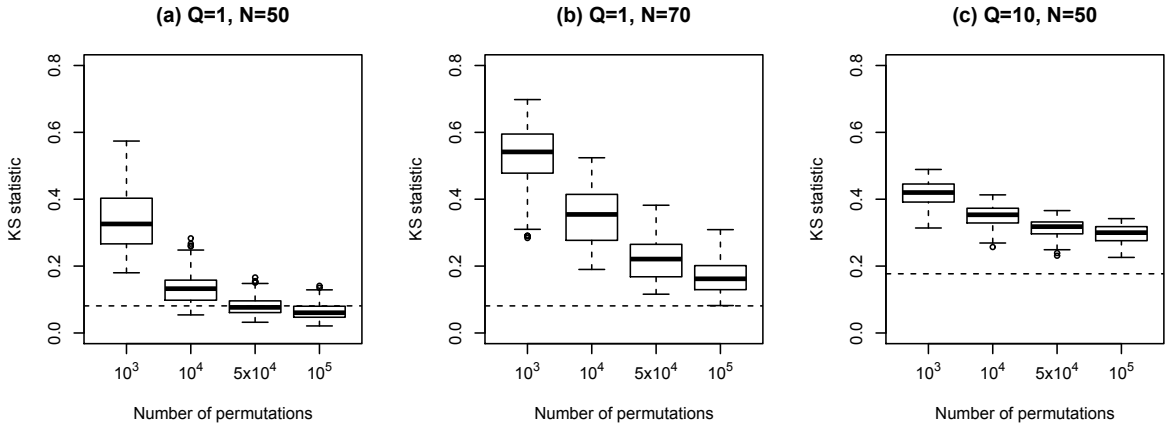

Figure 5.5: (a)-(b) Empirical distributions of the KS statistic quantifying the difference between the DBF permutation CDF, suitably transformed, and the ANOVA F CDF, for each set of Monte Carlo permutations. The dotted line represents the KS statistic comparing the approximate DBF CDF, suitably transformed, and the ANOVA F CDF. (c) Empirical distributions of the KS statistic quantifying the difference between the DBF permutation CDF, suitably transformed, and the Hotelling's T^2 CDF, for each set of Monte Carlo permutations. The dotted line represents the KS statistic comparing the approximate DBF CDF, suitably transformed, and the Hotelling's T^2 CDF.

5.5.2 Power Study with Functional Tests

Here we compare the DBF test against the Mantel test and two tests specifically designed for detecting differences between population curves (this problem is detailed in Section 9.2). These are the EDGE (Storey *et al.*, 2005b) and TN (Zhang *et al.*, 2010) tests, which have been proposed to test a null hypothesis of equality between population curves. This is equivalent to distance-based null hypothesis (2.4) with the

L_2 distance, because equality between curves equates to the area between them being zero. From this power study we aim to show that (i) DBF is competitive with existing methods for testing a null hypothesis of equality between curves, and (ii) for detecting other types of differences, DBF outperforms all methods, including the distance-based Mantel test.

We perform three Monte Carlo simulations in this endeavour, each with $B = 200$ runs. For each run 250 independent datasets of N curves across $G = 2$ groups of equal size are generated with 225 under the null hypothesis and 25 under the alternative hypothesis. For each experiment a different notion of distance is embraced: (i) area-preserving, in which the null curves have zero L_2 distances; (ii) shape-preserving, in which the null curves have zero Visual L_2 distances; (iii) curvature-preserving, in which the null curves have zero Curvature distances.

For a particular notion of distance, we simulate curves similar to those observed in real longitudinal gene expression datasets, such as those of the *M.tuberculosis* dataset described in Section 9.2.4, while respecting the chosen distance measure. We adopt a three-stage procedure to achieve this. First, true group curves $\{\mu_g(t)\}_{g=1}^2$ are defined for $t \in \tau = [0, 48]$ (to mimic the time-range of the *M.tuberculosis* data) using quadratic Bezier curves (Farin, 1992). In the second step, longitudinal observations are sampled from these curves at time-points $\boldsymbol{t} = (t_1, \ldots, t_S)^T$, to yield N S-dimensional longitudinal observation vectors. The third and final step consists of applying functional data analysis (FDA) techniques to model these vectors as curves (Ramsay and Silverman, 2006; Wu and Zhang, 2006), yielding a set of N curves. We describe each of these steps in detail below.

To begin, we use quadratic Bezier curves to generate $\{\mu_g(t)\}_{g=1}^2$. These curves are parameterized by a scalar $z \in [0, 1]$ and three two-dimensional control points; a start point $\boldsymbol{p}_S$, a middle point $\boldsymbol{p}_M$, and an end point $\boldsymbol{p}_E$. The first dimension of each coordinate represents time and the second dimension represents the curve value. The points $\boldsymbol{p}_S$ and $\boldsymbol{p}_E$ define the start and end of the curve being generated, with the middle coordinate $\boldsymbol{p}_M$ influencing the pattern of the curve between these points. The quadratic Bezier curve is defined by

$$\boldsymbol{b}(z; \boldsymbol{p}_S, \boldsymbol{p}_M, \boldsymbol{p}_E) = (1 - z^2)\boldsymbol{p}_S + 2(1 - z)z\boldsymbol{p}_M + z^2\boldsymbol{p}_E,$$

(Farin, 1992). By controlling the parameters we are able to generate realistic curves as typically observed in real gene expression datasets. Figure 5.6 shows a quadratic Bezier curve generated with randomly chosen control points $\boldsymbol{p}_S = (0,0)^T$, $\boldsymbol{p}_M = (10, 3.5)^T$ and $\boldsymbol{p}_E = (48, -0.65)^T$.

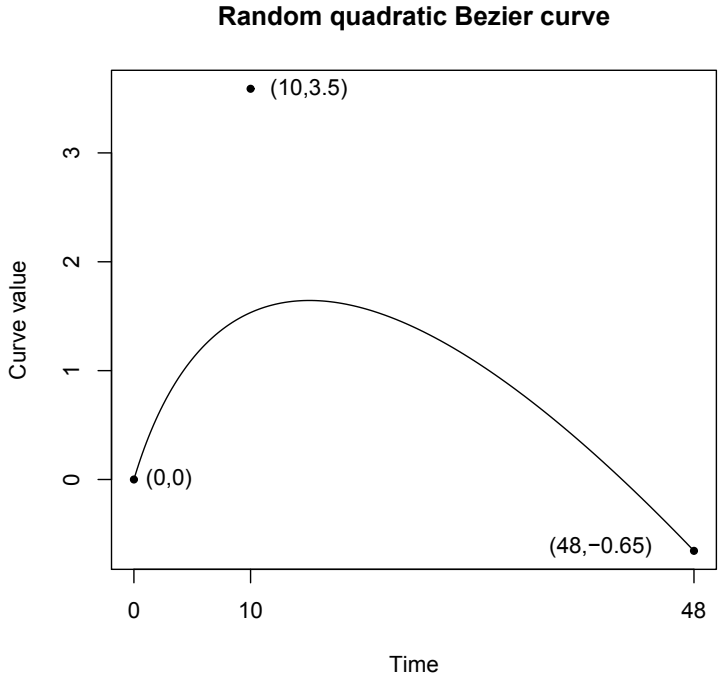

Figure 5.6: A quadratic Bezier curve with control points $\boldsymbol{p}_S = (0,0)^T$, $\boldsymbol{p}_M = (10, 3.5)^T$ and $\boldsymbol{p}_E = (48, -0.65)^T$ represented by the black points.

The Bezier curves representing the two true curves $\{\mu_g(t)\}_{g=1}^2$ for a given Monte Carlo dataset are generated as follows. For dataset $i = 1, \ldots, 250$ we let

$$\mu_g(t) = \boldsymbol{b}_i^{(g)}\left(z; (0,0)^T, \left(A_i^{(g)}, B_i^{(g)}\right)^T, \left(48, C_i^{(g)}\right)^T\right),$$

for $g = 1, 2$. In all cases, the start coordinate is taken to be $\boldsymbol{p}_S = (0,0)^T$ so that all curves have value 0 at time 0, and the first element of the end point $\boldsymbol{p}_E$ being 48 ensures that all the curves end at time 48. The parameters $A_i^{(g)}$, $B_i^{(g)}$ and $C_i^{(g)}$ are randomly chosen constants that determine specific features of the curves. The parameters are controlled in different ways, depending on the distance setting chosen for the experiment:

(i) Area-preserving distance settings: Under the null, the area between curves in the two groups is zero, and under the alternative, there is a large L_2 distance between the group curves. The generation of curve $\mu_1(t)$ under both hypotheses is carried out by assuming the following sampling distributions: $A_i^{(1)} \sim U_d(2, 47)$ (discrete Uniform distribution over $[2, 47]$), $B_i^{(1)} \sim U(-6, 6)$ and $C_i^{(1)} \sim U(-6, 6)$. This generates curves with peaks and troughs occurring at a range of time-points over τ and general up and down directions, representative of the type of patterns exhibited by real data. For H_0 datasets j, $\mu_2(t)$ is defined by the same coordinates as the group 1 curves, that is, $A_j^{(2)} = A_j^{(1)}$, $B_j^{(2)} = B_j^{(1)}$ and $C_j^{(2)} = C_j^{(1)}$, ensuring equality. For H_1 datasets l, we let $A_l^{(2)} \sim U_d(2, 47)$ but with the constraints that $A_l^{(2)} \neq A_l^{(1)}$, $B_l^{(2)} = B_l^{(1)} + U(-3.5, -1.5)$ and $C_l^{(2)} = C_l^{(1)} + U(1.5, 3.5)$. This ensures that under H_1 both curves have different expression values at similar time-points over τ, yielding large L_2 distances.

(ii) Shape-preserving distance settings: The $\mu_1(t)$ curves are generated using the area-preserving distance settings. Under the null hypothesis we simulate $\mu_2(t)$ to have the same shape as $\mu_1(t)$ but with the amplitude altered. For each H_0 dataset j, we let $A_j^{(2)} = A_j^{(1)}$, $B_j^{(2)} = K_j \times B_j^{(1)}$ and $C_j^{(2)} = K_j \times C_j^{(1)}$ where $K_j \sim U(1.4, 1.9)$, so that $\mu_2(t)$ is a scalar shift of $\mu_1(t)$, resulting in different expression values at the same time-points, but with the same overall shape. Under the alternative hypothesis of different scale-invariant shapes, we use the same procedure as in the area-preserving experiment to yield group curves having different shapes over τ, yielding large Visual L_2 distances.

(iii) Curvature-preserving distance settings: We simulate datasets such that the $\mu_1(t)$ curves have relatively large curvatures. This is achieved by generating curves with prominent peaks by letting $B_i^{(1)} \sim U(4, 6)$ and $C_i^{(1)} \sim U(-6, -2)$, and letting $A_i^{(1)} \sim U_d(2, 47)$ as before. Under the null hypothesis of no difference in curvature, the $\mu_2(t)$ curves can be either equal to $\mu_1(t)$, or reflections of $\mu_1(t)$ in the time axis. We thus simulate the $\mu_2(t)$ curves to be inversions of the $\mu_1(t)$ curves in 50% of the datasets under H_0, and let $\mu_2(t)$ equal $\mu_1(t)$ in the remaining 50% of datasets under H_0. For each H_0 dataset j, this is achieved by letting $A_j^{(2)} = A_j^{(1)}$, $B_j^{(2)} = M_j \times B_j^{(1)}$ and $C_j^{(2)} = M_j \times C_j^{(1)}$ where $M_j = 1$ and

$M_j = -1$ with respective probabilities of 0.5. Under the alternative hypothesis of different curvatures, we simulate low-curvature $\mu_2(t)$ curves by letting them be straight lines, that is, we let $A_l^{(2)} = 24$, $B_l^{(2)} = 0.5 \times C_l^{(1)}$ and $C_l^{(2)} = C_l^{(1)}$ for H_1 datasets l.

Having obtained the true curves $\{\mu_g(t)\}_{g=1}^2$ for a given dataset for the chosen distance measure, the second step involves sampling from these at time-points $\boldsymbol{t} = (t_1, \ldots, t_S)^T$ to yield $N/2$ S-dimensional observation vectors for each group. For group $g = 1, 2$, denote these by $\boldsymbol{y}_j^{(g)}$ for $j = 1, \ldots, N/2$ and $g = 1, 2$. These are generated via the model

$$\boldsymbol{y}_j^{(g)} = \mu_g(\boldsymbol{t}) + \boldsymbol{\epsilon}, \tag{5.13}$$

for $j = 1, \ldots, N/2$ where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_S)^T$ with $\epsilon_s \sim N(0, \sigma^2)$ for $s = 1, \ldots, S$ and $\sigma^2 \sim U(0.05, 1)$. The S-dimensional vector $\mu_g(\boldsymbol{t})$ contains the estimated values of the curve $\mu_g(t)$ at the time-points $t_1, \ldots, t_S$, and the elements of $\boldsymbol{\epsilon}$ provide elements of noise. All the vectors of both groups are collected together, and the notation altered so that we denote the N longitudinal observations across the two groups as $\{\boldsymbol{y}_i\}_{i=1}^N$, where $\boldsymbol{y}_i = \boldsymbol{y}_i^{(1)}$ for $i = 1, \ldots, N/2$ and $\boldsymbol{y}_i = \boldsymbol{y}_{i-N/2}^{(2)}$ for $i = N/2 + 1, \ldots, N$.

In the third and final step of the simulation procedure, these longitudinal observation vectors are represented as time-dependent curves. This is achieved by assuming they are noisy realizations of underlying true curves $\{z_i(t)\}_{i=1}^N$ defined for $t \in \tau$, which must be inferred from the data. This is the underlying premise of FDA methodology. In FDA, each curve is represented as a linear combination of K basis functions $\{\phi_k(t)\}_{k=1}^K$, all defined for $t \in \tau$, which are chosen depending on the characteristics of the observed data. Typically, a Fourier basis is preferred for periodic data, while B-splines offer a very flexible basis particularly suited to modeling non-periodic data (Ramsay and Silverman, 2006). The curve $z_i(t)$ can then be written as $z_i(t) = \boldsymbol{\phi}(t)^T \boldsymbol{c}_i$, where $\boldsymbol{\phi} = (\phi_1(t), \ldots, \phi_K(t))^T$ is the vector of basis functions and $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iK})^T$ are the corresponding basis expansion coefficients which are unknown. Methods of finding the optimal coefficients denoted $\{\hat{\boldsymbol{c}}_i\}_{i=1}^N$, leading to estimated curves $\{\hat{z}_i(t) = \boldsymbol{\phi}(t)^T \hat{\boldsymbol{c}}_i\}_{i=1}^N$, include weighted and penalized least-squares methods (Wu and Zhang, 2006) and semi-parametric mixed effect models (Storey *et al.*, 2005b; Berk and Montana, 2009; Aryee *et al.*, 2009; Stegle *et al.*, 2010). Here we use

cubic smoothing spline smoothing, which is a penalized least-squares approach using B-splines, to infer curves $\{\hat{z}_i(t)\}_{i=1}^{N}$ (details are provided in Appendix E). These are then taken as the N curves of our simulated dataset.

We use several combinations of N and S to generate datasets with different group sizes and using a different number of sampling points in creating the observed longitudinal data. We use $N = 6, 18$ and $S = 4, 9$, with respective time-points $\boldsymbol{t} = (0, 4, 8, 48)^T$ and $\boldsymbol{t} = (0, 6, 12, 18, 24, 30, 36, 42, 48)^T$. Figure 5.7 provides example curves simulated under both hypotheses for each distance setting for $N = 18$ and $S = 4$; these settings imitate the real data analyzed in Section 9.2.4.

For each of these distance settings and choice of N and S, the power of each method is computed using $B = 200$ Monte Carlo runs. These are reported in Table 5.2 for false positive rates of 1%, 5% and 10%. For the area-preserving distance settings, DBF is competitive with EDGE and TN in testing a null hypothesis of equality between curves for all N and S settings, as expected. As N and S increase, so does the power of all tests. For the shape-preserving distance settings, where the Visual L_2 distance is considered, DBF outperforms Mantel while TN and EDGE have very little power to detect the shape-related differences between groups. For the curvature-preserving distance settings, we see that again, DBF outperforms Mantel, with TN and EDGE not being able to detect the differences between groups. The better performance of DBF than Mantel is expected because DBF considers both within- and between-group distances rather than just between-group distances. It can be seen that Mantel suffers a large reduction in power for the curvature-preserving distance settings. This is because the Curvature distances are of low magnitude, masking the signal of difference between curves provided by the between-group distances. The DBF statistic detects this signal due its ratio formulation of between- to within-group distances. However, it can also be seen that the power of the DBF test for the dataset with $N = 6$ decreases as S increases. This is because the Curvature distance is very sensitive to perturbations in the curves, which can result from using more sampling points across τ. As N increases, a clearer signal of difference is exhibited and detected by DBF (as is the case with all testing procedures; as N increases so does their power).

Table 5.2: Power (with standard deviation) of the DBF, Mantel, TN and EDGE tests for false positive rates (FPR) of 1%, 5% and 10% in all three distance settings. The brackets $(6,4)$, $(6,9)$, $(18,4)$ and $(18,9)$ indicate the number of curves, N, and number of sampling time-points, S, used to obtain the results in form (N,S). For the area-preserving distance settings, DBF is competitive with EDGE and TN in testing a null hypothesis of equality for all (N,S) settings. EDGE and TN have no power to detect shape-related differences between the groups. In all cases DBF outperforms the Mantel test which only uses between-group distances and therefore has less power.

	Area-preserving			Shape-preserving			Curvature-preserving		
FPR (%)	1	5	10	1	5	10	1	5	10
DBF(6,4)	0.80 (0.09)	0.89 (0.07)	0.92 (0.06)	0.33 (0.11)	0.52 (0.10)	0.61 (0.10)	0.42 (0.11)	0.64 (0.00)	0.76 (0.08)
Mantel	0.73 (0.10)	0.83 (0.08)	0.87 (0.07)	0.20 (0.10)	0.40 (0.11)	0.51 (0.10)	0.05 (0.04)	0.17 (0.09)	0.29 (0.11)
TN	0.81 (0.10)	0.90 (0.07)	0.93 (0.06)	0.00 (0.00)	0.00 (0.01)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
EDGE	0.87 (0.08)	0.92 (0.06)	0.94 (0.05)	0.01 (0.02)	0.06 (0.05)	0.13 (0.05)	0.00 (0.00)	0.01 (0.02)	0.05 (0.03)
DBF(6,9)	0.94 (0.05)	0.97 (0.03)	0.98 (0.03)	0.55 (0.11)	0.71 (0.09)	0.78 (0.09)	0.33 (0.10)	0.51 (0.08)	0.62 (0.08)
Mantel	0.88 (0.07)	0.92 (0.03)	0.94 (0.03)	0.29 (0.11)	0.51 (0.12)	0.63 (0.11)	0.02 (0.03)	0.08 (0.06)	0.15 (0.08)
TN	0.93 (0.06)	0.96 (0.04)	0.97 (0.04)	0.00 (0.01)	0.01 (0.02)	0.02 (0.03)	0.00 (0.00)	0.00 (0.01)	0.02 (0.04)
EDGE	0.95 (0.04)	0.97 (0.03)	0.98 (0.03)	0.00 (0.00)	0.01 (0.02)	0.06 (0.04)	0.00 (0.00)	0.03 (0.03)	0.11 (0.04)
DBF(18,4)	0.93 (0.06)	0.97 (0.04)	0.97 (0.03)	0.64 (0.10)	0.74 (0.08)	0.79 (0.07)	0.86 (0.08)	0.96 (0.04)	0.98 (0.03)
Mantel	0.85 (0.08)	0.90 (0.07)	0.92 (0.06)	0.30 (0.11)	0.50 (0.12)	0.60 (0.10)	0.05 (0.05)	0.18 (0.09)	0.30 (0.11)
TN	0.93 (0.06)	0.97 (0.03)	0.98 (0.03)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
EDGE	0.96 (0.04)	0.98 (0.03)	0.98 (0.03)	0.05 (0.05)	0.15 (0.05)	0.20 (0.04)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
DBF(18,9)	0.99 (0.02)	1.00 (0.01)	1.00 (0.01)	0.80 (0.11)	0.88 (0.09)	0.90 (0.09)	0.58 (0.11)	0.74 (0.08)	0.79 (0.08)
Mantel	0.91 (0.06)	0.93 (0.05)	0.95 (0.05)	0.31 (0.11)	0.54 (0.12)	0.67 (0.11)	0.03 (0.04)	0.09 (0.06)	0.16 (0.07)
TN	0.98 (0.03)	0.99 (0.02)	0.99 (0.02)	0.00 (0.01)	0.01 (0.02)	0.02 (0.03)	0.00 (0.00)	0.00 (0.01)	0.02 (0.03)
EDGE	0.99 (0.02)	1.00 (0.01)	1.00 (0.01)	0.01 (0.00)	0.08 (0.02)	0.15 (0.04)	0.00 (0.00)	0.00 (0.01)	0.02 (0.02)

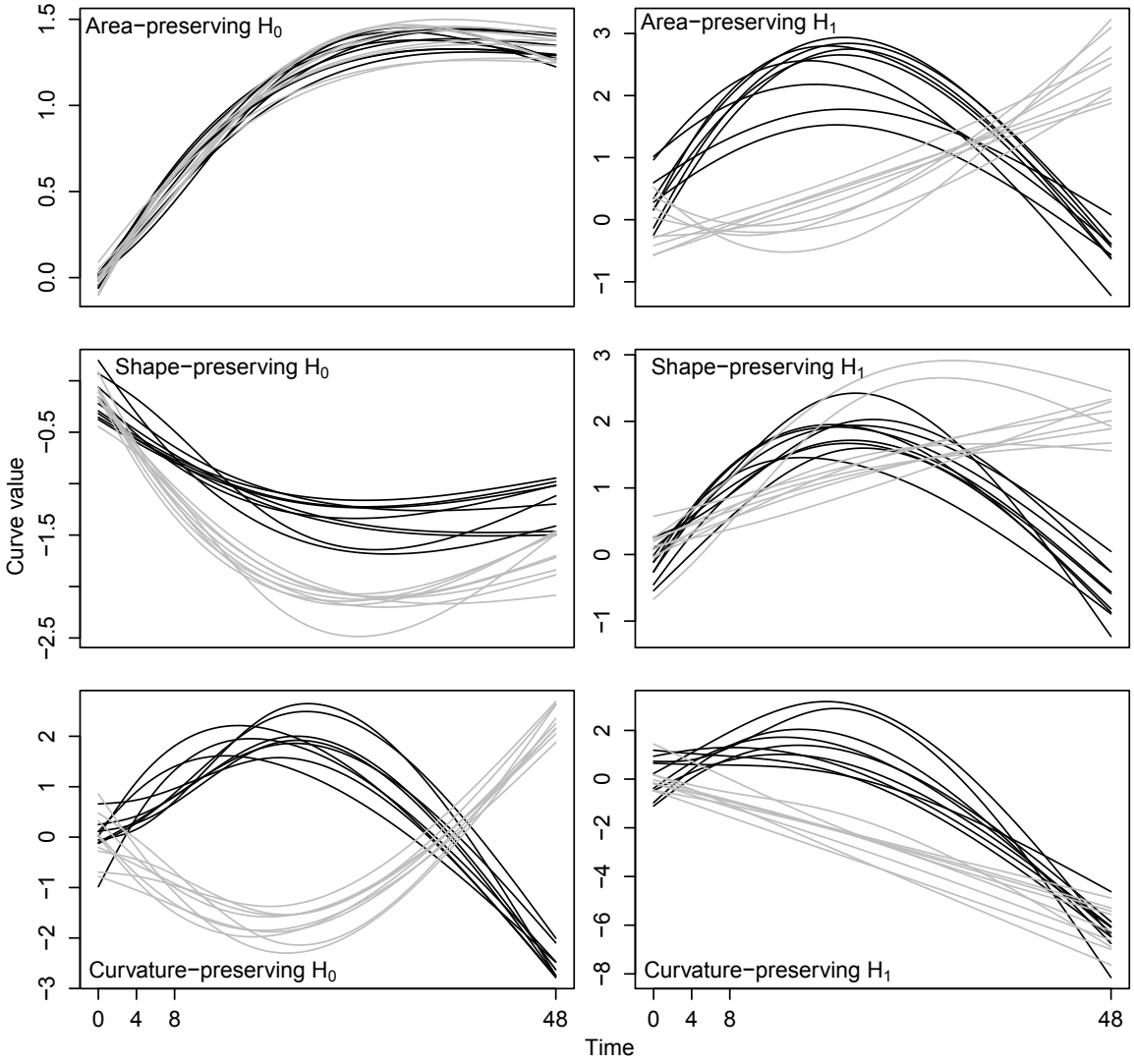

Figure 5.7: Examples of simulated curves for each of the distance settings with $N = 18$ and $S = 4$ under the null and alternative hypotheses. Curves in group 1 are black and those in group 2 are gray, with those on the left simulated under the null, and those on the right simulated under the alternative.

5.5.3 The Approximate Null Distribution of the DBF Statistic

In this section we illustrate how the approximate null distribution of the DBF statistic compares with the Monte Carlo permutation distribution for a number of data types and distances. In our setting, we explore a range of sample sizes and distance measures for simulated datasets, some of which are designed to mimic the real datasets intro-

duced in Section 5.4.1. For vectorial and real-valued data, we consider the Euclidean, Bray-Curtis, Manhattan and Maximum distances (see Appendix B.1 for details). For vectorial and discrete-valued data, we consider the IBS, Simple Matching, Sokal and Sneath, Rogers and Tanimoto I and Hamman I distances (see Appendix B.3 for details). For functional data we consider the L_2, Visual L_2 and Curvature distances (see Appendix B.2), and for graph-structured data we consider the Hamming distance (see Appendix B.4). On selection of data type, distance measure and number of samples N, the datasets are simulated as follows:

(i) Vectorial and real-valued data: 1000-dimensional vectors $\{\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,1000})^T\}_{i=1}^N$ are simulated such that $y_{iq} \sim N(0, 4)$ for $i = 1, \ldots, N$ and $q = 1, \ldots, 1000$. For the Bray-Curtis distance where positive values are required, we take absolute values.

(ii) Vectorial and discrete-valued data: 5-dimensional vectors $\{\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i5})^T\}_{i=1}^N$ are simulated based on the observations of the 153 control subjects from chromosome 1 of the ADNI dataset described in Section 8.2. N control subjects are randomly selected and their minor allele counts at 5 randomly chosen SNPs across the chromosome selected.

(iii) Functional data (curves): N curves $\{y_i(t)\}_{i=1}^N$ are simulated over the range $t \in [0, 48]$ by using the procedure detailed in Section 5.5.2 using quadratic Bezier curves and cubic smoothing splines. N Bezier curves are randomly generated, and N 1000-dimensional vectors are sampled from them at equally spaced points across $[0, 48]$ with standard Gaussian error. These are then smoothed via cubic smoothing splines to yield the N curves. This procedure generates random curves similar to those observed in real longitudinal datasets, such as those shown in Figure 5.1.

(iv) Graph-structured data: N undirected graphs $\{G_i = (V_i, E_i)\}_{i=1}^N$ are generated with vertex sets V_i and edge sets E_i. For the i^{th} graph the number of vertices is denoted $|V_i|$ and the number of edges connecting these vertices is denoted $|E_i|$. We set $\{V_i = V\}_{i=1}^N$ with $|V| = 15$, such that all graphs have a common vertex set comprised of 15 vertices. The edge sets $\{E_i\}_{i=1}^N$ are generated via the Erdős-Rényi model (Erdős and Rényi, 1960) such that $\{|E_i| = 94\}_{i=1}^N$, that is, each

graph is comprised of 94 edges. This procedure generates random graphs with $|V|$ vertices and M edges as follows. The maximum number of undirected edges is $N_V = \binom{|V|}{2}$, and there exist $\binom{N_V}{M}$ unique edge sets with M edges. An edge set is randomly selected from the unique edge sets, and together with the vertex set defines the random graph. Example graphs generated under this model are presented in Figure 5.8.

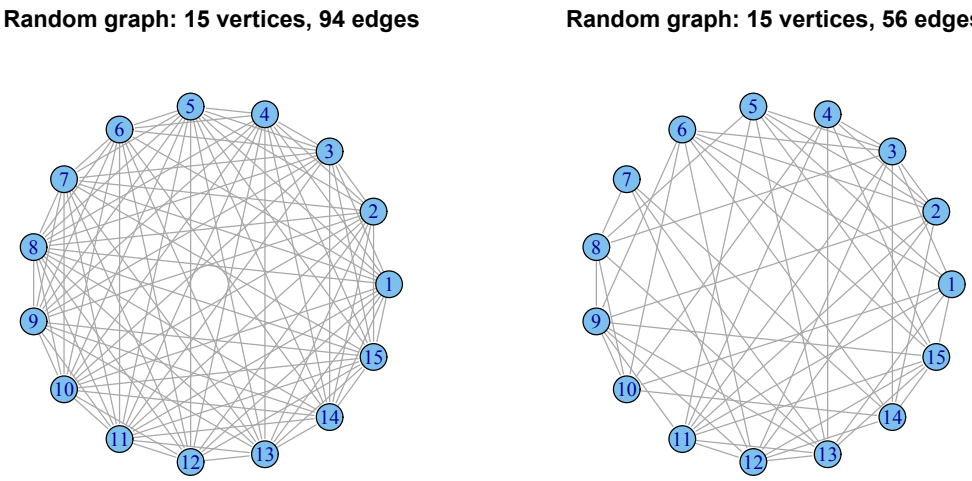

Figure 5.8: Two graphs generated under the Erdős-Rényi model which generates random graphs with a given number of vertices and edges. They are both comprised of 15 vertices, but one connects these vertices with 94 edges while the other uses 56 edges. The greater number of edges results in a graph of greater density.

We compare the theoretical and permutation p-values resulting from applying the DBF test under the null for the different distances applied to each data type. For $N = 10, 30, 100$ and $G = 2$, $B = 200$ Monte Carlo runs are performed, where for each run data is generated under the null, i.e., no group effect. For $N = 10$, all $N!$ permutations are used to compute the permutation p-value, but for $N = 30, 100$, a Monte Carlo set of 10^6 permutations is used. The theoretical and permutation p-values are computed, and the mean and standard deviation of the absolute difference between these for each combination of data type, distance measure and N are reported in Table 5.3. As expected, the absolute difference between the p-values decreases as N increases for each distance measure applied to each data type.

Table 5.3: Mean (and standard deviation) of the absolute differences between theoretical and permutation p-values of the DBF statistic under the null with 200 Monte Carlo runs for vectorial, SNP, functional (curve) and graph distances. For $N = 10$, all 10! permutations are used, and for $N = 30, 100$, 10^6 Monte Carlo permutations are used.

Data type	Distance measure	N = 10		N = 30		N = 100	
Vector, real-valued	Euclidean	0.0159	(0.0139)	0.000706	(0.000554)	0.000317	(0.000266)
	Bray-Curtis	0.0135	(0.0110)	0.000664	(0.000566)	0.000297	(0.000259)
	Manhattan	0.0141	(0.0127)	0.000565	(0.000453)	0.000314	(0.000254)
	Maximum	0.0165	(0.0142)	0.000675	(0.000550)	0.000314	(0.000231)
Vector, discrete-valued	IBS	0.0223	(0.0180)	0.00507	(0.00484)	0.00174	(0.00108)
	Simple Matching	0.0222	(0.0206)	0.00321	(0.00301)	0.00152	(0.000964)
	Sokal and Sneath	0.0217	(0.0187)	0.00551	(0.00509)	0.00393	(0.00220)
	RTI	0.0237	(0.0197)	0.00212	(0.00189)	0.000646	(0.000405)
	Hamman I	0.0211	(0.0201)	0.00324	(0.00317)	0.00158	(0.000988)
Functional	L_2	0.0267	(0.0256)	0.00870	(0.00803)	0.00595	(0.00573)
	Visual L_2	0.0370	(0.0332)	0.0130	(0.0118)	0.00885	(0.00841)
	Curvature	0.0515	(0.0502)	0.0262	(0.0238)	0.00880	(0.00975)
Graph-structured	Hamming	0.0194	(0.0151)	0.000837	(0.000753)	0.000196	(0.000228)

5.5.4 Power Study of Approximation and Permutation Approach

In this section we compare the power of the DBF test to reject the null hypothesis when using the approximate null distribution and Monte Carlo permutations. We perform a Monte Carlo experiment where vectorial and real-valued, and graph-structured data are generated under the alternative hypothesis, and the proportion of these rejected for given significance levels are monitored.

For this study we use 50 Monte Carlo runs, and for both data types generate 50 datasets under the alternative hypothesis. Each dataset is comprised of N observations simulated across two groups of equal size. The N vectorial observations $\{\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iQ})^T\}_{i=1}^N$ are generated with $y_{iq} \sim N(0, 4)$ for $i = 1, \ldots, N/2$ and $q = 1, \ldots, Q$, and $y_{iq} \sim N(4, 4)$ for $i = N/2+1, \ldots, N$ and $q = 1, \ldots, Q$, with $Q = 10$. The N graphs are generated under the Erdős-Rényi model with a common vertex set, V, with $|V| = 15$ but a different number of edges between the groups. In particular, we generate graphs $\{G_i = (V, E_i)\}_{i=1}^N$ such that $|E_i| = 94$ for $i = 1, \ldots, N/2$ and $|E_i| = 70$ for $i = N/2 + 1, \ldots, N$.

The DBF statistic is computed for each dataset using the Euclidean and Maximum distances for the vectorial observations, and the Hamming distance for the graph-structured observations. The p-value of each observed DBF statistic is then estimated via two approaches. The first is via the Pearson type III approximation, for which the proportion of p-values less than or equal to the significance levels of 0.1% and 0.01% are recorded. The mean power across all 50 Monte Carlo runs for each data type, distance measure and significance level are reported in Table 5.4 for $N = 14, 16, 18, 20, 22$. As expected, the power increases with N.

The second approach is via Monte Carlo permutations. A difficulty in performing a reliable power study using permutations is that permutation p-values suffer from sampling error which decreases as the number of permutations increases (see, for instance, Brown *et al.* (2001) and Phipson and Smyth (2010)). This suggests that one should use a very large (fixed) number of permutations for each setting, but this would be computationally infeasible. Instead, we consider using an unconstrained number of Monte Carlo permutations, and running as many permutations as required to achieve a power estimate with a given accuracy. In this way, we can indicate the order of the number of permutations required to achieve the given estimates of power via Monte

Carlo permutations.

We use the algorithm of Gandy and Rubin-Delanchy (2011) to achieve this. This algorithm estimates the power of a Monte Carlo test, and in addition gives a confidence interval around this estimate boasting a guaranteed coverage probability. It requires specification of a value which represents the maximum length of the resulting confidence interval, and the coverage probability, say 95% (i.e., 95% of the confidence intervals generated via this method will contain the true power). It then runs as many permutations as required to yield a power estimate with a confidence interval of length no greater than specified, and with the given coverage probability.

For each setting we run this algorithm and seek power estimates with confidence interval lengths bounded by twice the standard deviation of the corresponding estimate obtained via the approximation. That is, we consider one standard deviation on either side of the power estimate via the approximation as an empirical indication of the precision achieved. We monitor the number of Monte Carlo permutations required to obtain power estimates with such confidence intervals with a coverage probability of 95%. These results are also given in Table 5.4, where the power is stated alongside the confidence interval, and the number of Monte Carlo permutations required is stated on a separate line below the confidence interval.

We highlight two key aspects of these results. Firstly, while the power estimates improve with N, as expected, the number of Monte Carlo permutations varies between $O(10^7)$ to $O(10^{10})$ nonlinearly with N. One might expect that more permutations are required as N increases, but this is not the case with the algorithm used for this power study. The expected number of permutations depends on the length of the confidence interval sought (Gandy and Rubin-Delanchy, 2011), since greater precision is demanded when specifying a smaller length. This can be seen by the results of the Manhattan and Hamming distance settings with significance level 0.1% and $N = 20, 22$. The power estimates are similar, around $0.95 - 0.99$, but more permutations are required when the standard deviation of the estimate with the Pearson type III approximation is smaller (and hence the specified length of the confidence interval is smaller). Furthermore, the authors also show that the region of the confidence interval in the range $[0, 1]$ dictates the expected number of permutations. In particular, the algorithm requires more permutations when the true power is close to 0.5, since the

Table 5.4: Power of the DBF test at significance levels of 0.1% and 0.01% computed using the Pearson type III approximation, denoted Approx., (with standard deviation), and an unconstrained number of Monte carlo permutations, denoted Uncon., (with confidence interval end-points stated in square brackets). The confidence intervals are obtained with a 95% coverage probability, and the average number of required Monte Carlo permutations in millions is stated below the confidence interval for each distance measure, significance level and N.

Distance	Sig level (%)	Method	N = 14		16		18		20		22	
Manhattan	0.1	Approx.	0.578	(0.082)	0.792	(0.054)	0.896	(0.050)	0.958	(0.029)	0.981	(0.017)
		Uncon.	0.527	[0.457, 0.620]	0.783	[0.726, 0.833]	0.875	[0.822, 0.922]	0.952	[0.918, 0.976]	0.978	[0.956, 0.989]
				76.49		48.38		16.23		23.91		50.80
	0.01	Approx.	0.006	(0.011)	0.155	(0.046)	0.457	(0.056)	0.717	(0.067)	0.861	(0.042)
		Uncon.	0.009	[0.002, 0.024]	0.056	[0.020, 0.112]	0.567	[0.507, 0.619]	0.739	[0.662, 0.792]	0.871	[0.828, 0.912]
				433.2		859.5		1043		819.8		544.4
Maximum	0.1	Approx.	0.450	(0.069)	0.637	(0.066)	0.741	(0.072)	0.848	(0.045)	0.911	(0.045)
		Uncon.	0.363	[0.317, 0.363]	0.596	[0.532, 0.663]	0.739	[0.658, 0.799]	0.831	[0.782, 0.870]	0.915	[0.864, 0.953]
				83.45		57.24		23.01		34.48		11.56
	0.01	Approx.	0.070	(0.039)	0.232	(0.054)	0.410	(0.071)	0.589	(0.059)	0.714	(0.065)
		Uncon.	0.004	[0.000, 0.078]	0.030	[0.006, 0.113]	0.419	[0.336, 0.478]	0.602	[0.542, 0.659]	0.733	[0.666, 0.795]
				11.82		388.8		746.4		489.7		470.4
Hamming	0.1	Approx.	0.458	(0.061)	0.683	(0.064)	0.867	(0.060)	0.957	(0.023)	0.994	(0.012)
		Uncon.	0.473	[0.431, 0.552]	0.719	[0.652, 0.778]	0.880	[0.817, 0.934]	0.969	[0.940, 0.987]	0.994	[0.966, 0.999]
				142.6		31.20		16.07		32.41		38.93
	0.01	Approx.	0.099	(0.044)	0.280	(0.064)	0.532	(0.075)	0.745	(0.071)	0.902	(0.042)
		Uncon.	0.004	[0.000, 0.087]	0.000	[0.000, 0.128]	0.592	[0.514, 0.662]	0.815	[0.734, 0.873]	0.921	[0.871, 0.956]
				12.72		31.66		485.7		131.3		141.4

distribution of the p-values under the alternative tends to have greater mass around the threshold significance level (Gandy and Rubin-Delanchy, 2011). This can be observed with our results; for most settings the highest number of permutations is required for power estimates close to 0.5.

Secondly, in almost every setting the power estimates are similar to those of the approximation. In some cases where the power estimated via the approximation is not within the stated confidence interval, it is greater than that estimated by the large number of Monte Carlo permutations (for example, Hamming distance with significance level 0.01% and $N = 14, 16$). Thus, the approximation does not lose power when compared to running an unconstrained number of Monte Carlo permutations.

5.6 Summary

The DBF statistic suitable for testing null hypothesis (2.4) is derived based on an intuitive distance-based variance decomposition. It directly generalizes the Dempster trace criterion, such that the statistics are equal when the centered observations are vector-valued and the Euclidean distance is applied. It has also been shown that the DBF statistic is monotonically related to classical MANOVA statistics when specific Mahalanobis-like distances are applied.

For an observed DBF statistic, inference can be performed with or without Monte Carlo permutations. Without permutations, this requires approximating the discrete sampling distribution of the DBF statistic under the null by a suitably chosen continuous distribution. We showed that the permutation distribution of the DBF statistic depends on the permutation distribution of the between-group variability component of the distance-based variance decomposition. On presenting the skewed characteristics of the between-group variability for real biological datasets, we justified the use of the Pearson type III distribution to model its skewed nature. We then used its monotonic relationship with the DBF statistic to derive an approximate null distribution for the DBF statistic.

Simulation studies were used to present key aspects of the DBF test. For instance, the approximate null distribution of the DBF statistic was shown to approximate the known distributions of the ANOVA F and Hotelling's T^2 statistics. Furthermore, it was shown to approximate these distributions better than by using the number of

permutations typically used in practice, $O(10^5)$. For more general data types and distance measures, the permutation p-values under the null were shown to tend to the p-values arising from the approximate distribution as sample size increases.

Two power studies were additionally performed. In the first, the DBF test was shown to be competitive with the EDGE and TN methods when testing for equality between curves, which is a problem arising in the gene expression microarray time course literature (more details are given in Section 9.2). For detecting other types of differences between curves, the DBF test was shown to maintain power while EDGE and TN have no power. It was also shown that the Mantel test offers less power than the DBF test.

In the second power study, the power of the DBF test to reject the null hypothesis using the permutation and approximation approaches were compared for a range of distance measures and data types. It was shown that even for small sample sizes many millions of permutations would be required to achieve similar power estimates to those obtained via the approximation. Given that such a large number of permutations is required for the relatively small sample sizes considered (when compared to real datasets), these results provide empirical evidence of the computational advantage offered by using the approximation approach. Power of at least the same order can be achieved at much less computation cost by using the approximation. Thus the DBF test applied with the approximation is suitable in situations where many tests are required, such as for case-control GWA studies. Section 8.2 details such an analysis.

Chapter 6

Distance-Based Regression: the Pseudo F Test

In this chapter we derive an approximate null distribution for the pseudo F statistic described in Section 3.2.3 which quantifies the predictive relationship between a predictor matrix $\boldsymbol{X}$ and a distance matrix $\boldsymbol{\Delta}_{\mathcal{Y}}$. It is used to test null hypothesis (3.13), and we derive the null distribution such that the null hypothesis can be tested without permutations. First we show that the permutation distribution of the pseudo F statistic is monotonically related to the permutation distribution of a particular quantity featuring in the statistic, denoted H. Then we derive an approximate distribution of H by using expressions for the mean, variance and skewness which would be obtained by enumerating all permutations. The approximate distribution of the pseudo F statistic is then found based on this. Finally, we illustrate the applicability of the derived distribution for a range of distances and data types using simulated data and a real imaging genetics dataset.

6.1 Permutation Distribution of the Pseudo F Statistic

Recall from Section 3.2.3 that the permutation distribution of the pseudo F statistic under the null hypothesis is given by the set $\{\hat{F}_\pi\}_{\pi\in\Pi}$. For permutation π, $\hat{F}_\pi$ is defined by

$$\hat{F}_\pi = \frac{\hat{H}_\pi}{\mathrm{tr}(\boldsymbol{G}_{\mathcal{Y}}) - \hat{H}_\pi}, \tag{6.1}$$

where $\hat{H}_\pi = \text{tr}(\boldsymbol{H}\boldsymbol{G}_{\mathcal{Y},\pi})$, $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is the hat matrix computed from the $N \times M$ regressor matrix $\boldsymbol{X}$, and $\boldsymbol{G}_{\mathcal{Y},\pi}$ is the permuted centered inner product matrix $\boldsymbol{G}_\mathcal{Y}$ arising from $\boldsymbol{\Delta}_\mathcal{Y}$. Thus, permuted values of F are monotonically related to permuted values of the quantity $H = \text{tr}(\boldsymbol{H}\boldsymbol{G}_\mathcal{Y})$. In order to approximate the null distribution of F, we begin by approximating the null distribution of H.

6.2 The Approximate Null Distribution of H

The quantity $H = \text{tr}(\boldsymbol{H}\boldsymbol{G}_\mathcal{Y})$ is very similar to the quantity $B_\boldsymbol{\Delta} = \text{tr}\left(\boldsymbol{H}_c\boldsymbol{G}_\mathcal{Y}\right)$ which arises in distance-based analysis of variance (Chapter 5). Similarly to $B_\boldsymbol{\Delta}$, H is a weighted sum of squared distances whose distribution under the null is difficult to evaluate. We therefore approximate it by moment matching, and follow the approach adopted in Section 5.4.1.

In particular, we consider a Pearson type III approximation, and justify it as follows. From Sections 3.1.2 and 3.2.3 we know that the pseudo F statistic equals the classical F statistic (ignoring degrees of freedom divisors) when the response observations are scalar-valued, centered and the Euclidean distance measure is applied. In this case H equals the variance explained by the fitted regression model in the univariate linear regression framework, which is known to have the Chi-squared distribution under the null when errors are assumed to be normally distributed (see, for instance, Rencher (2002)). The Pearson type III distribution encompasses the Chi-squared distribution, which is a Gamma distribution, as a special case. Therefore using this distribution to approximate the null distribution of H means the Chi-squared distribution can be recovered in this special case.

For this approximation we require the mean, variance and skewness of the exact permutation distribution of H, which are given by

$$\mu_H = \frac{1}{N!}\sum_{\pi\in\Pi}\hat{H}_\pi, \quad \sigma_H^2 = \frac{1}{N!}\sum_{\pi\in\Pi}\hat{H}_\pi^2 - \mu_H^2 \quad \text{and} \quad \gamma_H = \frac{\frac{1}{N!}\sum_{\pi\in\Pi}\hat{H}_\pi^3 - 3\mu_H\sigma_H^2 - \mu_H^3}{\sigma_H^3},$$

respectively. We wish to use analytic manipulations of these expressions which only require specification of $\boldsymbol{H}$ and $\boldsymbol{G}_\mathcal{Y}$, and do not require performing the $N!$ permutations. The results of Kazi-Aoual *et al.* (1995) which were used for the corresponding

expressions for $B_{\boldsymbol{\Delta}}$ are not valid here, because while $\boldsymbol{H}$ and $\boldsymbol{G}_y$ are both square and symmetric, $\boldsymbol{H}$ is not centered. Centering $\boldsymbol{H}$ yields an asymmetric matrix, so again the results of Kazi-Aoual *et al.* (1995) are not valid.

To proceed we require results for the permutational mean, variance and skewness of H, which we derive in Section 6.2.1. For simplicity of notation in the following exposition, we drop the subscript associated with $\boldsymbol{G}_y$ so that $H = \operatorname{tr}(\boldsymbol{HG})$. H is thus comprised of the $N \times N$ matrices $\boldsymbol{H} = \{h_{ij}\}_{i,j=1}^{N}$ and $\boldsymbol{G} = \{g_{ij}\}_{i,j=1}^{N}$ satisfying the following properties: $\boldsymbol{H}$ is the projection matrix arising from the $N \times M$ regressor matrix $\boldsymbol{X}$ of full rank, i.e., it is symmetric, not centered and $\operatorname{tr}(\boldsymbol{H}) = M$, and $\boldsymbol{G}$ is symmetric and centered.

6.2.1 Permutational Mean, Variance and Skewness of H

The permutational mean, variance and skewness of H are given by

$$\mu_H = E_\Pi(H), \quad \sigma_H^2 = E_\Pi\left(H^2\right) - \mu_H^2, \quad \text{and} \quad \gamma_H = \frac{E_\Pi\left(H^3\right) - 3\mu_H\sigma_H^2 - \mu_H^3}{\sigma_H^3}, \tag{6.2}$$

respectively, where

$$E_\Pi(H) = \frac{1}{N!}\sum_{\pi\in\Pi} H_\pi, \quad E_\Pi\left(H^2\right) = \frac{1}{N!}\sum_{\pi\in\Pi} H_\pi^2, \quad \text{and} \quad E_\Pi\left(H^3\right) = \frac{1}{N!}\sum_{\pi\in\Pi} H_\pi^3, \tag{6.3}$$

are the first three permutational moments of H, with $H_\pi = \operatorname{tr}(\boldsymbol{HG}_\pi)$ and where $E_\Pi(\cdot)$ denotes the permutational expectation over all $N!$ permutations $\pi \in \Pi$.

To obtain explicit expressions for the quantities in (6.2), explicit expressions for the moments given in (6.3) are required. On expanding the moment expressions, we see that this equates to analytically evaluating the multiple summations

$$\begin{aligned}
E_\Pi(H) &= \sum_{i=1}^{N}\sum_{j=1}^{N} E_\Pi(g_{ij})\, h_{ij} \\
E_\Pi\left(H^2\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} E_\Pi(g_{ij}g_{kl})\, h_{ij}h_{kl} \\
E_\Pi\left(H^3\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} E_\Pi(g_{ij}g_{kl}g_{pq})\, h_{ij}h_{kl}h_{pq}.
\end{aligned}$$

This is achieved by decomposing each into weighted sums of summation operators applied to distinct combinations of the indices. We call these 'distinct index patterns', and the corresponding weight indicates the number of variations of index values, through symmetry and label-swapping, which are equivalent to the given distinct index pattern. For instance, $E_{\Pi}(H)$ is decomposed as

$$\begin{aligned} E_{\Pi}(H) &= \sum_{i=1}^{N} E_{\Pi}\left(g_{ii}\right) h_{ii} + \sum_{i \neq j} E_{\Pi}\left(g_{ij}\right) h_{ij} \\ &= E_{\Pi}\left(g_{ii}\right) \sum_{i=1}^{N} h_{ii} + E_{\Pi}\left(g_{ij}\right) \sum_{i \neq j} h_{ij}, \end{aligned}$$

that is, two components dictated by the distinct index patterns $i = j$ and $i \neq j$. We denote these by ii and ij, respectively, using distinct letters to indicate indices which are not equal. Both patterns have corresponding weights of 1 since there is only one way to obtain them. In each case the summation operator is directly applied to the elements of $\boldsymbol{H}$ with the given index pattern, while the expected value of the index pattern applied to the elements of $\boldsymbol{G}$ is a multiplicative constant. The expected values $E_{\Pi}\left(g_{ii}\right)$ and $E_{\Pi}\left(g_{ij}\right)$ represent the expected values of the diagonal and off-diagonal elements of $\boldsymbol{G}$, respectively, over all permutations.

We can express the decomposition of each multiple summation as

$$E_{\Pi}\left(H^{r}\right) = \sum_{i=1}^{N^{(r)}} w_{i}^{(r)} E_{\Pi}\left(p_{i}^{(r)}\left(\boldsymbol{G}\right)\right) \sum\left(p_{i}^{(r)}\left(\boldsymbol{H}\right)\right) \tag{6.4}$$

for $r = 1, 2, 3$, where

$$\begin{aligned}
N^{(r)} &= \text{number of distinct index patterns comprising the } r^{\text{th}} \text{ moment} \\
w_i^{(r)} &= \text{number of variations of index values equivalent to the } i^{\text{th}} \text{ distinct index pattern} \\
p_i^{(r)} &= i^{\text{th}} \text{ distinct index pattern} \\
p_i^{(r)}(\boldsymbol{A}) &= p_i^{(r)} \text{ with respect to elements of } N \times N \text{ matrix } \boldsymbol{A} \\
E_\Pi\left(p_i^{(r)}(\boldsymbol{A})\right) &= \text{expected value of } p_i^{(r)}(\boldsymbol{A}) \text{ over all } N! \text{ permutations in } \Pi \\
\sum\left(p_i^{(r)}(\boldsymbol{A})\right) &= \text{summation operator applied to } p_i^{(r)}(\boldsymbol{A}), \text{ summing over all non-equal indices.}
\end{aligned}$$

With this notation, we have for $r = 1$ that $N^{(1)} = 2$, $w_1^{(1)} = w_2^{(1)} = 1$, $p_1^{(1)} = ii$, $p_2^{(1)} = ij$. The evaluations of the expected values and summation operators applied to the elements of $\boldsymbol{G}$ and $\boldsymbol{H}$ are provided in the first two rows of Table F.1 in Appendix F. We provide the full derivations below to show the general approach used to obtain the required quantities.

For the quantities in terms of $\boldsymbol{H}$, $\left\{\sum\left(p_i^{(1)}(\boldsymbol{H})\right)\right\}_{i=1}^{2}$, we have

$$\begin{aligned}
\sum\left(p_1^{(1)}(\boldsymbol{H})\right) &= \sum_{i=1}^{N} h_{ii} \\
&= \text{tr}(\boldsymbol{H}) \\
&= M, \\
\sum\left(p_2^{(1)}(\boldsymbol{H})\right) &= \sum_{i \neq j} h_{ij} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} h_{ij} - \sum_{i=1}^{N} h_{ii} \\
&= \sum \boldsymbol{H} - M,
\end{aligned}$$

where $\sum \boldsymbol{H} = \sum_{i=1}^{N}\sum_{j=1}^{N} h_{ij}$, as required. For the expected values of the elements of $\boldsymbol{G}$, we note the following. $E_\Pi\left(p_1^{(1)}(\boldsymbol{G})\right) = E_\Pi(g_{ii})$, i.e., the expected value of the diagonal elements of $\boldsymbol{G}$. For all $N!$ permutations, the rows and columns of $\boldsymbol{G}$ are simultaneously permuted so that the diagonal elements remain in the diagonal

positions of $\boldsymbol{G}$. Thus, each diagonal element can only go into one of the N diagonal positions, that is, the i^{th} diagonal position of $\boldsymbol{G}$ takes each value in $\{g_{ii}\}_{i=1}^{N}$ with probability $1/N$. Hence

$$\begin{aligned} E_{\Pi}\left(p_1^{(1)}(\boldsymbol{G})\right) &= E_{\Pi}(g_{ii}) \\ &= \frac{1}{N}\sum_{i=1}^{N} g_{ii} \\ &= \frac{1}{N}\text{tr}(\boldsymbol{G}) \\ &= \frac{(N-1)!}{N!}\text{tr}(\boldsymbol{G}), \end{aligned}$$

as required. A similar argument is used for $E_{\Pi}\left(p_2^{(1)}(\boldsymbol{G})\right) = E_{\Pi}(g_{ij})$, i.e., the expected value of the off-diagonal elements of $\boldsymbol{G}$. There are $N(N-1)$ off-diagonal positions in $\boldsymbol{G}$, and these are filled with only the off-diagonal elements of $\boldsymbol{G}$ for all permutations. Thus the $(i,j)^{\text{th}}$ position of $\boldsymbol{G}$ takes each value in $\{g_{ij}\}_{i\neq j=1}^{N}$ with probability $1/(N(N-1))$, so that

$$\begin{aligned} E_{\Pi}\left(p_2^{(1)}(\boldsymbol{G})\right) &= E_{\Pi}(g_{ij}) \\ &= \frac{1}{N(N-1)}\sum_{i\neq j} g_{ij} \\ &= \frac{1}{N(N-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij} - \sum_{i=1}^{N} g_{ii}\right) \\ &= -\frac{1}{N(N-1)}\text{tr}(\boldsymbol{G}) \\ &= -\frac{(N-2)!}{N!}\text{tr}(\boldsymbol{G}), \end{aligned}$$

since $\boldsymbol{G}$ is centered.

From these expected value derivations we note that for any distinct index pattern with $N_i^{(1)}$ distinct indices,

$$E_{\Pi}\left(p_i^{(1)}(\boldsymbol{G})\right) = \frac{\left(N - N_i^{(1)}\right)!}{N!}\sum\left(p_i^{(1)}(\boldsymbol{G})\right),$$

where $N_1^{(1)} = 1$ and $N_2^{(1)} = 2$. This relationship can be generalized for $r = 2, 3,$

and shows that the expected values are computed by multiplying a constant to the evaluated summation over the distinct index pattern. Thus for any moment r and distinct index pattern $p_i^{(r)}$, the same summation operators are applied to the elements of $\boldsymbol{G}$ and $\boldsymbol{H}$, whereupon they each have their respective simplifications. For instance, for the first moment, the summations over the elements of $\boldsymbol{H}$ are evaluated and the fact that $\mathrm{tr}\,(\boldsymbol{H}) = M$ is used. The corresponding expected value $E_\Pi(\cdot)$ over the elements of $\boldsymbol{G}$ is evaluated by using the same summation operator applied to the elements of $\boldsymbol{G}$, applying the fact that $\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij} = 0$, and multiplying by the corresponding constant.

For the second and third moments, we have $N^{(2)} = 7$ and $N^{(3)} = 23$, and the corresponding quantities are also given Table F.1. Example derivations for $r = 2$ are also provided in Appendix F, in addition to examples of how the weights are derived.

The mean, variance and skewness of H are then accessible by substituting the required permutational expectation quantities given by (6.4) (in conjunction with Table F.1) into (6.2). For instance, the mean is given by

$$\begin{aligned}
\mu_H &= \sum_{i=1}^{N^{(1)}} w_i^{(1)} E_\Pi\left(p_i^{(1)}\left(\boldsymbol{G}\right)\right) \sum\left(p_i^{(1)}\left(\boldsymbol{H}\right)\right) \\
&= \left(1 \times \frac{(N-1)!}{N!}\mathrm{tr}\left(\boldsymbol{G}\right) \times M\right) + \left(1 \times -\frac{(N-2)!}{N!}\mathrm{tr}\left(\boldsymbol{G}\right) \times \left(\sum \boldsymbol{H} - M\right)\right) \\
&= \frac{(NM - \sum \boldsymbol{H})\,\mathrm{tr}\left(\boldsymbol{G}\right)}{N(N-1)}. \qquad (6.5)
\end{aligned}$$

The variance and skewness quantities are not easily simplified, so we do not include them here.

6.2.2 A Pearson Type III Approximation

The Pearson type III distribution is adopted to model the distribution of H given the exact mean, variance and skewness. On standardizing H by subtracting μ_H and dividing by σ_H, the distribution is parameterized by γ_H, as

$$H_s = \frac{H - \mu_H}{\sigma_H} \sim PT_{III}\left(\gamma_H\right).$$

The support of H_s is then given by $[-2/\gamma_H, \infty)$ if $\gamma_H > 0$, $(-\infty, -2/\gamma_H]$ if $\gamma_H < 0$, and $(-\infty, \infty)$ if $\gamma_H = 0$. We denote the CDF of H_s by $\mathcal{F}_{H_s}(h; \gamma_H)$, and the PDF of H_s by $f_{H_s}(h; \gamma_H)$, defined by

$$\frac{(2/\gamma_H)^{4/\gamma_H^2}}{\Gamma\left(4/\gamma_H^2\right)} \left(\frac{2+\gamma_H h}{\gamma_H}\right)^{(4-\gamma_H^2)/\gamma_H^2} \exp\left(-\frac{2(2+\gamma_H h)}{\gamma_H^2}\right)$$

for $\gamma_H > 0$ and $-2/\gamma_H \leq h < \infty$,

$$\frac{(-2/\gamma_H)^{4/\gamma_H^2}}{\Gamma\left(4/\gamma_H^2\right)} \left(\frac{-(2+\gamma_H)}{\gamma_H}\right)^{(4-\gamma_H^2)/\gamma_H^2} \exp\left(-\frac{2(2+\gamma_H h)}{\gamma_H^2}\right)$$

for $\gamma_H < 0$ and $-\infty < h \leq -2/\gamma_H$, and

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{h^2}{2}\right)$$

for $\gamma_H = 0$ and $-\infty < h < \infty$.

6.3 The Approximate Null Distribution of the Pseudo F Statistic

We wish to derive the null distribution of F in terms of the null distribution of H_s via the one-to-one function $h_1 : H_s \mapsto F$ defined by

$$h_1(H_s) = \frac{\mu_H + \sigma_H H_s}{\operatorname{tr}(\boldsymbol{G}) - \mu_H - \sigma_H H_s},$$

with inverse $h_1^{-1} : F \mapsto H_s$ defined by

$$h_1^{-1}(F) = \frac{(\operatorname{tr}(\boldsymbol{G}) - \mu_H) F - \mu_H}{\sigma_H (1+F)},$$

analogously to the approach of Section 5.4.2. The function h_1 is equal to the function h given by (5.6), except that the mean, variance and skewness correspond to the quantity H_s rather than $B^s_{\boldsymbol{\Delta}}$, and $\operatorname{tr}(\boldsymbol{G})$ is used in place of $T_{\boldsymbol{\Delta}}$ (recall they are equal).

To proceed as in Section 5.4.2, we require h_1 to be continuous over the support of H_s. Clearly it is not continuous at $\beta_1 = (\operatorname{tr}(\boldsymbol{G}) - \mu_H)/\sigma_H$, and we require the

discontinuity at $H_s = \beta_1$ to be in the positive plane. This can be shown as follows. We have from (6.5) that

$$\begin{aligned}\beta_1 &= \frac{1}{\sigma_H}\left(\operatorname{tr}(\boldsymbol{G}) - \frac{(NM - \sum \boldsymbol{H})\operatorname{tr}(\boldsymbol{G})}{N(N-1)}\right) \\ &= \frac{\operatorname{tr}(\boldsymbol{G})}{N(N-1)\sigma_H}\left(N(N-1) - NM + \sum \boldsymbol{H}\right),\end{aligned}$$

and since $\operatorname{tr}(\boldsymbol{G})$ and σ_H are positive, it remains to show that

$$N(N-1) - NM + \sum \boldsymbol{H} > 0.$$

To do this, recall assumption $N >> M$ regarding predictor matrix $\boldsymbol{X}$ (Section 3.1.2). From this, $N - 1 > M$, and multiplying on both sides by N yields $N(N-1) > NM$ so that

$$N(N-1) - NM > 0. \tag{6.6}$$

Furthermore, $\boldsymbol{H}$ is positive semi-definite since it has non-negative eigenvalues (see, for instance, Hoaglin and Welsch (1978)), thus $\sum \boldsymbol{H} = \mathbf{1}_N^T \boldsymbol{H} \mathbf{1}_N \geq 0$. Therefore, non-negative $\sum \boldsymbol{H}$ can be added to the left-hand side of (6.6) to yield

$$N(N-1) - NM + \sum \boldsymbol{H} > 0,$$

showing that $\beta_1 > 0$ as required.

It then follows that the arguments presented in Section 5.4.2 can be used in order to derive the null CDF and PDF of F, denoted $\mathcal{F}_F(\cdot; \mu_H, \sigma_H, \gamma_H)$ and $f_F(\cdot; \mu_H, \sigma_H, \gamma_H)$, respectively, in terms of the CDF and PDF of H_s. We provide them in the following proposition for the cases of negative and positive skewness; the case of zero skewness is ignored since in practice the skewness is not equal to zero exactly.

Proposition 2 *The approximate null CDF of the pseudo F statistic, F, can be written in terms of the CDF of the H_s statistic as*

$$\mathcal{F}_F(f; \mu_H, \sigma_H, \gamma_H) = \begin{cases} \mathcal{F}_{H_s}\left(h_1^{-1}(f); \gamma_H\right) - \mathcal{F}_{H_s}(\beta_1; \gamma_H) & -\infty < f < -1 \\ 1 + \mathcal{F}_{H_s}\left(h_1^{-1}(f); \gamma_H\right) - \mathcal{F}_{H_s}(\beta_1; \gamma_H) & \alpha_1 \leq f < \infty \end{cases}$$

for $\gamma_H > 0$*, where*

$$\alpha_1 = \frac{\gamma_H \mu_H - 2\sigma_H}{\gamma_H \left(tr(\boldsymbol{G}) - \mu_H\right) + 2\sigma_H},$$

$$\mathcal{F}_F\left(f; \mu_H, \sigma_H, \gamma_H\right) = \begin{cases} \mathcal{F}_{H_s}\left(h_1^{-1}(f); \gamma_H\right) - \mathcal{F}_{H_s}\left(\beta_1; \gamma_H\right) & -\infty < f \leq \alpha_1 \\ 1 + \mathcal{F}_{H_s}\left(h_1^{-1}(f); \gamma_H\right) - \mathcal{F}_{H_s}\left(\beta_1; \gamma_H\right) & -1 < f < \infty \end{cases}$$

for $\gamma_H < 0$ *and* $\alpha_1 < -1$*, and*

$$\mathcal{F}_F\left(f; \mu_H, \sigma_H, \gamma_H\right) = \begin{cases} 0 & -\infty < f \leq -1 \\ \mathcal{F}_{H_s}\left(h_1^{-1}(f); \gamma_H\right) & -1 < f \leq \alpha_1 \\ 1 & \alpha_1 < f < \infty \end{cases}$$

for $\gamma_H < 0$ *and* $\alpha_1 > -1$*.*

The approximate null PDF of F *can be written in terms of the PDF of* H_s *as*

$$f_F\left(f; \mu_H, \sigma_H, \gamma_H\right) = \frac{tr(\boldsymbol{G})}{\sigma_H (1+f)^2} f_{H_s}(h_1^{-1}(f); \gamma_H),$$

where the range of f *is given by the selected case of CDF.*

The proof that the CDF is a valid CDF is identical to that given for the DBF null distribution.

Having obtained the approximate null CDF of F, the p-value of an observed statistic, $\hat{F}$, can be approximated by $1 - \mathcal{F}_F\left(\hat{F}; \mu_H, \sigma_H, \gamma_H\right)$.

6.4 Simulation Experiments

In this section we illustrate how the approximate null distribution of the pseudo F statistic compares with the Monte Carlo permutation distribution for a number of data types and distances. In addition we demonstrate the applicability of the PDF by applying it to real neuroimaging genetics data.

6.4.1 The Approximate Null Distribution of the Pseudo F Statistic

We explore a range of sample sizes and distance measures for simulated response observations, and consider 6 predictor variables in each case. For vector-valued data we

consider the Euclidean, Person's correlation and Manhattan distances. For functional data we consider the L_2, Visual L_2 and Curvature distances, and for graph-valued data we consider the Hamming, Edit and MCS distances. On selection of data type, distance measure and number of samples N, the response observations are simulated exactly as in (i), (iii) and (iv) of Section 5.5.3. In each case the N observations $\{\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i6})^T\}_{i=1}^N$ comprising the rows of predictor matrix $\boldsymbol{X}$ are simulated such that $x_{im} \sim N(0, 4)$ for $i = 1, \ldots, N$ and $m = 1, \ldots, 6$. Thus there is expected to be no predictive relationship between $\boldsymbol{X}$ and the simulated response observations since they are simulated independently.

We compare the theoretical and permutation p-values resulting from applying the pseudo F test under the null for the different distances applied to each data type. For $N = 30, 60, 100$, $B = 200$ Monte Carlo runs are performed, where for each run the data is generated as described above. For all N, a Monte Carlo set of 10^6 permutations is used. The theoretical and permutation p-values are computed, and the mean and standard deviation of the absolute difference between these for each combination of data type, distance measure and N are reported in Table 6.1. As expected, the absolute difference between the p-values decreases as N increases for each distance measure applied to each data type.

6.4.2 Illustration of the Approximate Null Distribution of the Pseudo F Statistic with Real Data

We use a subset of the neuroimaging genetics data described in Section 8.3.1 to compare the approximate null distribution of the pseudo F statistic with that obtained by Monte Carlo permutations. This subset contains longitudinal MRI images observed on $N = 253$ subjects, which are represented as real-valued vectors, and discrete-valued SNPs genotyped in chromosome 1.

We apply three vectorial distance measures to the imaging data; the Euclidean, Pearson's correlation and Normalized Mutual Information (NMI) distances (see Appendix B.1 for details). For each distance we consider the pseudo F framework using two sets of SNPs as predictor variables. We use $M = 7$ and $M = 50$ contiguous SNPs (located side-by-side on the genome) as predictor variables, and use 10^6 Monte Carlo permutations to generate the null sampling distribution of the pseudo F statistic. The

Table 6.1: Mean (and standard deviation) of the absolute differences between theoretical and permutation p-values of the pseudo F statistic under the null with 200 Monte Carlo runs for vectorial, functional (curve) and graph distances. For all N, 10^6 Monte Carlo permutations are used.

Data type	Distance measure	N = 30		60		100	
Vector-valued	Euclidean	0.000449	(0.000280)	0.000351	(0.000301)	0.000294	(0.000245)
	Pearson's correlation	0.000425	(0.000243)	0.000328	(0.000259)	0.000282	(0.000206)
	Manhattan	0.000400	(0.000269)	0.000303	(0.000267)	0.000271	(0.000190)
Functional	L_2	0.00545	(0.00228)	0.00219	(0.00139)	0.000966	(0.000657)
	Visual L_2	0.00650	(0.00277)	0.00254	(0.00185)	0.000991	(0.000585)
	Curvature	0.00814	(0.00407)	0.00359	(0.00248)	0.00128	(0.000791)
Graph-valued	Hamming	0.000409	(0.000333)	0.000336	(0.000269)	0.000279	(0.000199)
	Edit	0.000367	(0.000299)	0.000306	(0.000262)	0.000243	(0.000170)
	MCS	0.000350	(0.000313)	0.000289	(0.000215)	0.000225	(0.000162)

approximate null distribution is computed via the proposed approach, and the PDF is superimposed on the permutation distributions generated. These are shown in Figure 6.1. Observe that the permutation distributions exhibit varying degrees of skewness, and in each case the approximation appears to capture the exhibited characteristics well.

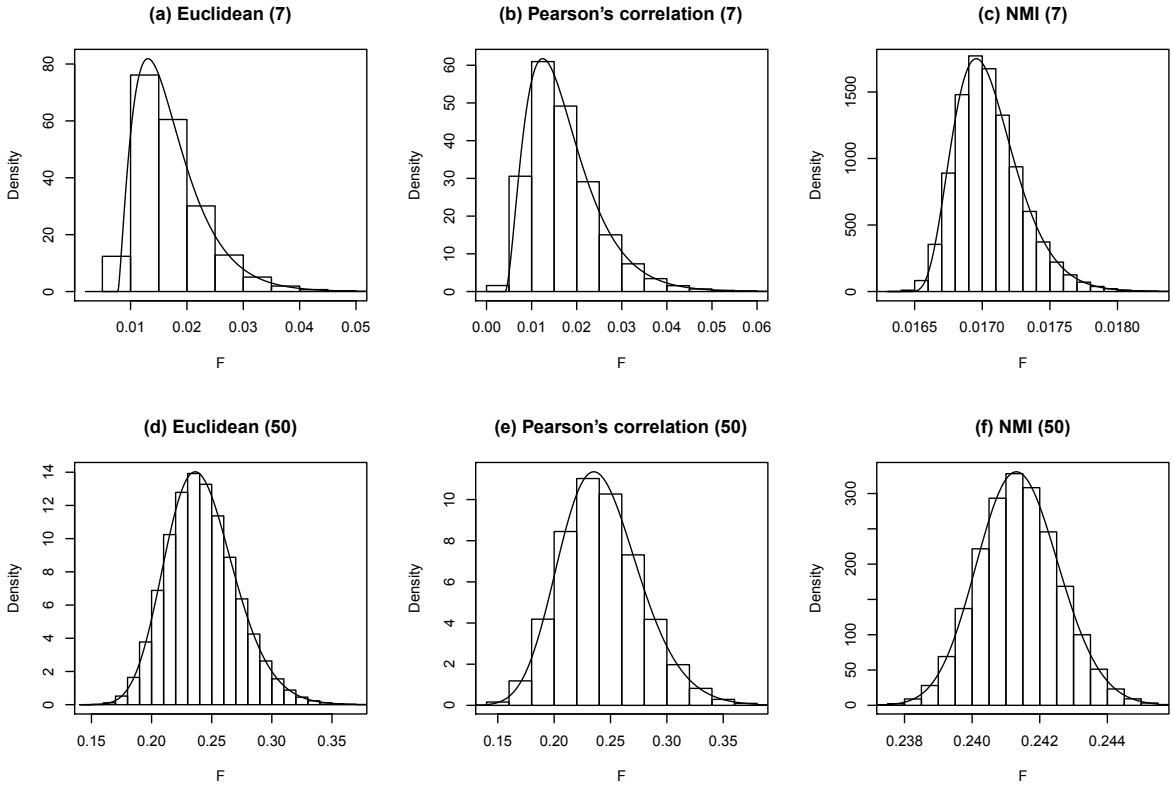

Figure 6.1: Sampling distributions of F obtained using 10^6 Monte Carlo permutations and the proposed approximate PDF. The Euclidean, Pearson's correlation and NMI distances are applied to the real and vector-valued imaging data, and a subset of M discrete-valued SNPs are used as predictor variables. (a)-(c) $M = 7$ SNPs are used. (d)-(f) $M = 50$ SNPs are used.

6.5 Summary

The approximate null distribution of the pseudo F statistic, F, was derived by using its monotonic relationship with the quantity H featuring in the statistic. For this quantity the first three exact permutational moments, i.e., the moments that would be obtained by enumerating all permutations, were derived. A Pearson type III distribution was

then used to approximate the null distribution of H given the first three moments. The approximate null distribution of the pseudo F statistic was then derived in terms of this approximate null distribution.

We demonstrate that the proposed distribution works well for a range of simulated data and using a subset of real imaging genetics data. In Section 8.3 we demonstrate that the pseudo F test with the null distribution approximation can be easily applied to GWA studies, where hundreds of thousands tests are required.

Chapter 7

Distance-Based Association: the GRV Test

In this chapter we propose the generalized RV (GRV) test suitable for testing null hypothesis (4.9). It is derived as a generalization of the RV test of Escoufier (1973) by first showing that the RV coefficient can be written in terms of the Euclidean distances between centered vector-valued observations. The Euclidean distances can then be replaced with any other distances to yield the GRV coefficient. We show that it is related to the dCor coefficient of Székely *et al.* (2007) when observations are vector-valued and a particular distance is used, and hence it can test null hypothesis (4.2) of independence between random vectors. For more general distance measures and data types, simulation studies are presented which demonstrate competitiveness with the distance-based standardized Mantel and PROTEST tests. An approximate distribution is also proposed, allowing inferences to be drawn without permutations for any distances applied. We demonstrate that the distribution works well for a range of simulated and real data.

7.1 A Distance Approach for the RV Coefficient

Consider the case where $\mathcal{X}$ and $\mathcal{Y}$ are real-valued random vectors with corresponding centered observations stored in $\boldsymbol{X}$ and $\boldsymbol{Y}$. Assume also that we are given the corresponding Euclidean distance matrices $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$. From principal coordinate analysis (Section 3.2.2), we know that $\mathcal{X}$ and $\mathcal{Y}$ can be represented by N-dimensional random

vectors $\tilde{\mathcal{X}} = (\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_N)^T$ and $\tilde{\mathcal{Y}} = (\tilde{\mathcal{Y}}_1, \ldots, \tilde{\mathcal{Y}}_N)^T$, respectively, with observations given by the principal coordinates $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$. By considering the derivation of RV with these vectors instead of $\mathcal{X}$ and $\mathcal{Y}$, we can show that the observed RV coefficient can be written directly in terms of Euclidean distances.

$\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ are such that $\text{cov}(\tilde{\mathcal{X}}_i, \tilde{\mathcal{X}}_i) = \lambda_{\mathcal{X},i}$ for $i = 1, \ldots, N$ and $\text{cov}(\tilde{\mathcal{X}}_i, \tilde{\mathcal{X}}_j) = 0$ for $i \neq j$, and similarly for $\tilde{\mathcal{Y}}$, within a multiplicative factor of $1/(N-1)$. The $2N \times 2N$ covariance matrix of $(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}})$ is thus given by

$$\begin{pmatrix} \frac{1}{N-1}\boldsymbol{\Lambda}_{\mathcal{X}} & \boldsymbol{\Sigma}_{\tilde{\mathcal{X}}\tilde{\mathcal{Y}}} \\ \boldsymbol{\Sigma}_{\tilde{\mathcal{Y}}\tilde{\mathcal{X}}} & \frac{1}{N-1}\boldsymbol{\Lambda}_{\mathcal{Y}} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{\tilde{\mathcal{X}}\tilde{\mathcal{Y}}} = \{\text{cov}(\tilde{\mathcal{X}}_i, \tilde{\mathcal{Y}}_j)\}_{i,j=1}^N$, and $\boldsymbol{\Sigma}_{\tilde{\mathcal{Y}}\tilde{\mathcal{X}}} = \boldsymbol{\Sigma}_{\tilde{\mathcal{X}}\tilde{\mathcal{Y}}}^T$. The RV coefficient between $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ is then defined as

$$\text{RV}(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}) = \frac{(N-1)^2 \text{tr}\left(\boldsymbol{\Sigma}_{\tilde{\mathcal{X}}\tilde{\mathcal{Y}}}\boldsymbol{\Sigma}_{\tilde{\mathcal{Y}}\tilde{\mathcal{X}}}\right)}{\sqrt{\text{tr}\left(\boldsymbol{\Lambda}_{\mathcal{X}}^2\right)\text{tr}\left(\boldsymbol{\Lambda}_{\mathcal{Y}}^2\right)}}.$$

The empirical RV coefficient is given by

$$\text{RV}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \frac{\text{tr}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\right)}{\sqrt{\text{tr}\left(\boldsymbol{\Lambda}_{\mathcal{X}}^2\right)\text{tr}\left(\boldsymbol{\Lambda}_{\mathcal{Y}}^2\right)}},$$

since $\boldsymbol{\Sigma}_{\tilde{\mathcal{X}}\tilde{\mathcal{Y}}}$ and $\boldsymbol{\Sigma}_{\tilde{\mathcal{Y}}\tilde{\mathcal{X}}}$ are estimated by $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}/(N-1)$ and $\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}/(N-1)$, respectively.

Recall that the centered inner product matrix arising from $\boldsymbol{\Delta}_{\mathcal{X}}$, $\boldsymbol{G}_{\mathcal{X}}$, satisfies $\boldsymbol{G}_{\mathcal{X}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T = \boldsymbol{U}_{\mathcal{X}}\boldsymbol{\Lambda}_{\mathcal{X}}\boldsymbol{U}_{\mathcal{X}}^T$ so that $\text{tr}\left(\boldsymbol{\Lambda}_{\mathcal{X}}^2\right) = \text{tr}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\right) = ||\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T||^2$, and similarly $\text{tr}\left(\boldsymbol{\Lambda}_{\mathcal{Y}}^2\right) = ||\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T||^2$. Then

$$\text{RV}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \frac{\text{tr}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\right)}{\left|\left|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\right|\right|\left|\left|\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T\right|\right|}. \tag{7.1}$$

By definition, $\boldsymbol{G}_{\mathcal{X}} = -\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}^2\boldsymbol{C}/2 = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T$ for centering matrix $\boldsymbol{C}$, but we also know that $\boldsymbol{G}_{\mathcal{X}} = \boldsymbol{X}\boldsymbol{X}^T$ in this case. It therefore follows that $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T = \boldsymbol{X}\boldsymbol{X}^T$, and similarly for $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$. Therefore $\text{RV}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \text{RV}(\boldsymbol{X}, \boldsymbol{Y})$, where we note that $\text{RV}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = 0$ is equivalent to $\text{RV}(\boldsymbol{X}, \boldsymbol{Y}) = 0$, so that no association is obtained when $\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{0}$.

Crucially, from this equality of $\text{RV}(\boldsymbol{X}, \boldsymbol{Y})$ and $\text{RV}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$, we observe that the RV

coefficient applied to the original centered data matrices is given by

$$\mathrm{RV}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathrm{tr}\left(\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}^2\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{Y}}^2\right)}{||\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}^2\boldsymbol{C}||||\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{Y}}^2\boldsymbol{C}||}$$

(since $\boldsymbol{C}\boldsymbol{C} = \boldsymbol{C}$). The RV coefficient can therefore be expressed solely in terms of the Euclidean distance matrices $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$.

7.2 The Generalized RV Coefficient

Now consider the case where different distance functions $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are applied to the pairwise observations of $\mathcal{X}$ and $\mathcal{Y}$ to yield $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$. Repeating as above, we can use the principal coordinates $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ as in (7.1). Since the terms $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T = \boldsymbol{G}_{\mathcal{X}}$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T = \boldsymbol{G}_{\mathcal{Y}}$ arise in the computation of the coefficient, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ do not need to be explicitly computed. This therefore means that any complex-valued components arising from semi-metric distance functions do not need to be explicitly handled, since $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ will be real-valued. As a computational advantage, this also means that spectral decompositions of $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ are not required. Thus, we define the generalized RV (GRV) coefficient as

$$\mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) = \frac{\mathrm{tr}\left(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}}\right)}{||\boldsymbol{G}_{\mathcal{X}}||\,||\boldsymbol{G}_{\mathcal{Y}}||}, \tag{7.2}$$

noting the implicit assumption $||\boldsymbol{G}_{\mathcal{X}}||\,||\boldsymbol{G}_{\mathcal{Y}}|| > 0$ which is always satisfied in practice; $||\boldsymbol{G}_{\mathcal{X}}|| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} g_{\mathcal{X}}^2(x_i, x_j)} > 0$ since $\boldsymbol{G}_{\mathcal{X}}$ contains real-valued elements which are not all trivially zero, and similarly for $\boldsymbol{G}_{\mathcal{Y}}$.

Similarly to the standardized Mantel coefficient, GRV can be thought of as a correlation coefficient. To see this, vectorize the matrices $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$, denoting the resulting N^2-dimensional vectors $\boldsymbol{g}_{\mathcal{X}}$ and $\boldsymbol{g}_{\mathcal{Y}}$, and consider the quantity $\mathrm{cor}(\boldsymbol{g}_{\mathcal{X}}, \boldsymbol{g}_{\mathcal{Y}})$. To compute this we require the mean and standard deviation of the values in $\boldsymbol{g}_{\mathcal{X}}$ and $\boldsymbol{g}_{\mathcal{Y}}$. The means are 0 since $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ are centered matrices. The

standard deviation of the elements in $\boldsymbol{g}_{\mathcal{X}}$ is given by

$$\begin{aligned}\sqrt{\frac{1}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{g_{\mathcal{X}}(x_i,x_j)}{||\boldsymbol{G}_{\mathcal{X}}||}\right)^2} &= \sqrt{\frac{1}{||\boldsymbol{G}_{\mathcal{X}}||^2\,(N^2-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}g_{\mathcal{X}}^2(x_i,x_j)} \\ &= \sqrt{\frac{||\boldsymbol{G}_{\mathcal{X}}||^2}{||\boldsymbol{G}_{\mathcal{X}}||^2\,(N^2-1)}} \\ &= \sqrt{\frac{1}{N^2-1}},\end{aligned}$$

and similarly for $\boldsymbol{g}_{\mathcal{Y}}$. Thus the correlation of interest is given by

$$\begin{aligned}\operatorname{cor}(\boldsymbol{g}_{\mathcal{X}},\boldsymbol{g}_{\mathcal{Y}}) &= \frac{1}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{g_{\mathcal{X}}(x_i,x_j)/|\boldsymbol{G}_{\mathcal{X}}||}{\sqrt{\frac{1}{N^2-1}}}\right)\left(\frac{g_{\mathcal{Y}}(y_i,y_j)/||\boldsymbol{G}_{\mathcal{Y}}||}{\sqrt{\frac{1}{N^2-1}}}\right) \\ &= \frac{1}{N^2-1}\frac{1}{\left(\frac{1}{N^2-1}\right)}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{g_{\mathcal{X}}(x_i,x_j)g_{\mathcal{Y}}(y_i,y_j)}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||} \\ &= \frac{1}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||}\sum_{i=1}^{N}\sum_{j=1}^{N}g_{\mathcal{X}}(x_i,x_j)g_{\mathcal{Y}}(y_i,y_j) \\ &= \frac{\operatorname{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})}{||\boldsymbol{G}_{\mathcal{X}}||\,||\boldsymbol{G}_{\mathcal{Y}}||} \\ &= \operatorname{GRV}(\boldsymbol{G}_{\mathcal{X}},\boldsymbol{G}_{\mathcal{Y}}).\end{aligned}$$

Consequently, GRV can be directly compared with the standardized Mantel coefficient. We see that the difference in these coefficients lies in the methods of standardization applied to the distances in each case. In standardized Mantel, the upper-triangular distances are subjected to a classical standardization, where their mean is subtracted and they are divided by their standard deviation. In GRV, however, all distance elements are considered, and they are squared, double-centered and normalized by dividing by their Frobenius norm.

7.3 Properties of the GRV Coefficient

The interpretation of GRV as a correlation coefficient indicates that it may range between -1 and 1. We show here that not all values in this range are attainable, and that negative values do not indicate association in the form of a linear correlation of

different sign (as with Pearson's correlation coefficient or standardized Mantel), but less association.

To begin, recall the relationship between the Frobenius distance and the RV coefficient given by (4.8). Analogously to this, the Frobenius distance between the scale invariant configurations $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$ is given by

$$d_F\left(\frac{\boldsymbol{G}_{\mathcal{X}}}{||\boldsymbol{G}_{\mathcal{X}}||}, \frac{\boldsymbol{G}_{\mathcal{Y}}}{||\boldsymbol{G}_{\mathcal{Y}}||}\right) = \sqrt{2\left(1 - \mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}})\right)}, \tag{7.3}$$

(replace $\boldsymbol{X}\boldsymbol{X}^T/||\boldsymbol{X}\boldsymbol{X}^T||$ and $\boldsymbol{Y}\boldsymbol{Y}^T/||\boldsymbol{Y}\boldsymbol{Y}^T||$ in (4.7) with $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$). From this we see that $d_F\left(\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||, \boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||\right) = 0$ suggests that perfect association is achieved when $\mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) = 1$, i.e., when

$$\frac{\boldsymbol{G}_{\mathcal{X}}}{||\boldsymbol{G}_{\mathcal{X}}||} = \frac{\boldsymbol{G}_{\mathcal{Y}}}{||\boldsymbol{G}_{\mathcal{Y}}||}. \tag{7.4}$$

This equality, however, can only be attained if $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ are both positive semi-definite (having non-negative diagonals), or both indefinite (having non-negative and negative values on the diagonals). These occur if $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are both metric, or semi-metric, respectively. When one distance function is metric and the other is semi-metric, perfect association cannot be attained, as the diagonals of $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ cannot be equal. To see this, consider the upper and lower bounds of the GRV coefficient, provided in the following proposition.

Proposition 3 *The bounds of the GRV coefficient for given centered inner product matrices $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ with ordered eigenvalues $\{\lambda_{\mathcal{X},i}\}_{i=1}^N$ and $\{\lambda_{\mathcal{Y},i}\}_{i=1}^N$, respectively, are given by*

$$\frac{\sum_{i=1}^N \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||} \leq GRV(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) \leq \frac{\sum_{i=1}^N \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||},$$

with $\sum_{i=1}^N \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i} \leq ||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||$.

Proof. First, consider the bounds of the quantity $\mathrm{tr}\left(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}}\right)$. We use the result of Lasserre (1995), which gives bounds for the trace of the product of two square Hermitian matrices (square, complex-valued, and equal to their conjugate transpose).

In terms of $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ (which are square and symmetric), the bounds are given by

$$\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1} \leq \text{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}}) \leq \sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i},$$

and since $||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| > 0$, we obtain

$$\frac{\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||} \leq \text{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) \leq \frac{\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||},$$

as required.

To show $\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i} \leq ||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||$, consider the following. From the Cauchy-Schwarz inequality, we have that

$$\begin{aligned} & \left(\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i}\right)^2 \leq \sum_{i=1}^{N} \lambda_{\mathcal{X},i}^2 \sum_{i=1}^{N} \lambda_{\mathcal{Y},i}^2 \\ \Rightarrow -\sqrt{\sum_{i=1}^{N} \lambda_{\mathcal{X},i}^2 \sum_{i=1}^{N} \lambda_{\mathcal{Y},i}^2} \leq & \quad \sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i} \quad \leq \sqrt{\sum_{i=1}^{N} \lambda_{\mathcal{X},i}^2 \sum_{i=1}^{N} \lambda_{\mathcal{Y},i}^2}, \end{aligned}$$

and it is easily shown that the term on the right-hand side equals $||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||$. Thus $\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i} \leq ||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||$, as required.

■

The upper bound on $\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i}$ ensures that the GRV coefficient does not exceed a value of 1. Thus the Frobenius distance given by (7.3) has a minimum value of 0, and the greater the distance value, the less associated $\mathcal{X}$ and $\mathcal{Y}$ are considered to be with respect to $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$.

The numerators of the upper and lower bounds of the GRV coefficient are sums of eigenvalue products, and so may be non-negative or negative depending on the sign of the eigenvalues. This in turn depends on the distance functions satisfying the metric property. We describe each of the three cases in turn: (i) both distance functions are metric, (ii) one distance function is metric and the other is semi-metric, and (iii) both distance functions are semi-metric.

7.3.1 Metric Distance Functions

If $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are metric then $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ are positive semi-definite and the ordered eigenvalues $\{\lambda_{\mathcal{X},i}\}_{i=1}^{N}$ and $\{\lambda_{\mathcal{Y},i}\}_{i=1}^{N}$ are non-negative. The summation in the lower bound contains the terms $\{\lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1}\}_{i=1}^{N}$, which are therefore non-negative, so that

$$\frac{\sum_{i=1}^{N}\lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||} \geq 0 \Rightarrow \mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) \geq 0.$$

The minimum value of 0 is attained when $\mathrm{tr}\left(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}}\right) = 0 \Rightarrow \mathrm{tr}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{T}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^{T}\right) = \mathrm{tr}\left(\tilde{\boldsymbol{Y}}^{T}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{T}\tilde{\boldsymbol{Y}}\right) = 0$. This occurs when the principal coordinates are orthogonal, i.e., $\tilde{\boldsymbol{X}}^{T}\tilde{\boldsymbol{Y}} = \boldsymbol{0}$, and indicates no association. The maximum value GRV can take is 1, since $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ have positive diagonal elements so that equality (7.4) can be attained. In this case the distance matrices are equal up to a positive scaling factor, and there is perfect association. Note also that when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered vector-valued observations with $\boldsymbol{X}^{T}\boldsymbol{Y} = \boldsymbol{0}$, and $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the Euclidean distance functions, then the GRV coefficient yields a value of 0 (as $\mathrm{tr}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{T}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^{T}) = \mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^{T}\boldsymbol{Y}\boldsymbol{Y}^{T}) = \mathrm{tr}(\boldsymbol{X}^{T}\boldsymbol{Y}\boldsymbol{Y}^{T}\boldsymbol{X}) = 0$).

7.3.2 Metric and Semi-Metric Distance Functions

Assume that $d_{\mathcal{X}}$ is metric and $d_{\mathcal{Y}}$ is semi-metric. Then only the ordered eigenvalues $\{\lambda_{\mathcal{X},i}\}_{i=1}^{N}$ are strictly non-negative, so that the summation in the lower bound may be negative. In this case, the GRV coefficient may attain negative values. From (7.3) we see that a negative GRV coefficient leads to a greater Frobenius distance between $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$, so that there is less association. The maximum Frobenius distance is attained for the minimum GRV value of

$$\frac{\sum_{i=1}^{N}\lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||},$$

which therefore indicates no association. The maximum attainable value of the GRV coefficient is not 1 in this case, since equality (7.4) cannot be attained. This is because the diagonals of $\boldsymbol{G}_{\mathcal{X}}$ are positive while the diagonals of $\boldsymbol{G}_{\mathcal{Y}}$ are both positive and

negative. The upper bound of the GRV coefficient is therefore given by

$$\frac{\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},i}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||} < 1,$$

so that perfect association cannot be attained, but larger values indicate greater association.

Recall that metric distance functions satisfy the triangle inequality so that distances with respect to $d_{\mathcal{X}}$ between any three observations satisfy the triangle inequality. The corresponding distances with respect to $d_{\mathcal{Y}}$ will not necessarily share this property as $d_{\mathcal{Y}}$ is semi-metric. Thus the inter-point relationships between all the distances in $\boldsymbol{\Delta}_{\mathcal{X}}$ will not match those in $\boldsymbol{\Delta}_{\mathcal{Y}}$ (for if they did $d_{\mathcal{Y}}$ would satisfy the triangle inequality). Heuristically then, the relationship between distances with respect to $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ cannot be the same, and so it is natural that perfect association cannot be attained.

7.3.3 Semi-Metric Distance Functions

If $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are semi-metric the ordered eigenvalues $\{\lambda_{\mathcal{X},i}\}_{i=1}^{N}$ and $\{\lambda_{\mathcal{Y},i}\}_{i=1}^{N}$ are both non-negative and negative, so that no association is indicated by a GRV value of

$$\frac{\sum_{i=1}^{N} \lambda_{\mathcal{X},i}\lambda_{\mathcal{Y},N-i+1}}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||},$$

which may be negative (this is when the maximum Frobenius distance is attained). For the maximum value of GRV, note that the diagonal elements of $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ are both positive and negative, so that there may exist two such matrices with equal diagonals. Hence there may exist a scenario in which equality (7.4) is attained, although it is not clear under what conditions this will happen. Thus a GRV value of 1 is attained, indicating perfect association.

7.4 Inference

Under null hypothesis (4.9) of no association between the distance matrices, the sampling distribution of the GRV coefficient is unknown. This is because the quantity $T = \text{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})$ in the numerator of the statistic is completely specified by the elements of the distance matrices $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$, which have unknown distributions. In

addition, some of the distance elements are correlated within each distance matrix, so T is a complicated expression involving sums and products of correlated and uncorrelated random variables.

The sampling distribution under the null can be generated by using permutations of one of the centered inner product matrices, $\boldsymbol{G}_{\mathcal{Y}}$, say. For each of N_π permutations $\pi \in \Pi$, the rows and columns of $\boldsymbol{G}_{\mathcal{Y}}$ are simultaneously permuted yielding $\boldsymbol{G}_{\mathcal{Y},\pi}$. This generates the set $\{\hat{\text{GRV}}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y},\pi})\}_{\pi\in\Pi}$ which defines the permutation distribution of $\text{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}})$. We note here that the bounds described for the GRV coefficient remain unchanged with permutations since the ordered eigenvalues of $\boldsymbol{G}_{\mathcal{Y},\pi}$ are equal to those of $\boldsymbol{G}_{\mathcal{Y}}$. Given an observed GRV coefficient, $\hat{\text{GRV}}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}})$, the empirical p-value under the null is found as

$$\frac{\#\left(\hat{\text{GRV}}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y},\pi}) \geq \hat{\text{GRV}}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}})\right)}{N_\pi},$$

as this is a right-tailed test; larger values of the statistic indicate greater association.

In order to approximate the p-value without expensive permutations, we adopt a moment matching approach where the exact null distribution which would be obtained if all $N!$ permutations were used is approximated by a continuous distribution. In particular, we approximate the null distribution by the same continuous distribution which has been used by Josse *et al.* (2008) for the RV coefficient; the Pearson type III distribution. To do this we require the mean, variance and skewness of the exact permutation distribution of T, given by

$$\mu_T = \frac{1}{N!}\sum_{\pi\in\Pi}\hat{T}_\pi, \quad \sigma_T^2 = \frac{1}{N!}\sum_{\pi\in\Pi}\hat{T}_\pi^2 - \mu_T^2, \quad \text{and} \quad \gamma_T = \frac{\frac{1}{N!}\sum_{\pi\in\Pi}\hat{T}_\pi^3 - 3\mu_T\sigma_T^2 - \mu_T^3}{\sigma_T^3},$$

respectively, where $\hat{T}_\pi = \text{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y},\pi})$ and Π contains all $N!$ permutations. Closed form expressions of these quantities are retrievable via the analytical results of Kazi-Aoual *et al.* (1995), requiring $\boldsymbol{G}_{\mathcal{X}}$ and $\boldsymbol{G}_{\mathcal{Y}}$ to be square, symmetric and centered (properties satisfied by definition). These are provided in Appendix C.

On obtaining the mean, variance and skewness of the exact permutation distribution of T, we standardize T by subtracting μ_T and dividing by σ_T. The Pearson type

III distribution can then be parameterized by γ_T:

$$T_s = \frac{T - \mu_T}{\sigma_T} \sim PT_{III}\left(\gamma_T\right).$$

Denote the CDF and PDF of T_s by $\mathcal{F}_{T_s}(t;\gamma_T)$ and $f_{T_s}(t;\gamma_T)$, respectively, for t in the support of random variable T_s. By assumption of this model, the support of T_s is given by $[-2/\gamma_T, \infty)$ if $\gamma_T > 0$, $(-\infty, -2/\gamma_T]$ if $\gamma_T < 0$, and $(-\infty, \infty)$ if $\gamma_T = 0$.

Using the above approximation for the distribution of T_s, we can obtain the approximate distribution of $\mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) = T/||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||$ by a simple transformation. Denote the CDF of the GRV coefficient by $\mathcal{F}_{\mathrm{GRV}}(\cdot;\gamma_T)$ and the PDF by $f_{\mathrm{GRV}}(\cdot;\gamma_T)$, then we have the following.

Proposition 4 *The approximate null CDF and PDF of the GRV coefficient can be written in terms of the CDF and PDF of the T_s statistic as*

$$\mathcal{F}_{GRV}(x;\gamma_T) = \mathcal{F}_{T_s}\left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T};\gamma_T\right),$$

and

$$f_{GRV}(x,\gamma_T) = \left(\frac{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||}{\sigma_T}\right) f_{T_s}\left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T};\gamma_T\right),$$

respectively.

Proof. The CDF of GRV is found as

$$\begin{aligned}
\mathcal{F}_{\mathrm{GRV}}(x;\gamma_T) &= P\left(\mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}) \leq x\right) \\
&= P\left(\frac{\sigma_T T_s + \mu_T}{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||} \leq x\right) \\
&= P\left(T_s \leq \frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T}\right) \\
&= \mathcal{F}_{T_s}\left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T};\gamma_T\right),
\end{aligned}$$

as required. This is a valid CDF since $\mathcal{F}_{T_s}(;\gamma_T)$ is a valid CDF.

The PDF is found by differentiation as

$$\begin{aligned} f_{\text{GRV}}(x, \gamma_T) &= \left| \frac{d}{dx} \left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T} \right) \right| f_{T_s} \left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T}; \gamma_T \right) \\ &= \left| \frac{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||}{\sigma_T} \right| f_{T_s} \left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T}; \gamma_T \right) \\ &= \left(\frac{||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}||}{\sigma_T} \right) f_{T_s} \left(\frac{x||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| - \mu_T}{\sigma_T}; \gamma_T \right), \end{aligned}$$

since $||\boldsymbol{G}_{\mathcal{X}}||||\boldsymbol{G}_{\mathcal{Y}}|| > 0$ and $\sigma_T > 0$.

■

The approximate p-value of an observed GRV coefficient $\hat{\text{GRV}}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}})$ is then found as $1 - \mathcal{F}_{\text{GRV}}(\hat{\text{GRV}}(\boldsymbol{G}_{\mathcal{X}}, \boldsymbol{G}_{\mathcal{Y}}); \gamma_T)$. Empirical results demonstrating how the p-values obtained in this matter compare with those obtained by Monte Carlo permutations are provided in Section 7.6.4.

7.5 Connection with the Distance Correlation Test

Although the GRV coefficient has been generalized from the correlation-based RV coefficient, we can show that for specific distance measures the GRV coefficient is related to the dCor statistic. Consequently, it can be applied to test for independence between real-valued random vectors $\mathcal{X}$ and $\mathcal{Y}$.

This connection arises when $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the square-rooted Euclidean distance measures applied to the pairwise combinations of rows of the centered observations $\boldsymbol{X}$ and $\boldsymbol{Y}$ of $\mathcal{X}$ and $\mathcal{Y}$, respectively. This is due to the following result which relates the squared distance covariance (dCov) to $\text{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})$.

Proposition 5 *Let $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$ be the square-rooted Euclidean distance matrices resulting from applying the square-rooted Euclidean distance functions $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ to the pairwise combinations of rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Then*

$$dCov^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{4tr(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})}{N^2}.$$

Proof. By definition of dCov, we have that

$$\text{dCov}^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\text{tr}\left(\boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{Y}}\right)}{N^2},$$

where $\boldsymbol{D}_{\mathcal{X}} = \boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}^2\boldsymbol{C}$ where $\boldsymbol{\Delta}_{\mathcal{X}}^2$ is the Euclidean distance matrix (since $\boldsymbol{\Delta}_{\mathcal{X}}$ is the square-rooted Euclidean distance matrix) and $\boldsymbol{C}$ is the centering matrix, and similarly for $\boldsymbol{D}_{\mathcal{Y}}$. Since the centered inner product matrix arising from $\boldsymbol{\Delta}_{\mathcal{X}}$ is $\boldsymbol{G}_{\mathcal{X}} = -\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{X}}^2\boldsymbol{C}/2$, it follows that $\boldsymbol{D}_{\mathcal{X}} = -2\boldsymbol{G}_{\mathcal{X}}$ and similarly for $\boldsymbol{D}_{\mathcal{Y}}$. Hence

$$\begin{aligned} \boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{Y}} &= 4\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}} \\ \Rightarrow \operatorname{tr}(\boldsymbol{D}_{\mathcal{X}}\boldsymbol{D}_{\mathcal{Y}}) &= 4\operatorname{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}}) \\ \Rightarrow \mathrm{dCov}^2(\boldsymbol{X},\boldsymbol{Y}) &= \frac{4\operatorname{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})}{N^2}, \end{aligned}$$

as required.

■

As a consequence of Proposition 5, $\mathrm{dVar}^2(\boldsymbol{X}) = 4\mathrm{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{X}})/N^2$, so that

$$\mathrm{dVar}^2(\boldsymbol{X}) = \frac{4}{N^2}\left|\left|\boldsymbol{G}_{\mathcal{X}}\right|\right|^2,$$

and similarly for $\mathrm{dVar}^2(\boldsymbol{Y})$. It then follows that the squared dCor statistic is given by

$$\begin{aligned} \mathrm{dCor}^2(\boldsymbol{X},\boldsymbol{Y}) &= \frac{\mathrm{dCov}^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{\mathrm{dVar}^2(\boldsymbol{X})\mathrm{dVar}^2(\boldsymbol{Y})}} \\ &= \frac{4\mathrm{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})/N^2}{\sqrt{16\left|\left|\boldsymbol{G}_{\mathcal{X}}\right|\right|^2\left|\left|\boldsymbol{G}_{\mathcal{Y}}\right|\right|^2/N^4}} \\ &= \frac{\mathrm{tr}(\boldsymbol{G}_{\mathcal{X}}\boldsymbol{G}_{\mathcal{Y}})}{\sqrt{||\boldsymbol{G}_{\mathcal{X}}||^2||\boldsymbol{G}_{\mathcal{Y}}||^2}} \\ &= \mathrm{GRV}(\boldsymbol{G}_{\mathcal{X}},\boldsymbol{G}_{\mathcal{Y}}). \end{aligned}$$

Hence the GRV coefficient equals the squared dCor statistic when using the square-rooted Euclidean distance measure. In this case it ranges between 0 and 1, taking the value 0 when dCor equals 0, i.e., when $\mathcal{X}$ and $\mathcal{Y}$ are independent. Similarly it takes the value of 1 when dCor takes the value of 1. Thus GRV can be used to test for independence between $\mathcal{X}$ and $\mathcal{Y}$, analogously to dCor. We provide empirical evidence of this in Section 7.6.5.

7.6 Simulation Experiments

In this section we report on a range of simulation experiments designed to demonstrate different aspects of the GRV test. In Sections 7.6.1 and 7.6.2 we demonstrate how the GRV test compares with the standardized Mantel test for specific cases regarding vector-valued observations. In Section 7.6.3 a power study is performed to demonstrate the competitiveness of the GRV test with standardized Mantel and PROTEST for vectorial, curve and genetic distance measures. For semi-metric distance functions, we also apply the RV test to the corresponding principal coordinates which have been corrected to be real-valued. We show that the GRV test achieves greater power than using the RV test with corrected principal coordinates. In Section 7.6.4 the approximate null GRV distribution is compared with the Monte Carlo permutation distribution, and the distribution is applied to real data to demonstrate its applicability. Finally, in Section 7.6.5 the GRV test is shown to be competitive with the dCor test when testing null hypothesis (4.2) of independence between random vectors.

7.6.1 Orthogonal Data Matrices: GRV and Standardized Mantel

Here we consider the setup where $\mathcal{X}$ and $\mathcal{Y}$ are P-dimensional real-valued random vectors satisfying $\{\text{cov}(\mathcal{X}_i, \mathcal{Y}_j) = 0\}_{i,j=1}^{N}$ and hence are not associated. We demonstrate that when applying the Euclidean distance function to pairwise centered observations of each, the standardized Mantel test incorrectly rejects the null hypothesis of no association while the GRV test does not. That is, it suffers from increased type I error rates.

A Monte Carlo experiment is performed with $B = 100$ Monte Carlo runs. For each run $N \times P$ data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are generated under a model of no association, and the GRV and standardized Mantel tests are applied. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are first generated to be orthogonal, and then $\boldsymbol{Y}$ is altered by adding noise to each element. The orthogonality is achieved by generating an $N \times N$ matrix $\boldsymbol{Z}$ with orthogonal columns, and using the first P columns as the columns of $\boldsymbol{X}$, and the subsequent P columns as the columns of $\boldsymbol{Y}$. $\boldsymbol{Z}$ is generated as the principal coordinate matrix arising from applying MDS to the Euclidean distance matrix containing distances between the rows of a random $N \times N$ Wishart matrix. $\boldsymbol{Y}$ is then replaced with $\boldsymbol{Y} + \boldsymbol{E}$, where the elements of $\boldsymbol{E}$ are random

observations from $N(0, \sigma^2)$, and is subsequently centered. The σ parameter controls the amount of noise added to $\boldsymbol{Y}$; as σ increases the signal of orthogonality between $\boldsymbol{X}$ and $\boldsymbol{Y}$ becomes less clear. The Euclidean distance matrices are obtained for each data matrix, and both tests are applied. The theoretical p-value approximation is used to obtain the p-values for the GRV test, and 10^4 Monte Carlo permutations are used for the standardized Mantel test.

The above experiment is performed with $N = 50$, $P = 10$, and for σ ranging from 0 to 5 in steps of 0.1. As the error components are increased, we monitor the average p-value obtained for each test. These are plotted in Figure 7.1. The standardized Mantel test obtains p-values less than the cutoff value of 0.05 for lower σ, indicating significant association. As σ increases so the signal of orthogonality between $\boldsymbol{X}$ and $\boldsymbol{Y}$ becomes less and less clear, small p-values yielding significant associations are still obtained. Eventually they rise above 0.05 to indicate no association, but at a slow rate. Conversely, the GRV test obtains mean p-values of 1 for all σ, indicating no associations. While GRV is expected to perform better than standardized Mantel in the case of orthogonal data, it is unclear how it consistently yields p-values of 1 such that no elements of the signal of orthogonality are masked by the added noise.

7.6.2 Correlated Distance Matrices: GRV and Standardized Mantel

Here we consider the setup where $\mathcal{X}$ and $\mathcal{Y}$ are real-valued random variables whose observations are correlated. We demonstrate that the GRV test has greater power to detect association than the standardized Mantel test.

We perform a Monte Carlo experiment with $B = 100$ runs, and for each run generate $N \times 1$ data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ under a model of association. First $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ is generated to have a clear difference between two subsets of the observations; $x_i \sim N(\mu, 1)$ for $i = 1, \ldots, N/2$ and $x_i \sim N(5\mu, 1)$ for $i = N/2+1, \ldots, N$, with $\mu \sim U(1, 2)$. Then the linear model $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e} = (e_1, \ldots, e_N)^T$ with $e_i \sim (0, \sigma^2)$ for $i = 1, \ldots, N$, is used to generate $\boldsymbol{y}$. We obtain the Euclidean distance matrices corresponding to $\boldsymbol{x}$ and $\boldsymbol{y}$, and apply the GRV and standardized Mantel tests. As before, p-values are obtained via the theoretical approximation for GRV, and with 10^4 Monte Carlo permutations for standardized Mantel.

The above experiment is performed with $N = 50$, and σ ranging from 0 to 11 in

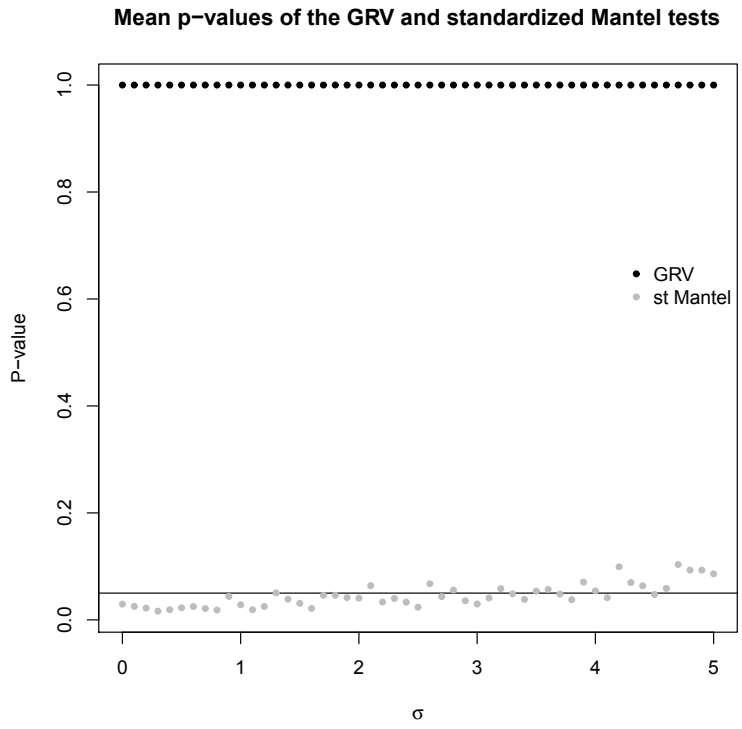

Figure 7.1: Mean p-value of each test after 100 Monte Carlo runs as σ increases from 0 to 5. The black line represents the p-value cutoff of 0.05, below which tests are deemed significant. The GRV test consistently yields large p-values of 1, indicating no association. The standardized Mantel test yields p-values less than 0.05 for smaller σ, indicating rejection of the null hypothesis in favour of the alternative hypothesis of association. As σ increases the p-values rise slowly, eventually indicating no association.

steps of 0.1. Generating the data in this way yields a bimodal distribution for the distances of the observations of $\mathcal{X}$, and the idea is to monitor how well this bimodal signal is detected in the distances of $\mathcal{Y}$ using each method. For example, with little noise ($\sigma = 0.1$) the bimodal characteristics of the distances in $\boldsymbol{\Delta}_{\mathcal{X}}$ are mirrored in $\boldsymbol{\Delta}_{\mathcal{Y}}$. Histograms of the standardized elements of these distance matrices, which are considered by the standardized Mantel statistic are shown in Figures 7.2 (a) and (b). The same bimodal characteristics are exhibited in both sets of distances, yielding a large standardized Mantel statistic, 0.9983, with associated permutation p-value of 0. Similarly, the distances are encoded in $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$ such that the same bimodal characteristics are exhibited, yielding a large GRV value of 0.9987 with associated p-value 7×10^{-13}; Figures 7.2 (c) and (d).

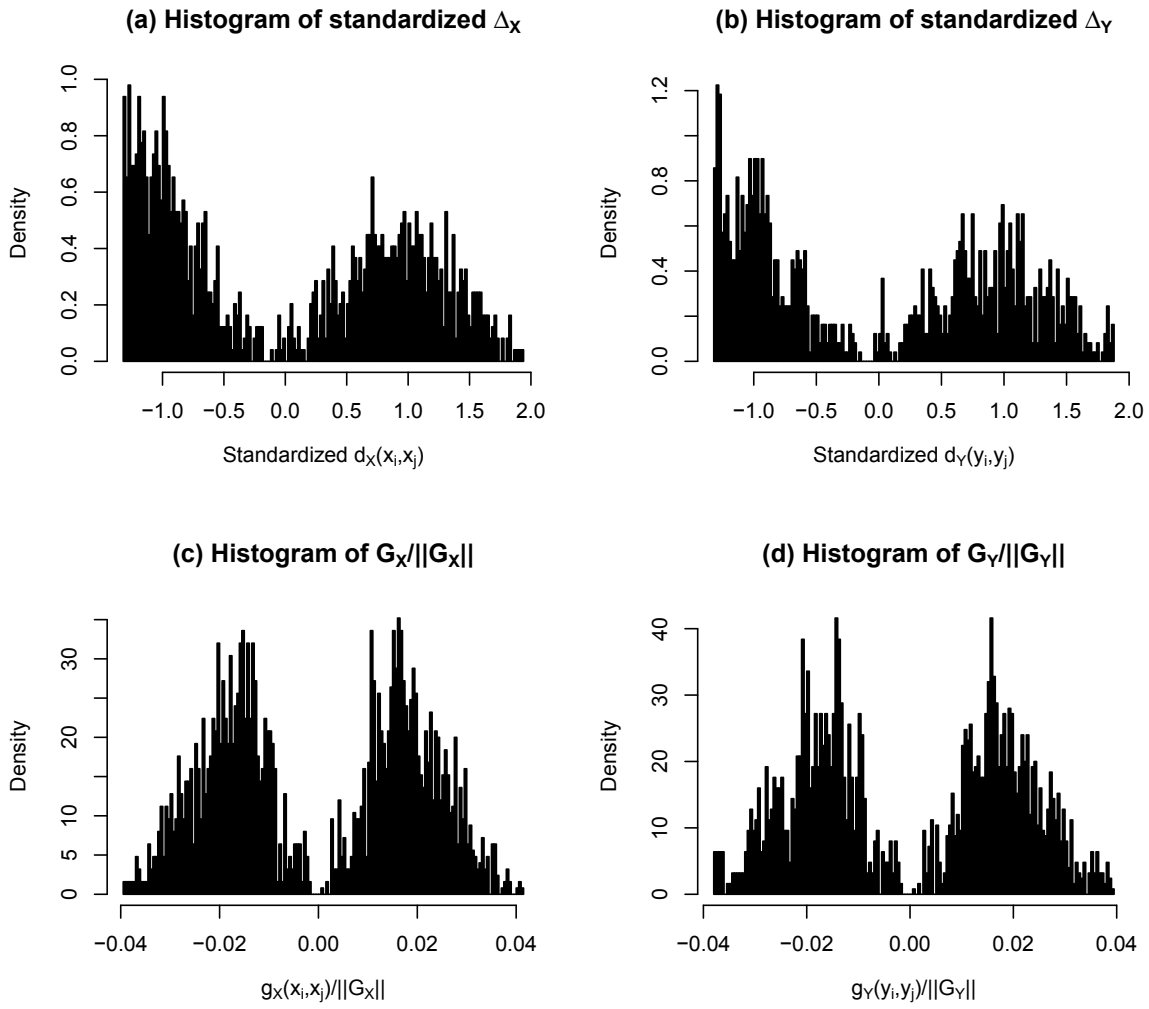

Figure 7.2: (a)-(b) Histogram of the standardized elements of the $N(N-1)/2$ upper triangular values of $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$, respectively. The standardized Mantel statistic depicting the correlation between these values is 0.9983 (p-value of 0 with 10^4 Monte Carlo permutations). (c)-(d) Histogram of the N^2 elements of $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$, respectively. The GRV statistic depicting the correlation between these values is 0.9987 (p-value of 7×10^{-13}).

However, as the noise increases, the bimodal characteristics in $\boldsymbol{\Delta}_{\mathcal{Y}}$ are masked, causing difficulties for both methods in detecting the association. Figure 7.3 shows the mean p-values obtained by applying each method as σ increases. The standardized Mantel test loses power to detect the association at a lower level of noise than the GRV test, as it attains higher p-values for lower σ (reaching the cutoff value of 0.05 for lower σ than GRV). As discussed, both methods are correlation coefficients applied

to different standardizations of the same distances. Thus the standardization adopted by GRV may be more beneficial in preserving any hidden signals than that used by standardized Mantel.

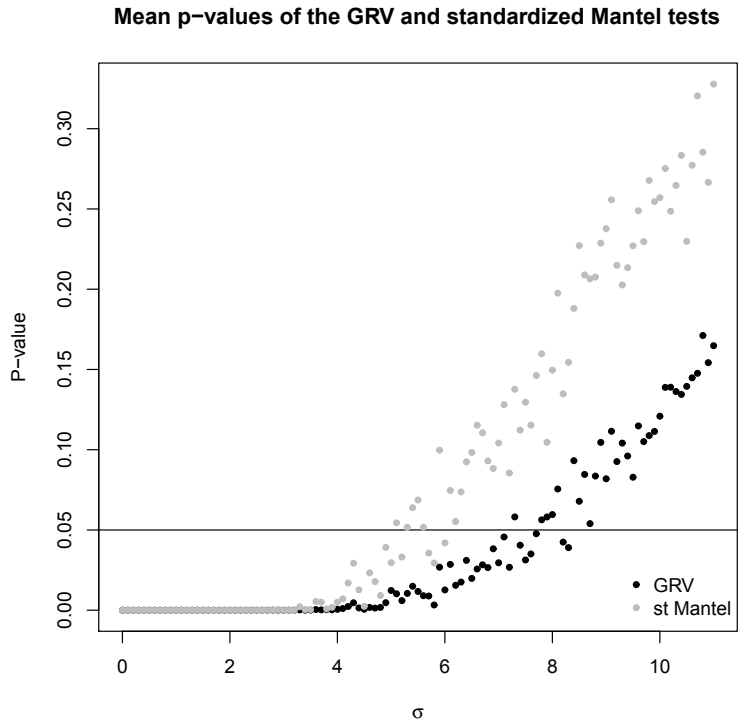

Figure 7.3: Mean p-value of each test after 100 Monte Carlo runs as σ increases from 0 to 11. The black line represents the p-value cutoff of 0.05, below which tests are deemed significant. For lower σ both tests yield small p-values, as expected. As σ increases, however, the standardized Mantel test yields higher p-values than the GRV test. This causes the standardized Mantel to lose power to detect the association for lower σ than GRV.

7.6.3 Power Study for Distance-Based Hypothesis

Two sets of simulations are performed to compare the power of the GRV test against the standardized Mantel and PROTEST approaches for testing (4.9). We demonstrate that GRV is a competitive test of no association between distance matrices. Furthermore, we consider semi-metric distances, and provide evidence that applying corrections to yield real-valued principal coordinates to be subsequently used can lead to a loss of power for a given test. In particular, we demonstrate that the RV test

applied to principal coordinates resulting from a correction is less powerful than the GRV test (which is essentially the RV test applied to the principal coordinates arising from applying no corrections).

The simulations are inspired by the application of detecting associations between paired data observed on SNPs and phenotypic variables. The idea is to simulate allele counts for two SNPs and generate phenotype responses dependent on the SNP observations via an additive model. The phenotypes are vector-valued in one simulation, and functional (curve-valued) in the other. We describe each simulation setting below, but first describe the common procedure for generating realistic SNP data.

The $N \times 2$ SNP data matrix $\boldsymbol{X}$ contains the N simulated minor allele counts at $P = 2$ SNPs, denoted $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T$ for $i = 1, \ldots, N$, with varying minor allele frequencies (MAFs). For SNP p, the MAF $m_p \sim U(0.1, 0.5)$ is generated, and $\{x_{ip}\}_{i=1}^N$ are simulated from the Multinomial distribution with probabilities $(1-m_p)^2$, $2m_p(1-m_p)$, and m_p^2 of observing 0, 1 and 2 minor alleles, respectively. The IBS distance measure is then applied to the simulated SNP data.

The $N \times Q$ phenotype data matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^T$ containing N Q-dimensional vector-valued observations is then generated as follows. Under the null hypothesis of no association, $\boldsymbol{y}_i = \boldsymbol{e}_i$ for $i = 1, \ldots, N$, where $\boldsymbol{e}_i \sim N_Q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_Q)^T$ and $\mu_q \sim U(0, 1)$ for $q = 1, \ldots, Q$, and $\boldsymbol{\Sigma}$ a random $Q \times Q$ Wishart matrix. Under the alternative hypothesis of association, the $N \times 1$ vector $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{1}_2 = (z_1, \ldots, z_N)^T$ containing the row sums of $\boldsymbol{X}$, i.e., the minor allele counts across the two SNPs, is computed. Then $\boldsymbol{y}_i = z_i \boldsymbol{1}_Q + \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is generated as in the case of no association. The Euclidean and Mahalanobis distance measures are applied to the simulated vectors.

The N functional phenotypes $\{y_i(t)\}_{i=1}^N$ with $t \in \tau$ are generated as observations of underlying true phenotype curves dependent on the minor allele counts across the two SNPs. In particular, for $\tau = [0, 5]$, quartic Bezier curves (Farin, 1992) are defined with common start and end points such that their characteristics in between are dependent on the minor allele count. As the minor allele count increases, the curves are simulated to rise faster, as shown in Figure 7.4. Denote these 5 true phenotype curves by $\{f_q(t)\}_{q=0}^4$, with the subscript q denoting the minor allele count.

Under the null hypothesis, all curves $\{y_i(t)\}_{i=1}^N$ are generated as random instances

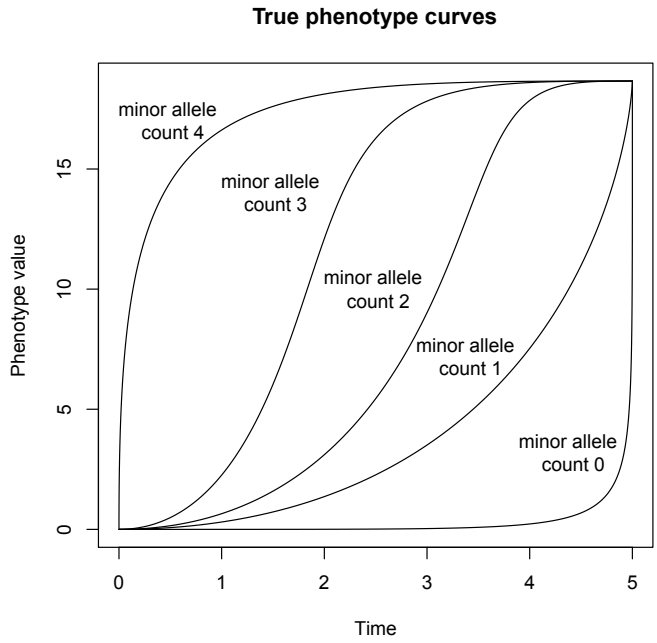

Figure 7.4: Simulated true phenotype curves defined over $\tau = [0, 5]$. All curves have the same start and end points, and they rise faster in between these points as the minor allele count increases.

of the same underlying true curve via a three-step process. In the simulation we take this curve to be $f_0(t)$, but here we describe the procedure of generating curve $y(t)$ as a random instance of curve $f_2(t)$, with an illustration provided in Figure 7.5 ($f_2(t)$ lends itself more nicely to a visual example than $f_0(t)$; it is represented by the gray line). First we simulate a longitudinal vector representing the value of $f_2(t)$ at the time-points $\boldsymbol{t} = (0, 1.25, 2.5, 3.75, 5)^T$, denoted $f_2(\boldsymbol{t})$ (the gray points in Figure 7.5). In the second step noise is added to these points in the form of random Normal observations. These new points, given by $f_2(\boldsymbol{t}) + \boldsymbol{e}$ where $\boldsymbol{e} = (e_1, \ldots, e_5)^T$ with $e_j \sim N(0, \sigma^2)$ for $j = 1, \ldots, 5$ and $\sigma \sim U(1, 4)$, represent noisy observations of the curve $y(t)$ at $\boldsymbol{t}$ (the black points in Figure 7.5). In the third step we use cubic smoothing spline smoothing to infer curve $y(t)$ from its observation points $f_2(\boldsymbol{t}) + \boldsymbol{e}$ (the black line in Figure 7.5). Thus, under the null hypothesis, N curves $\{y_i(t)\}_{i=1}^N$ are simulated in this manner, using $f_0(t)$ as the true underlying phenotype curve. Under the alternative hypothesis, curve $y_i(t)$ is simulated based on true underlying curve $f_{z_i}(t)$, so that it is dependent on the minor allele count across the SNPs. The L_2 and Visual L_2 distance measures are applied to the simulated curves.

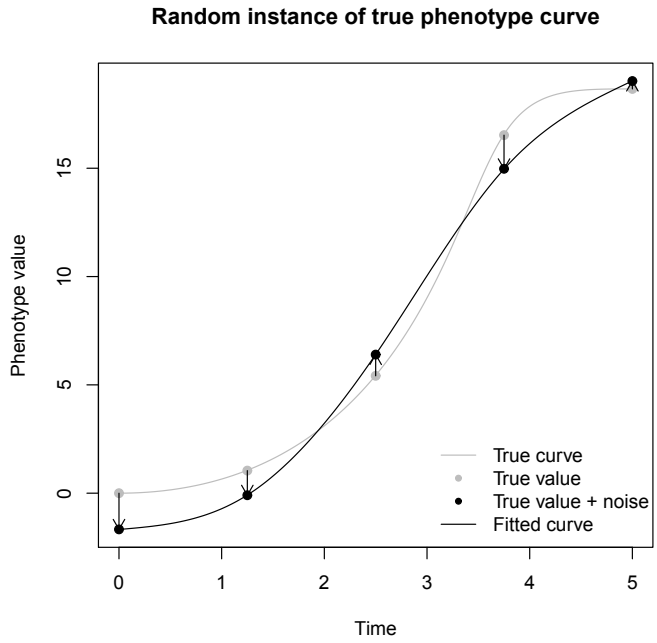

Figure 7.5: Simulation procedure for a random instance of a true phenotype curve. The value of the true curve defined over $\tau = [0, 5]$, represented by the gray line, is obtained at the time-points $\boldsymbol{t} = (0, 1.25, 2.5, 3.75, 5)^T$, represented by the gray points. Noise is added to these points, yielding new observation values, represented by the black points. A curve is fitted to these new points via cubic smoothing spline smoothing, represented by the black line. This resulting curve is the random instance of the true phenotype curve.

The comparison of the methods is then conducted as follows. For $N = 50, 100$, $B = 100$ Monte Carlo runs are performed, and each time 200 datasets are generated for both types of phenotype data ($Q = 10$ for the multivariate phenotypes). 180 of these are generated under the null hypothesis, and 20 under the alternative hypothesis. For each dataset the GRV, standardized Mantel and PROTEST coefficients are computed. A correction is also applied to the semi-metric distance matrices (the IBS and Visual L_2 distance matrices), and the RV coefficient is computed with the resulting real-valued principal coordinates. The power of each test is reported in Table 7.1 for false positive rates of 1%, 5% and 10%.

These results demonstrate that GRV is competitive with standardized Mantel and PROTEST for all false positive rates, and in particular, GRV exhibits more power than standardized Mantel in some cases (such as when using the Mahalanobis and Visual L_2 phenotype distances). Since both can be written as correlation coefficients,

the difference is due to the standardized distance elements used as inputs in each case; a classical standardization is applied by standardized Mantel, and a normalized double-centering is applied by GRV. To demonstrate this we take two datasets generated under the null hypothesis for $N = 50$, one for each type of phenotype. We plot the standardized upper triangular values of $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$, and the values of $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$, showing how the correlation between these values differs. These are given in Figure 7.6; (a) and (b) correspond to the multivariate phenotypes with the Mahalanobis distance, and (c) and (d) correspond to the functional phenotypes with the Visual L_2 distance. To see the difference in correlations more easily, linear regression lines are superimposed. The gradients of these lines are equal to the correlations, so the steeper the gradient, the greater the correlation. It is clear that in (b) and (d) the gradients are steeper than in (a) and (c), respectively, showing that the standardized distance elements in $\boldsymbol{G}_{\mathcal{X}}/||\boldsymbol{G}_{\mathcal{X}}||$ and $\boldsymbol{G}_{\mathcal{Y}}/||\boldsymbol{G}_{\mathcal{Y}}||$ are more correlated than the classically standardized elements of $\boldsymbol{\Delta}_{\mathcal{X}}$ and $\boldsymbol{\Delta}_{\mathcal{Y}}$.

In addition, GRV exhibits more power to reject the null hypothesis than RV, showing that applying a correction for semi-metric distances is not always beneficial. At least for these simulations, the results suggest that using the observed distances, and not altering them in any way, preserves the underlying signals of association.

7.6.4 The Approximate Null Distribution of the GRV Coefficient

We illustrate how the approximate null distribution of the GRV coefficient compares with the Monte Carlo permutation distribution for the distances used in the above simulations. For $N = 30, 60, 100$, $B = 200$ Monte Carlo runs are performed, where for each run data is simulated under the null hypothesis. The GRV coefficient is computed, and the corresponding p-value is computed via the Pearson type III approximation and by a Monte Carlo set of 10^6 permutations. The mean and standard deviation of the absolute difference between these for each N and for each phenotype distance are reported in Table 7.2. It can be seen that as N increases the differences between the p-values decrease.

As a further illustration of how the null distribution compares with the permutation distribution, we consider a subset of the imaging genetics data described in Section 8.3.1. This is the same data used in Section 6.4.2, but in the context of the GRV test

Table 7.1: Power (and standard deviation) of the GRV, standardized Mantel (st. Mantel), PROTEST and RV tests for false positive rates of $1\%, 5\%$ and 10%. GRV is competitive with standardized Mantel and PROTEST, and outperforms RV applied to real-valued principal coordinates arising from corrections of the semi-metric distances.

Phenotype distance	Test	$N = 50$ False positive rate (%) 1		5		10		$N = 100$ False positive rate (%) 1		5		10	
Euclidean	GRV	0.504	(0.151)	0.748	(0.111)	0.826	(0.092)	0.857	(0.093)	0.954	(0.041)	0.969	(0.040)
	st. Mantel	0.234	(0.120)	0.420	(0.113)	0.526	(0.112)	0.449	(0.152)	0.663	(0.112)	0.754	(0.091)
	PROTEST	0.529	(0.163)	0.763	(0.100)	0.845	(0.081)	0.697	(0.142)	0.875	(0.074)	0.925	(0.054)
	RV	0.458	(0.154)	0.674	(0.111)	0.777	(0.094)	0.695	(0.122)	0.843	(0.084)	0.885	(0.082)
Mahalanobis	GRV	0.879	(0.081)	0.950	(0.054)	0.964	(0.041)	0.992	(0.024)	0.997	(0.010)	0.998	(0.011)
	st. Mantel	0.397	(0.163)	0.703	(0.104)	0.796	(0.091)	0.799	(0.120)	0.959	(0.044)	0.978	(0.031)
	PROTEST	0.090	(0.074)	0.222	(0.104)	0.344	(0.121)	0.114	(0.080)	0.282	(0.120)	0.407	(0.124)
	RV	0.121	(0.101)	0.307	(0.124)	0.432	(0.141)	0.107	(0.084)	0.282	(0.124)	0.405	(0.135)
L_2	GRV	0.966	(0.040)	0.983	(0.035)	0.986	(0.031)	0.996	(0.015)	0.997	(0.012)	0.999	(0.014)
	st. Mantel	0.959	(0.054)	0.984	(0.031)	0.988	(0.024)	0.995	(0.025)	0.998	(0.016)	0.999	(0.011)
	PROTEST	0.709	(0.165)	0.884	(0.075)	0.919	(0.063)	0.870	(0.156)	0.972	(0.046)	0.979	(0.034)
	RV	0.936	(0.068)	0.964	(0.043)	0.969	(0.045)	0.967	(0.042)	0.972	(0.042)	0.974	(0.040)
Visual L_2	GRV	0.867	(0.096)	0.938	(0.054)	0.957	(0.047)	0.968	(0.045)	0.986	(0.035)	0.992	(0.023)
	st. Mantel	0.726	(0.135)	0.851	(0.080)	0.871	(0.085)	0.850	(0.106)	0.915	(0.063)	0.933	(0.057)
	PROTEST	0.113	(0.097)	0.220	(0.124)	0.315	(0.120)	0.151	(0.100)	0.282	(0.105)	0.343	(0.104)
	RV	0.465	(0.144)	0.640	(0.124)	0.728	(0.111)	0.301	(0.127)	0.475	(0.118)	0.579	(0.110)

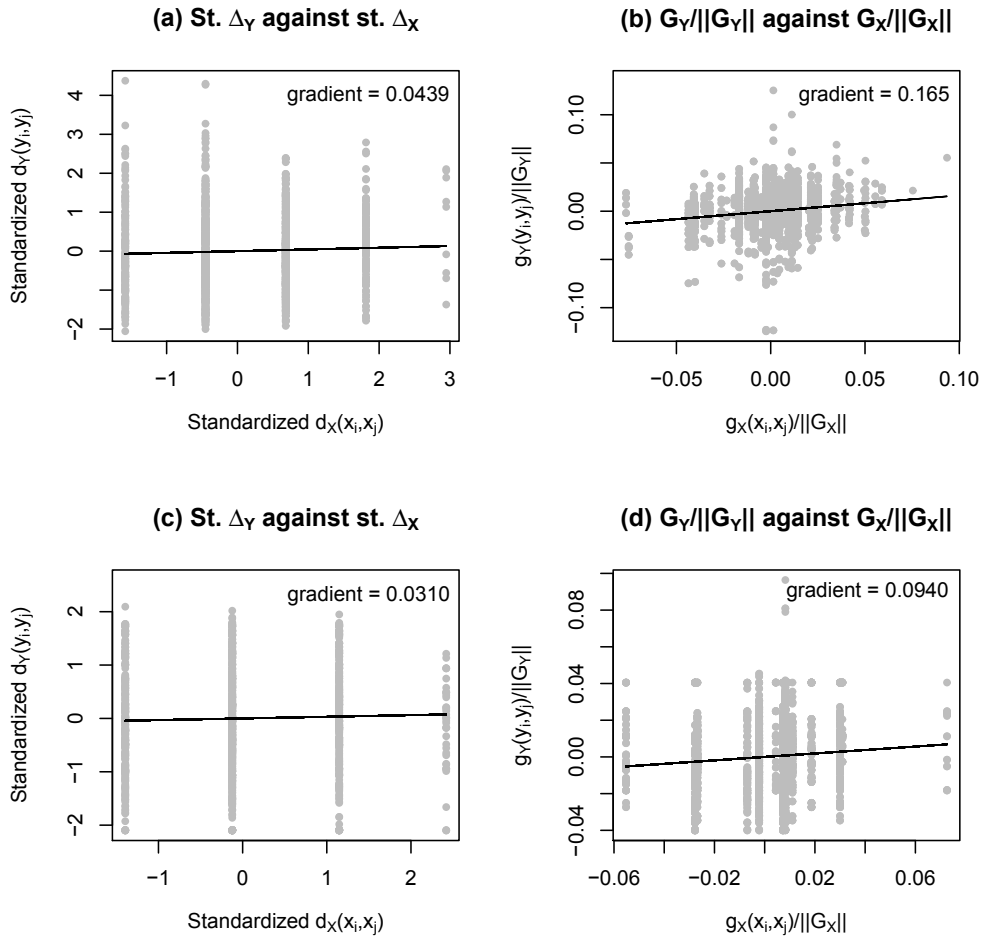

Figure 7.6: Standardized and normalized double-centered elements used in the correlation coefficient representation of the standardized Mantel and GRV coefficients, respectively (gray points). The linear regression lines (black lines) are superimposed indicating the strength of correlation between the values. (a)-(b) Multivariate phenotype with the Mahalanobis distance measure applied. (c)-(d) Functional phenotype with the Visual L_2 distance measure applied. For both phenotypes the standardization used in GRV yields a higher correlation than in standardized Mantel.

we describe it as follows. For the $N = 253$ subjects, $\boldsymbol{Y}$ is the data matrix containing the vector-valued imaging data, and $\boldsymbol{X}$ is the data matrix containing observations of P discrete-valued SNPs in chromosome 1. For the imaging data we apply the NMI distance, and for the SNP data we apply the IBS, Sokal and Sneath, and Rogers and Tanimoto I distances. We consider three sets of SNP data; N observations of $P = 3$, $P = 5$ and $P = 7$ contiguous SNPs. For each combination of genetic distance measure with the imaging distance and P, we obtain the approximate null distribution of the

Table 7.2: Mean (and standard deviation) of the absolute differences between theoretical and permutation p-values of the GRV coefficient under the null hypothesis with 200 Monte Carlo runs. The Euclidean and Mahalanobis distances are used for the multivariate phenotypes, and the L_2 and Visual L_2 distances are used for the functional phenotypes. 10^6 Monte Carlo permutations are used for the permutation p-values.

Phenotype distance	N = 10		30		100	
Euclidean	0.00373	(0.00351)	0.00311	(0.00239)	0.00285	(0.000219)
Mahalanobis	0.00207	(0.00139)	0.00100	(0.000581)	0.000920	(0.000529)
L_2	0.00671	(0.00603)	0.00570	(0.00432)	0.00467	(0.00309)
Visual L_2	0.00758	(0.00813)	0.00619	(0.00520)	0.00512	(0.00525)

GRV statistic and the permutation distribution using 10^6 Monte Carlo permutations. These are given in Figure 7.7, where we see that the approximate distribution provides a good fit for the often skewed permutation distribution.

7.6.5 Power Study for Independence Hypothesis

We compare the power of the dCor test and GRV test (with square-rooted Euclidean distances) to detect dependence between $\mathcal{X}$ and $\mathcal{Y}$, i.e., to test (4.2). We consider a range of univariate setups where Pearson's correlation test is also included for comparison, and a multivariate setup where RV and PROTEST are included.

Univariate Power Study

For $P = Q = 1$, we consider four examples of paired univariate distributions provided by Newton (2009); Figure 7.8 shows these four setups for $N = 500$. Each one is characterized by the general shape exhibited by the samples; w, parabola, hyperbola and independent clouds. The w, parabola and hyperbola setups are found via nonlinear relationships between $\mathcal{X}$ and $\mathcal{Y}$, and hence constitute datasets generated under the alternative hypothesis of dependence. For the independent clouds setup there is no dependence between $\mathcal{X}$ and $\mathcal{Y}$, so this constitutes a dataset generated under the null hypothesis of independence.

We perform a Monte Carlo experiment with $B = 1000$ runs, where for each run

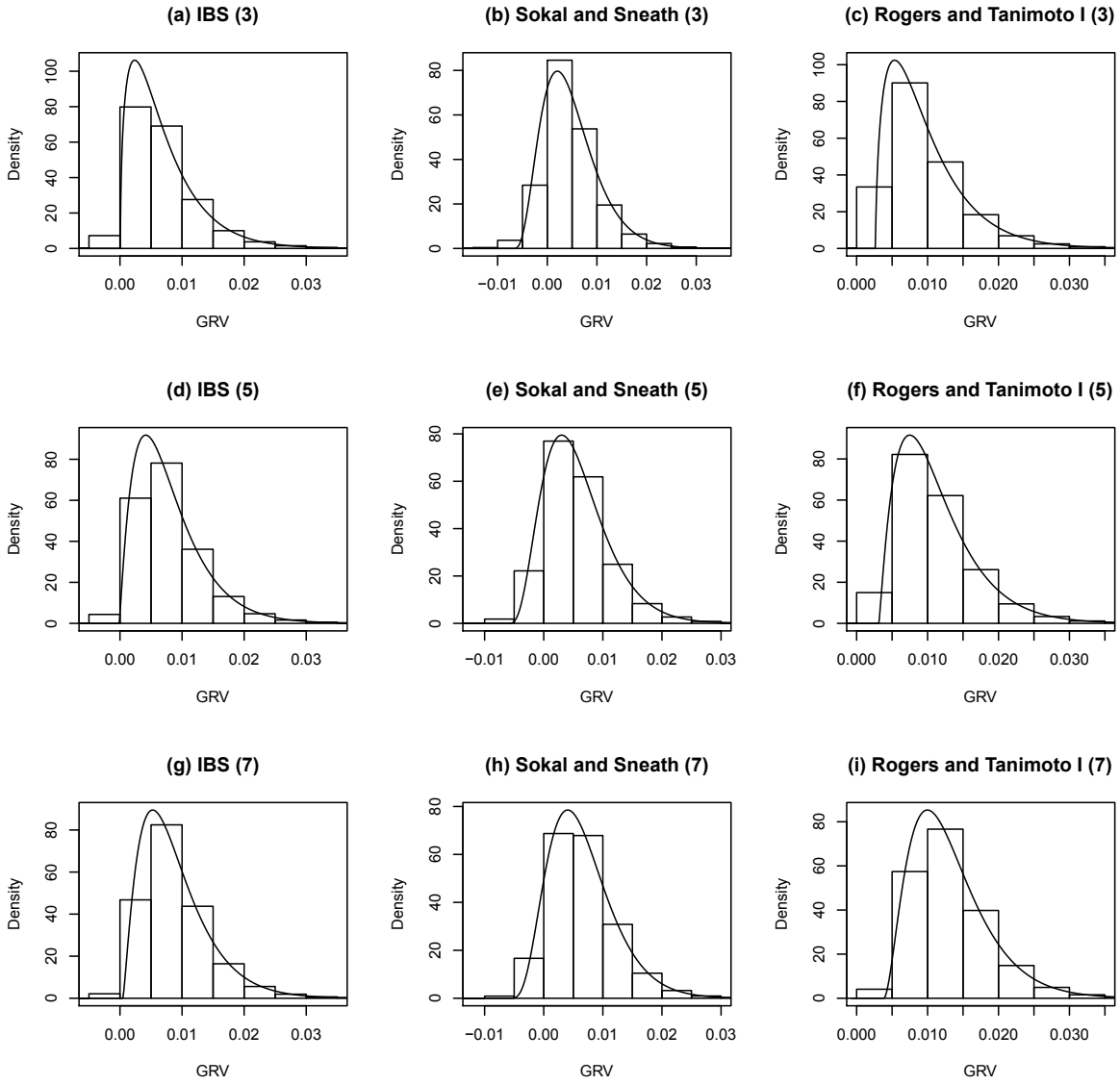

Figure 7.7: Sampling distributions of the GRV statistic obtained using 10^6 Monte Carlo permutations and the proposed approximate PDF. The NMI distance is applied to the real and vector-valued imaging data, and the IBS, Sokal and Sneath, and Rogers and Tanimoto I distances are applied to the observations of P discrete-valued SNPs. (a)-(c) $P = 3$ SNPs are used. (d)-(f) $P = 5$ SNPs are used. (g)-(i) $P = 7$ SNPs are used.

$N = \{25, 30, 35, \ldots, 95, 100\}$ samples are generated for each paired univariate setup. For each N and paired univariate setup, the dCor test is applied with 10^4 Monte Carlo permutations, the GRV test (with square-rooted Euclidean distances) is applied with the theoretical p-value approximation and Pearson's correlation test (denoted ρ) is applied with the theoretical p-value.

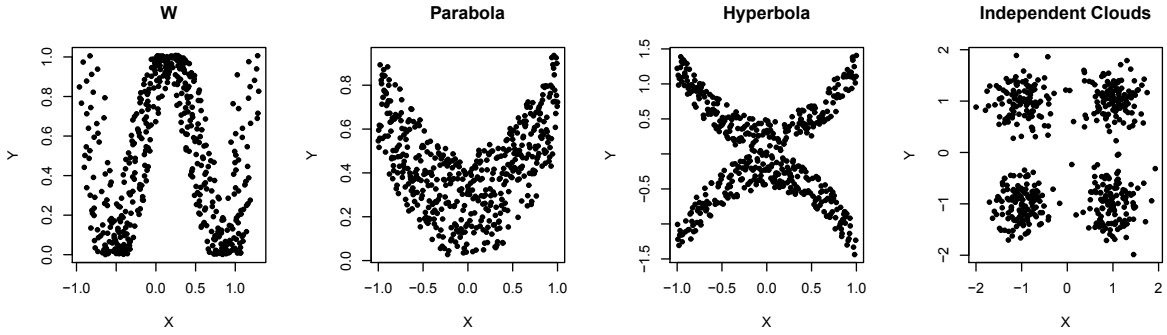

Figure 7.8: Scatter plot of 500 samples from each joint univariate distributional setup. The w, parabola and hyperbola setups are all obtained from a nonlinear relationship between $\mathcal{X}$ and $\mathcal{Y}$, whereas the independent clouds are obtained from an independent relationship between $\mathcal{X}$ and $\mathcal{Y}$.

The power to reject the null hypothesis at the 5% significance level is monitored for each method; Figure 7.9 shows the power as a function of sample size. For the w, parabola and hyperbola setups it is clear that dCor and GRV are competitive, with power increasing to 1 with N at almost identical rates. Note the much better performance than Pearson's correlation coefficient, as expected because Pearson's correlation does not detect the nonlinear dependence. For the independent clouds setup, all tests have a power of around 0.05, which is expected.

Multivariate Power Study

For $P = Q = 5$, we consider an example taken from Székely *et al.* (2007). For $N = \{25, 26, \ldots, 49, 50, 55, 60, \ldots, 95, 100\}$, and for each of $B = 1000$ Monte Carlo runs, we generate observations of $\mathcal{X}$ by $\boldsymbol{x}_i \sim N_5(\mathbf{0}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, N$, where $\boldsymbol{\Sigma}$ is a random 5×5 Wishart matrix. Observations of $\mathcal{Y}$ are generated as $\boldsymbol{y}_i = (\log(x_{i1}), \ldots, \log(x_{i5}))^T$ for $i = 1, \ldots, N$. We denote the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by $\boldsymbol{Y} = \log(\boldsymbol{X}^2)$. For each N the dCor and PROTEST tests are applied with 10^4 Monte Carlo permutations, and the RV and GRV (with square-rooted Euclidean distances) tests are applied with their respective theoretical p-value approximations.

We monitor the power of each test to reject the null hypothesis of independence at the 5% significance level; Figure 7.10 shows the power as a function of sample size. Here dCor and GRV behave identically, and outperform PROTEST and RV. Thus

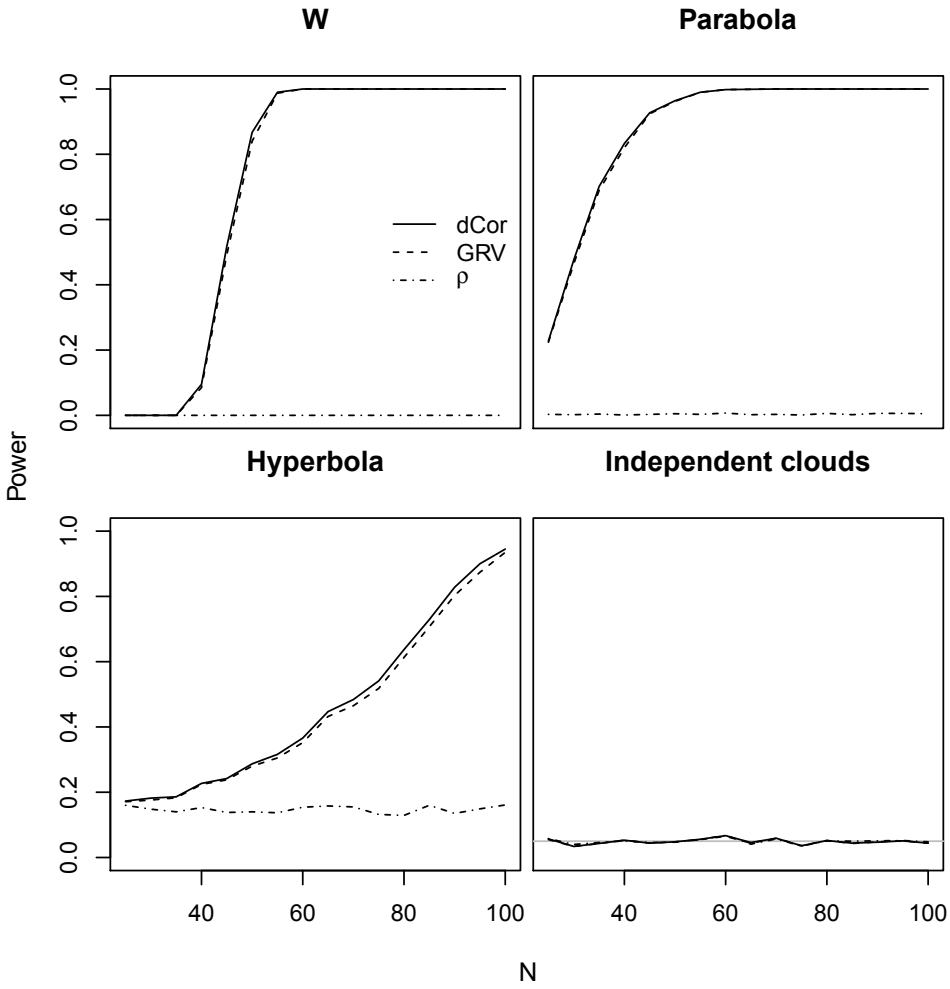

Figure 7.9: Power versus N for 1000 Monte Carlo runs for each paired univariate setup at the 5% significance level. dCor and GRV behave almost identically for increasing N, outperforming Pearson's correlation test which exhibits poor performance. For the independent clouds setup the gray line represents the power level of 5%, expected for all tests.

GRV is competitive with dCor, and it is interesting to note that its behaviour would have been identical to that of RV if it had been applied with Euclidean distances rather than square-rooted Euclidean distances.

7.7 Summary

The GRV test has been generalized from the RV test and been shown to be a versatile distance-based testing procedure for null hypothesis (4.9). When $\mathcal{X}$ and $\mathcal{Y}$

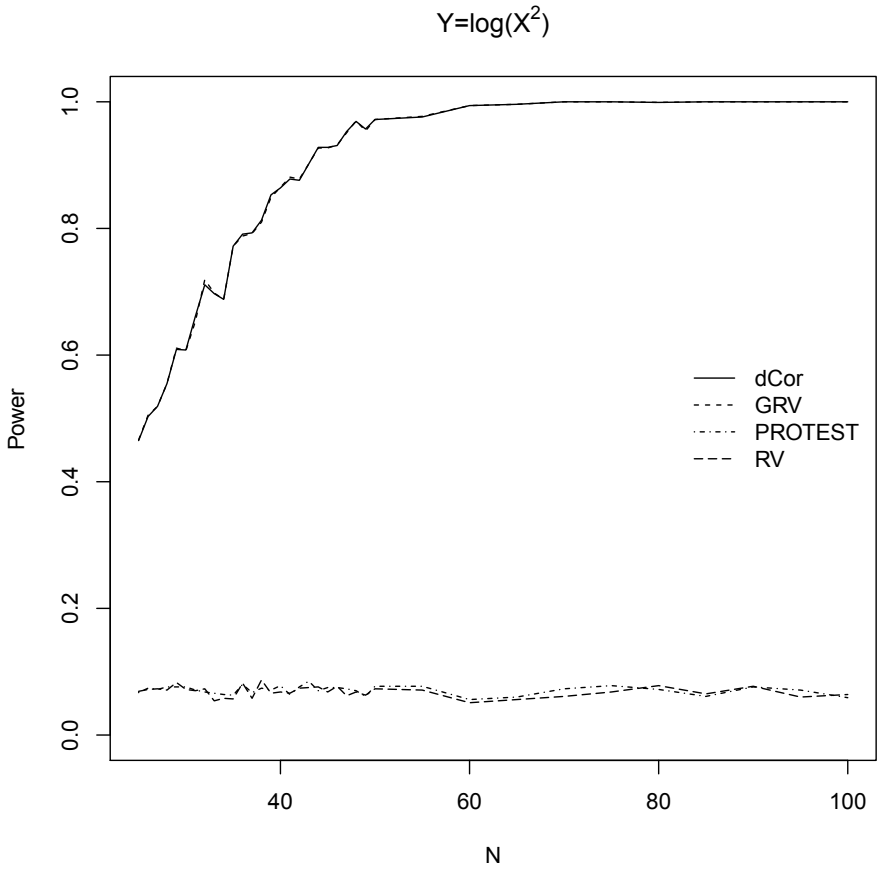

Figure 7.10: Power versus N for 1000 Monte Carlo runs for the paired multivariate setup $\boldsymbol{Y} = \log(\boldsymbol{X}^2)$ at the 5% significance level. dCor and GRV behave identically, and outperform PROTEST and RV which are unable to detect the nonlinear dependence.

are real-valued random vectors, both multivariate hypotheses (4.1) and (4.2) can be tested. When Euclidean distances are applied to centered observations of the random vectors, the GRV coefficient equals the RV coefficient, and null hypothesis (4.1) of no correlation between the variables comprising each vector can be tested. When the square-rooted Euclidean distance measure is used, the GRV coefficient equals the squared dCor coefficient, and hence can test null hypothesis (4.2) of independence between the random vectors.

For vector-valued $\mathcal{X}$ and $\mathcal{Y}$, the GRV test has been shown to counter the limitations of the standardized Mantel approach which have been highlighted by previous authors (see Section 4.2). In particular, for orthogonal data matrices, the GRV test does not incorrectly reject the null hypothesis of no association. We have also demonstrated

that the GRV test has higher power than the standardized Mantel test to detect association between correlated scalar-valued variables.

For non-vector-valued $\mathcal{X}$ and $\mathcal{Y}$ and corresponding distance measures, the GRV test was shown to be competitive with standardized Mantel and PROTEST. For semi-metric distance functions, we also demonstrated that using the RV test with principal coordinates arising from corrected distance matrices yields lower power than the GRV test (which is applied with the uncorrected distance matrices). This provides evidence that when testing for no association between distance matrices, applying corrections may not be beneficial, as suggested by Pekalska and Duin (2005).

An approximate null distribution was also proposed for the GRV coefficient, which can be applied for any distance measure. Through the connection between the GRV coefficient and dCor, this allows a test of independence between random vectors to be performed without permutations (permutations are required for the dCor test). In addition, through simulation we showed that the approximation works well for a selection of vectorial and curve distance measures. We also illustrated its applicability to a real imaging genetics dataset. In Sections 8.4 and 9.3.2 we demonstrate the full potential of the GRV test by applying it to two studies where many tests of no association between distance matrices are required.

Part III

Applications

Chapter 8

Genome-Wide Association Studies of Alzheimer's Disease

In this chapter we describe the use of GWA studies in the pursuit of genetic variants causative of Alzheimer's disease. We survey the existing distance-based approaches which have been applied in this endeavour. Three separate studies are then performed using the DBF test, the pseudo F test and the GRV test with data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Each distance-based method identifies well-known genetic variants, suggesting their validity in GWA studies.

8.1 A Brief Overview

Alzheimer's disease (AD) is a common neurodegenerative condition which causes sufferers to progressively lose their mental and physical functions (Bertram *et al.*, 2010). It is influenced by both environmental and genetic factors, and is thought to be moderately to highly heritable (Gatz *et al.*, 2006; Braskie *et al.*, 2011). In other words, many clues about the causes of AD lie within the genome, and GWA studies have been successfully performed to find the genetic variants across the genome associated with AD; an up to date list can be found on the AlzGene database at http://www.alzgene.org/.

In GWA studies, the genome is searched for genetic variants associated with disease risk (see, for example, Altshuler *et al.* (2008) and Pearson and Manolio (2008)). That is, genetic variation, which can be captured by observing SNPs, is related to disease

risk. Human SNPs are biallelic genetic markers which are comprised of a combination of two alleles; a major allele occurring more commonly in the study cohort, and a minor allele which is less common (this is considered to be the risk allele). The possible combinations are 'major, major', 'major, minor' and 'minor, minor'. The genotype of an individual at a given SNP is summarized by the discrete-valued minor allele count (i.e., the number of copies of the minor allele), and is thus represented by one element in $\{0, 1, 2\}$. These correspond to homozygotes for the major allele, heterozygotes and homozygotes for the minor allele, respectively. Typically, SNPs contributing to disease have large effects in aggregate but only small effects individually (Braskie *et al.*, 2011; Silver *et al.*, 2012). This has motivated multi-locus GWA studies, where multiple SNPs associated with disease are sought across the genome in a manner which nurtures possible joint effects (Hibar *et al.*, 2011). Examples include the regression approaches of Vounou *et al.* (2010), Silver *et al.* (2012), and Ge *et al.* (2012), which are adopted in favour of mass-univariate approaches which consider SNPs individually.

A common approach of scanning the genome in search of causative variants is to group SNPs together into SNP sets where it is plausible that some dependence exists between them. For example, they can be comprised of SNPs in the same gene or biological pathway (Mukhopadhyay *et al.*, 2010; Wu *et al.*, 2010; Yang *et al.*, 2009). Such groupings, however, ignore intergenic regions which may harbour useful information. These regions can be included by using a sliding window of fixed length which partitions the entire genome, chromosome-by-chromosome, into overlapping SNP sets (the window is moved one SNP at a time). Typically, window lengths of 2 to 9 SNPs have been used in application to GWA studies (Mathias *et al.*, 2006; Yang *et al.*, 2009). This yields SNP sets numbering in the hundreds of thousands to be analyzed for association with disease.

For the genome to be analyzed for variants associated with AD, a signature characteristic of AD is required with which to query the genome. This signature is generated by observing a phenotype of AD on the subjects in the study cohort. In traditional GWA studies this has been provided in the form of a dichotomous variable indicating case or control status of each subject, comprising 'case-control' GWA studies. Here, the classification of a subject is determined through clinical or cognitive assessment. However, such assessments are influenced by many factors unrelated to disease, such

as fatigue and anxiety of the subjects for instance (Braskie *et al.*, 2011), and can be misleading. In addition, subjects may not fall clearly into a particular group (Hibar *et al.*, 2011).

Recent interest has turned to considering quantitative imaging-derived signatures characteristic of AD, rather than crude dichotomous indicators (Braskie *et al.*, 2011; Hibar *et al.*, 2011; Ge *et al.*, 2012). In the case of AD, so-called 'neuroimaging phenotypes' are extracted from scans of the brain, such as those obtained via magnetic resonance imaging (MRI) or positron emission tomography (PET). The imaging data is typically represented by a very high-dimensional vector of voxels where each voxel represents the measurement of a 3-dimensional region of the brain measured in $O(\text{mm}^3)$. Such data provides visible clues of how the brain works differently in cases where, for example, a subject is suspected of having AD but the symptoms are not clearly evident through behavioural changes (Braskie *et al.*, 2011). Thus, imaging-derived phenotypes may lead to improved power in detecting causative genetic variants associated with AD (Hibar *et al.*, 2011).

The term 'imaging genetics' refers to the paradigm of seeking genetic variants across the genome which are associated with imaging phenotypes. GWA studies can be further categorized based on the manner in which the imaging phenotypes are considered. For instance, in 'candidate-phenotype' GWA studies, a set of voxels are preselected from all observed voxels and are held fixed while the genome is searched. An alternative approach is offered by 'brain-wide' GWA studies, where all available voxels are searched analogously to the SNPs comprising the genome.

In the following subsections separate reviews are provided of distance-based approaches which have been used in case-control and brain-wide GWA studies. To our knowledge no distance-based approaches have been considered for candidate-phenotype GWA studies.

8.1.1 Distance-Based Case-Control GWA Study Methods

Two prominent distance-based testing procedures derived specifically for case-control multi-locus GWA studies are KBAT (Mukhopadhyay *et al.*, 2010) and a logistic kernel-machine regression test which we denote LKMT (Wu *et al.*, 2010). For these approaches the genome is partitioned into SNP sets, and for each SNP set KBAT and

LKMT make use of similarities between observations rather than distances. This is achieved through the use of kernel functions $K(\cdot,\cdot)$ (Shawe-Taylor and Cristianini, 2004) which are closely related to distance measures, for instance, the IBS kernel function is equal to one minus the IBS distance measure.

KBAT (Mukhopadhyay *et al.*, 2010) tests for joint association of SNPs with a given disease by using similarities between subjects for each individual SNP as observations in separate classical ANOVA models. For each ANOVA model the classical within- and between-group variance quantities are computed. A combined within-group quantity is obtained by summing the within-group variance terms corresponding to all SNPs in the SNP set, and similarly for the between-group quantities. The KBAT test statistic is then formed as the ratio of these two quantities. Since the similarities are not normally distributed and are not all independent, significance of the KBAT statistic is assessed by permutations of the SNP observations across the groups. This approach suffers from two limitations. Firstly, joint association is modeled in an ad-hoc manner, as similarities between subjects are only measured for individual SNPs and then combined. That is, similarities are not computed using the information across all SNPs in the SNP set simultaneously. Secondly, due to the vast number of tests required to conduct a GWA study, the requirement to conduct permutations for inference will cause difficulties in implementation (the method was only presented with simulated data and an example with real data was not provided).

LKMT (Wu *et al.*, 2010) is a logistic regression approach which uses similarities between individuals based on all SNPs in the given SNP set (instead of individually) and does not require permutations. The approach consists of modeling the probability that a given subject is a case subject as a linear function of covariates such as age and sex, which can be referred to as non-SNP covariates, and a linear function of the similarities between that subject and all other subjects given observations of the SNPs. In this way, the joint effect of multiple SNPs, and possible nonlinear interactions between them, can be captured through the similarities and used to model case-control status.

Denote the case-control status of the N subjects by $\{s_i\}_{i=1}^N$, where $s_i = 0$ for controls and $s_i = 1$ for cases, the SNP set observations by $\{\boldsymbol{z}_i\}_{i=1}^N$, and the observations

of M non-SNP covariates by $\{\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iM})^T\}_{i=1}^N$. The model is then given by

$$\text{logit}[P(s_i = 1)] = \beta_0 + \sum_{m=1}^{M} \beta_m x_{im} + h(\boldsymbol{z}_i),$$

for $i = 1, \ldots, N$, where β_0 is an intercept, $\{\beta_m\}_{m=1}^M$ are regression coefficients, and $h(\boldsymbol{z}_i) = \sum_{j=1}^N \gamma_j K(\boldsymbol{z}_i, \boldsymbol{z}_j)$ for some constants $\{\gamma_j\}_{j=1}^N$. The null hypothesis of no SNP effect on case-control status, i.e., $H_0 : \boldsymbol{h} = \boldsymbol{0}$ where $\boldsymbol{h} = (h(\boldsymbol{z}_1), \ldots, h(\boldsymbol{z}_N))^T$, is tested by using a variance-component score statistic $Q = (\boldsymbol{s} - \hat{\boldsymbol{p}})^T \boldsymbol{K} (\boldsymbol{s} - \hat{\boldsymbol{p}})/2$, where $\boldsymbol{s} = (s_1, \ldots, s_N)^T$, $\boldsymbol{K} = \{K(\boldsymbol{z}_i, \boldsymbol{z}_j)\}_{i,j=1}^N$, and $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_N)^T$ is such that $\left\{\text{logit}\,(\hat{p}_i) = \hat{\beta}_0 + \sum_{m=1}^M \hat{\beta}_m x_{im}\right\}_{i=1}^N$. The distribution of Q under the null hypothesis is a mixture of Chi-squared distributions, which via the Satterthwaite method can be approximated by a scaled Chi-squared distribution. Inferences for each SNP set across the genome can then be drawn without permutations.

8.1.2 Distance-Based Brain-Wide GWA Study Methods

To our knowledge, brain-wide GWA studies of AD have not been performed where both the imaging data and the genetics data are subjected to distance-based representations. A very recent distance-based approach is the least squares kernel machine approach of Ge *et al.* (2012), which we denote LSKM, where similarities between samples are considered for multi-locus SNP sets obtained by grouping together SNPs in the same gene. That is, similarities are considered only for the genetics data, and not the imaging data.

This approach combines two areas: semi-parametric regression modeling and random field theory (RFT). The regression model is used to relate the scalar-valued observations of an imaging trait at a given voxel to non-SNP covariates linearly, and to the SNPs in the given SNP set nonlinearly. This nonlinear component is captured via similarities between SNP set observations obtained through the IBS kernel function, as in LKMT used for case-control multi-locus GWA studies. The RTF element of the approach is concerned with performing inference across the brain where multiple-testing corrections are required. The objective is to detect localized regions of high-intensity effects for individual voxels (voxel-wise inference), or spatial regions

represented by sets of contiguous voxels (cluster-wise inference), respectively, which exhibit association with the given SNP set.

The semi-parametric regression model is defined as follows. Define the N observations of voxel v by $\boldsymbol{y} = (y_{v1}, \ldots, y_{vN})^T$, the SNP set observations by $\{\boldsymbol{z}_i\}_{i=1}^N$, and the observations of M non-SNP covariates by $\left\{\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iM})^T\right\}_{i=1}^N$. Then

$$y_{vi} = \sum_{m=1}^{M} \beta_m x_{im} + h(\boldsymbol{z}_i) + \epsilon_{vi},$$

for $i = 1 \ldots, N$, where $\{\beta_m\}_{m=1}^M$ are regression coefficients, ϵ_{vi} are errors assumed to be distributed $N(0, \sigma_v^2)$ for unknown voxel-specific variance σ_v^2, and $h(\boldsymbol{z}_i)$ represents the nonlinear effect of the multiple SNPs in the SNP set determined through kernel function $K(\cdot, \cdot)$. The null hypothesis that the SNP measurements do not explain the measurements of voxel v is tested by using the score statistic

$$Q(v) = \frac{1}{2\hat{\sigma}_v^2} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{K} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right),$$

where $\boldsymbol{X}$ is the $N \times M$ matrix of non-SNP covariate measurements, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta_M})^T$ are the maximum likelihood estimates of the corresponding regression coefficients, $\hat{\sigma}_v^2$ is the maximum likelihood estimate of σ_v^2, and $\boldsymbol{K} = \{K(\boldsymbol{z}_i, \boldsymbol{z}_j)\}_{i,j=1}^N$. As with LKMT, the null distribution of $Q(v)$ is approximated by a scaled Chi-squared distribution.

For a given SNP set, the above model is applied independently to each voxel across the brain, yielding an observed statistic with a corresponding approximate p-value. To perform a voxel-wise analysis of the brain, a correction must be applied to these p-values. This is achieved by using notions from RFT, where the voxels are modeled as random elements comprising a random field. In particular, a familywise error corrected p-value is obtained which accounts for the volume and smoothness of the corresponding statistic mapped to this random field. Given the observed statistic $\hat{Q}(v)$ at voxel v, the corrected p-value is estimated by using permutations as follows. For every permutation π of N_π Monte Carlo permutations, the statistic value for each voxel is computed, and the maximum across all voxels stored. Thus for each permutation a 'maximal statistic' M_π is found as

$$M_\pi = \max_v |\hat{Q}_\pi(v)|,$$

where $\hat{Q}_\pi(v)$ is $Q(v)$ computed with the permuted kernel matrix $\boldsymbol{K}_\pi$ ($\boldsymbol{K}$ with rows and columns simultaneously permuted by π). The corrected p-value is then estimated by

$$\frac{\#(M_\pi \geq \hat{Q}_\pi(v))}{N_\pi},$$

and where $\#(M_\pi \geq \hat{Q}_\pi(v)) < 10$, a generalized Pareto distribution is used to approximate the tail of the permutation distribution such that small corrected p-values can be obtained more accurately. This greater accuracy for smaller p-values is required because a further Bonferroni correction is applied for the multiple-testing problem arising from considering multiple SNP sets across the genome.

The above corrected p-values can also be obtained for cluster-wise inference across the brain. Here, sets of contiguous voxels, which we call voxel sets, are considered across the brain. For each voxel set, the independently-derived statistic values of each voxel are combined (we omit the theoretical details). In doing so, practitioners are able to model spatial information exhibited by regions of the brain, which are represented by a multiple voxels, in terms of the multiple effects (and possible interactions) of SNPs in the chosen SNP set.

However, this is an ad-hoc approach to modeling spatial information of the brain because the predictive relationship of the SNP set is modeled individually for each voxel, and then combined. Perhaps modeling the explanatory relationship of a given SNP set directly on the voxel set, rather than on individual voxels, would be a suitable alternative. Furthermore, applying distances to the voxel data can potentially yield interesting patterns driven by distributed spatial patterns in the data, and these may be taken advantage of within the imaging genetics paradigm. We demonstrate this in Sections 8.3 and 8.4, where candidate-phenotype GWA studies are conducted with both the pseudo F test (distances only used for the imaging data), and the GRV test (distances used for both imaging and genetics data).

8.2 Case-Control Multi-Locus GWA Study of Alzheimer's Disease with the DBF Test

In this section we describe a case-control multi-locus GWA study of AD using the DBF test. On first describing the data, we discuss the choice of sliding window used

to obtain the SNP sets across the genome and which SNP distances to apply. We then describe the application of the DBF test to the data, present findings, and demonstrate the competitiveness of DBF with LKMT on a particular subset of the data.

8.2.1 Data Description

The data used is described in Vounou *et al.* (2010) and was obtained from the ADNI database (http://loni.ucla.edu/ADNI/). It consists of 254 subjects, 101 cases of AD and 153 controls, all genotyped at $316,348$ SNPs across chromosomes 1 to 22.

8.2.2 Choice of Sliding Window and SNP Distance Measure

We apply a sliding window of length 5, which is chosen somewhat arbitrarily. The key point here is that we wish to scan the genome and identify joint effects of SNPs which lie within close proximity to each other. If the window length is too large such that the window highlights a SNP set containing a small proportion of causative SNPs, their signal may be hidden by the non-causative SNPs. This may cause difficulty in locating the exact positions on the genome where the causative SNPs are located. A smaller window, however, will detect the signal more accurately, even if not all of the causative SNPs in the neighbourhood are highlighted within the window. Where causative SNPs are positioned side-by-side across the genome, their number is unknown and may differ at different locations. With a small window it is expected that the exact position of such causative SNPs will be more easily identifiable than by using a large window.

Having chosen a window length of 5, the N individuals are represented by discrete-valued 5-dimensional vectors in each SNP set. This results in a total number of $316,260$ SNP sets to be compared across the two populations.

Now we turn to the issue of which SNP distance measure to apply. Many measures exists (see, for instance, Selinski and Ickstadt (2005) and Appendix B.3). In an exploratory endeavour, we use five distance measures; the IBS, Simple Matching, Sokal and Sneath, Rogers and Tanimoto I, and Hamman I distances. The IBS distance is commonly used in GWA studies (Mukhopadhyay *et al.*, 2010; Wu *et al.*, 2010) and quantifies the difference in the proportion of risk alleles shared across the SNP set. That is, subjects are deemed less dissimilar if they have more risk alleles in common.

The Simple Matching, Sokal and Sneath, Rogers and Tanimoto I, and Hamman I distances all quantify differences between subjects based on the number of mismatches and matches in minor allele counts across the SNPs of the SNP set. The Simple Matching distance, for instance, considers the proportion of matches across the SNPs. The Sokal and Sneath, and Rogers and Tanimoto I distances consider two different ways of quantifying the ratio of mismatches to matches in minor allele counts across the SNPs. Subjects are deemed less dissimilar as this ratio increases. The Hamman I distance quantifies dissimilarity based on the difference in the number of matches and mismatches as a proportion of the number of SNPs.

8.2.3 Experimental Results

For each SNP set the DBF statistic and corresponding approximate p-value is computed using all five distance measures. For each distance measure this results in the simultaneous observation of $316,260$ p-values, and hence a large multiple-testing problem.

Declaring SNP sets as significant based on each individual p-value being below a stated cutoff value will yield an abundance of significant SNP sets. It is expected that only a few of these are truly significant, so multiple-testing corrections can be applied in pursuit of these truly significant SNP sets. Typically one of two approaches is deployed in this search. The first controls the familywise error rate by the well-known Bonferroni correction (Hochberg and Tamhane, 1987). That is, the probability that at least one SNP set is called significant when it is truly null (a false positive) is controlled. However, of the SNP sets called significant, the expected proportion of truly null SNP sets is unknown because the total number of truly null SNP sets is unknown. The second approach controls the false discovery rate, which is the rate at which SNP sets are truly null if they are called significant (Benjamini and Hochberg, 1995). Here, the expected proportion of truly null SNP sets of all of those called significant is known (it is directly controlled), even though the true total number of null SNP sets is still unknown. Thus it is expected to yield less truly null SNP sets than by controlling the familywise error rate. Approaches to control the false discovery rate include the Benjamini-Hochberg correction of Benjamini and Hochberg (1995) and the q-value approach of Storey and Tibshirani (2003).

In this case we control the familywise error rate by adopting a genome-wide significance threshold of 10^{-7}. Wu *et al.* (2010) state that such a threshold is very stringent and difficult to attain, so we adopt it here to show that with the proposed null sampling distribution such p-values can be attained.

In Figure 8.1 we provide a Manhattan plot which depicts the significant SNP sets across the entire genome for the Sokal and Sneath distance measure, showing the greatest effects in chromosomes 18 and 19. The results of all distance measures are summarized by the unique SNP and gene combinations identified; see Table 8.1. All significant SNPs are identified in chromosomes 18 and 19. In particular, chromosome 19 contains two genes, APOE and TOMM40, which are the major genetic variants found in many studies (see, for example, Braskie *et al.* (2011) and Shen *et al.* (2010)). Other reported genetic variants in chromosome 19 that overlap with our findings include APOC4, PVRL2 and CLPTM1 (Takei *et al.*, 2009; Yu *et al.*, 2007). The DCC gene has also been previously identified (Bredesen, 2009; Lourenco *et al.*, 2009).

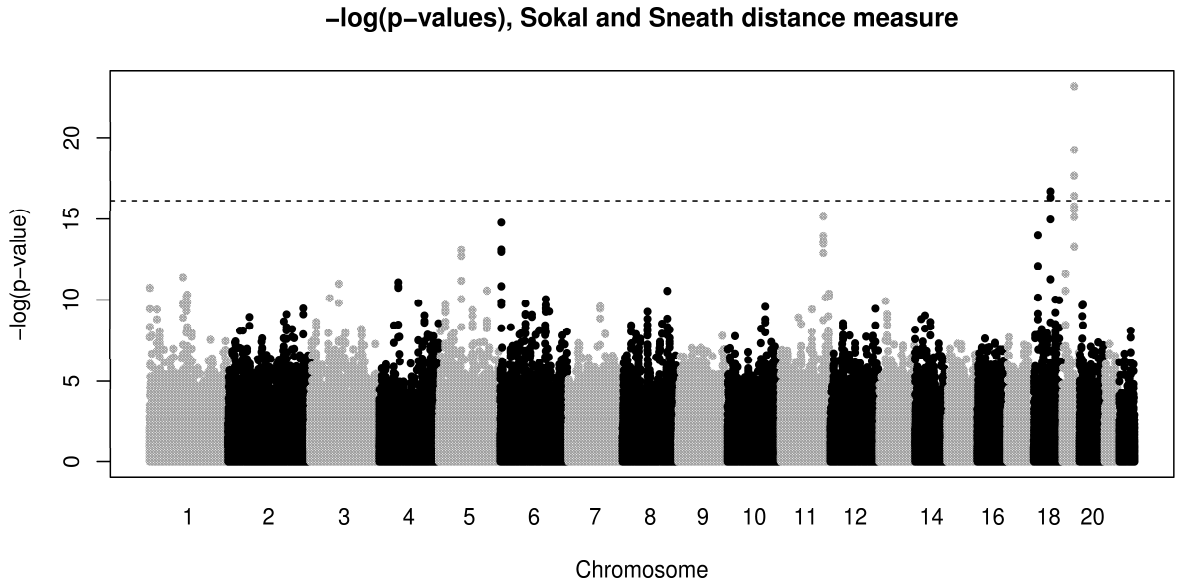

Figure 8.1: Manhattan plot of the -log(p-values) computed across the genome with the DBF test applied with the Sokal and Sneath genetic distance measure. Each point represents a window containing a multi-locus SNP set consisting of 5 contiguous SNPs. The dashed line represents the genome-wide significance threshold of -log(10^{-7}). The black and gray colours are used to distinguish between adjacent chromosomes.

We also compare our results with those obtained using LKMT. Since LKMT makes use of an approximate distribution rather than permutations, it is a direct competitor

of the DBF test in case-control GWA studies with no non-SNP covariates. We apply LKMT across the SNP sets of chromosome 19, which is the chromosome in which we obtained the smallest p-values. We apply the IBS kernel function and monitor the approximate p-values which result from LKMT. Figure 8.2 provides a visual comparison of the p-values obtained by both methods in this chromosome. Note that both methods identify the same SNPs at the significance threshold of 10^{-7}, i.e., the ones listed in Table 8.1. This provides evidence that DBF performs comparably with LKMT, as they both identify the well-known APOE and TOMM40 genes.

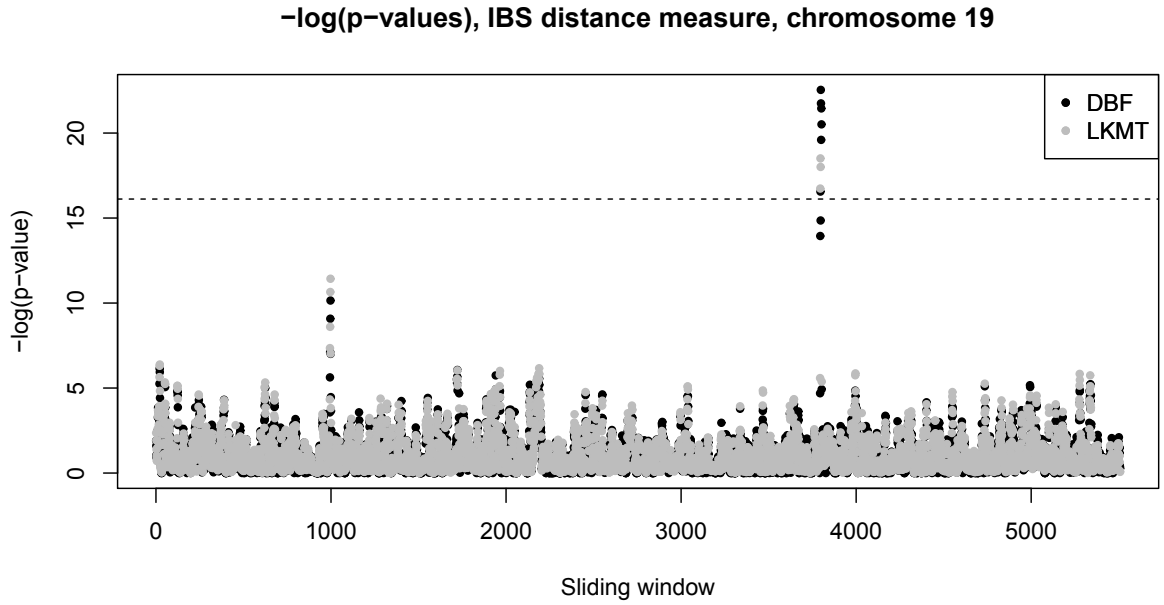

Figure 8.2: Manhattan plot of the -log(p-values) computed across chromosome 19 using the DBF and LKMT tests with the IBS distance measure and IBS kernel function, respectively. Each point represents a multi-locus SNP set consisting of 5 contiguous SNPs, and the dashed line represents the transformed genome-wide significance threshold of -log(10^{-7}).

8.3 Candidate-Phenotype Multi-Locus GWA Study of Alzheimer's Disease with the Pseudo F Test

In this section we describe a candidate-phenotype multi-locus GWA study of AD in which the pseudo F test is used with distances applied to the imaging data. The objective is to model the explanatory relationship of a multi-locus SNP set on a set of

voxels which have been preselected from all the available voxels. On first describing the data, we describe how the voxels have been selected. We then describe how a distance-based signature characteristic of AD can be derived from these voxels. The application of the pseudo F test to the data is then described and we present the findings.

8.3.1 Data Description

The data is described in Silver *et al.* (2012) and was obtained from the ADNI database. The original sample consists of 464 elderly subjects representing three groups; 99 have AD, 211 exhibit mild cognitive impairment (MCI), and 154 are healthy controls. They have been genotyped across the entire genome, and longitudinal MRI brain scans have been obtained. For this study we only consider the 253 AD and control samples.

Originally, the genetic markers observed on the subjects consist of SNPs and copy number variations (CNVs). Only SNPs in chromosomes 1 to 22 are considered for this study, and after pre-processing $434,271$ remain for analysis (Silver *et al.*, 2012).

The longitudinal MRI scans of the subjects' brains were observed at screening and followed up at 6, 12 and 24 months. To derive a neuroimaging phenotype from these, Silver *et al.* (2012) selected the subset of voxels deemed most discriminative between AD and control status. This subset was found via the following data-driven approach. For each subject a slope coefficient was obtained for each voxel representing the ventricular volumetric change across the 3 time-points relative to baseline, i.e., the initial scan. ANOVA was then performed at each voxel to quantify the difference between the average slope coefficients observed for the AD and control subjects, adjusting for age and sex. On applying a familywise error rate of 5% to the ANOVA p-values, a subset of $148,023$ voxels exhibited significant non-zero differences between average AD and control slope coefficients. The neuroimaging phenotype data matrix used in this study is comprised of the slope coefficients of each of the 253 subjects at the selected $148,023$ voxels.

8.3.2 Choice of Image Distance Measure

In order to derive a distance-based signature characteristic of AD given the observations of the preselected voxels, we choose a distance measure from a selection such that separation is exhibited between the AD and control samples.

Distances which have been applied in the imaging literature include the Euclidean distance, Pearson's correlation distance and the NMI distance (Michaels *et al.*, 1998; Holden *et al.*, 2000). We therefore consider these. In addition, we consider the Manhattan, Maximum, and Spearman's correlation distances. This selection of distance measures allows a range of possibly complex relationships to be captured across the observations of the voxels. For instance, Pearson's and Spearman's correlation distances capture positive linear relationships between the observations, while the NMI distance captures dependence between them.

Having obtained each of these distance matrices from the voxel data, we apply the DBF test to quantify the separation between the AD and control samples. In Figure 8.3 we provide 2-dimensional MDS plots of the samples showing the separation exhibited by each distance matrix, ranked in descending order of their respective DBF statistic values; Spearman's correlation (DBF=0.4504), Pearson's correlation (DBF=0.4351), Manhattan (DBF=0.2788), Euclidean (DBF=0.254), Maximum (DBF=0.1655) and NMI (DBF=0.005) (all are significant except for the NMI distance). The correlation distances exhibit the most separation, indicating a difference in the strength of positive linear relationships between the ventricular volumetric changes of the AD and control samples.

To see even more clearly how this separation is depicted, heatmaps of the normalized centered inner product matrices of the Spearman's correlation, Euclidean and NMI distance matrices are presented in Figure 8.4. We see that the clearest signal is indeed provided by the normalized centered inner product matrix arising from Spearman's correlation centered inner product matrix. We therefore retain the corresponding distance matrix and use it in the GWA study.

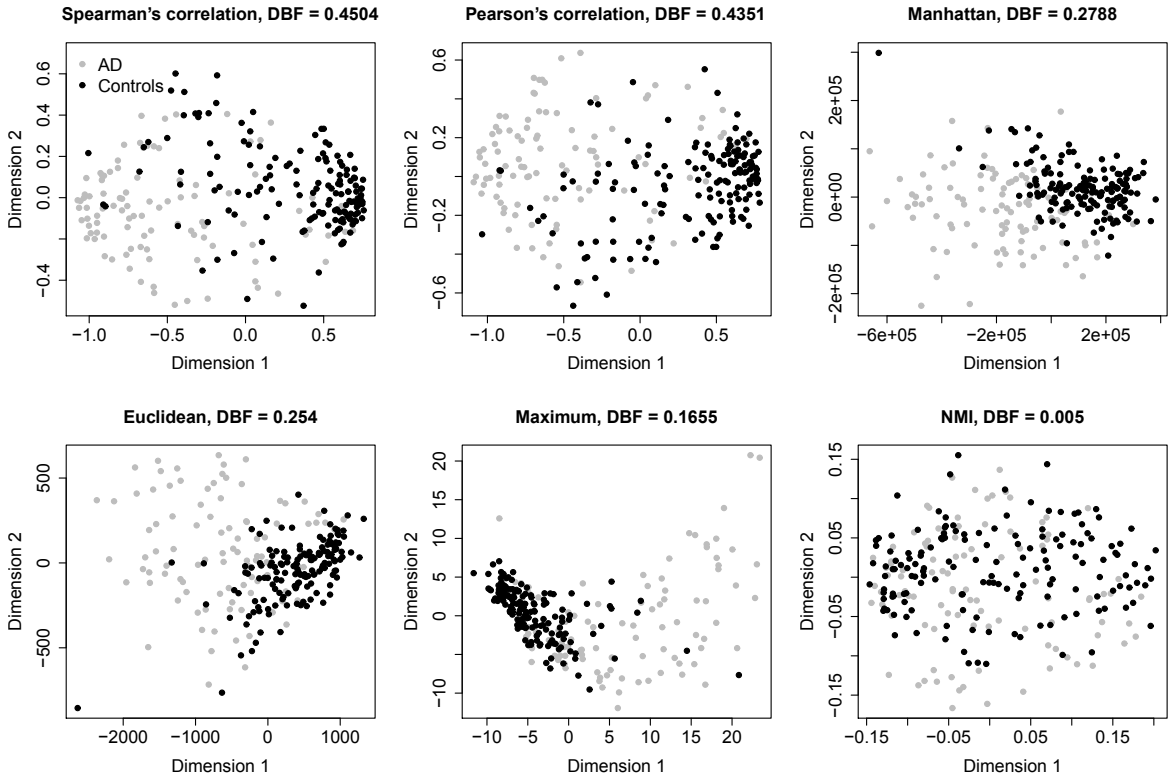

Figure 8.3: 2-dimensional MDS plots showing the separation exhibited by each of the distance measures between AD and control samples: Spearman's correlation, Pearson's correlation, Manhattan, Euclidean, Maximum and NMI. Spearman's and Pearson's correlation exhibit the most separation, as indicated by their DBF statistic values. The NMI distance exhibits the least separation, achieving the lowest DBF value.

8.3.3 Experimental Results

We apply a sliding window of 7 contiguous SNPs across the genome chromosome-by-chromosome. This results in 434,139 multi-locus SNP sets, with the observations of each forming a 253×7 predictor matrix of full rank (a sliding window of length 5 resulted in predictor matrices which were not all of full rank). The pseudo F test is applied to model the variation observed in the chosen image distance matrix in terms of each of the SNP set predictor matrices. The p-values have been presented in the Manhattan plot shown in Figure 8.5, showing that the main effects highlighted by a significance threshold of 10^{-7} are in chromosome 19.

To identify significant SNP sets, and hence SNPs, we apply multiple-testing corrections to the p-values and set a threshold significance level. The Bonferroni correction is applied and the familywise error rate controlled at 5%, and both the Benjamini-

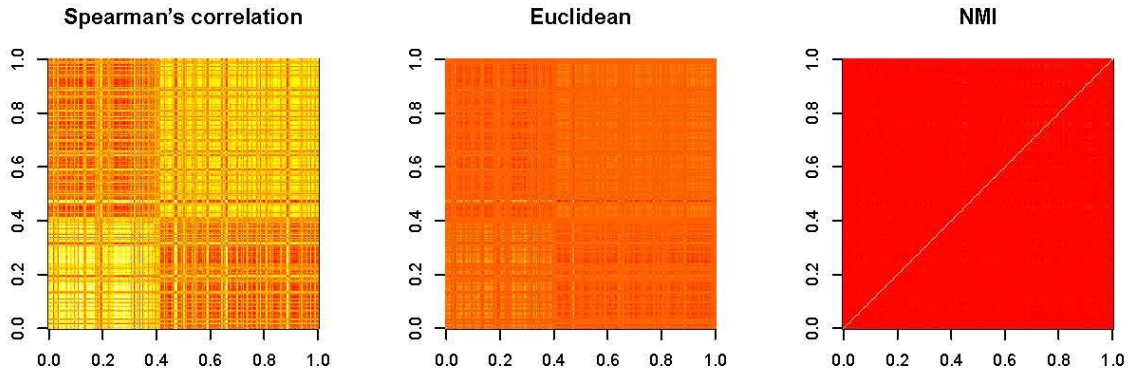

Figure 8.4: Heatmaps of the normalized centered inner product matrices arising from the Spearman's correlation, Euclidean and NMI distance matrices. The greatest separation is visible from Spearman's correlation distance. The Euclidean distance exhibits a much weaker separation, and in comparison the NMI distance exhibits no separation.

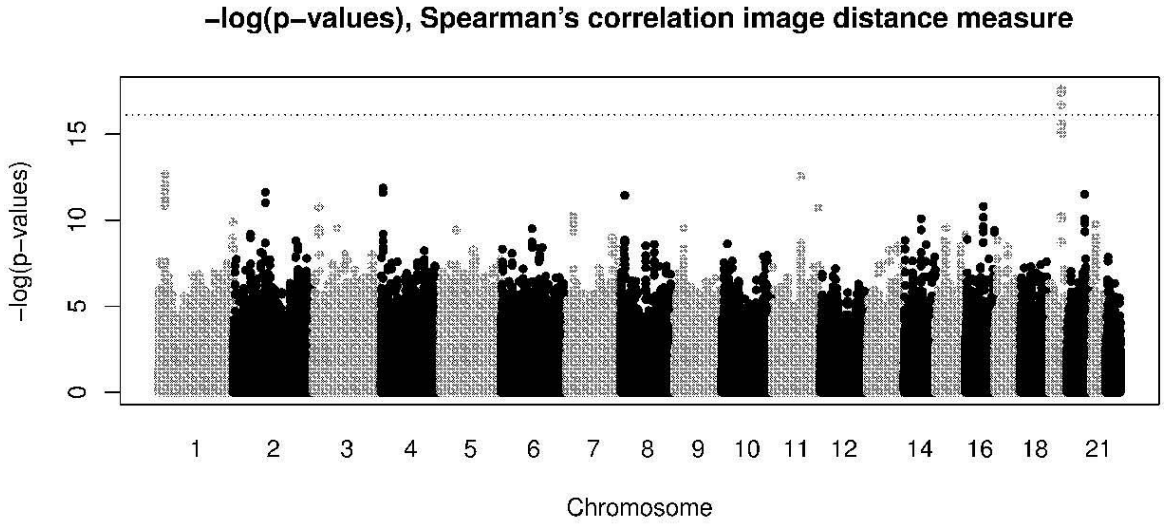

Figure 8.5: Manhattan plot of the -log(p-values) computed across the genome with the pseudo F test applied with the Spearman's correlation image distance measure. Each point represents a window containing a multi-locus SNP set consisting of 7 contiguous SNPs. The dashed line represents the genome-wide significance threshold of $-\log(10^{-7})$. The black and gray colours are used to distinguish between adjacent chromosomes.

Hochberg correction and the q-value approach of Storey and Tibshirani (2003) are applied and the false discovery rate controlled at 5%. With all corrections the significant SNP sets are located in chromosome 19. The results are summarized in Table 8.2, and indicate the SNPs identified are located in the well-known genes associated with

AD; APOE, TOMM40, APOC4, PVRL2 and CLPTM1. These overlap with the results of the DBF test described in Section 8.2. In addition, we also find rs10413089 which has been previously identified by Cervantes *et al.* (2011) via a fine mapping analysis of the APOE cluster genes (APOE, APOC1, APOC4, APOC2, and TOMM40).

8.4 Candidate-Phenotype Multi-Locus GWA Study of Alzheimer's Disease with the GRV Test

In this section we report on a candidate-phenotype multi-locus GWA study of AD conducted with the GRV test. Here a distance measure is used for both the imaging data and the genetics data. The imaging data and corresponding distance measure used are as described in Section 8.3. The SNP data is also as in Section 8.3, for which a range of distance measures and sliding window lengths are used in conjunction with the GRV test. We report on the findings and also provide some illustrative examples demonstrating how the GRV test works for this real dataset.

8.4.1 Experimental Results

Within the GRV testing framework the image centered inner product matrix remains fixed, and results from the Spearman's correlation distance matrix selected as in Section 8.3. We then apply three separate sliding windows across the genome; following Yang *et al.* (2009) we consider one of length 3, one of length 5 and one of length 7. For each SNP set obtained with each sliding window, the IBS, Sokal and Sneath, and Rogers and Tanimoto I distances are computed. This results in nine separate GWA studies, where the genome is partitioned into $434,227$ SNP sets containing 3 SNPs, $434,183$ SNP sets containing 5 SNPs, and $434,139$ SNP sets containing 7 SNPs. On applying the GRV test with each SNP distance measure, the corresponding number of p-values are obtained.

The three multiple-testing corrections used in Section 8.3 were also used in this study. To show the different SNP sets identified via these corrections, we present the Manhattan plot obtained by using a sliding window of length 3 and the Sokal and Sneath distance measure in Figure 8.6 (a). The dashed line represents the threshold value controlling the familywise error rate at 5%. Figure 8.6 (b) shows the equivalent

p-values after being corrected via the Benjamini-Hochberg procedure and 8.6 (c) shows the equivalent q-values after using the approach of Storey and Tibshirani (2003). In both of these plots the dashed line represents the threshold value controlling the false discovery rate at 5%. For the Bonferroni and q-value approaches the greatest effects are highlighted in chromosome 19, but for the Benjamini-Hochberg approach some effects are also located in chromosome 6.

The results of all distance measures and window lengths are summarized in Tables G.1, G.2 and G.3 in Appendix G, one for each setting of sliding window length. The majority of SNPs idenfitied are located in chromosome 19, and overlap with the findings of the DBF and pseudo F tests; they are located in genes APOE, TOMM40, APOC4, PVRL2 and CLPTM1. As with the pseudo F test approach, rs10413089 is also identified in chromosome 19. In addition to these previously reported variants, we identified several SNPs in chromosomes 1 and 6 as being associated with AD, none of which appear in the literature on Alzheimer's disease.

Finally, we give an illustrative example of how the GRV test works by considering two SNP sets of length 5, and applying the IBS genetic distance measure. One SNP set contains the apoe4 SNP, representing a well-known genetic variant associated with AD, and one contains the rs999562 SNP, representing a variant not known to be associated with AD. For each SNP set the normalized centered inner product matrix arising from the IBS distance matrix is obtained, and hierarchical clustering is applied to give some order to the samples given the observed distances (see, for instance, Venables and Ripley (2002)). The resulting clusters can be visualized using heatmaps or dendrograms; we provide the heatmaps in Figures 8.7 (a) and 8.7 (c). The GRV statistic and p-value indicating the strength of association between the SNP set and the neuroimaging phenotypes is also given, and it is seen that the SNP set containing apoe4 is associated with AD (GRV statistic is 0.104 and corresponding p-value is 1.27×10^{-9}) while the SNP set containing rs999562 is not (GRV statistic is 0.0151 and corresponding p-value is 0.0835).

For the genetic heatmaps arising from each SNP set, the samples are ordered differently as they depend on the separate clustering results. We compare these heatmaps with those arising from the neuroimaging phenotype distances upon applying the same ordering of samples. In this way the genetic and neuroimaging phenotype heatmaps

are directly comparable for each SNP set. The image heatmap given in Figure 8.7 (b) presents the samples ordered based on the clustering results arising from the SNP set containing apoe4. Similar patterns are visible in this heatmap and the genetic heatmap in Figure 8.7 (a), and this similarity is detected by the GRV test. In Figure 8.7 (d) the image heatmap is obtained by ordering the samples using the clustering results arising from the SNP set containing rs999562. Here we do not see clear similarities between the genetic and image heatmaps as in Figures 8.7 (a) and 8.7 (b), suggesting much weaker association, as depicted by the GRV test results.

The GRV statistic has been shown to measure the linear correlation between the elements of two normalized centered inner product matrices. Therefore the degree of similarity between the patterns exhibited by the imaging data and each SNP set can also be observed by looking at the respective scatter plots of the normalized centered inner product matrix elements arising from each. The two scatter plots are provided in Figures 8.8 (a) and (b). In (a) the elements arising from the neuroimaging phenotypes are plotted against the elements arising from the SNP set containing apoe4. The gradient of the superimposed regression line equals the GRV statistic value, and it is clear that there is a linear correlation between the elements of the respective normalized centered inner product matrices. Hence there is an association between the neuroimaging phenotypes and the SNP set containing apoe4. (b) provides the equivalent plot for the elements arising from the SNP set containing rs999562. The regression line has a much lower gradient than in (a), indicating a much weaker association.

8.5 Summary

Imaging genetics is a growing area in which the genetic variants associated with disease risk are sought using imaging phenotypes of disease. The imaging and genetics data are both high-dimensional, and each exhibit complex characteristics. Recent methods have adopted the idea of similarity through the use of kernel functions in order to capture the joint effects and possible interactions of multiple SNPs through similarities. For imaging data, however, the notion of similarity/distance has not been considered when the interest is in locating spatial regions (sets of voxels) of the brain which exhibit effects associated with genetic variation. Instead, ad-hoc procedures using independently-derived information from each voxel are adopted. We have shown

through the use of the pseudo F and GRV tests that all voxels in a given voxel set can be modeled simultaneously through the use of distances. In the applications presented the voxel set was chosen based on a discriminative analysis of all voxels. Regardless of how the voxel sets are defined, the distance-based pseudo F and GRV approaches can still be applied.

We have presented three separate GWA studies of AD; one case-control study and two candidate-phenotype studies. The case-control study was performed using the DBF test, resulting in well-known genetic variants being identified. We have also shown that the DBF test is competitive with the LKMT (when ignoring non-SNP covariates).

For the candidate-phenotype GWA study paradigm we have shown that the pseudo F and GRV tests can be successfully applied to find genetic variants associated with AD. The analyses described comprise the first known GWA studies of AD in which distances are applied to neuroimaging phenotypes. The results indicate that observations of multiple voxels can be considered simultaneously through the use of distances.

The pseudo F and GRV tests offer two different approaches to performing GWA analysis. The pseudo F test models multiple SNPs as predictors of the imaging data, whose variation is assessed through distances. In this regression framework, restrictions on the predictor matrix being of full rank can cause difficulty in its implementation. The GRV test, however, is symmetric in the sense that it does not model the predictive ability of one set of data on another, and so overcomes the limitations hindering the effective implementation of the pseudo F test in GWA studies. It also offers greater flexibility as any distance can be applied to each type of data.

The results indicate that the GRV test utilizing distances from both imaging and genetics data can yield potentially interesting insights when used in conjunction with GWA studies. The well-known SNPs/genes highlighted by the DBF and pseudo F tests were also highlighted by the GRV test, but in addition, other SNPs previously unidentified by other studies were also highlighted. This shows that there are potentially interesting insights to be gained by using distances applied to the imaging data, in addition to the genetics data.

Table 8.1: Significant SNPs and genes identified by the DBF test using each genetic distance measure and a genome-wide significance threshold of 10^{-7}. The chromosome in which the SNPs were identified are given, in addition to the p-value of the sliding window containing the given SNP. Where SNPs are present in more than one selected window, the minimum p-value of the windows is given.

Distance measure	SNP	Gene	Chromosome	P-value of window
IBS	rs2075650	TOMM40	19	1.626×10^{-10}
	rs8106922	TOMM40	19	3.600×10^{-10}
	rs5167	APOC4	19	4.844×10^{-10}
	apoe4	APOE	19	4.844×10^{-10}
	rs3760627	CLPTM1	19	1.237×10^{-9}
	rs405509	APOE	19	3.080×10^{-9}
	rs2075642	PVRL2	19	6.449×10^{-8}
	rs6859	PVRL2	19	6.449×10^{-8}
	rs157580	TOMM40	19	6.449×10^{-8}
Sokal and Sneath	apoe4	APOE	19	8.458×10^{-11}
	rs405509	APOE	19	8.458×10^{-11}
	rs2075650	TOMM40	19	8.458×10^{-11}
	rs8106922	TOMM40	19	8.458×10^{-11}
	rs157580	TOMM40	19	2.104×10^{-8}
	rs1222938	DCC	18	5.736×10^{-8}
	rs12960771	DCC	18	5.736×10^{-8}
	rs1560531	DCC	18	5.736×10^{-8}
	rs2960617	DCC	18	5.736×10^{-8}
	rs3862684	DCC	18	5.736×10^{-8}
	rs2075642	PVRL2	19	7.498×10^{-8}
	rs4803766	PVRL2	19	7.498×10^{-8}
	rs6859	PVRL2	19	7.498×10^{-8}
	rs17748116	DCC	18	8.212×10^{-8}
Rogers and Tanimoto I	rs157580	TOMM40	19	4.129×10^{-10}
	rs2075650	TOMM40	19	4.129×10^{-10}
	rs8106922	TOMM40	19	1.067×10^{-9}
	rs5167	APOC4	19	6.915×10^{-8}
	apoe4	APOE	19	6.915×10^{-8}
	rs405509	APOE	19	6.915×10^{-8}
Simple Matching, Hamman I	apoe4	APOE	19	2.738×10^{-10}
	rs405509	APOE	19	2.738×10^{-10}
	rs157580	TOMM40	19	2.738×10^{-10}
	rs2075650	TOMM40	19	2.738×10^{-10}
	rs8106922	TOMM40	19	2.738×10^{-10}

Table 8.2: Significant SNPs and genes identified by the pseudo F test applied with the Spearman's correlation image distance and a sliding window of length 7, with familywise error and false positive rates controlled at 5%. The chromosome in which the SNPs were identified are given, in addition to the p-value of the sliding window containing each SNP. Where SNPs are present in more than one selected window, the minimum p-value of the windows is given. The columns B (for Bonferroni), BH (for Benjamini-Hochberg) and Q (for q-value) indicate with which p-value correction the SNPs were identified.

SNP	Gene	Chromosome	P-value of window	P-value correction		
				B	BH	Q
rs2075650	TOMM40	19	2.282×10^{-8}	✓	✓	✓
rs8106922	TOMM40	19	2.282×10^{-8}	✓	✓	✓
rs405509	APOE	19	2.282×10^{-8}	✓	✓	✓
apoe4	APOE	19	2.282×10^{-8}	✓	✓	✓
rs439401	APOE	19	2.282×10^{-8}	✓	✓	✓
rs5167	APOC4	19	2.282×10^{-8}	✓	✓	✓
rs10413089		19	2.282×10^{-8}	✓	✓	✓
rs760114	CLPTM1	19	2.700×10^{-8}	✓	✓	✓
rs3760627	CLPTM1	19	2.786×10^{-8}	✓	✓	✓
rs157580	TOMM40	19	5.583×10^{-8}	✓	✓	✓
rs11668758		19	1.685×10^{-7}		✓	✓
rs387976		19	3.070×10^{-7}		✓	✓
rs6859	PVRL2	19	3.070×10^{-7}		✓	✓

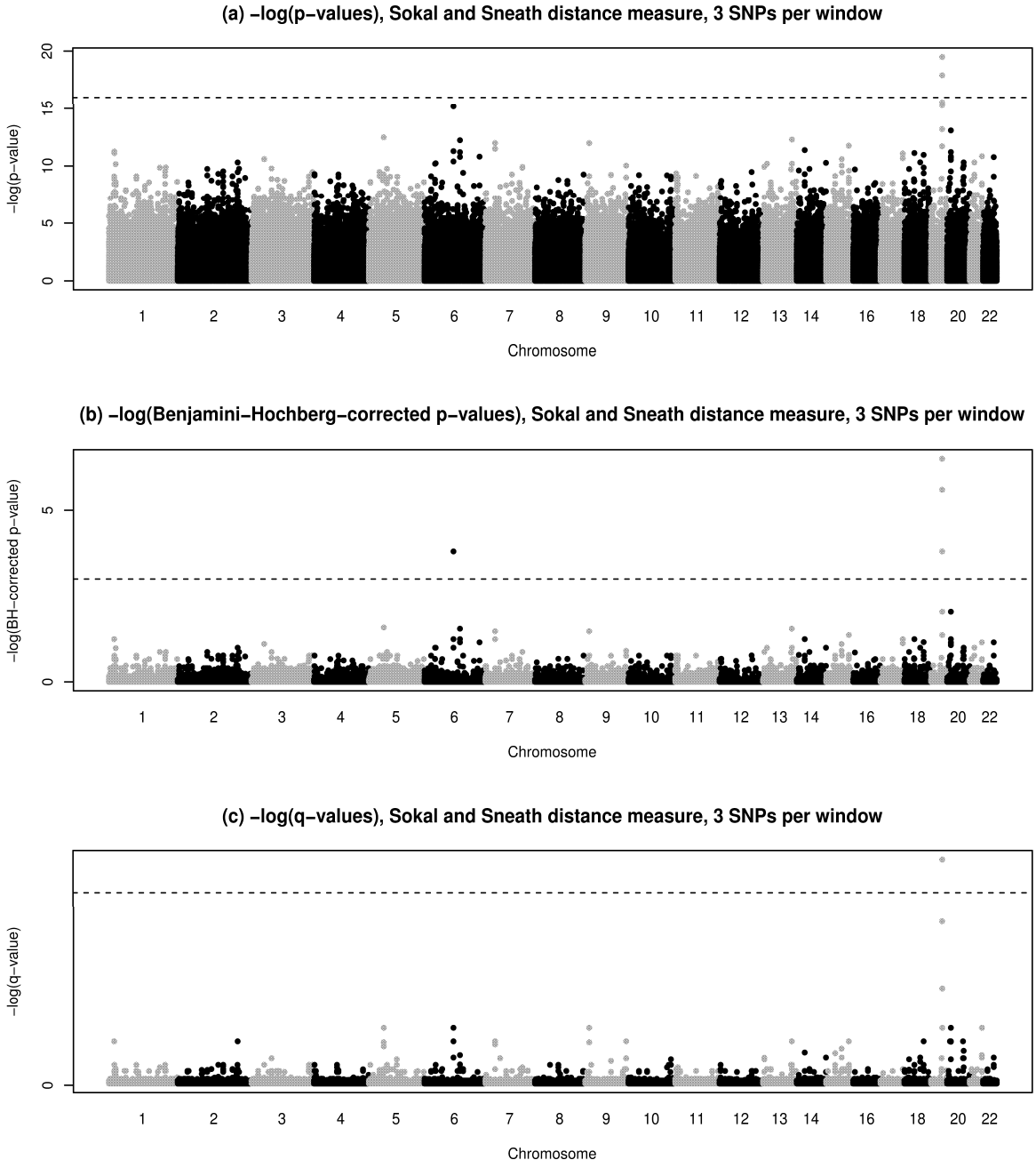

Figure 8.6: Manhattan plots of the -log(p-values) and adjusted -log(p-values) computed via the GRV test across the genome for the Sokal and Sneath genetic distance measure. Each point represents a window containing a multi-locus SNP set consisting of 3 adjacent SNPs. The black and gray colours are used to distinguish between adjacent chromosomes. (a) -log(p-values) with the dashed line representing the Bonferroni significance threshold of -log(0.05/434227) controlling the familywise error rate at 5%. (b) -log(Benjamini-Hochberg-corrected p-values) with the dashed line representing the significance threshold of -log(0.05) controlling the false discovery rate at 5%. (c) -log(q-values) with the dashed line representing the significance threshold of -log(0.05) controlling the false discovery rate at 5%.

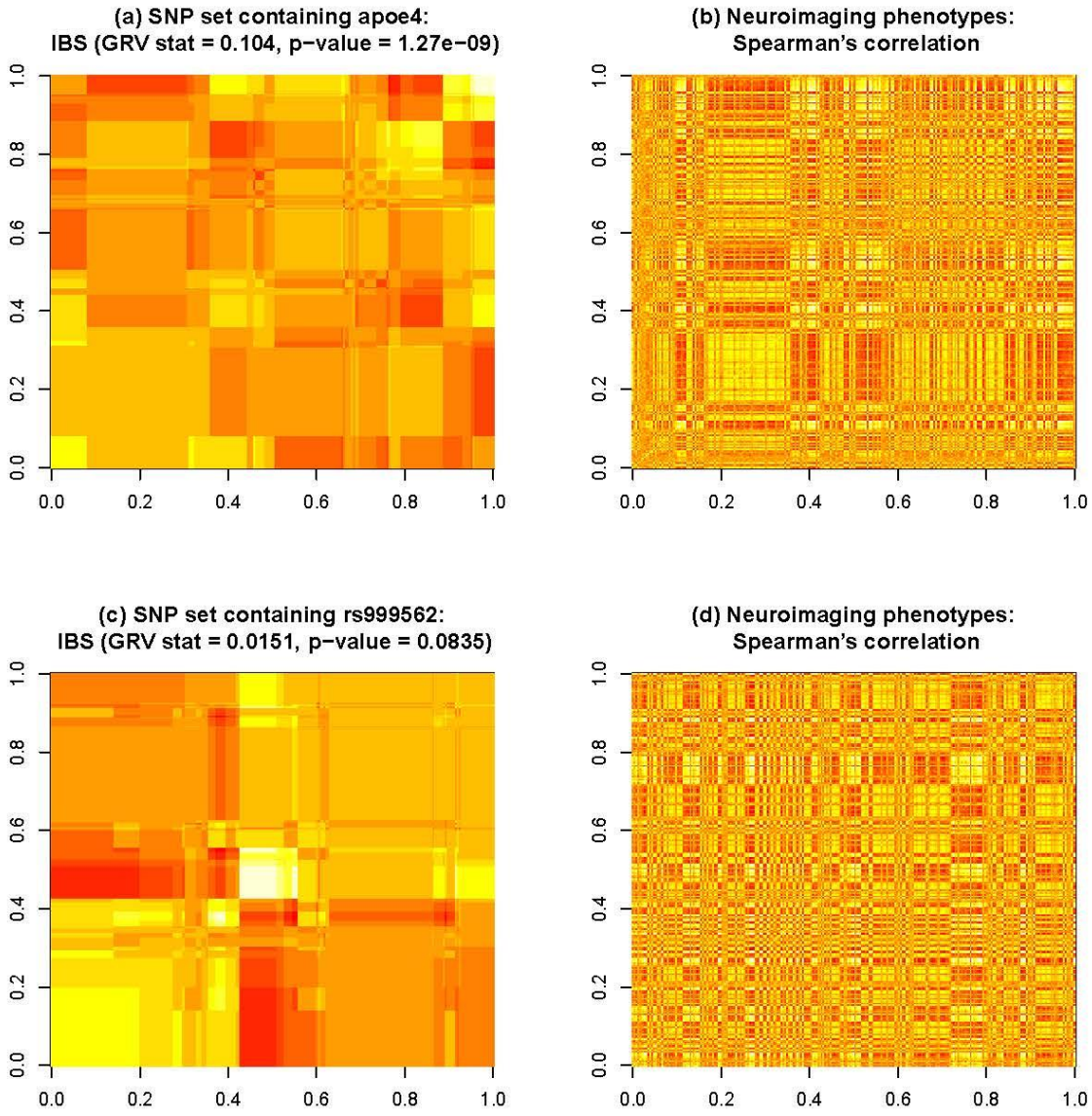

Figure 8.7: Heatmaps of the normalized centered inner product matrices arising from the IBS genetic distances and Spearman's correlation neuroimaging phenotype distances. (a) The genetic distances between samples of the SNP set containing apoe4 with samples ordered using the hierarchical clustering results. (b) The image distances with samples ordered using the clustering results of (a). (c) The genetic distances between samples of the SNP set containing rs999562 ordered using the hierarchical clustering results. (d) The image distances with samples ordered using the clustering results of (c). Similar clusters are visible in heatmaps (a) and (b), indicating a similarity in the patterns of variation exhibited by each. The similarity is quantified by the GRV test statistic of 0.104 with a corresponding p-value of 1.27×10^{-9}. Heatmaps (c) and (d) exhibit much less similarity, with the visible clusters in (c) not being observed in (d). This is quantified by the lower GRV test statistic of 0.0151 with a larger corresponding p-value of 0.0835.

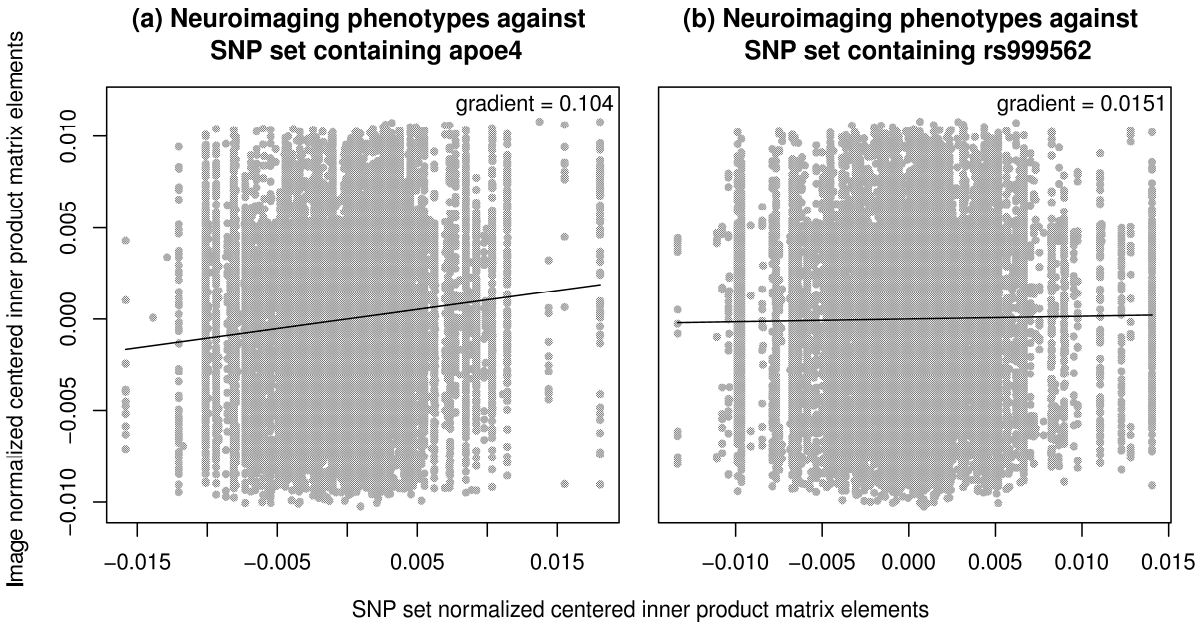

Figure 8.8: Scatter plots of the elements of the normalized centered inner product matrix arising from the neuorimaging phenotype distance matrix against the elements arising from the SNP set distance matrices (gray points). The linear regression lines (black lines) are superimposed indicating the strength of correlation between the values. The gradient of these lines equals the respective GRV statistic values. (a) Neuroimaging phenotypes against the SNP set containing apoe4. (b) Neuroimaging phenotypes against the SNP set containing rs999562. A stronger correlation is evident between the normalized centered inner product matrix elements arising from the neuroimaging phenotypes and the SNP set containing apoe4.

Chapter 9

Microarray Gene Expression Studies

In this chapter we introduce microarray gene expression studies, explaining two biological problems of interest where such data is used. These problems are the identification of genes whose expression differs over time between populations, and gene expression quantiative trait loci (eQTL) mapping where genes associated with SNPs are sought. For each of these problems we review existing approaches, highlighting areas for further development and how these can be addressed by the DBF and GRV tests, respectively. We then apply the DBF test to perform a differential analysis of a human immune cell *M.tuberculosis* dataset, the GRV test to perform an eQTL mapping of ovarian cancer, and present the findings of each study.

9.1 Gene Expression and Microarrays

In genomic experiments, researchers are interested in understanding the role of individual genes or collections of genes in achieving particular biological functions. For these studies, the biological variable of interest is gene expression. For instance, the interest may be in identifying genes which contribute to controlling a given infection, and this can be achieved by studying the effects of disease on gene expression. Alternatively, the interest may be in detecting genes which are differentially expressed between different populations or treatments. This can be achieved by studying the expression of genes within and between populations or treatments.

The starting point of many studies which seek to identify particular genes is the observation of expression levels of thousands of genes on a given cohort of biological replicates, such as cells. In the following paragraphs we describe what is meant by the term 'gene expression level', and how the expression levels of thousands of genes can be simultaneously observed using microarray technology.

Each replicate in the given cohort contains a copy of the complete DNA sequence, and genes are defined by small sections of this sequence. A gene is said to be expressed if there is a transfer of genetic information from the respective section of DNA of which it is comprised, to protein, which is used to perform biological functions within the replicate. This transfer is performed by messenger RNA (mRNA), and the level of gene expression is quantified by the abundance of this mRNA.

The relatively recent advent of microarray technology allows the simultaneous observation of mRNA abundance for thousands of genes using mRNA extracted from a biological replicate (see, for instance, Gibson and Muse (2004)). A microarray is a surface typically of the order of a few centimeters squared, containing many small deposits of DNA, referred to as 'transcripts' or 'probes'. Each probe corresponds to one gene, and multiple probes can correspond to the same gene. A process known as hybridization is then used to indicate which probes exhibit abundance of the mRNA extracted from the replicate; mRNA is said to be hybridized to the microarray. The probes exhibiting higher levels of abundance indicate the genes which are more expressed than others. Repeating for each replicate in the given cohort yields many observations of mRNA abundance for each probe, and hence each gene.

The ability to easily observe the expression levels of thousands of genes allows researchers to perform a wide range of experiments. A common set of experiments conducted are longitudinal microarray time course experiments, where repeated measurements of mRNA are extracted from all the available replicates at a relatively small number of time-points. Through hybridization to microarrays, the expression level of all genes can be observed over time, yielding time courses for each probe/gene. These time courses capture the temporal evolution of the genes' expression levels, and are typically compared between different populations in order to identify genes which are differentially expressed. A review of such studies is provided in Section 9.2, in addition to a differential analysis of human immune cells in response to the *M.tuberculosis*

infection.

Gene expression can also be used as phenotypes in GWA studies. In this case, the gene expressions are typically observed in addition to SNPs for each replicate, and the interest is in detecting individual SNPs or collections of SNPs which are associated with a gene, or multiple genes. This is commonly referred to as eQTL mapping, and we describe these studies in more detail in Section 9.3. An eQTL pathway analysis of ovarian cancer is also presented.

9.2 Longitudinal Microarray Time Course Studies

In longitudinal microarray experiments, the temporal evolution of expression levels of thousands of genes are monitored in an attempt to understand the dynamic processes that regulate them (Storey *et al.*, 2005b). A common aim of such studies is to compare gene expression profiles observed in different populations or under different experimental conditions, and to identify genes whose temporal profiles differ significantly.

The data produced by such longitudinal studies present several challenges for statistical analysis. Tests developed for cross-sectional data, such as the t test and its many modifications, are inadequate because they are only able to detect differential expression at individual time-points and they ignore the temporal dependencies that are typical of the experimental data (Storey *et al.*, 2005b). Models from classical time series analysis are also limited in scope as the time courses are generally very short, sampled at irregularly spaced time-points, and often contain missing data (Tai and Speed, 2005; Bar-Joseph *et al.*, 2003).

Over the last few years these issues have led to an increasing interest in the application of FDA techniques to model the longitudinal time courses as smooth curves (see, for instance, Ramsay and Silverman (2006) and Wu and Zhang (2006)). This results in a set of inferred curves representing the time courses of each probe, and these are subjected to testing in order to assess if there is a significant difference between the expression curves of each population.

The problem of detecting differentially expressed genes can be framed as a test of the null hypothesis of equality between curves belonging to different populations or groups. This problem has been considered in the non-parametric statistics literature and in the microarray literature. A brief review of methods in each field are given in

Sections 9.2.1 and 9.2.2, respectively.

9.2.1 Existing Methods in the Non-Parametric Statistics Literature

Existing tests of equality between curves can be categorized by the way in which the corresponding statistics use the information provided by the sample curves. For instance, some approaches explicitly use all curves in the statistic, while others only use the mean curve estimated for each group. In addition, some methods treat the curves as infinite-dimensional objects while others use finite but high-dimensional representations of the curves by discretizing them over a large number of time-points, i.e., the curves are vectorized.

Vectorization approaches include functional ANOVA (FANOVA) (Ramsay and Silverman, 2006), high-dimensional ANOVA (HANOVA) (Fan and Lin, 1998), and the graphical SiZer approach of Park and Kang (2008). The FANOVA approach of Ramsay and Silverman (2006) is proposed as a method for detecting differences between curves at a specific time-point. It applies the classical univariate ANOVA F test to the values of the curves at the given time-point, and where many time-points are of interest, the tests are applied independently. This approach therefore yields a multiple-testing problem which must be addressed.

The HANOVA approach of Fan and Lin (1998) considers using many time-points simultaneously within the corresponding statistic, as opposed to single time-points independently. For this approach the mean curves of two groups are represented by high-dimensional longitudinal vectors, and the vector of differences between these is obtained. Dimensionality reduction is performed to differentiate noise from actual signals in this vector of differences, and this is achieved by applying a discrete Fourier transform. The result is a representation in the frequency domain where high frequency components are discarded as noise. An adaptive Neyman statistic is then proposed which sums the remaining differences in the frequency domain, and is shown to have an asymptotic distribution.

Park and Kang (2008) use a graphical device called SiZer to visualize the differences between any pair of mean curves by discretizing them at many time-points. This method is based on local polynomial kernel smoothing (see Wu and Zhang (2006)), which is defined by a bandwidth which dictates the level of smoothing. SiZer lets

the bandwidth take a range of values, and for each resulting resolution computes the difference between the mean curves and the corresponding confidence intervals of the differences at the discretized time-points. If the confidence intervals do not contain zero for a given time-point and resolution, the curves are said to be different at that point in time and resolution. This method offers the practitioner a detailed visual analysis of where the differences occur and in what resolution, utilizing as much information as possible from the local polynomial kernel smoother.

For these high-dimensional vectorial approaches, better inference results from considering more time-points as more information regarding the temporal behaviour of the curves is used. Testing procedures which respect the infinite dimensionality of the curves are therefore expected to have more power to detect differences between groups. Approaches include the functional F test of Shen and Faraway (2004), and the L_2-based approaches of Cuevas *et al.* (2004) and Zhang *et al.* (2010).

Shen and Faraway (2004) generalize the single time-point FANOVA procedure of Ramsay and Silverman (2006). In FANOVA, the classical between- and within-group variance quantities are computed at a given time-point, and hence are both a function of time. Shen and Faraway (2004) generalize these quantities by integrating over all time-points in the given time-range. A functional F statistic is then defined by the ratio of these generalized between- and within-group quantities, and for a large sample size this statistic is shown to have an approximate F distribution under the null.

In a two-group setting, Cuevas *et al.* (2004) propose a statistic proportional to the L_2 distance between the group mean curves. This statistic rejects the null hypothesis for large values, and the authors give details of an asymptotic distribution. A similar statistic is proposed by Zhang *et al.* (2010), which we denote the TN statistic. It is proportional to the L_2 distance between the mean curves, and significance is assessed via permutations.

Many other approaches also exist, such as the approach of Behseta and Kass (2005) which considers detecting differences between curves by using their respective basis coefficients. Other methods consider comparing the residuals obtained by modeling the original longitudinal vectors via local polynomial kernel smoothers - the residuals are the differences between the fitted values at the observation time-points and the observed values. Such methods include those described by Neumeyer and Dette (2003),

Pardo-Fernández *et al.* (2007) and Hall and Van Keilegom (2007).

9.2.2 Existing Methods in the Microarray Literature

A widely used method for detecting differentially expressed genes is EDGE, proposed by Storey *et al.* (2005b). The longitudinal time courses arising from each probe are modeled via a functional mixed-effects model comprising of a mean curve and an additive replicate-specific effect at the observation time-points. Under the null hypothesis of equality between population curves, all time courses are modeled as coming from one population. Natural cubic splines are used, and deviation of each replicate from this curve is captured via the replicate-specific effects. Under the alternative hypothesis the time courses of different populations are modeled via separate models.

Under the null and alternative models the residuals are computed and an F-type statistic is computed as the ratio of two components; the difference in sum of squared residuals under the null and alternative, and the sum of squared residuals under the alternative. Larger values of this statistic indicate that separate modeling of the time courses arising from each population yields a better fit of the observed data. Thus larger values provide evidence of differential expression, and significance is assessed via application of the bootstrap procedure.

A functional hierarchical empirical Bayes approach has been proposed by Hong and Li (2006) for a two-group setting. The replicate time courses in a given gene are modeled via B-spline basis expansion, where the basis coefficients are comprised of a gene-specific component and a replicate-specific component. The replicate-specific component is equal to a Bernoulli random variable multiplied by a constant representing the difference between the coefficients for curves of different populations. The random variable dictates whether or not this difference exists in modeling a particular replicate. For instance, if the variable is zero-valued for all replicates, then this suggests there is no difference between populations. The posterior probability of a difference existing is computed as one minus the probability that the Bernoulli random variable is zero given all the data. A hierarchical model is specified to compute this probability, and an EM algorithm is used to obtain the required parameter estimates for this model. The probability associated with each gene is used as the statistic of differential expression, and significant genes are identified by ranking these probabilities

and applying a threshold cutoff.

Bar-Joseph *et al.* (2003) also consider a two-group setting, and model the two population curves in each gene as noisy realizations of the other under the null hypothesis. By setting one of the estimated curves to be a reference curve, optimal fitted values of the second curve are sought such that the probability of the curve being a noisy realization of the reference curve is maximized. This uses the L_2 norm of the difference between the two curves, and once the optimal fitted values are found, the Euclidean distance between these and the actual fitted values of the second curve is obtained. The statistic of differential expression is then taken to be a value proportional to this Euclidean distance, which is shown have a Chi-squared distribution under the null.

Other methods include PACE (Liu and Yang, 2009), which uses a functional mixed-effects model with replicate-specific effects modeled in terms of functional principal components, and the functional Bayes approach of Angelini *et al.* (2007). A comprehensive review of methods can be found in Coffey and Hinde (2011).

9.2.3 Limitations of Existing Methods for Microarray Applications

The underlying assumption shared by all methods in the non-parametric and microarray literature is that, under the null hypothesis of equality between curves, the area between them is zero. This is equivalent to the L_2 distance being zero, showing that existing methods either explicitly or implicitly test a null hypothesis of zero L_2 distances between population curves. The rest of this section describes how this is restrictive for longitudinal microarray studies, and we argue that using other distances can be beneficial in longitudinal microarray time course analysis.

The L_2 distance is only concerned with vertical distances between points taken on each curve, so expression profiles may not necessarily exhibit shape-based differences in the time-varying patterns of mRNA abundance even when having large L_2 distances. We illustrate this in Figure 9.1, where we consider several types of difference between two simulated gene curves. The solid vertical lines indicate the vertical distances considered when computing the L_2 distance. For both the A1 and A2 comparisons (top row) the L_2 distances are equal (indicated by $d_L = 3.24$ in the plots), indicating that the area between the two curves (shaded region) is the same in both cases. However, there is a clearly visible difference in their respective shapes. Specifi-

cally, halfway through the time course, the expression levels in A1 increasingly diverge as time progresses, whereas the expression levels in A2 both plateau.

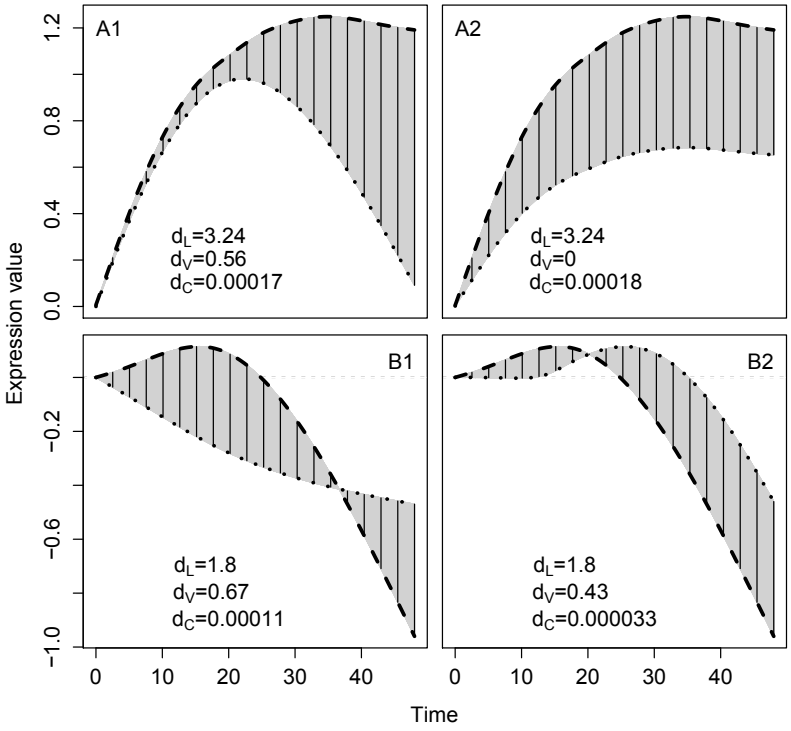

Figure 9.1: Four different comparisons between two simulated gene curves illustrating the effects of using the L_2 (d_L), Visual L_2 (d_V) and Curvature (d_C) distances. The curves in A1 and A2 have the same L_2 distances (represented by the shaded regions with vertical lines) despite clearly visible differences in the temporal gene expression patterns. Similarly, the curves in B1 and B2 have the same L_2 distances, although the curves in B1 have quite different time-varying behaviour while those in B2 have the same shape but are time-delayed. These shape-related differences are better captured by the Visual L_2 and Curvature distances.

For both the B1 and B2 comparisons (bottom row), we again see that the areas between the curves are the same ($d_L = 1.8$). Whereas the two curves in B2 have very similar shapes and have only undergone a time-shift, the curves in B1 have different time-varying patterns, resulting in shapes more different than just being time-shifted. Thus the L_2 distance is unable to identify similar temporal profiles that only differ due to delays on the time scale. Such time-shifts can be representative of expression responses which may be slower in one group than another, due to a time-lag in their transcription control, for example (Qian *et al.*, 2001).

These examples demonstrate that the L_2 distance may be unable to capture clearly visible differences in the shape of the expression profiles. Hence existing tests that focus on this distance are expected to have very little statistical power in detecting certain shape-related differences, as demonstrated by our simulations in Section 5.5.2. The Visual L_2 and Curvature distances can be deployed to capture such shape-related differences.

The Visual L_2 distance takes into account both vertical and horizontal distances between points on the curves once they have been made scale-invariant. In the A1 comparison in Figure 9.1, the curves have a larger Visual L_2 distance (indicated by $d_V = 0.56$) than in A2 ($d_V = 0$), since once the difference in amplitude has been removed the two curves have exactly the same shape. Also, the Visual L_2 distance is smaller for the curves represented in the B2 comparison than those in B1. This agrees with a visual exploration of the curves which clearly shows that the two temporal profiles are time-shifted, but their shapes are otherwise very similar.

The Curvature distance, on the other hand, quantifies the difference in smoothness of the curves, and unlike the L_2 and Visual L_2 distances, will yield a zero value if the curves are perfect reflections of each other in the time axis (for example, one having a peak and the other having a trough). Such inverted temporal profiles can indicate inhibitory relationships between the populations (Shi *et al.*, 2007). In this case using the Curvature distance will show that the gene curves are considered similar. In Figure 9.1, the Curvature distance is smaller in B2 than in B1 (indicated by comparing $d_C = 0.000033$ and $d_C = 0.00011$), showing that similarity in time-shifted curves can be detected.

9.2.4 Differential Analysis of Human Immune Cell *M.tuberculosis* Time Course Data with the DBF Test

Processes that may contribute to controlling *M.tuberculosis* infection may be highlighted by comparing the expression profiles of human phagocytic immune cells - macrophage and dendritic cells (denoted M/DCs) - that differ in their ability to limit bacterial growth (Tailleux *et al.*, 2008). In this section we describe a differential analysis performed on a sample of human immune cells which have been infected with *M.tuberculosis*. The analysis was published by us in Minas *et al.* (2011), and was

conducted using the DBF test with permutations.

Time course measurements of gene expressions were recorded at 0, 4, 18 and 48 hours after infection using Affymetrix U133A high-density oligonucleotide arrays. The observation of each type of cell at each time-point was repeated with human immune cells isolated from 9 healthy donors, yielding $N = 18$ samples. After pre-processing and removal of missing data, $10,995$ probes remained for the differential analysis (Tailleux *et al.*, 2008), each mapped to a particular gene (not necessarily uniquely). Temporal profiles of all genes were smoothed using cubic smoothing splines, after which they were normalized at baseline so that any differences detected were relative to pre-infection state. The DBF test was applied to each probe with the L_2, Visual L_2 and Curvature distances and using $24,310$ Monte Carlo permutations.

This results in the simultaneous estimation of $10,995$ p-values, forming a multiple-testing problem. We use the approach of Storey and Tibshirani (2003) to control the false discovery rate at 1%. This yields $3,201$ probes exhibiting a significant difference between the M/DCs in response to *M.tuberculosis* infection. The Venn diagram in Figure 9.2 presents a global view of how the distance measures identified these probes. While there is some overlap between the probes identified by different distance measures, as expected due to curve patterns exhibiting differences of different types (such as large areas and diverging over time), many probes are also identified uniquely by each distance measure.

The significant probes were then grouped into predefined gene ontology (GO) classifications of genes, resulting in a set of functional categories which were significantly enriched with each distance measure. In doing so, an overview of the pathways that are likely to be reprogrammed in dendritic cells compared to macrophages after infection is accessible. GO terms for membrane invagination (GO:0010324) and endocytosis (GO:0006897, GO:0016193, GO:0016196), the process whereby phagocytic cells (such as M/DCs) engulf foreign bodies (such as *M.tuberculosis* bacilli), significantly overlapped with genes recognised using the L_2 distance. Additionally, genes associated with the endosome (GO:0005768) and late endosome (GO:0005770), the membrane structures containing foreign bodies that are formed during endocytosis, were significantly enriched only using the Visual L_2 distance. Thus, the Visual L_2 measure identified subtle changes in gene expression between the cell types that did not rely

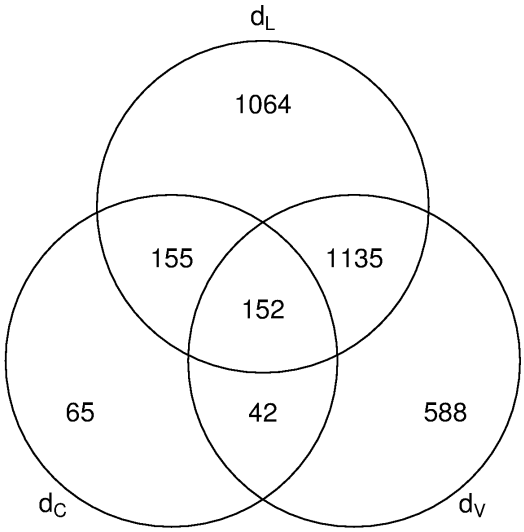

Figure 9.2: Venn diagram showing the overlap of significant probes identified by the L_2 (d_L), Visual L_2 (d_V) and Curvature (d_C) distance measures. While many probes are identified by the same distances, many probes are also identified uniquely by each distance measure.

on large differences in amplitude across time-points.

The biological significance of the results were also considered by looking at RAB GTPases, which are a subset of genes involved in intracellular trafficking. They are a family of small guanosine triphosphatases found on the surface of intracellular membranes that play integral roles in regulating their movement around the cell (Brumell and Scidmore, 2007). The retention of RAB5 and the failure to recruit RAB7 has been used to characterize the stalled development of the *M.tuberculosis*-containing phagosome (a phagosome is a compartment surrounding the given cell in which foreign bodies are digested and killed) (Brumell and Scidmore, 2007). Genes encoding RAB7A and RAB7L1 were identified to be differentially regulated between M/DCs using multiple measures (RAB7A with all measures, and RAB7L1 with the L_2 and Visual L_2 measures). Figure 9.3 displays the mean M/DC expression profiles for a selection of genes, of which RAB7A is the first. We see that there is a large difference in area between the two curves (L_2 distance), large scale-invariant differences in shape (Visual L_2 distance), and the macrophage curve changes shape much faster than the dendritic curve (Curvature distance).

RAB5B and RAB5C were only revealed to be divergently expressed using the L_2 distance, as was also the case for RAB22A (Figure 9.3 (b)) which has been implicated

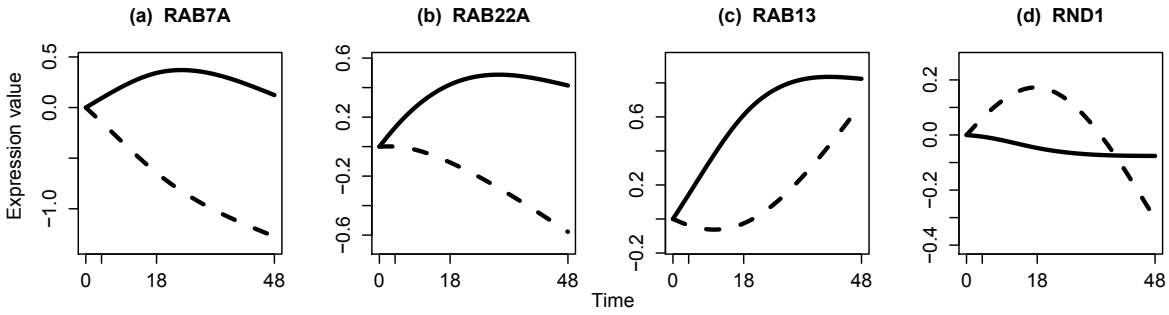

Figure 9.3: Mean macrophage (solid) and dendritic (dashed) expression profiles for genes identified by the DBF test with different distances. (a) RAB7A, identified with the L_2, Visual L_2 and Curvature distances. (b) RAB22A, identified with the L_2 distance. (c) RAB13, identified with the Visual L_2 distance. (d) RND1, identified with the Curvature distance.

in the reprogramming of *M.tuberculosis*-phagosome trafficking (Brumell and Scidmore, 2007). The differential expression of RAB13 (Figure 9.3 (c)) and RAB21 were only detected when considering the differences in the scale-invariant expression profiles as determined by the Visual L_2 distance. For these genes, the divergent pattern of gene expression over time indicates that distinct processes are impacting upon intracellular trafficking in macrophages compared to dendritic cells after mycobacterial infection. This therefore highlights pathways of interest for further investigation. RND1 (Figure 9.3 (d)) and RND3, Rho family GTPases, were only identified when considering the Curvature distance between the expression profiles. Here, large differences in the speed with which the profiles changed over time were captured, whereas the direct time-point comparisons of the L_2 distances did not detect significant differences in amplitude.

9.3 eQTL Mapping Studies

eQTL mapping refers to GWA studies in which gene expression phenotypes are used. Gene expression represents the phenotype 'most immediately connected to DNA sequence variation' (Rockman and Kruglyak, 2006), since gene expression is directly regulated by the DNA sequence (see, for instance, Gibson and Muse (2004)). Gene expression is therefore the bridge between an individuals genotype and phenotypes such as case-control status of disease or imaging-derived quantitative traits.

An eQTL is a SNP that influences the expression of a given gene, or set of genes.

The gene expressions are typically observed using microarrays, and hence thousands of probes can be used as quantitative traits. Due to the large number of genes represented by these traits which combine together in different ways to perform specific biological processes, eQTL mapping can yield insights into the genetic effects on the biological mechanisms underlying susceptibility to complex disease (Cookson *et al.*, 2009).

9.3.1 A Brief Review of Existing Methods

Many traditional approaches to eQTL mapping, such as those described by Stranger *et al.* (2005), DeCook *et al.* (2006) and Quigley *et al.* (2011), adhere to the 'single-SNP, single-trait' paradigm of GWA analysis. Association is inferred for each SNP-probe combination individually by using linear regression models in which the SNP is the independent variable and the probe is the dependent variable. These methods suffer from two main limitations. Firstly, genes, and therefore their respective expression traits, are known to function together in networks or pathways (Wessel *et al.*, 2007; Li *et al.*, 2010). Altered expression levels of a single gene can therefore induce altered expression in many of the genes within the same pathway, and single-trait analyses are unable to capture such combined actions. Secondly, considering SNPs individually ignores joint effects of multiple SNPs, and in particular, interactions between them cannot be captured (Stranger *et al.*, 2005).

Ad-hoc approaches have been adopted to account for multiple genes working in tandem. For instance, having performed the traditional single-SNP, single-trait analysis, clustering approaches have been used to group together probes/genes which appear to be influenced by the same SNP (Morley *et al.*, 2004; Quigley *et al.*, 2011). These approaches use either Pearson's correlation or the Euclidean distance applied to sample gene expressions as the notion of similarity/dissimilarity in the application of clustering. Thus the consideration of multiple genes is not considered in the actual GWA analysis.

This limitation has been highlighted by Wessel *et al.* (2007) and Li *et al.* (2010). In both cases, multiple genes are grouped together into pathways, and pathway analyses are described using the distance-based pseudo F test. The Euclidean distance matrix is obtained for sampled expressions of all genes in the pathway, and variation in this distance matrix is modeled in terms of a single SNP or multiple SNPs. Permutations

are used to assess significance. For the case of a single SNP, Li *et al.* (2010) also propose using traditional MANOVA techniques when the number of samples exceeds the number of probes. Here the sampled gene expressions are treated as vectors coming from different groups defined by the unique genotypes.

The effect of multiple SNPs on individual traits has also received much attention; so-called 'multiple-SNP, single-trait' GWA analyses are described by Storey *et al.* (2005a), McClurg *et al.* (2006) and Wu *et al.* (2008). In Storey *et al.* (2005a), a linear model is used where the variation in a single trait is explained by the additive effects of a pair of SNPs in addition to an interaction effect between them. The pairs of SNPs are selected across the genome using a stepwise procedure designed to be computationally efficient; using all possible pairs of SNPs yields a number of tests in the millions, even for a relatively small number of SNPs. In McClurg *et al.* (2006) and Wu *et al.* (2008), a sliding window approach is adopted to select multiple SNPs for analysis. Each window is comprised of 3 contiguous SNPs, and an ANOVA model is used where the observations of each trait are grouped based on the unique observations across the SNPs.

The distance-based pseudo F test with permutations has also been used to model the variation in a SNP distance matrix in terms of gene expression (Wessel and Schork, 2006). The gene expressions are treated as explanatory variables in the pseudo F regression framework, which may be deemed unintuitive since gene expression is influenced by DNA sequence variation rather than vice-versa. However, it indicates the utility of SNP distances in eQTL mapping, and this is in addition to separate studies which apply distance measures to gene expression.

It is clear that approaches are tending towards the paradigm of 'multiple-SNP, multiple-trait' GWA analysis. It is understood that SNPs within a SNP set can have joint effects and may interact, and it is also understood that many probes/genes can orchestrate a combined effect. SNPs and probes can be grouped together either as sets arising from a sliding window or by their existence in the same pathway. Throughout the literature there is evidence that applying distances to the samples of multiple SNPs and multiple probes, albeit in separate analyses, can be beneficial in yielding interesting biological insights.

To our knowledge, no such studies have been performed where distances are simul-

taneously applied to both data types. Where they are applied in the literature, it is clear that they offer a way to capture the variation in the given dataset without explicitly modeling complex behaviour (such as interactions of SNPs, or the combined effect of probes/genes in a pathway). Applying distances to both data types will therefore allow such complex behaviour to be accounted for when seeking associations.

9.3.2 An eQTL Pathway Analysis of Ovarian Cancer with the GRV Test

A major area of research in understanding ovarian cancer is determining the biological mechanisms underpinning the development of malignant and chemo-resistant cancer cells (Chapman-Rothe *et al.*, 2012). In this section we describe an eQTL pathway analysis of ovarian cancer using the GRV test. This is the first known eQTL pathway analysis in which distances are applied to both SNP and gene expression data.

The ovarian cancer data used for this analysis is described in Chapman-Rothe *et al.* (2012) and was obtained from the Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/). The dataset consists of 494 tumour samples, each genotyped at $906,600$ SNPs across chromosomes 1 to 22 of the genome, and mRNA samples have been obtained for each at $22,277$ probes across the genome (after pre-processing). The SNPs and gene expression probes were independently mapped to genes using annotation information obtained from the BioMart database (http://www.ensembl.org/), then separately grouped into $4,119$ pathways taken from the Consensus Pathway Database (http://cpdb.molgen.mpg.de/). The interest is in detecting the pathways for which there is an association between SNPs and gene expression traits (probes).

On using the IBS, Sokal and Sneath and Rogers and Tanimoto I distances for the SNP data, and the Euclidean and Pearson's correlation distances for the gene expression data, we use the GRV test in two ways. An overall pathway analysis was first conducted, followed by individual pathway analyses. These separate analyses are described in turn below.

For the overall pathway analysis the objective was to obtain an overall view of association between all SNPs and all probes. For the SNP data the three genetic distance matrices were obtained, and for the gene expression data the two gene expression distance matrices were obtained. On obtaining the corresponding centered

inner product matrices, the GRV test was applied to the six combinations of SNP and gene expression distances. Each test was significant, with p-values being below 10^{-14}. We plot the scatter plots of the normalized centered inner product matrices for each combination in Figure 9.4, showing the associations detected. The superimposed black lines are the regression lines whose gradients equal the respective GRV statistic values. All combinations yield association between the SNPs and gene expressions, suggesting that many of the pathways will also exhibit association.

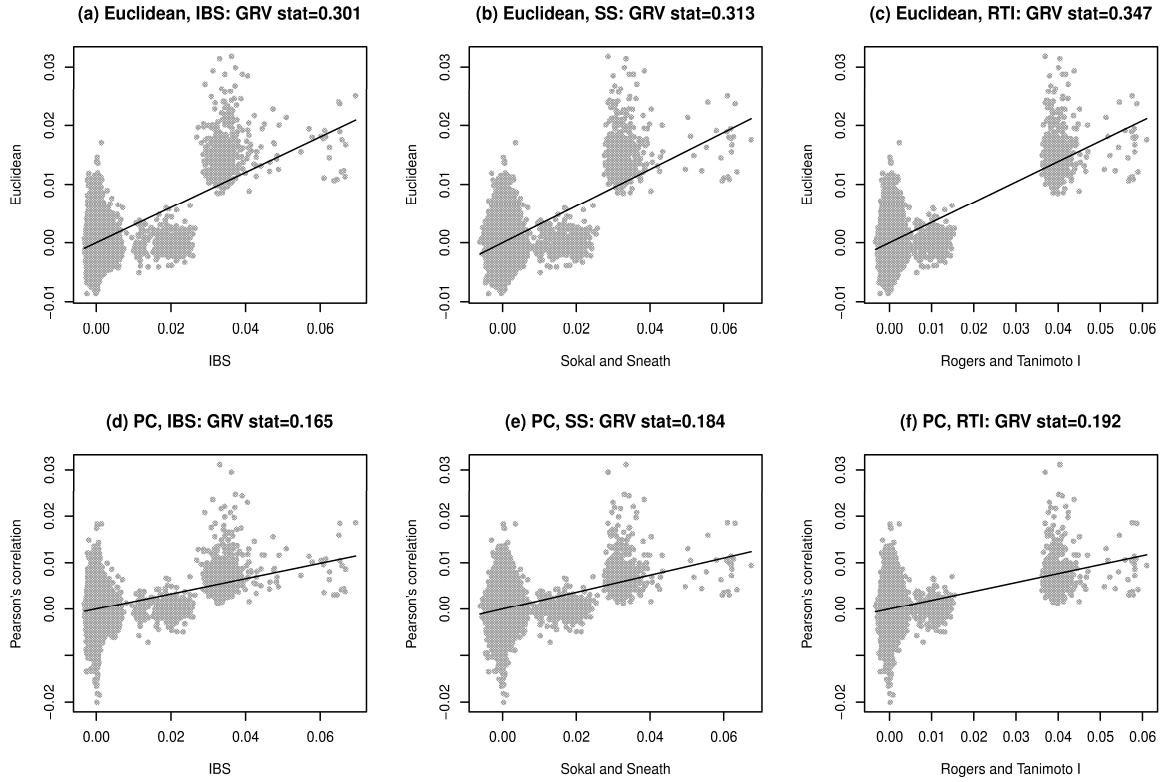

Figure 9.4: Scatter plots of the elements of the normalized centered inner product matrix arising from the gene expression distance matrix against the elements arising from the SNP distance matrix (gray points). The linear regression lines (black lines) indicate the strength of correlation between the values. The gradient of these lines equals the respective GRV statistic values. (a)-(c) Euclidean distances applied to the sampled gene expressions with the IBS, Sokal and Sneath (SS) and Rogers and Tanimoto I (RTI) distances applied to the observed SNPs. (d)-(f) Pearson's correlation (PC) distance applied to the sampled gene expressions with the IBS, Sokal and Sneath and Rogers and Tanimoto I distances applied to the observed SNPs.

The GRV test was then applied to each pathway individually, using all six combinations of SNP and gene expression distance measures. This resulted in $4,119$ p-values for each of the six combinations of distances. On applying the Benjamini-Hochberg

multiple-testing correction and controlling the false discovery rate at 0.1%, many pathways were identified for each combination of distance measures; see Table 9.1. Of these pathways, 575 were identified by all combinations, and each combination uniquely identified a much smaller subset of pathways (shown in brackets in Table 9.1). This demonstrates that the GRV approach applied with different distance combinations captures different types of association.

Table 9.1: Number of pathways identified for each SNP-gene expression distance combination on applying the GRV test to the ovarian cancer data. The number in the brackets refers to the number of pathways which were uniquely identified by the given combination of SNP and gene expression distance.

Gene expression distance \ SNP distance	IBS	Sokal and Sneath	Rogers and Tanimoto I
Euclidean	1015 (1)	1976 (19)	1982 (29)
Pearson's correlation	662 (0)	1532 (2)	1550 (6)

Many of the pathways implicating genes or biological processes as being associated with ovarian cancer identified via the GRV test have been previously identified in other studies. Of the 575 pathways identified by all measures, for instance, the VEGF signaling pathway is well-known (Trinh *et al.*, 2009; Dhillon *et al.*, 2007). It promotes ovarian cancer progression, and has been the target of successful chemotherapeutic agents such as Bevacizumab (Burger *et al.*, 2007). The MAPK signaling pathway is also well-known (Trinh *et al.*, 2009; Dhillon *et al.*, 2007), in addition to the JAK STAT pathway (Liongue *et al.*, 2012), which is of clinical importance in ovarian cancer; STAT1 has been shown to control chemotherapy resistance of ovarian cancer cells (Stronach *et al.*, 2011). Other well-known pathways identified include gap-filling DNA repair synthesis and ligation in GG-NER and TC-NER (Shuck *et al.*, 2008). The uniquely identified pathways for the distance combinations have also been previously identified. For instance, for the Euclidean and IBS distance a pathway involving the ErBb2 gene was identified, which is a well-known gene associated with cancer (Yu and Hung, 2000). For the Euclidean and Sokal and Sneath combination, BMP signalling

and regulation has been identified (Shepherd *et al.*, 2010). The well-known FGFR4 gene (French *et al.*, 2012) was identified with the Euclidean and Rogers and Tanimoto I distances. On using Pearson's correlation distance with the Sokal and Sneath and Rogers and Tanimoto I distances, the IL-9 signalling pathway and point mutants of the FGFR1 gene have been been previously identified (Hodge *et al.*, 2005; Rand *et al.*, 2005). These results therefore suggest the validity of the GRV approach in eQTL pathway analysis.

9.4 Summary

The level of mRNA abundance for a given gene represents its expression level, and the expression level of thousands of genes can be simultaneously obtained using microarrays. This has allowed researchers to conduct a range of exploratory studies, of which two have been discussed in this chapter; longitudinal microarray time course studies and eQTL mapping studies.

Longitudinal microarray time course studies were discussed in Section 9.2. The interest is in detecting genes exhibiting differential expression between populations or treatments, and this can be framed as a test of the null hypothesis of equality between curves for each gene. Traditional approaches to this problem have been reviewed in Section 9.2.1, detailing approaches used in non-parametric statistics, and in Section 9.2.2, detailing methods used in the microarray literature. The inherent limitation that all methods only detect L_2 distances between curves is highlighted in Section 9.2.3, where we demonstrate that the L_2 distance can miss shape-related differences which can be captured by other distances such as the Visual L_2 and Curvature distances.

This observation is also supported by the differential analysis of the human immune cell *M.tuberculosis* data presented in Section 9.2.4. The results demonstrate that the deployment of shape-based distances in the DBF test can lead to meaningful biological insights. Such distances may be desired in scenarios where large changes in amplitude of gene expression, as captured by the L_2 distance, are not a prerequisite, or where the differential actions of expression profiles are of interest.

eQTL mapping studies were described in Section 9.3. In such studies the aim is to identify eQTLs, or SNPs, which are associated with gene expression; essentially, gene expression phenotypes are used in GWA studies. Traditional approaches are

described in Section 9.3.1, where it has been highlighted that SNPs can exhibit joint effects on the expression level of a single probe/gene and multiple probes/genes. In the literature distances have been adopted for gene expression and SNPs separately, demonstrating the usefulness of distances in eQTL mapping. For the paradigm of identifying association between multiple SNPs and multiple probes/genes, however, no studies have been performed where distances are applied to both SNPs and gene expressions simultaneously. We have presented the first study of this kind in Section 9.3.2, using the GRV test to identify pathways for which SNPs and gene expression are associated in ovarian cancer. The findings overlap with previous studies, suggesting the validity of the fully distance-based GRV approach in eQTL mapping.

Chapter 10

Conclusions and Further Work

In this thesis we have considered three statistical problems arising in the bioinformatics literature, and have focused on the distance-based setting for each. These problems and existing approaches have been reviewed in Chapters 2, 3 and 4, where limitations of distance-based methods have been highlighted. A recurring limitation is that in application to real datasets, computationally expensive permutation testing procedures are used. Such procedures yield p-value estimates which are plagued by sampling errors introduced by the relatively small number of permutations typically applied. The overall contribution of this thesis is the proposal of approximate null distributions for tests of these problems which allow computationally cheap estimation of p-values for a variety of distance measures and data types. We summarize marginal contributions for each problem below.

For the problems of detecting differences between groups and detecting association between variables, we have proposed new statistics, the DBF and GRV statistics (Chapters 5 and 7, respectively), with corresponding permutation procedures for estimating p-values. For each we have proposed approximations to the permutation distribution which would be obtained by enumerating all permutations, and have demonstrated the applicability of the resulting approximate null distributions for a range of distances and data types. Furthermore, competitiveness with existing approaches for the respective problems have been demonstrated. Finally, we have demonstrated that the proposed distributions facilitate the effective implementation of each test in bioinformatics applications. In Chapter 8 both tests have been applied to GWA studies of Alzheimer's disease, and in Chapter 9 the GRV test has been applied to an eQTL

mapping study of ovarian cancer.

For the problem of detecting predictive relationships between variables, we have reviewed the distance-based pseudo F test in Chapter 3, which is routinely used in bioinformatics applications. In Chapter 6 we have proposed an approximation to the permutation distribution which would be obtained by enumerating all permutations. This required the analytical derivation of the moments of this permutation distribution, and simulations were used to show the resulting approximate null distribution is applicable for a range of distance measures and data types. In Chapter 8 the pseudo F test with the approximate null distribution has been applied to a GWA study of Alzheimer's disease.

This thesis has provided a snapshot of the way distance-based approaches can be applied effectively in bioinformatics. However, the full potential of such methods has yet to be realized, as each specific biological problem brings with it new challenges requiring specialized uses of the distance-based testing procedures. To illustrate this, we provide some suggestions for further work which require extending the methods described:

(i) Distance metric learning within the DBF, pseudo F and GRV tests: Distance metric learning is an area of machine learning concerned with finding an optimal distance measure for a given problem (see, for instance, Xiang *et al.* (2008) and Ying and Li (2012)), such as clustering microarray gene expression data (Xiong and Chen, 2006).

The problem is typically considered for N Q-dimensional vector-valued observations $\{\boldsymbol{y}_i\}_{i=1}^N$, and the interest is in finding the symmetric $Q \times Q$ weighting matrix $\boldsymbol{A}$ such that using the Mahalanobis-like distance measure

$$d(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sqrt{(\boldsymbol{y}_i - \boldsymbol{y}_j)^T \boldsymbol{A} (\boldsymbol{y}_i - \boldsymbol{y}_j)}$$

maximizes a given objective function. Seeking the optimal weights in $\boldsymbol{A}$ provides a method of incorporating relationships between variables which is data-driven.

For vector-valued observations, an objective function to be maximized for clustering can be formulated from the DBF statistic by using the above distance, and numerically solving for the optimal weighting matrix. This involves using

optimization methods such as those described in Kiers (2002), Boyd and Vandenberghe (2004) and Xiang *et al.* (2008). The testing procedure can then be applied to assess the significance of the observed statistic obtained on finding the optimal weighting matrix. Similarly, objective functions can be derived from the pseudo F and GRV statistics to find optimal distances which maximize predictive effects, and association, respectively.

(ii) Accounting for the effects of population stratification in case-control GWA studies within the pseudo F testing framework: In case-control GWA studies, population stratification refers to the confounding effect where differences in allele frequencies of SNPs observed on case and control individuals are due to ethnicity, for instance, instead of association with disease risk (Thomas and Witte, 2002; Price *et al.*, 2010). This leads to increased type I errors in GWA studies, since SNPs can be identified as causative of disease when in fact they are due to underlying structures within the cohort (Li *et al.*, 2009).

Li *et al.* (2009) propose an adjusted pseudo F test to identify causative SNPs while accounting for population stratification effects, given the genetic distance matrix $\boldsymbol{\Delta}_{\mathcal{Y}}$ and predictor variables including possible confounding variables and case-control status. Let the M predictor variables be partitioned such that $\mathcal{X} = (\mathcal{X}^1; \mathcal{X}^2) = (\mathcal{X}_1, \ldots, \mathcal{X}_{M_1}; \mathcal{X}_{M_1+1}, \ldots, \mathcal{X}_M)$ with $M_1 + M_2 = M$, where the interest is in testing for no effect of $\mathcal{X}^2$ on response variable $\mathcal{Y}$ (with distances in $\boldsymbol{\Delta}_{\mathcal{Y}}$), while adjusting for the effects of $\mathcal{X}^1$. In Li *et al.* (2009), $\mathcal{X}^1$ is vector-valued and comprised of the possible confounding variables, such as self-declared ethnicity etc., and $\mathcal{X}^2$ is the scalar-valued case-control status variable.

Given N observations of $\mathcal{X}$, the predictor matrix is given by $\boldsymbol{X} = (\boldsymbol{X}_1; \boldsymbol{X}_2)$ where $\boldsymbol{X}_1 \in \mathbb{R}^{N \times M_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{N \times M_2}$. The following regression model is then defined,

$$\tilde{\boldsymbol{Y}} = \boldsymbol{X}_1 \boldsymbol{B}_1 + \boldsymbol{X}_2 \boldsymbol{B}_2 + \boldsymbol{E},$$

where $\tilde{\boldsymbol{Y}}$ is the $N \times N$ principal coordinate matrix arising from a principal coordinate analysis of $\boldsymbol{\Delta}_{\mathcal{Y}}$, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are the $M_1 \times N$ and $M_2 \times N$ matrices of regression coefficients, respectively, and $\boldsymbol{E}$ is the $N \times N$ matrix of errors in the

model. The null hypothesis to be tested is expressed as

$$H_0 : \boldsymbol{B}_2 = \mathbf{0},$$

and the adjusted pseudo F statistic used to test this is defined as

$$F(\mathcal{X}^2|\mathcal{X}^1) = \frac{\text{tr}\left(\left(\boldsymbol{H} - \boldsymbol{H}_1\right)\boldsymbol{G}_{\mathcal{Y}}\right)}{\text{tr}\left(\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\boldsymbol{G}_{\mathcal{Y}}\right)},$$

where $\boldsymbol{G}_{\mathcal{Y}} = -\boldsymbol{C}\boldsymbol{\Delta}_{\mathcal{Y}}^2\boldsymbol{C}/2$, $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$, and $\boldsymbol{H}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T$. Larger values of this statistic provide evidence against the null since the term $\text{tr}\left(\left(\boldsymbol{H} - \boldsymbol{H}_1\right)\boldsymbol{G}_{\mathcal{Y}}\right)$ in the numerator quantifies the variability in $\boldsymbol{G}_{\mathcal{Y}}$ explained after accounting for the effects of $\boldsymbol{X}_1$.

Given an observed value of the test statistic, $\hat{F}(\mathcal{X}^2|\mathcal{X}^1)$, inference is performed using permutations. For N_π Monte Carlo permutations $\pi \in \Pi$, the set $\{\hat{F}_\pi(\mathcal{X}^2|\mathcal{X}^1)\}_{\pi\in\Pi}$ is generated where

$$\begin{aligned}\hat{F}_\pi(\mathcal{X}^2|\mathcal{X}^1) &= \frac{\text{tr}\left(\left(\boldsymbol{H} - \boldsymbol{H}_1\right)\boldsymbol{G}_{\mathcal{Y},\pi}\right)}{\text{tr}\left(\left(\boldsymbol{I}_N - \boldsymbol{H}\right)\boldsymbol{G}_{\mathcal{Y},\pi}\right)} \\ &= \frac{\hat{H}_\pi - \hat{H}_{1,\pi}}{\text{tr}\left(\boldsymbol{G}_{\mathcal{Y}}\right) - \hat{H}_\pi},\end{aligned}$$

where $\hat{H}_\pi = \text{tr}\left(\boldsymbol{H}\boldsymbol{G}_{\mathcal{Y},\pi}\right)$ and $\hat{H}_{1,\pi} = \text{tr}\left(\boldsymbol{H}_1\boldsymbol{G}_{\mathcal{Y},\pi}\right)$ are the permuted values of $H = \text{tr}\left(\boldsymbol{H}\boldsymbol{G}_{\mathcal{Y}}\right)$ and $H_1 = \text{tr}\left(\boldsymbol{H}_1\boldsymbol{G}_{\mathcal{Y}}\right)$, respectively. The p-value is then computed as the proportion of the N_π permuted statistics greater than or equal to the observed $\hat{F}_\pi(\mathcal{X}^2|\mathcal{X}^1)$, i.e.,

$$\frac{\#\left(\hat{F}_\pi(\mathcal{X}^2|\mathcal{X}^1) \geq \hat{F}(\mathcal{X}^2|\mathcal{X}^1)\right)}{N_\pi}.$$

Li *et al.* (2009) use $O(10^3)$ permutations for $N > 2000$ samples to estimate these p-values, which is extremely low.

The methodology derived in Chapter 6 can be applied to approximate the null distribution of $F(\mathcal{X}^2|\mathcal{X}^1)$, such that p-values can be estimated without permutations. The approximate null distributions of H and H_1 can be obtained separately

by applying the proposed results for the permutational moments and using the Pearson type III approximation. The problem then consists of combining these to approximate the null distribution of $F(\mathcal{X}^2|\mathcal{X}^1)$, and this is encompassed within the vast problem of obtaining the distribution of algebraic manipulations of random variables (Springer, 1979).

(iii) Distance-based variable selection within the pseudo F regression framework: A common problem in linear regression consists of identifying the subset of predictor variables that 'best' explains variation exhibited by the response variable. In candidate-phenotype GWA studies, such an approach can highlight the subset of causative SNPs of a given set of SNPs which best explains the variation in the quantitative phenotype (see, for instance, Vounou *et al.* (2010)).

A traditional technique used where the response variable is scalar-valued is the iterative 'stepwise procedure' (see, for instance, Rencher (2002)), which uses the classical F statistic in addition to the partial F statistic (the F statistic adjusted to account for the effects of a given predictor variable or set of predictor variables). As a first attempt at a distance-based variable selection method, we can directly generalize the stepwise regression approach to work within the pseudo F regression framework, since the pseudo F statistic is a generalization of the classical F statistic and the adjusted F statistic is a generalization of the partial F statistic.

Appendix A

Proof Regarding Hat Matrix in Regression Setting

Consider the $N \times M$ predictor matrix $\boldsymbol{X}$ with $N = M$, and assume it is of full rank, i.e., $\text{rank}(\boldsymbol{X}) = N$. We prove that the corresponding hat matrix $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ equals the $N \times N$ identity matrix $\boldsymbol{I}_N$.

To proceed we require the following well-known properties of $\boldsymbol{H} = \{h_{ij}\}_{i,j=1}^N$:

(i) $\text{rank}(\boldsymbol{H}) = \text{rank}(\boldsymbol{X})$

(ii) the N eigenvalues of $\boldsymbol{H}$, $\{\lambda_i\}_{i=1}^N$, are either 0 or 1

(iii) $\text{rank}(\boldsymbol{H})$ equals the number of non-zero eigenvalues of $\boldsymbol{H}$

(iv) $0 \leq h_{ii} \leq 1$ (follows from the fact that $\boldsymbol{H}$ is idempotent)

(v) $h_{ij}^2 \leq h_{ii}(1 - h_{ii})$ for $i \neq j$

(Hoaglin and Welsch, 1978; Dodge and Hadi, 1999). From (i) we have that $\text{rank}(\boldsymbol{H}) = N$, and using this with (iii) we have that all $\{\lambda_i\}_{i=1}^N$ are non-zero. Therefore, from (ii) we have that $\lambda_i = 1$ for all $i = 1, \ldots, N$. Now, from matrix theory we have that $\text{tr}\,(\boldsymbol{H}) = \sum_{i=1}\lambda_i$, so that $\text{tr}\,(\boldsymbol{H}) = N$. But we also have that $\text{tr}\,(\boldsymbol{H}) = \sum_{i=1}^N h_{ii}$, so that $\sum_{i=1}^N h_{ii} = N$. Using this with (iv), we observe that $h_{ii} = 1$ for all $i = 1, \ldots, N$. Substituting $h_{ii} = 1$ into (v) yields $h_{ij} = 0$ for all $j \neq i$. Therefore $\boldsymbol{H} = \boldsymbol{I}_N$, as required.

Appendix B

Examples of Distance Measures

Here we present a selection of distance measures which can be applied to measure the dissimilarities between objects of different types.

B.1 Distance Measures for Vectors

Assume two P-dimensional real-valued vectors $\boldsymbol{x} = (x_1, \ldots, x_P)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_P)^T$. Many measures exist (see, for example, Pekalska and Duin (2005)), of which a few are provided in Table B.1, along with their ranges and properties, i.e., whether they are metric or semi-metric. These include the Euclidean, Manhattan, Maximum, Bray-Curtis, Mahalanobis, Pearson's correlation and the Cosine angle distances.

Each distance captures a different aspect of dissimilarity between vector-valued objects. For example, the Euclidean distance provides the length of the line segment connecting $\boldsymbol{x}$ to $\boldsymbol{y}$ in P-dimensional Euclidean space, which is the shortest distance between the two points. The Manhattan distance on the other hand, considers the length between the points with respect to only their projections on the P orthogonal axes.

The Bray-Curtis measure is of interest in ecological applications (Legendre and Legendre, 1998), where it was originally proposed for data comprised of integer-valued counts. It is a weighted Manhattan distance which provides a measure of the proportion of difference between the values of two vectors across all P values.

A greater Pearson's correlation distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ indicates a weaker positive linear relationship between the vectors. This is also highlighted by the Cosine

Table B.1: Commonly encountered distance measures for vector-valued objects. The (M) or (SM) by each distance name indicates whether it is metric or semi-metric.

Distance	Notation	Definition	Range
Euclidean (M)	$d_E(\boldsymbol{x}, \boldsymbol{y})$	$\sqrt{(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})}$	$[0, \infty)$
Manhattan (SM)	$d_{MAN}(\boldsymbol{x}, \boldsymbol{y})$	$\sum_{p=1}^{P} \lvert x_p - y_p \rvert$	$[0, \infty)$
Maximum (SM)	$d_{MAX}(\boldsymbol{x}, \boldsymbol{y})$	$\max_p \{\lvert x_p - y_p \rvert\}$	$[0, \infty)$
Bray-Curtis (SM)	$d_{BC}(\boldsymbol{x}, \boldsymbol{y})$	$\frac{\sum_{p=1}^{P} \lvert x_p - y_p \rvert}{\sum_{p=1}^{P} (x_p + y_p)}$	$[0, \infty)$
Mahalanobis (SM)	$d_{MAH}(\boldsymbol{x}, \boldsymbol{y})$	$\sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{S}^{-1} (\boldsymbol{x}-\boldsymbol{y})}$, $\boldsymbol{S}$ a $P \times P$ covariance matrix, $P < N$	$[0, \infty)$
Pearson's correlation (SM)	$d_{PC}(\boldsymbol{x}, \boldsymbol{y})$	$1 - \frac{\sum_{p=1}^{P} (x_p - \bar{x})(y_p - \bar{y})}{\sum_{p=1}^{P} (x_p - \bar{x})^2 \sum_{p=1}^{P} (y_p - \bar{y})^2}$, $\bar{x} = \frac{1}{P}\sum_{p=1}^{P} x_p, \quad \bar{y} = \frac{1}{P}\sum_{p=1}^{P} y_p$	$[0, 2]$
Cosine angle (SM)	$d_{PC}(\boldsymbol{x}, \boldsymbol{y})$	$1 - \frac{\boldsymbol{x}^T \boldsymbol{y}}{\lvert\lvert\boldsymbol{x}\rvert\rvert \, \lvert\lvert\boldsymbol{y}\rvert\rvert}$, $\lvert\lvert\boldsymbol{x}\rvert\rvert = \sqrt{\sum_{p=1}^{P} x_p^2}, \quad \lvert\lvert\boldsymbol{y}\rvert\rvert = \sqrt{\sum_{p=1}^{P} y_p^2}$	$[0, 2]$

angle distance, which considers the cosine of the angle between the two vectors. If the vectors 'point' in opposite directions, their Cosine angle dissimilarity is greatest. That is, the are negatively correlated.

A weakness of Pearson's correlation distance is that Pearson's correlation coefficient is sensitive to outliers in the data. A version of this coefficient which has been proposed to overcome this limitation is Spearman's correlation. The idea is to apply Pearson's correlation to the ranks of the elements of the vectors, rather than the actual values. In particular, let $\boldsymbol{x}_r = (x_{r1}, \ldots, x_{rP})^T$ and $\boldsymbol{y}_r = (y_{r1}, \ldots, y_{rP})^T$ be the vectors containing the ranks of the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively in ascending order (highest value given rank 1). That is, x_{rp} is the rank of x_p, and similarly for y_{rp}. If several elements of a given vector are equal, they are assigned a rank equal

to the mean of their respective positions in the list of ascending values. For example, for vector $(0.1, 0.4, 0.4, 0.5, -31)^T$, their respective positions are $(4, 3, 2, 1, 5)^T$ or $(4, 2, 3, 1, 5)^T$, so that the ranks are given by $(4, 2.5, 2.5, 1, 5)^T$. Spearman's correlation distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is thus given by

$$d_{SC}(\boldsymbol{x}, \boldsymbol{y}) = d_{PC}(\boldsymbol{x}_r, \boldsymbol{y}_r),$$

which ranges between 0 and 2 and is semi-metric.

Distances like Pearson's correlation, Spearman's correlation and the Cosine angle can only detect linear relationships between vectors. A distance which can detect any type of dependence between the vectors, not just linear, would be more widely applicable. Such a distance is provided by using the information-theoretic notion of normalized mutual information (NMI).

NMI is a measure of dependence between two random variables. In our setting, the P elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ are considered to be observations of the random variables $\mathcal{X}$ and $\mathcal{Y}$, respectively. NMI uses the idea of information entropy, or entropy, of a random variable, which is a measure of the uncertainty associated with it. The entropy of a random variable can be estimated by using the probability mass function (PMF) found by considering a histogram of the observations. For example, let $p_{\mathcal{X}}(\cdot)$ denote the PMF of $\boldsymbol{x}$ found by considering the histogram of the elements of $\boldsymbol{x}$ with M bins. That is, $p_{\mathcal{X}}(i)$ gives the proportion of the elements $\{x_p\}_{p=1}^{P}$ in the i^{th} bin. We follow Priness *et al.* (2007) and use the integer value of $\sqrt{P}$ as M. Then the entropy of $\mathcal{X}$ is estimated as

$$\mathrm{E}(\mathcal{X}) = -\sum_{m=1}^{M} p_{\mathcal{X}}(m) \log\left(p_{\mathcal{X}}(m)\right),$$

and similarly for $\mathcal{Y}$ with PMF $p_{\mathcal{Y}}(\cdot)$. The joint entropy of $\mathcal{X}$ and $\mathcal{Y}$ is found by considering the joint PMF, denoted $\{p_{\mathcal{X}\mathcal{Y}}(i,j)\}_{i,j=1}^{M}$, found by considering a two-dimensional histogram of $\boldsymbol{x}$ and $\boldsymbol{y}$. It is then estimated by

$$\mathrm{E}(\mathcal{X}, \mathcal{Y}) = -\sum_{m=1}^{M}\sum_{n=1}^{M} p_{\mathcal{X}\mathcal{Y}}(m,n) \log\left(p_{\mathcal{X}\mathcal{Y}}(m,n)\right).$$

The mutual information (MI) between $\mathcal{X}$ and $\mathcal{Y}$ is then estimated by

$$\mathrm{MI}(\mathcal{X}, \mathcal{Y}) = \mathrm{E}(\mathcal{X}) + \mathrm{E}(\mathcal{Y}) - \mathrm{E}(\mathcal{X}, \mathcal{Y}),$$

and is bounded from below by 0 but not from above. The NMI is therefore used to achieve an upper bound of 1, found by dividing the MI by the larger of the two individual entropies, that is,

$$\mathrm{NMI}(\mathcal{X}, \mathcal{Y}) = \frac{\mathrm{E}(\mathcal{X}) + \mathrm{E}(\mathcal{Y}) - \mathrm{E}(\mathcal{X}, \mathcal{Y})}{\max\{\mathrm{E}(\mathcal{X}), \mathrm{E}(\mathcal{Y})\}},$$

(Michaels *et al.*, 1998). Thus, $\mathrm{NMI}(\mathcal{X}, \mathcal{Y})$ takes the value of 0 if there is no dependence between $\mathcal{X}$ and $\mathcal{Y}$, and the value of 1 if there is maximum dependence. The NMI distance measure is then defined as

$$d_{NMI}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \mathrm{NMI}(\mathcal{X}, \mathcal{Y}),$$

so that maximum dependence equates to minimum distance. This distance is bounded by 0 and 1 and is semi-metric. An advantage of this distance over others is that it is robust with respect to missing values (Priness *et al.*, 2007).

B.2 Distance Measures for Curves

Assume two time-dependent curves $x(t)$ and $y(t)$ defined over the same time-range τ.

The L_2 distance represents the area between the curves, and hence the magnitude of the difference between them (Ferraty and Vieu, 2006; Salem *et al.*, 2010). It is defined by

$$d_L(x, y) = \left(\int_\tau (x(t) - y(t))^2 \, dt \right)^{\frac{1}{2}},$$

is metric and is bounded from below by 0.

The curvature distance quantifies the difference in the rate of change between two curves (Ferraty and Vieu, 2006), and is defined by

$$d_C(x, y) = \left| \int_\tau (x''(t))^2 \, dt - \int_\tau (y''(t))^2 \, dt \right|,$$

where $x''(t)$ denotes the second derivative of $x(t)$, and similarly for $y(t)$. This distance is semi-metric, and is bounded from below by 0. It is not dependent on magnitude, as with the L_2 distance, but solely on the rate of change of the curves.

The Visual L_2 distance quantifies the difference in the scale-invariant shape between curves, analogously to the difference detected by the human eye (Marron and Tsybakov, 1995). Whereas the L_2 distance compares the vertical distance between curves, the Visual L_2 distance considers both vertical and horizontal comparisons. The curves are initially scaled both in time and magnitude, so that their values range between 0 and 1 in time-range $\tau = [0, 1]$; denote these by $x^s(t)$ and $y^s(t)$ where $t \in [0, 1]$. They are then represented as infinite sets of points in the two-dimensional plane, denoted $p_x = \{(t, x^s(t)) \mid t \in [0, 1]\}$ and $p_y = \{(t, y^s(t)) \mid t \in [0, 1]\}$. The visual L_2 distance is then defined by

$$d_V(x, y) = \left(\int_0^1 d_{xy}^2(t)dt + \int_0^1 d_{yx}^2(t)dt \right)^{\frac{1}{2}},$$

where $d_{xy}(t)$ is the minimum Euclidean distance between the point $x^s(t)$ and all points p_y representing y^s, and $d_{yx}(t)$ is the minimum Euclidean distance between the point $y^s(t)$ and all points p_x. Note that $d_{xy}(t)$ is not necessarily equal to $d_{yx}(t)$. This distance is semi-metric and bounded from below by 0.

Other distances include procedures based on landmarks, such as comparing the location of the maxima of the curves as in Cerioli *et al.* (2003). A rank correlation between two curves has also been defined which is equal to 1 if and only if the curves are similar (Heckman and Zamar, 2000), which can be converted into a distance measure. Halima *et al.* (2005) propose a distance measure which extends this rank correlation idea by combining it with the locations of the maxima of the curves. Some measures have also been proposed based on ideas from mathematical morphology which is a branch of mathematics based on set theory, integral geometry and lattice algebra used to analyze spatial structures. Dissimilarity measures include comparing morphological covariance (Epifanio, 2008) and morphological spatial size distributions (Ayala *et al.*, 2008). Parui and Majumder (1983) define a range of dissimilarities based on considering open curves, a finite sequence of equally-spaced points on the curves, and use notions such as the length of the curves etc. Curves can also be represented by a

vector of their moments (Epifanio, 2008), so that multivariate distances such as the Euclidean distance can be applied to these vectors in order to compare the curves.

B.3 Distance Measures for SNPs

Assume two P-dimensional vectors $\boldsymbol{x} = (x_i, \ldots, x_P)^T$ and $\boldsymbol{y} = (y_i, \ldots, y_P)^T$ with discrete-valued elements representing minor allele counts at P SNPs.

The identity-by-state (IBS) distance measure is commonly used (Wessel and Schork, 2006; Wu *et al.*, 2010; Mukhopadhyay *et al.*, 2010), giving a summary measure of the difference in proportion of risk alleles shared across the SNPs. It considers each individual SNP directly, and is defined as

$$d_{IBS}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{1}{2P} \sum_{p=1}^{P} s(x_p, y_p),$$

where $s(x_p, y_p) = 0$ if $x_p = 0$ and $y_p = 2$, or if $x_p = 2$ and $y_p = 0$, $s(x_p, y_p) = 1$ if $x_p = 1$ and $y_p \neq 1$, or if $y_p = 1$ and $x_p \neq 1$, and $s(x_p, y_p) = 2$ if $x_p = y_p$. This distance takes values between 0 and 1 and is semi-metric. Weighted versions of this distance exist where a weight is attached to each of the P SNPs depending on properties such as functional significance or frequency of the minor allele (Wessel and Schork, 2006; Li *et al.*, 2009).

Genetic distances have also been proposed based on the contingency table containing the frequency that each combination of minor allele counts occurs over the SNPs (Selinski and Ickstadt, 2005); see Table B.2.

Table B.2: Contingency table containing the frequency of a given combination of minor allele count between $\boldsymbol{x}$ and $\boldsymbol{y}$ over the P SNPs. m_{kl} is the frequency of $\boldsymbol{x}$ having k minor alleles and $\boldsymbol{y}$ having l minor alleles.

$\boldsymbol{x}$ \ $\boldsymbol{y}$	0	1	2
0	m_{00}	m_{01}	m_{02}
1	m_{10}	m_{11}	m_{12}
2	m_{20}	m_{21}	m_{22}

The key statistics in this table are the number of complete matches of the minor allele counts, $m^+ = \sum_{k=0}^{2} m_{kk}$, and the number of mismatches, $m^- = P - m^+$, where the total number of possible matches is P. Based on these quantities, the following 'matching coefficient' distance measures can be defined; the Simple Matching distance

$$d_{SM}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{m^+}{P},$$

the Sokal and Sneath distance

$$d_{SS}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{m^+}{m^+ + \frac{1}{2}m^-},$$

and the Rogers and Tanimoto I distance

$$d_{RTI}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{m^+}{m^+ + 2m^-}.$$

There is also the Hamman I similarity measure

$$s_{HI}(\boldsymbol{x}, \boldsymbol{y}) = \frac{m^+ - m^-}{P},$$

which can be transformed into a distance measure as follows. Assume N P-dimensional minor allele count vectors $\{\boldsymbol{x}_i\}_{i=1}^N$; this is required in order to normalize the magnitude of the similarities to the range $[0, 1]$. The Hamman I distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is then given by

$$d_{HI}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{s^*(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\max\limits_{i,j}\{s^*(\boldsymbol{x}_i, \boldsymbol{x}_j)\}},$$

where $s^*(\boldsymbol{x}_i, \boldsymbol{x}_j) = s_{HI}(\boldsymbol{x}_i, \boldsymbol{x}_j) + |\min\limits_{i,j}\{s_{HI}(\boldsymbol{x}_i, \boldsymbol{x}_j)\}|$. This takes values between 0 and 1 and is semi-metric.

Each of these distance measures focuses on a different aspect of the SNP data. The Simple Matching distance, for instance, considers only the proportion of direct matches across the SNPs. The Sokal and Sneath and Rogers and Tanimoto I distances quantify a ratio of mismatches to matches. The Hamman I distance takes a different approach, and considers the difference between the matches and mismatches as a proportion of the number of SNPs.

B.4 Distance Measures for Graphs

Assume two undirected graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ defined with the same vertex set such that $V_1 = V_2 = V$. Each graph can be represented by the symmetric $|V| \times |V|$ matrix $\boldsymbol{A}^{(i)} = \left\{a_{kl}^{(i)}\right\}_{k,l=1}^{|V|}$, for $i = 1, 2$, called the adjacency matrix, which contains binary elements representing the presence of an edge between any pair of vertices. That is, the $(k, l)^{th}$ element equals 1 if an edge connects the k^{th} and l^{th} vertices, and equals 0 if there is no edge between them. The diagonal values are 0, since no vertex is connected to itself by an edge.

The Hamming distance (Hamming, 1950) captures the number of edges not shared by graphs G_1 and G_2. It can be quantified in terms of the adjacency matrices as

$$d_H(G_1, G_2) = \sum_{k=1}^{|V|} \sum_{l=1}^{|V|} \left| a_{kl}^{(1)} - a_{kl}^{(2)} \right|.$$

This takes values greater than 0 and is semi-metric.

The Graph Edit distance is obtained by applying the Levenshtein distance (Levenshtein, 1966), initially proposed for strings of symbols or letters, to capture the minimum number of edits required to transform G_1 into G_2. The edits include edge deletions, insertions, and substitutions. A substitution involves simultaneously deleting an edge and inserting an edge, and is considered to be one edit (note that some variations of the distance count substitutions as two edits, since it is comprised of a deletion and an insertion). The distance is formulated via a recursive algorithm, which we define in terms of $\boldsymbol{a}^{(1)} = \left\{a_k^{(1)}\right\}_{k=1}^{|V|^2}$ and $\boldsymbol{a}^{(2)} = \left\{a_k^{(2)}\right\}_{k=1}^{|V|^2}$, the vectorized adjacency matrices, as follows. Define the $(|V|^2 + 1) \times (|V|^2 + 1)$ matrix $\boldsymbol{W} = \{w_{kl}\}_{k,l=1}^{|V|^2+1}$ such that $\{w_{k1} = k - 1\}_{k=1}^{|V|^2+1}$ and $\{w_{1l} = l - 1\}_{l=1}^{|V|^2+1}$. Then for $k, l = 2, \ldots, |V|^2 + 1$,

$$w_{kl} = \min\left\{w_{k-1,l} + 1, w_{k,l-1} + 1, w_{k-1,l-1} + \left| a_{k-1}^{(1)} - a_{l-1}^{(2)} \right|\right\},$$

where $|\cdot|$ is the absolute operator. The graph edit distance is given by $d_{GE}(G_1, G_2) = w_{|V|^2+1,|V|^2+1}$, that is, the $(|V|^2+1, |V|^2+1)^{\text{th}}$ element of $\boldsymbol{W}$. This takes values greater than 0 and is semi-metric.

The Maximum Common Subgraph (MCS) distance (Bunke and Shearer, 1998; Fernández and Valiente, 2001) considers the MCS between two graphs. A subgraph

which is common to G_1 and G_2 is a graph whose vertices and edges are contained within the vertex and edge sets of each graph. The MCS of G_1 and G_2 is the subgraph of largest size contained within both graphs, where the size of a graph $G = (V, E)$ is defined as $|G| = |V| + |E|$, that is, the total number of vertices and edges comprising the graph. There exist many algorithms for finding the MCS of two graphs (Bunke *et al.*, 2002), but since graphs G_1 and G_2 are assumed to have the same number of vertices we use a simple procedure to locate the common edges from the adjacency matrices. The MCS distance is then defined as

$$d_{MCS}(G_1, G_2) = 1 - \frac{|\text{MCS}(G_1, G_2)|}{\max\{|G_1|, |G_2|\}}.$$

This takes values between 0 and 1 and is semi-metric.

Appendix C

Permutational Moment Results for the Trace of a Matrix Product

Here we describe the results of Kazi-Aoual *et al.* (1995) which give the closed form expressions for the first three permutational moments of the trace of a matrix product. In particular, consider the statistic $T = \mathrm{tr}(\boldsymbol{AB})$ where the $N \times N$ matrices $\boldsymbol{A} = \{a_{ij}\}_{i,j=1}^{N}$ and $\boldsymbol{B} = \{b_{ij}\}_{i,j=1}^{N}$ are centered, symmetric, and real-valued. For all $N!$ permutations Π where each $\pi \in \Pi$ is a one-to-one mapping of the set $\{1, \ldots, N\}$ to itself, $T_\pi = \mathrm{tr}(\boldsymbol{AB}_\pi)$ where $\boldsymbol{B}_\pi$ denotes the matrix $\boldsymbol{B}$ with rows and columns simultaneously permuted by π. The set $\{T_\pi\}_{\pi\in\Pi}$ then contains all $N!$ permuted values of T, and has a mean, variance and skewness given by

$$\mu_T = \frac{1}{N!}\sum_{\pi\in\Pi} T_\pi, \quad \sigma_T^2 = \frac{1}{N!}\sum_{\pi\in\Pi} T_\pi^2 - \mu_T^2, \quad \text{and} \quad \gamma_T = \frac{\frac{1}{N!}\sum_{\pi\in\Pi} T_\pi^3 - 3\mu_T\sigma_T^2 - \mu_T^3}{\sigma_T^3},$$

respectively.

The closed form expressions of these quantities are retrievable via the analytical

results of Kazi-Aoual *et al.* (1995), given as follows. Define the quantities

$$\begin{aligned}
A_1 &= \operatorname{tr}(\boldsymbol{A}) & B_1 &= \operatorname{tr}(\boldsymbol{B}) \\
A_2 &= \operatorname{tr}(\boldsymbol{AA}) & B_2 &= \operatorname{tr}(\boldsymbol{BB}) \\
A_3 &= \operatorname{tr}(\boldsymbol{A}^2) & B_3 &= \operatorname{tr}(\boldsymbol{B}^2) \\
A_4 &= \operatorname{tr}(\boldsymbol{AAA}) & B_4 &= \operatorname{tr}(\boldsymbol{BBB}) \\
A_5 &= \operatorname{tr}(\boldsymbol{A}^3) & B_5 &= \operatorname{tr}(\boldsymbol{B}^3) \\
A_6 &= \textstyle\sum \boldsymbol{A}^3 & B_6 &= \textstyle\sum \boldsymbol{B}^3 \\
A_7 &= \boldsymbol{d}_A^T \boldsymbol{d}_{AA} & B_7 &= \boldsymbol{d}_B^T \boldsymbol{d}_{BB} \\
A_8 &= \boldsymbol{d}_A^T \boldsymbol{A} \boldsymbol{d}_A & B_8 &= \boldsymbol{d}_B^T \boldsymbol{B} \boldsymbol{d}_B,
\end{aligned}$$

where $\boldsymbol{A}^k = \{a_{ij}^k\}_{i,j=1}^N$ for $k > 1$, $\sum \boldsymbol{A}^3 = \sum_{i=1}^N \sum_{j=1}^N a_{ij}^3$, $\boldsymbol{d}_A = (a_{11}, \ldots, a_{NN})^T$, $\boldsymbol{d}_{AA} = ((\boldsymbol{AA})_{11}, \ldots, (\boldsymbol{AA})_{NN})^T$and similarly for $\boldsymbol{B}$. The mean and variance are then given by

$$\begin{aligned}
\mu_T = \frac{A_1 B_1}{N-1} \quad \text{and} \quad \sigma_T^2 \;=\; & \frac{2\left((N-1)\,A_2 - A_1^2\right)\left((N-1)\,B_2 - B_1^2\right)}{(N-1)^2(N+1)(N-2)} \\
& + \left[\frac{(N(N+1)A_3 - (N-1)\,(A_1^2 + 2A_2))}{(N+1)N(N-1)(N-2)(N-3)}\right. \\
& \left. \times \left(N(N+1)B_3 - (N-1)\left(B_1^2 + 2B_2\right)\right)\right],
\end{aligned}$$

respectively. For the skewness we first provide the expression for the third moment of T, i.e., $\frac{1}{N!}\sum_{\pi\in\Pi} T_\pi^3$, which is given by

$$
\begin{aligned}
&\frac{(N-6)!}{N!}\left[N^2(N+1)\left(N^2+15N-4\right)A_5B_5-8\left(A_4B_8+A_8B_4\right)\left(3N^2-15N+24\right)\right.\\
&+4\left(N^4-8N^3+19N^2-4N-16\right)A_6B_6+24\left(N^2-N-4\right)\left(A_6B_8+A_8B_6\right)\\
&+6\left(N^4-8N^3+21N^2-6N-24\right)A_8B_8+12\left(N^4-N^3-8N^2+36N-48\right)A_7B_7\\
&+12\left(N^3-2N^2+9N-12\right)\left(A_1A_3B_7+B_1B_3A_7\right)\\
&+3\left(N^4-4N^3-2N^2+9N-12\right)A_1B_1A_3B_3\\
&+24\left(\left(N^3-3N^2-2N+8\right)\left(A_7B_6+A_6B_7\right)\right.\\
&\left.+\left(N^3-2N^2-3N+12\right)\left(A_7B_8+A_8B_7\right)\right)\\
&+12\left(N^2-N+4\right)\left(A_1A_3B_6+B_1B_3A_6\right)\\
&+6\left(2N^3-7N^2-3N+12\right)\left(A_1A_3B_8+B_1B_3A_8\right)\\
&-2N(N-1)\left(N^2-N+4\right)\left(\left(2A_6+3A_8\right)B_5+\left(2B_6+3B_8\right)A_5\right)\\
&-3N(N-1)^2(N+4)\left(\left(A_1A_3+4A_7\right)B_5+\left(B_1B_3+4B_7\right)A_5\right)\\
&+2N(N-1)(N-2)\left(\left(A_1^3+6A_1A_2+8A_4\right)B_5+\left(B_1^3+6B_1B_2+8B_4\right)A_5\right)\\
&+A_1^3\left(\left(N^3-9N^2+23N-14\right)B_1^3+6(N-4)B_1B_2+8B_4\right)\\
&+6A_1A_2\left(\left(N-4\right)B_1^3+\left(N^3-9N^2+24N-14\right)B_1B_2+4(N-3)B_4\right)\\
&+8A_4\left(B_1^3+3(N-3)B_1B_2+\left(N^3-9N^2+26N-22\right)B_4\right)\\
&-16\left(A_1^3B_6+A_6B_1^3\right)-6\left(A_1A_2B_6+B_1B_2A_6\right)\left(2N^2-10N+16\right)\\
&-8\left(A_4B_6+A_6B_4\right)\left(3N^2-15N+16\right)-\left(A_1^3B_8+B_1^3A_8\right)\left(6N^2-30N+24\right)\\
&-6\left(A_1A_2B_8+B_1B_2A_8\right)\left(4N^2-20N+24\right)-(N-2)\left\{24\left(A_1^3B_7+B_1^3A_7\right)\right.\\
&+6\left(A_1A_2B_7+B_1B_2A_7\right)\left(2N^2-10N+24\right)+8\left(A_4B_7+A_7B_4\right)\left(3N^2-15N+24\right)\\
&+\left(3N^2-15N+6\right)\left(A_1^3B_1B_3+B_1^3A_1A_3\right)\\
&+6\left(A_1A_2B_1B_3+B_1B_2A_1A_3\right)\left(N^2-5N+6\right)\\
&\left.\left.+48\left(A_4B_1B_3+B_4A_1A_3\right)\right\}\right].
\end{aligned}
$$

The skewness is then given by

$$\gamma_T=\frac{\frac{1}{N!}\sum_{\pi\in\Pi}T_\pi^3-3\mu_T\sigma_T^2-\mu_T^3}{\sigma_T^3}.$$

Appendix D

Proof of CDF Results for the DBF Statistic

D.1 Derivation of the CDF of DBF for Positive Skewness

First consider the case where $-\infty < f < -1$. By inspection of Figure 5.4 (a), we see that we need only consider the relationship between $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ where $B^s_{\boldsymbol{\Delta}} > \beta$. We thus have that

$$\begin{aligned}
\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) &= P(F_{\boldsymbol{\Delta}} \leq f;\mu_B,\sigma_B,\gamma_B) \\
&= P\left(\beta < B^s_{\boldsymbol{\Delta}} \leq h^{-1}(f);\gamma_B\right) \\
&= P\left(B^s_{\boldsymbol{\Delta}} \leq h^{-1}(f);\gamma_B\right) - (B^s_{\boldsymbol{\Delta}} \leq \beta;\gamma_B) \\
&= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)\,.
\end{aligned}$$

Now consider the case where $\alpha \leq f < \infty$. From Figure 5.4 (a) we see that we must consider the relationship between $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ for $B^s_{\boldsymbol{\Delta}} < \beta$, while adding the

cumulative component of all values of $F_{\boldsymbol{\Delta}}$ for which $B^s_{\boldsymbol{\Delta}} > \beta$. That is,

$$\begin{aligned}
\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) &= P\left(\alpha \leq F_{\boldsymbol{\Delta}} \leq f;\mu_B,\sigma_B,\gamma_B\right) + P(-\infty < F_{\boldsymbol{\Delta}} < -1;\mu_B,\sigma_B,\gamma_B) \\
&= P\left(\frac{-2}{\gamma_B} \leq B^s_{\boldsymbol{\Delta}} \leq h^{-1}(f);\gamma_B\right) + P\left(\beta < B^s_{\boldsymbol{\Delta}} < h^{-1}(-1);\gamma_B\right) \\
&= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\frac{-2}{\gamma_B};\gamma_B\right) + P\left(\beta < B^s_{\boldsymbol{\Delta}} < \infty;\gamma_B\right) \\
&= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\frac{-2}{\gamma_B};\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right),
\end{aligned}$$

and since $-2/\gamma_T \leq B^s_{\boldsymbol{\Delta}} < \infty$, we have $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(-2/\gamma_B;\gamma_B) = 0$ and $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\infty;\gamma_B) = 1$, so that

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) = 1 + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right).$$

Thus we have that the CDF of $F_{\boldsymbol{\Delta}}$ is given by

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B) = \begin{cases} \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right) & -\infty < f < -1 \\ 1 + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right) & \alpha \leq f < \infty \end{cases}$$

for $\gamma_B > 0$, as required.

Next we show that this is a valid CDF by showing that the following conditions are satisfied:

(i) The limit of $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f)$ as f tends to $-\infty$ from the right is 0, and as f tends to ∞ from the left is 1. That is

$$\lim_{f\to-\infty^+}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B)\right] = 0 \quad \text{and} \quad \lim_{f\to\infty^-}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B)\right] = 1.$$

These follow because

$$\begin{aligned}
\lim_{f\to-\infty^+}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B)\right] &= \lim_{b\to\beta^+}\left[\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(b;\gamma_B)\right] - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) \\
&= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B) \\
&= 0,
\end{aligned}$$

and

$$\begin{aligned}\lim_{f\to\infty^-}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right)\right] &= 1+\lim_{b\to\beta^-}\left[\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(b;\gamma_B)\right]-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)\\ &= 1+\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)\\ &= 1.\end{aligned}$$

(ii) $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f)$ is a monotone, non-decreasing function of f. That is, for $f_1 < f_2$, $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_1) \leq \mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_2)$.

For $-\infty < f_1 < f_2 < -1$, we have that

$$\begin{aligned}\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_1;\mu_B,\sigma_B,\gamma_B\right) &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_1);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right)\\ \mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_2;\mu_B,\sigma_B,\gamma_B\right) &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_2);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right),\end{aligned}$$

so that $\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_1;\mu_B,\sigma_B,\gamma_B\right)-\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_2;\mu_B,\sigma_B,\gamma_B\right)$ is equal to

$$\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_1);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_2);\gamma_B\right).$$

This is negative since $h^{-1}(f_1) < h^{-1}(f_2)$ and $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(b;\gamma_B)$ is a non-decreasing, monotone function of b (as it is a valid CDF). Hence $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_1) \leq \mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_2)$, as required.

For $\alpha \leq f_1 < f_2 < \infty$, we have that

$$\begin{aligned}\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_1;\mu_B,\sigma_B,\gamma_B\right) &= 1+\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_1);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right)\\ \mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_2;\mu_B,\sigma_B,\gamma_B\right) &= 1+\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_2);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right),\end{aligned}$$

so that $\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_1;\mu_B,\sigma_B,\gamma_B\right)-\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_2;\mu_B,\sigma_B,\gamma_B\right)$ is equal to

$$\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_1);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f_2);\gamma_B\right).$$

As before, this is negative since $h^{-1}(f_1) < h^{-1}(f_2)$ and $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(b;\gamma_B)$ is non-decreasing and monotone. Hence $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_1) \leq \mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_2)$, as required.

Finally, let $f_1 = -1$ and $f_2 = \alpha$. Then

$$\begin{aligned}\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_1;\mu_B,\sigma_B,\gamma_B) &= 1-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)\\ \mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_2;\mu_B,\sigma_B,\gamma_B) &= 1+\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\frac{-2}{\gamma_B};\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B),\end{aligned}$$

and since $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(-2/\gamma_B;\gamma_B)=0$, we have that $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_1)\leq\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_2)$ holds at the discontinuity.

(iii) $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f)$ is continuous from the right, that is

$$\lim_{\epsilon\to 0^+}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f+\epsilon;\mu_B,\sigma_B,\gamma_B)\right]=\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B).$$

For $-\infty < f < -1$ we have that

$$\begin{aligned}\lim_{\epsilon\to 0^+}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f+\epsilon;\mu_B,\sigma_B,\gamma_B)\right] &= \lim_{\epsilon\to 0^+}\left[\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f+\epsilon);\gamma_B\right)\right]-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)\\ &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\lim_{\epsilon\to 0^+}\left[h^{-1}(f+\epsilon)\right];\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B).\end{aligned}$$

Since

$$\begin{aligned}h^{-1}(f+\epsilon) &= \frac{(T_{\boldsymbol{\Delta}}-\mu_B)(f+\epsilon)-\mu_B}{\sigma_B(1+f+\epsilon)}\\ &= \frac{(T_{\boldsymbol{\Delta}}-\mu_B)f-\mu_B}{\sigma_B(1+f+\epsilon)}+\frac{(T_{\boldsymbol{\Delta}}-\mu_B)\epsilon}{\sigma_B(1+f+\epsilon)}\\ \Rightarrow \lim_{\epsilon\to 0^+}\left[h^{-1}(f+\epsilon)\right] &= \lim_{\epsilon\to 0^+}\left[\frac{(T_{\boldsymbol{\Delta}}-\mu_B)f-\mu_B}{\sigma_B(1+f+\epsilon)}+\frac{(T_{\boldsymbol{\Delta}}-\mu_B)\epsilon}{\sigma_B(1+f+\epsilon)}\right]\\ &= \frac{(T_{\boldsymbol{\Delta}}-\mu_B)f-\mu_B}{\sigma_B(1+f)}\\ &= h^{-1}(f),\end{aligned}$$

it follows that

$$\begin{aligned}\lim_{\epsilon\to 0^+}\left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f+\epsilon;\mu_B,\sigma_B,\gamma_B)\right] &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(\beta;\gamma_B)\\ &= \mathcal{F}_{F_{\boldsymbol{\Delta}}}(f;\mu_B,\sigma_B,\gamma_B),\end{aligned}$$

and for $\alpha \leq f < \infty$ a similar argument yields

$$\begin{aligned}\lim_{\epsilon \to 0^+} \left[\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f+\epsilon;\mu_B,\sigma_B,\gamma_B\right)\right] &= 1 + \lim_{\epsilon \to 0^+} \left[\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f+\epsilon);\gamma_B\right)\right] - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right) \\ &= 1 + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right) \\ &= \mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right),\end{aligned}$$

as required.

Thus $\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right)$ is a valid CDF for $\gamma_B > 0$.

D.2 Derivation of the CDF of DBF for Negative Skewness

First consider the case where $-\infty < f \leq \alpha$. By inspection of Figure 5.4 (b), we see that we need only consider the relationship between $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ where $B^s_{\boldsymbol{\Delta}} > \beta$. We thus have that

$$\begin{aligned}\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right) &= P(F_{\boldsymbol{\Delta}} \leq f;\mu_B,\sigma_B,\gamma_B) \\ &= P\left(\beta < B^s_{\boldsymbol{\Delta}} \leq h^{-1}(f);\gamma_B\right) \\ &= P\left(B^s_{\boldsymbol{\Delta}} \leq h^{-1}(f);\gamma_B\right) - (B^s_{\boldsymbol{\Delta}} \leq \beta;\gamma_B) \\ &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right).\end{aligned}$$

Now consider the case where $-1 < f < \infty$. From Figure 5.4 (b) we see that we must consider the relationship between $F_{\boldsymbol{\Delta}}$ and $B^s_{\boldsymbol{\Delta}}$ for $B^s_{\boldsymbol{\Delta}} < \beta$, while adding the cumulative component of all values of $F_{\boldsymbol{\Delta}}$ for which $B^s_{\boldsymbol{\Delta}} > \beta$. That is,

$$\begin{aligned}\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right) &= P\left(-1 < F_{\boldsymbol{\Delta}} \leq f;\mu_B,\sigma_B,\gamma_B\right) + P(-\infty < F_{\boldsymbol{\Delta}} \leq \alpha;\mu_B,\sigma_B,\gamma_B) \\ &= P\left(-\infty < B^s_{\boldsymbol{\Delta}} \leq h^{-1}(f);\gamma_B\right) + P\left(\beta < B^s_{\boldsymbol{\Delta}} \leq \frac{-2}{\gamma_B};\gamma_B\right) \\ &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(-\infty;\gamma_B\right) + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\frac{-2}{\gamma_B};\gamma_B\right) \\ &\quad -\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right),\end{aligned}$$

and since $-\infty < B^s_{\boldsymbol{\Delta}} \leq -2/\gamma_B$, we have $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(-\infty;\gamma_B\right) = 0$ and $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(-2/\gamma_B;\gamma_B\right) = 1$, so that

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right) = 1 + \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right) - \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right).$$

Thus we have that the CDF of $F_{\boldsymbol{\Delta}}$ is given by

$$\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right)=\begin{cases}\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right) & -\infty<f\leq\alpha\\ 1+\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(h^{-1}(f);\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right) & -1<f<\infty\end{cases}$$

for $\gamma_B<0$ and $\alpha<-1$, as required.

We now show that this is a valid CDF by showing that the three required properties are satisfied.

(i) This is the same as in Appendix D.1.

(ii) As discussed in Appendix D.1, $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}(b;\gamma_B)$ is a valid CDF and so it is a non-decreasing, monotone function of b. Since $F_{\boldsymbol{\Delta}}$ is an increasing function of $B^s_{\boldsymbol{\Delta}}$ on both sides of β, i.e., for the ranges $-\infty<f\leq\alpha$ and $-1<f<\infty$, $\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right)$ is also a non-decreasing, monotone function in these ranges.

Let $f_1=\alpha$ and $f_2=-1$. Then

$$\begin{aligned}\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_1;\mu_B,\sigma_B,\gamma_B\right) &= \mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\frac{-2}{\gamma_B};\gamma_B\right)-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right)\\ \mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f_2;\mu_B,\sigma_B,\gamma_B\right) &= 1-\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(\beta;\gamma_B\right),\end{aligned}$$

and since $\mathcal{F}_{B^s_{\boldsymbol{\Delta}}}\left(-2/\gamma_B;\gamma_B\right)=1$, we have that $\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_1)\leq\mathcal{F}_{F_{\boldsymbol{\Delta}}}(f_2)$ holds at the discontinuity.

(iii) This is the same as in Appendix D.1.

Therefore $\mathcal{F}_{F_{\boldsymbol{\Delta}}}\left(f;\mu_B,\sigma_B,\gamma_B\right)$ is a valid CDF.

Appendix E

Cubic Smoothing Spline Smoothing

Assume N objects have been observed at S time-points $t_1, \ldots, t_S$ over time-range $\tau = [t_1, t_S]$, giving rise to the longitudinal observations $\{\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iS})^T\}_{i=1}^N$ where y_{is} represents the observation of the i^{th} object at time-point t_s. We wish to model the longitudinal observations as noisy realizations of curves $\{z_i(t)\}_{i=1}^N$ defined for $t \in \tau$ via the regression model $\boldsymbol{y}_i = z_i(\boldsymbol{t}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_S)^T$ with $\epsilon_s \sim N(0, \sigma^2)$ for $s = 1, \ldots, S$ where σ^2 is unknown.

Cubic smoothing spline smoothing describes the approach whereby the curves are represented as a linear combination of K B-spline basis functions $\{\phi_k(t)\}_{k=1}^K$ and the resulting curve estimate is penalized via its roughness. The roughness of $z_i(t)$ is quantified by its curvature, $\int_\tau (z_i''(t))^2 dt$, where $z_i''(t)$ denotes the second derivative of $z_i(t)$. Under basis function expansion, $z_i(t) = \boldsymbol{\phi}(t)^T \boldsymbol{c}_i$, where $\boldsymbol{\phi} = (\phi_1(t), \ldots, \phi_K(t))^T$ and $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iK})^T$, and the curvature can be written in matrix form as $\boldsymbol{c}_i^T \boldsymbol{R} \boldsymbol{c}_i$ where $\boldsymbol{R} = \int_\tau \boldsymbol{\phi}''(t) \boldsymbol{\phi}''(t)^T dt$. $\boldsymbol{R}$ is a $K \times K$ matrix with $(i,j)^{\text{th}}$ element $\int_\tau \phi_i''(t) \phi_j''(t) dt$, that is, the inner product of the second derivative of the basis functions $\phi_i(t)$ and $\phi_j(t)$. The optimal $\hat{\boldsymbol{c}}_i$ is then found by solving the penalized least-squares optimization

$$\min_{c_i, \lambda_i} \left\{ (\boldsymbol{y}_i - \boldsymbol{\Phi} \boldsymbol{c}_i)^T (\boldsymbol{y}_i - \boldsymbol{\Phi} \boldsymbol{c}_i) + \lambda_i \boldsymbol{c}_i^T \boldsymbol{R} \boldsymbol{c}_i \right\},$$

where $\boldsymbol{\Phi} = (\boldsymbol{\phi}(t_1), \ldots, \boldsymbol{\phi}(t_S))^T$ is the $S \times K$ design matrix containing the values of the basis functions evaluated at the observation time-points and λ_i is a positive smoothing

parameter. From standard least-squares theory, the minimizing $\hat{\boldsymbol{c}}_i$ is found as

$$\hat{\boldsymbol{c}}_i = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda_i \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}_i,$$

and λ_i is chosen optimally via procedures such as generalized cross-validation in order to trade-off the curve of best fit of the data with the smoothest curve (Ramsay and Silverman, 2006). This leads to curve estimates $\{\hat{z}_i(t) = \boldsymbol{\phi}(t)^T\hat{\boldsymbol{c}}_i\}_{i=1}^N$, where each longitudinal observation is represented by a curve which is as smooth as possible while capturing the exhibited variational patterns.

Appendix F

Derivations for the Permutational Moments of H

Here we give all the quantities required for the permutational moments of H. Note that where the distinct index patterns refer to the indices of two or three matrices being multiplied, a comma is used to separate every two which refer to one matrix.

Table F.1: Expressions for the quantities associated with the i^{th} component of the decomposition of the r^{th} permutational moment of H, with $r = 1, 2, 3$. Here, $\boldsymbol{G}^k = \{g_{ij}^k\}_{i,j=1}^N$ and $\sum \boldsymbol{G}^k = \sum_{i=1}^N \sum_{j=1}^N g_{ij}^k$ for $k = 2, 3$, $\boldsymbol{d}_G = (g_{11}, \ldots, g_{NN})^T$, $\boldsymbol{w} = \left(\sum_{i=1}^N g_{1i}^2, \ldots, \sum_{i=1}^N g_{Ni}^2\right)^T$, $\boldsymbol{H}^k = \{h_{ij}^k\}_{i,j=1}^N$ for $k = 2, 3$, $\sum \boldsymbol{H} = \sum_{i=1}^N \sum_{j=1}^N h_{ij}$, $\sum \boldsymbol{H}^3 = \sum_{i=1}^N \sum_{j=1}^N h_{ij}^3$, $\boldsymbol{d}_H = (h_{11}, \ldots, h_{NN})^T$, $\boldsymbol{d}_H^2 = (h_{11}^2, \ldots, h_{NN}^2)^T$, $\boldsymbol{v}_1 = \left(\sum_{i=1}^N h_{1i}, \ldots, \sum_{i=1}^N h_{Ni}\right)^T$, $\boldsymbol{v}_1^2 = \left(\left(\sum_{i=1}^N h_{1i}\right)^2, \ldots, \left(\sum_{i=1}^N h_{Ni}\right)^2\right)^T$, $\boldsymbol{v}_2 = \left(\sum_{i=1}^N h_{1i}^2, \ldots, \sum_{i=1}^N h_{Ni}^2\right)^T$, $\sum \boldsymbol{H}\boldsymbol{v}_1 = \sum_{i=1}^N (\boldsymbol{H}\boldsymbol{v}_1)_i$ and $\sum \boldsymbol{H}(\boldsymbol{v}_1^2) = \sum_{i=1}^N (\boldsymbol{H}(\boldsymbol{v}_1^2))_i$.

r	i	$p_i^{(r)}$	$w_i^{(r)}$	$E_\Pi\left(p_i^{(r)}(\boldsymbol{G})\right)$	$\sum\left(p_i^{(r)}(\boldsymbol{H})\right)$
1	1	ii	1	$\frac{(N-1)!}{N!}\text{tr}(\boldsymbol{G})$	M
	2	ij	1	$-\frac{(N-2)!}{N!}\text{tr}(\boldsymbol{G})$	$\sum \boldsymbol{H} - M$
2	1	ii, ii	1	$\frac{(N-1)!}{N!}\text{tr}(\boldsymbol{G}^2)$	$\text{tr}(\boldsymbol{H}^2)$
	2	ii, ij	4	$-\frac{(N-2)!}{N!}\text{tr}(\boldsymbol{G}^2)$	$\boldsymbol{d}_H^T\boldsymbol{v}_1 - \text{tr}(\boldsymbol{H}^2)$
	3	ii, jj	1	$\frac{(N-2)!}{N!}\left(\text{tr}(\boldsymbol{G})^2 - \text{tr}(\boldsymbol{G}^2)\right)$	$M^2 - \text{tr}(\boldsymbol{H}^2)$

Continued on next page

Table F.1 – *Continued from previous page*

r	i	$p_i^{(r)}$	$w_i^{(r)}$	$\pi_i^{(r)}(\boldsymbol{G})$	$\sigma_i^{(r)}(\boldsymbol{H})$
	4	ij, ij	2	$\frac{(N-2)!}{N!}\left(\mathrm{tr}\left(\boldsymbol{GG}\right) - \mathrm{tr}\left(\boldsymbol{G}^2\right)\right)$	$M - \mathrm{tr}\left(\boldsymbol{H}^2\right)$
	5	ii, kj	2	$\frac{(N-3)!}{N!}\left(\mathrm{tr}\left(\boldsymbol{G}^2\right) - \mathrm{tr}\left(\boldsymbol{G}\right)^2\right.$	$M\sum \boldsymbol{H} - M^2 - 2\boldsymbol{d}_H^T\boldsymbol{v}_1$
				$\left.+2\mathrm{tr}\left(\boldsymbol{H}^2\right)\right)$	$+2\mathrm{tr}\left(\boldsymbol{H}^2\right)$
	6	ij, kj	4	$\frac{(N-3)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^2\right) - \mathrm{tr}\left(\boldsymbol{GG}\right)\right)$	$\sum \boldsymbol{H} - M - 2\boldsymbol{d}_H^T\boldsymbol{v}_1$
					$+2\mathrm{tr}\left(\boldsymbol{H}^2\right)$
	7	ij, kl	1	$\frac{(N-4)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{GG}\right) + \mathrm{tr}\left(\boldsymbol{G}\right)^2\right.$	$\left(\sum \boldsymbol{H}\right)^2 + 8\boldsymbol{d}_H^T\boldsymbol{v}_1$
				$\left.-6\mathrm{tr}\left(\boldsymbol{G}^2\right)\right)$	$-\left(\sum \boldsymbol{H}\right)(2M+4)$
					$-6\mathrm{tr}\left(\boldsymbol{H}^2\right) + M\left(2+M\right)$
3	1	ii, ii, ii	1	$\frac{1}{N}\mathrm{tr}\left(\boldsymbol{G}^3\right)$	$\mathrm{tr}\left(\boldsymbol{H}^3\right)$
	2	ii, ii, ij	6	$-\frac{(N-2)!}{N!}\mathrm{tr}\left(\boldsymbol{G}^3\right)$	$\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1 - \mathrm{tr}\left(\boldsymbol{H}^3\right)$
	3	ii, ii, jj	3	$\frac{(N-2)!}{N!}\left(\mathrm{tr}\left(\boldsymbol{G}\right)\mathrm{tr}\left(\boldsymbol{G}^2\right)\right.$	$M\mathrm{tr}\left(\boldsymbol{H}^2\right) - \mathrm{tr}\left(\boldsymbol{H}^3\right)$
				$\left.-\mathrm{tr}\left(\boldsymbol{G}^3\right)\right)$	
	4	ii, ij, ij	12	$\frac{(N-2)!}{N!}\left(\boldsymbol{d}_G^T\boldsymbol{w} - \mathrm{tr}\left(\boldsymbol{G}^3\right)\right)$	$\boldsymbol{d}_H^T\boldsymbol{v}_2 - \mathrm{tr}\left(\boldsymbol{H}^3\right)$
	5	ii, ij, jj	6	$\frac{(N-2)!}{N!}\left(\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G - \mathrm{tr}\left(\boldsymbol{G}^3\right)\right)$	$\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H - \mathrm{tr}\left(\boldsymbol{H}^3\right)$
	6	ij, ij, ij	4	$\frac{(N-2)!}{N!}\left(\sum \boldsymbol{G}^3 - \mathrm{tr}\left(\boldsymbol{G}^3\right)\right)$	$\sum \boldsymbol{H}^3 - \mathrm{tr}\left(\boldsymbol{H}^3\right)$
	7	ii, ik, ij	12	$\frac{(N-2)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^3\right) - \boldsymbol{d}_G^T\boldsymbol{w}\right)$	$2\mathrm{tr}\left(\boldsymbol{H}^3\right) - \boldsymbol{d}_H^T\boldsymbol{v}_2$
	8	ii, ii, kj	3	$\frac{(N-3)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^3\right)\right.$	$2\mathrm{tr}\left(\boldsymbol{H}^3\right) - 2\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
				$\left.-\mathrm{tr}\left(\boldsymbol{G}\right)\mathrm{tr}\left(\boldsymbol{G}^2\right)\right)$	$+\mathrm{tr}\left(\boldsymbol{H}^2\right)\left(\sum \boldsymbol{H} - M\right)$
	9	ii, ik, jj	12	$\frac{(N-3)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^3\right) - \boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G\right.$	$M\left(\boldsymbol{d}_H^T\boldsymbol{v}_1 - \mathrm{tr}\left(\boldsymbol{H}^2\right)\right)$
				$\left.-\mathrm{tr}\left(\boldsymbol{G}\right)\mathrm{tr}\left(\boldsymbol{G}^2\right)\right)$	$-\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H - \left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
					$+2\mathrm{tr}\left(\boldsymbol{H}^3\right)$
	10	ik, ij, ij	24	$\frac{(N-3)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^3\right) - \boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\boldsymbol{v}_1^T\boldsymbol{v}_2 + 2\mathrm{tr}\left(\boldsymbol{H}^3\right)$
				$\left.-\sum \boldsymbol{G}^3\right)$	$-\boldsymbol{d}_H^T\boldsymbol{v}_2 - \sum \boldsymbol{H}^3$
					$-\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
	11	ii, kj, ij	24	$\frac{(N-3)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^3\right) - \boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1 + 2\mathrm{tr}\left(\boldsymbol{H}^3\right)$
				$\left.-\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G\right)$	$-\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1 - \boldsymbol{d}_H^T\boldsymbol{v}_2$
					$-\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H$
	12	ii, jj, kk	1	$\frac{(N-3)!}{N!}\left(\mathrm{tr}\left(\boldsymbol{G}\right)^3 + 2\mathrm{tr}\left(\boldsymbol{G}^3\right)\right.$	$2\mathrm{tr}\left(\boldsymbol{H}^3\right) + M\left(M^2\right.$
				$\left.-3\mathrm{tr}\left(\boldsymbol{G}^2\right)\mathrm{tr}\left(\boldsymbol{G}\right)\right)$	$\left.-3\mathrm{tr}\left(\boldsymbol{H}^2\right)\right)$

Continued on next page

Table F.1 – *Continued from previous page*

r	i	$p_i^{(r)}$	$w_i^{(r)}$	$\pi_i^{(r)}(\boldsymbol{G})$	$\sigma_i^{(r)}(\boldsymbol{H})$
	13	ii, jk, jk	6	$\frac{(N-3)!}{N!}\left(2\text{tr}\left(\boldsymbol{G}^3\right) - 2\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$2\text{tr}\left(\boldsymbol{H}^3\right) - 2\boldsymbol{d}_H^T\boldsymbol{v}_2$
				$+\text{tr}\left(\boldsymbol{G}\right)\left(\sum \boldsymbol{G}^2 - \text{tr}\left(\boldsymbol{G}^2\right)\right))$	$+M\left(M - \text{tr}\left(\boldsymbol{H}^2\right)\right)$
	14	ij, ik, kj	8	$\frac{(N-3)!}{N!}\left(\text{tr}\left(\boldsymbol{GGG}\right) + 2\text{tr}\left(\boldsymbol{G}^3\right)\right.$	$M + 2\text{tr}\left(\boldsymbol{H}^3\right) - 3\boldsymbol{d}_H^T\boldsymbol{v}_2$
				$\left.-3\boldsymbol{d}_G^T\boldsymbol{v}_2\right)$	
	15	ii, ij, kl	12	$\frac{(N-4)!}{N!}\left(-6\text{tr}\left(\boldsymbol{G}^3\right)\right.$	$\sum \boldsymbol{H}\left(\boldsymbol{d}_H^T\boldsymbol{v}_1 - \text{tr}\left(\boldsymbol{H}^2\right)\right)$
				$+2\boldsymbol{d}_G^T\boldsymbol{w} + 2\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G$	$-2\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right) + 6\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
				$\left.+\text{tr}\left(\boldsymbol{G}^2\right)\text{tr}\left(\boldsymbol{G}\right)\right)$	$-2\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1 + 2\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H$
					$+M\left(\text{tr}\left(\boldsymbol{H^2}\right) - \boldsymbol{d}_H^T\boldsymbol{v}_1\right)$
					$+2\boldsymbol{d}_H^T\boldsymbol{v}_2 - 6\text{tr}\left(\boldsymbol{H}^3\right)$
	16	ij, ik, il	8	$\frac{(N-4)!}{N!}\left(-6\text{tr}\left(\boldsymbol{G}^3\right) + 3\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\sum \boldsymbol{H}\left(\boldsymbol{v}_1^2\right) - 6\text{tr}\left(\boldsymbol{H}^3\right)$
				$\left.+2\sum \boldsymbol{G}^3\right)$	$-3\boldsymbol{v}_1^T\boldsymbol{v}_2 - 3\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right)$
					$+6\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1 + 3\boldsymbol{d}_H^T\boldsymbol{v}_2$
					$+2\sum \boldsymbol{H}^3$
	17	ii, jj, kl	3	$\frac{(N-4)!}{N!}\left(-6\text{tr}\left(\boldsymbol{G}^3\right)\right.$	$\sum \boldsymbol{H}\left(M^2 - \text{tr}\left(\boldsymbol{H}^2\right)\right)$
				$+5\text{tr}\left(\boldsymbol{G}^2\right)\text{tr}\left(\boldsymbol{G}\right)$	$+M\left(5\text{tr}\left(\boldsymbol{H}^2\right) - 4\boldsymbol{d}_H^T\boldsymbol{v}_1\right.$
				$\left.+2\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G - \text{tr}\left(\boldsymbol{G}\right)^3\right)$	$\left.-M^2\right) + 4\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
					$-6\text{tr}\left(\boldsymbol{H}^3\right) + +2\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H$
	18	ii, jk, jl	12	$\frac{(N-4)!}{N!}\left(-6\text{tr}\left(\boldsymbol{G}^3\right) + 3\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$2M\left(2\text{tr}\left(\boldsymbol{H}^2\right) - \boldsymbol{d}_H^T\boldsymbol{v}_1\right)$
				$\text{tr}\left(\boldsymbol{G}\right)\left(2\text{tr}\left(\boldsymbol{G}^2\right) - \sum \boldsymbol{G}^2\right)$	$+M\left(\sum \boldsymbol{H}\boldsymbol{v}_1 - M\right)$
				$\left.+2\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G\right)$	$+3\boldsymbol{d}_H^T\boldsymbol{v}_2 + 2\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H$
					$-2\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1 - 6\text{tr}\left(\boldsymbol{H}^3\right)$
					$-\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right) + 4\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
	19	ij, ij, kl	6	$\frac{(N-4)!}{N!}\left(-6\text{tr}\left(\boldsymbol{G}^3\right) + 4\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\sum \boldsymbol{H}\left(M - \text{tr}\left(\boldsymbol{H}^2\right)\right)$
				$+\text{tr}\left(\boldsymbol{G}\right)\left(\text{tr}\left(\boldsymbol{G}^2\right) - \sum \boldsymbol{G}^2\right)$	$-6\text{tr}\left(\boldsymbol{H}^3\right) + 4\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
				$\left.\sum \boldsymbol{G}^2\right)$	$-4\boldsymbol{v}_1^T\boldsymbol{v}_2 + 4\boldsymbol{d}_H^T\boldsymbol{v}_2$
					$+M\left(\text{tr}\left(\boldsymbol{H}^2\right) - M\right)$
					$+2\sum \boldsymbol{H}^3$
	20	ik, ij, lj	24	$\frac{(N-4)!}{N!}\left(-6\text{tr}\left(\boldsymbol{G}^3\right) + 5\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\boldsymbol{v}_1^T\boldsymbol{H}\boldsymbol{v}_1 - 6\text{tr}\left(\boldsymbol{H}^3\right)$
				$\left.+\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G - \text{tr}\left(\boldsymbol{GGG}\right)\right.$	$+4\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1 + 5\boldsymbol{d}_H^T\boldsymbol{v}_2$

Continued on next page

Table F.1 – *Continued from previous page*

r	i	$p_i^{(r)}$	$w_i^{(r)}$	$\pi_i^{(r)}(\boldsymbol{G})$	$\sigma_i^{(r)}(\boldsymbol{H})$
				$+\sum \boldsymbol{G}^3)$	$-2\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1+\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H$
					$+\sum \boldsymbol{H}^3-M-\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right)$
					$-2\boldsymbol{v}_1^T\boldsymbol{v}_2$
	21	ii,jk,lp	3	$\frac{(N-5)!}{N!}\left(24\mathrm{tr}\left(\boldsymbol{G}^3\right)-8\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\sum \boldsymbol{H}\left(M\sum \boldsymbol{H}-4\boldsymbol{d}_H^T\boldsymbol{v}_1\right.$
				$+2\mathrm{tr}\left(\boldsymbol{G}\right)\left(\sum \boldsymbol{G}^2-5\mathrm{tr}\left(\boldsymbol{G}^2\right)\right)$	$+4\mathrm{tr}\left(\boldsymbol{H}^2\right)-2M^2)$
				$\left.-8\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G+\mathrm{tr}\left(\boldsymbol{G}\right)^3\right)$	$+8\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right)-24\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
					$-8\boldsymbol{d}_H^T\boldsymbol{v}_2+8\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1$
					$-8\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H+12M\boldsymbol{d}_H^T\boldsymbol{v}_1$
					$+M^2(2+M)+24\mathrm{tr}\left(\boldsymbol{H}^3\right)$
					$-10M\mathrm{tr}\left(\boldsymbol{H}^2\right)$
					$-4M\sum \boldsymbol{H}\boldsymbol{v}_1$
	22	ij,ik,lp	12	$\frac{(N-5)!}{N!}\left(24\mathrm{tr}\left(\boldsymbol{G}^3\right)-16\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\sum \boldsymbol{H}\left(\sum \boldsymbol{H}-2\boldsymbol{d}_H^T\boldsymbol{v}_1\right.$
				$-4\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G+2\mathrm{tr}\left(\boldsymbol{G}\boldsymbol{G}\boldsymbol{G}\right)$	$-M+2\mathrm{tr}\left(\boldsymbol{H}^2\right))$
				$+\mathrm{tr}\left(\boldsymbol{G}\right)\left(\sum \boldsymbol{G}^2-\mathrm{tr}\left(\boldsymbol{G}^2\right)\right)$	$+10\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right)-24\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
				$\left.-4\sum \boldsymbol{G}^3\right)$	$-16\boldsymbol{d}_H^T\boldsymbol{v}_2+8\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1$
					$-4\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H-2M\mathrm{tr}\left(\boldsymbol{H}^2\right)$
					$-2\sum \boldsymbol{H}\left(\boldsymbol{v}_1^2\right)+10\boldsymbol{v}_1^T\boldsymbol{v}_2$
					$-4\mathrm{tr}\left(\boldsymbol{H}^3\right)-4\boldsymbol{v}_1^T\boldsymbol{H}\boldsymbol{v}_1$
					$+M\left(2\boldsymbol{d}_H^T\boldsymbol{v}_1+M+2\right.$
					$\left.-\sum \boldsymbol{H}\boldsymbol{v}_1\right)+24\mathrm{tr}\left(\boldsymbol{H}^3\right)$
	23	ij,kl,pq	1	$\frac{(N-6)!}{N!}\left(-120\mathrm{tr}\left(\boldsymbol{G}^3\right)+72\boldsymbol{d}_G^T\boldsymbol{w}\right.$	$\sum \boldsymbol{H}\left(\left(\sum \boldsymbol{H}\right)^2+24\boldsymbol{d}_H^T\boldsymbol{v}_1\right.$
				$+6\mathrm{tr}\left(\boldsymbol{G}\right)\left(3\mathrm{tr}\left(\boldsymbol{G}^2\right)-\sum \boldsymbol{G}^2\right)$	$+3M(M+2)-18\mathrm{tr}\left(\boldsymbol{H}^2\right)$
				$+24\boldsymbol{d}_G^T\boldsymbol{G}\boldsymbol{d}_G+16\sum \boldsymbol{G}^3$	$\left.-3\sum \boldsymbol{H}\left(M+4\right)\right)$
				$-8\mathrm{tr}\left(\boldsymbol{G}\boldsymbol{G}\boldsymbol{G}\right)-\mathrm{tr}\left(\boldsymbol{G}\right)^3$	$+144\left(\boldsymbol{d}_H^2\right)^T\boldsymbol{v}_1$
					$-72\boldsymbol{d}_H^T\left(\boldsymbol{v}_1^2\right)+72\boldsymbol{d}_H^T\boldsymbol{v}_2$
					$+24\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{d}_H+16\sum \boldsymbol{H}^3$
					$-48\boldsymbol{d}_H^T\boldsymbol{H}\boldsymbol{v}_1+24\boldsymbol{v}_1^T\boldsymbol{H}\boldsymbol{v}_1$
					$+M\left(12\sum \boldsymbol{H}\boldsymbol{v}_1\right.$
					$-24\boldsymbol{d}_H^T\boldsymbol{v}_1-M^2-6M$

Continued on next page

Table F.1 – *Continued from previous page*

r	i	$p_i^{(r)}$	$w_i^{(r)}$	$\pi_i^{(r)}(\boldsymbol{G})$	$\sigma_i^{(r)}(\boldsymbol{H})$
					$-8 + 18\text{tr}\left(\boldsymbol{H}^2\right))$
					$-48\boldsymbol{v}_1^T\boldsymbol{v}_2 + 16\sum \boldsymbol{H}\left(\boldsymbol{v}_1^2\right)$

The quantities given in Table F.1 were checked empirically on simulated datasets for $N = 7, 8, 9$ as they were derived. That is, all $N!$ permutations were enumerated to ensure the expressions for the expected values with respect to $\boldsymbol{G}$ were correct, in addition to all summations with respect to $\boldsymbol{H}$.

We also demonstrate how certain quantities are derived for the second permutational moment of $H = \text{tr}\left(\boldsymbol{HG}\right)$, where $\boldsymbol{H} = \{h_{ij}\}_{i,j=1}^N$ and $\boldsymbol{G} = \{g_{ij}\}_{i,j=1}^N$ are $N \times N$ matrices satisfying the following properties: $\boldsymbol{H}$ is the projection matrix arising from the $N \times M$ regressor matrix $\boldsymbol{X}$ of full rank, i.e., it is symmetric, not centered and $\text{tr}\left(\boldsymbol{H}\right) = M$, and $\boldsymbol{G}$ is symmetric and centered.

We begin by considering how the weights $w_2^{(2)}$ and $w_4^{(2)}$ are derived. Weight $w_2^{(2)}$ corresponds to pattern $p_2^{(2)} = ii, ij$, for which $E_\Pi\left(p_2^{(2)}\left(\boldsymbol{G}\right)\right) = E_\Pi\left(g_{ii}g_{ij}\right)$ and $\sum\left(p_2^{(2)}\left(\boldsymbol{H}\right)\right) = \sum_{i\neq j} h_{ii}h_{ij}$. Due to symmetry of $\boldsymbol{G}$ and $\boldsymbol{H}$, we have

$$\begin{aligned}
E_\Pi\left(p_2^{(2)}\left(\boldsymbol{G}\right)\right)\sum\left(p_2^{(2)}\left(\boldsymbol{H}\right)\right) &= E_\Pi\left(g_{ii}g_{ij}\right)\sum_{i\neq j} h_{ii}h_{ij} \\
&= E_\Pi\left(g_{ii}g_{ji}\right)\sum_{i\neq j} h_{ii}h_{ji} \\
&= E_\Pi\left(g_{ij}g_{ii}\right)\sum_{i\neq j} h_{ij}h_{ii} \\
&= E_\Pi\left(g_{ji}g_{ii}\right)\sum_{i\neq j} h_{ji}h_{ii}.
\end{aligned}$$

Each of these summations is distinct, because swapping the indices of any given summation, i.e., setting $i \to j$ and $j \to i$, does not yield any other summation. For instance, swapping i and j in the first summation variation $E_\Pi\left(g_{ii}g_{ij}\right)\sum_{i\neq j} h_{ii}h_{ij}$ yields $E_\Pi\left(g_{jj}g_{ji}\right)\sum_{i\neq j} h_{jj}h_{ji}$, which is not equal to any of the other three summation variations. Thus the corresponding weight, $w_2^{(2)}$, is 4.

Weight $w_4^{(2)}$ corresponds to pattern $p_4^{(2)} = ij, ij$, for which $E_\Pi\left(p_4^{(2)}\left(\boldsymbol{G}\right)\right) = E_\Pi\left(g_{ij}^2\right)$

and $\sum\left(p_4^{(2)}(\boldsymbol{H})\right)=\sum_{i\neq j}h_{ij}^2$. Due to symmetry we have

$$\begin{aligned}E_\Pi\left(p_4^{(2)}(\boldsymbol{G})\right)\sum\left(p_4^{(2)}(\boldsymbol{H})\right) &= E_\Pi\left(g_{ij}g_{ij}\right)\sum_{i\neq j}h_{ij}h_{ij}\\ &= E_\Pi\left(g_{ij}g_{ji}\right)\sum_{i\neq j}h_{ij}h_{ji}\\ &= E_\Pi\left(g_{ji}g_{ij}\right)\sum_{i\neq j}h_{ji}h_{ij}\\ &= E_\Pi\left(g_{ji}g_{ji}\right)\sum_{i\neq j}h_{ji}h_{ji},\end{aligned}$$

In this case, however, only the variations

$$E_\Pi\left(g_{ij}g_{ij}\right)\sum_{i\neq j}h_{ij}h_{ij}\quad\text{and}\quad E_\Pi\left(g_{ij}g_{ji}\right)\sum_{i\neq j}h_{ij}h_{ji}$$

are distinct, since the other two variations are found from these by swapping i and j. The corresponding weight of pattern $p_4^{(4)}$, $w_4^{(4)}$, is therefore 2.

Now we show how the four quantities $\sum\left(p_2^{(2)}(\boldsymbol{H})\right)$, $E_\Pi\left(p_2^{(2)}(\boldsymbol{G})\right)$, $\sum\left(p_6^{(2)}(\boldsymbol{H})\right)$ and $E_\Pi\left(p_6^{(2)}(\boldsymbol{G})\right)$ are derived. Consider first the two quantities associated with pattern $p_2^{(2)}$. We have

$$\begin{aligned}\sum\left(p_2^{(2)}(\boldsymbol{H})\right) &= \sum_{i\neq j}h_{ii}h_{ij}\\ &= \sum_{i=1}^N\sum_{j=1}^N h_{ii}h_{ij}-\sum_{i=j=1}^N h_{ii}h_{ij}\\ &= \sum_{i=1}^N h_{ii}\left(\sum_{j=1}^N h_{ij}\right)-\sum_{i=1}^N h_{ii}^2\\ &= \boldsymbol{d}_H^T\boldsymbol{v}_1-\operatorname{tr}\left(\boldsymbol{H}^2\right),\end{aligned}$$

where $\boldsymbol{d}_H=(h_{11},\ldots,h_{NN})^T$, $\boldsymbol{v}_1=\left(\sum_{i=1}^N h_{1i},\ldots,\sum_{i=1}^N h_{Ni}\right)^T$ and $\boldsymbol{H}^2=\{h_{ij}^2\}_{i,j=1}^N$,

and

$$\begin{aligned} E_{\Pi}\left(p_2^{(2)}\left(\boldsymbol{G}\right)\right) &= E_{\Pi}\left(g_{ii}g_{ij}\right) \\ &= \frac{(N-2)!}{N!}\sum\left(p_2^{(2)}\left(\boldsymbol{G}\right)\right) \\ &= \frac{(N-2)!}{N!}\sum_{i\neq j} g_{ii}g_{ij} \\ &= \frac{(N-2)!}{N!}\left(\sum_{i=1}^{N} g_{ii}\left(\sum_{j=1}^{N} g_{ij}\right) - \sum_{i=1}^{N} g_{ii}^2\right) \\ &= -\frac{(N-2)!}{N!}\mathrm{tr}\left(\boldsymbol{G}^2\right), \end{aligned}$$

since $\boldsymbol{G}$ is centered, where $\boldsymbol{G}^2 = \{g_{ij}^2\}_{i,j=1}^{N}$. Now consider the two quantities associated with the pattern $p_6^{(2)} = ij, kj$. We have

$$\begin{aligned} \sum\left(p_6^{(2)}\left(\boldsymbol{H}\right)\right) &= \sum_{i\neq j\neq k} h_{ij}h_{kj} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} h_{ij}h_{kj} - \sum_{i=1}^{N} h_{ii}^2 - \sum_{i\neq j} h_{ij}^2 - 2\sum_{i\neq j} h_{ii}h_{ji}, \end{aligned}$$

where

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} h_{ij}h_{kj} &= \sum_{j=1}^{N}\left(\sum_{i=1}^{N} h_{ij}\right)\left(\sum_{k=1}^{N} h_{kj}\right) \\
&= \boldsymbol{v}_1^T\boldsymbol{v}_1 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\boldsymbol{H}\boldsymbol{H}\right)_{ij} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} h_{ij} \\
&= \sum \boldsymbol{H}, \\
\sum_{i=1}^{N} h_{ii}^2 &= \mathrm{tr}\left(\boldsymbol{H}^2\right), \\
\sum_{i\neq j} h_{ij}^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N} h_{ij}^2 - \sum_{i=1}^{N} h_{ii}^2 \\
&= \mathrm{tr}\left(\boldsymbol{H}\boldsymbol{H}\right) - \mathrm{tr}\left(\boldsymbol{H}^2\right) \\
&= \mathrm{tr}\left(\boldsymbol{H}\right) - \mathrm{tr}\left(\boldsymbol{H}^2\right) \\
&= M - \mathrm{tr}\left(\boldsymbol{H}^2\right), \\
\sum_{i\neq j} h_{ii}h_{ji} &= \boldsymbol{d}_H^T\boldsymbol{v}_1 - \mathrm{tr}\left(\boldsymbol{H}^2\right),
\end{aligned}$$

so that

$$\sum\left(p_6^{(2)}\left(\boldsymbol{H}\right)\right) = \sum \boldsymbol{H} - M - 2\boldsymbol{d}_H^T\boldsymbol{v}_1 + 2\mathrm{tr}\left(\boldsymbol{H}^2\right),$$

as required. Lastly, we have

$$\begin{aligned}
E_\Pi\left(p_6^{(2)}\left(\boldsymbol{G}\right)\right) &= E_\Pi\left(g_{ij}g_{kj}\right) \\
&= \frac{(N-3)!}{N!}\sum\left(p_6^{(2)}\left(\boldsymbol{G}\right)\right) \\
&= \frac{(N-3)!}{N!}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} g_{ij}g_{kj} - \sum_{i=1}^{N} g_{ii}^2 - \sum_{i\neq j} g_{ij}^2 - 2\sum_{i\neq j} g_{ii}g_{ji}\right),
\end{aligned}$$

where

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} g_{ij}g_{kj} &= 0, \\
\sum_{i=1}^{N} g_{ii}^2 &= \mathrm{tr}\left(\boldsymbol{G}^2\right), \\
\sum_{i\neq j} g_{ij}^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}^2 - \sum_{i=1}^{N} g_{ii}^2 \\
&= \mathrm{tr}\left(\boldsymbol{G}\boldsymbol{G}\right) - \mathrm{tr}\left(\boldsymbol{G}^2\right), \\
\sum_{i\neq j} g_{ii}g_{ji} &= \sum_{i=1}^{N}\sum_{j=1}^{N} g_{ii}g_{ji} - \sum_{i=1}^{N} g_{ii}^2 \\
&= -\mathrm{tr}\left(\boldsymbol{G}^2\right),
\end{aligned}$$

so that

$$E_{\Pi}\left(p_6^{(2)}\left(\boldsymbol{G}\right)\right) = \frac{(N-3)!}{N!}\left(2\mathrm{tr}\left(\boldsymbol{G}^2\right) - \mathrm{tr}\left(\boldsymbol{G}\boldsymbol{G}\right)\right),$$

as required.

Appendix G

Results of Candidate-Phenotype Multi-Locus GWA Studies of Alzheimer's Disease

Table G.1: Significant SNPs and genes identified for each genetic distance measure on using the GRV test with a sliding window of length 3 and familywise error and false positive rates controlled at 5%. The chromosome in which the SNPs were identified are given, in addition to the p-value of the sliding window containing each SNP. Where SNPs are present in more than one selected window, the minimum p-value of the windows is given. The columns B (for Bonferroni), BH (for Benjamini-Hochberg) and Q (for q-value) indicate with which p-value correction the SNPs were identified.

Distance measure	SNP	Gene	Chromosome	P-value of window	P-value correction		
					B	BH	Q
IBS	apoe4	APOE	19	1.131×10^{-9}	✓	✓	✓
	rs439401	APOE	19	1.131×10^{-9}	✓	✓	✓
	rs5167	APOC4	19	1.131×10^{-9}	✓	✓	✓
	rs405509	APOE	19	8.201×10^{-9}	✓	✓	✓
	rs157580	TOMM40	19	1.764×10^{-7}		✓	✓
	rs2075650	TOMM40	19	1.764×10^{-7}		✓	✓
	rs8106922	TOMM40	19	1.764×10^{-7}		✓	✓
Sokal and Sneath	rs157580	TOMM40	19	3.438×10^{-9}	✓	✓	✓
	rs2075650	TOMM40	19	3.438×10^{-9}	✓	✓	✓
	rs8106922	TOMM40	19	3.438×10^{-9}	✓	✓	✓
	apoe4	APOE	19	1.676×10^{-8}	✓	✓	✓
	rs439401	APOE	19	1.676×10^{-8}	✓	✓	✓
	rs5167	APOC4	19	1.676×10^{-8}	✓	✓	✓
	rs405509	APOE	19	1.769×10^{-7}		✓	✓
	rs9352023		6	2.575×10^{-7}		✓	
	rs7774274		6	2.575×10^{-7}		✓	
	rs9446996		6	2.575×10^{-7}		✓	
Rogers and Tanimoto I	rs157580	TOMM40	19	1.582×10^{-7}		✓	✓
	rs2075650	TOMM40	19	1.582×10^{-7}		✓	✓
	rs8106922	TOMM40	19	1.582×10^{-7}		✓	✓
	apoe4	APOE	19	1.947×10^{-7}		✓	✓
	rs439401	APOE	19	1.947×10^{-7}		✓	✓
	rs5167	APOC4	19	1.947×10^{-7}		✓	✓
	rs405509	APOE	19	2.130×10^{-7}		✓	✓

Table G.2: Significant SNPs and genes identified for each genetic distance measure on using the GRV test with a sliding window of length 5 and familywise error and false positive rates controlled at 5%. The chromosome in which the SNPs were identified are given, in addition to the p-value of the sliding window containing each SNP. Where SNPs are present in more than one selected window, the minimum p-value of the windows is given. The columns B (for Bonferroni), BH (for Benjamini-Hochberg) and Q (for q-value) indicate with which p-value correction the SNPs were identified.

Distance measure	SNP	Gene	Chromo-some	P-value of window	P-value correction		
					B	BH	Q
IBS	rs157580	TOMM40	19	1.272×10^{-9}	✓	✓	✓
	rs2075650	TOMM40	19	1.272×10^{-9}	✓	✓	✓
	rs8106922	TOMM40	19	1.272×10^{-9}	✓	✓	✓
	rs405509	APOE	19	1.272×10^{-9}	✓	✓	✓
	apoe4	APOE	19	1.272×10^{-9}	✓	✓	✓
	rs439401	APOE	19	1.372×10^{-9}	✓	✓	✓
	rs5167	APOC4	19	4.850×10^{-9}	✓	✓	✓
	rs10413089		19	4.850×10^{-9}	✓	✓	
	rs3760627	CLPTM1	19	1.867×10^{-7}		✓	
	rs12124893		1	3.936×10^{-7}		✓	
	rs2526839		1	3.936×10^{-7}		✓	
	rs6695214		1	3.936×10^{-7}		✓	
	rs7538876		1	3.936×10^{-7}		✓	
	rs1204897		1	3.936×10^{-7}		✓	
Sokal and Sneath	rs157580	TOMM40	19	3.615×10^{-10}	✓	✓	✓
	rs2075650	TOMM40	19	3.615×10^{-10}	✓	✓	✓
	rs8106922	TOMM40	19	3.615×10^{-10}	✓	✓	✓
	rs405509	APOE	19	3.615×10^{-10}	✓	✓	✓
	apoe4	APOE	19	3.615×10^{-10}	✓	✓	✓
	rs439401	APOE	19	2.928×10^{-9}	✓	✓	✓
	rs5167	APOC4	19	9.262×10^{-8}	✓	✓	✓
Rogers and Tanimoto I	rs157580	TOMM40	19	1.879×10^{-9}	✓	✓	✓
	rs2075650	TOMM40	19	1.879×10^{-9}	✓	✓	✓
	rs8106922	TOMM40	19	1.879×10^{-9}	✓	✓	✓
	rs405509	APOE	19	1.879×10^{-9}	✓	✓	✓
	apoe4	APOE	19	1.879×10^{-9}	✓	✓	✓
	rs439401	APOE	19	4.317×10^{-9}	✓	✓	✓
	rs5167	APOC4	19	1.948×10^{-7}		✓	✓

Table G.3: Significant SNPs and genes identified for each genetic distance measure on using the GRV test with a sliding window of length 7 and familywise error and false positive rates controlled at 5%. The chromosome in which the SNPs were identified are given, in addition to the p-value of the sliding window containing each SNP. Where SNPs are present in more than one selected window, the minimum p-value of the windows is given. The columns B (for Bonferroni), BH (for Benjamini-Hochberg) and Q (for q-value) indicate with which p-value correction the SNPs were identified.

Distance measure	SNP	Gene	Chromo-some	P-value of window	P-value correction		
					B	BH	Q
IBS	rs157580	TOMM40	19	4.276×10^{-10}	✓	✓	✓
	rs2075650	TOMM40	19	4.276×10^{-10}	✓	✓	✓
	rs8106922	TOMM40	19	4.276×10^{-10}	✓	✓	✓
	rs405509	APOE	19	4.276×10^{-10}	✓	✓	✓
	apoe4	APOE	19	4.276×10^{-10}	✓	✓	✓
	rs439401	APOE	19	4.276×10^{-10}	✓	✓	✓
	rs5167	APOC4	19	4.276×10^{-10}	✓	✓	✓
	rs10413089		19	7.581×10^{-10}	✓	✓	✓
	rs6859	PVRL2	19	3.127×10^{-9}	✓	✓	✓
	rs387976		19	4.206×10^{-9}	✓	✓	✓
	rs3760627	CLPTM1	19	1.881×10^{-7}		✓	
Sokal and Sneath	rs157580	TOMM40	19	5.948×10^{-10}	✓	✓	✓
	rs2075650	TOMM40	19	5.948×10^{-10}	✓	✓	✓
	rs8106922	TOMM40	19	5.948×10^{-10}	✓	✓	✓
	rs405509	APOE	19	5.948×10^{-10}	✓	✓	✓
	apoe4	APOE	19	5.948×10^{-10}	✓	✓	✓
	rs439401	APOE	19	5.948×10^{-10}	✓	✓	✓
	rs5167	APOC4	19	5.948×10^{-10}	✓	✓	✓
	rs387976		19	3.529×10^{-9}	✓	✓	✓
	rs6859	PVRL2	19	3.399×10^{-9}	✓	✓	✓
	rs10413089		19	2.100×10^{-8}	✓	✓	✓
Rogers and Tanimoto I	rs157580	TOMM40	19	1.813×10^{-9}	✓	✓	✓
	rs2075650	TOMM40	19	1.813×10^{-9}	✓	✓	✓
	rs8106922	TOMM40	19	1.813×10^{-9}	✓	✓	✓
	rs405509	APOE	19	1.813×10^{-9}	✓	✓	✓
	apoe4	APOE	19	1.813×10^{-9}	✓	✓	✓
	rs439401	APOE	19	1.813×10^{-9}	✓	✓	✓
	rs5167	APOC4	19	1.813×10^{-9}	✓	✓	✓
	rs10413089		19	3.035×10^{-9}	✓	✓	✓
	rs387976		19	3.610×10^{-9}	✓	✓	✓
	rs6859	PVRL2	19	3.610×10^{-9}	✓	✓	✓

References

Altshuler, D., Daly, M., and Lander, E. (2008). Genetic Mapping in Human Disease. *Science*, **322**(5903), 881–888.

Anderson, M. J. (2001). A New Method for Non-Parametric Multivariate Analysis of Variance. *Austral Ecology*, **26**(1), 32–46.

Angelini, C., De Canditiis, D., Mutarelli, M., and Pensky, M. (2007). A Bayesian Approach to Estimation and Testing in Time-Course Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **6**(1).

Arenas, C. and Cuadras, C. (2004). Comparing Two Methods for Joint Representation of Multivariate Data. *Communications in Statistics - Simulation and Computation*, **33**(2), 415–430.

Aryee, M. J., Gutirrez-Pabello, J. A., Kramnik, I., Maiti, T., and Quackenbush, J. (2009). An Improved Empirical Bayes Approach to Estimating Differential Gene Expression in microarray Time-Course Data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics*, **10**(1), 409–418.

Ayala, G., Gaston, M., Leon, T., and Mallor, F. (2008). Measuring Dissimilarity Between Curves by Means of Their Granulometric Size Distributions. *Functional and Operatorial Statistics*, page 35.

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D. K., and Jaakkola, T. (2003). Comparing the Continuous Representation of Time-Series Expression Profiles to Identify Differentially Expressed Genes. *Proceedings of the National Academy of Sciences*, **100**(18), 10146–10151.

Barton, D. E. and Dennis, K. E. (1952). The Conditions Under Which Gram-Charlier and Edgeworth Curves are Positive Definite and Unimodal. *Biometrika*, **39**(3/4), 425–427.

Beckmann, L., Thomas, D. C., Fischer, C., and Chang-Claude, J. (2005). Haplotype Sharing Analysis Using Mantel Statistics. *Human Heredity*, **59**(2), 67–78.

Behseta, S. and Kass, R. E. (2005). Testing Equality of Two Functions Using BARS. *Statistics in Medicine*, **24**(22), 3523–34.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Berk, M. and Montana, G. (2009). Functional Modeling of Microarray Time Series with Covariate Curves. *Statistica*, **2**(3), 153–177.

Berk, M., Hemingway, C., Levin, M., and Montana, G. (2012). *Advanced Statistical Methods for the Analysis of Large Data-Sets*, chapter Longitudinal Analysis of Gene Expression Profiles Using Functional Mixed-Effects Models, pages 56–67. Springer.

Berry, K. J. and Mielke, P. W. (1983). Moment Approximations as an Alternative to the F Test in Analysis of Variance. *British Journal of Mathematical and Statistical Psychology*, **36**(2), 202–206.

Bertram, L., Lill, C., and Tanzi, R. (2010). The Genetics of Alzheimer Disease: Back to the Future. *Neuron*, **68**(2), 270–281.

Bingham, N. H. and Fry, J. (2010). *Regression: Linear Models in Statistics*. Springer-Verlag London Limited r2010.

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling. Theory and Applications.* Springer Science+Business Media, Inc., second edition.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Braskie, M. N., Ringman, J. M., Thompson, P. M., *et al.* (2011). Neuroimaging Measures as Endophenotypes in Alzheimers Disease. *International Journal of Alzheimer's disease*, **2011**, 490140.

Bredesen, D. E. (2009). Neurodegeneration in Alzheimer's Disease: Caspases and Synaptic Element Interdependence. *Molecular Neurodegeneration*, **4**, 27.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, **16**(2), 101–117.

Brumell, J. H. and Scidmore, M. A. (2007). Manipulation of Rab GTPase Function by Intracellular Bacterial Pathogens. *Microbiology and Molecular Biology Reviews*, **71**(4), 636–652.

Bunke, H. and Shearer, K. (1998). A Graph Distance Metric Based on the Maximal Common Subgraph. *Pattern Recognition Letters*, **19**(3), 255–259.

Bunke, H., Foggia, P., Guidobaldi, C., Sansone, C., and Vento, M. (2002). A Comparison of Algorithms for Maximum Common Subgraph on Randomly Connected Graphs. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 85–106.

Burger, R. A., Sill, M. W., Monk, B. J., Greer, B. E., and Sorosky, J. I. (2007). Phase II Trial of Bevacizumab in Persistent or Recurrent Epithelial Ovarian Cancer or Primary Peritoneal Cancer: a Gynecologic Oncology Group Study. *Journal of Clinical Oncology*, **25**(33), 5165–5171.

Cerioli, A., Laurini, F., and Corbellini, A. (2003). Functional Cluster Analysis of Financial Time Series. In *Proceedings of the Meeting of Classification and Data Analysis Group of the Italian Statistical Society (CLADAG 2003)*, pages 107–110. Springer.

Cervantes, S., Samaranch, L., Vidal-Taboada, J. M., Lamet, I., Bullido, M. J., Frank-García, A., Coria, F., Lleó, A., Clarimón, J., Lorenzo, E., Alonso, E., Sánchez-Juan, P., Rodríguez-Rodríguez, Combarros, O., Rosich, M., Vilella, E., and Pastor, P. (2011). Genetic Variation in *APOE* Cluster Region and Alzheimer's Disease Risk. *Neurobiology of Aging*, **32**(11), 2107.e7–2107.e17.

Chapman-Rothe, N., Curry, E., Zeller, C., Liber, D., Stronach, E., Gabra, H., Ghaem-Maghami, S., and Brown, R. (2012). Chromatin H3K27me3/H3K4me3 Histone Marks Define Gene Sets in High-Grade Serous Ovarian Cancer that Distinguish Malignant, Tumour-Sustaining and Chemo-Resistant Ovarian Tumour Cells. *Oncogene*.

Coffey, N. and Hinde, J. (2011). Analyzing Time-Course Microarray Data Using Functional Data Analysis- A Review. *Statistical Applications in Genetics and Molecular Biology*, **10**.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping Complex Disease Traits with Global Gene Expression. *Nature Reviews Genetics*, **10**(3), 184–194.

Cuadras, C. (2008). Distance-Based Association and Multi-Sample Tests for General Multivariate Data. *Advances in Mathematical and Statistical Modeling*, pages 61–71.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An Anova Test for Functional Data. *Computational Statistics and Data Analysis*, **47**(1), 111–122.

DeCook, R., Lall, S., Nettleton, D., and Howell, S. H. (2006). Genetic Regulation of Gene Expression During Shoot Development in Arabidopsis. *Genetics*, **172**(2), 1155–1164.

Dempster, A. P. (1960). A Significance Test for the Separation of two Highly Multivariate Small Samples. *Biometrics*, **16**(1), 41–50.

Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP Kinase Signalling Pathways in Cancer. *Oncogene*, **26**(22), 3279–3290.

Dodge, Y. and Hadi, A. S. (1999). Simple Graphs and Bounds for the Elements of the Hat Matrix. *Journal of Applied Statistics*, **26**(7), 817–823.

Dow, M. M., Cheverud, J. M., and Friedlaender, J. S. (1987). Partial Correlation of Distance Matrices in Studies of Population Structure. *American Journal of Physical Anthropology*, **72**(3), 343–352.

Epifanio, I. (2008). Shape Descriptors for Classification of Functional Data. *Technometrics*, **50**(3), 284–294.

Erdős, P. and Rényi, A. (1960). On the Evolution of Random Graphs.

Escoufier, Y. (1973). Le Traitement des Variables Vectorielles. *Biometrics*, pages 751–760.

Fan, J. and Lin, S. K. (1998). Test of Significance When Data are Curves. *Journal of the American Statistical Association*, **93**(443), 1007–1021.

Farin, G. (1992). *Curves and Surfaces for Computer Aided Geometric Design*. Academic Press, third edition.

Fernández, M. L. and Valiente, G. (2001). A Graph Distance Metric Combining Maximum Common Subgraph and Minimum Common Supergraph. *Pattern Recognition Letters*, **22**(6-7), 753–758.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science+Business Media, Inc.

French, D., Lin, B., Wang, M., Adams, C., Shek, T., Hötzel, K., Bolon, B., Ferrando, R., Blackmore, C., Schroeder, K., Rodriguez, L. A., Hristopoulos, M., Venook, R., Ashkenazi, A., and Desnoyers, L. R. (2012). Targeting FGFR4 Inhibits Hepatocellular Carcinoma in Preclinical Mouse Models. *PloS One*, **7**(5), e36713.

Fujikoshi, Y., Himeno, T., and Wakaki, H. (2004). Asymptotic Results of a High Dimensional MANOVA Test and Power Comparison When the Dimension is Large Compared to the Sample Size. *Journal of the Japan Statistical Society*, **34**(1), 19–26.

Gandy, A. and Rubin-Delanchy, P. (2011). An Algorithm to Compute the Power of Monte Carlo Tests with Guaranteed Precision. *Arxiv preprint arXiv:1110.1248*.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J., Berg, S., Fiske, A., and Pedersen, N. (2006). Role of Genes and Environments for Explaining Alzheimer Disease. *Archives of General Psychiatry*, **63**(2), 168–174.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. (2012). Increasing Power for Voxel-Wise Genome-Wide Association Studies: The Random Field Theory, Least Square Kernel Machines and Fast Permutation Procedures. *NeuroImage*, pages 858–873.

Gibson, G. and Muse, S. V. (2004). *A Primer of Genome Science*. Sinauer Associates, Inc., second edition.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*, **53**(3-4), 325–338.

Gower, J. C. (1971). Statistical Methods of Comparing Different Multivariate Analyses of the Same Data. In *Mathematics in the Archaeological and Historical Sciences* (F. R. Hodson *et al.*, eds). pages 138–149.

Gower, J. C. and Krzanowski, W. J. (1999). Analysis of Distance for Structured Multivariate Data and Extensions to Multivariate Analysis of Variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**(4), 505–519.

Halima, B., Buddana, A., *et al.* (2005). Functional Clustering Algorithm for High-Dimensional Proteomics Data. *Journal of Biomedicine and Biotechnology*, (2), 80–86.

Hall, P. and Van Keilegom, I. (2007). Two-sample Tests in Functional Data Analysis Starting from Discrete Data. *Statistica Sinica*, **17**(4), 1511.

Hamming, R. (1950). Error Detecting and Error Correcting Codes. *The Bell System Technical Journal*, **29**(2), 147–160.

Heckman, N. and Zamar, R. (2000). Comparing the Shapes of Regression Functions. *Biometrika*, **87**(1), 135.

Heywood, J. S. (1991). Spatial Analysis of Genetic Variation in Plant Populations. *Annual Review of Ecology and Systematics*, **22**, 335–355.

Hibar, D., Kohannim, O., Stein, J., Chiang, M., and Thompson, P. (2011). Multilocus Genetic Analysis of Brain Images. *Frontiers in genetics*, **2**.

Hoaglin, D. and Welsch, R. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*, **32**(1), 17–22.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*, volume 21. Wiley Online Library.

Hodge, D. R., Hurt, E. M., and Farrar, W. L. (2005). The role of il-6 and stat3 in inflammation and cancer. *European Journal of Cancer*, **41**(16), 2502–2512.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, **19**(3), 293–325.

Holden, M., Hill, D., Denton, E., Jarosz, J., Cox, T., Rohlfing, T., Goodey, J., and Hawkes, D. (2000). Voxel Similarity Measures for 3-D Serial MR Brain Image Registration. *IEEE Transactions on Medical Imaging*, **19**(2), 94–102.

Hong, F. and Li, H. (2006). Functional Hierarchical Models for Identifying Genes with Different Time-Course Expression Profiles. *Biometrics*, **62**(2), 534–544.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, **28**(3/4), 321–377.

Jackson, D. (1995). PROTEST: A PROcrustean Randomization TEST of Community Environment Concordance. *Écoscience*, **2**(3), 297–303.

Johnson, N. (1949). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, **36**(1/2), 149–176.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., second edition.

Josse, J., Pagès, J., and Husson, F. (2008). Testing the Significance of the RV Coefficient. *Computational Statistics and Data Analysis*, **53**(1), 82–91.

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J. D. (1995). Refined Approximations to Permutation Tests for Multivariate Inference. *Computational Statistics and Data Analysis*, **20**(6), 643–656.

Kiers, H. A. L. (2002). Setting up Alternating Least Squares and Iterative Majorization Algorithms for Solving Various Matrix Optimization Problems. *Computational Statistics & Data Analysis*, **41**(1), 157–170.

Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., and Shmulevich, I. (2009). Fewer Permutations, More Accurate P-values. *Bioinformatics*, **25**(12), i161–i168.

Kourouklis, S. and Moschopoulos, P. (1985). On the Distribution of the Trace of a Noncentral Wishart Matrix. *Metron*, **43**(2), 85–92.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis: a User's Perspective*. Oxford University Press.

Lasserre, J. (1995). A Trace Inequality for Matrix Product. *IEEE Transactions on Automatic Control*, **40**(8), 1500–1501.

Legendre, P. and Anderson, M. J. (1999). Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments. *Ecological Monographs*, **69**(1), 1–24.

Legendre, P. and Fortin, M. (2010). Comparison of the Mantel Test and Alternative Approaches for Detecting Complex Multivariate Relationships in the Spatial Analysis of Genetic Data. *Molecular Ecology Resources*, **10**(5), 831–844.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology*, volume 20. Elsevier, second edition.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics-Doklady*, **10**, 707–710.

Li, Q., Wacholder, S., Hunter, D. J., Hoover, R. N., Chanock, S., Thomas, G., and Yu, K. (2009). Genetic Background Comparison Using Distance-Based Regression, With Applications in Population Stratification Evaluation and Adjustment. *Genetic Epidemiology*, **33**(5), 432–441.

Li, S., Lu, Q., and Cui, Y. (2010). A Systems Biology Approach for Identifying Novel Pathway Regulators in eQTL Mapping. *Journal of Biopharmaceutical Statistics*, **20**(2), 373–400.

Liongue, C., O'Sullivan, L. A., Trengove, M. C., and Ward, A. C. (2012). Evolution of JAK-STAT Pathway Components: Mechanisms and Role in Immune System Development. *PLoS One*, **7**(3), e32777.

Liu, X. and Yang, M. C. K. (2009). Identifying Temporally Differentially Expressed Genes Through Functional Principal Components Analysis. *Biostatistics*, **10**(4), 667–679.

Lord, L. D., Allen, P., Expert, P., Howes, O., Lambiotte, R., McGuire, P., Bose, S. K., Hyde, S., and Turkheimer, F. (2011). Characterization of the Anterior Cingulate's Role in the At-Risk Mental State Using Graph Theory. *NeuroImage*, **56**, 1531–1539.

Lourenco, F. C., Galvan, V., Fombonne, J., Corset, V., Llambi, F., Müller, U., Bredesen, D. E., and Mehlen, P. (2009). Netrin-1 Interacts with Amyloid Precursor Protein and Regulates Amyloid-β Production. *Cell Death and Differentiation*, **16**(5), 655–663.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, **27**(2), 209–220.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Academic Press.

Marron, J. S. and Tsybakov, A. B. (1995). Visual Error Criteria for Qualitative Smoothing. *Journal of the American Statistical Association*, **90**(430), 499–507.

Mathias, R., Gao, P., Goldstein, J. L., Wilson, A. F., Pugh, E. W., Furbert-Harris, P., Dunston, G. M., Malveaux, F. J., Togias, A., Barnes, K. C., Beaty, T. H., and K, H. S. (2006). A Graphical Assessment of P-values from Sliding Window Haplotype Tests of Association to Identify Asthma Susceptibility Loci on Chromosome 11q. *BMC genetics*, **7**(1), 38.

McArdle, B. and Anderson, M. (2001). Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. *Ecology*, **82**(1), 290–297.

McClurg, P., Pletcher, M. T., Wiltshire, T., and Su, A. I. (2006). Comparative Analysis of Haplotype Association Mapping Algorithms. *BMC Bioinformatics*, **7**(1), 61.

Michaels, G., Carr, D., Askenazi, M., Fuhrman, S., Wen, X., and Somogyi, R. (1998). Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. In *Pacific Symposium on Biocomputing*, volume 3, pages 42–53.

Mielke, P. and Berry, K. (2007). *Permutation Methods: A Distance Function Approach*. Springer Science+Business Media, LLC.

Mielke, P., Berry, K., and Johnson, E. (1976). Multi-Response Permutation Procedures for a Priori Classifications. *Communications in Statistics-Theory and Methods*, **5**(14), 1409–1424.

Minas, C., Waddell, S. J., and Montana, G. (2011). Distance-Based Differential Analysis of Gene Curves. *Bioinformatics*, **27**(22), 3135–3141.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine learning*, **52**(1), 91–118.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004). Genetic Analysis of Genome-Wide Variation in Human Gene Expression. *Nature*, (7001), 743–747.

Mukhopadhyay, I., Feingold, E., Weeks, D. E., and Thalamuthu, A. (2010). Association Tests Using Kernel-Based Measures of Multi-Locus Genotype Similarity Between Individuals. *Genetic Epidemiology*, **34**(3), 213–221.

Neumeyer, N. and Dette, H. (2003). Nonparametric Comparison of Regression Curves: An Empirical Process Approach. *Annals of Statistics*, **31**, 880–920.

Newton, M. (2009). Introducing the Discussion Paper by Székely and Rizzo. *The Annals of Applied Statistics*, **3**(4), 1233–1235.

Pardo-Fernández, J. C., Van Keilegom, I., and González-Manteiga, W. (2007). Testing for the Equality of k Regression Curves. *Statistica Sinica*, **17**(3), 1115.

Park, C. and Kang, K. (2008). SiZer Analysis for the Comparison of Regression Curves. *Computational Statistics and Data Analysis*, **52**, 3954–3970.

Parui, D. and Majumder, D. (1983). Shape Similarity Measures for Open Curves. *Pattern Recognition Letters*, **1**(3), 129–134.

Pearson, E. (1963). Some Problems Arising in Approximating to Probability Distributions, Using Moments. *Biometrika*, **50**(1/2), 95–112.

Pearson, K. (1895). Contributions to the mathematical theory of evolution- ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, **186**, 343–414.

Pearson, K. (1901). Mathematical Contributions to the Theory of Evolution.- X. Supplement to a Memoir on Skew Variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **197**(287-299), 443–459.

Pearson, T. and Manolio, T. (2008). How to Interpret a Genome-Wide Association Study. *The Journal of the American Medical Association*, **299**(11), 1335–1344.

Pekalska, E. and Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific Co. Pte. Ltd.

Peres-Neto, P. R. and Jackson, D. A. (2001). How Well do Multivariate Data Sets Match? The Advantages of a Procrustean Superimposition Approach Over the Mantel Test. *Oecologia*, **129**(2), 169–178.

Phipson, B. and Smyth, G. K. (2010). Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology*, **9**, 39.

Price, A., Zaitlen, N., Reich, D., and Patterson, N. (2010). New Approaches to Population Stratification in Genome-Wide Association Studies. *Nature Reviews Genetics*, **11**(7), 459–463.

Priness, I., Maimon, O., and Ben-Gal, I. (2007). Evaluation of Gene-Expression Clustering via Mutual Information Distance Measure. *BMC Bioinformatics*, **8**(1), 111–123.

Qian, J., Filhart, M. D., Yu, J., Haiyuan, L., and Gerstein, M. (2001). Beyond Synexpression Relationships: Local Clustering of Time-Shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. *Journal of Molecular Biology*, **314**(5), 1053–1066.

Quigley, D. A., To, M. D., Kim, I. J., Lin, K. K., Albertson, D. G., Sjolund, J., Pérez-Losada, J., and Balmain, A. (2011). Network Analysis of Skin Tumor Progression Identifies a Rewired Genetic Architecture Affecting Inflammation and Tumor Susceptibility. *Genome Biology*, **12**(1), R5.

Ramsay, J. O. and Silverman, B. W. (2006). *Functional Data Analysis*. Springer Science+Business Media, LLC, second edition.

Rand, V., Huang, J., Stockwell, T., Ferriera, S., Buzko, O., Levy, S., Busam, D., Li, K., Edwards, J. B., Eberhart, C., Murphy, K. M., Tsiamouri, A., Beeson, K., Simpson, A. J. G., , Venter, J. C., Riggins, G. J., , and Strausberg, R. L. (2005). Sequence Survey of Receptor Tyrosine Kinases Reveals Mutations in Glioblastomas. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(40), 14344–14349.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures. *Bioinformatics*, **19**(3), 368–375.

Reiss, P. T., Stevens, M. H. H., Shehzad, Z., Petkova, E., and Milham, M. P. (2009). On Distance-Based Permutation Tests for Between-Group Comparisons. *Biometrics*.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons, Inc., second edition.

Robert, P. and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Applied Statistics*, pages 257–265.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of Global Gene Expression. *Nature Reviews Genetics*, **7**(11), 862–872.

Rubinov, M. and Sporns, O. (2010). Complex Network Measures of Brain Connectivity: Uses and Interpretations. *NeuroImage*, **52**(3), 1059–1069.

Salem, R. M., O'Connor, D. T., and Schork, N. J. (2010). Curve-Based Multivariate Distance Matrix Regression Analysis: Application to Genetic Association Analyses Involving Repeated Measures. *Physiol. Genomics*, **42**(2), 236–247.

Schneider, J. and Borlund, P. (2007). Matrix Comparison, Part 2: Measuring the Resemblance Between Proximity Measures or Ordination Results by Use of the Mantel and Procrustes Statistics. *Journal of the American Society for Information Science and Technology*, **58**(11), 1596–1609.

Schork, N. J. and Zapala, M. A. (2012). Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis. *Frontiers in Genetics*, **3**(190), 1–10.

Schott, J. R. (2007). Some High-Dimensional Tests for a One-Way MANOVA. *Journal of Multivariate Analysis*, **98**(9), 1825–1839.

Selinski, S. and Ickstadt, K. (2005). Similarity Measures for Clustering SNP Data. *Technical Report / Universitt Dortmund, SFB 475 Komplexittsreduktion in Multivariaten Datenstrukturen*.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., Foroud, T., Pankratz, N., Moore, J. H., Sloan, C. D., Huentelman, M. J., Craig, D. W., DeChairo, B. M., Potkin, S. G., Jack Jr, C. R., Weiner, M. W., Saykin, A. J., and the Alzheimer's Disease Neuroimaging Initiative (2010). Whole Genome Association Study of Brain-Wide Imaging Phenotypes for Identifying Quantitative Trait Loci in MCI and AD: A Study of the ADNI Cohort. *NeuroImage*, **53**(3), 1051–1063.

Shen, Q. and Faraway, J. (2004). An F Test for Linear Models with Functional Responses. *Statistica Sinica*, **14**(4), 1239–1258.

Shen, Y., Lin, Z., and Zhu, J. (2011). Shrinkage-Based Regularization Tests for High-Dimensional Data with Application to Gene Set Analysis. *Computational Statistics and Data Analysis*, **55**, 2221–2233.

Shepherd, T. G., Mujoomdar, M. L., and Nachtigal, M. W. (2010). Constitutive Activation of BMP Signalling Abrogates Experimental Metastasis of OVCA429 Cells via Reduced Cell Adhesion. *Journal of Ovarian Research*, **3**(1), 1–14.

Shi, Y., Mitchell, T., and Bar-Joseph, Z. (2007). Inferring Pairwise Regulatory Relationships from Multiple Time Series Datasets. *Bioinformatics*, **23**(6), 755–763.

Shuck, S. C., Short, E. A., and Turchi, J. J. (2008). Eukaryotic Nucleotide Excision Repair: From Understanding Mechanisms to Influencing Biology. *Cell Research*, **18**(1), 64–72.

Silver, M., Janousova, E., Hua, X., Thompson, P. M., and Montana, G. (2012). Identification of Gene Pathways Implicated in Alzheimer's Disease Using Longitudinal Imaging Phenotypes with Sparse Regression.

Solomon, H. and Stephens, M. (1977). Distribution of a Sum of Weighted Chi-Square Variables. *Journal of the American Statistical Association*, **72**(360), 881–885.

Solomon, H. and Stephens, M. (1978). Approximations to Density Functions Using Pearson Curves. *Journal of the American Statistical Association*, **73**(361), 153–160.

Springer, M. D. (1979). *The Algebra of Random Variables*. Wiley New York.

Srivastava, M. S. (2007). Multivariate Theory for Analyzing High Dimensional Data. *Journal of the Japan Statistical Society*, **37**(1), 53–86.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A Robust Bayesian Two-Sample Test for Detecting Intervals of Differential Gene Expression in Microarray Time Series. *Journal of Computational Biology*, **17**(3), 355–367.

Storey, J. and Tibshirani, R. (2003). Statistical Significance for Genomewide Studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9441–9445.

Storey, J. D., Akey, J. M., and Kruglyak, L. (2005a). Multiple Locus Linkage Analysis of Genomewide Expression in Yeast. *PLoS Biology*, **3**(8), e267.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005b). Significance Analysis of Time Course Microarray Experiments. *Proceedings of the National Academy of Sciences*, **102**(36), 12837–12842.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S. E., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2005). Genome-Wide Associations of Gene Expression Variation in Humans. *PLoS Genetics*, **1**(e78), 695–704.

Stronach, E. A., Alfraidi, A., Rama, N., Datler, C., Studd, J. B., Agarwal, R., Guney, T. G., Gourley, C., Hennessy, B. T., Mills, G. B., Mai, A., Brown, R., Dina, R., and H., G. (2011). HDAC4-Regulated STAT1 Activation Mediates Platinum Resistance in Ovarian Cancer. *Cancer research*, **71**(13), 4412–4422.

Szabo, A., Boucher, K., Jones, D., Tsodikov, A. D., Klebanov, L. B., and Yakovlev, A. Y. (2003). Multivariate Exploratory Tools for Microarray Data Analysis. *Biostatistics*, **4**(4), 555–567.

Székely, G. J. and Rizzo, M. L. (2009). Brownian Distance Covariance. *The Annals of Applied Statistics*, **3**(4), 1236–1265.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, **35**(6), 2769–2794.

Tai, Y. and Speed, T. (2005). Statistical Analysis of Microarray Time Course Data. *DNA Microarrays*, pages 257–279.

Tailleux, L., Waddell, S. J., Pelizzola, M., Mortellaro, A., Withers, M., Tanne, A., Castagnoli, P. R., Gicquel, B., Stoker, N. G., Butcher, P. D., Foti, M., and Neyrolles, O. (2008). Probing Host Pathogen Cross-Talk by Transcriptional Profiling of Both *Mycobacterium tuberculosis* and Infected Human Dendritic Cells and Macrophages. *PLoS ONE*, **3**(1), e1403.

Takei, N., Miyashita, A., Tsukie, T., Arai, H., Asada, T., Imagawa, M., Shoji, M., Higuchi, S., Urakami, K., Kimura, H., Kakita, A., Takahashi, H., Tsuji, S., Kanazawa, I., Ihara, Y., Odani, S., Kuwano, R., and the Japanese Genetic Study

Consortium for Alzheimer Disease (2009). Genetic Association Study in and Around the APOE in Late-Onset Alzheimer Disease in Japanese. *Genomics*, **93**(5), 441–448.

Thomas, D. C. and Witte, J. S. (2002). Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? *Cancer Epidemiology Biomarkers & Prevention*, **11**(6), 505–512.

Torgerson, W. S. (1952). Multidimensional Scaling: I. Theory and Method. *Psychometrika*, **17**(4), 401–419.

Trinh, X. B., Tjalma, W. A. A., Vermeulen, P. B., Van den Eynden, G., Van der Auwera, I., Van Laere, S. J., Helleman, J., Berns, E. M. J. J., Dirix, L. Y., and van Dam, P. A. (2009). The VEGF Pathway and the AKT/mTOR/p70S6K1 Signalling Pathway in Human Epithelial Ovarian Cancer. *British Journal of Cancer*, **100**(6), 971–978.

Tsai, C.-A. and Chen, J. J. (2009). Multivariate Analysis of Variance Test for Gene Set Analysis. *Bioinformatics*, **25**(7), 897–903.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer Science+Business Media, LLC, fourth edition.

Vounou, M., Nichols, T. E., Montana, G., and the Alzheimer's Disease Neuroimaging Initiative (2010). Discovering Genetic Associations with High-Dimensional Neuroimaging Phenotypes: A Sparse Reduced-Rank Regression Approach. *NeuroImage*, **53**(3), 1147–1159.

Wallace, D. (1958). Asymptotic Approximations to Distributions. *The Annals of Mathematical Statistics*, **29**(3), 635–654.

Weisberg, S. (1985). *Applied Linear Regression*. John Wiley & Sons, Inc., second edition.

Wessel, J. and Schork, N. J. (2006). Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis. *The American Journal of Human Genetics*, **79**(5), 792–806.

Wessel, J., Zapala, M. A., and Schork, N. J. (2007). Accommodating Pathway Information in Expression Quantitative Trait Locus Analysis. *Genomics*, **90**(1), 132–142.

Wu, C., Delano, D. L., Mitro, N., Su, S., Janes, J., McClurg, P., Batalov, S., Welch, G. L., Zhang, J., Orth, A. P., Walker, J. R., Glynne, R. J., Cooke, M. P., Takahashi, J. S., Shimomura, K., Kohsaka, A., Bass, J., Saez, E., Wiltshire, T., and Su, A. I. (2008). Gene Set Enrichment in eQTL Data Identifies Novel Annotations and Pathway Regulators. *PLoS Genetics*, **4**(5), e1000070.

Wu, H. and Zhang, J. T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed Effects Modeling Approaches*. John Wiley & Sons, Inc.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D., and Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. *The American Journal of Human Genetics*, **86**(6), 929–942.

Xiang, S., Nie, F., and Zhang, C. (2008). Learning a Mahalanobis Distance Metric for Data Clustering and Classification. *Pattern Recognition*, **41**(12), 3600–3612.

Xiong, H. and Chen, X. (2006). Kernel-based distance metric learning for microarray data classification. *BMC Bioinformatics*, **7**(299).

Yang, H. C., Liang, Y. J., Wu, Y. L., Chung, C. M., Chiang, K. M., Ho, H. Y., Ting, C. T., Lin, T. H., Sheu, S. H., Tsai, W. C., Chen, J. H., Leu, H. B., Yin, W. H., Chiu, T. Y., Chen, C. I., Fann, C. S. J., Wu, J. Y., Lin, T. N., Lin, S. J., Chen, Y. T., Chen, J. W., and Pan, W. H. (2009). Genome-Wide Association Study of Young-Onset Hypertension in the Han Chinese Population of Taiwan. *PLoS One*, **4**(5), e5459.

Ying, Y. and Li, P. (2012). Distance Metric Learning with Eigenvalue Optimization. *The Journal of Machine Learning Research*, **13**, 1–26.

Yu, C. E., Seltman, H., Peskind, E. R., Galloway, N., Zhou, P. X., Rosenthal, E., Wijsman, E. M., Tsuang, D. W., Devlin, B., and Schellenberg, G. D. (2007). Comprehensive Analysis of APOE and Selected Proximate Markers for Late-Onset Alzheimer's Disease: Patterns of Linkage Disequilibrium and Disease/Marker Association. *Genomics*, **89**(6), 655–665.

Yu, D. and Hung, M. C. (2000). Overexpression of ErbB2 in Cancer and ErbB2-Targeting Strategies. *Oncogene*, **19**(53), 6115.

Zapala, M. A. and Schork, N. J. (2006). Multivariate Regression Analysis of Distance Matrices for Testing Associations Between Gene Expression Patterns and Related Variables. *Proceedings of the National Academy of Sciences*, **103**(51), 19430–19435.

Zhang, C., Peng, H., and Zhang, J. (2010). Two Samples Tests for Functional Data. *Communications in Statistics - Theory and Methods*, **39**(4), 559–578.