Dimension Reduction and High-Dimensional Data: Estimation and Inference with Application to Genomics and Neuroimaging

Maxime Turgeon

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec
April 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

This thesis is dedicated to my loving wife, Geneviève, and our family.

Acknowledgements

First and foremost, I would like to thank my co-supervisors, Dr. Celia Greenwood and Dr. Aurélie Labbe. As Aurélie told me at our very first meeting, a good supervisory team is more important than a good research topic. Their support, patience and incredible availability were key ingredients in making this thesis possible. Celia's vast knowledge of genomics and statistical genetics, combined with Aurélie's attention to details, made them an excellent supervisory team, better than any PhD candidate could hope for. As I continue to grow as a researcher, their thoughtful comments and advice will keep shaping my career and make me a better statistician and mentor to my own students.

I want to acknowledge the financial support I received from the Fonds de recherche Nature et Technologies du Québec, the Ludmer Centre for Neuroinformatics and Mental Health, and the Faculty of Medicine, the latter of which funded a one-week research stay at Stanford University. I also want to thank Celia for financially supporting my travelling to conferences. I also want to acknowledge the use of Calcul Québec's scientific computing services, along with the computing facility at the Lady Davis Institute for Medical Research. All the simulations and data analyses presented in this thesis were performed on one of these two computing platforms.

I want to thank the Department of Epidemiology, Biostatistics and Occupational Health for believing that someone with a background in arithmetic geometry could be a successful PhD candidate. To Dr. James Hanley, Dr. Andrea Benedetti, Dr. Olli Saarela, Dr. Erica Moodie, you have taught me so much and made me a better statistician. The strong (bio)statistical foundation I now have is due to your dedication and that of your colleagues from the Department of Statistics

To my brothers in arms, Sahir Bhatnagar and Kevin McGregor, I am grateful for all the conversations we have had over the years and for indulging my constant sharing of statistical

blog posts. I have learned tremendously from your insistence on details, tough questions, and clarity.

Finally, this thesis would not have been possible without the constant support of my wife, Geneviève, and our family. To them, I am eternally grateful.

Preface & Contribution of Authors

This thesis contains new research in addition to an introductory overview of methodological approaches to dimension reduction of high-dimensional datasets.

Chapters 1 and 2 contain an original introduction and summary of selected methodological approaches to dimension reduction of high-dimensional datasets. These chapters were written entirely by Maxime Turgeon (MT). They were both edited and corrected by Celia MT Greenwood (CMTG) and Aurélie Labbe (AL).

The idea for the study in Chapter 3 was conceived by Karim Oualkacha (KO) and AL. The proof of the main theorem was written by MT, and all methodological work, writing, programming, and analyses were carried out by MT with CMTG and AL as advisors, editors (also contributing small amounts of content directly) and troubleshooters.

Chapter 4 was conceptualized by MT. MT carried out the methodological work, writing, programming, and analysis. AL and CMTG served as advisors and editors. Helpful discussions with Stepan Grinek (SG) also helped shape early versions of the manuscript.

For Chapter 5, MT and CMTG contributed to the conceptualization of the data analysis; Kathleen Klein (KK) contributed to the data curation; MT contributed to the formal analysis and the writing of the original draft; all authors contributed to editing the draft.

Chapter 6, which concludes the thesis, was written by MT and edited by CMTG and AL.

Abstract

Recent technological advances in many domains including both genomics and brain imaging have led to an abundance of high-dimensional and correlated data being routinely collected. A widespread analytical goal in these fields is to investigate the relationships between, on the one hand, a group of genomic markers or anatomical brain measurements and, on the other hand, a set of clinical variables or phenotypes. To leverage the correlation within each set of measurements, and to improve the interpretability of a measure of the association, one can use *dimension reduction* techniques: one, or both, group of variables can be summarised by a small set of latent features that summarise the structure of interest and capture association through an appropriately chosen statistic. But the high-dimensionality of contemporary datasets brings many computational and theoretical challenges, and most classical multivariate methods cannot be used directly.

This thesis is comprised primarily of three manuscripts that investigate the issues related to measuring association in high dimensional datasets. In the first manuscript, I explore the optimality properties of a dimension reduction method known as *Principal Component of Explained Variance* (PCEV). This method seeks a linear combination of the outcome variables that maximises the proportion of variance explained by a set of covariates of interest. I then explain how PCEV can be extended to a computationally simple and efficient estimation strategy for high-dimensional outcomes ($p > n$) that relies on a "block-independence" assumption. In the second manuscript, I study the problem of inference with high-dimensional datasets: given two datasets Y and X, with one or both being high-dimensional, how can we perform a test of association in a computationally efficient way? Specifically, I look at the set of multivariate methods that can be described as a *double Wishart problem*; PCEV, Canonical Correlation Analysis (CCA), and Multivariate Analysis of Variance (MANOVA) are all examples of double Wishart problems. I show that valid high-dimensional p-values can be derived using an empirical estimator of the null distribution. This is achieved by performing

a small number of permutations, and then fitting a location-scale family of the Tracy-Widom distribution of order 1 to the test statistics computed from the permuted data. Finally, in the third manuscript, I apply the concepts developed in the two other manuscripts to a data analysis of targeted custom capture bisulfite methylation data. I show how PCEV can be used in conjunction with the ideas in the second manuscript to test for a region-level association between the methylation levels of CpG dinucleotides and levels of anti-citrullinated protein antibody (ACPA), an antigen thought to be a predictor of rheumatoid arthritis onset. In this study, the CpG dinucleotides are naturally grouped by design, and several of these groups contain a number of methylation measurements that is larger than the sample size.

Abrégé

Les avancées technologiques récentes dans plusieurs domaines, dont la génomique et la neuro-imagerie, ont contribué à une abondance de données de grande dimension et corrélées. Un objectif analytique commun à ces domaines est l'étude des relations entre, d'une part, un groupe de marqueurs génomiques ou des mesures anatomiques du cerveau, et d'autre part, un ensemble de variables cliniques ou de phénotypes. Pour mettre à profit la corrélation entre ces ensembles de mesures, et pour améliorer l'interprétabilité des mesures d'association, on peut utiliser des méthodes de *réduction dimensionnelle*: un groupe de variables (ou les deux) peut être représenté par un petit ensemble de variables latentes qui récapitule la structure d'intérêt et capture l'association via une statistique choisie. Or, la grande dimension des jeux de données contemporains amène de nombreux défis computationnels et théoriques, et la majorité des techniques multivariées classiques ne peuvent être utilisées directement.

Cette thèse est composée principalement de trois manuscrits qui étudient les enjeux reliés aux mesures d'association entre jeux de données de grande dimension. Dans le premier manuscrit, j'explore les propriétés optimales d'une méthode de réduction dimensionnelle connue sous le nom de *Composante Principale de Variance Expliquée* (PCEV). Cette méthode recherche la combinaison linéaire des variables réponses qui maximise la proportion de la variance qui peut être expliquée par les covariables d'intérêt. Ensuite, j'explique comment PCEV peut être généralisé à une stratégie d'estimation efficace et computationellement simple pour les données de grande dimension ($p > n$). Cette généralisation repose sur une hypothèse d'indépendance "par bloc". Dans le second manuscrit, j'étudie le problème d'inférence pour les jeux de données de grande dimension: étant donné deux jeux de données Y et X potentiellement de grande dimension, comment peut-on obtenir un test d'association de manière computationnellement efficace? Plus précisément, le manuscrit porte sur l'ensemble des méthodes multivariées pouvant être décrites comme un *problème Wishart double*; PCEV, l'Analyse en Corrélations Canoniques (CCA), et l'Analyse de Vari-

ance Multivarié (MANOVA) sont tous des exemples de problèmes Wishart double. Je montre comment des valeurs-p valides et de grande dimension peuvent être obtenues en utilisant un estimateur empirique de la distribution nulle. Cet estimateur peut être construit en commençant par un petit nombre de permutations, et ensuite en calibrant une famille position-échelle de la distribution Tracy-Widom d'ordre 1 en utilisant les statistiques de test calculées sur les données permutées. Finalement, dans le troisième manuscrit, j'utilise les concepts développés dans les deux premiers manuscrits pour analyser des données de méthylation obtenues par séquençage ciblé. Je démontre comment PCEV peut être combiné aux idées développées dans le second manuscrit pour tester l'association entre les niveaux de méthylation de l'ADN et les niveaux de l'anticorps pour les protéines anti-citrullinées (ACPA), un antigène soupçonné d'être un prédicteur de la polyarthrite rhumatoïde, au niveau des régions génomiques. Dans cette étude, les dinucléotides CpG sont naturellement regroupés, et plusieurs de ces groupes comportent plus de mesures de méthylation que d'échantillons.

Table of contents

List of Tables

List of Figures

Abbreviations

ACPA anti-citrullinated protein antibody

AD Alzheimer's disease

ADNI Alzheimer's Disease Neuroimaging Initiative

Aβ amyloid beta

ARCTIC Assessment of Risk for Colorectal Tumors in Canada

CCA Canonical Correlation Analysis

CDF Cumulative Distribution Function

CRAN Comprehensive R Archive Network

DMR Differentially Methylated Region

eQTL expression Quantitative Trait Loci

EVD Eigenvalue Decomposition

GLRM Generalized Low-Rank Models

ICA Independent Component Analysis

KEGG Kyoto Encyclopedia of Genes and Genomes

LDA Linear Discriminant Analysis

LRT Likelihood Ratio Test

MANOVA Multivariate Analysis of Variance

meQTL methylation Quantitative Trait Loci

MLE Maximum Likelihood Estimation

MRI Magnetic Resonance Imaging

NHST Null Hypothesis Significance Testing

NIPALS Nonlinear Iterative Partial Least Squares

PCA Principal Component Analysis

PCEV Principal Component of Explained Variance

PCH Principal Component of Heritability

PCR Principal Component Regression

PET Positron Emission Tomography

PLS Partial Least Squares

PMD Penalized Matrix Decomposition

RA Rheumatoid Arthritis

RDA Redundancy Analysis

SAVE Sliced Average Variance Estimate

SCoTLASS Simplified Component Technique-Lasso

SDR Sufficient Dimension Reduction

SIR Sliced Inverse Regression

SNP Single Nucleotide Polymorphism

SVD Singular Value Decomposition

UIT Union-Intersection Test

VIF Variable Importance Factor

VIP Variable Importance on Projection

Chapter 1

Introduction

Technological advances over the last decades have created a world drowning in data. As part of the "big data" revolution, businesses and technology companies are collecting vast amounts of transactional data on their customers and employees. By gathering as much information as possible and building data science teams to make sense of it, these companies hope to gain important insights into their business processes that can translate into a competitive advantage. This arms race was encapsulated in an October 2012 article by the *Harvard Business Review* that called data science the "sexiest job of the 21st century".

A similar data revolution has occurred in biological sciences. This revolution was spearheaded by the development of high-throughput technologies. Examples of these technologies include microarrays and next-generation sequencing in genomics, and imaging technologies in neuroscience. Due to these technological advancements, we are now able to collect vast amounts of measurements for each experimental unit. Similarly, researchers hope that this vast amount of biological and molecular data can be leveraged to further our understanding of the aetiology of complex diseases

Unlike information systems that easily record transactional data in real-time and at a low cost, the technologies used in biological sciences still have significant costs attached to them.

As a result, a typical study will often limit data collection to a few hundred individuals, with relatively few studies having larger sample sizes. This has led to the well-known "large p– small n" problem; that is, the number of measurements can be much larger than the number of experimental units. In statistics, this is generally referred to as *high-dimensional data* problems, and these problems have generated a lot of interest and research over the last decades.

A few examples may help illustrate the points above. Using microarrays and next-generation sequencing platforms, researchers can measure gene expression for the whole genome, for tens of thousands of genes and transcripts [Ozsolak and Milos, 2011, Zhao et al., 2014]. Using these expression levels, they are then often interested in studying the relationship between gene expression and disease propensity or other phenotypes. If a given gene is associated with a disease, this gene becomes a potential target for a drug. Gene expression data can also be used to investigate associations with genetic data; these studies are known as expression Quantitative Trait Loci (eQTL) studies. A similar situation arises with DNA methylation data. DNA methylation is the most commonly studied epigenetic mark, meaning that it is a chemical modification of the DNA molecule that do not change the nucleotide sequence. Like gene expression, DNA methylation can be measured over the entire genome, at millions of CpG dinucleotides. It explains a large part of tissue heterogeneity, and in some cases, it controls access to the parts of the genome that can be transcribed and translated. Therefore, evidence of association between methylation levels and a phenotype can again become a potential treatment target. Similarly, methylation Quantitative Trait Loci (meQTL) studies look at the association between DNA methylation levels and genetic data on Single Nucleotide Polymorphisms (SNPs).

The field of neuroimaging has gone through parallel technological changes as genomics, and accordingly there are parallels in the type of questions being posed and the methodological challenges encountered. Researchers are interested in the relationship between anatomical

features of the brain (e.g. cortical thickness, grey matter volume) and cognitive measurements. Structural Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET) can be used to quantify structural features at tens of thousands of locations in the human brain [Klein et al., 2009]. Researchers then look for evidence of association between brain locations and phenotypes of interest. This evidence can in turn further our understanding of cognitive processes, as well as the aetiology of neurodegenerative diseases.

A newly emerging field, called *imaging genetics*, combines the two fields of genomics and neuroimaging [Hariri et al., 2006]. Specifically, imaging genetics investigates the molecular basis of brain structure and function. For example, researchers may be interested in the association between structural brain features and genetic variation. To this end, they could correlate MRI measurements with SNP data. As another example, they could investigate the relationship between the same brain features and gene expression.

Historically, biological hypotheses about the relationship between a disease and a genomic (or brain) feature were approached in a targeted manner, i.e. a set of candidate markers for a particular disease or phenotype were selected *a priori*. However, with the recent technological advances, these hypotheses are now generally tested in an agnostic manner, i.e. the number of potential candidates is mostly limited by the resolution and output of the technology being used. Accordingly, the issue of multiple testing has become central to statistical genetics and neuroimaging, with considerable work looking for improvements to the classical Bonferroni correction. Specifically, researchers generally test for association between a molecular marker–or a brain region–and the phenotype of interest *one at a time*, and they then need to account for multiple testing when assessing statistical significance. The limitations of these approaches have been addressed in the literature [Visscher et al., 2017]: all such tests of pairwise association have limited power to detect weak effects, even if they are widespread; the correlation among the molecular markers and among the phenotypes is largely ignored; and the computational burden is typically high.

These problems are compounded when there are multiple phenotypes, and when we perform these pairwise tests for two high-dimensional datasets (e.g. brain measurements and gene expression). From a biological standpoint, some natural groupings of these measurements often arise. For example, some genes are co-regulated and act together as part of gene pathways. The methylation status of nearby CpG dinucleotides tend to be highly correlated. Correlation patterns between anatomical brain features are strongly driven by spatial distance and neural connectivity. Translating this into statistical terms, we can construe these observed biological phenomena as arising from a set of common latent processes. If we could test for association between these latent variables and a phenotype of interest, we could dramatically reduce the multiple testing burden and increase the signal-to-noise ratio. In turn, this would lead to higher statistical power to detect biological associations.

In this thesis, I look specifically at a class of biological hypotheses that I call *omnibus*: hypotheses that pertain to the relationship between two sets of molecular, imaging, or phenotype features. From a statistical perspective, I look at omnibus hypotheses that are amenable to a global association test: the association of interest is whether most or all variables in one set demonstrate association with most or all variables in the other set. The approach I take builds on the natural correlation present in these datasets. It uses dimension reduction methods to project the information contained in a set of variables onto a set of latent variables of lower dimension. By summarising the information and leveraging the correlation within these datasets, the objective is to increase power (compared to pairwise tests) while keeping the computational burden relatively low.

The thesis is structured as follows: in the next chapter, I review the literature on dimension reduction methods in high-dimensional data, with a focus on methods commonly used in genomics and neuroimaging. I provide a general framework that includes these commonly used dimension reduction methods, and I describe how sparsity has been the focus of dimension reduction in high dimensions. I also review the literature on null hypothesis testing

in the context of double Wishart problems, which underlie several multivariate approaches. The following three chapters present the three manuscripts that form the original contributions of this thesis. First, I present a structured covariance estimator that leads to a simple extension of the Principal Component of Explained Variance (PCEV) methodology to high-dimensional data. The computational advantages are discussed, and I show how the resulting approach (deemed "by block") results in a higher powered approach than other common high-dimensional methods. This manuscript has already been published [Turgeon et al., 2018b]. Second, I present an empirical estimator of the null distribution of the largest root statistic of double Wishart problems. I then show how this empirical estimator can be used to perform high-dimensional inference using dimension reduction methods like Canonical Correlation Analysis (CCA) and PCEV. Third, I apply the concepts developed in the two other manuscripts to a data analysis of targeted custom capture methylation data. I show how PCEV can be used in conjunction with the ideas in the second manuscript to test for association between groups of CpG dinucleotides and levels of anti-citrullinated protein antibody (ACPA), an antigen thought to be a predictor of rheumatoid arthritis onset. Finally, the last chapter summarises the contributions of the thesis and discusses future research avenues.

Chapter 2

Literature Review

As described in the introduction, this thesis focuses on (biological) *omnibus hypotheses*: we are looking for evidence of a relationship between two sets of measures on the same sample of individuals. More specifically, I will look at dimension reduction methods that try to address these omnibus questions. Motivation comes from both statistical genetics and neuroimaging, and therefore we expect high-dimensional, correlated variables for most types of measures. I will look at both the issue of *estimation*, i.e. how to reduce the dimension of the data, and the issue of *inference*, i.e. how can the reduced data provide evidence for a biological relationship. This chapter provides a literature review of these two topics. Section 2.1 addresses the first point. I start with an overview of common dimension reduction methods used in both genomics and neuroimaging. I then present some existing estimation frameworks, and I argue for a new one that encompasses most methods that are typically used in neuroimaging and genomics; I call this new framework *iterative optimisation*. I then survey existing approaches to reducing high-dimensional datasets. There is a strong emphasis on sparsity, and I cover the theoretical justification for such emphasis.

In Section 2.2, I focus the attention on inference in the context of *double Wishart problems*. I review that many multivariate approaches fall under this framework, and I show how

inference can be performed by looking at the distribution of the largest root to such problems. I then review several approaches to computing the distribution of this largest root under a null hypothesis of interest, and I highlight their limitations.

2.1 Estimation

In what follows, let $\mathbf{y}$ and $\mathbf{x}$ be a p-dimensional and q-dimensional random vector, respectively. Moreover, let $\mathbf{Y}$ be an $n \times p$ matrix and $\mathbf{X}$, an $n \times q$ matrix. The rows of these two matrices represent realisations of $\mathbf{y}$ and $\mathbf{x}$, respectively, measured on the same n experimental units. We are interested in reducing the dimension of either $\mathbf{Y}$ or $\mathbf{X}$ (or both), while describing their potential relationship.

2.1.1 Examples of dimension reduction methods

I start by providing an overview of several common dimension reduction methods. These methods have been repeatedly used in genomics and neuroimaging studies to summarise multivariate signals.

Principal Component Analysis

The canonical example of a dimension reduction method is Principal Component Analysis (PCA). Indeed, this approach is almost as old as modern statistics, with early examples coming from Pearson [1901] and Hotelling [1933]. I will describe two different approaches to PCA.

Pearson's approach to PCA is rooted in geometry. Indeed, Pearson was looking for the optimal linear manifold approximation to the data. He argued that this can be achieved by minimizing the *reconstruction error*. More specifically, let U be a $k \times p$ orthogonal matrix, with $k \leq p$. We can thus project our dataset $\mathbf{Y} = (y_1, \ldots, y_n)^T$ onto a k-dimensional linear

subspace:

$$y_i \to U y_i.$$

Conversely, given a $p \times k$ orthogonal matrix W, we get an embedding of $\mathbb{R}^k$ into the larger space $\mathbb{R}^p$:

$$z \to W z, \qquad z \in \mathbb{R}^k.$$

By chaining these two transformations, we can compare the original data y_i with its reconstruction:

$$\|y_i - WUy_i\|^2. \tag{2.1}$$

If we are interested in minimizing the reconstruction error 2.1, it turns out that we can focus on $U = W^T$ [Shalev-Shwartz and Ben-David, 2014, Lemma 23.1]. To sum up, we are interested in finding the orthogonal matrix W that minimizes the reconstruction error

$$\min_W \sum_{i=1}^{n} \|y_i - WW^T y_i\|^2.$$

This would lead to the best k-dimensional linear manifold approximation of the data. If we let $A = \mathbf{Y}^T\mathbf{Y}$, we can show that the reconstruction error is minimized exactly when W is the matrix whose columns are the eigenvectors corresponding to the k largest eigenvalues of A [Shalev-Shwartz and Ben-David, 2014, Theorem 23.2].

The other approach to PCA that I wish to highlight is due to Hotelling. He takes a different approach than Pearson, looking for directions in the data with largest variance. In other words, we are looking for the linear combination $w^T\mathbf{y}$ with highest variance:

$$\max_w \mathrm{Var}(w^T\mathbf{y}).$$

However, this optimisation problem is ill-defined: for every $w \neq 0$, $\mathrm{Var}((2w)^T\mathbf{y}) > \mathrm{Var}(w^T\mathbf{y})$, and therefore there is no maximum. This problem can be alleviated by normalizing the

objective function:

$$\max_w \frac{\mathrm{Var}(w^T \mathbf{y})}{w^T w}.$$

The solution to this and similar problems will be central to this thesis, and therefore I encapsulate the solution in the following Proposition.

Proposition 1. *Let A be a symmetric matrix and B a positive-definite matrix, where both A and B are of dimension p. Consider the* Rayleigh quotient

$$Q(w) = \frac{w^T A w}{w^T B w}.$$

The maximum value attained by this quotient corresponds to the largest root of the following determinantal equation:

$$\det(A - \lambda B) = 0, \tag{2.2}$$

and the vector that maximises $Q(w)$ is the corresponding eigenvector.

Proof. First, we note that for any scalar C, we have

$$\begin{aligned} Q(Cw) &= \frac{(Cw)^T A (Cw)}{(Cw)^T B (Cw)} \\ &= \frac{C^2 w^T A w}{C^2 w^T B w} \\ &= \frac{w^T A w}{w^T B w} \\ &= Q(w). \end{aligned}$$

In other words, the objective function $Q(w)$ is constant along one-dimensional subspaces of $\mathbb{R}^p$. Therefore, to maximise $Q(w)$, we can restrict our search to a sphere around the origin. For convenience, we select the sphere of radius 1 in the Mahalanobis distance induced by B. That is, maximising $Q(w)$ is equivalent to the following constrained optimisation problem:

$$\max_w w^T A w, \quad \text{subject to } w^T B w = 1.$$

We can rewrite this using the method of Lagrange multipliers: define

$$\mathcal{L}(w, \lambda) = w^T A w - \lambda(w^T B w - a).$$

The critical points of $\mathcal{L}(w, \lambda)$ can be found via differentiation:

$$\begin{aligned} \frac{\partial \mathcal{L}(w, \lambda)}{\partial w} = 0 &\implies 2w^T A - 2\lambda w^T B = 0 \\ &\implies Aw = \lambda B w. \end{aligned}$$

Since the equation $Aw = \lambda B w$ is equivalent to Equation 2.2 above, this completes the proof. □

For PCA, since $\text{Var}(w^T \mathbf{y}) = w^T \text{Var}(\mathbf{y}) w$, we simply take $A = \text{Var}(\mathbf{y})$ and $B = I_p$. Hence, we can see that Hotelling's formulation of PCA reduces to a classical eigenvalue problem. We can obtain further components by repeating the above procedure with $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{y}\tilde{w}\tilde{w}^T$, where $\tilde{w} = \text{argmax}\, Q(w)$.

PCA is ubiquitous in science and machine learning. It is used in genetics to correct for population stratification in association studies [Tiwari et al., 2008]. Combined with linear regression, it is also a very popular way of extending regression to ill-conditioned covariate matrices. Moreover, it is a very common preprocessing step with high-dimensional data. An excellent reference about PCA, its extensions and applications, is the book by Jolliffe [2002].

There are several ways one can use PCA to measure the association between two sets of variables $\mathbf{Y}$ and $\mathbf{X}$. First, a single component could be extracted from both sets and then a univariate test of association between these two components could be performed. In the presence of confounders, this univariate test could even be performed within a regression framework for adjustment. If one wanted to extract more than one component from either

$\mathbf{X}$ or $\mathbf{Y}$, they could then also be used within a (potentially multivariate) regression framework to test for association. However, the main PCA principle remains: dimension reduction of both sets of variables is done completely independently from the other set, and therefore there is no guarantee that the extracted components are relevant to the global association.

Independent Component Analysis

As we can see with both approaches to PCA, we obtain components that are uncorrelated. Of course, this can be strictly weaker than requiring independence when the components are not normally distributed. Motivated by this difference, Independent Component Analysis (ICA) looks for a decomposition of the dataset into *maximally independent* linear components. Two excellent sources for ICA are a review paper by Hyvärinen and Oja [2000], as well as a book by the first author [Hyvärinen et al., 2009].

ICA is better described as a family of methods with a common goal: given a multivariate variable $\mathbf{Y}$, find a decomposition

$$\mathbf{Y} = \mathbf{W}\mathbf{Z}$$

where $\mathbf{W}$ is a deterministic matrix and $\mathbf{Z}$ is a multivariate random variable, such that the components of $\mathbf{Z}$ are statistically independent. It turns out that for identifiability purposes, we must also impose a non-gaussian assumption: at most one column of $\mathbf{Z}$ can have a normal distribution; otherwise, the mixing matrix $\mathbf{W}$ is not identifiable [Comon, 1994]. This turns out to be a blessing in disguise, as it gives us better numerical criteria to optimise. Specifically, instead of looking for independent components explicitly, we can look for components that are as "non-gaussian" as possible.

One approach to identifying non-gaussian components is via the *kurtosis*. Kurtosis is a classical statistical measure defined as

$$\text{kurt}(X) = E(X^4) - 3(E(X^2))^2.$$

Since the fourth moment of a normal random variable is equal to $3(E(X^2))^2$, we can see that its kurtosis is zero; therefore, we can use the kurtosis as a measure of gaussianity. One could therefore look for linear combinations of the data that maximise the empirical kurtosis. However, since it depends on the fourth empirical moment, it is highly variable and very sensitive to outliers.

A second approach to ICA uses *entropy*. Entropy is an information-theoretic metric that measures the amount of information contained in a random variable. It is defined as follows:

$$H(X) = -\int p(x) \log p(x) dx,$$

where p is the density function of X. It is a well-known result of information theory that normal random variables maximise entropy among continuous distributions with a fixed variance (see for example [Cover and Thomas, 2012, Chapter 12]). If we define *negentropy* $J(X)$ as the difference in entropy between a random variable X and a normal random variable N with the same covariance structure, i.e.

$$J(X) = H(N) - H(X),$$

then $J(X)$ can also be seen as a measure of gaussianity. By maximising $J(X)$, we are in effect maximising non-gaussianity.

However, estimating negentropy is difficult: using the definition directly would require an estimate of the density function of X. Under some assumptions, it is possible to estimate the negentropy using higher-order moments, for example:

$$J(X) \approx \frac{1}{12} E(X^3)^2 + \frac{1}{48} \text{kurt}(X)^2$$

(see Jones and Sibson [1987]). Again, we run into similar problems as above, and in general the domain of applicability of the above approximation is rather limited.

Hyvärinen [1998] proposed new approximations of the negentropy based on a maximum-entropy principle; this principle leads to the following general approximation form:

$$J(X) \approx \sum_{i=1}^{K} \alpha_i \left(E(g_i(X)) - E(g_i(Z))\right)^2 .$$

In the equation above, the α_i's are positive constants, the functions g_i are nonquadratic transformations, the variable X is assumed to be standardized, and Z is a standard normal variable. For example, we could use a single nonquadratic transformation g and weight $\alpha_1 = 1$ to approximate the negentropy:

$$J(X) \propto \left(E(g(X)) - E(g(Z))\right)^2 .$$

Two common choices for g are

$$g(u) = \log \cosh u, \quad g(u) = -\exp(-u^2/2).$$

The real breakthrough for ICA came with the introduction of the FastICA algorithm by Hyvärinen [1999]. It is a fixed-point iteration scheme that seeks to maximise the negentropy as defined above using the approximation based on nonquadratic transformations. Let $\dot{g}$ be the derivative of the transformation g. For estimating a single component, the FastICA alternates between two steps until convergence:

1. Let $\tilde{w} = E(z\dot{g}(w^T z)) - E(\ddot{g}(w^T z))w$;
2. Normalise $w = \tilde{w}/\|\tilde{w}\|$.

Finally, a third approach to ICA uses *mutual information*. The mutual information of a set of variables $\mathbf{y} = (y_1, \ldots, y_p)$ is defined as

$$I(\mathbf{y}) = \sum_{i=1}^{p} H(y_i) - H(\mathbf{y}).$$

Unpacking the definition of entropy, we see that mutual information is equivalent to the Kullback-Leibler divergence between the joint density and the product of the marginal densities. Therefore, unlike the two previous settings, this third setting is explicitly looking for components that are as maximally independent as possible.

If we assume that $\mathbf{Y} = \mathbf{WZ}$ is such that $\mathbf{W}$ is invertible and the columns of $\mathbf{Z}$ are uncorrelated with unit variance, we can show that [Cover and Thomas, 2012]

$$I(\mathbf{Y}) = C - \sum_{i=1}^{p} J(y_i),$$

for some constant C. In other words, by minimizing the mutual information, we are also maximizing the non-gaussianity of the components.

ICA can be, and typically is, used as a stand-in for PCA when the non-gaussianity assumption for the latent features seems more natural [Mantini et al., 2010, Castro et al., 2016]. Therefore, when studying the relationship between two sets of variables, it suffers from the same drawbacks as PCA: the dimension of the datasets is reduced without using their dependence relationship. ICA is most commonly seen in neuroimaging, psychometrics, and finance, and it is rarely used in genomics (but for an example, see Teschendorff et al. [2011]).

Redundancy Analysis

As highlighted above, PCA and ICA are dimension reduction techniques that reduce the dimension of a single dataset, without using any information about the possible relationship between two datasets. Redundancy Analysis (RDA) is a first step towards reducing dimension while capturing the dependency between $\mathbf{Y}$ and $\mathbf{X}$. The best reference on the topic is Legendre and Legendre [2012, Chapter 11].

Consider the following setting: assume that the relationship between $\mathbf{y}$ and $\mathbf{x}$ can be repre-

sented via a linear model:

$$\mathbf{y} = \mathbf{Byx} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{B}$ is a $p \times q$ matrix of regression coefficients, and $\mathbf{e} \sim N_p(0, \mathbf{V}_R)$ is a vector of residual errors. RDA seeks a linear combination $w^T\mathbf{y}$ that maximises the *redundancy index*:

$$r(w) = \text{Var}(w^T\mathbf{Bx}).$$

Note that for an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$, we have

$$\begin{aligned} r(w) &= \text{Var}(w^T\hat{\mathbf{B}}\mathbf{x}) \\ &= w^T \text{Var}(\hat{\mathbf{B}}\mathbf{x})w \\ &= w^T \text{Var}(\hat{\mathbf{y}})w. \end{aligned}$$

Therefore, we can see that maximising the redundancy index is equivalent to performing PCA on the fitted values $\hat{\mathbf{y}}$.

RDA was first developed in the field of psychometrics as an alternative to Canonical Correlation Analysis (CCA) [Van Den Wollenberg, 1977]. It has gained some traction in the field of ecology, but it is essentially unheard of in the fields of genomics and neuroimaging.

Principal Component of Explained Variance

This dimension reduction approach will be presented in more detail in Chapter 3. Its presentation is repeated here using notation designed to highlight similarities and differences with the other dimension reduction methods discussed in this section. Principal Component of Explained Variance (PCEV) is very similar to RDA[1]. It starts from the same assumption about the relationship between $\mathbf{y}$ and $\mathbf{x}$. However, instead of maximising $r(w)$, it seeks a

[1]Some researchers even conflate the two methods; for example, see Lin et al. [2012]

linear combination $w^T\mathbf{y}$ that maximises the *proportion* of explained variance:

$$R^2(w) = \frac{\text{Var}(w^T\mathbf{Bx})}{\text{Var}(w^T\mathbf{y})}.$$

In contrast to RDA, PCEV was actually developed within the field of genetics, under the name of *Principal Component of Heritability (PCH)* [Ott and Rabinowitz, 1999]. Indeed, in family studies, we can typically decompose the total variance of a phenotype into a component coming from the genetic model and a residual component. The ratio of the genetic variation over the total variation is known as the heritability (in the broad sense) [Visscher et al., 2008]. In Turgeon et al. [2018b], I advocated for a different name, for two main reasons: 1) the concept of heritability can be misleading in the context of population studies, and 2) I showed that the domain of applicability is actually much broader than data about genetic variation. Therefore, I shall henceforth only use the term PCEV for this dimension reduction method.

In Proposition 1, we saw that the maximum value of $R^2(w)$ corresponds to the largest solution to a generalized eigenvalue problem. As I show next, to find the corresponding eigenvector, we can use a series of Eigenvalue Decompositions (EVDs). Recall the notation of Proposition 1: A is a symmetric matrix and B a positive-definite matrix, where both A and B are of dimension p. First, we perform an eigenvalue decomposition of B:

$$B = \Phi_B \Lambda_B \Phi_B^T.$$

Since B was assumed positive definite, all diagonal elements of Λ_B are positive (and the off-diagonal elements are zero), and therefore the matrix $\tilde{\Phi}_B = \Phi_B \Lambda_B^{-1/2}$ is well-defined. Now, consider the matrix

$$\tilde{A} = \tilde{\Phi}_B^T A \tilde{\Phi}_B.$$

It is easy to see that it is also symmetric, and therefore we can decompose it as

$$\tilde{A} = \Phi_A \Lambda \Phi_A^T,$$

where Φ_A is orthogonal and Λ diagonal. Let $\Phi = \Phi_B \Lambda_B^{-1/2} \Phi_A$. We have

$$\begin{aligned}
\Phi^T A \Phi &= \Phi_A^T (\Lambda_B^{-1/2} \Phi_B^T) A (\Phi_B \Lambda_B^{-1/2}) \Phi_A \\
&= \Phi_A^T \tilde{\Phi}_B^T A \tilde{\Phi}_B \Phi_A \\
&= \Phi_A^T \tilde{A} \Phi_A \\
&= \Lambda.
\end{aligned}$$

Simultaneously, we also have

$$\begin{aligned}
\Phi^T B \Phi &= \Phi_A^T \Lambda_B^{-1/2} (\Phi_B^T B \Phi_B) \Lambda_B^{-1/2} \Phi_A \\
&= \Phi_A^T \Lambda_B^{-1/2} \Lambda_B \Lambda_B^{-1/2} \Phi_A \\
&= \Phi_A^T \Phi_A \\
&= I.
\end{aligned}$$

In other words, we have a system of linear equations

$$\Phi^T A \Phi = \Lambda, \quad \Phi^T B \Phi = I.$$

This is equivalent to the generalized eigenvalue problem $A\Phi = B\Phi\Lambda$. In other words, the column of Φ corresponding to the largest diagonal element of Λ maximises the corresponding Rayleigh quotient.

When B is positive semidefinite, we can still solve the generalized eigenvalue problem by using a truncated EVD. For more details, see Appendix B.1.

Canonical Correlation Analysis

In both RDA and PCEV, the role of $\mathbf{y}$ and $\mathbf{x}$ is asymmetric: we first started with the assumption that their relationship follows a linear model, with $\mathbf{x}$ as covariates and $\mathbf{y}$ as response variables. In contrast, CCA treats $\mathbf{y}$ and $\mathbf{x}$ symmetrically. It seeks two linear combinations $w^T\mathbf{y}$ and $v^T\mathbf{x}$ that are maximally correlated with each other:

$$\max_{w,v} \operatorname{Corr}(w^T\mathbf{y}, v^T\mathbf{x})^2.$$

To solve this optimisation problem, let $\tilde{w} = \operatorname{Var}(\mathbf{y})^{1/2}w, \tilde{v} = \operatorname{Var}(\mathbf{x})^{1/2}v$. We can then write

$$\begin{aligned}\operatorname{Corr}(w^T\mathbf{y}, v^T\mathbf{x})^2 &= \frac{(w^T \operatorname{Cov}(\mathbf{y}, \mathbf{x})v)^2}{(w^T \operatorname{Var}(\mathbf{y})w)(v^T \operatorname{Var}(\mathbf{x})v)} \\ &= \frac{(\tilde{w}^T \operatorname{Var}(\mathbf{y})^{-1/2} \operatorname{Cov}(\mathbf{y}, \mathbf{X}) \operatorname{Var}(\mathbf{x})^{-1/2}\tilde{v})^2}{(\tilde{w}^T\tilde{w})(\tilde{v}^T\tilde{v})}\end{aligned}$$

Let $P = \operatorname{Var}(\mathbf{y})^{-1/2} \operatorname{Cov}(\mathbf{y}, \mathbf{x}) \operatorname{Var}(\mathbf{x})^{-1/2}$. By the Cauchy-Schwartz inequality, we have

$$(\tilde{w}^T P\tilde{v})^2 \leq (\tilde{w}^T PP^T\tilde{w})\tilde{v}^T\tilde{v},$$

and therefore we have

$$\operatorname{Corr}(w^T\mathbf{y}, v^T\mathbf{x})^2 \leq \frac{\tilde{w}^T PP^T\tilde{w}}{\tilde{w}^T\tilde{w}} = \frac{w^T \operatorname{Cov}(\mathbf{y}, \mathbf{x}) \operatorname{Var}(\mathbf{x})^{-1} \operatorname{Cov}(\mathbf{x}, \mathbf{y})w}{w^T \operatorname{Var}(\mathbf{y})w}.$$

In other words, as for PCEV, to optimise $\operatorname{Corr}(w^T\mathbf{y}, v^T\mathbf{X})^2$, we need to solve a generalized eigenvalue problem, where $A = \operatorname{Cov}(\mathbf{y}, \mathbf{x}) \operatorname{Var}(\mathbf{x})^{-1} \operatorname{Cov}(\mathbf{x}, \mathbf{y})$ and $B = \operatorname{Var}(\mathbf{y})$ (cf. Proposition 1 and the preceding subsection).

Partial Least Squares

Partial Least Squares (PLS) was first introduced by Wold [1985] as a way to improve prediction accuracy in the context of linear regression with high multicollinearity. The starting point is the *power method* for PCA: to compute the first principal component of a matrix $M = \mathbf{Y}^T\mathbf{Y}$, we start with a random vector v_0, and then iteratively normalise the product $M^i v_0$ until convergence. As I described earlier, PCA completely ignores the relationship between $\mathbf{Y}$ and $\mathbf{X}$ when reducing the dimension of $\mathbf{Y}$. However, Wold suggested a modification of the power method that incorporates information about $\mathbf{X}$. This algorithm is called Nonlinear Iterative Partial Least Squares (NIPALS), and it is presented here as Algorithm 1.

Algorithm 1 NIPALS

Center and scale both $\mathbf{X}$ and $\mathbf{Y}$; set $E_1 = \mathbf{X}$, $F_1 = \mathbf{Y}$. Set convergence limit ϵ.
for $i = 1$ to I **do**
  Randomly initialize u_i.
  repeat
    Set $w_i = E_i^T u_i / \|E_i^T u_i\|$.
    Set $t_i = E_i w_i$.
    Set $c_i = F_i^T t_i / \|F_i^T t_i\|$.
    Set $u_i = F_i c_i$.
  until $\|t_{i,\text{new}} - t_{i,\text{old}}\| < \epsilon$
  Set $b_i = t_i^T u_i$ and $p_i = E_i^T t_i$.
  Set $E_{i+1} = E_i - t_i p_i^T$ and $F_{i+1} = F_i - b_i t_i c_i^T$.
end for

As described in Frank and Friedman [1993], the solution obtained from Algorithm 1 also solves the following optimisation problem:

$$\max_{u,c} \operatorname{Var}(u^T\mathbf{y}) \operatorname{Corr}(u^T\mathbf{y}, c^T\mathbf{x}) \operatorname{Var}(c^T\mathbf{x}).$$

Intuitively, we are therefore jointly solving two PCA problems and one CCA problem, corresponding to the three terms in the expression above. We can therefore see PLS as a compromise between Principal Component Regression (PCR) and CCA.

For more details about PLS, its history, extensions and methods, see Abdi [2010].

Linear Discriminant Analysis

Finally, I will look at a slightly different example. Linear Discriminant Analysis (LDA) has a dual purpose: it is a dimension reduction approach, but also a classification method. Accordingly, the linear combinations of $\mathbf{Y}$ that LDA seeks are those that maximise the separation between the different classes. More precisely, assume that we observe n_k p-dimensional variables $\mathbf{Y}_k$, for $k = 1, \ldots, K$. In other words, the observations belong to one of K groups, and we can encode this via the vector $\mathbf{X}$, where $X_i = k$ whenever observation i belongs to group k. Let $\hat{\mathbf{Y}}_k$ and S_k be the sample mean and covariance of the k-th group, respectively, and let $\hat{\mathbf{Y}}$ be the overall sample mean. Consider the following two matrices:

$$S_B = \sum_{k=1}^{K} (\hat{\mathbf{Y}}_k - \hat{\mathbf{Y}})(\hat{\mathbf{Y}}_k - \hat{\mathbf{Y}})^T,$$
$$S_W = \sum_{k=1}^{K} S_k.$$

Similarly to PCEV and CCA, LDA considers a Rayleigh quotient:

$$D(w) = \frac{w^T S_B w}{w^T S_W w}.$$

The value $\hat{w}$ that maximises $D(w)$ is called the *first discriminant*. A total of $K - 1$ discriminants can be extracted by adding the constraint that each subsequent $\tilde{w}$ that maximises $D(w)$ must also be S_W-orthogonal to the previously extracted discriminants.

By imposing distributional assumptions on $\mathbf{Y}$, we can then use these discriminants to construct an optimal classifier. For more details, see Friedman et al. [2001].

2.1.2 Existing frameworks

In the previous section, I surveyed several common dimension reduction approaches. To facilitate the discussion in the later sections of this chapter, it would be helpful to think of these methods as part of a general framework. To this end, I start by presenting two general frameworks for dimension reduction commonly encountered in the literature: Generalized Low-Rank Models (GLRM), and Sufficient Dimension Reduction (SDR). I show that they have different scopes, and I argue that these frameworks do not quite capture the methods that we typically encounter in genomics and neuroimaging. Based on this, I present a third framework called *iterative optimisation*.

Generalized Low-Rank Models

A recent paper by Udell et al. [2016] provided a very general framework describing low-rank matrix approximations. This framework encompasses many popular matrix decomposition methods, e.g. Singular Value Decomposition (SVD), regularized PCA, non-negative matrix factorization, but it also includes more surprising unsupervised approaches like k-means clustering. Suppose we wish to approximate $\mathbf{Y} = (Y_{ij})$ by a matrix $\tilde{\mathbf{Y}}$ of rank $k \leq \min(n, p)$; for example, the matrix $\tilde{\mathbf{Y}}$ could be the reconstruction (in p dimensions) of the data $\mathbf{Y}$ after reducing its dimension to k. This is equivalent to approximating $\mathbf{Y}$ by the matrix product $W^T Z$, where W is a matrix of dimension $k \times n$ and Z is of dimension $k \times p$. In its more general form, the GLRM corresponds to the following optimisation problem:

$$\min \sum_{(i,j)\in\Omega} L_{ij}(w_i^T z_j, Y_{ij}) + \sum_{i=1}^{m} r_i(w_i) + \sum_{j=1}^{n} \tilde{r}_j(z_j), \tag{2.4}$$

where

- $\Omega \subseteq \{1, \ldots, n\} \times \{1, \ldots, p\}$ is a set of pairs (i, j) corresponding to entries of Y;
- $L_{ij} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ are given loss functions for $i = 1, \ldots, n$ and $j = 1, \ldots, p$;

- $r_i, \tilde{r}_j : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ are regularizing functions;
- w_i is the i-th column of W, and z_j is the j-th column of Z.

The minimisation problem in Equation 2.4 is very general, and therefore I will go through some examples in the hope of improving clarity. The simplest example is that of truncated SVD; it can be shown that it is the solution to the following optimisation problem:

$$\min \left\| \mathbf{Y} - W^T Z \right\|_F^2; \tag{2.5}$$

for an early instance of this result, see Eckart and Young [1936]. Here, I use $\|\cdot\|_F^2$ to denote the squared Frobenius norm; this matrix norm corresponds to the sum of the squares of the matrix entries. Therefore, to relate truncated SVD to the framework 2.4, we simply note that both regularisation functions are identically zero, and the loss functions are all equal to the quadratic loss.

An interesting feature of the Frobenius norm is that, if $\mathbf{Y}$ contains missing values, we can simply omit them from the sum. In this way, we can actually use truncated SVD as a data imputation method. Without missing values, the solution to Equation 2.4 involves only elementary linear algebra. However, in the presence of missing data, the optimisation problem becomes more computationally involved [Troyanskaya et al., 2001]. By requiring the loss function to be separable over the entries of $\mathbf{Y}$ as above, GLRM seeks to retain this ability to impute missing data.

Another advantage of customising the loss function to the entries of $\mathbf{Y}$ is that each entry could potentially receive different weights. An example of GLRM that has this feature is weighted PCA [Srebro and Jaakkola, 2003]. Yet another advantage of an entry-specific loss function is to allow for matrix decompositions when $\mathbf{Y}$ contains heterogeneous data types: for example, if some columns of $\mathbf{Y}$ contain continuous data and other columns, binary data, one could choose a mixture of quadratic and logistic loss [Schein et al., 2003].

For more examples of GLRMs, and for details about fitting procedures, I refer the reader to Udell et al. [2016].

Even though GLRM is very general, it only performs dimension reduction on a single dataset $\mathbf{Y}$. For example, we can recover PCA by choosing regularisation functions $r_i, \tilde{r}_j$ that are identically zero and setting $L_{ij}(w_i^T z_j, Y_{ij}) = \left(Y_{ij} - w_i^T z_j\right)^2$. However, none of the other methods above fit this framework.

Sufficient Dimension Reduction

Sufficient Dimension Reduction provides an elegant framework for dimension reduction by providing a natural definition. First, I assume that $p = 1$, i.e. $\mathbf{y}$ is a single random variable; the setting where $p > 1$ will be addressed later in this section. SDR considers linear transformations of $\mathbf{x}$ that contain all the information about the joint distribution of $\mathbf{x}$ and $\mathbf{y}$. More precisely, it seeks a $q \times r$ matrix B (with $r \leq q$) such that

$$\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid B^T \mathbf{x}. \tag{2.6}$$

It is not hard to see that if B satisfies the condition above, so does BA for any non-singular $r \times r$ matrix A [Li, 2018, Chapter 2.2]. Therefore, the identifiable parameter in SDR is not the matrix B itself, but rather its column space $\text{span}(B)$. A subspace V of $\mathbb{R}^q$ is called an *SDR subspace* if $V = \text{span}(B)$ for a matrix B that satisfies Property 2.6.

Several well-known regression models satisfy the property 2.6. Generalized Linear Models are such examples: for a given link function g, we have

$$g\left(\mathrm{E}\left(\mathbf{y} \mid \mathbf{x}\right)\right) = \beta \mathbf{x}.$$

In other words, the relationship between $\mathbf{y}$ and $\mathbf{x}$ is encapsulated in the linear predictor

$\beta\mathbf{x}$[2]. Semiparametric regression models, such as the Multiple Index Model, also satisfy the conditional independence property above [Yin et al., 2008].

Under very weak conditions on the distribution of $\mathbf{x}$, the intersection of two SDR subspaces is again an SDR subspace [Yin et al., 2008]. This result leads to the following definition:

DEFINITION 1. Let $\mathbf{y}$ and $\mathbf{x}$ be as above, and let $\mathfrak{A}$ be the set of all corresponding SDR subspaces; since $\mathbb{R}^q$ is an SDR subspace, the set $\mathfrak{A}$ is non-empty. The subspace

$$\mathcal{S}_{\mathbf{y}|\mathbf{x}} = \bigcap_{V \in \mathfrak{A}} V$$

is called the *central SDR subspace*.

Therefore, the ultimate goal of SDR is to estimate the central SDR subspace. This can be construed as an analogue to finding a minimal sufficient statistic in Fisherian statistics.

The first example of a statistical methodology whose explicit goal was to estimate the central SDR subspace is Sliced Inverse Regression (SIR) [Li, 1991]. This method relies on the observation that, whereas the semiparametric regression problem $\mathrm{E}(\mathbf{y} \mid \mathbf{x})$ suffers from the curse of dimensionality, the inverse regression $E(\mathbf{x} \mid \mathbf{y})$ can be approached through univariate smoothing techniques. The connection between this inverse regression and SDR is summarised in the following result:

THEOREM 1. [Li, 1991, Theorem 3.1] Let B be such that $\mathcal{S}_{\mathbf{y}|\mathbf{x}} = \mathrm{span}(B)$, and assume that $\mathbf{x}$ follows an elliptical distribution. The centred inverse regression curve $E(\mathbf{x} \mid \mathbf{y}) - E(\mathbf{x})$ is contained in the linear subspace spanned by $B^T\Sigma_x$, where Σ_x denotes the covariance matrix of $\mathbf{x}$.

[2]Strictly speaking, this is not quite strong enough to lead to conditional independence. Accordingly, this weaker form is called *Sufficient Dimension Reduction for Conditional Mean*. For more details, see Cook et al. [2002]

In other words, by estimating the centred inverse regression curve and looking at the subspace it spans, we obtain a "lower bound" on the central subspace. To estimate the centred inverse regression curve, Li [1991] advocates for a simple method that relies on computing the sample mean of $\mathbf{x}$ over slices of the range of $\mathbf{y}$. But he also concedes that any smoothing technique could be used.

The result above only guarantees that the centred inverse regression curve is contained in the central SDR. Indeed, this inclusion may be strict: as shown by Cook and Weisberg [1991], in the presence of symmetric dependence, SIR will only uncover a proper subspace of $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$. Cook and Weisberg suggested a new method called Sliced Average Variance Estimate (SAVE) that relies on second order relationships between $\mathbf{x}$ and $\mathbf{y}$ that can circumvent the issues arising from symmetric dependence. Moreover, under some mild conditions on the joint distribution of $\mathbf{y}$ and $\mathbf{x}$, SAVE is guaranteed to recover the whole central SDR subspace [Li, 2018, Theorem 5.3].

The ideas behind SIR and SAVE have led to a plethora of methodological papers. Inverse regression can be tackled parametrically or through the use of kernel-machine regression [Zhu et al., 1996]. Hypothesis tests have also been developed to estimate the dimension of the central SDR subspace [Li, 2018, Chapter 9]. For high-dimensional data, Li [2007] suggests using shrinkage estimation to obtain a sparse representation B_s of the central SDR subspace. Alternatively, Cook et al. [2007] show how to perform SDR without the need of inverting large matrices, extending SIR and SAVE to high-dimensional settings.

Some common dimension reduction approaches can be seen to fit the SDR framework. PCR, whereby PCA is performed on the vector $\mathbf{x}$ and then followed by regression of $\mathbf{y}$ on a selected number of principal components, is such an example [Adragni and Cook, 2009]; univariate PLS regression provides another example [Helland, 1990]. To see this, let S_X be the sample covariance estimate for the covariates $\mathbf{x}$, and let S_{XY} be the sample estimator of the covariance $\text{Cov}(\mathbf{x}, \mathbf{y})$ (recall that we assumed $\mathbf{y}$ has dimension one, and therefore

S_{XY} is a vector as well). Consider regressing $\mathbf{y}$ on $\mathbf{x}$ in two steps: first, we reduce the dimension of $\mathbf{x}$ to $r \leq q$ using a $q \times r$ projection matrix B; second, we regress $\mathbf{y}$ on the reduced covariates $B^T\mathbf{x}$. Now, if we let B be the first r eigenvectors of S_X, we obtain PCR; if we let $B = (S_{XY}, S_X S_{XY}, \ldots, S_X^{r-1} S_{XY})$, we obtain PLS [Helland, 1990]. Under some distributional assumptions on $\mathbf{x}$, the subspace span(B) obtained from PCR is a consistent estimator for an SDR subspace $\mathcal{S}$ [Cook et al., 2008]; we obtain a similar consistency result for PLS under some assumptions on the relationship between Cov($\mathbf{x}, \mathbf{y}$) and the eigenstructure of Cov($\mathbf{x}$) [Naik and Tsai, 2000].

However, there is no obvious link between SDR and other common multivariate analysis methods such as CCA, LDA and Multivariate Analysis of Variance (MANOVA). This is related to the fact that, even though Property 2.6 can easily be extended to a multivariate $\mathbf{y}$, methods such as SIR and SAVE cannot be easily adapted to this multivariate setting. Indeed, they both suffer from the curse of dimensionality that they were trying to avoid in the first place by performing inverse regression.

2.1.3 Iterative optimisation framework: a new perspective on dimension reduction

The two frameworks described above have a number of advantages and limitations. GLRM is very flexible and computationally mature; however, it is only restricted to dimension reduction of a single set of variables at a time, with no obvious way of using the relationship between two sets $\mathbf{Y}$ and $\mathbf{X}$. SDR is theoretically very satisfactory, with a natural approach to dimension reduction; however, the extension to multivariate $\mathbf{Y}$ brings computational challenges, and it is not clear how the theoretical framework leads to an estimation framework. Moreover, both frameworks exclude some very common dimension reduction techniques: GLRM does not include PLS, and SDR does not include CCA (cf. Table 2.1). For these reasons, I present here an alternative framework. This framework includes all

Method	GLRM	SDR	Iterative Optimisation
PCA	**Yes**	**Yes**[3]	**Yes**
ICA	No	No	No
RDA	No	No	**Yes**
PCEV	No	No	**Yes**
CCA	No	No	**Yes**
PLS	No	**Yes**	**Yes**
LDA	No	No	**Yes**

Table 2.1: Breakdown of which dimension reduction methods are included within the GLRM, SDR, and Iterative Optimisation frameworks.

methods presented in Section 2.1.1, except ICA.

For the purpose of this thesis, I will mainly consider dimension reduction performed as a series of optimal linear projections onto a one-dimensional subspace. More precisely, I am looking for a p-dimensional vector $\tilde{w}_1$, such that for a given function

$$f : \mathbb{R}^p \to \mathbb{R},$$

we have

$$\tilde{w}_1 = \underset{w}{\operatorname{argmax}}\, f(w^T \mathbf{y}; \mathbf{x}, \theta), \quad \text{subject to } \|w\|_M = 1,$$

where $\|\cdot\|_M$ is the norm induced by the Malahanobis distance corresponding to a positive definite matrix M (cf. Proposition 1). The subsequent components can be extracted by repeating the optimisation with $\tilde{\mathbf{y}} = \mathbf{y} - \tilde{w}_1 \tilde{w}_1^T$, which induces an orthogonality constraint (again with respect to the matrix M).

Some further comments are required:

- We present the framework as a maximisation one, but this is done without loss of generality. Indeed, we could easily replace a function f' that needs to be minimised by its negative $-f'$ and turn this into a minimisation problem.
- We allow for the function f to depend on another set of variables $\mathbf{x}$ and thus take

[3]More specifically, PCR and not PCA is an example of SDR.

advantage of the relationship between $\mathbf{y}$ and $\mathbf{x}$. We also allow for extra parameters θ that potentially describe this relationship. For example, θ could be the regression parameter describing the linear relationship between $\mathbf{y}$ and $\mathbf{x}$ in RDA and PCEV.

- The same framework could be used to perform dimension reduction of both $\mathbf{y}$ and $\mathbf{x}$: first, replace $\mathbf{y}$ by $(\mathbf{y}, \mathbf{x})$ and extend the domain of f to $\mathbb{R}^{p+q}$. If the matrix M used to define the Malahanobis distance is block-diagonal, with blocks M_1, M_2 of dimension p, q, respectively, then we have

$$\|(w, v)\|_M = \|w\|_{M_1} + \|v\|_{M_2}.$$

This framework is quite general, and indeed it includes most linear regression frameworks. It also includes all common dimension reduction methods seen in neuroimaging and genomics presented in Section 2.1.1, with the exception of ICA. Indeed, since the goal of ICA is (joint) independence between the components, it is intrinsically a joint dimension reduction approach. FastICA does provide an algorithm for minimising the negentropy that computes a single component at a time; however, the algorithm is presented as a fixed-point iterative scheme, and it is not clear that the solution also optimises a single criterion. Therefore, I do not consider ICA within the iterative optimisation framework.

For all other dimension reduction methods discussed above, Table 2.2 provides the corresponding objective function f.

2.1.4 Dimension reduction and high-dimensional data

By showing how the above methods fit the iterative optimisation framework, I was able to highlight another common trait: these methods all rely on the estimation of sample covariance matrices. However, when the number of features p is larger than the sample size n, these matrices are ill-conditioned. Moreover, these sample covariance matrices are no

Method	Objective function
PCA	$f(w^T\mathbf{y}) = \mathrm{Var}(w^T\mathbf{y})$
RDA	$f(w^T\mathbf{y}) = \mathrm{Var}(w^T B^T\mathbf{x})$
PCEV	$f(w^T\mathbf{y}) = \frac{\mathrm{Var}(w^T B^T\mathbf{x})}{\mathrm{Var}(w^T\mathbf{y})}$
CCA	$f([u,v]^T[\mathbf{y},\mathbf{x}]) = \mathrm{Corr}(u^T\mathbf{y}, v^T\mathbf{x})$
PLS	$f([u,c]^T[\mathbf{y},\mathbf{x}]) = \mathrm{Var}(u^T\mathbf{y})\,\mathrm{Corr}(u^T\mathbf{y}, c^T\mathbf{x})\,\mathrm{Var}(c^T\mathbf{x})$
LDA	$f(w^T\mathbf{y}) = \frac{w^T S_B w}{w^T S_W w}$

Table 2.2: Objective function for the dimension reduction methods described in Section 2.1.1.

longer consistent estimators of the corresponding true covariance matrices for large p. This can be seen, for example, as a consequence of the work by Marčenko and Pastur [1967] on the distribution of the eigenvalues of the sample covariance. This inconsistency result also transfers to the eigenvectors:

THEOREM 2 (Johnstone and Lu [2009]). Assume that we have n observations $\{y_i\}_{i=1}^n$ from a p-dimensional model of the form

$$\mathbf{y} = \nu\rho + \sigma Z,$$

where $\rho \in \mathbb{R}^p$, $\nu \sim N(0,1)$, and $Z \sim N_p(0, I)$. Let $\hat{\rho}$ be the eigenvector corresponding to the largest root of the sample covariance matrix. Furthermore, assume that the limiting signal-to-noise ratio is bounded away from zero:

$$\lim_{n\to\infty} \|\rho\|^2/\sigma^2 = \omega > 0.$$

Let $c = \lim_{n\to\infty} p_n/n$ be the limit of the ratio of the dimension to the sample size; note that we let the dimension grow with the sample size. Then almost surely

$$\frac{\hat{\rho}^T\rho}{\|\hat{\rho}\|\|\rho\|} \to \frac{(\omega^2 - c)_+}{\omega^2 + c\omega}, \quad \text{as } n \to \infty.$$

In particular, $(\omega^2 - c)_+/(\omega^2 + c\omega) < 1$ if and only if $c > 0$, and therefore $\hat{\rho}$ is a consistent estimator of ρ if and only if $p/n \to 0$.

As an extreme case of the above result, when $c \geq \omega^2$, the vectors $\hat{\rho}$ and ρ are asymptotically orthogonal. In other words, the estimator $\hat{\rho}$ contains no information whatsoever about its corresponding estimand ρ. This result also extends to multicomponent factor analytic model Paul [2007].

Sparsity constraints

The solution for recovering consistency in high dimensions is through a *sparsity* constraint: we generally assume that the number of latent variables is much smaller than the dimension of the observed variables to which they give rise. It is difficult to overstate the importance of Theorem 2; indeed, and as we shall see, it underlies most approaches to high-dimensional dimension reduction. Under their one-component model, Johnstone and Lu [2009] suggest the following algorithm for sparse PCA:

1. Transform the data using a wavelet transformation, where sparsity is more likely to hold (under some conditions);
2. Retain the k most variable transformed variables;
3. Perform PCA on the remaining variables;
4. Filter out the noise in principal components using hard thresholding;
5. Return to the original representation of the data.

However, it is not clear that the wavelet transform is always the best transformation for achieving sparsity. Indeed, it is also not clear how the optimal transformation can be obtained.

In Jolliffe et al. [2003], the authors also argue that sparsity can improve the interpretation

of the derived principal components. To this aim, they introduce Simplified Component Technique-Lasso (SCoTLASS). Drawing inspiration from Tibshirani [1996], they propose to solve the PCA optimization problem

$$\min \operatorname{Var}(w_i^T \mathbf{Y}), \quad \text{subject to } w_i^T w_i = 1, w_i^T w_j = 0 \text{ for } i \neq j$$

with an extra L_1-condition on the loading vectors:

$$\sum_{\ell=1}^{p} |w_{i\ell}| \leq t$$

for some tuning parameter $t > 0$. Although this optimization is appealing in its simplicity, it is no longer a convex optimization problem, and therefore it converges slowly and is prone to converging to local optima.

Zou et al. [2006] improves on both Jolliffe *et al.* and Johnstone & Lu by proposing a computationally efficient algorithm for recovering sparse principal components. Their approach leverages Pearson's geometric view of PCA. Specifically, they start by recasting PCA as a ridge regression problem:

THEOREM 3 ([Zou et al., 2006]). Let y_i be the i-th row of the matrix $\mathbf{Y}$. Suppose we are considering the first k principal components. For any $\lambda > 0$, let

$$(\hat{W}, \hat{U}) = \underset{W,U}{\operatorname{argmin}} \sum_{i=1}^{n} \|y_i - WU^T y_i\|^2 + \lambda \sum_{j=1}^{k} \|u_j\|^2 \tag{2.7}$$
$$\text{subject to } W^T W = I_k,$$

where u_j is the j-th column of U. Then $\hat{u}_j$ is proportional to the j-th loading vector for the j-th principal component, $j = 1, \ldots, k$.

If we add the constraint $W = U$ and set $\lambda = 0$, we recover Pearson's formulation of PCA.

Therefore, we can construe the addition of the ridge penalty as a counterbalance to the relaxation of the matrix equality constraint $W = U$, and we can still recover (up to a multiplicative constant) the original principal components. To obtain a sparse PCA algorithm, the authors suggest adding an L_1-penalty to Equation 2.7. They also provide an alternating minimisation algorithm that is computationally efficient.

Sparse dimension reduction beyond PCA

Witten et al. [2009] go even one step further than Zou et al. [2006]. Their starting point is the observation that SVD can be used to construct optimal rank r approximations of a matrix. More specifically, let ℓ denote the rank of $\mathbf{Y}$, and let U, D, V be such that

$$Y = UDV^T,$$

with D diagonal. If we let u_k, v_k be the k-th column of U, V, respectively, and d_k be the k-th element on the diagonal of D, we have

$$\sum_{k=1}^{r} d_k u_k v_k^T = \underset{\hat{\mathbf{Y}}}{\operatorname{argmin}} \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_F^2,$$

where $\hat{\mathbf{Y}}$ ranges over all rank $r \leq \ell$ matrices of dimension $n \times p$, and $\|\cdot\|_F^2$ is the squared Frobenius norm. In the case of $r = 1$, the minimisation problem above is equivalent to

$$\max_{u,v} u^T \mathbf{Y} v, \quad \text{subject to } \|u\|_2^2 = 1, \|v\|_2^2 = 1. \tag{2.8}$$

The authors thus define the rank-1 Penalized Matrix Decomposition (PMD) as an extension of Equation 2.8 in which they add two penalty terms P_1 and P_2:

$$\max_{u,v} u^T \mathbf{Y} v, \quad \text{subject to } \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, P_1(u) \leq c_1, P_2(v) \leq c_2. \tag{2.9}$$

They provide an efficient alternating optimisation algorithm that relies on the biconvexity of Equation 2.9.

When P_1, P_2 are chosen as L_1-penalties on u, v, rank-1 PMD can be seen as yet another approach to sparse PCA. Indeed, the authors show how they can recover Jolliffe et al. [2003]'s SCoTLASS, and they thus provide an efficient algorithm for its computation. They also show how their algorithm relates to Zou et al. [2006]'s sparse PCA.

Furthermore, PMD can also be used to obtain a sparse version of CCA; this is achieved by performing PMD on $\mathbf{X}^T\mathbf{Y}$ (assuming $\mathbf{X}$ and $\mathbf{Y}$ have mean zero), and by replacing the constraints $\|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1$ by $u^T\mathbf{X}^T\mathbf{X}u \leq 1, v^T\mathbf{X}^T\mathbf{X}v \leq 1$. Witten et al. [2009] also contains a discussion of other approaches to sparse CCA, along with a comparison of these approaches with PMD.

Inspired by Zou et al. [2006], Chun and Keleş [2010] proposed a sparse approach to PLS. Their ideas are similar: they recast the PLS optimization problem by augmenting the original equation with the addition of a second parameter, and then they regularise this new equation by adding an L_2 constraint. A sparse solution is then obtained by adding a second penalty term, an L_1-penalty. More precisely, they aim to solve the following optimization problem:

$$(\hat{w}, \hat{c}) = \underset{w,c}{\operatorname{argmin}} \left\{ -\kappa w^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}w + (1-\kappa)(c-w)^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}(c-w) + \lambda_1\|c\|_1 + \lambda_2\|c\|_2 \right\}$$
$$\text{subject to } w^Tw = 1,$$

where κ is a tuning parameter.

Similarly to Zou et al. [2006], Chun and Keleş [2010] show that this optimisation problem can be solved efficiently by alternating between optimising over w and c separately.

Fang et al. [2014b] proposed a similar solution to obtain sparse loadings for PCEV. They suggest optimising the following objective function:

$$R^2_\lambda(w) = \frac{\text{Var}(w^T\mathbf{BX})}{\text{Var}(w^T\mathbf{Y}) + \lambda\|w\|_1^2}.$$

Under a sparsity assumption on the loadings, the authors prove a consistency result about this regularized PCEV. However, this objective function is difficult to optimise. For this reason, Fang et al. [2014b] suggest approximating it with

$$\tilde{R}^2_\lambda(w) = \text{Var}(w^T\mathbf{Y}) + \lambda\|w\|_1^2 + \gamma\left(\text{Var}(w^T\mathbf{BX}) - 1\right)^2.$$

This objective function is still non-convex, but the authors provide an efficient optimisation algorithm based on coordinate descent.

Sparsity through structured covariance estimation

The different approaches to sparsity described above all use a similar approach: they add an L_1 constraint to the optimisation problem to obtain sparse loadings. Crucially, which components should be zero is not set *a priori*, and indeed it is learned from the data. There is another approach to sparsity that does not add a penalty term to the objective function: sparsity via structured estimation. We give two examples, both due to Bickel & Levina. In Bickel et al. [2008], the authors study banded estimators of the covariance matrix: if $M = (m_{ij})$ is a p-dimensional matrix, for any integer $0 \leq k < p$, we can define a banding operator via

$$B_k(M) = (m_{ij} \cdot \mathbb{1}(|i-j| \leq k)).$$

In other words, some of the off-diagonal entries are set to zero. If $\hat{\Sigma}$ is the sample covariance matrix, Bickel & Levina suggest to estimate the population covariance using $\hat{\Sigma}_k = B_k(M)$. They then prove consistency results in the operator norm; this norm is equal to the largest

eigenvalue of the matrix. Therefore, these optimality results also translate to the principal components derived from $\hat{\Sigma}_k$. In turn, this leads to good performance when PCA is performed on these estimators.

Alternatively, Bickel and Levina [2008] argue for the use of a *thresholding* operator on the covariance matrix: for any $s \geq 0$, we define

$$T_s(M) = (m_{ij} \cdot \mathbb{1}(|m_{ij}| \geq s)).$$

In other words, every entry of M whose absolute value is less than s is set to zero. Similar to above, the authors suggest to estimate the population covariance using $\hat{\Sigma}_s = T_s(M)$. Again, consistency results in the operator norm are derived.

As we have seen above, a large number of methodological solutions to dimension reduction with high dimensional data build on the success of lasso regression [Tibshirani, 1996]. They impose an L_1-constraint on the loading vectors, thus ensuring a sparse solution to the corresponding optimisation problem. Other approaches enforce sparsity via structure (e.g. thresholding and banding). Bickel et al. [2008] provide further examples of this approach.

Similarly, analogues of ridge regression have been used as well to address issues of high dimensionality; for example, see Vinod [1976], Leurgans et al. [1993], Ramsay [2006], Wang et al. [2007a], Allen et al. [2013], Aflalo and Kimmel [2017], Wang and Huang [2017]. Through regularisation of the loadings, these methods address some of the limitations of the classical methods; however, they typically do not have the same interpretability advantages as their sparse counterparts.

One important detail I have yet to mention is that all methods above depend on one or more tuning parameters. Often, these tuning parameters control the amount of regularisation occurring within the optimization problem. In other cases, like in Bickel & Levina's

approaches, the tuning parameter controls the amount of banding, or at which point small values in the estimate matrices should be set to zero. All authors recommend some form of resampling to estimate these parameters; for example, in sparse PCA, cross-validation can be used to find the value of λ that minimizes the (out-of-sample) reconstruction error. For more details, see Friedman et al. [2001, Chapter 7].

Tuning these parameters using resampling imposes an additional computational burden. This burden can become especially acute when we use dimension reduction over hundreds or thousands of sets of variables.

In the first manuscript, presented in Chapter 3, I provide an example of a dimension reduction approach to high-dimensional data that has a much lower computational cost. The method presented is another example of sparsity via structured estimation. I achieve this goal by taking advantage of the natural correlation structure present in genomics and neuroimaging data, as well as some properties of the PCEV optimisation problem under block independence. Crucially, my approach does not require any tuning parameter, and it is therefore computationally efficient.

2.2 Inference

Next, I turn our attention to the issue of assessing the evidence of an association between $\mathbf{Y}$ and $\mathbf{X}$. Within the context of dimension reduction, this is often reformulated in terms of the relationship between components $w^T\mathbf{Y}$ and $u^T\mathbf{X}$. To this aim, I will focus our discussion to the framework of Null Hypothesis Significance Testing (NHST). This is by far the most common inferential framework used in genetic epidemiology and neuroimaging.

On the simpler end of the spectrum, methods like PCA and ICA that reduce dimension of each set of variables independently of one another can often rely on classical, univariate tests of association. For example, the set $\mathbf{Y}$ could be reduced to a single component $w^T\mathbf{Y}$

through PCA. We could then use univariate linear regression and an F-test to assess the significance of the relationship between $w^T\mathbf{Y}$ and $\mathbf{X}$. However, there is no guarantee that these tests will have any power to detect an association between the two sets of variables. This is due to the dimension reduction being performed without using any information about the relationship. With high-throughput technologies, where technical artifacts and tissue-heterogeneity can drive most of the variance in a dataset, it can become almost impossible to detect any association between $\mathbf{Y}$ and $\mathbf{X}$.

At the other end of the spectrum, for high-dimensional variants of CCA and PLS, the analytical distribution of the test statistics can sometimes be difficult to derive, or even approximate. Typically, they can only be derived with strong distributional assumptions on $\mathbf{Y}$ and $\mathbf{X}$, and they are often restricted to the classical setting of $p/n \to 0$. This problem is especially acute with models that seek sparse loadings for the components (cf. Section 2.1.4). In these settings, the most convenient approach is often to use a permutation strategy [Nichols and Holmes, 2002]. Briefly, a permutation test is a resampling approach to estimating the null distribution of our test statistic. The observation labels of $\mathbf{Y}$ or $\mathbf{X}$ (but not both) are permuted multiple times, and for each iteration the test statistic is computed on the scrambled data. The validity of the procedure relies on the fact that under the null hypothesis of no association between $\mathbf{Y}$ and $\mathbf{X}$, we lose no information whatsoever by breaking the link between observation i in $\mathbf{Y}$ and observation i in $\mathbf{X}$. As one can imagine, these tests are very versatile, due to their minimal assumptions. However, this flexibility comes with a heavy computational burden. This is especially true in the context of multiple hypothesis tests, where the number of permutations needs to be large enough to allow a significance assessment even after correction for multiple testing.

In this thesis, I focus on inference performed under the double Wishart setting. This setting provides a very powerful inferential framework for some of the dimension reduction methods discussed above, but also for other multivariate techniques. I first give a short review of

Union-Intersection Test (UIT), which forms the main link between dimension reduction methods and the double Wishart setting. I then give a general overview of this setting, with a short collection of examples. Finally, I review the different approaches to estimating the null distribution of the test statistic.

2.2.1 Union-Intersection Tests

In many hypothesis-testing problems in multivariate analysis, there is no uniformly most powerful (unbiased) test. Accordingly, there has been several approaches advocated for NHST, each with their advantages and disadvantages. Two of the most common approaches are the classical Likelihood Ratio Test (LRT) and the UIT. The former test is well-known from introductory mathematical statistics courses: we look at the ratio of the maximum likelihood under an alternative and a null hypothesis, and large values of this ratio provide evidence against the null hypothesis. The latter approach does not require an alternative hypothesis, but it is probably not as familiar to the average reader. I start with a definition:

DEFINITION 2 ([Mardia et al., 1979, Chapter 5]). A UIT for the null hypothesis H_0 is a test whose rejection region R can be written as a union of rejection regions R_a for component hypotheses H_{0a}, and where H_0 can be written as

$$H_0 = \cap_a H_{0a}.$$

To help clarify this definition, I will give an example following Mardia et al. [1979]. Let $\mathbf{Y}$ be an $n \times p$ matrix whose rows are normally distributed $N_p(\mu, \Sigma)$. We assume that Σ is known, and we are interested in the null hypothesis $H_0 : \mu = 0$. For any p-dimensional, non-random vector w, consider $w^T\mathbf{Y}$. We know that it also follows a (univariate) normal distribution $N(w^T\mu, w^T\Sigma w)$. The null hypothesis H_0 induces a similar null hypothesis for all nonzero

vector w:

$$H_{0,w} : w^T \mu = 0.$$

Moreover, if $H_{0,w}$ is true for all $w \neq 0$, we can infer H_0 must be true. In other words, we can view H_0 as the *intersection* of several univariate null hypotheses $H_{0,w}$:

$$H_0 = \cap_{w \neq 0} H_{0,w}.$$

Now, since $w^T \mathbf{Y}$ is univariate and normally distributed, we can easily build a test for $H_{0,w}$ by looking at the ratio $z_w = \frac{w^T \mathbf{Y}}{w^T \Sigma w}$. We can then build a rejection region

$$R_w = \{z_w \mid z_w^2 > c^2\}$$

for some critical value c. By combining these multiple rejection regions, we can get a rejection region for the multivariate null hypothesis H_0 as a *union* of the rejection regions R_w:

$$R = \cup_{w \neq 0} R_w.$$

We can simplify this further by noting that we do not reject the null hypothesis if and only if $z_w^2 \leq c^2$ for all $w \neq 0$; this is equivalent to not rejecting H_0 if and only if

$$\max_w z_w^2 \leq c^2.$$

Of course, more complicated settings will lead to a more complex rejection region for the UIT. But the same principles apply broadly.

It is worth noting that the above maximisation highlights how natural UIT can be in the context of dimension reduction. Indeed, we can see that z_w^2 above could be replaced by the criterion being maximised in both PCEV and CCA. For each possible component $w^T \mathbf{Y}$,

we get a univariate test, and the solution to the dimension reduction problem corresponds to when $R^2(w)$ or $\text{Corr}(u^T\mathbf{Y}, v^T\mathbf{X})^2$ is maximised. Following the UIT framework described above, we see that we can construct a global test statistic by looking at the maximum of $R^2(w)$ or $\text{Corr}(u^T\mathbf{Y}, v^T\mathbf{X})^2$ over all possible linear combinations.

2.2.2 Double Wishart Problems

As I discussed in Section 2.1.1, both PCEV and CCA can be performed by optimising a Rayleigh quotient, where the numerator and denominator depend on the matrices $\mathbf{Y}$ and $\mathbf{X}$. Moreover, Proposition 1 described how the optimisation of this Rayleigh quotient is related to a determinantal equation of the form

$$\begin{aligned} \det(A - \lambda B) = 0, \text{ or} \\ \det(A - \lambda(A + B)) = 0. \end{aligned} \tag{2.10}$$

For both PCEV and CCA, assuming the null hypothesis of no association between $\mathbf{Y}$ and $\mathbf{X}$ holds, we can derive two important consequences:

- A and B are Wishart-distributed with the same scale matrix Σ;
- A and B are independent.

These two consequences are at the heart of what is called the *double Wishart problem*. More generally, we have the following definitions.

DEFINITION 3. 1. Let $\mathbf{Z}$ be an $n \times p$ normal data matrix: each row is an independent observation from $N_p(0, \Sigma)$. A $p \times p$ matrix $A = \mathbf{Z}^T\mathbf{Z}$ is then said to have a *Wishart distribution* $A \sim W_p(\Sigma, m)$.

2. Let $A \sim W_p(\Sigma, m)$ be independent of $B \sim W_p(\Sigma, n)$. Then the largest root $\hat{\lambda}$ of Equation 2.10 is called the *largest root statistic* and a random variate having this

distribution is denoted $\theta_1(p, m, n)$, or $\theta_{1,p}$ for short.

We could have a similar definition for the distribution $\theta_{i,p}, i = 1, \ldots, \min(p, n)$ of all roots of Equation 2.10. However, for the purpose of this thesis, I will only focus on the largest root.

The UITs for PCEV and CCA correspond to this largest root statistic. Moreover, note that we can see both PCA and RDA as limiting cases of the double Wishart problem, where B converges to a (non-stochastic) identity matrix. For this reason, the limiting case is usually referred to as the *single Wishart problem*. The distribution of its largest root corresponds to the distribution of the largest eigenvalue of a Wishart matrix.

The double Wishart problem goes beyond dimension reduction, and indeed it underlies many multivariate techniques (e.g. MANOVA, tests of equality of covariances). Johnstone [2009] contains many examples of double Wishart problems, and I also review some of them in Chapter 4.

2.2.3 Largest Root Distribution

It turns out that hypothesis testing performed using LRT typically uses test statistics constructed using all roots from the determinantal equation 2.10 (e.g. Wilk's Lambda in MANOVA); whereas testing performed using UIT typically uses only the extreme eigenvalues (e.g. Roy's Largest Root in MANOVA), as seen above. Accordingly, I will begin our discussion of the largest root distribution with the joint distribution of all roots to the determinantal equation 2.10 (in its second form). This distribution is given by [Muirhead, 2009b]:

$$f(\theta_1, \ldots, \theta_p) = C \prod_{i=1}^{p} (1 - \theta_i)^{(m-p-1)/2} \theta_i^{(n-p-1)/2} \prod_{i<j} |\theta_i - \theta_j|,$$

where

$$C = \frac{\pi^{p^2/2}}{\Gamma_p\left(\frac{1}{2}p\right)} \frac{\Gamma_p\left(\frac{1}{2}(n+m)\right)}{\Gamma_p\left(\frac{1}{2}n\right)\Gamma_p\left(\frac{1}{2}m\right)}.$$

Recall that $\Gamma_p(\cdot)$ is the *multivariate Gamma function*:

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(a + (1-j)/2).$$

This result has been known for several decades, and indeed it is a simple consequence of a change of coordinates. However, going from the joint distribution of the roots to the marginal distribution of the *largest* root increases the level of complexity tremendously. In turn, this complexity has led to many computational challenges. This has already been discussed elsewhere (see for example Johnstone [2008]), but for the sake of completeness I review it here and add some other, more recent, developments.

The first important step was provided by Constantine [1963], who showed that the marginal distribution can be expressed in terms of a hypergeometric function with matrix argument:

$$P(\theta_{1,p} < x) = C_{1,p} x^{pm/2} \, {}_2F_1\left(\frac{m}{2}, \frac{-(n+p+1)}{2}; \frac{m+p+1}{2}; xI\right),$$

where

$$C_{1,p} = \frac{\Gamma_p^{(1)}\left(\frac{m+n}{2}\right)\Gamma_p^{(1)}\left(\frac{p+1}{2}\right)}{\Gamma_p^{(1)}\left(\frac{m+p+1}{2}\right)\Gamma_p^{(1)}\left(\frac{n}{2}\right)}.$$

When $n+p+1$ is a positive, even integer, $P(\theta_{1,p} < x)$ can actually be expressed as a terminating series involving zonal polynomials [Muirhead, 2009b, Corollary 10.6.9].

Koev and Edelman [2006] showed how the hypergeometric functions with a matrix argument can be efficiently computed by leveraging recursion relations of Jack functions (which are a generalization of zonal polynomials). However, these approximations are relatively accurate only for small values of p, n, m. For more information about the algorithmic and computational approaches to estimating hypergeometric functions with a matrix argument,

see Dumitriu and Koev [2008] and the references therein.

Using a different approach, Gupta and Richards [1985] used results from De Bruijn [1955] to show that the hypergeometric function evaluated at a scalar multiple of the identity matrix is related to the Pfaffian of a matrix whose entries are expressible in terms of simpler, classical hypergeometric functions. Recall that the Pfaffian of a skew-symmetric matrix A is the polynomial $\mathcal{P}$ that satisfies the relationship $\mathcal{P}(A)^2 = \det(A)$. This result was then used to show that the marginal distribution $P(\theta_{1,p} < x)$ can itself be expressed as the Pfaffian of a matrix whose entries are double integrals.

This theoretical result was recently implemented by Butler and Paige [2011]. The authors give a series expansion of the double integrals, and they also provide numerical results about their implementation. Unfortunately, just as the work of Koev and Edelman [2006], the computational approach is limited to small values of p, n, m. Furthermore, in some cases, these computations can take several hours, making them impractical for large datasets such as is common neuroimaging and genomic studies.

Building on the work of De Bruijn [1955] and Gupta and Richards [1985], and also on his previous work on the largest eigenvalue of Wishart matrices [Chiani, 2014], Chiani [2016] exploited the Pfaffian relationship to provide a computationally efficient and exact way of computing the distribution $P(\theta_{1,p} < x)$. His approach, based on recursions, is summarised in the following theorem:

THEOREM 4 ([Chiani, 2016, Theorem 1]). Let $\mathcal{B}(x; \alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$ be the *incomplete beta function*. The CDF of the largest root $\theta_{1,p}$ is proportional to the Pfaffian of a skew-symmetric matrix:

$$P(\theta_{1,p} < x) = C\sqrt{\det(A(x))},$$

where C is given above. When p is even, the elements of the $p \times p$ skew-symmetric matrix

$A(x)$ are given by

$$a_{ij}(x) = \mathcal{E}(x; m+j, m+i) - \mathcal{E}(x; m+i, m+j), \quad i, j = 1, \ldots, p,$$

where

$$\mathcal{E}(x; a, b) = \int_0^x t^{\alpha-1}(1-t)^n \mathcal{B}(t; b, n+1)dt.$$

When p is odd, the elements of the $(p+1) \times (p+1)$ skew-symmetric matrix $A(x)$ are given by as above for $1 \leq i, j \leq p$, and

$$\begin{aligned} a_{i,p+1}(x) &= \mathcal{B}(x; m+i, n+1), \quad i = 1, \ldots, p \\ a_{p+1,j}(x) &= -a_{j,p+1}(x), \quad j = 1, \ldots, s \\ a_{p+1,p+1}(x) &= 0. \end{aligned}$$

Note that $a_{ij}(x) = -a_{ji}(x)$ and $a_{ii}(x) = 0$.

Moreover, the elements $a_{ij}(x)$ can be computed iteratively, starting from the beta function, without numerical integration of series expansion (see Algorithm 2).

However, one caveat that Chiani fails to mention is that his algorithm is numerically unstable for large values of p. This is due to the numerical underflow and overflow problems arising from the computation of both the constant C and the determinant of A. To my knowledge, accurate calculations require arbitrary-precision arithmetic, which can be computationally demanding. This is the approach I used when implementing Algorithm 2 in the `R` package `rootWishart` using the `boost` and `eigen` C++ libraries (see also Chapter 4).

In a landmark paper, Johnstone [2008] showed that accuracy and computational efficiency for the largest root distribution could be attained by first transforming the largest root and then estimating its distribution using the Tracy-Widom distribution of order 1. His main

Algorithm 2 CDF of the largest eigenvalue of Roy's test

Input: s, m, n, x
Output: $P(\theta_{1,p} < x)$
  $\mathbf{A} = \mathbf{0}$
  for $i = 1 \to p$ **do**
    $b_i = 0.5\mathcal{B}(x; m+i, n+1)^2$
    for $j = i \to p-1$ **do**
      $b_{j+1} = \frac{(m+j)b_j - \mathcal{B}(x; 2m+i+j, 2n+2)}{m+j+n+1}$
      $a_{i,j+1} = \mathcal{B}(x; m+i, n+1)\mathcal{B}(x; m+j+1, n+1) - 2b_{j+1}$
    end for
  end for
  if p is odd **then**
    for $i = 1 \to p$ **do**
      $a_{i,p+1} = \mathcal{B}(x; m+i, n+1)$
    end for
  end if
  $\mathbf{A} = \mathbf{A} - \mathbf{A}^T$
  $C = \pi^{0.5p} \prod_{k=1}^{p} \frac{\Gamma(0.5(k+2m+2n+p+2))}{\Gamma(0.5k)\Gamma(0.5(k+2m+1))\Gamma(0.5(k+2n+1))}$
  return $P(\theta_{1,p} < x) = C\sqrt{|\mathbf{A}|}$

result is given below:

THEOREM 5 ([Johnstone, 2008]). Assume $\mathbf{A} \sim W_p(\Sigma, m)$ and $\mathbf{B} \sim W_p(\Sigma, n)$ are independent, with Σ positive-definite. Let λ be the largest root of Equation 2.10. As $p, m, n \to \infty$, we have

$$\frac{\operatorname{logit} \lambda - \mu}{\sigma} \xrightarrow{\mathcal{D}} TW(1),$$

where $TW(1)$ is the Tracy-Widom distribution of order 1 (cf. Tracy and Widom [1996]), and μ, σ are defined as follows:

$$\mu = 2 \log\left(\tan\left(\frac{\varphi + \gamma}{2}\right)\right)$$
$$\sigma^3 = \frac{16}{(m+n+1)^2}\left(\sin^2(\varphi+\gamma)\sin\varphi\sin\gamma\right)^{-1},$$

where

$$\sin^2(\gamma/2) = \frac{\min(p,n) - 1/2}{m+n+1}$$
$$\sin^2(\varphi/2) = \frac{\max(p,n) - 1/2}{m+n+1}.$$

The Tracy-Widom distribution of order 1 is defined by

$$F_1(s) = \exp\left\{-\frac{1}{2}\int_s^\infty q(x) + (x-s)q^2(x)dx\right\}, \quad s \in \mathbb{R},$$

where q solves the nonlinear Painlevé II differential equation:

$$\ddot{q}(x) = xq(x) + 2q^3(x);$$

asymptotically, we have $q(x) \sim Ai(x)$, where Ai is the Airy function [Hastings and Mcleod, 1980]. This distribution was found by Tracy and Widom [1996] as the limiting law of the largest eigenvalue of an $n \times n$ Gaussian symmetric matrix. The distribution can be computed numerically to the desired degree of accuracy, or it can itself be estimated using a scaled and shifted gamma distribution [Chiani, 2014].

However, a limitation shared by both Chiani [2016] and Johnstone [2008] is that they do not apply to settings where the dimension p of the Wishart matrix A is larger than its degrees of freedom m. Unfortunately, this is almost always the case with high-dimensional data. In Chapter 4, I show that the result from Johnstone [2008] can still be leveraged in high dimension to construct an empirical estimator of the largest root distribution. I provide numerical evidence that after transformation, the largest root distribution can still be approximated by a Tracy-Widom distribution, and I show how using a small number of permutations, we can estimate the location-scale parameters of this distribution. In

this way, we are still able to construct valid p-values in high dimensions while keeping the computational burden relatively low.

In this chapter, I reviewed some of the most frequently used dimension reduction methods in genomics and neuroimaging. I presented two general frameworks that included only subsets of these methods. Therefore, I suggested using a third one. This framework, that I call *iterative optimisation*, also naturally extends to a regularised setting by using a regularised objective function f. In this way, it also accommodates several of the high-dimensional approaches described above. Chapter 4 introduces an empirical estimator based on the Tracy-Widom distribution to compute valid high-dimensional p-values. Finally, in Chapter 5, I apply these methods to the analysis of DNA methylation and anti-citrullinated protein antibody (ACPA) levels in individuals from a nested case-control study.

Chapter 3

Principal component of explained variance: an efficient and optimal data dimension reduction framework for association studies

Preamble to Manuscript 1. As I discussed in Chapter 2, most approaches to dimension reduction in high-dimensional data involve some form of regularisation. The amount of regularisation, or penalisation, is typically controlled by at least one parameter that needs to be selected by the user. The approach normally recommended for tuning this parameter is cross-validation: for a range of values, we compute a cross-validated loss function (e.g. mean squared error, log-likelihood), and we select the parameter value that minimises this loss.

Although this approach has desirable theoretical properties, it imposes two important burdens on the user:

- The cross-validation can be computationally intensive.

- Sample sizes in genomic and neuroimaging studies can be small, making the data splitting unstable.

The first point becomes particularly acute when we want to perform hypothesis testing and control for multiple tests. Indeed, the null distribution of the test statistics is often unknown for regularised estimators, and therefore researchers usually use permutation tests to obtain p-values. And when performing multiple tests, as is often the case in genomic and neuroimaging studies, the number of permutations required to reject any null hypothesis at a given significance level increases with the number of tests.

The purpose of the manuscript presented in this chapter was to develop a dimension reduction method that did not require any tuning parameter. As it turns out, I was able to prove that PCEV can be broken down in two stages when the covariance matrix is block-diagonal: 1) perform PCEV on each block separately; 2) then perform PCEV again on the components obtained at the first stage. This approach was partly motivated by the correlation patterns observed in DNA methylation data: the methylation levels of nearby CpG dinucleotides are highly correlated [Eckhardt et al., 2006]. Moreover, as described in detail below, this two-stage approach is quite robust to violations of this block-diagonal assumption, and therefore it can be used much more broadly.

Principal component of explained variance: an efficient and optimal data dimension reduction framework for association studies

Maxime Turgeon[1,2,3], Karim Oualkacha[4], Antonio Ciampi[1,3], Golsa Dehghan[1], Brent W. Zanke[5], Andréa L. Benedet[6], Pedro Rosa-Neto[6,7,8], Celia MT. Greenwood[1,2,3,9,10], Aurélie Labbe[1,3,7,11] for the Alzheimer's Disease Neuroimaging Initiative[12].

[1] *Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2] *Lady Davis Institute for Medical Research, Montréal*

[3] *Ludmer Centre for Neuroinformatics and Mental Health, Montréal*

[4] *Département de mathématiques, Université du Québec à Montréal*

[5] *Arctic Diagnostics Inc.*

[6] *Translational Neuroimaging Laboratory, McGill Center for Studies in Aging, Montréal, Canada*

[7] *Department of Psychiatry, McGill University*

[8] *Department of Neurology and Neurosurgery, McGill University*

[9] *Department of Human Genetics, McGill University*

[10] *Department of Oncology, McGill University*

[11] *Douglas Mental Health University Institute, Verdun*

[12] *Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:* `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

Paper published in *Statistical Methods in Medical Research* (2018). 27(5):1331-1350. DOI 10.1177/0962280216660128.

Abstract

The genomics era has led to an increase in the dimensionality of the data collected to investigate biological questions. In this context, dimension-reduction techniques can be used to summarize high-dimensional signals into low-dimensional ones, to further test for association with one or more covariates of interest. This paper revisits one such approach, previously known as Principal Component of Heritability and renamed here as *Principal Component of Explained Variance* (PCEV). As its name suggests, the PCEV seeks a linear combination of outcomes in an optimal manner, by maximising the proportion of variance explained by one or several covariates of interest. By construction, this method optimises power but limited by its computational complexity, it has unfortunately received little attention in the past. Here, we propose a general analytical PCEV framework that builds on the assets of the original method, i.e. conceptually simple and free of tuning parameters. Moreover, our framework extends the range of applications of the original procedure by providing a computationally simple strategy for high-dimensional outcomes, along with exact and asymptotic testing procedures that drastically reduce its computational cost. We investigate the merits of the PCEV using an extensive set of simulations. Furthermore, the use of the PCEV approach will be illustrated using three examples taken from the epigenetics and brain imaging areas.

3.1 Introduction

In the omics era, a considerable amount of data is now routinely collected to investigate the relationships between a set of covariates of interest (X) and genetic, molecular, or clinical outcomes (Y). As such, a substantial proportion of the methodological research developed in the last decade has focused on the numerous statistical challenges and computational issues highlighted by the joint analysis of such high-dimensional correlated data. If we imagine a spectrum of models indexed by dimensionality, at one end, we have a model that attempts to accommodate all variables at once. This may lead to results that are difficult to interpret. At the other extreme of this spectrum, we have univariate models which treat each variable separately while ignoring the others. This approach may miss subtle signals arising from complex biological interactions and correlations. An intermediate approach would therefore either identify *a priori* relevant groups of variables Y on which to perform the analysis or reduce the dimensionality of the problem by summarising the data into meaningful components. A popular example of dimension reduction is Principal Component Analysis (PCA). This method seeks linear combinations of the original data that explain the maximum amount of variance.

In contrast to studies with high-dimensional *covariates*, here we focus on studies where the goal is to investigate the association between a set of (possibly high dimensional) correlated *outcomes* and one or more covariates of interest. In this context, several methods integrating both data reduction and association analysis simultaneously have been developed. For example, in a variant of Principal Component Regression (PCR), a PCA analysis is performed on the set of outcomes Y, producing a smaller number of components. These components are then used in an association test with the covariates of interest X. Although being widely used in practice, this method has very poor power in cases where the outcomes showing the greatest variability—hence those captured by the first principal components—are not the ones associated with the covariates of interest. Since outcomes strongly influenced by the

environment could easily have much larger variability than outcomes influenced by a Single Nucleotide Polymorphism (SNP), this is of particular concern in genetic studies. Therefore, PCA in this context is likely to find components where environmentally driven traits have the highest weights (since they are highly variable), whereas the weights associated with genetically-controlled traits may be very small if they are less variable across individuals. If an association analysis is then performed between such components $\tilde{Y}$ and genotype data X, it is likely that no association will be detected. Since the data reduction step in PCR is performed independently of the covariates, the low power in such situations is not surprising; as a consequence, other methods have been developed to jointly perform both data reduction and association analysis. For example, Partial Least Squares (PLS) regression [Abdi, 2010], Canonical Correlation Analysis (CCA) [Härdle and Simar, 2007], and Linear Discriminant Analysis (LDA) [Friedman, 1989], are widely used component-based methods that can be employed to find, simultaneously, the "best" components describing two sets of variables (e.g. outcomes and covariates). This is achieved by maximising the association between them. The optimisation criterion for association differs between these methods: while CCA is based on maximising correlations, PLS optimises covariances. Methods of this type are very powerful by construction and have been extended in several ways to accommodate high-dimensional variables using regularisation techniques or sparsity measures. However, such extensions heavily depend on tuning parameters, and formal testing procedures are currently lacking.

Although PLS and CCA have been gaining popularity in genomics, these methods were originally developed in other fields and mainly for prediction purposes. Ott and Rabinowitz [1999] developed a closely related method implementing similar ideas to PLS and CCA, specifically for a genetic context. Using family data and with the aim of performing linkage analysis between multiple correlated outcomes Y and SNP genotypes X, the idea was to find the best linear combination of outcomes (i.e. a component) maximising the heritability at the SNP tested. This method, termed Principal Component of Heritability (PCH), has

unfortunately received little attention although it was later extended [Wang et al., 2007b, Klei et al., 2008, Fang et al., 2014a] to the more general setting of population-based studies and high-dimensional outcomes based on regularisation or sparsity techniques. Since heritability is simply defined in PCH as the proportion of variance in the outcomes Y explained by the SNP covariate, it can be seen as a method belonging to the family of PLS and CCA based on another criterion, i.e. the proportion of variance in the outcome variables explained by the covariates. Since the term "heritability" can be misleading in the context of population-based studies, and since the concept underlying PCH can also be used in a general setting outside genetics, we have renamed this approach Principal Component of Explained Variance (PCEV). We note that this method is also closely related to dual-scaling [Nishisato, 1995], originally introduced in the psychometrics literature.

In this paper, we present a completely new analytical framework based on the PCEV concept that copes smoothly with data of very high dimension at the outcome level, includes valid hypothesis tests, leads to interpretable results, and yet is computationally efficient. This approach is implemented in an `R` package, `pcev`, available on the Comprehensive R Archive Network (CRAN). Specifically, we first show that unlike the competing, similar approaches discussed earlier, PCEV has mathematical properties that allow inclusion of very high-dimensional outcomes without the need to rely on variable selection strategies that require selecting tuning parameters. Secondly, we show that PCEV can be extremely computationally efficient, unlike its competitors, by developing an exact testing procedure that does not rely on permutations. Therefore, this approach is entirely feasible for use in genome-wide studies (or studies with very high-dimensional response variables), such as we frequently encounter today.

We then investigate the merits of the PCEV using an extensive set of simulations. In the large n—small p setting, we focus our discussion on methods that do not require tuning parameters and those for which a testing framework is fully developed. For this reason,

we discarded from our comparison methods such as sparse PLS [Chun and Keleş, 2010], regularized CCA [Vinod, 1976, Leurgans et al., 1993], and the projection regression model proposed by Lin et al. [2012]. On the other hand, in the high-dimensional setting (i.e. small n—large p setting), we compare our extension of PCEV to commonly used approaches, such as lasso [Tibshirani, 1996] and sparse PLS (sPLS); these two methods do involve tuning parameters. We show that PCEV outperforms all these methods, while also being much more computationally efficient. Finally, the use of the PCEV approach is illustrated in three settings: i) using DNA methylation data derived from whole genome bisulfite sequencing (Y), we test if a region near the BLK gene on chromosome 8 is differentially methylated in B-cells compared to T-cells or monocytes; ii) using DNA methylation microarray data (Y), we perform a genome-wide gene-based association analysis of methylation with respect to cigarette smoking status; and iii) using brain imaging traits derived from [18F]Florbetapir PPositron Emission Tomography (PET) scans (Y), we perform an association test between amyloid-β accumulation and two sets of covariates: Alzheimer's disease status and a set of SNPs located near the APOE gene.

3.2 Method

The general methodological framework aims to simultaneously test a (possibly large) set of phenotypes or outcomes $\mathbf{Y}$, against a set of covariates $\mathbf{X}$. The method is evaluated through an extensive simulation study and then applied to three independent datasets.

3.2.1 General theoretical PCEV framework

We consider the following setting: let $\mathbf{Y}$ be a multivariate phenotype of dimension p (e.g. methylation values at p CpG dinucleotides, or brain imaging measures at p locations in the brain), let $\mathbf{X}$ be a q-dimensional vector of covariates of interest (e.g. smoking, cell type or SNPs) and let $\mathbf{C}$ be an r-dimensional vector of confounders (e.g. age or sex). We assume

that the relationship between $\mathbf{Y}$ and $\mathbf{X}$ can be represented via a linear model

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{\Gamma}\mathbf{C} + \mathbf{E}, \tag{3.1}$$

where $\mathbf{B}$ and $\mathbf{\Gamma}$ are $p \times q$ and $p \times r$ matrices of regression coefficients for the covariates of interest and confounders, respectively, and $\mathbf{E} \sim N_p(0, \mathbf{V}_R)$ is a vector of residual errors. This model assumption allows us to decompose the total variance of $\mathbf{Y}$, conditional on $\mathbf{C}$, as follows:

$$\begin{aligned}
\operatorname{Var}(\mathbf{Y} \mid \mathbf{C}) &= \operatorname{Var}(\mathbf{B}\mathbf{X} \mid \mathbf{C}) + \operatorname{Var}(\mathbf{E}) \\
&= \mathbf{B} \operatorname{Var}(\mathbf{X} \mid \mathbf{C}) \mathbf{B}^T + \mathbf{V}_R \\
&= \mathbf{V}_M + \mathbf{V}_R.
\end{aligned}$$

where $\mathbf{V}_M = \mathbf{B} \operatorname{Var}(\mathbf{X} \mid \mathbf{C}) \mathbf{B}^T$ is the *model* variance component and $\mathbf{V}_R$ is the *residual* variance component. PCEV seeks a linear combination of outcomes, $\mathbf{w}^T\mathbf{Y}$, which maximises the ratio $h^2(\mathbf{w})$ of variance being explained by the covariates $\mathbf{X}$:

$$\mathbf{w}_{\text{PCEV}} := \operatorname*{argmax}_{\mathbf{w}} h^2(\mathbf{w}),$$

where

$$h^2(\mathbf{w}) = \frac{\operatorname{Var}(\mathbf{w}^T\mathbf{B}\mathbf{X} \mid \mathbf{C})}{\operatorname{Var}(\mathbf{w}^T\mathbf{Y} \mid \mathbf{C})} = \frac{\mathbf{w}^T\mathbf{V}_M\mathbf{w}}{\mathbf{w}^T(\mathbf{V}_M + \mathbf{V}_R)\mathbf{w}}.$$

The original PCH, as presented by Ott and Rabinowitz [1999] relied on the same variance decomposition and the same optimization problem described above. However, genetic data from families were used to estimate the genetic variance component $\mathbf{V}_M$. In contrast, Klei et al. [2008] and Lin et al. [2012] extended the idea to population data using a linear model assumption.

It has been shown that $\mathbf{w}_{\text{PCEV}}$ is the solution to the generalised eigenvector problem $\mathbf{V}_M\mathbf{w} =$

$\lambda \mathbf{V}_R \mathbf{w}$ [Ott and Rabinowitz, 1999], and therefore standard linear algebraic results can be used to get a closed form solution $\mathbf{w}_{\text{PCEV}}$. Although there are some similarities between PCA and PCEV since both methods seek a linear combination of outcomes optimising a given criterion, we recall that PCA reduces the dimension of $\mathbf{Y}$ by looking for a linear combination of its components with maximal variance, independently of the covariates $\mathbf{X}$. Furthermore, an important difference between PCA and PCEV is in the number of components one can extract. While the number of components to select in PCA is usually left to the user, and the maximum number of extracted components is bounded above by p, the maximum number of components that can be extracted for PCEV is bounded above by the number of covariates q. Therefore, if we are only interested in one covariate, only one PCEV can be extracted; this follows from considering the rank of the matrix $\mathbf{V}_R^{-1}\mathbf{V}_M$.

3.2.2 PCEV with high-dimensional data

When the number of response variables p is larger than the sample size n, a naïve implementation of PCEV will fail. To ensure the uniqueness of the solution to the maximisation process, the invertibility of the residual matrix $\mathbf{V}_R$ is required; therefore, an accurate estimation of the matrix $\mathbf{V}_R$ requires $n > p$. This limitation has led to the introduction of regularisation techniques for the estimation of $\mathbf{w}$, reminiscent of ridge regression [Hoerl and Kennard, 1970] and lasso [Tibshirani, 1996]. We stress once again that these methods require parameters that are computationally expensive to compute. For this reason, we propose a novel alternative, namely a *block approach* to the estimation of PCEV. Assume we can partition $\mathbf{Y}$ into blocks (or clusters) such that the number of components in a given block is small enough, i.e. smaller than n. We can then perform PCEV and get a linear combination $\tilde{Y}_j$ of the traits belonging to the jth block, for each block $j = 1, \ldots, b$. We then obtain a new multivariate pseudo-phenotype $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_b)$, where each $\tilde{Y}_j$ is of dimension one. We can then perform PCEV again on $\tilde{\mathbf{Y}}$. Since the result is a linear combination of linear combinations, it is itself a linear combination of the original traits $\mathbf{Y}$. Although one might

think that this stepwise approach is an *ad-hoc* extension of the original PCEV approach, it has nonetheless a very appealing and relevant mathematical property, described in the following result:

THEOREM 1. Assume one can partition the outcomes $\mathbf{Y}$ into blocks in such a way that blocks are uncorrelated (i.e. outcomes lying in different blocks are uncorrelated). Then the linear combination (PCEV) obtained from the traditional approach and that obtained from the stepwise block approach described above are equal.

The proof of this result is given in Appendix A.1. Of course, if such a partition does not exist, there will generally be a difference between the two estimation procedures. However, through a series of simulations and analyses of data sets, we will show that the discrepancy is small. Therefore, comparable conclusions can be achieved even in the presence of some dependence between the blocks.

3.2.3 Test of significance

The PCEV methodology was first presented in the context of family-based studies, and the lack of a proper significance testing framework may have hindered its adoption in population-based studies. Klei et al. [2008] discussed such a framework, but their approach relies on computationally intensive sample splitting and resampling. In fact, here we are able to show that there is an analytic test of the null hypothesis $H_0 : h^2(\mathbf{w}) = 0$ that requires no resampling. This test is based on the largest eigenvalue λ of the matrix $\mathbf{V}_R^{-1}\mathbf{V}_M$, a test statistic used in multivariate analysis of variance [Everitt and Dunn, 1991]. When $\mathbf{X}$ consists of a single covariate, it is also known as the Wilks statistics; it can be shown that under the null hypothesis

$$\left(\frac{n-p-1}{p}\right)\lambda \sim F(p, n-p-1),$$

where $F(\nu_1, \nu_2)$ is the Fisher-Snedecor distribution, with degrees of freedom ν_1 and ν_2 (see Appendix A.2). The two matrices $\mathbf{V}_R$ and $\mathbf{V}_M$ are typically estimated during the classical PCEV process, which then allows us to compute the test statistic. When multiple covariates of interest are included in the model (e.g. multiple indicator variables for cell type composition, or multiple SNPs in a given genomic region), the statistic λ is better known as Roy's largest root test statistic (also called Roy's union-intersection test statistics) [Rencher and Christensen, 2012]. The asymptotic distribution of a suitable transformation of λ was derived by Johnstone [2008]. More details are given in Appendix A.3.

We note that the Wilks and Roy's largest root test statistics both rely on the assumption of normality of the outcomes, and the former also requires $n - p - 1 > 0$ (corresponding to the second degree of freedom of the F test). If these assumptions are not satisfied, permutation tests provide an adequate control of the Type I error and can be used as an alternative.

3.2.4 Variable importance

The PCEV framework described above allows reduction of the multiple testing burden by performing only a single test for a set of outcomes. If the test is significant, it is of great interest to identify the set of outcomes contributing the most to the global association detected. Here, we use the Variable Importance on Projection (VIP) defined for outcome j as $\text{VIP}_j = \text{Cor}(Y_j, Y_{\text{PCEV}})$, where $Y_{\text{PCEV}} = \mathbf{w}_{\text{PCEV}}^T(Y_1, \ldots, Y_n)$. This VIP measure can be signed (as a correlation) or unsigned (in absolute value). In the case of PCEV-block, the VIP measure is defined in the same way, i.e correlation between the *original* outcomes and the final PCEV component. We note that the VIP should be used as a means of ranking the contributions of each individual outcome to the overall association; as such, the actual values obtained are not necessarily interpretable. We provide examples of how to use the VIP values in the Results and Discussion sections below.

3.2.5 Defining blocks

Theorem 1 allows us to extend PCEV to the setting where there are more response variables than observations ($p \gg n$), through the use of blocks. In practice, the researcher needs to define these blocks prior to analysis, and correlation-based clustering methods (e.g. hierarchical and k-means clustering) can be used to obtain blocks that are minimally correlated with one another. However, in some settings, prior knowledge about the response variables can be leveraged in defining blocks. For example, DNA methylation at neighbouring CpG dinucleotides is known to be highly correlated [Eckhardt et al., 2006]. Therefore, in studies of DNA methylation, CpG nucleotides could be grouped based on distance. In the data analyses below, we give examples of both approaches.

3.2.6 Simulation study

Power and type I error of the PCEV approach are evaluated through extensive simulations assuming a sample size $n = 500$. In all simulations, a single continuous covariate $X \sim N(0, 1)$ is simulated and no confounder variables are included in the analysis. The multivariate outcomes $\mathbf{Y}$ are simulated under model (3.1) using a multivariate normal distribution of the residuals with variances equal to 1 and a block correlation structure made of five blocks of equal size. This correlation structure in $\mathbf{V}_R$ is governed by parameters ρ_w, defined as the within-block correlation and ρ_b, defined as the between-block correlation. The regression coefficients B_k $(k = 1, \ldots, p)$, corresponding to the kth outcome, are chosen to match the values of h^2 for each outcome, i.e. the proportion of variance explained by the covariate X, according to the relationship $B_k^2 = h^2/(1 - h^2)$. When multiple outcomes are associated with the variable X, we assume that each outcome explains the same proportion of variance h^2.

We compare the PCEV approach with a PCR analysis and a PCEV-block approach. The PCR approach was selected for comparison because it does not involve any calibration (e.g. sparse methods typically require a tuning parameter) and because it provides a convenient framework for significance testing. In the PCEV approach, p-values are computed using the

asymptotic test described above. In PCR, the first principal component of $\mathbf{Y}$ is extracted and then tested for association with the covariate. Finally, the PCEV-block approach is applied by using the same blocking structure as in the simulated data (i.e. five blocks). This analytical strategy is chosen to evaluate the performance of the PCEV-block approach in the situation where the mathematical property of independence between blocks is verified. Furthermore, we chose scenarios where the PCEV-block approach is not necessary ($p \ll n$), allowing us to compare the performance of the PCEV-block with respect to the original PCEV.

We investigate the performance of the PCEV approach by varying several simulation parameters, namely p (the number of outcomes), ρ_w, ρ_b and h^2 under five main scenarios described in Table 3.1. For each scenario, 500 simulated datasets were generated, and power was computed for $p = 20, 50, 100, 200, 300, 400$, $\rho_w = 0, 0.5, 0.7$ and $\rho_b = 0, 0.5, 0.7$ with $\rho_w \geq \rho_b$. Scenario 0 is designed to investigate the type I error rate of each method; therefore, none of the outcomes are associated with X. Scenarios $1 - 3$ are designed to examine power under various settings. Finally, scenario 4 evaluates the robustness of the PCEV-block approach with respect to the independence of blocks assumption. In this case, PCEV-block is applied using three strategies to define the blocks: i) $b = 5$ blocks are used, where blocks coincide with the simulation design; ii) $b = 10$ blocks are used, where the five blocks in the previous strategy are each split in two; iii) $b = 10$ blocks are used, where blocks are chosen at random. These strategies are labeled PCEV_{b1}, PCEV_{b2} and PCEV_{random}, respectively.

Table 3.1: Parameters used for the simulations. p represents the dimension of the outcome vector $\mathbf{Y}$; ρ_w and ρ_b represent the within-block and between-block correlation in the simulated outcomes (with $\rho_b < \rho_w$); h^2 represents the heritability of each outcome associated with X.

Scenario	h^2	Outcomes associated with X
0	0	None
1	1%	10 outcomes associated: $Y_1, \ldots, Y_5$ and $Y_{p-4}, \ldots, Y_p$
2	0.1%	25% of the outcomes in first two blocks
3	0.1%	25% of the outcomes in all blocks
4	0.1%	25% of the outcomes in all blocks

Finally, we also compare PCEV-block to other, traditional high-dimensional methods, namely lasso (as implemented in the R package `glmnet` [Friedman et al., 2010]) and sparse PLS (sPLS) [Chun and Keleş, 2010]. Regularized CCA (rCCA) is excluded due to its prohibitively high computational time (see Table 3.2). The simulation scenario here is slightly different than described above: the sample size is fixed at $n = 100$ and the number of response variables takes values $p = 100, 200, 300, 400$, and 500. The correlation structure between these variables varies in the same way as the other scenarios above. However, the association structure resembles that of Scenario 3: the response variables are partitioned into 10 blocks, and 25% of the variables in each block are truly associated with X with h^2 =1%.

Table 3.2: Average running times (in milliseconds) over 100 runs for four high-dimensional methods as a function of the number of outcomes p. The running times for lasso, sparse PLS and regularized CCA include the selection of a tuning parameter using 10-fold cross-validation and a grid of length 100.

	p				
	100	200	300	400	500
PCEV	10.68	19.33	29.61	42.17	57.64
lasso	567.79	515.75	558.58	635.14	539.76
sPLS	5418.79	8614.59	10295.60	12188.78	14582.90
rCCA	23723.56	91503.15	206320.93	378216.06	654003.12

Both lasso and sPLS require a tuning parameter. This parameter is selected using 10-fold cross-validation and with a grid of length 100. Moreover, both methods require a "reversed" approach, i.e. the multivariate response vector $\mathbf{Y}$ needs to be treated as the covariate vector, and X needs to be treated as the response variable. Finally, to compute power and perform hypothesis testing, we use the correlation between X and the predicted values $\hat{\beta}^T\mathbf{Y}$ as our test statistic for sPLS. For lasso, we compute a Likelihood Ratio Test (LRT) statistic comparing the selected model to the null model (i.e. including only an intercept). The null distribution of these test statistics is estimated using 500 permutations. The p-value for both PCEV-block approaches is also computed using 500 permutations.

In this high-dimensional simulation scenario, we also investigate the impact of block choice

on the analysis. To this end, we compare two PCEV-block approaches, one where the 10 true blocks are known *a priori*, and another where the response variables are randomly assigned to one of 10 blocks.

3.2.7 Datasets

All datasets used in this paper are publicly available or available upon request. The bisulphite sequencing data is included in the `R` package `pcev`. The Assessment of Risk for Colorectal Tumors in Canada (ARCTIC) methylation data have been deposited in dbGAP under accession number [phs000779.v1.p1]. Finally, the Alzheimer's Disease Neuroimaging Initiative (ADNI) data is archived in a secured and encrypted system provided by the LONI Image Data Archive (IDA). Applying for access to the data requires the submission of an online application form.

Analysis of bisulfite sequencing data

The dataset contains measurements of DNA methylation levels derived from bisulphite sequencing around the BLK gene, located on chromosome 8. A total of 40 different samples were analysed from three cell types: B cells (8 samples), T cells (19 samples), and monocytes (13 samples). These samples are derived from whole blood collected on a cohort of healthy individuals from Sweden. Data were sequenced on the Illumina HiSeq2000 system. Missing values are imputed using the mean of neighbouring sites and the imputed data is available in the `R` package `pcev`. The region analysed contains 24,068 CpG sites, from which we removed the duplicate sites and analyzed the 25% most variable sites (which coincide with the sites having the largest depth), for a total of 5,986 sites spanning a region of 2.5 Mb. Methylation levels at each CpG sites were measured using the logit of the methylated proportions. To apply the PCEV-block approach, we took advantage of the natural clustering as a function of physical distance between CpG sites on the DNA strand. We clustered CpG sites such that sites within 500kb were grouped together. In order to obtain clusters with less than

30 sites (to ensure $p < n$), we subsequently broke large clusters into smaller ones based on genomic distance. We obtained a total of 983 blocks of size ranging from 1 to 30 sites. Since this dataset is too large to use the classical PCEV method, we only used the PCEV-block framework to test the association between methylation levels and cell type (dichotomous variable, testing B-cells versus others).

Gene-based analysis of 450K Illumina methylation data

This analysis focuses on a sample of 1035 individuals who served as controls for the ARCTIC study [Cotterchio et al., 2000, Zanke et al., 2007]. Methylation at 485,512 CpG sites was measured in stored lymphocyte samples using the Infinium Human Methylation450 BeadChip. The CpG sites were allocated to 20,041 genes using the UCSC gene annotation; we also considered sites located 2kB upstream from the transcription starting site and downstream of the 3' end. With this definition, the genes under consideration contained anywhere between 2 and 288 CpGs, allowing us to perform PCEV using both the block and the classical approach since the number of CpGs per gene was substantially smaller than the number of individuals. Methylation data was normalised using Functional Normalisation [Fortin et al., 2014] and further adjusted to account for cell type mixture using the reference-based method proposed by Houseman et al. [2012] After normalisation, correction and exclusion of the X and Y chromosomes, we were left with 169,239 CpGs, located in 18,969 genes. In this analysis, we are interested in the association between methylation at the gene level and cigarette smoking status (which is dichotomous). By performing a gene-based analysis, the epigenome-wide significance threshold was lowered. Therefore, more associated genes were expected to be detected.

Analysis of brain imaging data

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investi-

gator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial Magnetic Resonance Imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

The analysis performed was based on data acquired from 340 participants from ADNI GO/2, for whom both genetic and PET data were available. [18F]Florbetapir PET imaging was employed to assess brain amyloid beta (Aβ) protein load using PET standardised uptake value ratios (SUVR) in 96 brain regions. These 96 Aβ levels represent the phenotypes of interest in our analysis. Genotype data was derived from the Illumina OmniQuad array [Potkin et al., 2009]. Association of Aβ levels with both diagnosis (Alzheimer versus others) and with 20 SNPs located in the PVRL2-TOMM40-APOE region on chromosome 19 were investigated. Phenotypes were adjusted for gender, age and education level directly within the PCEV framework.

3.3 Results

3.3.1 Simulation study

As one can see in Figure 3.1, Type I error for PCEV is well controlled and is not influenced by the number of variables nor by the correlation between outcomes. Here we have used the Wilks' test for the classical PCEV, and performance is as expected, even when the number of responses is as large as 400. The same figure, with 95% confidence intervals included, appears as Supplementary Figure A.1, showing that the observed variation can be explained by Monte Carlo errors. Supplementary Table A.1 shows that the Type I error for Wilks' test is also well controlled at significance levels as low as 10^{-4}.

Figures 3.2, 3.3, and 3.4 illustrate the power of the PCEV, PCEV-block, and PCR approaches in the next three simulated scenarios. In all cases, we observe that power of PCR is very

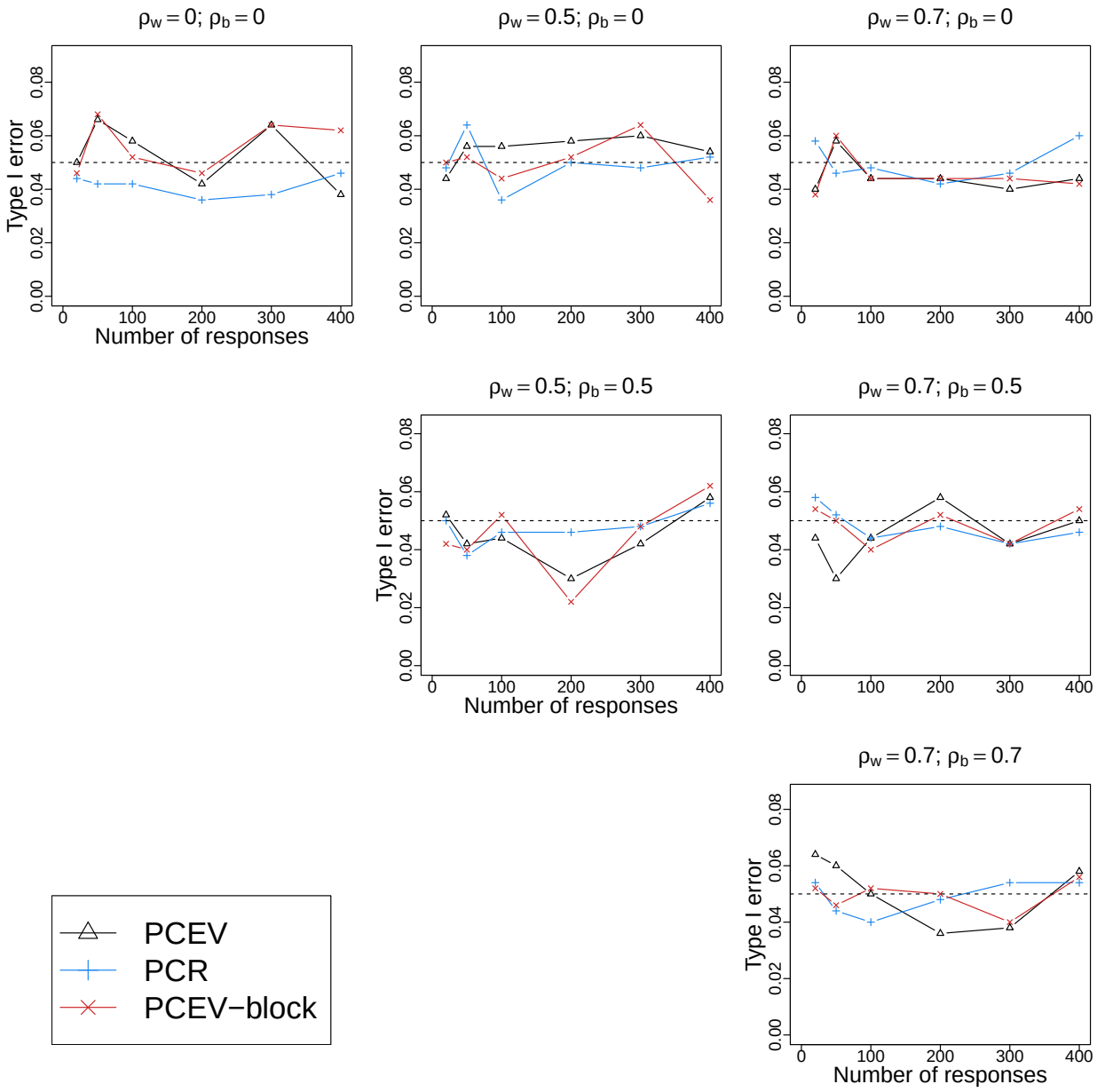
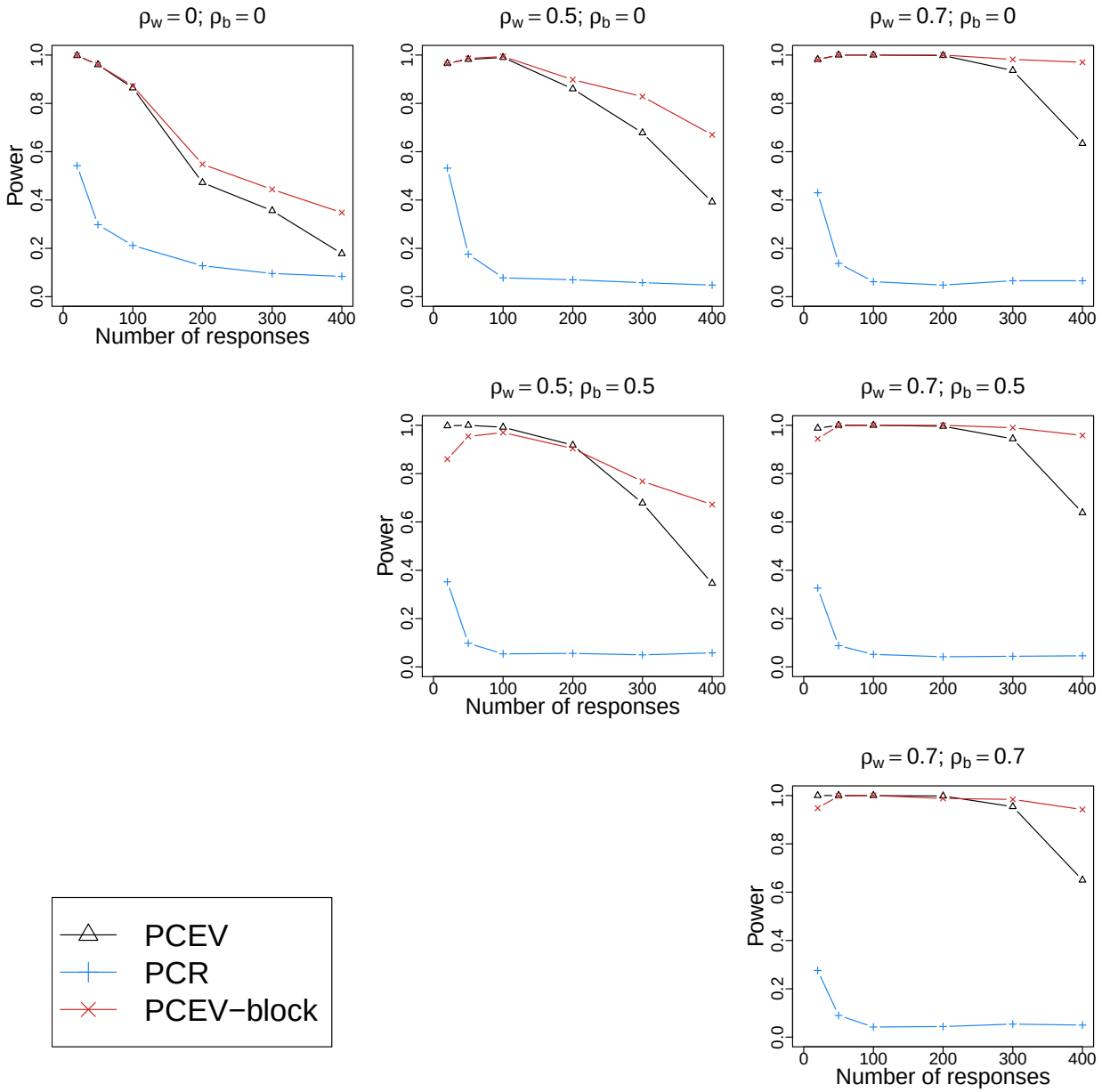


Figure 3.1: Type I error as a function of the correlation parameters ρ_w and ρ_b, and the number of responses p.

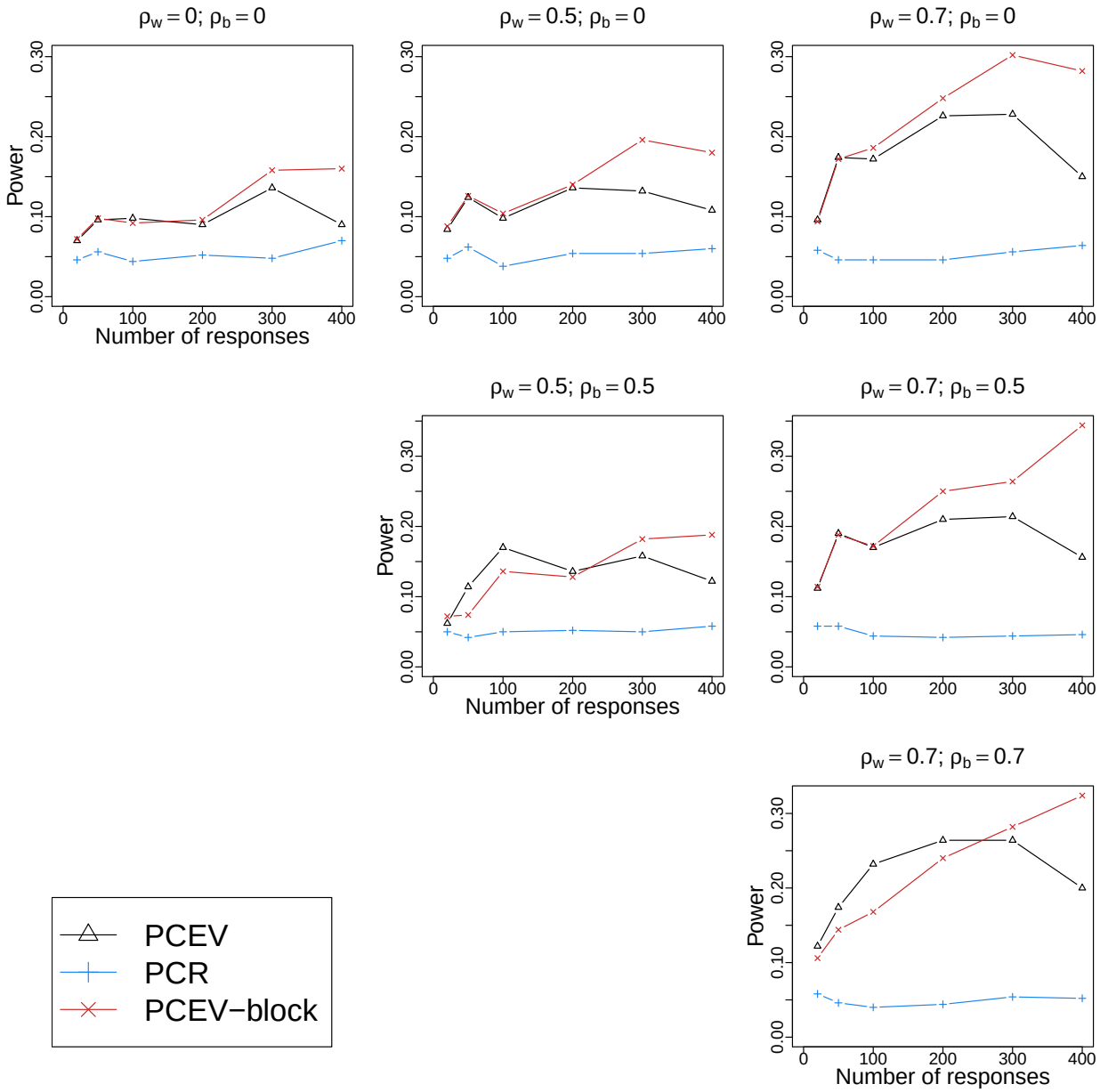
low compared to the PCEV approaches. This is as expected, since PCR is not optimal with respect to identifying components maximally related to the covariate X. In all figures, power can also be seen to increase with correlation between the outcomes. Such behaviour is also expected, since we can think of correlation as spreading out the signal across multiple outcomes Therefore, this signal is easier to detect with the PCEV methods.

Figure 3.2: Power of PCEV for scenario 1 as a function of the number of variables p and the correlation parameters ρ_w and ρ_b.



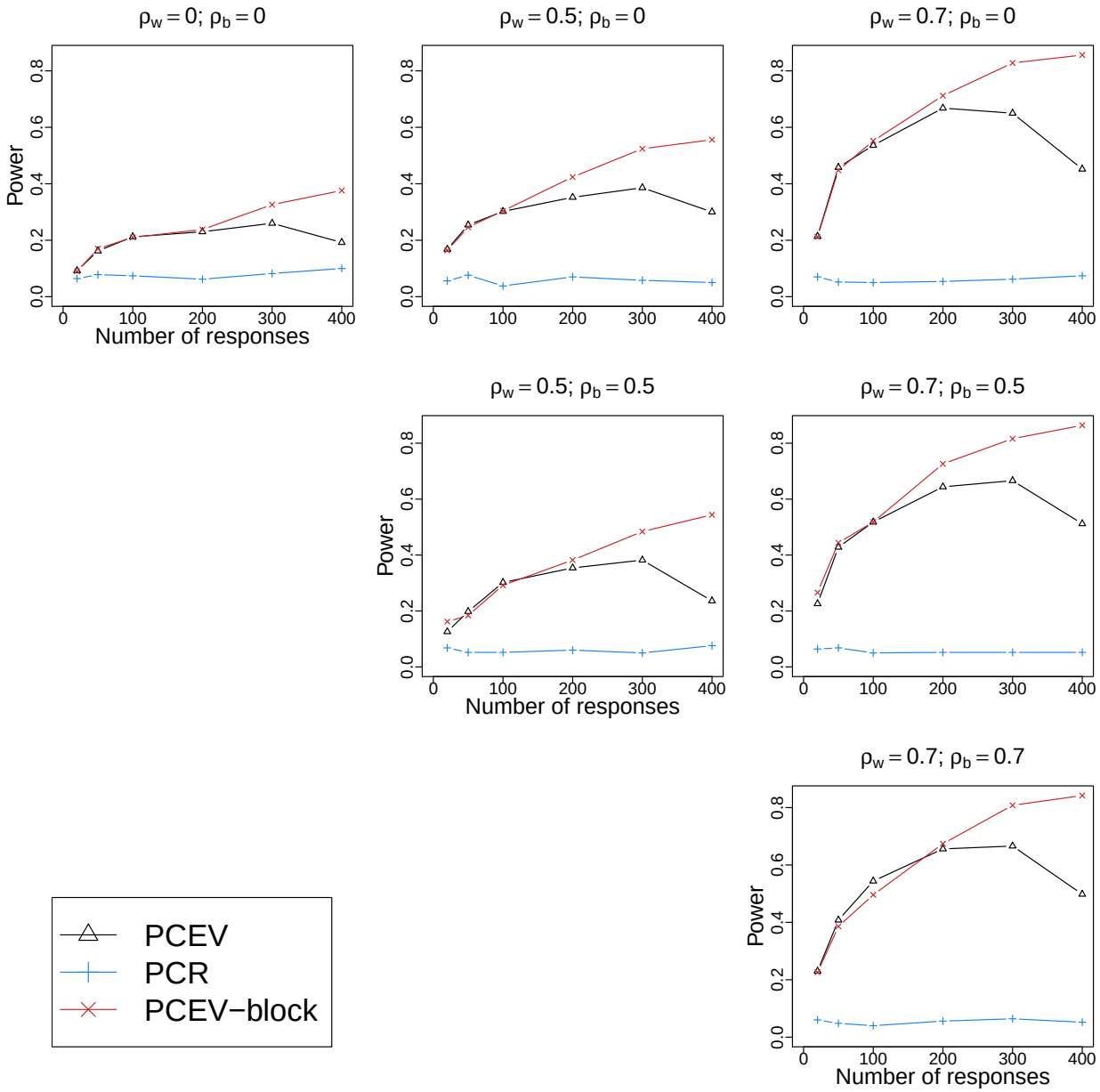
However, the relationships between power and patterns of associated responses are more nuanced. Increasing the number of outcomes, i.e. increasing p, has a detrimental effect on power when the number of outcomes associated with X is low (scenario 1, Figure 3.2), and this is true for any of the methods that we explored. However, when the signal to

Figure 3.3: Power of PCEV for scenario 2 as a function of the number of variables p and the correlation parameters ρ_w and ρ_b.



noise ratio is kept constant as p increases (scenarios 2-3-4), power tends to increase with p up to a point, and then decrease afterwards. This inflection point seems to be where the number of outcome variables is large enough that the residual variance matrices become ill-conditioned. Once such a state is reached, the power of the PCEV approach decreases quickly

Figure 3.4: Power of PCEV for scenario 3 as a function of the number of variables p and the correlation parameters ρ_w and ρ_b.



(Figures 3.3 and 3.4) with larger values of p. Crucially, however, the PCEV-block approach is not affected by this ill-conditioning phenomenon, since the residuals are estimated within each block. Therefore, each residual variance matrix is estimated from a much smaller number of outcomes. We also note that varying the between-block correlation ρ_B in the

simulation has a small impact on power, as it increases correlation between variables but does not spread the signal across outcomes. On the contrary, increasing the within-block correlation ρ_W increases power, as expected.

Figure 3.5 illustrates the sensitivity of the PCEV-block approach with respect to how the blocks are chosen. As can be seen, power is reduced when the blocks are more correlated with each other. This is as expected, since the PCEV-block approach relies mathematically on the independence between blocks. However, as we will see in the data analysis section, despite the reduction in power, the VIP measures are very robust to misspecifications of the independence assumption.

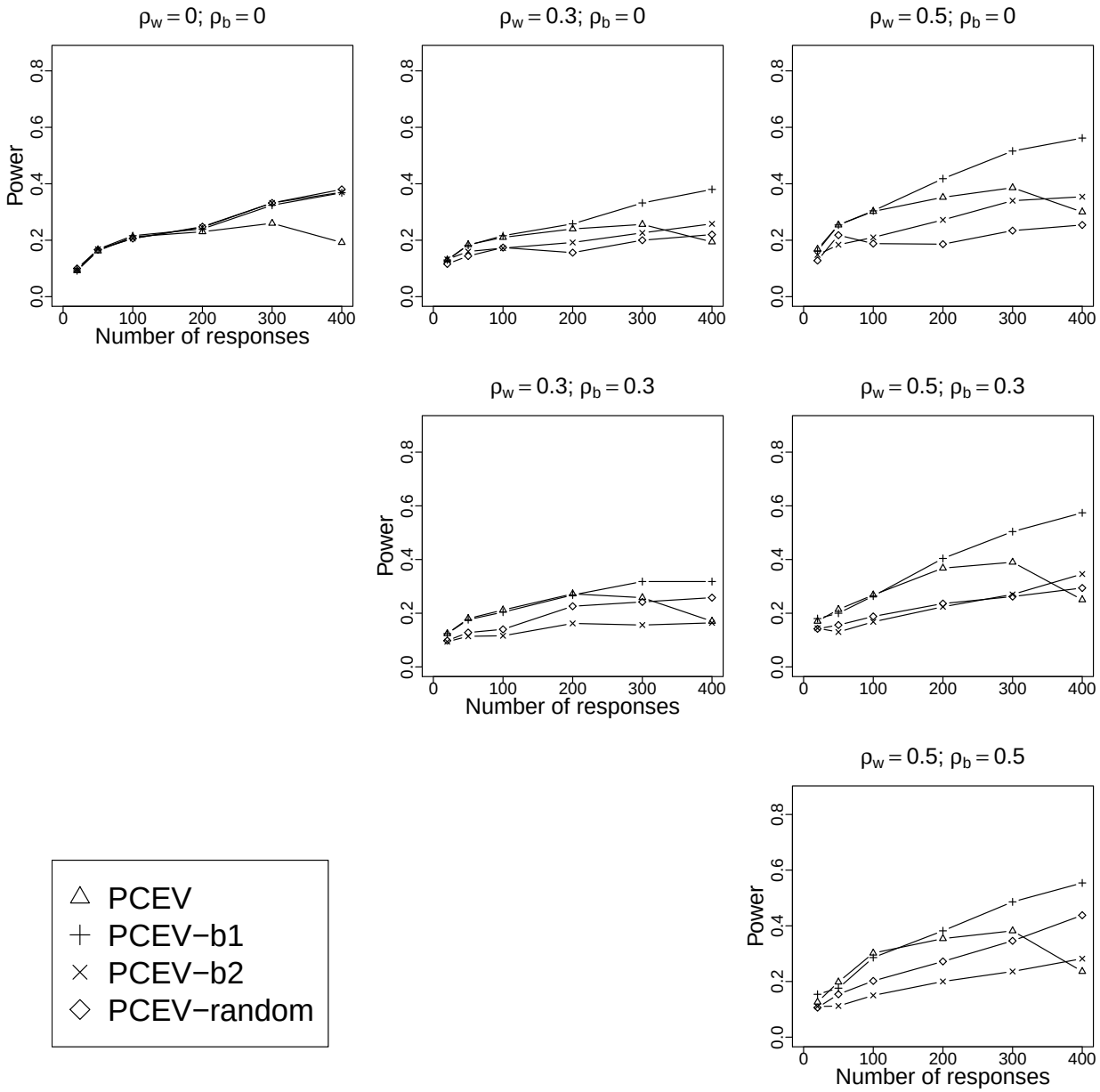
Finally, in Figure 3.6, we see that PCEV-block has better power than both lasso and sPLS. Moreover, we see that choosing the blocks randomly has little impact on power and we essentially get the same performance as when the blocks are known *a priori*. Moreover, Table 3.2 shows that PCEV is more computationally efficient than the other methods, with substantial benefit in some cases.

3.3.2 Data analysis

Analysis of bisulfite sequencing data

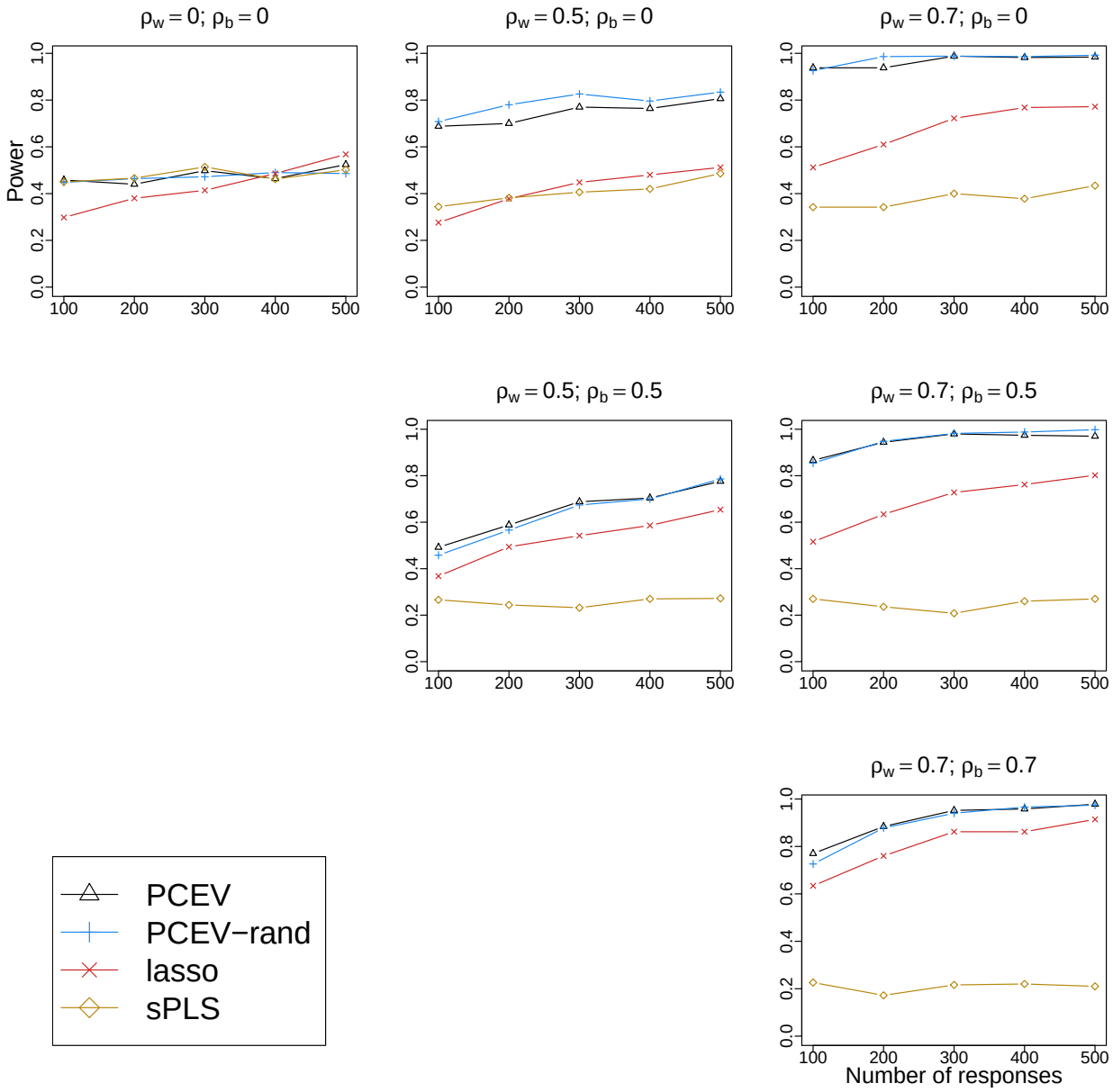
The genomic region we have analysed near the BLK gene is known to be hypomethylated in B-cells, compared to other cell types [Miceli-Richard et al., 2015]. Since we have only 40 samples and over 5,986 sites in our region of interest, it is impossible to perform the regular PCEV test. Therefore, we use the PCEV-block framework, and we expect to capture the region's association using only a single statistical test. Note that the methylation levels in this dataset show only a mild level of correlation: 50% of the CpG sites pairs have a correlation smaller than 0.15 and 99% of the pairs have a correlation smaller than 0.50. As a result, the assumption of independence between blocks is not strongly violated in this example.

Figure 3.5: Power of PCEV for scenario 4 as a function of the number of variables p and the correlation parameters ρ_w and ρ_b.



The p-value obtained with PCEV-block for all 5,986 sites simultaneously was 6×10^{-5}; this p-value was computed using 100,000 permutations. This result can be considered highly significant since only one test was performed. Furthermore, the results obtained are in excellent agreement with other approaches to analysis. For instance, Figure 3.7a shows the results us-

Figure 3.6: Power of PCEV for the high-dimensional scenario as a function of the number of variables p and the correlation parameters ρ_w and ρ_b.

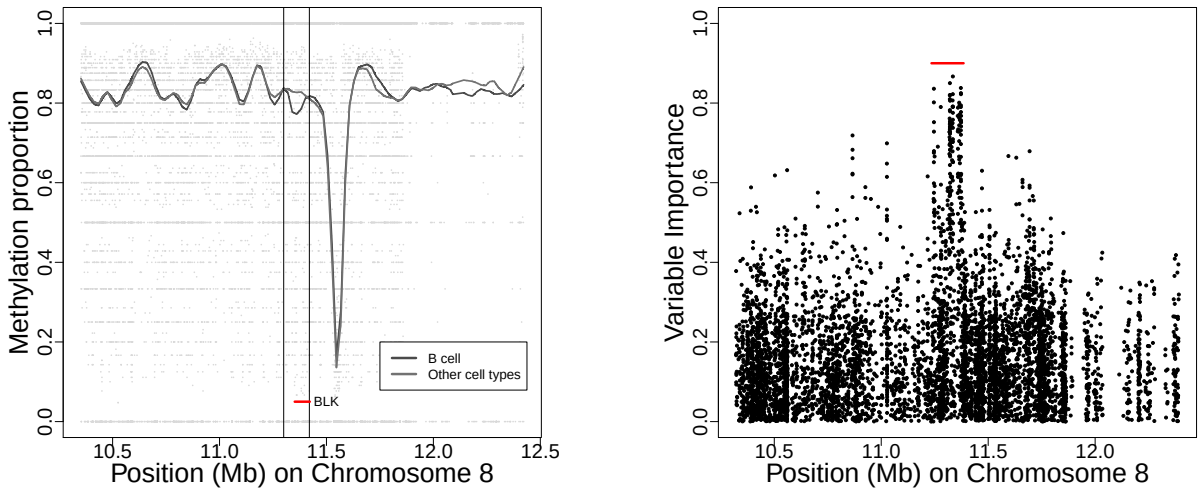


ing local linear regression to obtain a smoothed curve for each cell type in the region studied. As one can see, there is a region (delimited by vertical bars) comprising the BLK gene and a small upstream region where B-cells are differentially methylated compared to the two other cell types. Univariate regression analyses also confirmed this hypothesis (Supplementary

Figure A.3). Using the VIP measure, we are also able to identify which CpG sites contribute most to the association obtained (Figure 3.7b). Note that the region delimited in red in Figure 3.7b is identical to the region delimited by vertical bars in Figure 3.7a. Furthermore, there is a strong relationship between VIP measures and univariate p-values (Supplementary Figure A.4, left panel). Signed VIP measures can also be used to detect the direction of the association, and we see in Supplementary Figure A.4 (right panel) that such measures are highly correlated with the univariate slope regression coefficients.

Figure 3.7: Methylation sequencing data: analysis of the BLK region. (a) Analysis using local linear regression (LOWESS) on methylation values smoothed using BSmooth [Hansen et al., 2012]. B cells are hypomethylated at the differentially methylation region (DMR) delimited by vertical lines. (b) VIP (unsigned) measures for each CpG site.

(a) (b)



Gene-based analysis of 450K Illumina methylation data

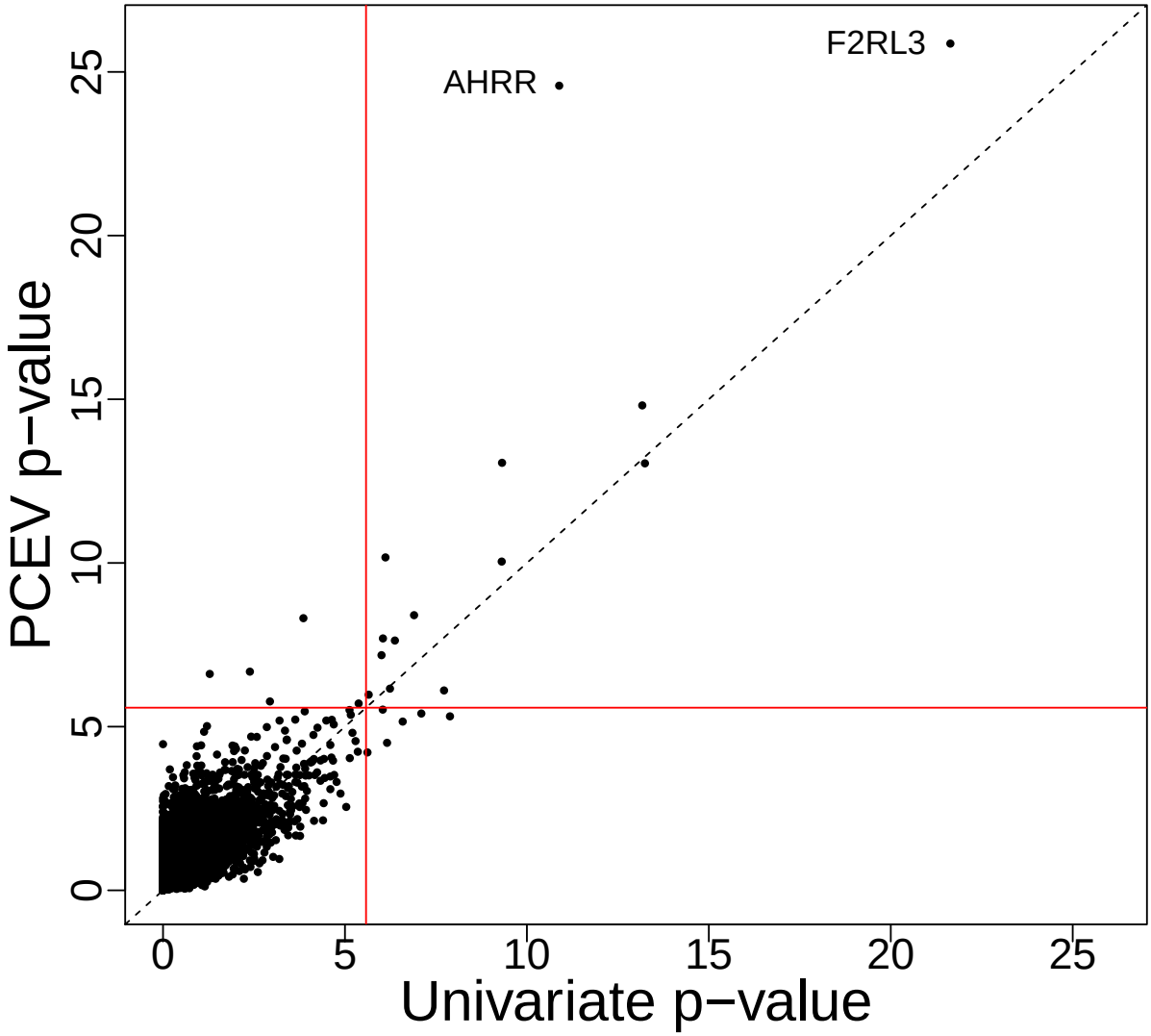
Analysis of the ARCTIC data was undertaken gene by gene, after selecting probes in or near each gene. In Figure 3.8, we show a scatter plot comparing the results of PCEV for each gene (without using the block version), and a gene level summary of the univariate p-values, where both are given on the negative log scale. The gene-level summary of the univariate p-values is defined as the minimum p-value among the CpGs tested in the gene, corrected

for the number of independent CpG sites in the gene, i.e. the minimum p-value is multiplied by the estimated number of independent CpGs. The effective number of independent tests was estimated using the method proposed by Gao et al. [2008]. By examining the smallest p-values, or the top of this scatter plot, Figure 3.8 suggests that PCEV tends to enhance power. This is particularly true for the genes with the strongest association. In Figure 3.9, we compare the VIP values obtained from both the classical PCEV and the block approach, for four different genes known to be associated with cigarette smoking: F2RL3, AHRR, RARA, and GNG12 [Breitling et al., 2011, Lee and Pausova, 2013]. The PCEV-block approach proceeded by defining three blocks in each of these genes, and by allocating CpG sites to blocks in a linear fashion (for example, if a gene contained 30 sites, the first 10 were assigned to one block, the next 10 to a second block and so on). We purposely chose a non-optimal block definition so that we could evaluate the robustness of the block method to a poor choice of block (here, blocks were not chosen to be independent). As we can see, the VIP values provide information that is very similar to the univariate p-values across these four genes.

Analysis of brain imaging data

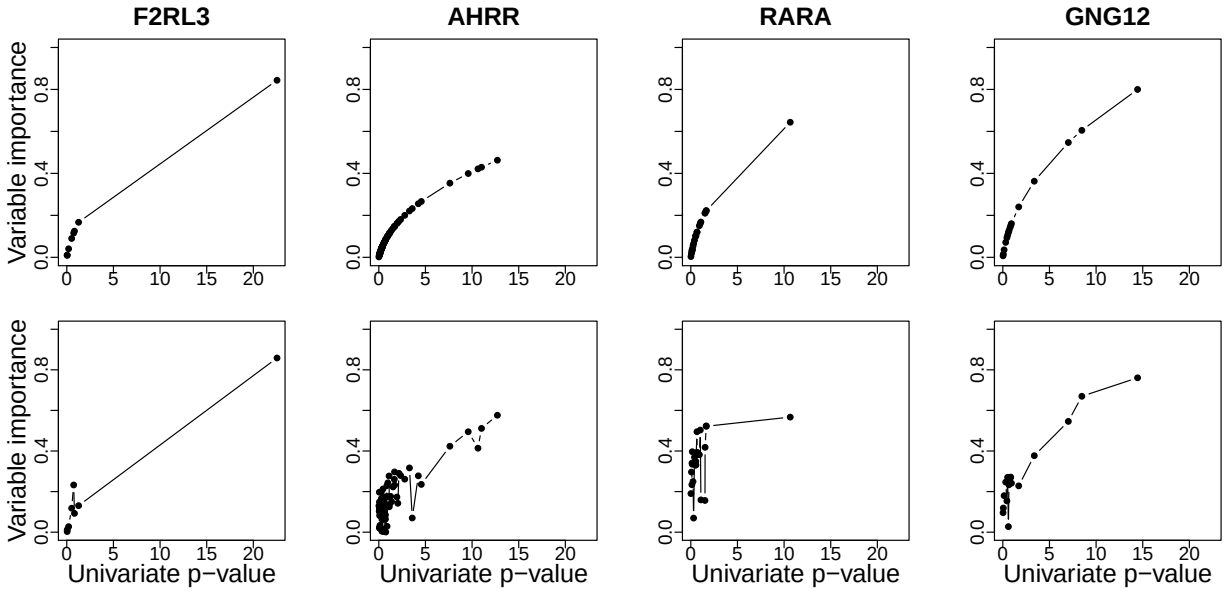
This dataset illustrates an example where $p \ll n$; therefore, no block strategy was necessary to perform a PCEV analysis. However, we performed both analyses (with and without the blocks) for the sake of comparison. In this data, there is a very high level of correlation between all brain regions, as shown in Supplementary Figure A.5. Such high and extensive correlation makes clustering regions into blocks extremely challenging, and we did not succeed in identifying well-separated clusters with any reasonable level of confidence. However, we present the results of the PCEV-block approach using 10 blocks obtained using a hierarchical clustering technique. Hence, we note that this dataset has several interesting features allowing us to i) compare results between PCEV and PCEV-block, and ii) evaluate the performance of the PCEV-block strategy, using the traditional PCEV approach

Figure 3.8: Comparison of p-values obtained using the gene-based PCEV approach (without block) and a univariate analysis on the ARCTIC data. Red lines correspond to significance threshold.



as a gold standard, in a situation where clusters are very poorly chosen and highly correlated with each other. Analysis of the Aβ accumulation against disease status (Alzheimer disease versus others) revealed a significant association regardless of the testing procedure chosen (see Table 3.3). Furthermore, Figure 3.10 illustrates the good agreement between

Figure 3.9: Comparison of the variable importance (VIP) measures for four genes, F2RL3, AHRR, RARA, and GNG12, known to be associated with cigarette smoking in the analysis of the ARCTIC data. For each gene, the top panel shows the VIP obtained from the PCEV without block, while the bottom panel shows the VIP obtained from the PCEV with block.



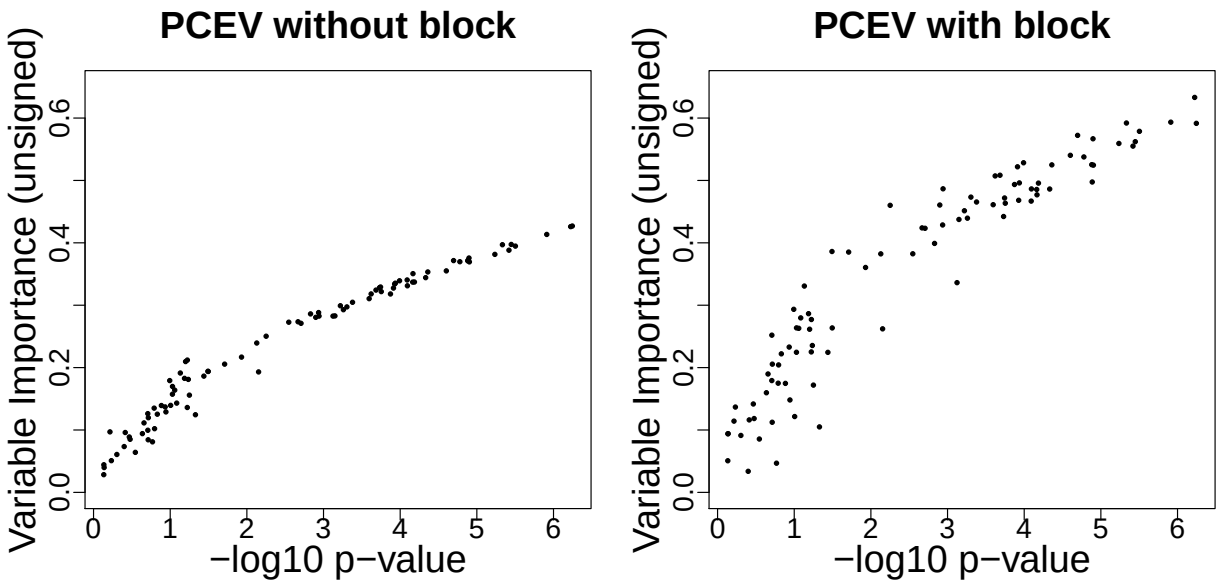
variable importance measures computed using the classical PCEV approach and the block approach. We also note the monotonic relationship between univariate p-values (obtained from a simple regression analysis between Aβ levels in each brain region and disease status) and VIPs.

Table 3.3: P-values for the joint association between amyloid-β accumulation and disease status. Permutation tests were performed using 100,000 permutations.

	PCEV	PCEV with blocks
Exact test	8.13×10^{-5}	—
Permutation test	2×10^{-5}	5×10^{-5}

To further assess robustness to the choice of blocks, we defined 500 random partitions of the 96 brain regions into ten blocks of similar size. PCEV-block was then performed using each of these random partitions. In all cases, the p-values obtained (using 100,000 permutations) were less than 10^{-5}. Moreover, the VIP measures were highly consistent. For each of the random partitions, we looked at the brain regions receiving 1 of the 10 largest VIP values.

Figure 3.10: PCEV variable importance measures versus univariate p-values (negative log scale) for the association between amyloid-β accumulation and Alzheimer's disease status.



We then compared this list with that coming from the original analysis (i.e. the original PCEV approach, without any block). All such lists coming from the random partitions shared at least 7 brain regions with the original analysis, and 95% shared at least 8 brain regions. Moreover, the same brain region (i.e. brain stem) consistently received the highest VIP value.

Finally, we also tested the global association between the Aβ levels and 20 SNPs in a region around the APOE gene. The p-value obtained was 0.0117; since only one test was performed, this is significant at level $\alpha = 0.05$. To measure the contribution of each SNP to the overall association, we looked at two different measures. First, we computed the Pearson correlation between each SNP and the estimated PCEV. Second, we computed a p-value using a LRT. We note that these two methods provide distinct and complementary information: the correlation is a marginal measure of association, whereas the LRT p-value measures the incremental value of a SNP to the model containing all SNPs but the current one. These values, along with the name of the gene containing each SNP, appear in Table 3.4.

Table 3.4: P-values and correlations for the joint association between amyloid-β accumulation and 20 SNPs located near the APOE gene. Correlations are computed with respect to a SNP and the PCEV component. Likelihood Ratio Test (LRT) p-values were computed using the multivariate linear model. The SNPs are ordered from highest correlation value to lowest. The position corresponds to the location of the SNP on chromosome 19.

SNP id	Position	Gene	Correlation	LRT p-value
rs769449	45410002	APOE	0.536	0.023
rs157582	45396219	TOMM40	0.496	0.554
rs2075650	45395619	TOMM40	0.445	0.413
kgp12169129	—	—	0.377	0.709
kgp7807118	—	—	0.371	0.168
kgp658335	—	—	0.343	0.966
rs157580	45395266	TOMM40	0.278	0.657
rs1160985	45403412	TOMM40	0.275	0.689
kgp2187574	—	—	0.262	0.184
kgp9081044	—	—	0.241	0.803
rs3729640	45381917	PVRL2	0.136	0.087
rs445925	45415640	APOC1	0.096	0.050
rs439401	45414451	APOE	0.090	0.664
rs8104483	45372354	PVRL2	0.090	0.104
rs387976	45379060	PVRL2	0.088	0.580
kgp5805962	—	—	0.084	0.605
kgp7618482	—	—	0.065	0.769
rs11879589	45373276	PVRL2	0.064	0.454
kgp3326341	—	—	0.058	0.985
kgp347852	—	—	0.025	0.356

3.4 Discussion

In this article, we have revisited a dimension-reduction approach, PCEV, which has unfortunately received little attention in statistical genetics or biostatistics in general. We showed how PCEV is well-suited for multivariate association studies. This is due to its optimality with respect to capturing the association between a multidimensional phenotype and a set of covariates.

We also presented a hypothesis-test framework which relies on an asymptotic result but no resampling. In particular, we presented two analytic tests that can be used with the traditional single-block PCEV: Wilks' and Roy's tests. The former should only be used when

there is a single covariate and when $n > p + 1$, while the latter can be used in much more general settings (even when $p \gg n$). Note also that Wilks' test is exact, whereas Roy's uses an asymptotic result. Hence, the latter test is based on an approximation to the null distribution when the sample size is large enough. These asymptotic results mean that PCEV can be computationally extremely fast, which makes it feasible for use in large scale pipelines of multivariate analyses. For the block approach to PCEV, the independence of blocks assumption violates the distributional assumption necessary for Johnstone's approximation [Johnstone, 2008] to be valid; for this reason, we have opted for a permutation procedure in order to test for association. Finally, although our discussion has focused on testing the first PCEV component (which is the only component computable when X is a single covariate), this framework can also be applied to several PCEV components independently, when multiple covariates are analysed simultaneously.

Our framework also allows for the direct inclusion of confounding variables. In the presence of confounding, this feature leads to unbiased estimates of the matrix $\mathbf{B}$ of regression coefficients and therefore to correct inference about the association between the outcomes and the covariates. However, we note that covariates (which we denoted X in our model) and confounders (which we denoted Z) play a very different role in the analysis: the covariates should be of scientific interest to the analysis. For example, in our analysis of Aβ accumulation and Alzheimer's disease, we were not interested in the effect of age, gender, and education level on Aβ accumulation in the brain. However, since these three variables are known to be associated with both Aβ levels and with Alzheimer's disease, we included them as confounders in our PCEV analysis.

We also introduced a useful metric, the VIP measures, to help researchers decompose the global signal with respect to each phenotype. This metric measures the *marginal* importance of each variable, in contrast to their *incremental* importance with respect to the other variables. This fact is highlighted in Figures 3.9 and 3.10, where we show that the VIP measures

are a monotone function of the univariate p-values. For this reason, we recommend the use of the VIP values in a more qualitative approach: extreme values, as well as large gaps in the distribution, are more informative than the actual values.

Given that there are likely to be many possible block partitions, understanding the impact of poor block choice is of interest. In general, for a fixed number of blocks of constant size, minimizing the correlation between the blocks will lead to maximal power. On the other hand, when blocks are not chosen in an optimal way, simulation results and data analyses below show that the loss in power is minimal.

We compared the performance of PCEV to PCR, which is a very popular approach for region-based analyses [Lindenmayer et al., 1995, Kherif et al., 2002, Livshits et al., 2002, Arya et al., 2002, Rowe and Hoffmann, 2006, Teipel et al., 2007, Formisano et al., 2008] . As expected, our simulation results show there is no guarantee that the first principal component is at all associated with the covariates. PCR had very low power to detect association, especially when compared to both the classical and block approaches to PCEV. Hence, in general, we discourage the use of PCR for region-based analyses of multidimensional phenotypes.

In a truly high-dimensional simulation scenario ($p \gg n$), we compared PCEV-block to both lasso and sPLS. As expected, the computational time is much faster for PCEV than the other competing methods—in fact, PCEV is 10 times faster than lasso, its fastest competitor. Moreover, we also showed that power is higher for PCEV in almost all scenarios, with only one exception when there is no correlation between any of the variables. In this context, all four methods give similar results. The lower power of lasso and sPLS seen with other correlation structures is likely due to small effect sizes between the covariate and individual response variables. The lower power of lasso could also be a consequence of its poor performance in the presence of correlated features [Efron et al., 2004]. As can be seen in Supplementary Figure A.2, neither lasso nor sPLS were successful in selecting the truly associated variables. On the other hand, PCEV was able to capture the global signal even

when individual associations were small. We note that PCEV has many other advantages over lasso and sPLS: i) it does not require a "reverse" analysis (where the role of $\mathbf{Y}$ and X are reversed); ii) confounders can be directly included in the analytical framework; and iii) it can easily handle more than one covariate of any type (i.e. binary, categorical, or continuous).

For comparison purposes, we decided to use lasso and sPLS instead of similar procedures that incorporate more structured penalties, such as sparse group lasso [Simon et al., 2013] or sparse group PLS [Liquet et al., 2016]. These methods make use of external information to define groups of similar variables that are more likely that have similar regression coefficients. The penalty then uses this information to shrink the coefficients toward a common mean. However, this new information needs to be incorporated via the addition of a second tuning parameter. While these methods would likely perform better than lasso and sPLS in terms of variable selection, it is not clear that this would lead to better power. In any case, the simulation shows that when PCEV ignores the block information and the blocks are selected randomly (PCEV-rand), the performance is similar to when the block information is incorporated in the estimation.

In all our simulation scenarios, we included only one covariate, and we did not include any confounders. However, the effects of these factors on power and Type I error are well documented; this follows from our assumption of a linear model. For example, including more covariates will decrease power of all methods, while the effect of adding confounders to the analysis will depend on the strength of the correlation between the covariates and the confounders. We therefore decided not to present simulation scenario re-addressing these questions [Pearl, 2009].

Our analysis of the bisulfite sequencing data reveals how useful the block approach can be in the presence of high-dimensional data. Indeed, most multivariate methods do not give meaningful results for such a dataset, due to the discrepancy between the sample size and

the number of variables. In contrast, the block PCEV approach was able to recapitulate known results about the potential existence of a Differentially Methylated Region (DMR) around the BLK gene. As shown in Figure 3.7b (right panel), PCEV-block captures the association between 5,986 variables and the cell-type covariate in a single test, despite the fact that there are only 40 subjects. Moreover, the VIP values were able to accurately pinpoint the source of the signal among the 6,000 CpG sites (Figure 3.7a). Supplementary Figure A.4 (left panel) shows that VIP correlates well with nominal p-values from univariate tests, and Supplementary Figure A.4 (right panel) indicates that signed-VIP values correlate with univariate regression coefficients measuring the association between each variable and cell type.

With the ARCTIC dataset, our aim was to confirm the results of the association between methylation and cigarette smoking shown in Breitling et al. [2011]. Two genes showed particularly strong associations in Figure 3.8, F2RL3 (PCEV p-value: $\approx 10^{-26}$), and AHRR (PCEV p-value: $\approx 10^{-25}$), and both are previously reported smoking meQTL loci [Breitling et al., 2011, Lee and Pausova, 2013]. By pooling all CpG sites at or near each gene, our power to detect these associations was greater than what we could have achieved with an adjusted univariate approach. After using the VIPs to decompose the multivariate signals at these two genes (Figure 3.9), it is evident that for F2RL3, the signal is mostly driven by one CpG site, whereas for AHRR, there seem to be five sites contributing to the overall signal. In the latter case, this "pooling of forces" probably leads to the power increase.

In the context of the brain imaging study, we were able to investigate the effect of wide-ranging correlations on the block approach to PCEV. We showed that, even though the independence of block assumption was clearly violated, the p-value obtained using permutations was comparable to that obtained from a classical approach to PCEV (i.e. without blocks) and an exact test. Furthermore, the VIP factors were similar for both approaches.We also looked at the impact of choosing blocks randomly, and we showed that the results were

similar to those obtained from an analysis using the original PCEV framework (i.e. without any block). Therefore, we see that the block approach is quite robust to violations of the assumption contained in Theorem 1.

We also provided an example of using PCEV with multiple covariates. We analysed the joint association between Aβ accumulation and 20 SNPs in the PVRL2-TOMM40-APOE region on chromosome 19, and we obtained a single, significant p-value. In assessing the contribution of each SNP to this global result, we determined that the most important SNP was `rs769449`, which is located in the APOE gene. This SNP is a well-known risk variant for Alzheimer's disease [Cruchaga et al., 2013]. We also uncovered evidence of linkage disequilibrium in this genomic region: a few SNPs located in the TOMM40 gene show evidence of marginal association (as measured using correlations), but none of them are significant when using a Likelihood Ratio test. This is consistent with results in the literature [Yu et al., 2007, Bu, 2009].

In summary, we have shown how PCEV is particularly well-suited for finding multivariate signals with widespread association with a set of covariates. There is a fine balance to strike when performing a multivariate association test with a very high-dimensional phenotype: although we want to include multiple correlated response variables and thus borrow strength across phenotypes, at the same time we would also like to retain interpretability of the multivariate signal. Suppose, to give an extreme example, that in the context of a genome-wide methylation association study, an analyst decided to include all the CpG sites available on a microarray as an enormous set of outcomes, and then to test for global association, at the genome level, with a covariate. Rejecting the null hypothesis in such a scenario would be of no help whatsoever in targeting the source of this signal. This example also illustrates the fact that PCEV is not a variable selection method, and our simulations showed that when the signal is very sparse and the signal-to-noise ratio is low, power is greatly reduced. In contrast, if the analyst decides to severely limit the number of outcomes jointly analyzed,

then targeting the source of any identified associations will be straightforward, but there will be little benefit over univariate testing. In general, choosing an appropriate set of outcomes for joint analysis, i.e. finding the right balance between interpretability and power, is an important consideration that should consider the context and the biology. We feel that region-based analyses—e.g. where regions could be parts of the brain, genes, pathways or some other external data partitioning—is where this method should be considered and can shine.

3.5 Software

An `R` package called `pcev`, implementing both the block and the classical approach to PCEV, is currently available on both CRAN (cran.r-project.org/) and GitHub (github.com/GreenwoodLab/pcev).

Acknowledgements

The authors want to thank the reviewers for their helpful comments and constructive suggestions.

Computations were made on the supercomputer Mammouth-parallèle 2 from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's

Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work was also supported by the Natural Sciences and Engineering Research Council of Canada (AL), the Fonds de recherche Santé (chercheur boursier, AL) and Nature et Technologies (Doctoral award, MT) du Québec. The authors acknowledge the Ludmer Centre for Neuroinformatics and Mental Health for financial support, and they also thank Dr. Tomi Pastinen for providing the bisulfite sequencing data.

Chapter 4

A Tracy-Widom Empirical Estimator For Valid P-values With High-Dimensional Datasets

Preamble to Manuscript 2. While working on the first manuscript, I experimented with several ways of regularising the estimation of the residual variance matrix, i.e. the matrix appearing in the denominator of the Rayleigh quotient corresponding to PCEV. One of these attempts involved replacing the sample covariance estimator of the residuals by the linear shrinkage estimator of Ledoit and Wolf [2004]. To compute p-values, I was using a permutation framework. However, the Ledoit-Wolf estimator was a location-scale transform of the sample covariance matrix, and I eventually recognised a parallel between the Ledoit-Wolf methodology and the transformation of the largest root appearing in Johnstone [2008]. As a consequence, I started to investigate whether a different location-scale transformation of the largest root could be obtained for PCEV combined with the Ledoit-Wolf estimator. Such a discovery would yield an analytical approximation of the null distribution of the test statistic, freeing us from the permutation inferential framework.

However, an analytical description of such a transformation seemed intractable, and I eventually decided to estimate these parameters using a small number of permutations. This number of permutations was significantly less than that that would be required for the full permutation procedure, especially in the context of multiple testing. Simulation studies then confirmed that this empirically-derived transformation yielded valid p-values.

Around the same time, I started a collaboration with Stepan Grinek on using PCEV as a visualisation tool for sequencing data. While reading the pattern recognition literature, Stepan had discovered that a truncated variant of Singular Value Decomposition (SVD) was routinely used to solve generalised eigenvalue problems when the matrices are not invertible. One consequence of this discovery was another way to perform PCEV with high-dimensional data.

However, I soon realised that the result of Johnstone [2008] was not applicable to this high-dimensional setting. Indeed, his result contains an explicit condition on the rank of the matrices in order to guarantee that the joint distribution of the roots of the generalised eigenvalue problem is regular. In the high-dimensional setting, this rank condition no longer holds.

Again, we could have computed valid p-values using a permutation procedure, but I saw an opportunity to test my empirical estimator in this singular setting. There is plenty of evidence in the literature that the Tracy-Widom distribution should still be the "right" limiting distribution. And indeed, simulations showed that the empirical estimator yielded valid p-values in this setting as well.

Finally, through several iterations of the manuscript, I recognised that the methods could be applied much more broadly than just within the PCEV context. Indeed, it could be applied to any *double Wishart problem*, which covers PCEV, but also CCA, Multivariate Analysis of Variance (MANOVA), and many other methods from multivariate analysis. The manuscript presented in this thesis reflects this larger scope.

A Tracy-Widom Empirical Estimator For Valid P-values With High-Dimensional Datasets

Maxime Turgeon[1], Celia MT. Greenwood[1], Aurélie Labbe[2].

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2]*Department of Decision Sciences, HEC Montréal*

Manuscript submitted to the *Journal of the American Statistical Association.*

Abstract

Recent technological advances in many domains including both genomics and brain imaging have led to an abundance of high-dimensional and correlated data being routinely collected. Classical multivariate approaches like Multivariate Analysis of Variance (MANOVA) and Canonical Correlation Analysis (CCA) can be used to study relationships between such multivariate datasets. Yet, special care is required with high-dimensional data, as the test statistics may be ill-defined and classical inference procedures break down.

In this work, we explain how valid p-values can be derived for these multivariate methods even in high dimensional datasets. Our main contribution is an empirical estimator for the largest root distribution of a singular double Wishart problem; this general framework underlies many common multivariate analysis approaches. From a small number of permutations of the data, we estimate the location and scale parameters of a parametric Tracy-Widom family that provides a good approximation of this distribution. Through simulations, we show that this estimated distribution also leads to valid p-values that can be used for high-dimensional inference. We then apply our approach to a pathway-based analysis of the association between DNA methylation and disease type in patients with systemic auto-immune rheumatic diseases.

4.1 Introduction

Consider the following scenario: your research team received a small grant to study the relationship between anatomical brain features and a set of clinical phenotypes. For the past several months, you've been painstakingly enrolling patients in the study and collecting neuroimaging data on them. The subject-matter expert on your team has already grouped the millions of voxels into regions of interest. A given region can now be represented by a dataset Y of dimension $n \times p$, for n patients, and the clinical phenotypes by a dataset X of dimension $n \times q$. Prior to the study, you had identified Canonical Correlation Analysis (CCA) as a potential analytical tool; it would allow the extraction of maximally correlated components from both datasets, and the overall relationship between Y and X could be summarised with a series of canonical correlations. In the classical low-dimensional setting ($p < n, q < n$), you could also test for significance of these canonical correlations using Rao's statistic [Mardia et al., 1979]. However, despite your colleague's excellent work, most brain regions still contain more features or measurements than your overall sample size ($p > n$); in other words, you are no longer in the classical inference setting, but in a high-dimensional setting. Building on your knowledge of linear algebra and matrix decomposition, you know that you can still extract the canonical components and canonical correlations using a truncated Eigenvalue Decomposition (EVD). But one last obstacle remains: the null distribution of the first canonical correlation is unknown in this high-dimensional setting. Although you could rely on a permutation strategy to obtain a p-value, you are also aware that such a procedure is very computationally expensive if you want to reach a high level of precision for the estimated p-value, or if you want to analyze multiple regions in the brain.

In this article, we provide a fast computational approach for estimating the null distribution of the first canonical correlation in such high dimensional setting.

However, the contribution of this paper extends well beyond this CCA example provided

above. More generally, a cursory look at the table of contents of recent volumes in both neuroimaging and genomics journals reveals a strong bias towards multivariate analysis methods such as Principal Component Analysis (PCA), Multivariate Analysis of Variance (MANOVA), CCA, Principal Component of Explained Variance (PCEV), and Linear Discriminant Analysis (LDA); for a small yet broad sample, see Park et al. [2017], Zhao et al. [2017], Hao et al. [2017], Pesonen et al. [2017], Gossmann et al. [2018], Fraiman and Fraiman [2018], Happ and Greven [2018], Yang et al. [2018], Turgeon et al. [2018b]. This is hardly surprising, given the evolving nature of technological capabilities data and the complex underlying biological processes that are now measurable. As described in our scenario above, using truncated matrix decomposition, we can often perform dimension reduction even in high dimensions. But beyond dimension reduction, many classical multivariate approaches also aim at summarizing the relationship between two datasets Y and X; in the list above, CCA, MANOVA and PCEV all have this common goal. Furthermore, this subset of methods also provide a unified way of performing null hypothesis significance testing: they all rely on the largest root of a *double Wishart problem*. Specifically, the strength of the association between Y and X is measured in terms of the magnitude of the largest solution to the following determinantal equation:

$$\det\left(\mathbf{B} - \lambda(\mathbf{A} + \mathbf{B})\right) = 0. \tag{4.1}$$

where A and B are two independent random matrices, both following a Wishart distribution with the same scale matrix. The definition of A and B is formulated under the null hypothesis and is method-specific. Several specific examples will be given later in this paper. From the distribution of this largest root, we can then compute a p-value for the null hypothesis of interest.

More formally, in high-dimensional settings, when the sample size is smaller than the number of measurements, both matrices $\mathbf{A}$ and $\mathbf{B}$ can have singular distributions. This singularity

leads to both computational challenges for estimation and theoretical challenges for inference. On the one hand, common estimators in the non-singular case can be ill-conditioned (or even undefined) for singular problems; on the other hand, classical asymptotic convergence results rely on large sample sizes and therefore may not directly apply to high-dimensional settings.

In this work, we are interested in multivariate analysis involving two datasets, $\mathbf{Y}$ and $\mathbf{X}$, such that the dimension of one or both matrices may be much larger than the sample size n. We posit that proper high-dimensional inference in several multivariate statistical methods such as CCA, MANOVA and PCEV, can be attained by studying the *singular* double Wishart problem described above. Our main contribution is an empirical estimator of the distribution of the largest root that is applicable to the analysis of high-dimensional data. This estimate provides valid p-values by fitting a location-scale family of distributions to a small number of permutations of the original data. By a theorem of Johnstone [2008], this family of distributions is known to provide an excellent approximation in the non-singular case, and we provide empirical evidence that this good performance extends to the singular case.

The rest of the article is structured as follows: in Section 4.2, we provide some detailed examples of double Wishart problems; these examples are later used to illustrate our approach. In Section 4.3, we describe our approach to approximating the distribution function of its largest root using an empirical estimator. Then, in Section 4.4, we investigate the numerical accuracy of the approximation and show how it leads to valid p-values. We further illustrate our approach through an analysis of the association between DNA methylation and disease type in patients with systemic auto-immune rheumatic diseases. Finally, in Section 4.6, we explain how our empirical estimator can be extended to accommodate linear shrinkage covariance estimators within the double Wishart setting. Our approach has already been implemented in two `R` packages: `pcev` and `covequal`; both packages are available on `CRAN`.

Notation

In what follows, X and Y will denote $n \times q$ and $n \times p$ matrices, respectively. We also write $\mathbf{A} \sim W_p(\Sigma, n)$ when $\mathbf{A}$ is Wishart-distributed with parameters p, n and scale matrix Σ. Recall that this is equivalent to having an $n \times p$ matrix X where each row is independently drawn from a multivariate normal $N_p(0, \Sigma)$ and such that $\mathbf{A} = X^T X$.

4.2 Examples of double Wishart problems

As stated above, the developments in our paper build on theory associated with double Wishart problems. Therefore, we start here by giving four examples of well known multivariate tests that are each double Wishart problems in order to emphasize the number of different applications of the theoretical results to follow. In each case, if the two matrices $\mathbf{A}$ and $\mathbf{B}$ are singular or ill-conditioned, then the theoretical results developed for double Wishart problems no longer apply. The four examples are one-way MANOVA, a test of equality of covariance matrices, CCA, and PCEV. The first example is given for its simplicity and historical importance; the other three examples are used later to illustrate our approach.

4.2.1 MANOVA

Suppose that we have a set of n independent observations $\{Y_{ik}\}$, where $Y_{ik} \sim N_p(\mu_k, \Sigma)$ denotes the i-th observation of the k-th group, with $i = 1, \ldots, n_k$, $k = 1, \ldots, K$. In MANOVA, we are interested in the null hypothesis of equality of means $H_0 : \mu_1 = \cdots = \mu_k$. First, for each group, we can form the sample mean $\bar{Y}_k$ and covariance matrix S_k. We then compute two basic quantities: 1) the within-group sum of squares $W = \sum_k n_k S_k$; and 2) the between-group sum of squares $B = \sum_k n_k (\bar{Y}_k - \bar{Y})(\bar{Y}_k - \bar{Y})^T$, where $\bar{Y} = n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ik}$ is the overall mean. Under H_0, the matrices W and B are independent and Wishart-distributed, with $W \sim W_p(\Sigma, n - K)$ and $B \sim W_p(\Sigma, k - 1)$. The union-intersection test of H_0 uses

the largest root of $W^{-1}B$ as a test statistic or, equivalently, that of $(W+B)^{-1}B$ [Mardia et al., 1979, Chapter 12]. In the notation of Equation 4.1, we therefore can express the union-intersection test as a double Wishart problem with $\mathbf{A} = W$ and $\mathbf{B} = B$. This test statistic is also known as Roy's largest root, and it is one of the standard tests in classical MANOVA.

4.2.2 Test of covariance equality

Suppose that independent samples X, Y from two multivariate normal distributions $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$ lead to covariance estimates S_1, S_2 which are independent and Wishart distributed on n_1, n_2 degrees of freedom, respectively. For example, we could take S_i to be the maximum likelihood estimate of the covariance matrix based on $n_i + 1$ observations. We are interested in testing the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$; this can done using the largest root of the determinantal equation 4.1 as a test statistic for which we set $\mathbf{A} = n_1 S_1$ and $\mathbf{B} = n_2 S_2$ [Anderson, 2003, Chapter 10].

4.2.3 CCA

Suppose we have two datasets X, Y of dimension $n \times q$ and $n \times p$, respectively. Recall that CCA seeks to find linear combinations of X and Y that are maximally correlated with each other. Assume each row (X_i, Y_i) of (X, Y) has joint distribution $N_{q+p}(0, \Sigma)$, where Σ has the form

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_Y \end{pmatrix}.$$

For simplicity, we first assume $\Sigma_X = \Sigma_Y$ (which implies that $q = p$). Under our normality assumption, a hypothesis test for independence between X and Y is equivalent to a hypothesis test for $\Sigma_{XY} = 0$.

Write $P = Y(Y^TY)^{-1}Y^T$ and let $P^\perp = I - P$. Using these quantities, we can define

$$\mathbf{A} = X^T P^\perp X,$$
$$\mathbf{B} = X^T P X.$$

Under our assumptions, we have $\mathbf{A} \sim W_q(\Sigma_X, n-p)$ and $\mathbf{B} \sim W_q(\Sigma_X, p)$. We can test the null hypothesis $H_0 : \Sigma_{XY} = 0$ using as a test statistic the largest root of the double Wishart problem corresponding to the pair of matrices $\mathbf{A}, \mathbf{B}$ above [Mardia et al., 1979, Chapter 10]. We note that this largest root corresponds to the square of the first canonical correlation between X and Y.

The computation above can be generalized to the case when $\Sigma_X \neq \Sigma_Y$ (and $q \neq p$) and with nonzero mean; for details, see Johnstone [2009].

4.2.4 PCEV

Similar to CCA, PCEV can also be used to simultaneously perform dimension reduction and test for association between two multivariate samples X, Y of dimension $n \times q$ and $n \times p$, respectively. However, whereas CCA seeks to maximise the correlation between linear combinations of Y and X, PCEV seeks the linear combination of Y whose proportion of variance explained by X is maximised. Specifically, we assume that the relationship between X and Y can be represented via a linear model:

$$Y = \Gamma X + E,$$

where Γ is a $p \times q$ matrix of regression coefficients for the covariates of interest, and $E \sim N_p(0, \Sigma_R)$ is a vector of residual errors. This model assumption allows us to decompose the

total variance of Y as the sum of variance explained by the model and residual variance:

$$\mathrm{Var}(Y) = \Sigma_M + \Sigma_R,$$

where $\Sigma_M = \Gamma \,\mathrm{Var}(X)\Gamma^T$. PCEV seeks a linear combination of outcomes, $\mathbf{w}^T Y$, that maximises the proportion $R^2(\mathbf{w})$ of variance being explained by the covariates X:

$$\mathbf{w}_{\mathrm{PCEV}} = \underset{\mathbf{w}\in\mathbf{R}^p:\|\mathbf{w}\|=1}{\mathrm{argmax}} \; R^2(\mathbf{w}),$$

where

$$R^2(\mathbf{w}) = \frac{\mathrm{Var}(\mathbf{w}^T\Gamma X)}{\mathrm{Var}(\mathbf{w}^T Y)} = \frac{\mathbf{w}^T\Sigma_M\mathbf{w}}{\mathbf{w}^T(\Sigma_M+\Sigma_R)\mathbf{w}}. \tag{4.2}$$

Then testing the null hypothesis $H_0 : \Gamma = 0$ is performed using as a test statistic the largest root of the determinantal equation 4.1 for which we set $\mathbf{A} = \Sigma_R$ and $\mathbf{B} = \Sigma_M$. Turgeon et al. [2018b] have further details on this dimension-reduction technique.

For further examples of double Wishart problems, we refer the reader to Johnstone [2008, 2009].

4.3 Empirical Estimator of the Distribution of the Largest Root

As we can see from the examples above, many null hypothesis significance tests in multivariate analysis follow the same two steps 1) computing the largest root of equation 4.1; and 2) from its distribution under the null hypothesis, compute a p-value. For the first point, we may have to use a truncated version of Singular Value Decomposition (SVD) (or EVD) when both $\mathbf{A}$ and $\mathbf{B}$ are singular; this singularity is common in high-dimensional datasets. In essence, these matrix decompositions are restricted to the subspace spanned by the eigenvectors corresponding to the non-zero eigenvalues; the mathematical details are reviewed in

Section B.1 of the Supplementary Materials. In this section, we focus on the second point. We show how the distribution of the largest root can be accurately approximated in the singular setting.

4.3.1 Known results in the non-singular setting

In the non-singular setting, the distribution of the largest root to the determinantal equation 4.1 is well studied [Mardia et al., 1979, Muirhead, 2009a]. To provide a theoretical underpinning to the remainder of the section, we highlight two approaches to computing this distribution: an exact approach, and an asymptotic result.

First, Chiani [2016] described an explicit algorithm for computing the Cumulative Distribution Function (CDF) of the largest root λ. Building on his earlier work [Chiani, 2014], he proposed a new expression that relates the CDF to the Pfaffian of a skew-symmetric matrix. He also provided a set of recursive equations that provide a fast and efficient way to compute this matrix. However, this matrix quickly becomes very large when the parameters of the Wishart distribution increase, leading to both computational instability and numerical overflow problems. And yet, this matrix can only be computed in the non-singular setting.

The second approach we wish to highlight uses results from random matrix theory to derive an approximation to the marginal distribution. Specifically, Johnstone [2008] showed that after a suitable transformation of the largest root, its distribution can be approximated by the Tracy-Widom distribution of order 1. His result, using the notation of this manuscript, is given below:

THEOREM 2. [Johnstone [2008]] Assume $\mathbf{A} \sim W_p(\Sigma, m)$ and $\mathbf{B} \sim W_p(\Sigma, n)$ are independent, with Σ positive-definite. Let λ be the largest root of Equation 4.1. As $p, m, n \to \infty$, we have

$$\frac{\text{logit}\,\lambda - \mu}{\sigma} \xrightarrow{\mathcal{D}} TW(1),$$

where $TW(1)$ is the Tracy-Widom distribution of order 1 (cf. Tracy and Widom [1996]), and μ, σ are defined as follows:

$$\mu = 2\log\left(\tan\left(\frac{\varphi+\gamma}{2}\right)\right)$$
$$\sigma^3 = \frac{16}{(m+n+1)^2}\left(\sin^2(\varphi+\gamma)\sin\varphi\sin\gamma\right)^{-1},$$

where

$$\sin^2(\gamma/2) = \frac{\min(p,n)-1/2}{m+n+1}$$
$$\sin^2(\varphi/2) = \frac{\max(p,n)-1/2}{m+n+1}.$$

4.3.2 Tracy-Widom Empirical Estimator

Unfortunately, in the singular setting, the marginal distribution of the largest root is not as well-understood. Crucially, the results from both Johnstone [2008] and Chiani [2016] depend on the non-singularity of the matrix $\mathbf{A}$, and therefore they cannot readily be applied to the singular setting.

As highlighted in the introduction, a common approach to computing p-values when the null distribution is unknown is to use a permutation procedure. But high precision requires a large number of permutations, and therefore is computationally burdensome. In this section, we show that we can drastically reduce the number of permutations by using Johnstone's theorem above as inspiration.

We suggest an empirical estimator for approximating the CDF of the largest root. Indeed, we propose to estimate the distribution using a location-scale family of the Tracy-Widom distribution of order 1 indexed by two parameters (μ, σ). Algorithm 3 below describes our

approach.

Algorithm 3 Tracy-Widom empirical estimator of the largest root distribution.

1: For a double Wishart problem associated with matrices $\mathbf{Y}$ and $\mathbf{X}$:
2: **for** $k = 1$ to K **do**
3: Perform a permutation procedure on $\mathbf{Y}$ and $\mathbf{X}$.
4: Compute the largest root $\lambda^{(k)}$ of the corresponding double Wishart problem.
5: **end for**
6: Transform the roots $\lambda^{(1)}, \ldots, \lambda^{(K)}$ to the logit scale.
7: Using a fitting procedure, find the estimates $\hat{\mu}$ and $\hat{\sigma}$ of the location and scale parameters, respectively.
8: Approximate the CDF of the largest root using the CDF of $\hat{\sigma}TW(1) + \hat{\mu}$.

A few comments are required:

- As we will show in Section 4.4, the number of permutations K can be chosen to be relatively small while retaining good performance.
- The appropriate permutation procedure on Line 3 depends on the particular double Wishart problem being studied. For a test of association between $\mathbf{Y}$ and $\mathbf{X}$, we typically permute the rows of $\mathbf{Y}$ and keep those of $\mathbf{X}$ fixed. The test of equality of covariance matrices requires a different strategy, and we give all the relevant details in Section 4.4.2 below.
- Line 7 of Algorithm 3 refers to a fitting procedure. In Section 4.4.1, we investigate four different approaches: Method of moments; Maximum Likelihood Estimation; and Maximum Goodness-of-Fit estimation [Luceño, 2006] using the Anderson-Darling statistic and a modified version that gives more weight to the right tail. Details about this latter approach, including computational formulas of these statistics, are given in Section B.2 of the Supplementary Materials.

Therefore, we propose this location-scale transformation of the Tracy-Widom distribution as a suitable parametric family for estimating the distribution of the largest root in singular double Wishart problems.

4.4 Simulation results

To investigate the performance of our empirical estimator, we performed two simulation studies: 1) we compared the distribution obtained from our empirical estimator to an empirically generated cumulative distribution function (CDF), as described below, for different number of permutations K and using four different fitting strategies; 2) we compared p-values derived from the estimated distribution to those obtained through a permutation procedure. The first simulation study is aimed at assessing the performance of our empirical estimator over the whole range of the distribution; the second simulation study specifically looks at the upper tail of the distribution and therefore at the validity of the resulting p-values.

4.4.1 Comparison to the true distribution

Since the distribution function of the largest root distribution is not available analytically, we cannot use an analytical function as the benchmark for the "true" CDF. Therefore, an estimate of the true distribution was derived computationally. Specifically, we started by generating 1000 pairs of singular Wishart variates $\mathbf{A} \sim W_p(\Sigma, m), \mathbf{B} \sim W_p(\Sigma, n)$ as follows: each singular Wishart variate was generated by first generating a sample of multivariate normal variates $Z_1, \ldots, Z_m \sim N_p(0, \Sigma)$, and then defining $\mathbf{A} = \sum Z_i^T Z_i$; the Wishart variate $\mathbf{B}$ was generated similarly. For each pair of Wishart variates, we then computed its corresponding largest root using truncated SVD (cf. Section B.1 of the Supplementary Materials). From these 1000 largest roots, we calculated the empirical CDF for the marginal distribution: this estimate was considered our benchmark for assessing the performance of our approach.

For our simulations, we fixed the degrees of freedom $m = 96$ and $n = 4$. In the context of a one-way MANOVA, this would correspond to four distinct populations and a total sample size of 100; in the context of CCA, this would correspond to one set of variates of dimension 4 (the dimension of the other is p), and a total sample size of 99. We also fixed the scale

matrix $\Sigma = I_p$, but we varied the parameter $p = 500, 2000$.

To compute the Tracy-Widom empirical estimator, we sampled K of these largest roots (with $K = 25$ or 100) and fitted the location-scale family as described in Algorithm 3. Finally, we looked at four different fitting strategies: the method of moments (MM); Maximum Likelihood Estimation (MLE); and Maximum Goodness-of-Fit estimation [Luceño, 2006] using the Anderson-Darling statistic (AD) and a modified version that gives more weight to the right tail (ADR).

The simulation results appear in Figure 4.1. As we can see, with a larger number of samples K, all four methods estimate the distribution of the largest root reasonably well; on the other hand, for a smaller value of K, the method of moments clearly outperforms the other fitting strategies. For this reason, unless otherwise stated, we use the method of moments in the remainder of this article.

4.4.2 Comparison of p-values

Next, we used our empirical estimator to compute p-values for three double Wishart problems: a test of equality of covariances, CCA, and PCEV. In all settings, we performed 100 simulations, and we fixed the sample size at $n = 100$. We also used $K = 100$ permutations to fit the Tracy-Widom empirical estimator using the method of moments. Finally, we compared our approach to a permutation procedure with 500 permutations. As a reference, we summarise the parameters for all simulation scenarios appear in Table 4.1.

Method	Dimension q of X	Dimension p of Y	Association structure	Association parameter
Equality of covariance	$200, 300, 400, 500$	$200, 300, 400, 500$	$\Sigma_X = I_q$ Autoregressive Σ_Y	$\rho = 0, 0.2, 0.5$
CCA	$200, 300, 400, 500$	Fixed at 20	$(\Sigma_{XY})_{ii} = \rho$ for $i = 1, 2$, $(\Sigma_{XY})_{ij} = 0$ otherwise	$\rho = 0, 0.2, 0.5$
PCEV	Fixed at 1	$200, 300, 400, 500$	Linear model	$R^2 = 0\%, 1\%, 5\%$

Table 4.1: Values of the parameters for all simulation scenarios

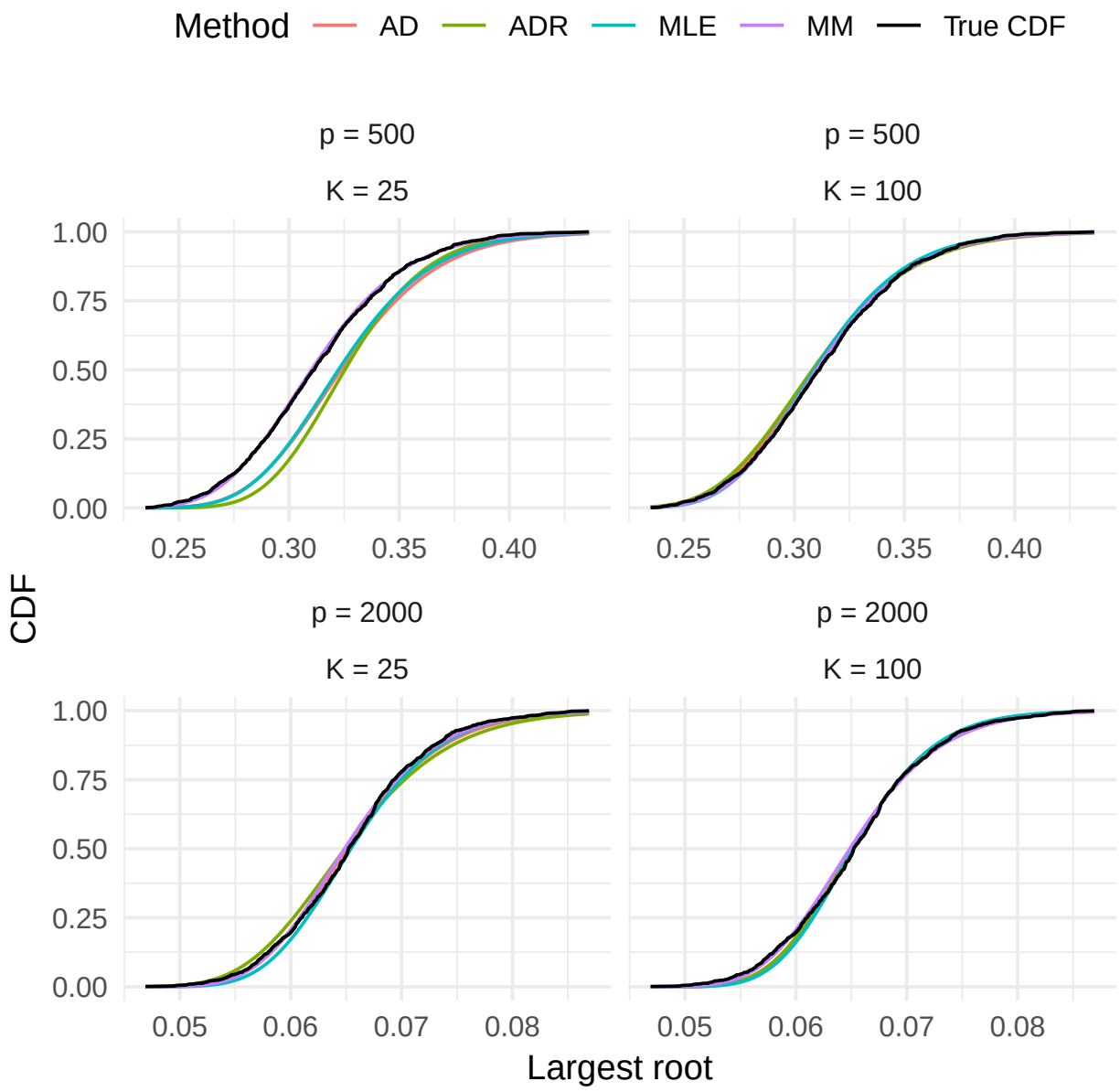


Figure 4.1: **Approximation to the CDF:** Tracy-Widom empirical estimator, using four different fitting strategies, compared to the "true CDF" (derived through computational means). AD: Anderson-Darling; ADR: right-tail-weighted Anderson-Darling; MLE: Maximum Likelihood Estimation; MM: Method of Moments.

Test of equality of covariances

For the test of equality of covariances, we simulated two datasets $X \sim N_p(0, I)$ and $Y \sim N_p(0, \Sigma)$, both with $n = 100$ observations. We selected an autocorrelation structure for the covariance matrix Σ, with $\text{Cov}(Y_i, Y_j) = \rho^{|i-j|}$. We varied two parameters: 1) the dimension $p = 200, 300, 400, 500$ of both X and Y; and 2) the correlation parameter $\rho = 0, 0.2, 0.5$. Note that $\rho = 0$ corresponds to the null hypothesis of the same covariance, and $\rho = 0.2, 0.5$ correspond to alternative hypotheses.

The permutation procedure for testing the equality of covariance matrices started by centring both X and Y. Then, the observations were permuted *between* X and Y: that is, a valid permutation would sample 50 observations from both X and Y to create a permuted X, and the remaining 50 observations from X and Y were used to create a permuted Y. The results are summarised using QQ-plots (see Figure 4.2). The computations were performed using the `R` package `covequal`.

The simulation results appear in Figure 4.2. As we can for these QQ-plots, there is excellent agreement between the p-values obtained from a permutation procedure and those obtained from the Tracy-Widom empirical estimator.

CCA

For CCA, we again simulated two datasets $X \sim N_q(0, I)$ and $Y \sim N_p(0, I)$, both with $n = 100$ observations, and with fixed $q = 20$. We selected an exchangeable structure with parameter ρ for the covariance matrix $\text{Cov}(X, Y)$. We again varied two parameters: 1) the dimension $p = 200, 300, 400, 500$ of X; and 2) the correlation parameter $\rho = 0, 0.2, 0.5$. As above, the value $\rho = 0$ corresponds to the null hypothesis of no association between X and Y, and the values $\rho = 0.2, 0.5$ correspond to alternative hypotheses. Finally, the permutation procedure consisted in permuting the rows of X and keeping those of Y fixed.

The simulation results appear in Figure 4.3. As above, there is excellent agreement between

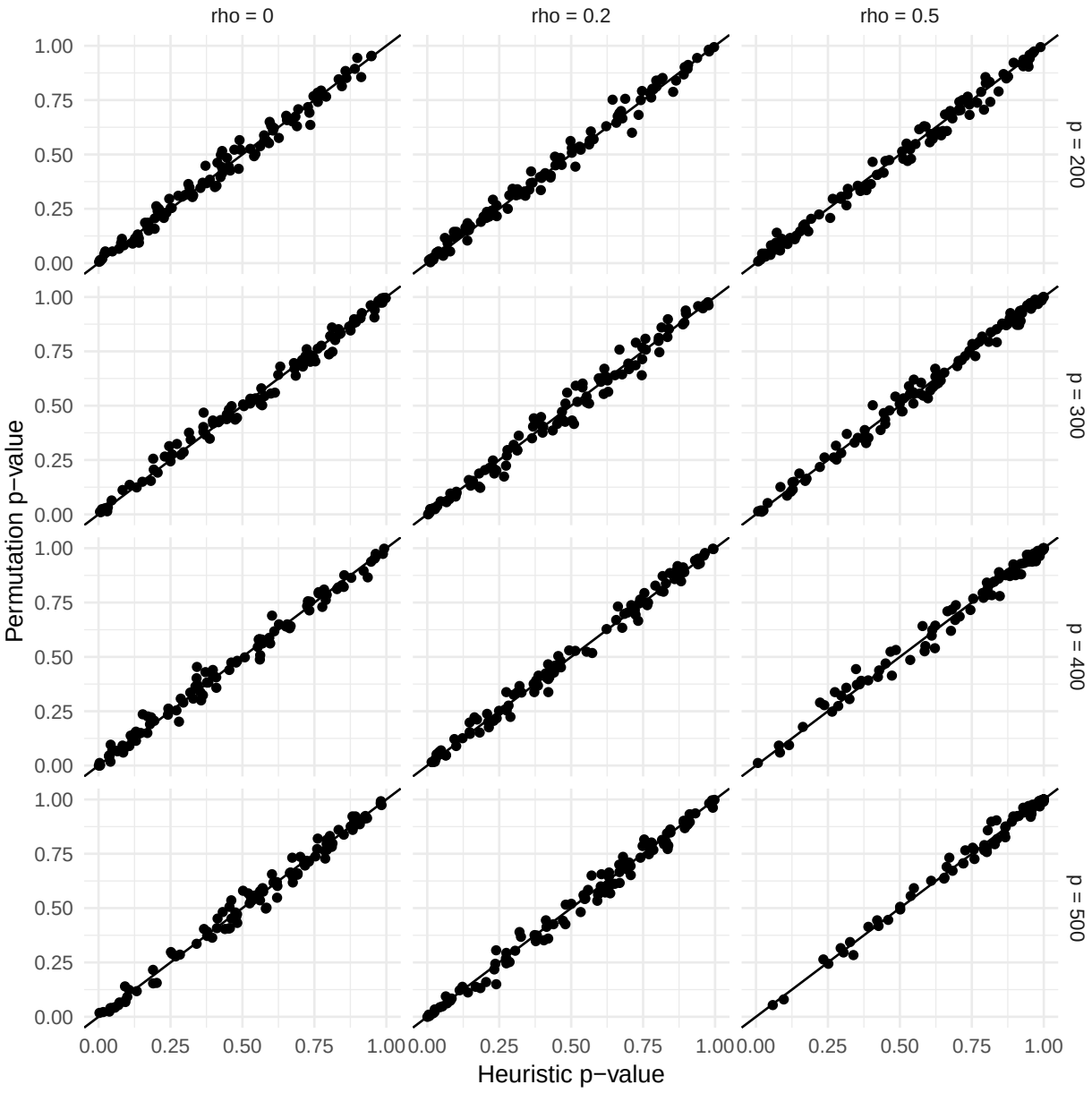


Figure 4.2: **Equality of covariance:** QQ-plots comparing the p-values obtained from a permutation procedure to those obtained from the Tracy-Widom empirical estimator.

the p-values obtained from a permutation procedure and those obtained from the Tracy-Widom empirical estimator.

PCEV

For PCEV, we looked at the following high-dimensional simulation scenario: we fixed the number of observations $n = 100$ and a balanced binary covariate X. We then varied the number of response variables $p = 200, 300, 400, 500$, and fixed the covariance structure of the error term $\Sigma_R = I_p$. We induced an association between X and the first 50 response variables in Y. This association was controlled by the parameter $R^2 = 0\%, 1\%, 5\%$; this parameter is related to the univariate regression coefficient β through the following relationship:

$$\beta^2 = \frac{R^2}{1 - R^2}.$$

As above, we summarised the results using QQ-plots (Figure 4.4). The computations were performed using the `R` package `pcev` [Turgeon et al., 2018b]; the methodology presented here is part of the package starting with version 2.1.

The simulation results appear in Figure 4.4. Once again, there is excellent agreement between the p-values obtained from a permutation procedure and those obtained from the Tracy-Widom empirical estimator.

As we can see, our Tracy-Widom empirical estimator yields valid p-values in a variety of high-dimensional scenarios that include both null and alternative hypotheses. For further simulation results, see Section B.3 of the Supplementary Materials.

4.5 Data analysis

To showcase our ideas in the context of a real analysis of a high-dimensional dataset, we decided to look at the association between DNA methylation and disease type in patients with

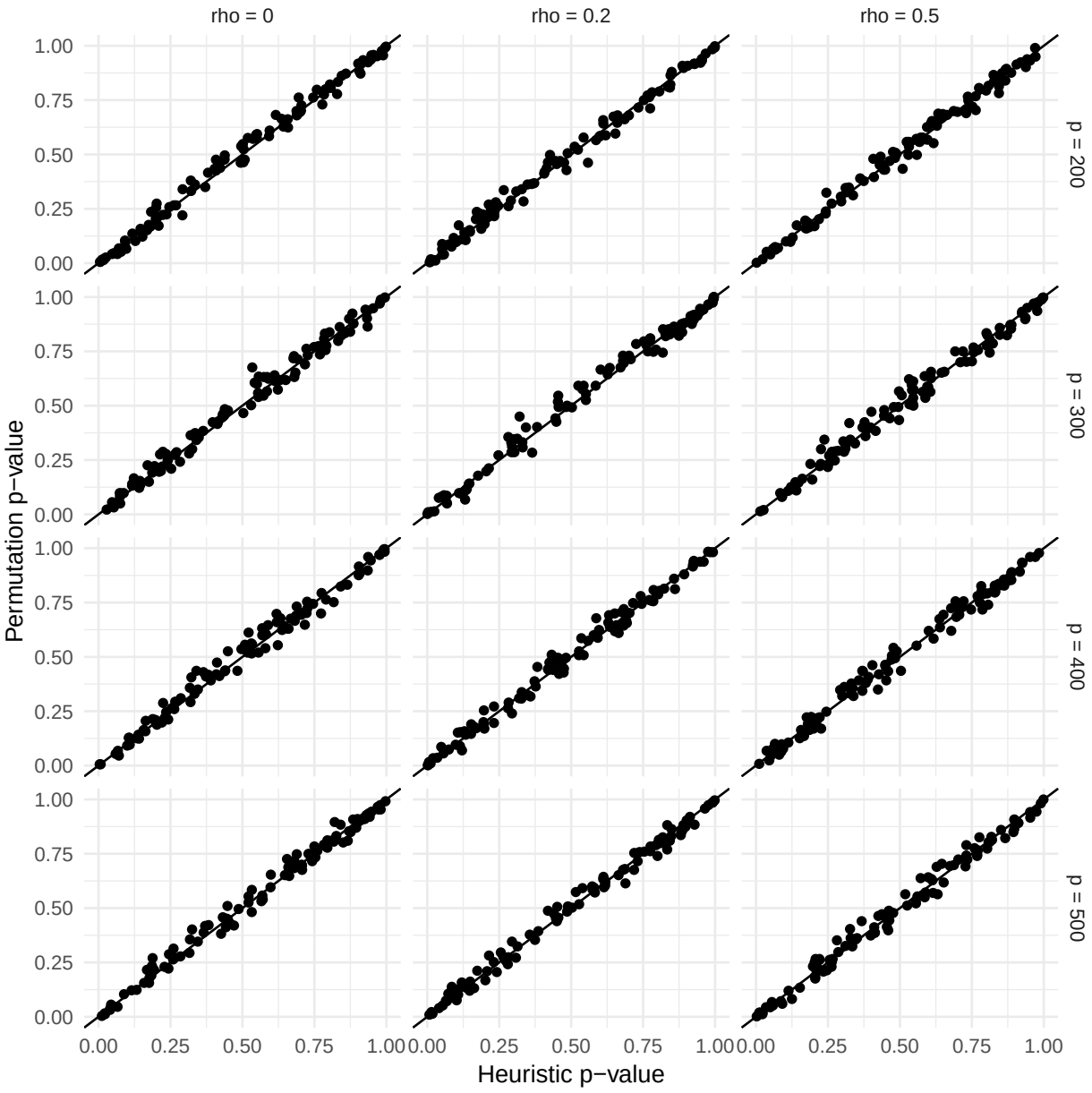


Figure 4.3: **CCA:** QQ-plots comparing the p-values obtained from a permutation procedure to those obtained from the Tracy-Widom empirical estimator.

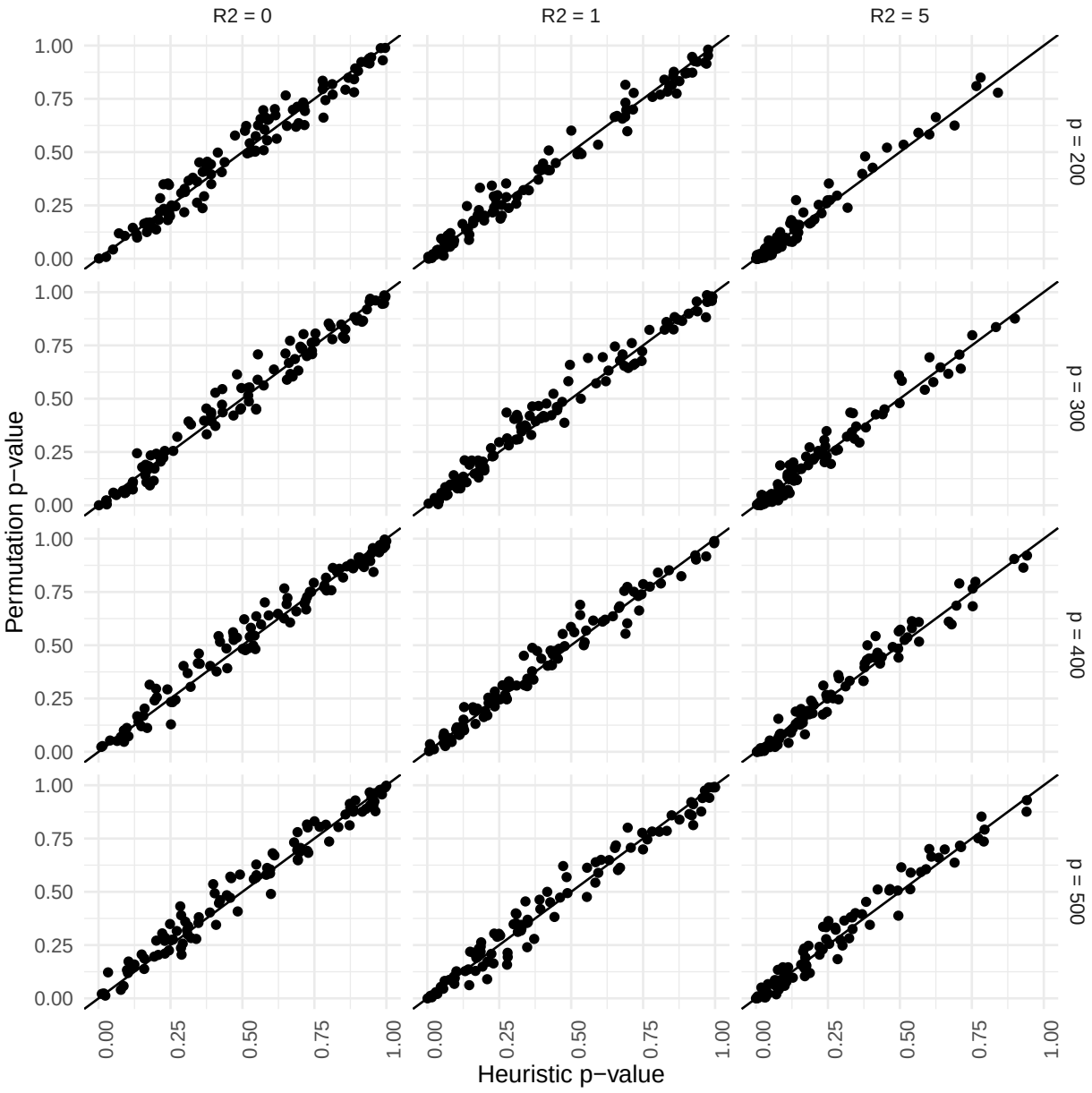


Figure 4.4: **PCEV:** QQ-plots comparing the p-values obtained from a permutation procedure to those obtained from the Tracy-Widom empirical estimator.

four systemic auto-immune rheumatic diseases: Scleroderma, Rheumatoid arthritis, Systemic lupus erythematosus, and Myositis. DNA methylation is an epigenetic mark, meaning that it is a chemical modification of the DNA that does not alter the nucleotide sequence [Baylin, 2005]. It is known to be associated with changes in RNA transcription, and it is therefore correlated with gene expression.

The DNA methylation used for this analysis was measured prior to treatment on T-cell samples from 28 patients using the Illumina 450k platform [Hudson et al., 2017]. Baseline characteristics of the patients appear in Table 4.2.

	Scleroderma (n = 14)	Other diseases (n = 14)
Age Mean (sd)	48 (16)	52 (14)
Female (%)	50%	71%

Table 4.2: Baseline characteristics

We opted to test for differential methylation between scleroderma and the other three diseases *at the pathway level*: from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa et al., 2017], we extracted the list of genes included in their manually curated list of molecular pathways; these pathways correspond to networks of genes interacting and reacting together as part of a given biological process. For each of the 320 pathways, we then identified all CpG dinucleotides contained in at least one gene of this pathway. All CpG dinucleotides mapped to a given pathway were analysed jointly. The extraction of gene lists was performed using the `R` package `KEGGREST` [Tenenbaum, 2017].

For each pathway, we thus have two datasets: an $n \times p$ matrix Y that contains the methylation values at all p CpG dinucleotides (where p ranges from 39 to 21,640 over the 320 pathways) and an $n \times 1$ matrix X indicating whether an individual has scleroderma. Recall that $n = 28$, and therefore all pathways lead to a high-dimensional dataset. The analysis performed had two steps: first, we did a test of equality for the covariance matrices between both disease groups; then we used PCEV to test for differential methylation between these two groups.

The PCEV analysis included both age and sex as possible confounders [El-Maarri et al., 2007, Horvath, 2013]. For both steps, the Tracy-Widom empirical estimator was computed using 50 permutations of the data.

Since we repeated the same analysis independently for all 320 pathways, we need to account for multiple comparison. However, since a given gene may appear in multiple pathways, the 320 hypothesis tests are not independent; therefore, a naive Bonferroni correction would be too conservative. To estimate the effective number of independent tests, we looked at the average number of pathways in which a given CpG dinucleotide appears. Overall, 134,941 CpG dinucleotides were successfully matched to at least one of 320 KEGG pathways. On average, each dinucleotide appeared in 4.5 pathways; this leads to effectively 70 independent hypothesis tests. For a nominal family-wise error rate of $\alpha = 0.05$, an appropriate Bonferroni-corrected significance threshold is therefore given by 7.14×10^4.

We compared the p-value obtained from our procedure above to that obtained from a permutation procedure; the latter is a common approach when the null distribution of a test statistic is unknown but has the disadvantage of being computationally expensive. Given our significance threshold, we determined that we needed to perform at least 10,000 permutations in order to be able to assess significance.

In Table 4.3, we present the five most significant pathways, with the top pathway achieving overall significance for the test of differential methylation. None of the covariance equality tests yielded a significant p-value; to improve clarity, we omitted these results from the table.

For the most significant pathway (i.e. Glutamatergic synapse), we computed the Variable Importance Factor (VIF) and compared them to the p-values obtained from a univariate approach where each CpG dinucleotide is tested individually against the disease outcome [Turgeon et al., 2018b]. The VIF is defined as the correlation between the individual response variables and the principal component maximising the proportion of variance. The compari-

KEGG code	Description	# CpGs	Tracy-Widom P-value	Permutation P-value
path:hsa00120	Glutamatergic synapse	225	1.91×10^{-4}	7.00×10^{-4}
path:hsa03040	Ras signaling pathway	2119	1.33×10^{-3}	1.40×10^{-3}
path:hsa03450	Circadian rhythm	267	1.52×10^{-3}	1.00×10^{-4}
path:hsa00563	Histidine metabolism	394	1.59×10^{-3}	3.00×10^{-4}
path:hsa04380	Pathogenic E. coli infection	2312	1.65×10^{-3}	5.20×10^{-3}

Table 4.3: **Data analysis**: Five most significant pathways based on our empirical estimator. The Tracy-Widom p-values are compared to p-values obtained from a permutation procedure.

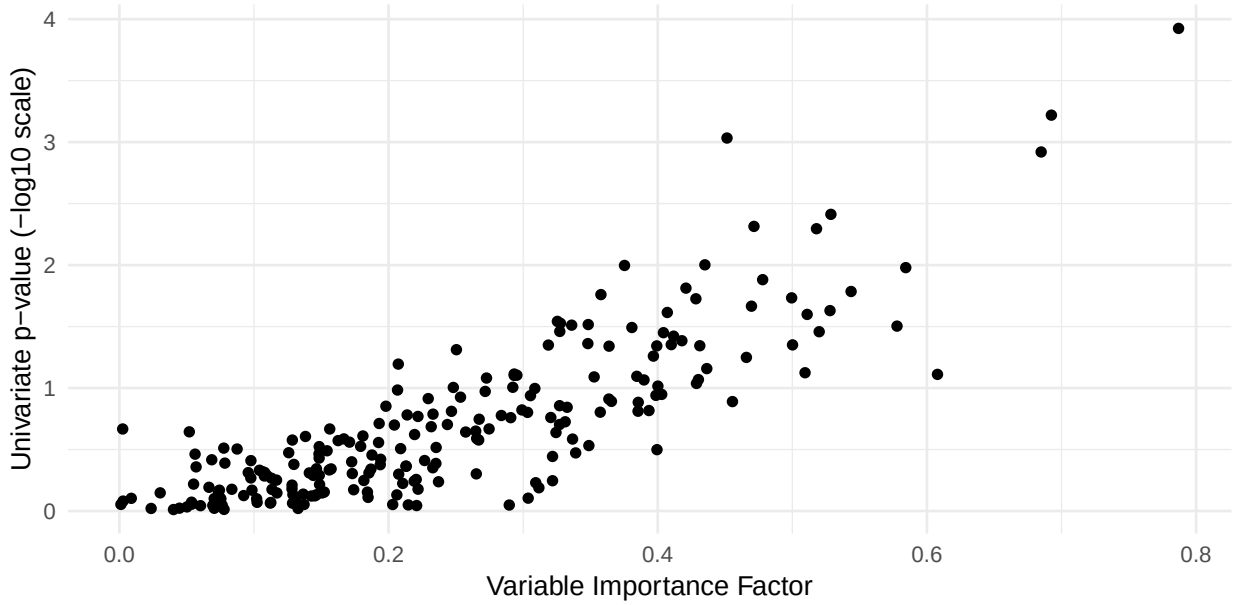


Figure 4.5: **path:hsa00120—Glutamatergic synapse**: Comparison of Variable Importance Factor and 225 univariate p-values for the most significant pathway.

son is presented in Figure 4.5. As previously showed in the literature [Turgeon et al., 2018b], there is some degree of agreement between these two measures of association. Moreover, we can see evidence that the overall association between this pathway and our disease indicator is driven by a few CpG dinucleotides.

4.6 Extension to linear shrinkage covariance estimators

In Section 4.4, we presented graphical evidence that our proposed Tracy-Widom empirical estimator provides a good approximation of the distribution of the largest root of the de-

terminantal equation 4.1. As discussed in Section 4.2, the estimates of the matrices $\mathbf{A}, \mathbf{B}$ appearing in this equation often involve high-dimensional covariance matrices. However, a common problem with such high-dimensional covariance matrices is their instability [Pourahmadi, 2013, Chapter 1]. As a result, the power of statistical tests derived from the double Wishart problem decreases as the dimension of $\mathbf{A}$ and $\mathbf{B}$ increases. However, it would seem that the Tracy-Widom empirical estimator relies on the assumption that $\mathbf{A}$ and $\mathbf{B}$ are Wishart-distributed, and it is not clear *a priori* that this estimator can be applied with other, more efficient estimators of the underlying high-dimensional covariances matrices.

One strategy for improving the stability of a covariance estimator is to use a shrinkage estimator. One such estimator for the covariance matrix was introduced by Ledoit and Wolf [2004]. In this section, we show that by replacing the matrix $\mathbf{A}$ by a linearly shrunk version $\mathbf{A}^*$ in Equation 4.1, our Tracy-Widom empirical estimator still provides a good approximation of the distribution of the largest root.

Let $\mathbf{A} \sim W_p(\Sigma, n)$, $S = \frac{1}{n}\mathbf{A}$ and I be the $p \times p$ identity matrix. Ledoit and Wolf [2004] look for an optimal linear combination $\Sigma^* = \rho_1 I + \rho_2 S$ to estimate the population covariance matrix; the optimality criterion is described in the following result:

THEOREM 3. [Ledoit and Wolf [2004]] Let $\|\cdot\|^2$ be the squared Frobenius norm, and let $\langle \cdot, \cdot \rangle$ be its corresponding inner product. Consider the optimization problem:

$$\min_{\rho_1, \rho_2} \mathrm{E}(\|\Sigma^* - \Sigma\|^2), \quad \text{such that} \quad \Sigma^* = \rho_1 I + \rho_2 S,$$

where the coefficients ρ_1, ρ_2 are nonrandom. Its solution verifies:

$$\Sigma^* = \frac{\beta^2 \mu}{\delta^2} I + \frac{\alpha^2}{\delta^2} S,$$
$$\mathrm{E}(\|\Sigma^* - \Sigma\|^2) = \frac{\alpha^2 \beta^2}{\delta^2},$$

where

$$\mu = \langle \Sigma, I \rangle, \alpha^2 = \|\Sigma - \mu I\|^2, \beta^2 = \mathrm{E}(\|S - \Sigma\|^2), \text{ and } \delta^2 = \mathrm{E}(\|S - \mu I\|^2).$$

Furthermore, Ledoit and Wolf [2004] provide consistent estimators for all quantities $\mu, \alpha^2, \beta^2, \delta^2$ under mild conditions that hold true for normal random variables, and therefore hold true in our setting. The resulting estimator for Σ is denoted S^*, and we thus get a linearly shrunk $\mathbf{A}^* = nS^*$.

To assess the performance of our Tracy-Widom empirical estimator under this extended setting, we repeated the simulations from Section 4.4.1 but by substituting the matrix $\mathbf{A}$ with its linearly shrunk equivalent $\mathbf{A}^*$. The results appear in Figure 4.6, and they are very similar to the earlier results; we get good agreement even with small values of K, and the method of moments provides a better approximation than the other fitting procedures.

4.7 Discussion

In this work, we investigated the singular double Wishart problem, which arises in the multivariate analysis of high-dimensional datasets. We presented an empirical estimator of the distribution of the largest root that is simple, efficient, and valid for high-dimensional data. Through simulation studies, we showed how our approach leads to a good approximation of the true distribution, and we also showed that it leads to valid p-values. Finally, using a pathway-based approach, we analysed the relationship between DNA methylation and disease type in patients with systemic auto-immune rheumatic diseases. Our analysis used the empirical estimator in two settings: a test for the equality of covariance matrices, and with the dimension-reduction method known as PCEV.

The empirical estimator we presented fills a gap in high-dimensional multivariate analysis. Many common methods, such as MANOVA and CCA, fit into the framework of double

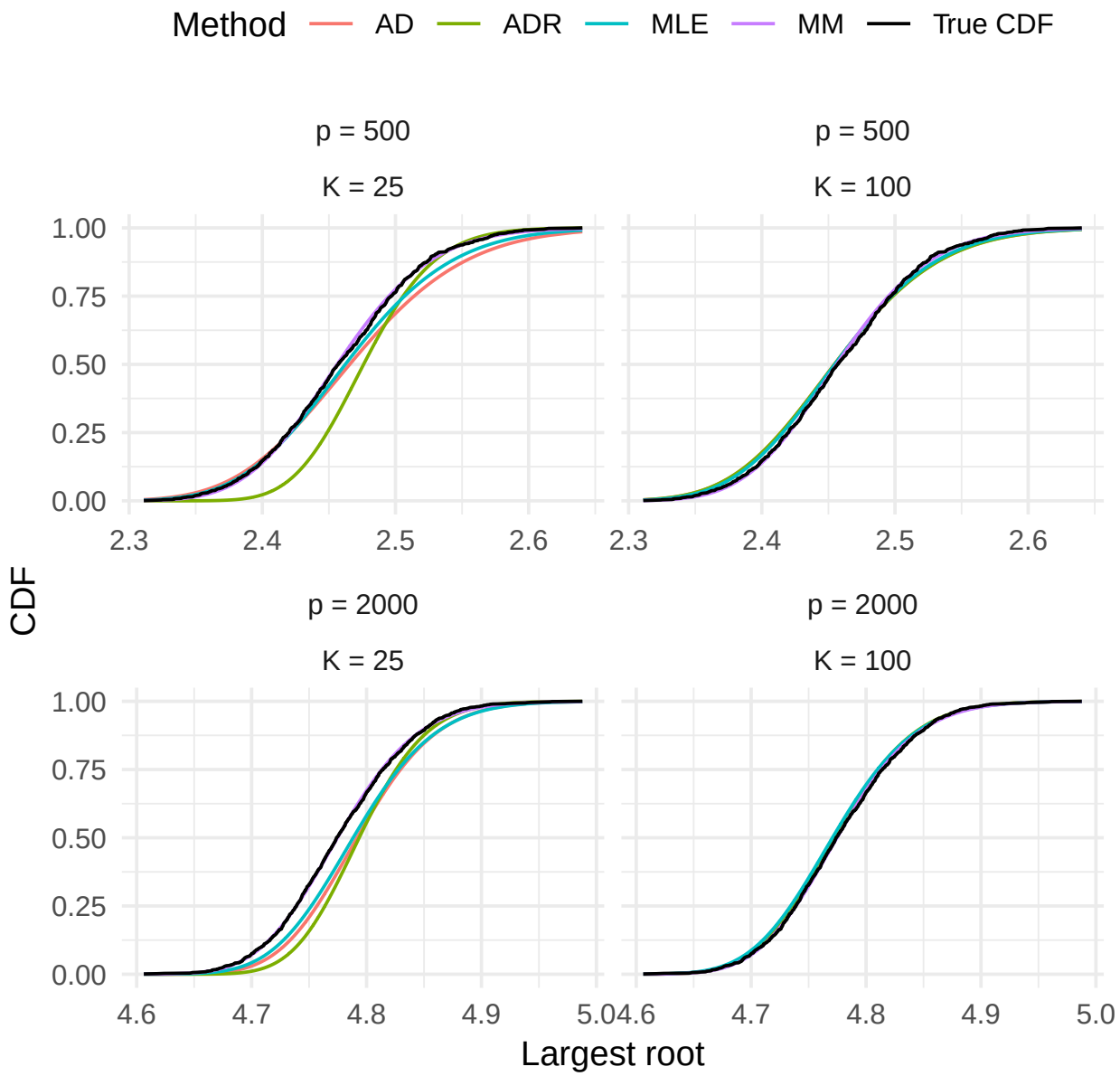


Figure 4.6: **Approximation to the CDF under a linearly shrunk A:** Heuristic, using four different values for the number of permutations, compared to the true CDF

Wishart problems. However, classical hypothesis testing breaks down in high dimensions, and therefore analysts often rely on computationally intensive resampling techniques to perform valid inference. For example, genomic studies often require significance thresholds of the order of 10^{-6} and lower in order to correct for multiple testing. In this context, a permutation procedure would require at least 1 million resamples. By relying on results from random matrix theory, we can drastically cut down the required number of permutations. Since we only need to estimate two parameters from a location-scale family, good approximation is achieved with less than 100 permutations. Critically, this number is independent of the number of tests performed, and therefore the computation time is reduced by several orders of magnitude.

We motivated our empirical estimator of the CDF using an approximation theorem of Johnstone [2008]. Our approach is further motivated by several results from random matrix theory that suggest a central-limit-type theorem for random matrices, with the Tracy-Widom distribution replacing the normal distribution [Deift, 2007]. More evidence in support of our empirical estimator is given from the study of the largest root of Wishart variates. On the one hand, Srivastava [2007] and Srivastava and Fujikoshi [2006] showed that the asymptotic distribution of $\lambda_{\max}$ is approximately equal to the distribution of the largest eigenvalue of a scaled Wishart matrix. On the other hand, several results analogous to Johnstone's theorem were also derived for Wishart distributions. Indeed, it has been shown that if λ_1 is the largest root of a Wishart variate, then there exists μ, σ, both functions of the parameters of the Wishart distribution, such that the ratio $\frac{\lambda_1 - \mu}{\sigma}$ converges in distribution to $TW(1)$ [Johansson, 2000, Johnstone, 2001, El Karoui, 2003, Tracy and Widom, 2009].

Finally, the results from Section 4.6 show that the domain of applicability of our empirical estimator extends beyond that of a strict double Wishart problem. Specifically, we showed that we can replace the matrix $\mathbf{A}$ in Equation 4.1 by a linearly shrunk estimator based on earlier work by Ledoit and Wolf [2004]. This approach can easily be applied to multivariate

analysis approaches for which $\mathbf{A}$ is explicitly the covariance matrix of a multivariate random variable. Examples include the test of equality of covariance matrices and PCEV. However, more care is required in order to use these results with CCA.

Acknowledgements

Computations were made on the supercomputer Mammouth-parallèle 2 from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec—Nature et technologies (FRQ-NT).

The methylation data used for the data analysis was kindly provided by Marie Hudson, Sasha Bernatsky, Ines Colmegna, and Tomi Pastinen.

Finally, the authors would also want to thank Stepan Grinek for bringing to our attention the machine learning and pattern recognition literature on the use of truncated SVD.

Chapter 5

Region-based analysis of DNA methylation and anti-citrullinated protein antibody levels using Principal Component of Explained Variance

Preamble to Manuscript 3. One of my supervisors, Celia Greenwood, recently collaborated on a study that looked at the association between DNA methylation levels and anti-citrullinated protein antibody (ACPA) levels in individuals without clinical manifestations of Rheumatoid Arthritis (RA) [Shao et al., 2018]. ACPA is an antibody thought to be a predictor of RA onset. The statistical analysis performed was standard: CpG nucleotides were tested for significance with ACPA status (positive or negative) one at a time. However, the bisulfite sequencing platform used naturally grouped the CpG dinucleotides into small genomic regions. Indeed, these regions were explicitly targeted by the sequence capturing process. And as I described in the preamble to Chapter 3, correlation between the methylation levels of CpG dinucleotides is largely driven by base-pair distance. As a consequence,

significant correlation between the methylation levels within a region was ignored, reducing the efficiency of the univariate tests. Hence, the DNA methylation dataset of Shao et al. [2018] seemed an excellent opportunity to use PCEV and perform global association tests at the region level.

When I started to analyse the data, it became clear that the empirical estimator described in Chapter 4 would be required for some of the regions. Indeed, some regions contained more CpG dinucleotides than observations. Moreover, due to the presence of missing data induced from samples with zero reads, several observations needed to be removed from these analyses, making the high-dimensional problem more acute. Therefore, the analysis presented in this chapter applies many of the ideas discussed in the previous chapters. Results of this analysis are part of ongoing investigations with colleagues.

Region-based analysis of DNA methylation and anti-citrullinated protein antibody levels using Principal Component of Explained Variance

Maxime Turgeon[1], Xiaojian Shao[2], Kathleen Oros Klein[3], Ines Colmegna[4], Marie Hudson[4], Tomi Pastinen[2,5], Sasha Bernatsky[3,4], Aurélie Labbe[6], Celia M.T. Greenwood[1,2,3,7].

[1]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

[2]Department of Human Genetics, McGill University

[3]Lady Davis Institute for Medical Research, Jewish General Hospital

[4]Department of Medicine, McGill University

[5]Genome Québec Innovation Centre

[6]Department of Decision Sciences, HEC Montréal

[7]Department of Oncology, McGill University

Abstract

A recent nested case-control study looked at the association between DNA methylation and anti-citrullinated protein antibody (ACPA) levels, a key serological marker of rheumatoid arthritis (RA) risk. The methylation levels were measured using a novel targeted sequencing technique that naturally grouped the CpG dinucleotides into genomic regions; however, the association tests were performed one CpG dinucleotide at a time, without taking this grouping into account. In this article, we show how Principal Component of Explained Variance (PCEV) can be used to jointly analysed the ACPA levels and the methylation levels of CpG dinucleotides within a given region. PCEV is a dimension reduction method that seeks the linear combination of the methylation values that maximises the proportion of variance explained by the ACPA levels. The estimation procedure also takes into account additional variables and potential confounders (e.g. age, sex, cigarette smoking, cell-type composition). We build on recent work that provides a computationally efficient framework for computing p-values, even when the number of CpG dinucleotides within a region exceeds the number of observations. We show that our joint analysis uncovers differentially methylated regions (DMRs) that were missed by the univariate approach.

5.1 Motivation

Rheumatoid arthritis (RA) is an autoimmune disorder that primarily manifests itself with inflammation of the synovial joints [Carmona et al., 2010]. There are multiple genetic and environmental risk factors associated with RA. Furthermore, there is evidence of gene-environment interactions in the aetiology of the disease [Brennan and Silman, 1994, Karlson et al., 2010]. In turn, there is evidence that the role of environmental factors in the progression of RA is mediated through epigenetic factors, such as DNA methylation [Liu et al., 2013, Petronis, 2010]. DNA methylation is a chemical change of the DNA molecule that does not alter its sequence information. Methylation can repress gene transcription, and therefore it is related to one of the main mechanisms for cell-type diversity Goll and Bestor [2005], Klose and Bird [2006].

Due to its slow progression, the exact date of RA onset is sometimes difficult to establish [Raza and Gerlag, 2014]. This is critical: early treatment targeting remission is associated with improved long term functional outcomes (reduction of joint damage and lower risk of disability).

A molecular marker that has gained a lot of attention over the last decade is the *anti-citrullinated protein antibody* (ACPA). Citrullination is a biochemical process whereby the structure of a protein is slightly altered at a specific amino-acid residue (i.e. arginine). However, this small local change can have larger impacts on the folding patterns of the protein [Nicholas et al., 2017]. In turn, this change can become antigenic and hence elicit an immune response from the body; this process then gives rise to elevated levels of ACPA. Several cohort studies have shown that elevated levels of ACPA can be detected early in patients with RA, even before any clinical manifestation [Forslind et al., 2004, Mateen et al., 2016]. Understanding the relationship between methylation and ACPA levels in individuals without clinical manifestations of RA could inform on causal pathways, which could in turn lead to novel treatment targets.

To assess the association between ACPA levels and DNA methylation, a nested case-control study was recently performed on individuals randomly selected from the CARTaGENE cohort (`cartagene.qc.ca`). Firstly, ACPA status was determined for the set of randomly sampled individuals, and then DNA methylation was measured among two subsamples of individuals: those with either high or low ACPA status. Methylation levels were measured using a novel next-generation sequencing technique called *methylation capture sequencing* (MCC-Seq) [Shao et al., 2018]. The sequencing methodology targeted regions of the genome that were deemed by the authors to be potentially relevant to the study of autoimmune diseases, and therefore relevant to the study of RA. Although Shao et al. identified many CpG dinucleotides with evidence of an association with ACPA status (defined as high or low levels), their analysis was based on a univariate approach, i.e. analysing one CpG at a time, and therefore they did not leverage the natural correlation within these pre-defined regions. Ignoring this information may have resulted in lower statistical power than a multivariate strategy.

In this article, we use a multivariate approach to improve statistical power. Specifically, we use Principal Component of Explained Variance (PCEV) to analyse all CpG dinucleotides within a targeted region jointly. PCEV is a dimension reduction method whose objective is to find the linear combination of the response variables (e.g. methylation levels) that maximises the proportion of variance explained by the covariates (e.g. ACPA levels). We build on recent work by Turgeon et al. [2018a] to provide a hypothesis testing framework that applies to all regions, even when the number of CpG dinucleotides p is greater than the number of observations n.

The rest of this article is structured as follows: in the next section, we describe in detail our motivating dataset, and we cover the necessary methodological concepts alluded to in this introduction. Results are presented in Section 5.3 and discussed in Section 5.4. All methods used in this manuscript have been implemented in the `R` package `pcev`.

5.2 Methods

In this section, we give a high-level description of MCC-Seq so that the inputs to the analysis are clear. We then review PCEV and describe how it is used in our given context.

5.2.1 Data

DNA methylation is an epigenetic mark, i.e. a chemical modification of DNA that does not alter the sequence information. A common way to measure DNA methylation is via bisulfite sequencing: for every CpG covered by the sequencing library, and for each biological sample, we obtain a certain number M of methylated reads and a certain number U of unmethylated reads. The methylation level can then be represented as the proportion $\pi = M/(U + M)$ of methylated reads.

Using collected sera from a random subset of CARTaGENE subjects, the original study identified 69 ACPA positive subjects (ACPA levels above 20 units), representing around 2% of the randomly sampled individuals. The ACPA positive individuals were then matched for age, sex and smoking status to ACPA negative subjects (ACPA levels less 20 units). Whole blood from ACPA positive (N=63) and negative subjects (N=66) was used to extract DNA for the analyses. Proportions of the common blood cell types were available for all individuals in the dataset.

DNA methylation in the selected samples from CARTaGENE was measured using a recently developed high-throughput technique called methylation capture sequencing (MCC-Seq). For complete details, we refer the reader to Allum et al. [2015] and Cheung et al. [2017]. As opposed to whole-genome bisulfite sequencing, which essentially measures methylation levels at all CpG dinucleotides in the genome, MCC-Seq was developed to target pre-specified regions of the genome. In our setting, the regions were selected to cover areas of the genome that are relevant to immune diseases, e.g. gene promoters, methylation footprint regions observed in blood, and blood-cell-lineage specific enhancer regions [Shao et al., 2018].

The dataset included 4,219,158 CpG dinucleotides contained in 438,843 genomic regions. We removed all CpG dinucleotides for which coverage was zero for at least 50% of the individuals. After data cleaning, we were left with 179,701 CpG dinucleotides located in 23,350 regions.

5.2.2 Principal Component of Explained Variance

As described above, the methylation data is naturally grouped into genomic regions. Next, we explain how we can test for global association between methylation levels within a region and ACPA levels. Let M_{ijk}, U_{ijk} be the number of methylated and unmethylated reads, respectively, for individual i at CpG j in region k. Let $N_{ijk} = M_{ijk} + U_{ijk}$ be the total number of reads (also known as the *coverage* or *depth*), and let $\pi_{ijk} = M_{ijk}/N_{ijk}$ be the methylated proportion.

As the notation suggests, the depth N_{ijk} varies between CpG dinucleotides (for a fixed individual) and between individuals (for a fixed CpG dinucleotide). Accordingly, a larger value of N_{ijk} translates into more certainty around the methylated proportion π_{ijk}. Moreover, the variance of π_{ijk} also depends on the number of methylated reads M_{ijk}, leading to *heteroscedasticity*. To mitigate this effect on PCEV, we follow Park and Wu [2016], and we let $Y_{ijk} = \arcsin(2\pi_{ijk} - 1)$ be the transformed methylated proportions. Under a beta-binomial model for the proportion π_{ijk}, Park & Wu showed that the variance of Y_{ijk} only depends on the number of reads N_{ijk} and not on the mean of the underlying model, i.e. it is a variance-stabilising transformation. The residual variance can readily be accommodated by the PCEV model, as discussed next.

Let $\mathbf{Y}_k$ be the $n \times p_k$ matrix defined by $(\mathbf{Y}_k)_{ij} = Y_{ijk}$; hence, we have $n = 129$ samples and p_k CpG dinucleotides in region k. Let $\mathbf{X}$ be the n-dimensional vector containing the binary indicator for whether an individual is positive or negative for ACPA. Finally, let $\mathbf{Z}$ be the matrix containing the intercept, as well as data on additional variables and potential

confounders. We assume that the relationship between $\mathbf{X}$ and $\mathbf{Y}_k$ can be represented via a linear model:

$$\mathbf{Y}_k = \beta_k^T \mathbf{X} + \Gamma_k \mathbf{Z} + \mathbf{E}_k,$$

where β_k, Γ_k are the matrices of regression coefficients, and $\mathbf{E}_k$ is the matrix of residuals. PCEV seeks a linear combination of the outcome variables, $\mathbf{w}^T Y_k$, that maximises the proportion $R^2(\mathbf{w})$ of variance being explained by the covariate $\mathbf{X}$:

$$\hat{\mathbf{w}}_k = \underset{\mathbf{w} \in \mathbf{R}^{p_k} : \|\mathbf{w}\| = 1}{\operatorname{argmax}} R^2(\mathbf{w}),$$

where

$$R^2(\mathbf{w}) = \frac{\operatorname{Var}(\mathbf{w}^T \beta_k^T \mathbf{X})}{\operatorname{Var}(\mathbf{w}^T \beta_k^T \mathbf{X}) + \operatorname{Var}(\mathbf{w}^T \mathbf{E}_k)}. \tag{5.1}$$

In the equation above, the numerator corresponds to the variance of the covariates, and the denominator corresponds to the variance of the response variable (after adjusting for $\mathbf{Z}$). Turgeon et al. [2018b] have further details on this dimension-reduction technique.

To account for the residual variance alluded to in the previous paragraph, we assume that the data-generating process giving rise to the depth N_{ijk} does not depend on individual i. As a consequence, the residual variance can be estimated as part of the diagonal of the residual variance matrix $\mathbf{E}$.

As is common with sequencing data, the coverage N_{ijk} can sometimes be identically zero. In turn, this implies that the corresponding methylated proportion π_{ijk} is unknown. To fit the linear model even in the context of missing data, we opted for a *complete-case analysis*, i.e. for a given region k, only observations without missing values were retained. Therefore, the effective sample size for region k was $n_k \leq n$.

When $p_k > n_k$, we use a truncated Eigenvalue Decomposition (EVD) to maximise $R^2(\mathbf{w})$ [Turgeon et al., 2018a]. Let $V_M = \operatorname{Var}(\mathbf{w}^T \beta_k^T \mathbf{X}), V_T = \operatorname{Var}(\mathbf{w}^T \beta_k^T \mathbf{X}) + \operatorname{Var}(\mathbf{w}^T \mathbf{E}_k)$. Let r be the rank of V_T, with $r \leq p_k$. From ordinary EVD, we know there exists an orthogonal matrix T

of the same dimension as V_T such that

$$D = T^T V_T T$$

is a diagonal matrix with exactly r nonzero values on the diagonal. Let $D_{[r]}$ be the diagonal matrix obtained from D by keeping only those nonzero elements, and let $T_{[r]}$ be the $p_k \times r$ matrix defined by keeping only the columns of T corresponding to the nonzero elements of the diagonal of D. If we write $\tilde{T} = T_{[r]} D_{[r]}^{-1/2}$, we get

$$\tilde{T}^T V_T \tilde{T} = I_r.$$

We can then use ordinary EVD a second time and diagonalise $\tilde{T}^T V_M \tilde{T}$ using an orthogonal transformation T'. Therefore, if we let $S = \tilde{T} T'$, we get

$$S^T V_M S = \Lambda$$
$$S^T V_T S = I_r,$$

where Λ is again a diagonal matrix. The largest diagonal element of Λ is equal to the maximum of $R^2(\mathbf{w})$. Moreover, the column of S corresponding to this largest diagonal element is the vector $\mathbf{w}$ at which $R^2(\mathbf{w})$ is maximised.

As additional variables in $\mathbf{Z}$, we included sex, age, cigarette smoking status, and cell-type composition. These variables are all known to be associated with both methylation [Zhang et al., 2011, Horvath, 2013, Lee and Pausova, 2013, McGregor et al., 2016] and rheumatoid arthritis [Carmona et al., 2010].

To test for a global association between a region's methylation levels and ACPA levels, we use $\lambda = \max R^2(\mathbf{w})$ as a test statistic. Under the null hypothesis of no association, we can approximate the distribution of λ (after a logit transformation) using a location-scale

		ACPA positive (N=63)	ACPA negative (N=66)
	Age, mean (sd)	55.6 (8.1)	54.9 (7.5)
	Female, N (%)	39 (61.9%)	42 (63.6%)
Smoking Status, N (%)	Current	14 (22.2%)	13 (19.7%)
	Past	22 (34.9%)	26 (39.4%)
	Never	27 (42.9%)	27 (40.9%)

Table 5.1: Characteristics of individuals in the nested case-control study.

variant of the Tracy-Widom distribution of order 1 [Johnstone, 2008]. When $n_k \geq p_k$, these location and scale parameters can be computed exactly. When $n < p_k$, Turgeon et al. [2018a] showed how these parameters can be estimated using a small (i.e. around 50) number of permutations. Therefore, under both the classical and the high-dimensional scenario, we have a computationally efficient way of obtaining p-values. In other words, we do not need to rely on a full permutation procedure to perform inference.

5.3 Results

In Table 5.1, we can see a summary of the baseline characteristics of our subjects, providing evidence that matching was successful.

We performed PCEV independently on these 23,350 regions. The median number of regions per chromosome was 1034, with a minimum of 386 and a maximum of 2101. On average, a region contained 8 CpGs, with a minimum of 1 and a maximum of 127. However, after removing CpG sites with too much missing data and then observations with missing methylation values, the number of available subjects per region was typically lower than 129. On average, there were 48 observations per regions, with a minimum of 8 and a maximum of 128. Overall, 2519 regions had more CpGs than observations; this corresponds to 11% of all regions. These regions can be considered as high-dimensional datasets, and therefore we decided to use the method in Turgeon et al. [2018a] to efficiently compute p-values. For

all other regions, p-values were computed using the approximation to the null distribution described in Johnstone [2008] (see also Turgeon et al. [2018b]).

Given the number of regions, a Bonferroni correction for a Family-Wise Error Rate (FWER) of $\alpha = 0.05$ yields a cutoff point of 2.14×10^{-6}. At this level, 1062 regions were deemed statistically significant. A Manhattan plot depicting the results is presented in Figure 5.1. Note that the smallest p-value that can be estimated using the implementation of the Tracy-Widom distribution in `R` corresponds to 3.4×10^{-8}. This limitation explains why several p-values are aligned at the top of the graph. Further details on the top 25 most significant results appear in Table 5.2 (the regions are arranged according to the value of the test statistic). Moreover, the Supplementary Material contains information on the most frequent gene ontology terms appearing among the significant regions.

To better understand the overall significance of the association between methylation levels within a region and the levels of ACPA, we can look at the *variable importance factors* (VIF). These factors are defined as the absolute value of the (Pearson) correlation between a variable and the estimated PCEV component. As shown in [Turgeon et al., 2018b], the VIF serves as a proxy of the univariate association test, and it can therefore be used to rank the contribution of each CpG to the overall association. In Figure 5.2, we can see the VIF plots for the top 4 most significant regions.

We compared our results to a univariate approach. Specifically, we fitted a linear regression between the (transformed) methylation levels Y_{ijk} and the ACPA levels for each CpG dinucleotide. We also included the same additional variables as for the PCEV analysis (i.e. age, sex, cigarette smoking status, cell type composition). As for the PCEV analysis, we performed a complete-case analysis: for a given CpG dinucleotide, we removed all subjects with missing data. Of the 175,399 CpG dinucleotides in the data, 42 were significant (Bonferroni-corrected threshold for FWER of $\alpha = 0.05$: 2.85×10^{-7}). Figure 5.3 shows a Manhattan plot summarising these results. The 42 significant CpG dinucleotides were located in 5 regions.

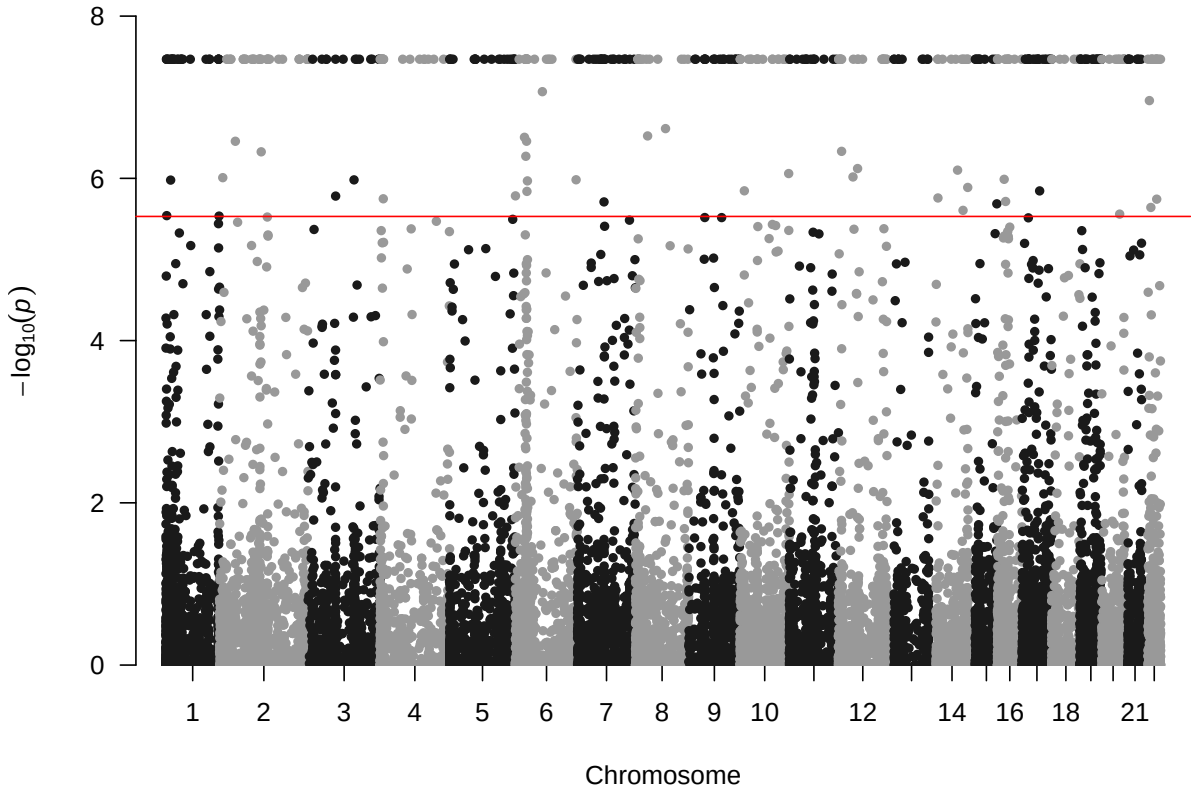


Figure 5.1: Manhattan plot of the results. The red horizontal line corresponds to a Bonferroni-corrected FWER of $\alpha = 0.05$.

Chromosome	Position	# CpGs	# Obs.	Test statistic	P-value
chr18	77638264	16	22	50.14	$< 3.4 \times 10^{-8}$
chr14	95893508	7	13	49.79	$< 3.4 \times 10^{-8}$
chr6	30095549	15	21	49.76	$< 3.4 \times 10^{-8}$
chr9	97109858	10	16	49.63	$< 3.4 \times 10^{-8}$
chr22	16404641	8	14	49.42	$< 3.4 \times 10^{-8}$
chr16	2030427	15	21	49.41	$< 3.4 \times 10^{-8}$
chr2	131010441	9	15	49.32	$< 3.4 \times 10^{-8}$
chr2	162270916	12	18	48.62	$< 3.4 \times 10^{-8}$
chr13	19240446	14	20	47.90	$< 3.4 \times 10^{-8}$
chr15	20104458	5	10	47.88	$< 3.4 \times 10^{-8}$
chr18	77616688	11	17	46.87	$< 3.4 \times 10^{-8}$
chr11	1268962	8	14	46.75	$< 3.4 \times 10^{-8}$
chr2	70369949	2	60	46.64	$< 3.4 \times 10^{-8}$
chr10	46096070	13	19	45.85	$< 3.4 \times 10^{-8}$
chr1	1840867	5	11	45.42	$< 3.4 \times 10^{-8}$
chr20	62917260	10	15	45.33	$< 3.4 \times 10^{-8}$
chr2	172970966	11	17	45.31	$< 3.4 \times 10^{-8}$
chr4	120133845	2	58	45.11	$< 3.4 \times 10^{-8}$
chr2	241585756	8	14	45.01	$< 3.4 \times 10^{-8}$
chr19	58446424	15	20	44.67	$< 3.4 \times 10^{-8}$
chr13	19172150	8	13	44.45	$< 3.4 \times 10^{-8}$
chr19	53324227	16	22	44.39	$< 3.4 \times 10^{-8}$
chr6	25652562	9	14	44.39	$< 3.4 \times 10^{-8}$
chr11	20229567	6	11	44.10	$< 3.4 \times 10^{-8}$
chr5	1641001	5	11	43.86	$< 3.4 \times 10^{-8}$

Table 5.2: Top 25 most significant regions ordered by test statistic: the position corresponds to the left-most CpG dinucleotide (relative to the p arm of the chromosome). We show the number of CpGs analysed as well as the number of individuals included in the analysis of the region, after removing missing data. All p-values correspond to the lowest resolution possible with the Tracy-Widom empirical estimator.

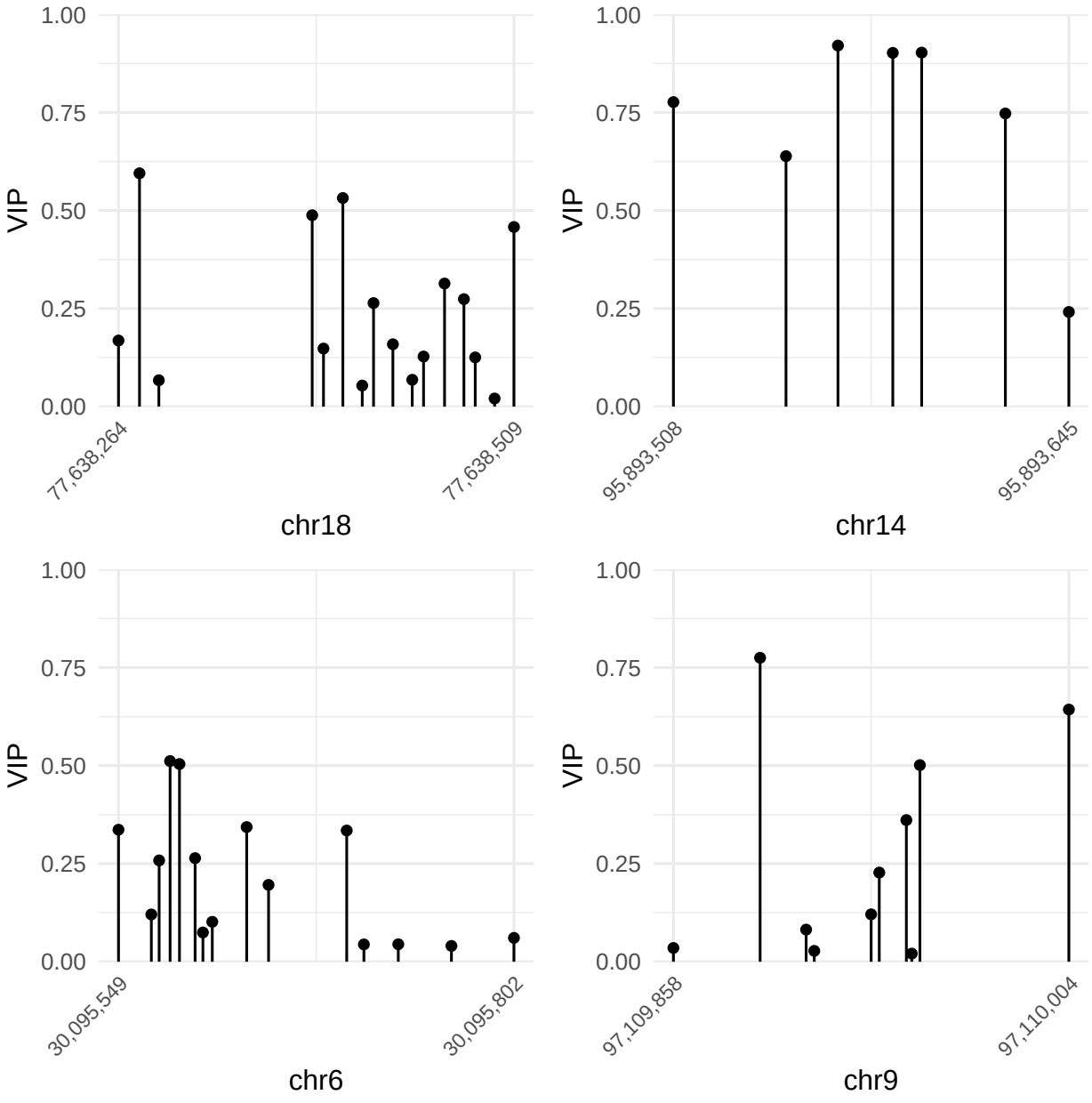


Figure 5.2: Variable Importance Factor (VIF) plots for the top 4 most significant regions.

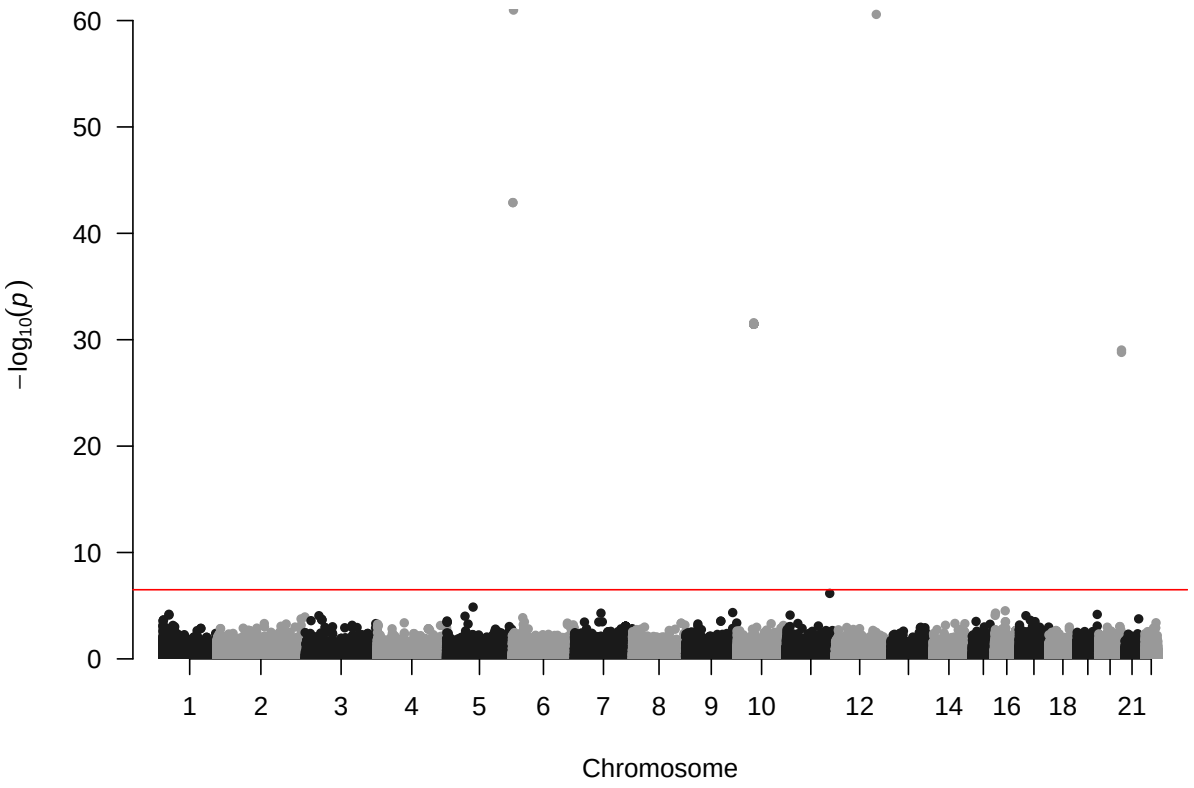


Figure 5.3: Manhattan plot of the univariate results. The red horizontal line corresponds to a Bonferroni-corrected FWER of $\alpha = 0.05$.

However, none of the PCEV p-values corresponding to these regions are significant. Finally, we also compared the univariate and the PCEV p-values (see Figure 5.4). In order to obtain a single p-value for each region, we took the minimal univariate p-value and applied a Bonferroni correction for the number of CpG dinucleotides in the region. Points appearing below the diagonal corresponds to regions for which the PCEV p-value was smaller than the univariate p-value.

Finally, Figure 5.5 shows the VIFs for the 5 regions containing CpG dinucleotides that were deemed significant by the univariate approach. Note that after accounting for missing data within the PCEV framework, methylation levels for some CpG dinucleotides were constant across samples, and therefore VIFs could not be measured. This explains why Figure 5.5

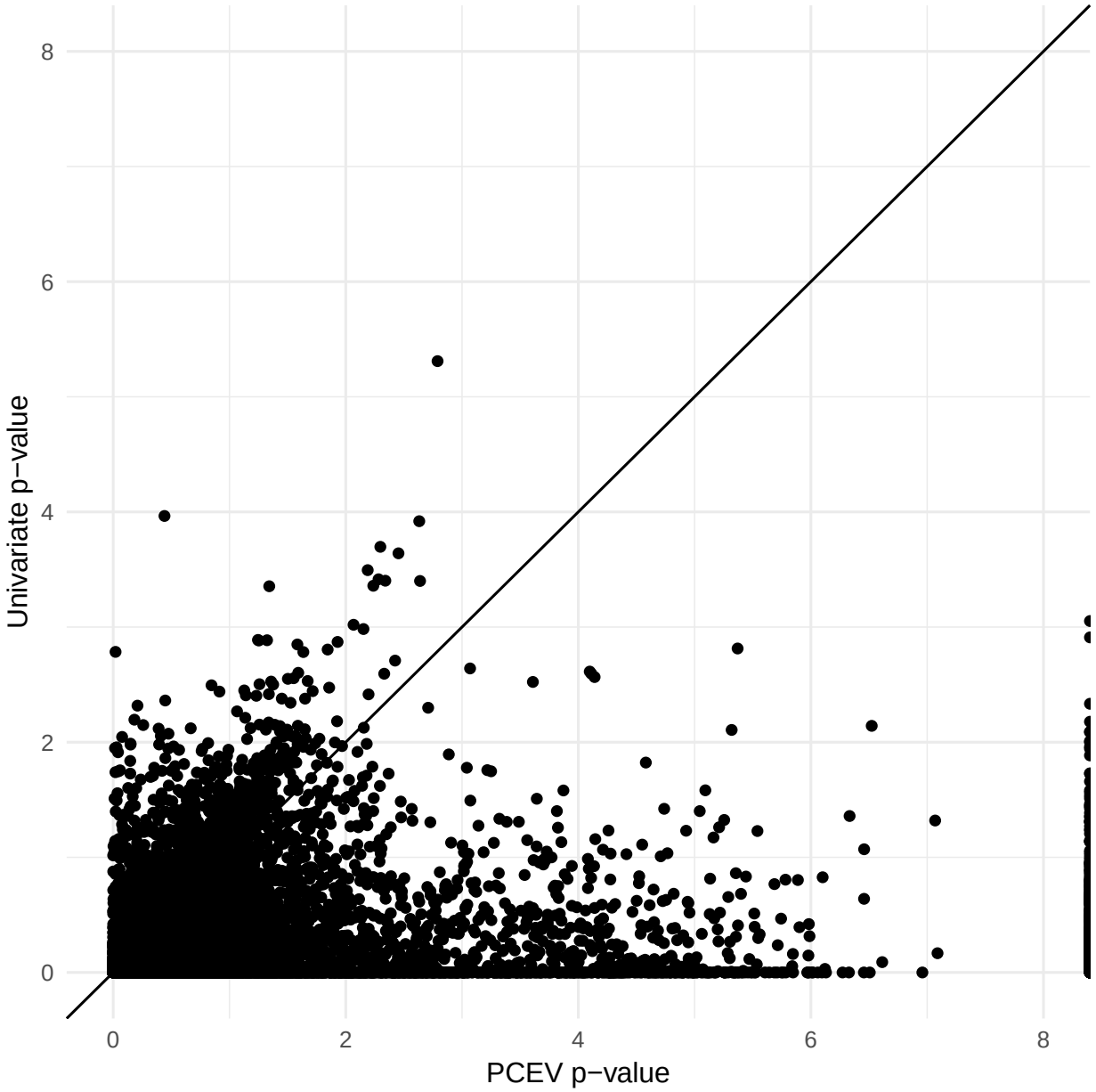


Figure 5.4: Comparison of PCEV and univariate p-values. Both methods used a complete-case analysis.

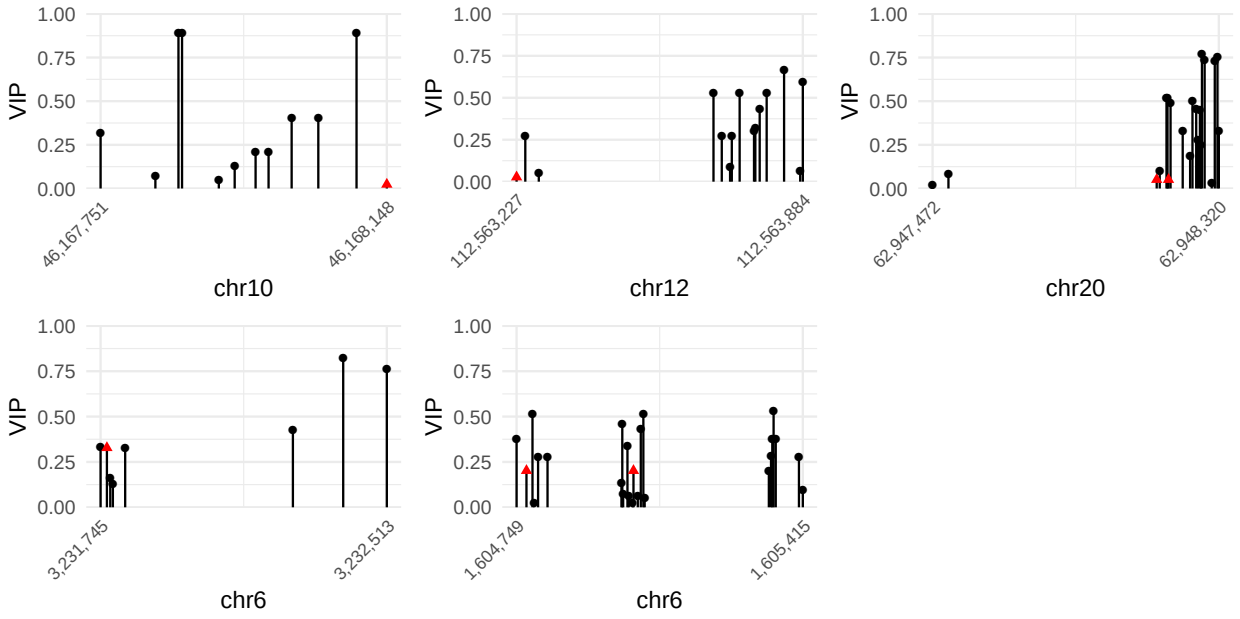


Figure 5.5: Variable Importance Factor (VIF) plots for the 5 regions containing CpG dinucleotides with significant univariate p-values (marked as red triangles).

contains less than 42 red triangles.

5.4 Discussion

In this article, we showed how the dimension reduction method PCEV can be used to perform region-based analyses of DNA methylation. Our example focused the association between DNA methylation and ACPA levels in asymptomatic subjects, where elevated ACPA is thought to be a precursor state for RA. The methylation levels were measured using a novel sequencing technique, and the CpG dinucleotides were naturally grouped into genomic regions. Multivariate analysis methods like PCEV can help leverage the natural correlation patterns between neighbouring CpG dinucleotides to obtain efficient test statistics for region-level association tests.

In DNA methylation analyses, there are a number of additional variables for which we typically need to control. On the one hand, matching was used to control for sex, age, and cigarette smoking status. On the other hand, the multivariate linear regression framework

on which PCEV is built was used to control for differences in blood-cell type composition. Nonetheless, sex, age, and cigarette smoking status were also added to the regression equation to increase efficiency, as they help explain a significant portion of the observed variability in DNA methylation levels. Adding them also helps account for any residual imbalances after matching.

Figure 5.1 highlights a computational limitation of our approach. The smallest p-value that can be measured corresponds to 3.4×10^{-8}. This is a consequence of the maximum accuracy level that was chosen when implementing the Tracy-Widom distribution in the package `RMTstat`. Indeed, the Tracy-Widom distribution can be defined as the solution to a certain differential equation, and numerical methods were used to approximate it to the desired level of accuracy. In the context of our study, this resolution still allowed us to make significance statement even after a Bonferroni correction. Therefore, the only impact this limitation has on our analysis is aesthetic. However, for studies requiring a higher resolution, there are two possible ways of addressing this limitation: 1) obtain a more accurate approximation of the Tracy-Widom distribution using the same numerical approach used for the development of the `RMTstat` package; 2) use a result of Chiani [2014] to approximate the Tracy-Widom distribution using a scaled and shifted gamma distribution. Alternatively, when computational efficiency is not a concern (e.g. when the number of regions is small), a permutation procedure can also be used to estimate p-values at any desired level of accuracy.

By looking at the VIF plots, we can see how the individual contribution of CpG dinucleotides to the overall significance can change from one region to the next. For example, the top row of Figure 5.2 gives two examples of regions where multiple CpGs have a similar contribution. On the contrary, the bottom row highlights regions where a couple CpG seems to be driving the overall association. This contrast is key to understanding the usefulness of multivariate methods: multiple, correlated moderate contributions can be combined to attain overall significance. A univariate approach would potentially have dismissed these moderate

correlations as being non-significant.

When comparing our approach to a univariate analysis, Figure 5.4 shows that the p-values obtained by PCEV tended to be smaller. This is expected: PCEV leverages the correlation pattern within a region to increase efficiency and statistical power. Unfortunately, due to the amount of missing data, there was no overlap between the significant regions identified by the two methods.

For both the univariate and PCEV analyses, we decided to perform a complete-case analysis, i.e. remove all observations with missing data. However, this approach has different consequences on the two methods: for a given region, if observation 1 is missing methylation value A, and observation 2 is missing methylation value B, both observations will be removed from the PCEV analysis, whereas for the univariate analysis, each observation can still be used for the analysis of any CpG for which it has a methylation level. As a consequence, the PCEV analysis tended to remove more observations from the analysis than the univariate approach. In turn, this contributed to some of the discrepancies in the results of the two approaches. For example, after restricting the PCEV analysis to complete observations only, the methylation levels of some of the remaining CpG dinucleotides were constant; that is, all remaining observations had the same methylation proportions for some CpG dinucleotides. This phenomenon occurred for some of the CpG dinucleotides deemed significant by the univariate analysis, and as a result, they could not contribute to the overall significance of the corresponding region. This is also illustrated in Figure 5.5, where less than 42 significant CpG dinucleotides are highlighted.

The complete-case analysis is a very common approach to handling missing data, and it corresponds to the default behaviour of the `R` function `lm`. Implicitly, we are making the assumption that the data is *missing completely at random* (MCAR) [Rubin, 1976]. Indeed, without this assumption, the estimators of the regression parameters are potentially biased for their corresponding population-level parameters. In the context of high-throughput tech-

nologies, an argument could be made that the MCAR assumption holds, at least partially. Indeed, some level of technical defects is expected from these technologies, and in turn this can lead to missing data. However, it is not clear that these technical errors account for all observed missing data. Therefore, more research is required to understand how missing data in next-generation sequencing studies departs from the MCAR assumption. Moreover, future work includes extending PCEV beyond the MCAR assumption. For example, imputation could be used to address some of these limitations; however, it is not clear how the imputation uncertainty could be folded into the PCEV uncertainty in a computationally efficient way (i.e. minimising the use of resampling techniques).

As we have highlighted in this study, multivariate analysis can have an important impact on statistical power when analysing correlated variables. Targeted sequencing methods like MCC-Seq are particularly well-suited to these methods, since they explicitly target specific genomic regions. On the other hand, since genetic distance is an important driver of correlation between methylation levels, multivariate methods could also be used to analyse data from any studies involving methylation sequencing. We hope that the present example encourages researchers to pursue this avenue further.

Acknowledgements

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: the Fonds de recherche Nature et Technologies (Doctoral award, MT) du Québec. The authors acknowledge the Ludmer Centre for Neuroinformatics and Mental Health for additional financial support.

Chapter 6

Conclusion

6.1 Summary

In this thesis, I presented three manuscripts, corresponding to Chapters 3, 4, and 5. Together, these manuscripts described specific approaches to dimension reduction with high-dimensional datasets, both from the point of view of estimation (i.e. estimating the dimension reduction parameters) and from the point of view of inference (i.e. how the extracted components provide evidence of global associations). As discussed in Chapter 2, a key assumption of many dimension reduction methods for high-dimensional data is *sparsity*: a small number of latent features give rise to the high-dimensional dataset. Indeed, most high-dimensional approaches to dimension reduction discussed in Chapter 2 leveraged this sparsity assumption. Chapter 3 proposes a different approach to PCEV motivated by this sparsity principle and implemented through structured estimation of the covariance matrix. Specifically, I showed that, if the response variables can be partitioned into blocks such that variables between different blocks are uncorrelated, PCEV can be performed in two stages: first, PCEV is performed one block at a time, and second, PCEV can be performed again on the obtained components. Through extensive simulations, I demonstrated that this block approach has

higher power than common high-dimensional approaches, and I showed that the inference results are quite robust to violations of the "block-independence" assumption. Critically, the block approach is a rare example of a high-dimensional approach that does not require any tuning parameter. This characteristic translates into high computational efficiency.

In Chapter 4, I looked at the double Wishart problem in the context of high-dimensional datasets. The largest root of double Wishart problems is commonly used as a test statistic for multivariate methods, such as MANOVA, CCA, and PCEV. As reviewed in Chapter 2, the distribution of the largest root under the null hypothesis of no association between the two datasets $\mathbf{Y}, \mathbf{X}$ has been extensively studied in the regular case, i.e. in the classical case where $p, q < n$. However, in high dimensions, very little is known about the distribution. The manuscript presented in Chapter 4 is one of the first to investigate the singular case and propose an efficient inferential strategy in high dimensions. Specifically, I provided numerical evidence that, after a suitable transformation, the largest root approximately follows a Tracy-Widom distribution of order 1. Building on that evidence, I showed that using a small number of permutations, we could estimate the location-scale parameters of this approximating distribution. As a consequence, we can compute valid p-values for high-dimensional inference for all double Wishart problems.

Finally, in Chapter 5, I showed how the ideas of the two previous manuscripts can be used for region-based analyses of DNA methylation levels. Specifically, I studied the association between DNA methylation levels and ACPA levels in individuals without any clinical manifestation of RA. The methylation levels were measured using a recently developed targeted custom capture sequencing method. The CpG dinucleotides were naturally grouped into regions when designing the molecular assay, and I showed that these regions could be used to perform a global association test between methylation and ACPA. For about 10% of the analysed regions, there were more CpG dinucleotides than observations, and therefore I used the results from Chapter 4 to compute p-values. This region-based approach uncovered many

Differentially Methylated Regions (DMRs) that were missed by the first analysis.

I also developed several `R` packages while carrying out the work in these three chapters. The package `pcev` implements the different versions of PCEV used in this thesis: the classical version (i.e. $n \geq p$), the block approach from Chapter 3, and the singular approach from Chapter 4. The package `covequal` implements the empirical estimator from Chapter 4 to perform a test of equality of covariance matrices with high-dimensional data. The package `rootWishart` implements the algorithms of Chiani [2014] and Chiani [2016]. The required arbitrary-precision linear algebra is implemented using the `C++` libraries `boost` and `eigen`. All three packages have been released on Comprehensive R Archive Network (CRAN) and are actively maintained.

6.1.1 Limitations

In Chapter 3, through extensive simulations, I demonstrated that inference is robust to violations of the block-independence assumption. However, estimation is not quite as robust. In other words, the PCEV component estimated using the block approach will vary depending on the strength of the correlation between the blocks. As seen in the simulation results of Chapter 3, this variability of the component has a limited impact on the Variable Importance on Projection (VIP) values, but it could have more important impacts on other downstream uses of the estimated component, e.g. if the component were used as part of a structural equation model.

In Chapter 4, I provided an empirical estimator that allows computing p-values in high dimensions. However, this estimator does not address any of the limitations coming from the decreasing signal-to-noise ratio that results from an increasing number of variables. The extension of the empirical estimator to the Ledoit-Wolf linear shrinkage covariance estimator partly addresses this issue, but in general, there still needs to be a cost-benefit analysis of computational efficiency versus statistical power when opting for the Tracy-Widom empirical

estimator.

In Chapter 5, the limitations of PCEV due to missing data are in the forefront. I opted for a complete-case analysis, but this approach can significantly decrease statistical power and even lead to biased estimates.

More generally, this thesis focused on the relationship between two sets of variables $\mathbf{Y}$ and $\mathbf{X}$. Accordingly, I excluded all multivariate methods developed for the analysis of three sets or more. These methods are often grouped under the umbrella terms "data integration" or "data fusion". Specific examples include generalized canonical correlation analysis [Horst, 1961, Kettenring, 1971] and multiblock factor analysis [Eslami et al., 2013]. However, I would like to highlight that some extensions of these data integration methods to high-dimensional datasets have followed a parallel development to what I described in Section 2.1.4. That is, regularised approaches have been proposed based on the sparsity principle. For example, see Tenenhaus and Tenenhaus [2014].

Furthermore, I restricted the discussion in this thesis to linear approaches to dimension reduction. That is, all dimension reductions were obtained through a linear projection of $\mathbf{Y}$ or $\mathbf{X}$. There is a vast literature on nonlinear dimension reduction, also known as *manifold learning*. Examples include kernel approaches to PCA and CCA, Locally Linear Embedding, Self-Organising Maps, and t-Distributed Stochastic Neighbour Embedding [Lee and Verleysen, 2007]. The focus of manifold learning is somewhat different than dimension reduction as presented here. Indeed, whereas this thesis is interested in making inference about the relationship between $\mathbf{Y}$ and $\mathbf{X}$, manifold learning is often seen as an exploratory or visualisation technique. For this reason, I excluded these approaches from consideration for this thesis.

6.2 Future work

The main chapters of this thesis highlight at least two potential avenues for future research. First, all three manuscripts presented multivariate analyses over subsets of the high-dimensional data: a gene-based analysis of the Assessment of Risk for Colorectal Tumors in Canada (ARCTIC) dataset; a pathway-based analysis of methylation data from patients with systemic auto-immune diseases; and a region-based analysis of methylation levels and ACPA levels. In all three settings, I used a Bonferroni correction to account for multiple testing across the analysed sets. This approach is likely conservative. Indeed, the correlation between methylation levels in different genomic segments or between pathways induces correlation between the corresponding test statistics. In Chapter 4, I used a heuristic to estimate the effective number of independent tests; this was achieved by looking at the average number of pathways under which a CpG dinucleotide was classified. However, more work is required for a systematic approach to estimating this effective number of independent tests. For PCEV and CCA, this amounts to studying the correlation between largest roots of related double Wishart problems. Currently, very little is known about these correlation patterns.

Another avenue for future work was mentioned in Chapter 5. Currently, in the presence of missing data, PCEV can only be used for a complete-case analysis. A natural compromise would be to first impute the missing data, and then perform PCEV. Future work would look into the effect of imputation on the overall properties of PCEV.

Finally, I am interested in understanding in what capacity nonlinear dimension reduction can help improve our understanding of the aetiology of complex diseases. Indeed, given the complex architecture of some of these diseases, where complex cascades and interactions of multiple factors are likely to occur, nonlinear dimension reduction could potentially capture latent structures that are more relevant to the biological question. I am currently conducting preliminary work addressing these questions in the context of predictive modelling.

6.3 Concluding remarks

High-dimensional data has driven many methodological advancements over the last decades. Dimension reduction methods can leverage the correlation between multiple phenotypes or molecular markers. In this way, they help decrease the multiple testing burden, which is particularly acute with high-dimensional datasets generated by high-throughput technologies. Accordingly, always improving high-dimensional approaches for multivariate analytical methods will be crucial if we want to make sense of all the biological data being generated daily. Robust, sound, and efficient statistical methods will continue to be important for improving our knowledge of the aetiology of complex diseases.

Appendices

APPENDIX A

Appendix to Manuscript 1

A.1 Appendix 1: proof of Theorem 1

The proof of Theorem 1 will follow from the proof of the following, more general result.

THEOREM 4. Let $\mathbf{A}$ and $\mathbf{B}$ be two positive semidefinite matrices of dimension $p \times p$, with $\mathbf{B}$ invertible. Assume $\mathbf{B}$ is block diagonal:

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{B}_b \end{pmatrix},$$

and decompose A similarly:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1b} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{b1} & \cdots & \mathbf{A}_{bb} \end{pmatrix}.$$

Then the maximisation of the expression

$$h(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}},$$

can be decomposed in $b+1$ distinct maximisations, as follows:

1. Maximise the expression
$$h_i(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A}_{ii} \mathbf{w}}{\mathbf{w}^T \mathbf{B}_i \mathbf{w}},$$
for $i = 1, \ldots, b$; let $\mathbf{u}_i$ be the solution of the i-th maximisation.

2. Define the following matrices:
$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{u}_1^T \mathbf{A}_{11} \mathbf{u}_1 & \cdots & \mathbf{u}_1^T \mathbf{A}_{1b} \mathbf{u}_b \\ \vdots & \ddots & \vdots \\ \mathbf{u}_b^T \mathbf{A}_{b1} \mathbf{u}_1 & \cdots & \mathbf{u}_b^T \mathbf{A}_{bb} \mathbf{u}_b \end{pmatrix}, \quad \tilde{\mathbf{B}} = \begin{pmatrix} \mathbf{u}_1^T \mathbf{B}_1 \mathbf{u}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{u}_b^T \mathbf{B}_b \mathbf{u}_b \end{pmatrix}.$$
Maximise the expression
$$H(\mathbf{w}) = \frac{\mathbf{w}^T \tilde{\mathbf{A}} \mathbf{w}}{\mathbf{w}^T \tilde{\mathbf{B}} \mathbf{w}},$$
and denote the maximiser of $H(\mathbf{w})$ by $\mathbf{v} = (v_1, \ldots, v_b)$.

3. The maximiser of the original expression $h(\mathbf{w})$ is thus given by $\mathbf{w} = (v_1 \mathbf{u}_1, \ldots, v_b \mathbf{u}_b)$.

Proof. We start by solving the original maximisation problem. Let $\mathbf{L}$ be a square-root of $\mathbf{B}$, so that $\mathbf{B} = \mathbf{L}\mathbf{L}^T$. We thus have
$$\begin{aligned} h(\mathbf{w}) &= \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \\ &= \frac{\mathbf{w}^T (\mathbf{L}\mathbf{L}^{-1}) \mathbf{A} (\mathbf{L}^{-T}\mathbf{L}^T) \mathbf{w}}{\mathbf{w}^T \mathbf{L}\mathbf{L}^T \mathbf{w}} \\ &= \frac{(\mathbf{w}^T \mathbf{L}) \mathbf{L}^{-1} \mathbf{A} \mathbf{L}^{-T} (\mathbf{L}^T \mathbf{w})}{(\mathbf{w}^T \mathbf{L})(\mathbf{L}^T \mathbf{w})}. \end{aligned}$$

Therefore, if $\mathbf{c}$ is an eigenvector of $\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T}$ associated to its largest eigenvalue, then $\mathbf{w} := \mathbf{L}^{-T}\mathbf{c}$ maximises $h(\mathbf{w})$.

Since $\mathbf{B}$ (and thus $\mathbf{L}$) is block diagonal, we can be explicit about the matrix $\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T}$; namely, we have

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T} = \begin{pmatrix} \mathbf{L}_1^{-1}\mathbf{A}_{11}\mathbf{L}_1^{-T} & \cdots & \mathbf{L}_1^{-1}\mathbf{A}_{1b}\mathbf{L}_b^{-T} \\ \vdots & \ddots & \vdots \\ \mathbf{L}_b^{-1}\mathbf{A}_{b1}\mathbf{L}_1^{-T} & \cdots & \mathbf{L}_b^{-1}\mathbf{A}_{bb}\mathbf{L}_b^{-T} \end{pmatrix},$$

where $\mathbf{L}_i$ is the i-th block of $\mathbf{L}$.

Now, we repeat the same analysis but with the first b maximisations. For a fixed i, we have

$$\begin{aligned} h_i(\mathbf{w}) &= \frac{\mathbf{w}^T\mathbf{A}_{ii}\mathbf{w}}{\mathbf{w}^T\mathbf{B}_i\mathbf{w}} \\ &= \frac{\mathbf{w}^T(\mathbf{L}_i\mathbf{L}_i^{-1})\mathbf{A}_{ii}(\mathbf{L}_i^{-T}\mathbf{L}_i^T)\mathbf{w}}{\mathbf{w}^T\mathbf{L}_i\mathbf{L}_i^T\mathbf{w}} \\ &= \frac{(\mathbf{w}^T\mathbf{L}_i)\mathbf{L}_i^{-1}\mathbf{A}_{ii}\mathbf{L}_i^{-T}(\mathbf{L}_i^T\mathbf{w})}{(\mathbf{w}^T\mathbf{L}_i)(\mathbf{L}_i^T\mathbf{w})}. \end{aligned}$$

Similarly as above, if $\mathbf{c}_i$ is an eigenvector of $\mathbf{L}_i^{-1}\mathbf{A}_{ii}\mathbf{L}_i^{-T}$ associated to its largest eigenvalue, then $\mathbf{u}_i := \mathbf{L}_i^{-T}\mathbf{c}_i$ maximises $h_i(\mathbf{w})$; let $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_b)$ be the concatenation of all these vectors, and note that its length is equal to p.

The final maximisation takes more care. First, we note that

$$\begin{aligned} \tilde{\mathbf{B}} &= \begin{pmatrix} \mathbf{u}_1^T\mathbf{B}_1\mathbf{u}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{u}_b^T\mathbf{B}_b\mathbf{u}_b \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{u}_1^T\mathbf{L}_1\mathbf{L}_1^T\mathbf{u}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{u}_b^T\mathbf{L}_b\mathbf{L}_b^T\mathbf{u}_b \end{pmatrix} \\ &= \operatorname{diag}\left(\|\mathbf{L}_i^T\mathbf{u}_i\|\right)^2. \end{aligned}$$

We have thus found a square root for the $b \times b$ diagonal matrix $\tilde{\mathbf{B}}$.

As above, the maximisation of $H(\mathbf{w})$ is related to an eigenvalue problem. In this case, we are looking for an eigenvector of the matrix

$$\operatorname{diag}\left(\|\mathbf{L}_i^T \mathbf{u}_i\|\right)^{-1} \cdot \tilde{\mathbf{A}} \cdot \operatorname{diag}\left(\|\mathbf{L}_i^T \mathbf{u}_i\|\right)^{-1} = \left[\|\mathbf{L}_i^T \mathbf{u}_i\|^{-1} (\mathbf{u}_i^T \mathbf{A}_{ij} \mathbf{u}_j) \|\mathbf{L}_j^T \mathbf{u}_j\|^{-1}\right]_{ij}.$$

Let $\mathbf{d} = (d_1, \ldots, d_b)$ be an eigenvector of this matrix associated to its largest eigenvalue. It finally follows that

$$\mathbf{v} = (v_1, \ldots, v_b) = \operatorname{diag}\left(\|\mathbf{L}_i^T \mathbf{u}_i\|\right)^{-1} \mathbf{d} = \left(\|\mathbf{L}_1^T \mathbf{u}_1\|^{-1} d_1, \ldots, \|\mathbf{L}_b^T \mathbf{u}_b\|^{-1} d_b\right)$$

maximises $H(\mathbf{w})$.

Recall our claim that $\mathbf{w} = (v_1 \mathbf{u}_1, \ldots, v_b \mathbf{u}_b)$ maximises the original criterion

$$h(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}}.$$

We can write

$$\begin{aligned}
\mathbf{w} &= \begin{pmatrix} v_1\mathbf{u}_1 \\ \vdots \\ v_b\mathbf{u}_b \end{pmatrix} \\
&= \begin{pmatrix} \|\mathbf{L}_1^T\mathbf{u}_1\|^{-1}d_1\mathbf{u}_1 \\ \vdots \\ \|\mathbf{L}_b^T\mathbf{u}_b\|^{-1}d_b\mathbf{u}_b \end{pmatrix} \\
&= \begin{pmatrix} \|\mathbf{c}_1\|^{-1}d_1\mathbf{L}_1^{-T}\mathbf{c}_1 \\ \vdots \\ \|\mathbf{c}_b\|^{-1}d_b\mathbf{L}_b^{-T}\mathbf{c}_b \end{pmatrix} \\
&= \mathbf{L}^{-T}\begin{pmatrix} \|\mathbf{c}_1\|^{-1}d_1\mathbf{c}_1 \\ \vdots \\ \|\mathbf{c}_b\|^{-1}d_b\mathbf{c}_b \end{pmatrix}.
\end{aligned}$$

Therefore, to finish the proof, we need to show that $\tilde{\mathbf{w}} = (\|\mathbf{c}_1\|^{-1}d_1\mathbf{c}_1, \ldots, \|\mathbf{c}_b\|^{-1}d_b\mathbf{c}_b)$ is an eigenvector of $\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T}$ associated to its largest eigenvalue. In other words, we need to show that $\tilde{\mathbf{w}}$ maximises the ratio

$$\tilde{h}(\mathbf{w}) = \frac{\mathbf{w}^T(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T})\mathbf{w}}{\mathbf{w}^T\mathbf{w}}.$$

We first compute the denominator:

$$\begin{aligned}
\tilde{\mathbf{w}}^T\tilde{\mathbf{w}} &= (\|\mathbf{c}_1\|^{-1}d_1\mathbf{c}_1,\ldots,\|\mathbf{c}_b\|^{-1}d_b\mathbf{c}_b)^T(\|\mathbf{c}_1\|^{-1}d_1\mathbf{c}_1,\ldots,\|\mathbf{c}_b\|^{-1}d_b\mathbf{c}_b)\\
&= \sum_{i=1}^{b}\|\mathbf{c}_i\|^{-2}d_i^2\mathbf{c}_i^T\mathbf{c}_i\\
&= \sum_{i=1}^{b}\|\mathbf{c}_i\|^{-2}d_i^2\|\mathbf{c}_i\|^2\\
&= \sum_{i=1}^{b}d_i^2\\
&= \mathbf{d}^T\mathbf{d}.
\end{aligned}$$

On the other hand, the numerator is given by

$$\begin{aligned}
\tilde{\mathbf{w}}^T(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T})\tilde{\mathbf{w}} &= \sum_{i=1}^{b}\sum_{j=1}^{b}\|\mathbf{c}_i\|^{-1}d_i\mathbf{c}_i^T\mathbf{L}_i^{-1}\mathbf{A}_{ij}\mathbf{L}_j^{-T}\|\mathbf{c}_j\|^{-1}d_j\mathbf{c}_j\\
&= \sum_{i=1}^{b}\sum_{j=1}^{b}\|\mathbf{L}_i^T\mathbf{u}_i\|^{-1}d_i(\mathbf{u}_i^T\mathbf{L}_i)\mathbf{L}_i^{-1}\mathbf{A}_{ij}\mathbf{L}_j^{-T}\|\mathbf{L}_j^T\mathbf{u}_j\|^{-1}d_j(\mathbf{L}_j^T\mathbf{u}_j)\\
&= \sum_{i=1}^{b}\sum_{j=1}^{b}d_i\|\mathbf{L}_i^T\mathbf{u}_i\|^{-1}(\mathbf{u}_i^T\mathbf{A}_{ij}\mathbf{u}_j)\|\mathbf{L}_j^T\mathbf{u}_j\|^{-1}d_j\\
&= d^T\left[\|\mathbf{L}_i^T\mathbf{u}_i\|^{-1}(\mathbf{u}_i^T\mathbf{A}_{ij}\mathbf{u}_j)\|\mathbf{L}_j^T\mathbf{u}_j\|^{-1}\right]_{ij}d.
\end{aligned}$$

In other words, the ratio $\tilde{h}(\tilde{\mathbf{w}})$ evaluated at $\tilde{\mathbf{w}}$ is equal to

$$\tilde{h}(\tilde{\mathbf{w}}) = \frac{\mathbf{d}^T\left[\|\mathbf{L}_i^T\mathbf{u}_i\|^{-1}(\mathbf{u}_i^T\mathbf{A}_{ij}\mathbf{u}_j)\|\mathbf{L}_j^T\mathbf{u}_j\|^{-1}\right]_{ij}\mathbf{d}}{\mathbf{d}^T\mathbf{d}}.$$

But since $\mathbf{d}$ is an eigenvector for the matrix $\left[\|\mathbf{L}_i^T\mathbf{u}_i\|^{-1}(\mathbf{u}_i^T\mathbf{A}_{ij}\mathbf{u}_j)\|\mathbf{L}_j^T\mathbf{u}_j\|^{-1}\right]_{ij}$ associated to its largest eigenvalue, we know that this ratio is maximised. This concludes the proof. □

As a consequence of this theorem, we have proven the validity of the block approach. Note

that since the maximisation of the two ratios

$$\frac{\mathbf{w}^T\mathbf{V}_M\mathbf{w}}{\mathbf{w}^T(\mathbf{V}_M+\mathbf{V}_R)\mathbf{w}}, \qquad \frac{\mathbf{w}^T\mathbf{V}_M\mathbf{w}}{\mathbf{w}^T\mathbf{V}_R\mathbf{w}}$$

is equivalent, it follows that the independence assumption needed for the block approach can be made either conditional on the covariates $\mathbf{X}$ or unconditionally. Moreover, because of the generality of the theorem, we can deduce that the block approach also works for extracting further components; one simply needs to regress the first component on $\mathbf{Y}$ and perform PCEV on the residuals to get the second component, and so on.

A.2 Appendix 2: Wilks test of significance

Let X be a single covariate, and consider the multivariate regression of X on a p-dimensional response vector $\mathbf{Y} = (Y_1, \ldots, Y_p)$. It is well known that

$$F := \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)} \sim F(p, n-p-1).$$

Moreover, note that we can rewrite F as follows:

$$\begin{aligned} F &= \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)} \\ &= \frac{R^2/p}{(1-R^2)/(n-p-1)} \\ &= \left(\frac{R^2}{1-R^2}\right) \cdot \left(\frac{n-p-1}{p}\right). \end{aligned}$$

It remains to show that $\lambda = R^2/(1-R^2)$. But since R^2 is the coefficient of multivariate correlation between X and the p variables $Y_1, \ldots, Y_p$, R^2 is defined as the maximum correlation between X and any linear combination of $\mathbf{Y}$:

$$R^2 := \max_w \mathrm{Cor}(X, w^T\mathbf{Y}).$$

Therefore, R^2 must be equal to the first canonical correlation between $\mathbf{Y}$ and X. Hence, we have

$$R^2 = \frac{\lambda}{1+\lambda} \Leftrightarrow \lambda = \frac{R^2}{1-R^2}.$$

This completes the proof. □

A.3 Appendix 3: Roy's largest root test statistic

In the hypothesis test framework developed above, we presented two tests based on the largest eigenvalue λ of the matrix $\mathbf{V}_R^{-1}\mathbf{V}_M$, which is also the largest root of the determinantal equation

$$\det(\mathbf{V}_M - \lambda \mathbf{V}_R) = 0.$$

Under the assumption that there is only one covariate, we can derive the exact distribution of the ratio $\lambda/(1+\lambda)$; as noted in Appendix 2, this results holds as long as $n > p+1$, where n is the sample size and p, the number of response variables.

More generally, Johnstone [2009] derived an approximation to the distribution of λ, after a suitable transformation. The result, using the notation of this article, is given below:

THEOREM 5. [Johnstone, 2008] As $n, q, p \to \infty$, we have

$$\frac{\log \lambda - \mu}{\sigma} \xrightarrow{\mathcal{D}} TW(1),$$

where $TW(1)$ is the Tracy-Widom distribution of order 1 [Tracy and Widom, 1996], and μ, σ are defined as follows: let $s = \min(p, q)$ and write

$$M = \frac{|\, p - q \,| - 1}{2}, \qquad N = \frac{n - q - p - 2}{2}, \qquad L = 2(s + M + N) + 1.$$

Next, define

$$\sin^2(\gamma/2) = \frac{s - 1/2}{L}$$
$$\sin^2(\varphi/2) = \frac{s + 2M + 1/2}{L}.$$

Finally, we set

$$\mu = 2\log\left(\tan\left(\frac{\varphi+\gamma}{2}\right)\right)$$
$$\sigma^3 = \frac{16}{n}\left(\sin^2(\varphi+\gamma)\sin\varphi\sin\gamma\right)^{-1}.$$

Under the scenario where there is only one covariate and $n > p+1$, we recommend using the Wilks' Lambda test, since it typically has higher power than the test based on Johnstone's approximation. In any case, Theorem 5 can be used in much more generality, even when $p \gg n$.

A.4 Appendix 4: Relationship to other multivariate approaches

A.4.1 Notations

To discuss the relationship of PCEV with other existing methods, one needs to introduce some notations and background. Assume that the vectors Y and X are measured on the same sampling unit $i, (i = 1, \ldots, n)$, and denote $\mathbf{Y}$ and $\mathbf{X}$ the matrices $n \times p$ and $n \times q$ of the observed data, respectively. Without loss of generality, we will assume that $\mathbf{Y}$ and $\mathbf{X}$ are centred matrices. Model (1) in the article leads to a multivariate regression model between $\mathbf{Y}$ and $\mathbf{X}$. Thus, one can write

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where the matrix $\mathbf{B}$ is the coefficients of the model.

The least square estimator of $\mathbf{B}$ is given by

$$\hat{\mathbf{B}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}, \tag{A.1}$$

where $\mathbf{S}_{xx}$ and $\mathbf{S}_{xy}$ are block forms of the overall sample covariance matrix $\mathbf{S}$ of the vector $(y_1, \ldots, y_p, x_1, \ldots, x_q)$, with $\mathbf{S}$ given as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}. \tag{A.2}$$

One can also write the total sum of squares and products as

$$\begin{aligned} \mathbf{Y}^T\mathbf{Y} &= \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y} + \hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{E} + \mathbf{H}, \end{aligned} \tag{A.3}$$

where

$$\mathbf{E} = \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y},$$

is the overall sum of squares and products that is unexplained by $\mathbf{X}$ (i. e. residuals) and

$$\mathbf{H} = \hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y},$$

is the overall regression sum of squares and products matrix (i. e. variance explained by $\mathbf{X}$). Notice also that following (A.1), one can write $\mathbf{E}$ and $\mathbf{H}$ in terms of the blocks form of the overall sample covariance matrix $\mathbf{S}$ in (A.2) as follow:

$$\mathbf{E} = \mathbf{S}_{yy} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}, \tag{A.4}$$

$$\mathbf{H} = \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}. \tag{A.5}$$

A.4.2 PCEV and its relationship to CCA

Recall that PCEV seeks a linear combination of outcomes, $\mathbf{w}^T\mathbf{Y}$, which maximises the ratio $h^2(\mathbf{w})$ of variance being explained by $\mathbf{X}$. Thus, following (A.3), the heritability $h^2(\mathbf{w})$ defined in section 2.1 of the paper can be written as

$$h^2(w) = \frac{w^T\mathbf{H}w}{w^T(\mathbf{H}+\mathbf{E})w},$$

If one substitutes $\mathbf{E}$ and $\mathbf{H}$ by their values from equations (A.4) and (A.5), then one can write

$$h^2(\mathbf{w}) \;=\; \frac{\mathbf{w}^T\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{w}}{\mathbf{w}^T\mathbf{S}_{yy}\mathbf{w}}. \tag{A.6}$$

Thus, the vector $\mathbf{w}$ that maximises $h^2(\mathbf{w})$ in (A.6) can be obtained as the first eigenvector of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$, which is the first canonical direction on $\mathbf{Y}$. The maximum heritability in this case is the largest squared canonical correlation, r_1^2.

Therefore, even if PCEV seeks the best linear combination of $\mathbf{Y}$, since it maximises the ratio of the variance explained to total variance, it implicitly looks for the best linear combination of $\mathbf{X}$ with maximum correlation with $\mathbf{w}_{PCEV}^T\mathbf{Y}$.

A.4.3 PCEV and its relationship to LDA and one-way MANOVA

Assume that we have K groups (e. g. factor with K levels) and assume that the $p \times 1$ vector $\mathbf{Y}$ is measured on the sampling unit $i, (i = 1, \ldots, n)$ for each group $k = 1, \ldots, K$, and denote $\mathbf{Y}$ the matrices $nK \times p$ of the observed data. Notice that here we deal with a simple balanced case where n the number of units at each group is the same. All the results remain the same for the unbalanced case.

In order to establish a relationship between PCEV and LDA (and one-way MANOVA), one needs to rewrite the MANOVA model as a multivariate regression model. To do so, one needs to assume first $K-1$ dummy variables $x_k, k = 1 \ldots, (K-1)$, defined as

$$x_k = \begin{cases} 1 & \text{for subjects belonging to group k} \\ 0 & \text{elsewhere} \end{cases}$$

Then, one can write the MANOVA model as the multivariate regression of $\mathbf{Y}$ on the dummy variables $x_k, k = 1 \ldots, (K-1)$. One can write,

$$\mathbf{Y} = \mathbf{XB} + E.$$

Following equation (A.3), one can verify that $\mathbf{E}$ and $\mathbf{H}$ are exactly the within sum of squares and products and the between sum of squares and products, respectively.

Again, from the definition of heritability, one can see that PCEV searches for direction on the column space of $\mathbf{Y}$ that maximises the ratio of the variation explained by the groups to the total variation. That is, one can write

$$h^2(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{SSB}\mathbf{w}}{\mathbf{w}^T(\mathbf{SSB}+\mathbf{SSW})\mathbf{w}},$$

where **SSB** and **SSW** are the within sum of squares and products and the between sum of squares and products, respectively. Thus, the vector w_{PCEV} that maximises the heritability is exactly the first linear discriminant function. In the MANOVA context, $h^2(\mathbf{w}_{PCEV})$ is termed *Roy's largest root test*. Another interpretation of $h^2(\mathbf{w}_{PCEV})$ is as the maximum squared correlation r_1^2, between the first discriminant function and the best linear combination of the $K-1$ dummy variables. The later interpretation can be concluded from the relationship between PCEV and CCA.

Note that the heritability or *Roy's largest root test* statistic is equal to

$$h^2(w_{PCEV}) \quad = \quad \frac{\lambda_1}{1+\lambda_1}, \tag{A.7}$$

where λ_1 is the largest eigenvalue of $\mathbf{SSW}^{-1}\mathbf{SSB}$. For $K > 2$, its distribution is difficult to determine but an approximation can be used to calculate an exact p-value. SAS uses Davies approximation to calculate the p-value. In the case of only two groups ($K = 2$), λ_1 is proportional to the Hoteling statistic which is the generalisation of the squared of the t-test statistic in the univariate case. Thus λ_1 can be transformed to a Fischer exact-test statistic, and can be used instead of $h^2(\mathbf{w}_{PCEV})$.

A.5 Appendix 5: Supplementary figures and results

Table A.1: Type I error of PCEV as a function of the within-block correlation ρ_W and the number of responses p, for several significance levels.

Within-block correlation	Significance level	Number of response variables					
		$p = 20$	$p = 50$	$p = 100$	$p = 200$	$p = 300$	$p = 400$
$\rho_W = 0$	0.1	0.09922	0.09978	0.09958	0.10139	0.09971	0.09954
	0.05	0.05044	0.04908	0.04996	0.05057	0.05033	0.05089
	0.01	0.01015	0.00994	0.01004	0.01008	0.00970	0.01030
	0.005	0.00550	0.00498	0.00512	0.00503	0.00473	0.00530
	0.001	0.00108	0.00108	0.00110	0.00084	0.00096	0.00109
	0.0005	0.00055	0.00058	0.00050	0.00033	0.00053	0.00069
	0.0001	0.00007	0.00011	0.00011	0.00005	0.00009	0.00016
$\rho_W = 0.5$	0.1	0.09902	0.10070	0.09958	0.10139	0.09971	0.09954
	0.05	0.04960	0.05010	0.04996	0.05057	0.05033	0.05089
	0.01	0.00993	0.00979	0.01004	0.01008	0.00970	0.01030
	0.005	0.00497	0.00495	0.00512	0.00503	0.00473	0.00530
	0.001	0.00093	0.00078	0.00110	0.00084	0.00096	0.00109
	0.0005	0.00043	0.00043	0.00050	0.00033	0.00053	0.00069
	0.0001	0.00009	0.00011	0.00011	0.00005	0.00009	0.00016
$\rho_W = 0.7$	0.1	0.09977	0.10118	0.09943	0.09863	0.10001	0.09977
	0.05	0.05022	0.05105	0.04874	0.04915	0.04981	0.05036
	0.01	0.01005	0.00994	0.00974	0.01022	0.00952	0.01013
	0.005	0.00490	0.00504	0.00485	0.00532	0.00462	0.00490
	0.001	0.00109	0.00108	0.00101	0.00113	0.00094	0.00080
	0.0005	0.00049	0.00064	0.00054	0.00064	0.00039	0.00036
	0.0001	0.00013	0.00011	0.00012	0.00010	0.00013	0.00010

Table A.2: Running times relative to PCEV for four high-dimensional methods as a function of the number of outcomes p. There is no correlation between the response variables (results are similar for other correlation structures).

	p				
	100	200	300	400	500
PCEV	1	1	1	1	1
lasso	53	27	19	15	9
sPLS	507	446	348	289	253
rCCA	2221	4733	6968	8969	11347

Figure A.1: Type I error as a function of the correlation parameters ρ_w and ρ_b, and the number of responses p.

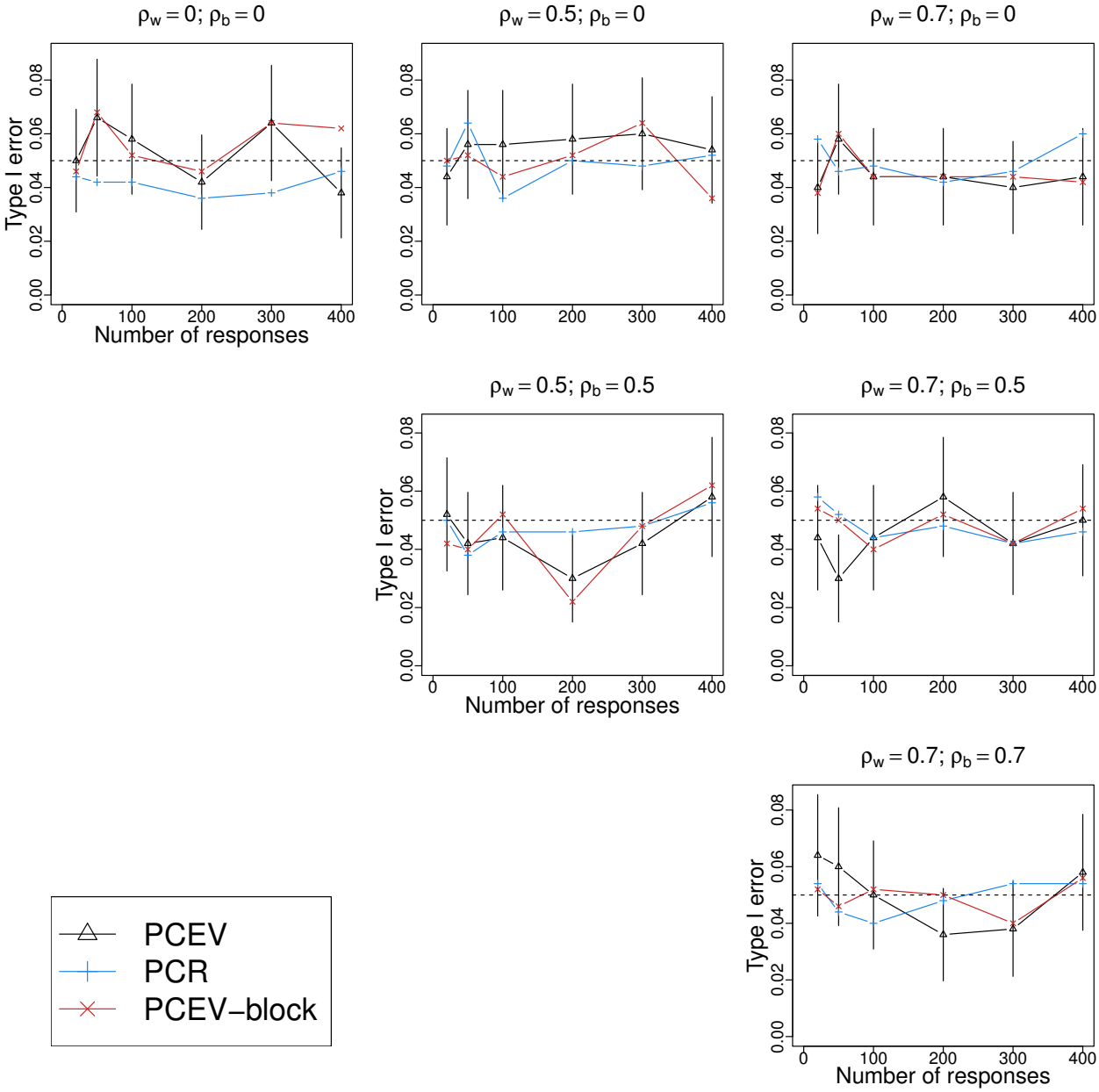


Figure A.2: True positive rate as a function of the number of selected variables in the high-dimensional simulation scenario. Panels are arranged by sample size ($p = 100, 200, 300, 400, 500$) and correlation structure (each pair corresponds to the within-block ρ_w and between-block ρ_b correlation, respectively).

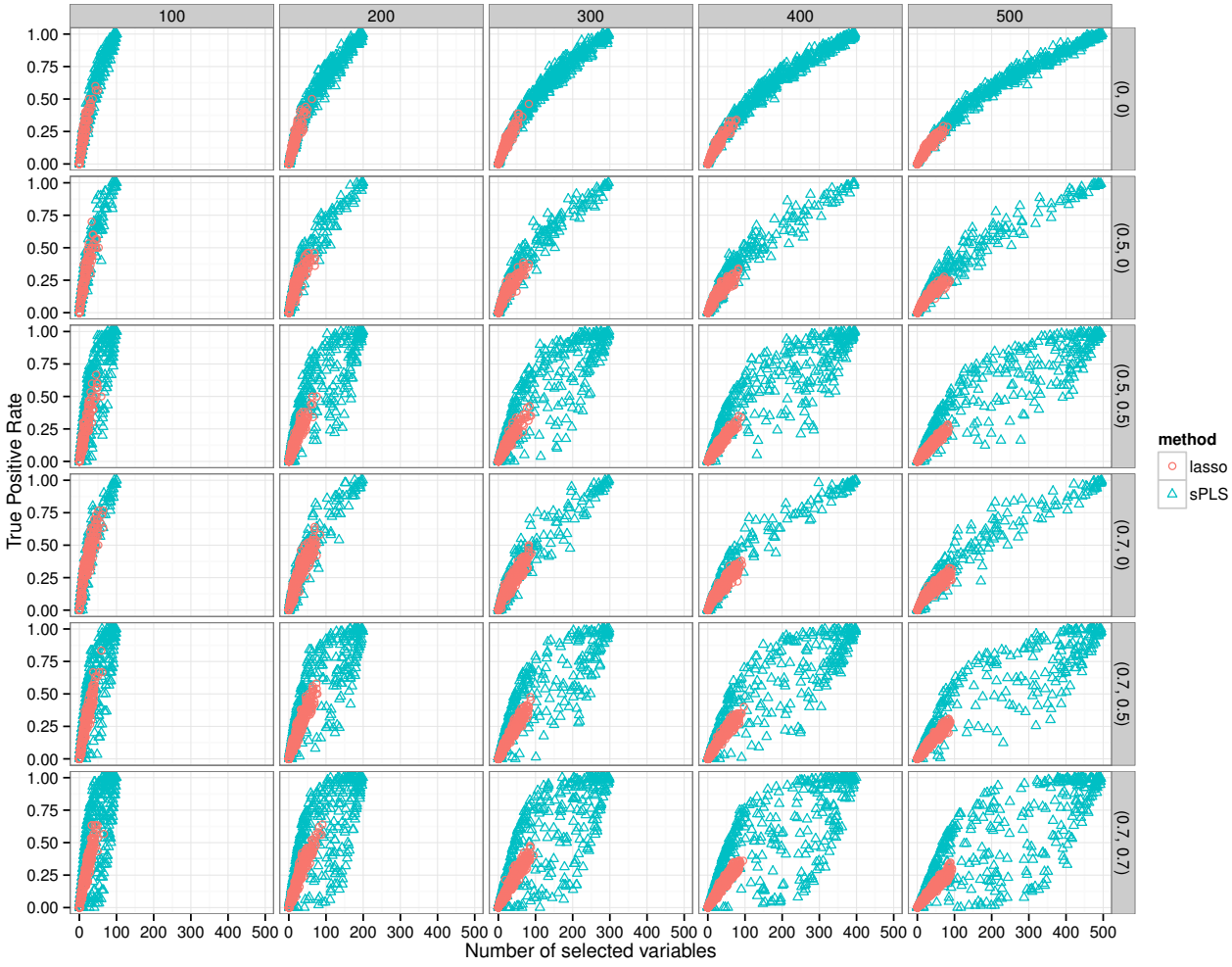


Figure A.3: Methylation sequencing data: Univariate analysis results (-log 10 pvalues) for a genomic region near the BLK gene on chromosome 8.

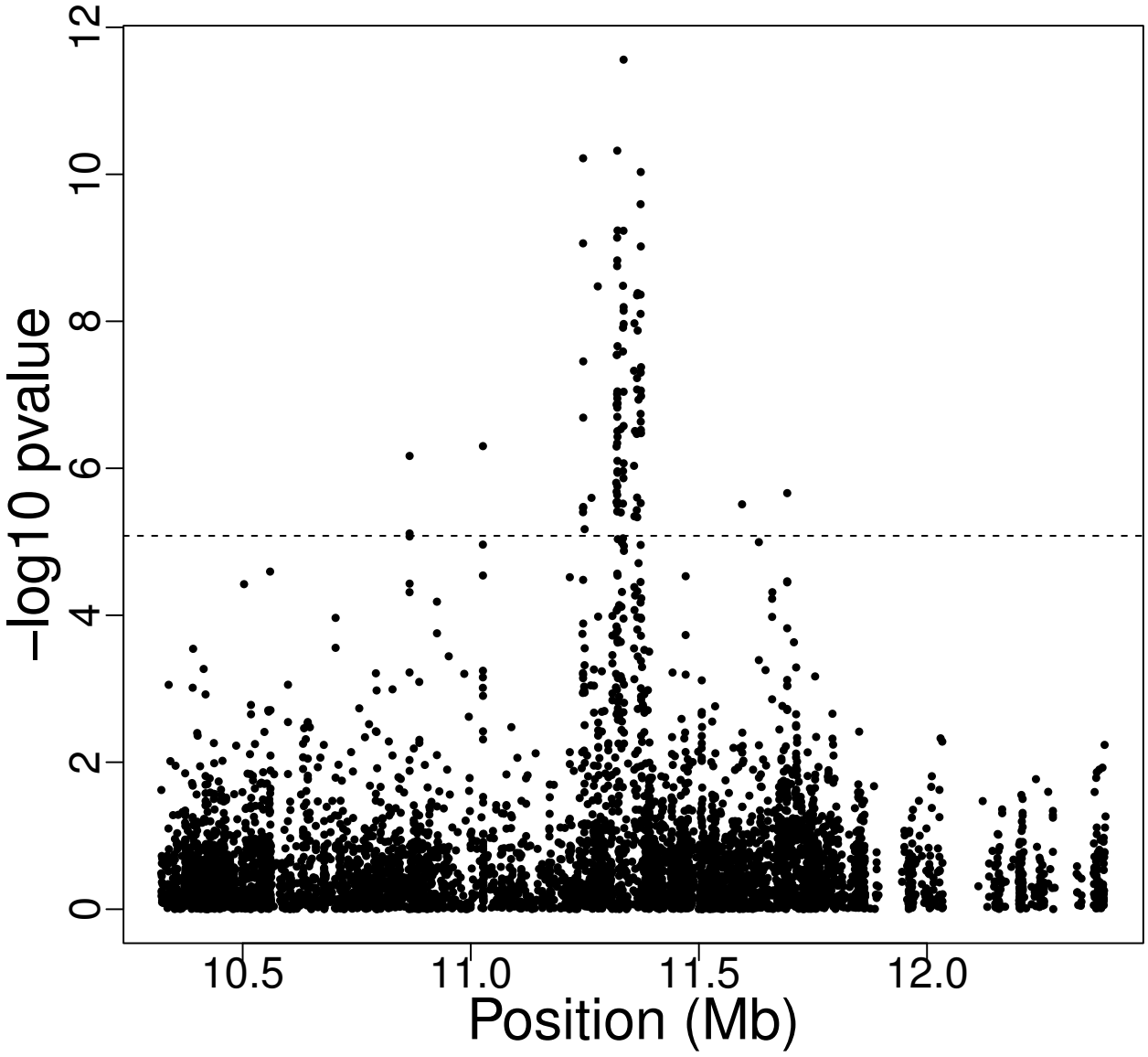


Figure A.4: Methylation sequencing data: relationship between VIP, univariate regression coefficients and univariate p-values

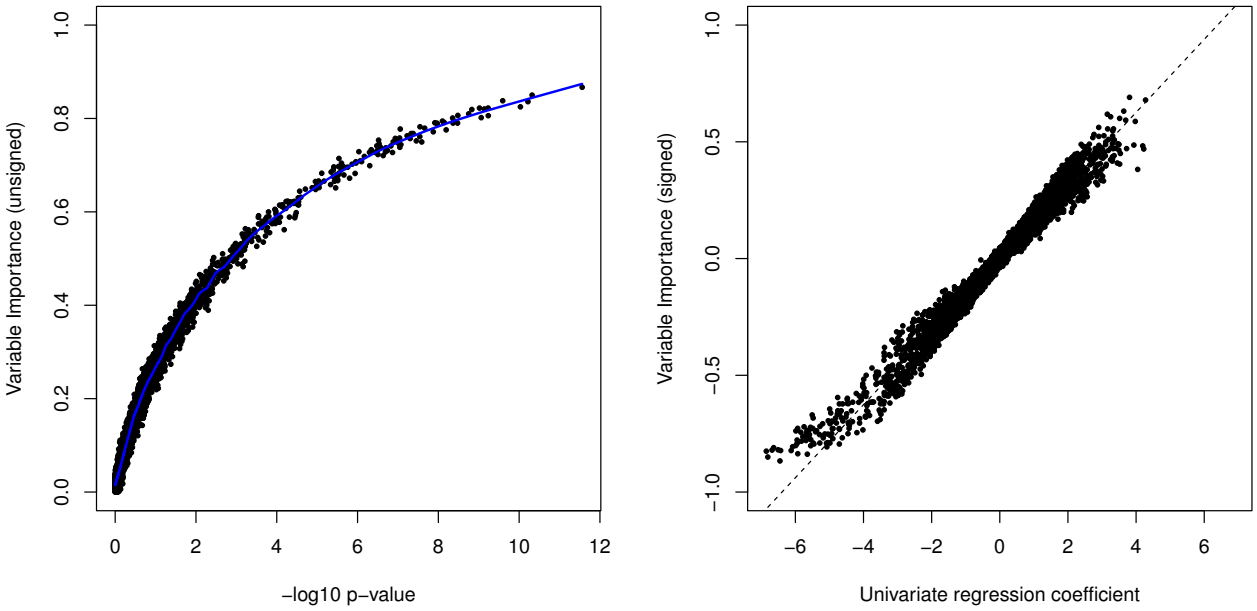
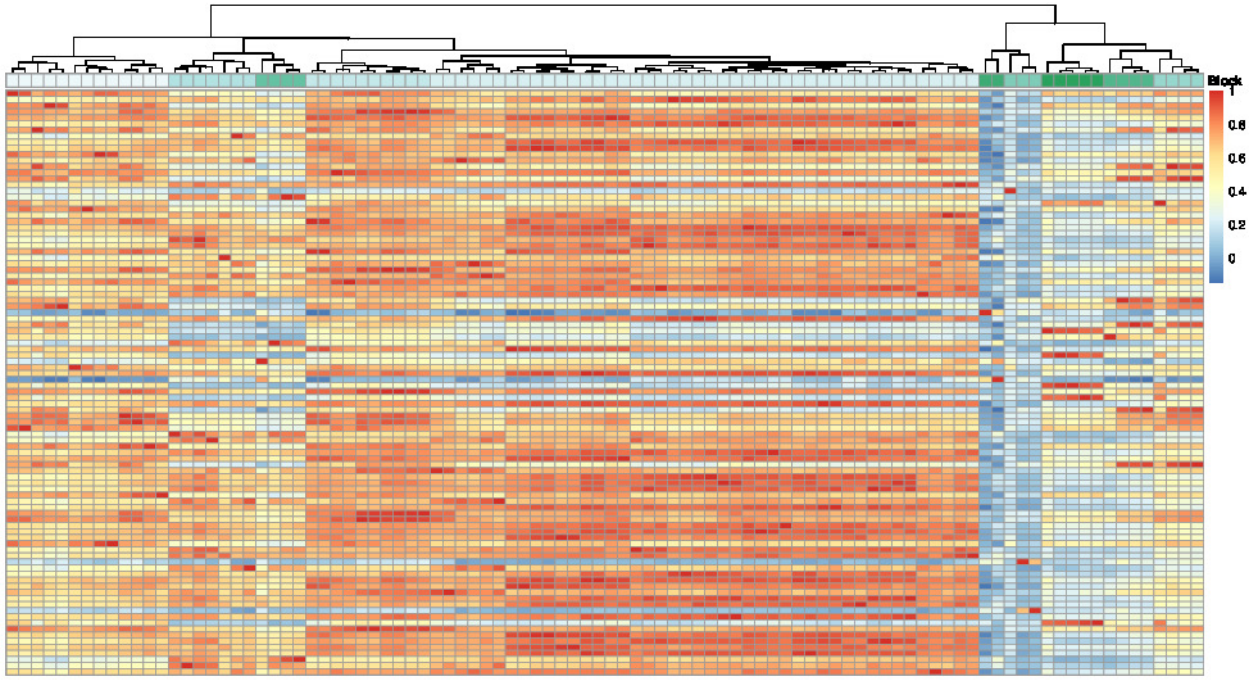


Figure A.5: ADNI study: correlation map of the 96 regions of the brain. The 10 blocks obtained by hierarchical clustering for the PCEV-block approach can be identified using the dendrogram at the top of the figure.



APPENDIX B

Appendix to Manuscript 2

B.1 Computation of the largest root in singular settings

In high-dimensional settings, both matrices $\mathbf{A}$ and $\mathbf{B}$ often have a singular distribution. If the estimate of $\mathbf{A}$ is not invertible, special care is required in order to compute the largest root. More precisely, we have to use a truncated variant of the eigenvalue decomposition (EVD) or singular value decomposition (SVD). This computational trick is well-known in the machine learning community [Udell et al., 2016], yet it does not seem to have received similar attention in the statistical literature. Therefore we review this procedure below.

Truncated EVD proceeds as follows: let $\hat{A}, \hat{B}$ be realizations of the estimators $\mathbf{A}, \mathbf{B}$, respectively. Let r be the rank of $\hat{A}$. From ordinary SVD, we know there exists an orthogonal matrix T of the same dimension as $\hat{A}$ such that

$$D := T^T \hat{A} T$$

is a diagonal matrix with exactly r nonzero values on the diagonal. Let $D_{[r]}$ be the diagonal matrix obtained from D by keeping only the nonzero elements, and let $T_{[r]}$ be the $n \times r$ matrix defined by keeping only the columns of T corresponding to the nonzero elements of

the diagonal of D. If we write $\tilde{T} = T_{[r]} D_{[r]}^{-1/2}$, we get

$$\tilde{T}^T \hat{A} \tilde{T} = I_r.$$

We can then use ordinary SVD a second time and diagonalize $\tilde{T}^T \hat{B} \tilde{T}$ using an orthogonal transformation T'. Therefore, if we let $S := \tilde{T} T'$, we get

$$S^T \hat{B} S = \Lambda$$
$$S^T \hat{A} S = I_r,$$

where Λ is a diagonal matrix. One can check that the largest diagonal element of Λ is the largest root of the determinantal equation in the main manuscript. We further note that, in the context of PCEV, the column of S corresponding to this largest root maximises the proportion of variance explained.

Truncated SVD proceeds similarly, *mutatis mutandis*.

B.2 Fitting procedures

As mentioned in the main manuscript, we look at four different fitting strategies to derive the estimates $\hat{\mu}$ and $\hat{\sigma}$ of our main algorithm:

1. Method of moments;
2. Maximum Likelihood Estimation;
3. Maximum Goodness-of-Fit estimation [Luceño, 2006] with the Anderson-Darling test statistic;
4. Maximum Goodness-of-Fit estimation with a modified Anderson-Darling statistic that gives more weight to the right tail.

As its name suggests, Maximum Goodness-of-Fit estimation seeks the values of the parameters that maximise a goodness-of-fit criterion.

Recall that the Anderson-Darling statistic is a defined as

$$A^2 = \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x),$$

where $F_n(x)$ is the empirical distribution function for a sample of size n and $F(x)$ is the CDF of target distribution. The modified Anderson-Darling statistic in 4. above gives more weight to the right tail of $F(x)$:

$$AR^2 = \int \frac{(F_n(x) - F(x))^2}{1 - F(x)} dF(x).$$

Luceño [2006] also provides computational formulas for these two statistics: let $X_1, \ldots, X_n$ be a sample of size n drawn from a distribution with CDF $F(x)$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics, and define $Z_i = F(X_{(i)})$ be the value of the CDF at the i-th order statistic. Using the fact that the empirical CDF is a step function, one can show that

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left(\log Z_i + \log(1 - Z_{n+1-i}) \right),$$
$$AR^2 = \frac{n}{2} - 2 \sum_{i=1}^{n} Z_i - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \log(1 - Z_{n+1-i}).$$

B.3 Further simulation results

B.3.1 Comparison to the true distribution

For the supplementary materials, we looked at the same simulation scenario as in the main manuscript, but we looked at more values for the parameters. We varied the dimension $p = 500, 1000, 1500, 2000$ and the number of roots used to estimate the distribution $K =$

$25, 50, 75, 100$. Finally, we looked at an exchangeable correlation structure with parameter $\rho = 0, 0.2, 0.5$. Figures B.1, B.2, and B.3 correspond to $\rho = 0, 0.2, 0.5$, respectively.

B.3.2 Comparison to the true distribution–Linear shrinkage estimator

As in the main manuscript, we repeated the above simulation scenario, but with the linear shrinkage estimator for the Wishart matrix A. Figures B.4, B.5, and B.6 correspond to $\rho = 0, 0.2, 0.5$, respectively.

B.3.3 Comparison of p-values

PCEV

We follow the same simulation scenario as in the main manuscript, with one exception: the residual covariance matrix Σ_R has an exchangeable structure with parameter $\rho = 0.5$. The results appear in Figure B.7, and they are consistent with those in the main manuscript.

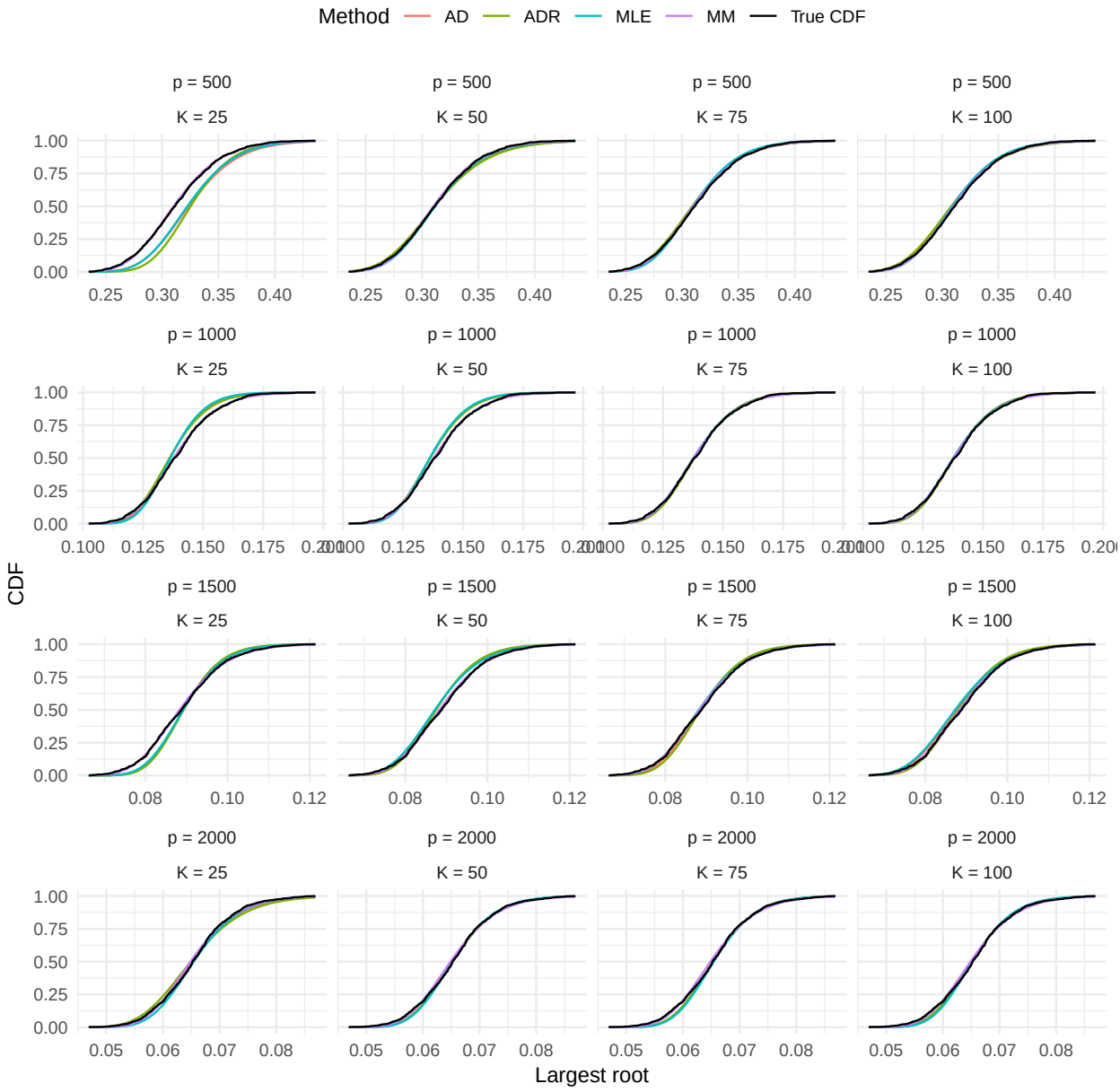


Figure B.1: **Approximation to the CDF:** Heuristic, using four different values for the number of permutations, compared to the true CDF. The covariance parameter is $\rho = 0$.

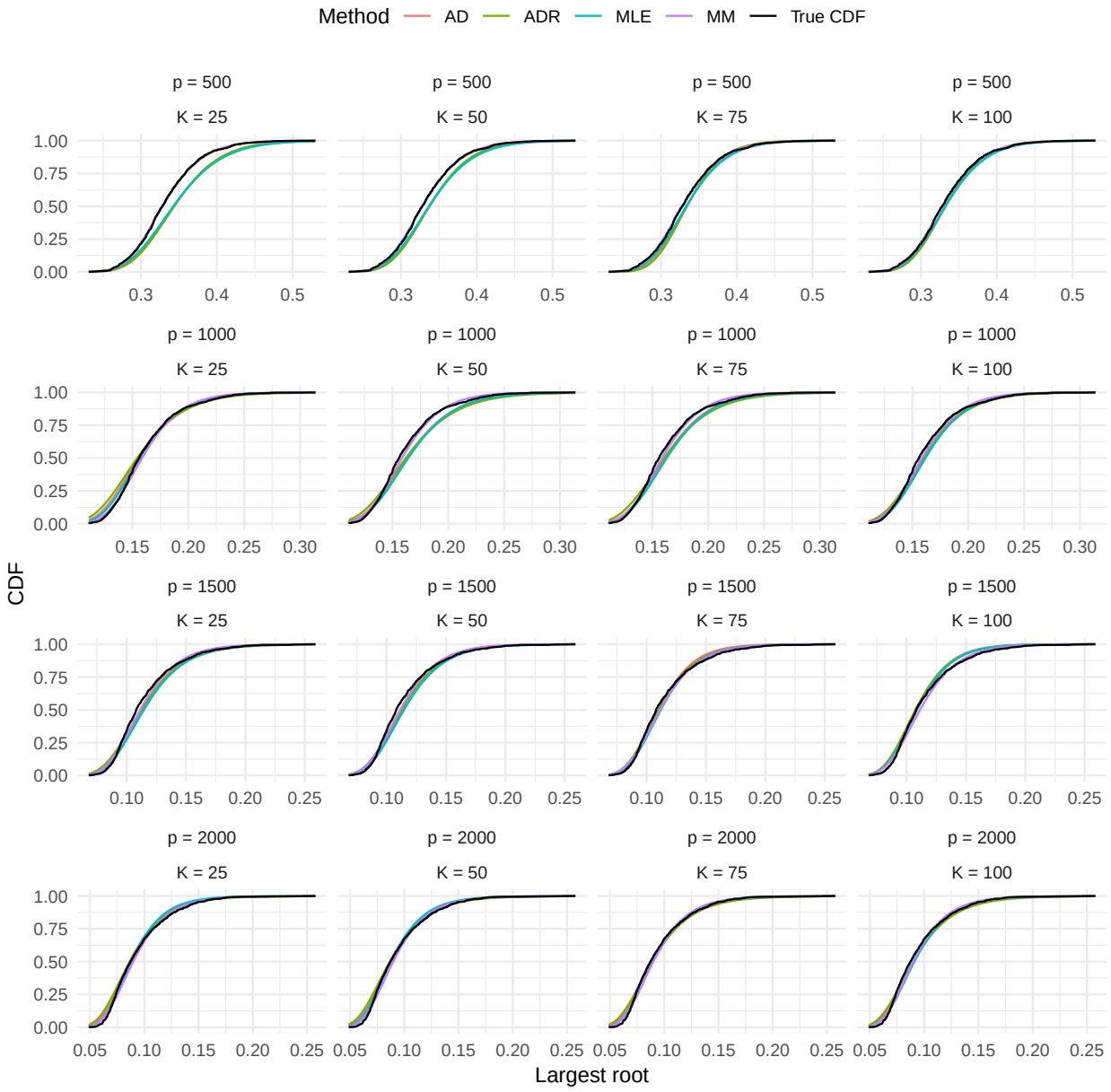


Figure B.2: **Approximation to the CDF:** Heuristic, using four different values for the number of permutations, compared to the true CDF. The covariance parameter is $\rho = 0.2$.

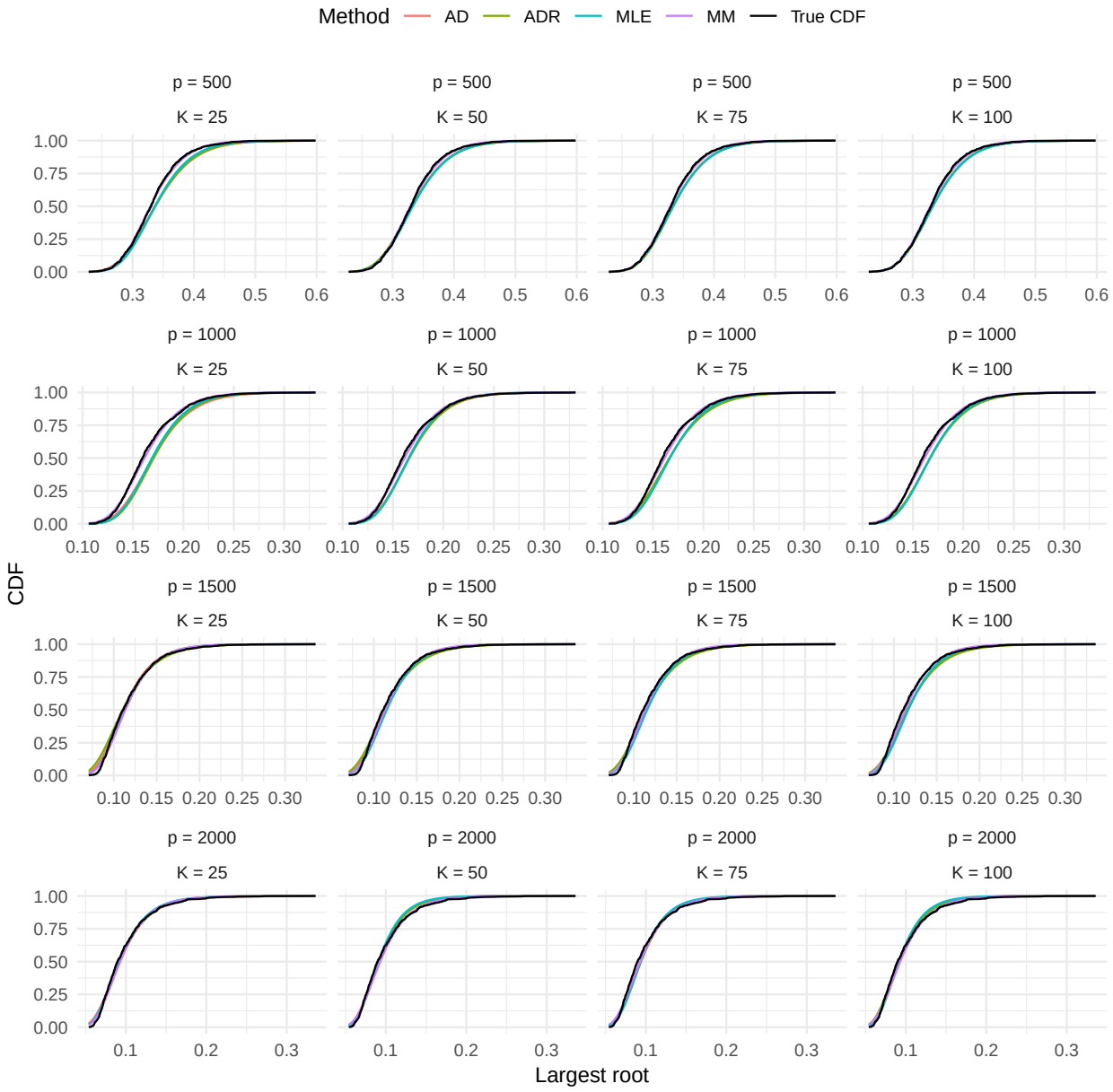


Figure B.3: **Approximation to the CDF:** Heuristic, using four different values for the number of permutations, compared to the true CDF. The covariance parameter is $\rho = 0.5$.

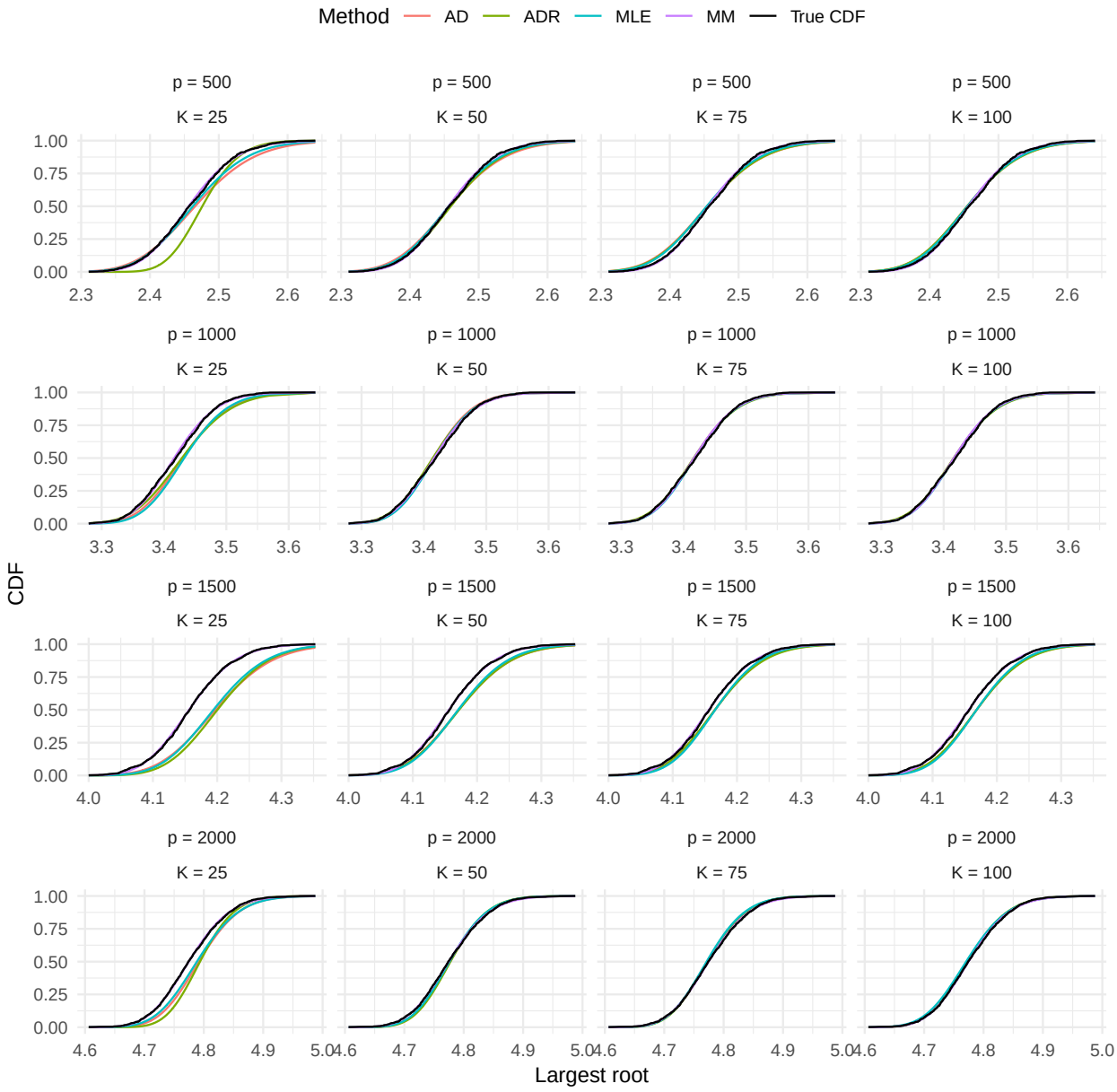


Figure B.4: **Approximation to the CDF:** Heuristic, using four different values for the number of permutations, compared to the true CDF. The covariance parameter is $\rho = 0$.

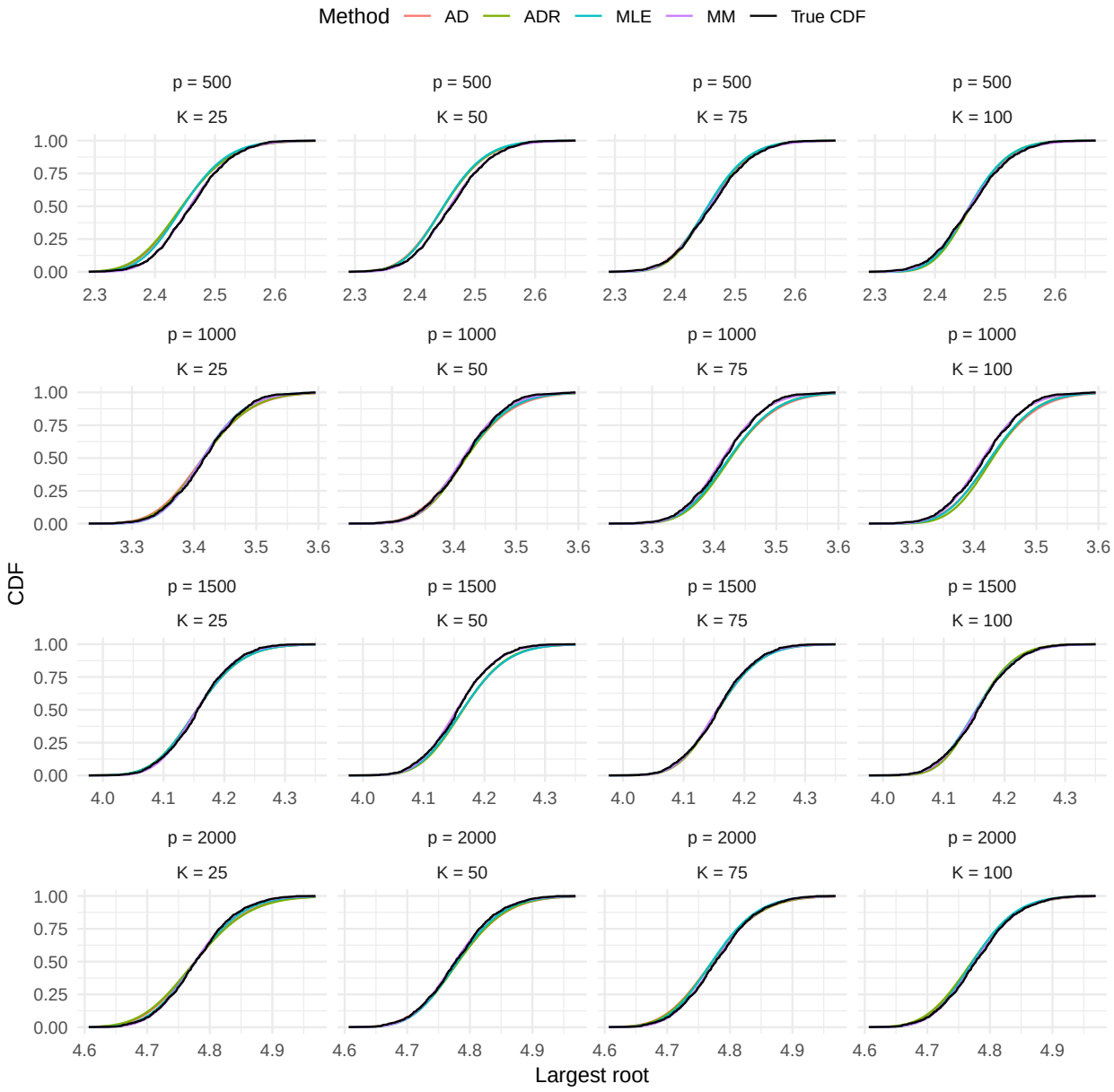


Figure B.5: **Approximation to the CDF:** Heuristic, using four different values for the number of permutations, compared to the true CDF. The covariance parameter is $\rho = 0.2$.

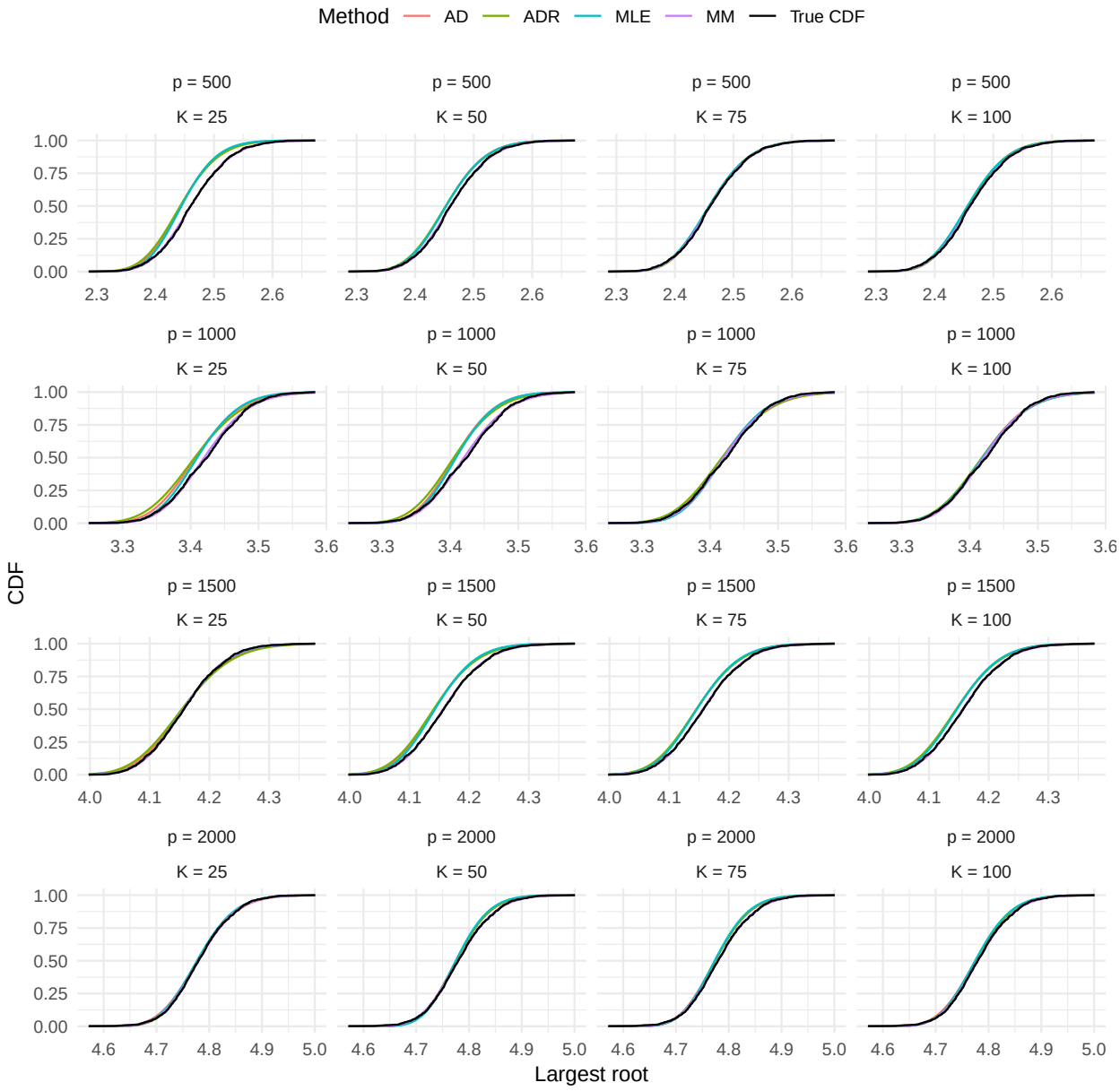


Figure B.6: **Approximation to the CDF:** Heuristic, using four different values for the number of permutations, compared to the true CDF. The covariance parameter is $\rho = 0.5$.

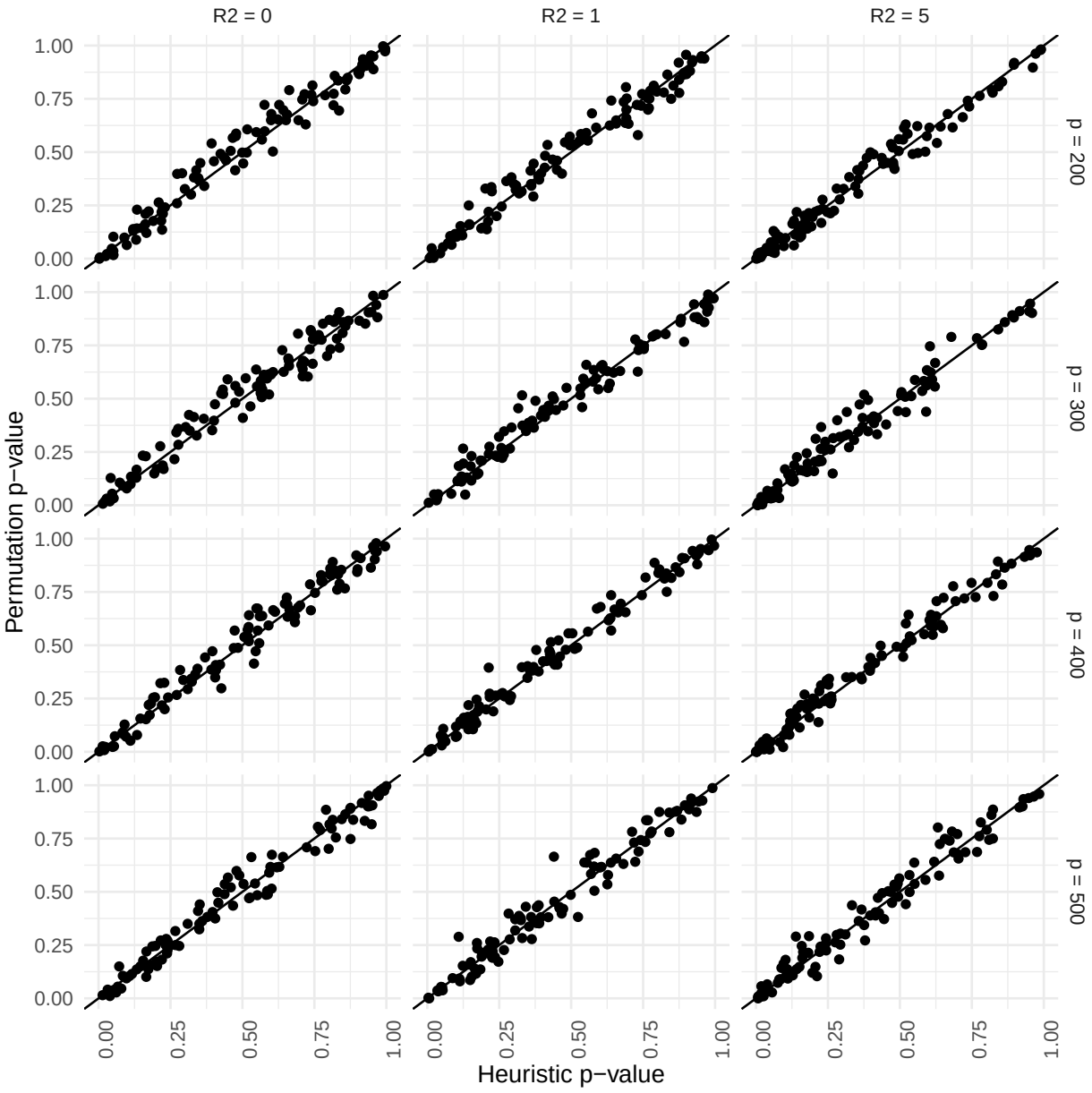


Figure B.7: **PCEV with $\rho = 0.5$:** QQ-plots comparing the p-values obtained from a permutation procedure to those obtained from the Tracy-Widom empirical estimator.

APPENDIX C

Appendix to Manuscript 3

C.1 Gene Ontology Analysis

To gain insight into the main results, we performed a Gene Ontology (GO) Analysis of the significant regions found by PCEV. First, we found all the genes overlapping the region we analysed. Then, for each gene, we retrieved the corresponding GO terms. Table C.1 shows the 20 most frequent GO terms among the significant regions.

The annotations were retrieved using the `biomaRt` package from *Bioconductor* [Durinck et al., 2005, 2009].

GO Identifier	GO Term
GO:0005515	Protein Binding
GO:0005829	Cytosol
GO:0005634	Nucleus
GO:0005886	Plasma Membrane
GO:0005737	Cytoplasm
GO:0005654	Nucleoplasm
GO:0000981	DNA-Binding Transcription Factor Activity, RNA Polymerase II-Specific
GO:0016021	Integral Component of Membrane
GO:0046872	Metal Ion Binding
GO:0070062	Extracellular Exosome
GO:0016020	Membrane
GO:0005887	Integral Component of Plasma Membrane
GO:0005524	ATP Binding
GO:0003677	DNA Binding
GO:0004674	Protein Serine/Threonine Kinase Activity
GO:0045944	Positive Regulation of Transcription by RNA Polymerase II
GO:0005509	Calcium Ion Binding
GO:0007165	Signal Transduction
GO:0000122	Negative Regulation of Transcription by RNA Polymerase II
GO:0003723	RNA Binding

Table C.1: Top 20 most frequent Gene Ontology (GO) terms appearing among the regions deemed significant by PCEV.

References

Herve Abdi. *Partial least square regression, projection on latent structure regression, PLS-Regression*. Wiley Interdisciplinary Reviews: Computational Statistics, 2010.

Kofi P Adragni and R Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

Yonathan Aflalo and Ron Kimmel. Regularized principal component analysis. *Chinese Annals of Mathematics, Series B*, 38(1):1–12, 2017.

Genevera I Allen, Christine Peterson, Marina Vannucci, and Mirjana Maletić-Savatić. Regularized partial least squares with an application to NMR spectroscopy. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(4):302–314, 2013.

Fiona Allum, Xiaojian Shao, Frédéric Guénard, Marie-Michelle Simon, Stephan Busche, Maxime Caron, John Lambourne, Julie Lessard, Karolina Tandre, Åsa K Hedman, et al. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nature communications*, 6:7211, 2015.

Theodore W Anderson. *An introduction to multivariate statistical analysis*. Wiley, 2003.

Rector Arya, John Blangero, Ken Williams, Laura Almasy, Thomas D Dyer, Robin J Leach, Peter O'Connell, Michael P Stern, and Ravindranath Duggirala. Factors of insulin resis-

tance syndrome–related phenotypes are linked to genetic locations on chromosomes 6 and 7 in nondiabetic mexican-americans. *Diabetes*, 51(3):841–847, 2002.

Stephen B Baylin. DNA methylation and gene silencing in cancer. *Nature clinical practice Oncology*, 2:S4–S11, 2005.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

Peter J Bickel, Elizaveta Levina, et al. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

Lutz P Breitling, Rongxi Yang, Bernhard Korn, Barbara Burwinkel, and Hermann Brenner. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *The American Journal of Human Genetics*, 88(4):450–457, 2011.

Paul Brennan and Alan J Silman. An investigation of gene-environment interaction in the etiology of rheumatoid arthritis. *American journal of epidemiology*, 140(5):453–460, 1994.

Guojun Bu. Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. *Nature Reviews Neuroscience*, 10(5):333–344, 2009.

Ronald W Butler and Robert L Paige. Exact distributional computations for roy's statistic and the largest eigenvalue of a wishart distribution. *Statistics and Computing*, 21(2): 147–157, 2011.

Loreto Carmona, Marita Cross, Ben Williams, and Marissa Lassere. Rheumatoid arthritis. *Best Practice & Research in Clinical Rheumatology*, 24(6):733–745, 2010.

Eduardo Castro, R Devon Hjelm, Sergey M Plis, Laurent Dinh, Jessica A Turner, and Vince D Calhoun. Deep independence network analysis of structural brain imaging: application to schizophrenia. *IEEE transactions on medical imaging*, 35(7):1729–1740, 2016.

Warren A Cheung, Xiaojian Shao, Andréanne Morin, Valérie Siroux, Tony Kwan, Bing Ge, Dylan Aïssi, Lu Chen, Louella Vasquez, Fiona Allum, et al. Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome biology*, 18(1):50, 2017.

Marco Chiani. Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy–Widom distribution. *Journal of Multivariate Analysis*, 129:69–81, 2014.

Marco Chiani. Distribution of the largest root of a matrix for roy's test in multivariate analysis of variance. *Journal of Multivariate Analysis*, 143:467–471, 2016.

Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

AG Constantine. Some non-central distribution problems in multivariate analysis. *The Annals of Mathematical Statistics*, 34(4):1270–1285, 1963.

R Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

R Dennis Cook, Bing Li, et al. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3):569–584, 2007.

R Dennis Cook, Liliana Forzani, et al. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501, 2008.

Michelle Cotterchio, Gail McKeown-Eyssen, Heather Sutherland, Giao Buchan, Melyssa Aronson, Alexandra M Easson, Jeannette Macey, Eric Holowaty, and Steven Gallinger. Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis Can*, 21(2):81–86, 2000.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Carlos Cruchaga, John SK Kauwe, Oscar Harari, Sheng Chih Jin, Yefei Cai, Celeste M Karch, Bruno A Benitez, Amanda T Jeng, Tara Skorupa, David Carrell, et al. GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron*, 78 (2):256–268, 2013.

NG De Bruijn. On some multiple integrals involving determinants. *J. Indian Math. Soc*, 19: 133–151, 1955.

Percy Deift. Universality for mathematical and physical systems. In Marta Sanz-Solé, Javier Soria, Juan Luis Varona, and Joan Verdera, editors, *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*, pages 125–152, 2007.

Ioana Dumitriu and Plamen Koev. Distributions of the extreme eigenvaluesof beta–jacobi random matrices. *SIAM Journal on Matrix Analysis and Applications*, 30(1):1–6, 2008.

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440, 2005.

Steffen Durinck, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4:1184–1191, 2009.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K Rakyan, John Attwood, Matthias Burger, John Burton, Tony V Cox, Rob Davies, Thomas A Down, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, 38(12):1378–1385, 2006.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Noureddine El Karoui. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and $p/n \to \infty$. *arXiv preprint math/0309355*, 2003.

Osman El-Maarri, Tim Becker, Judith Junen, Syed Saadi Manzoor, Amalia Diaz-Lacava, Rainer Schwaab, Thomas Wienker, and Johannes Oldenburg. Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. *Human genetics*, 122(5):505–514, 2007.

Aida Eslami, El Mostafa Qannari, Achim Kohler, and Stéphanie Bougeard. Analyses factorielles de données structurées en groupes d'individus. *Journal de la Société Française de Statistique*, 154(3):44–57, 2013.

Brian Everitt and Graham Dunn. *Applied Multivariate Data Analysis*. Edward Arnold. London, 1991.

Yixin Fang, Yang Feng, and Ming Yuan. Regularized principal components of heritability. *Computational Statistics*, 29(3-4):455–465, 2014a.

Yixin Fang, Yang Feng, and Ming Yuan. Regularized principal components of heritability. *Computational Statistics*, 29(3-4):455–465, 2014b.

Elia Formisano, Federico De Martino, and Giancarlo Valente. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magnetic resonance imaging*, 26(7):921–934, 2008.

Kristina Forslind, Monica Ahlmén, Kerstin Eberhardt, Ingiäld Hafström, and Björn Svensson. Prediction of radiological outcome in early rheumatoid arthritis in clinical practice: role of antibodies to citrullinated peptides (anti-CCP). *Annals of the rheumatic diseases*, 63(9):1090–1095, 2004.

Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia MT Greenwood, and Kasper D Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol*, 15(11):503, 2014.

Daniel Fraiman and Ricardo Fraiman. An ANOVA approach for statistical comparisons of brain networks. *Scientific reports*, 8(1):4746, 2018.

LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via. *Journal of Statistical Software*, 33(1):1–22, 2010.

J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

Xiaoyi Gao, Joshua Starmer, Eden R Martin, et al. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology*, 32(4):361, 2008.

Mary Grace Goll and Timothy H Bestor. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, 74:481–514, 2005.

Alexej Gossmann, Pascal Zille, Vince Calhoun, and Yu-Ping Wang. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. *IEEE Transactions on Medical Imaging*, 2018.

Rameshwar D Gupta and Donald St P Richards. Hypergeometric functions of scalar matrix argument are expressible in terms of classical hypergeometric functions. *SIAM journal on mathematical analysis*, 16(4):852–858, 1985.

Kasper D Hansen, Benjamin Langmead, Rafael A Irizarry, et al. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13 (10):R83, 2012.

Xiaoke Hao, Chanxiu Li, Jingwen Yan, Xiaohui Yao, Shannon L Risacher, Andrew J Saykin, Li Shen, Daoqiang Zhang, and Alzheimer's Disease Neuroimaging Initiative. Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. *Bioinformatics*, 33(14):i341–i349, 2017.

Clara Happ and Sonja Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, pages 1–11, 2018.

Wolfgang Härdle and Léopold Simar. Canonical correlation analysis. *Applied Multivariate Statistical Analysis*, pages 321–330, 2007.

Ahmad R Hariri, Emily M Drabant, and Daniel R Weinberger. Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biological psychiatry*, 59(10):888–897, 2006.

Stuart P Hastings and John Bryce Mcleod. A boundary value problem associated with the second painlevé transcendent and the korteweg-de vries equation. *Archive for Rational Mechanics and Analysis*, 73(1):31–51, 1980.

Inge S Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, pages 97–114, 1990.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Paul Horst. Relations amongm sets of measures. *Psychometrika*, 26(2):129–149, 1961.

Steve Horvath. DNA methylation age of human tissues and cell types. *Genome biology*, 14 (10):3156, 2013.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Eugene A Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

Marie Hudson, Sasha Bernatsky, Ines Colmegna, Maximilien Lora, Tomi Pastinen, Kathleen Klein Oros, and Celia MT Greenwood. Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis. *Epigenetics*, pages 1–8, 2017.

Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Neural Information Processing Systems*, volume 10, pages 273–279, 1998.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. Independent component analysis. In *Natural Image Statistics*, pages 151–175. Springer, 2009.

Kurt Johansson. Shape fluctuations and random matrices. *Communications in mathematical physics*, 209(2):437–476, 2000.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

Iain M Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Annals of statistics*, 36(6):2638, 2008.

Iain M Johnstone. Approximate null distribution of the largest root in multivariate analysis. *The annals of applied statistics*, 3(4):1616, 2009.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

Ian T Jolliffe. *Principal Component Analysis*. Springer, 2002.

Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics*, 12(3): 531–547, 2003.

M Chris Jones and Robin Sibson. What is projection pursuit? *Journal of the Royal Statistical Society. Series A (General)*, pages 1–37, 1987.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45 (D1):D353–D361, 2017.

Elizabeth W Karlson, Shun-Chiao Chang, Jing Cui, Lori B Chibnik, Patricia A Fraser, Immaculata De Vivo, and Karen H Costenbader. Gene–environment interaction between HLA-DRB1 shared epitope and heavy cigarette smoking in predicting incident rheumatoid arthritis. *Annals of the rheumatic diseases*, 69(01):54–60, 2010.

Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

Ferath Kherif, Jean-Baptiste Poline, Guillaume Flandin, Habib Benali, Olivier Simon, Stanislas Dehaene, and Keith J Worsley. Multivariate model specification for fMRI data. *Neuroimage*, 16(4):1068–1083, 2002.

Lambertus Klei, Diana Luca, B Devlin, and Kathryn Roeder. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic epidemiology*, 32(1):9–19, 2008.

Arno Klein, Jesper Andersson, Babak A Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E Christensen, D Louis Collins, James Gee, Pierre Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, 46(3):786–802, 2009.

Robert J Klose and Adrian P Bird. Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences*, 31(2):89–97, 2006.

Plamen Koev and Alan Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75(254):833–846, 2006.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

Ken WK Lee and Zdenka Pausova. Cigarette smoking and DNA methylation. *Frontiers in genetics*, 4, 2013.

P Legendre and L Legendre. *Numerical ecology*, volume 24. Elsevier, 2012.

Sue E Leurgans, Rana A Moyeed, and Bernard W Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 725–740, 1993.

Bing Li. *Sufficient Dimension Reduction*. Chapman and Hall/CRC, 2018.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Lexin Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

Ja-an Lin, Hongtu Zhu, Rebecca Knickmeyer, Martin Styner, John Gilmore, and Joseph G Ibrahim. Projection regression models for multivariate imaging phenotype. *Genetic epidemiology*, 36(6):631–641, 2012.

J-P Lindenmayer, Ruth Bernstein-Hyman, Sandra Grochowski, and Nigel Bark. Psychopathology of schizophrenia: initial validation of a 5-factor model. *Psychopathology*, 28(1):22–31, 1995.

Benoît Liquet, Pierre Lafaye de Micheaux, Boris P Hejblum, and Rodolphe Thiébaut. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1):35–42, 2016.

Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2):142, 2013.

G Livshits, A Roset, K Yakovenko, S Trofimov, and E Kobyliansky. Genetics of human body size and shape: body proportions and indices. *Annals of human biology*, 29(3):271–289, 2002.

Alberto Luceño. Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis*, 51(2):904–917, 2006.

Dante Mantini, L Marzetti, M Corbetta, GL Romani, and C Del Gratta. Multimodal integration of fmri and eeg data for high spatial and temporal resolution analysis of brain networks. *Brain topography*, 23(2):150–158, 2010.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

K Mardia, J Kent, and JM Bibby. *Multivariate analysis*. Academic Press, London, 1979.

Somaiya Mateen, Atif Zafar, Shagufta Moin, Abdul Qayyum Khan, and Swaleha Zubair. Understanding the role of cytokines in the pathogenesis of rheumatoid arthritis. *Clinica chimica acta*, 455:161–171, 2016.

Kevin McGregor, Sasha Bernatsky, Ines Colmegna, Marie Hudson, Tomi Pastinen, Aurélie Labbe, and Celia MT Greenwood. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome biology*, 17(1):84, 2016.

Corine Miceli-Richard, Shu-Fang Wang-Renault, Saida Boudaoud, Florence Busato, Céline Lallemand, Kevin Bethune, Rakiba Belkhir, Gaétane Nocturne, Xavier Mariette, and Jörg

Tost. Overlap between differentially methylated DNA regions in blood B lymphocytes and genetic at-risk loci in primary Sjögren's syndrome. *Ann Rheum Dis*, 0(0):1–8, 2015.

Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009a.

Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009b.

Prasad Naik and Chih-Ling Tsai. Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):763–771, 2000.

Anthony P Nicholas, Sanjoy K Bhattacharya, and Paul R Thompson. *Protein deimination in human health and disease*. Springer, 2017.

Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

Shizuhiko Nishisato. Optimization and data structure: Seven faces of dual scaling. *Annals of Operations Research*, 55(2):345–359, 1995.

Jürg Ott and Daniel Rabinowitz. A principal-components approach based on heritability for combining phenotype information. *Human heredity*, 49(2):106–111, 1999.

Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87, 2011.

Yeonhee Park, Zhihua Su, and Hongtu Zhu. Groupwise envelope models for imaging genetic analysis. *Biometrics*, 73(4):1243–1253, 2017.

Yongseok Park and Hao Wu. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453, 2016.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge university press, 2009.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

Maiju Pesonen, Jaakko Nevalainen, Steven Potter, Somnath Datta, and Susmita Datta. A combined PLS and negative binomial regression model for inferring association networks from next-generation sequencing count data. *IEEE/ACM transactions on computational biology and bioinformatics*, 2017.

Arturas Petronis. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, 465(7299):721, 2010.

Steven G Potkin, Guia Guffanti, Anita Lakatos, Jessica A Turner, Frithjof Kruggel, James H Fallon, Andrew J Saykin, Alessandro Orro, Sara Lupoli, Erika Salvi, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for alzheimer's disease. *PloS one*, 4(8):e6501, 2009.

Mohsen Pourahmadi. *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons, 2013.

James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

Karim Raza and Danielle M Gerlag. Preclinical inflammatory rheumatic diseases: an overview and relevant nomenclature. *Rheumatic Disease Clinics*, 40(4):569–580, 2014.

Alvin Rencher and William Christensen. *Methods of Multivariate Analysis*. John Wiley and Sons, 2012.

Daniel B Rowe and Raymond G Hoffmann. Multivariate statistical analysis in fMRI. *Engineering in Medicine and Biology Magazine, IEEE*, 25(2):60–64, 2006.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Andrew I Schein, Lawrence K Saul, and Lyle H Ungar. A generalized linear model for principal component analysis of binary data. In *Aistats*, volume 3, page 10, 2003.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Xiaojian Shao, Marie Hudson, Inés Colmegna, Celia MT Greenwood, Marvin Fritzler, Philip Awadalla, Tomi Pastinen, and Sasha Bernatsky. Rheumatoid arthritis relevant DNA methylation changes identified in ACPA-positive asymptomatic individuals using methylome capture sequencing. Submitted, 2018.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 720–727, 2003.

Muni S Srivastava. Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society*, 37(1):53–86, 2007.

Muni S Srivastava and Yasunori Fujikoshi. Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis*, 97(9):1927–1940, 2006.

Stefan J Teipel, Christine Born, Michael Ewers, Arun LW Bokde, Maximilian F Reiser, Hans-Jürgen Möller, and Harald Hampel. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *Neuroimage*, 38(1): 13–24, 2007.

Dan Tenenbaum. *KEGGREST: Client-side REST access to KEGG*, 2017. R package version 1.14.1.

Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of operational research*, 238(2):391–403, 2014.

Andrew E Teschendorff, Joanna Zhuang, and Martin Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 2011.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Hemant K Tiwari, Jill Barnholtz-Sloan, Nathan Wineinger, Miguel A Padilla, Laura K Vaughan, and David B Allison. Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Human heredity*, 66 (2):67–86, 2008.

Craig A Tracy and Harold Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

Craig A Tracy and Harold Widom. The distributions of random matrix theory and their applications. In *New trends in mathematical physics*, pages 753–765. Springer, 2009.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Maxime Turgeon, Celia MT Greenwood, and Aurélie Labbe. A Tracy-Widom empirical estimator for valid p-values with high-dimensional datasets. Submitted, 2018a.

Maxime Turgeon, Karim Oualkacha, Antonio Ciampi, Hanane Miftah, Golsa Dehghan, Brent W Zanke, Andréa L Benedet, Pedro Rosa-Neto, Celia MT Greenwood, Aurélie Labbe, et al. Principal component of explained variance: an efficient and optimal data

dimension reduction framework for association studies. *Statistical methods in medical research*, 27(5):1331–1350, 2018b.

Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

Arnold L Van Den Wollenberg. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219, 1977.

Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of econometrics*, 4(2):147–166, 1976.

Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era—concepts and misconceptions. *Nature reviews genetics*, 9(4):255, 2008.

Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

Wen-Ting Wang and Hsin-Cheng Huang. Regularized principal component analysis for spatial data. *Journal of Computational and Graphical Statistics*, 26(1):14–25, 2017.

Yuanjia Wang, Yixin Fang, and Man Jin. A ridge penalized principal-components approach based on heritability for high-dimensional data. *Human heredity*, 64(3):182–191, 2007a.

Yuanjia Wang, Yixin Fang, and Jin Man. A ridge penalized principal-components approach based on heritability for high-dimensional data. *Human Heredity*, 64:182–191, 2007b.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Herman Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.

Biao Yang, Jinmeng Cao, Tiantong Zhou, Li Dong, Ling Zou, and Jianbo Xiang. Exploration of neural activity under cognitive reappraisal using simultaneous eeg-fmri data and kernel canonical correlation analysis. *Computational and mathematical methods in medicine*, 2018, 2018.

Xiangrong Yin, Bing Li, and R Dennis Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99 (8):1733–1757, 2008.

Chang-En Yu, Howard Seltman, Elaine R Peskind, Nichole Galloway, Peter X Zhou, Elisabeth Rosenthal, Ellen M Wijsman, Debby W Tsuang, Bernie Devlin, and Gerard D Schellenberg. 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, 2007.

Brent W Zanke, Celia MT Greenwood, Jagadish Rangrej, Rafal Kustra, Albert Tenesa, Susan M Farrington, James Prendergast, Sylviane Olschwang, Theodore Chiang, Edgar Crowdy, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics*, 39(8):989–994, 2007.

Fang Fang Zhang, Roberto Cardarelli, Joan Carroll, Kimberly G Fulda, Manleen Kaur, Karina Gonzalez, Jamboor K Vishwanatha, Regina M Santella, and Alfredo Morabia. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics*, 6(5):623–629, 2011.

Feng Zhao, Lishan Qiao, Feng Shi, Pew-Thian Yap, and Dinggang Shen. Feature fusion via hierarchical supervised local CCA for diagnosis of autism spectrum disorder. *Brain imaging and behavior*, 11(4):1050–1060, 2017.

Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Com-

parison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644, 2014.

Li-Xing Zhu, Kai-Tai Fang, et al. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.