Decomposing the RV coefficient to identify genetic markers associated with changes in brain structure

by

JinCheol Choi

B.Sc., Kyungpook National University, 2013

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2018

Approval

Name:	**JinCheol Choi**
Degree:	**Master of Science (Statistics)**
Title:	***Decomposing the RV coefficient to identify genetic markers associated with changes in brain structure***
Examining Committee:	**Chair:** Dr. Boxin Tang Professor

Dr. Brad McNeney
Senior Supervisor
Associate Professor _____

Dr. Joan Hu
Supervisor
Professor _____

Dr. Jinko Graham
Internal Examiner
Professor
Statistics and Actuarial Science _____

Date Defended: 13 April 2018

Abstract

Alzheimer's disease (AD) is a chronic neurodegenerative disease that causes memory loss and decline in cognitive abilities; it is the sixth leading cause of death in the United States, affecting an estimated 5 million Americans and 747,000 Canadians. A recent study of AD pathogenesis (Szefer et al., 2017) used the RV coefficient to measure linear association between multiple genetic variants and multiple measurements of structural changes in the brain, using data from Alzheimer's Disease Neuroimaging Initiative (ANDI). The authors decomposed the RV coefficient into contributions from individual variants and displayed these contributions graphically. In this project, we investigate the properties of such a "contribution plot" in terms of an underlying linear model, and discuss estimation of the components of the plot when the correlation signal may be sparse. The contribution plot is applied to genomic and brain imaging data from the ADNI-1 study, and to data simulated under various scenarios.

Keywords: Alzheimer's disease; Alzheimer's Disease Neuroimaging Initiative; RV coefficient; Genetic association; Multivariate linear association

Dedication

To my beloved parents and older brother, who are always supportive of me with their unconditional and eternal love.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Brad McNeney, for his thoughtful guidance and academic support, without which this project would not have been able to be possibly done. At the same time, my thank goes to Dr. Jinko Graham for providing me with an interesting idea about the contribution plot and the ADNI-1 genomic and brain imaging data for this project. Without them, I could not have even initiated this project in the first place.

I extend my gratitude to Dr. Boxin Tang and Dr. Joan Hu for their generously agreeing to spend their valuable time to be examining committees for my defense. Also, I deeply thank to all staffs and faculty members in Department of Statistics and Actuarial Science for their passionate dedication and commitment to the department and students, expecially the professors whose course greatly broadened my insight and knowledge in statistics.

In addition, I cannot be more grateful to all my friends and academic sisters and brothers, especially Trevor Thomson, Yuping Yang, Khalif Halani, Tian Li, Charlie Zhou, Jiying Wen, Dilshani Induruwage, Lillian Lin, Michelle Thiessen, and Grace G. Hsu, for their being with me throughout my time at SFU, which has been the most memorable and academically productive chapter in my life.

Last but foremost, I want to deliver my deepest appreciation to my beloved family that was always there to make my life brighter everyday with their unconditional support and eternal love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder. As a type of dementia, it is a neurological dysfunction that is irreversible, neurodegenerative and progressive, causing memory loss and the decline of cognitive function. The disease is considered a complex disease driven by a combination of genetic and environmental factors, and it usually occurs in older people. The Alzheimer's Association reported that more than 5 million Americans may suffer from the disorder and AD is ranked as sixth as a cause of mortality in the United States of America (Alzheimer's Association, 2017).

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal, multi-site study that started in 2004 to understand the onset, progression, and etiology of AD. The ADNI objectives are: (1) the development of optimized and uniform standards for obtaining longitudinal magnetic resonance imaging (MRI) and positron emission tomography (PET) data on subjects who have AD and mild cognitive impairment (MCI) as well as cognitively normal (CN) elderly controls across multiple centers, (2) the development of methods to assess treatment effects in these subjects, (3) the establishment of accessible data repositories with diverse types of information including longitudinal changes in brain structure and metabolism, cognitive function, and biomarkers in these subjects, and (4) the acquirement of biological and pathogenic interpretation of MCI and AD through statistical analysis using the aforementioned data with genetic data as predictors (Weiner et al., 2013).

Association studies using the cohort data from ADNI have been conducted on a genome-wide scale to investigate genetic variants that are associated with AD. In 2011, two large genome-wide association studies (GWAS) that included ADNI data in their analyses reported candidate susceptibility genes (*MS4A6A/MS4A4E*, *EPHA1*, *CD33* and *CD2AP*) for AD using meta-analyses that systematically combined results from multiple studies (Hollingworth et al., 2011; Naj et al., 2011). In addition, a recent GWAS (Lee et al., 2017) found 6 genes associated with AD, including *PDS5B*, *ADARB2*, *BDH1*, *ST6Gal1*, *RAB20*, and *SPSB1*. The study stated that *PDS5B* and *ADARB2* are directly related to AD, and

identified two single nucleotide polymorphisms (SNPs) within these genes as possible influences of cognitive function. The other four genes, *ADARB2*, *BDH1*, *ST6Gal1*, *RAB20*, and *SPSB1*, are thought to be indirectly related to AD with significant SNPs linked to other risk factors relevant to AD like aging.

Many different correlation coefficients have been introduced to measure the association between two multivariate datasets. One of the most popular coefficients is the R Vector (RV) coefficient that measures the linear relationship between two data sets. It can be viewed as a unifying tool for linear multivariate statistical methods, since many major multivariate data analytic applications, such as multivariate regression, canonical correlation analysis, and principal component analysis, can be construed as the search for linear transformations of two original matrices that maximize the RV coefficient under certain constraints. It is an estimate of a population quantity called the vector correlation coefficient (Josse and Holmes, 2014).

Szefer et al. (2017) used the RV coefficient to summarize the relationships between SNPs in AD linkage regions and rates of change for 28 brain regions of interest. The analyses included a SNP selection phase and a validation phase. On the validation data they performed a test of the null hypothesis that the multivariate correlation coefficient, ρ_V (see equation 2.1), is equal to 0 vs the alternative hypothesis $\rho_V \neq 0$ and rejected the null hypothesis. Following the significance test, they decomposed the RV coefficient into contributions from each SNP and plotted the result to explore the relative contribution of each SNP to the RV coefficient. In this project, we further develop this "contribution plot" to evaluate the linear effect of each predictor variable to the overall linear dependence with multiple response variables, and to identify predictor variables that drive the multivariate linear association with response variables of interest in high-dimensional data.

An outline of the project is as follows. In Chapter 2, we define the RV coefficient and the contribution plot. In Chapter 3, we describe the ADNI-1 data and the results of our analysis. In Chapter 4, we summarize simulations to further explore the characteristics and behaviors of the contribution plot under a variety of circumstances. Finally, we conclude with summaries and discussions in Chapter 5. Calculation of the RV coefficient when subjects are unequally weighted (e.g., to account for biased sampling) are given in Appendix A. Supplementary descriptions of the RV coefficient including properties, problems, and hypothesis test, are encompassed in Appendix B along with our additional comments on the RV coefficient.

Chapter 2

Methods

In this chapter, we define the RV coefficient and its population counterpart, the multivariate correlation coefficient ρ_V. We also decompose ρ_V into contributions from each SNP marker, and study the form of such contributions under a multivariate linear model for brain phenotypes given genomic data. Finally, we discuss shrinkage estimation that may be useful when the correlation signal is sparse. By sparse we mean few non-zero pairwise correlations between genotypes and phenotypes.

2.1 The Multivariate Correlation and RV Coefficients

Our development follows Section 2 of Josse and Holmes (2016). Let $X = (X_1, \ldots, X_p)$ denote a random vector of p genotypes and $Y = (Y_1, \ldots, Y_q)$ denote a random vector of q phenotypes. Contrary to the convention in Statistics, we define these as row vectors. A measure of population correlation between X and Y is (Escoufier, 1973)

$$\rho_V(X,Y) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q}\mathrm{Cov}^2(X_k, Y_l)}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{Cov}^2(X_k, X_l)\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}^2(Y_k, Y_l)}} \tag{2.1}$$

where Cov() denotes population covariance. The coefficient ρ_V may be viewed as an extension of the squared population correlation to the multivariate setting. It can be shown that $0 \leq \rho_V \leq 1$. This and other properties of ρ_V are discussed in Appendix B.

Suppose we have n independent and identically distributed realizations of X and Y, arranged row-wise as column centred data matrices $\mathbf{X}(n \times p)$ and $\mathbf{Y}(n \times q)$, respectively. Let $X_{\cdot k}$ denote the kth column of $\mathbf{X}$; i.e., the vector of genotypes for marker k. Similarly, let $Y_{\cdot l}$ denote the lth column of $\mathbf{Y}$; i.e., the vector of measurements for phenotype l. The multivariate correlation coefficient in equation (2.1) can be estimated by the RV coeffi-

cient, obtained by replacing population covariances such as $\mathrm{Cov}(X_k, Y_l)$ by their sample counterparts $\mathrm{cov}(X_{\cdot k}, Y_{\cdot l})$:

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q}\mathrm{cov}^2(X_{\cdot k}, Y_{\cdot l})}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{cov}^2(X_{\cdot k}, X_{\cdot l})\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{cov}^2(Y_{\cdot k}, Y_{\cdot l})}} \tag{2.2}$$

Appendix A discusses alternative forms of the RV coefficient.

In addition to estimation of $\rho_V(X, Y)$, we may test the hypothesis $H_0 : \rho_V(X, Y) = 0$ *versus* $H_1 : \rho_V(X, Y) > 0$. A common approach to testing is to reject H_0 for large values of RV, with p-values obtained by comparing RV to an approximate permutation null distribution. Approximations to the permutation null distribution may be obtained analytically (Kazi-Aoual et al., 1995), or by Monte Carlo. In the case of Monte Carlo approximation, the estimated permutation null distribution is obtained by randomly permuting either the genotypes or the phenotypes of the n subjects and re-calculating RV for each permutation.

2.2 Contributions to the RV Coefficient

Szefer et al. (2017) found a significant association between the genotypes and phenotypes in their validation study. After the significant finding, they decomposed the RV coefficient into components due to each genetic marker and plotted these contributions *versus* marker location, to look for markers that might be driving the association. From equation (2.2), the contribution of the kth marker to the RV coefficient is proportional to

$$\hat{\mathscr{C}}_k = \sum_{l=1}^{q}\mathrm{cov}^2(X_{\cdot k}, Y_{\cdot l}). \tag{2.3}$$

The notation $\hat{\mathscr{C}}_k$ reflects the fact that the contribution of marker k to the RV coefficient is an estimate of a corresponding contribution to $\rho_V(\mathbf{X}, \mathbf{Y})$:

$$\mathscr{C}_k = \sum_{l=1}^{q}\mathrm{Cov}^2(X_k, Y_l). \tag{2.4}$$

The covariances that comprise $\mathscr{C}_k$ can be derived under a linear model for the association between X and Y. Such a model is consistent with the fact that the RV coefficient measures the linear relationship between two multidimensional data sets. Assume the multivariate

multiple regression model

$$Y = X\mathbf{B} + E \tag{2.5}$$

where $\mathbf{B}$ is a $p \times q$ matrix of regression parameters, and E is a row vector of q error terms assumed to be independent of X. The errors are assumed to follow a multivariate normal distribution, $MVN(0, \Sigma_{q\times q})$ where $\Sigma_{q\times q}$ is the covariance matrix. Component-wise, equation (2.5) is

$$Y_l = \sum_{k'=1}^{p} \beta_{k'l} X_{k'} + E_l \tag{2.6}$$

for $1 \leq l \leq q$.

Using equation (2.6), and the fact that $\mathrm{Cov}(X_k, E_l) = 0$ for all k and l, by independence of X and E, we can rewrite

$$\begin{aligned}
\mathrm{Cov}(X_k, Y_l) &= \mathrm{Cov}(X_k, \sum_{k'=1}^{p} \beta_{k'l} X_{k'} + E_l) \\
&= \sum_{k'=1}^{p} \beta_{k'l} \mathrm{Cov}(X_k, X_{k'}) + \mathrm{Cov}(X_k, E_l) \\
&= \sum_{k'=1}^{p} \beta_{k'l} \mathrm{Cov}(X_k, X_{k'}) \\
&= \beta_{kl} \mathrm{Var}(X_k) + \sum_{k'\neq k} \beta_{k'l} \mathrm{Cov}(X_k, X_{k'}),
\end{aligned} \tag{2.7}$$

where Var() denotes variance, and hence

$$\begin{aligned}
\mathscr{C}_k &= \sum_{l=1}^{q} \mathrm{Cov}^2(X_k, Y_l) \\
&= \sum_{l=1}^{q} \left\{ \beta_{kl} \mathrm{Var}(X_k) + \sum_{k'\neq k} \beta_{k'l} \mathrm{Cov}(X_k, X_{k'}) \right\}^2 .
\end{aligned} \tag{2.8}$$

Equation (2.8) shows that $\mathscr{C}_k$ depends on not only the regression coefficients, but also the variance of X_k and the covariances between X_k and the other components of X. Some simplification of the contributions is obtained by scaling each X_k by its standard deviation, so that the variance terms become one and covariances become correlations. In what follows we assume such scaling and use the notation SD() for population standard deviation, sd() for sample standard deviation, Cor() for population correlation and cor() for sample correlation. Dividing both sides of equation (2.6) by $\mathrm{SD}(Y_l)$, and defining $Y_l^* = Y_l/\mathrm{SD}(Y_l)$ and $X_k^* =$

$X_k/\mathrm{SD}(X_k)$ yields

$$Y_l^* = \sum_{k=1}^{p} \beta_{kl}^* X_k^* + E_l^* \tag{2.9}$$

where $E_l^* = E_l/SD(Y_l)$ and $\beta_{kl}^* = \beta_{kl} SD(X_k)/SD(Y_l)$, $k = 1, \ldots, p$, $l = 1, \ldots, q$, are the regression coefficients of the model for the standardized data. The interpretation of β_{kl}^* is the expected change in Y_l^* for a one unit change in X_k^* holding all other $X_{k'}^*$ fixed, or the expected SD change in Y_l for a one SD change in X_k holding all other $X_{k'}$ fixed. The contribution of marker k to $\rho_V(X^*, Y^*)$ is then

$$\mathscr{C}_k^* = \sum_{l=1}^{q} \left\{ \beta_{kl}^* + \sum_{k' \neq k} \beta_{k'l}^* \mathrm{Cor}(X_k^*, X_{k'}^*) \right\}^2, \tag{2.10}$$

which depends on the regression coefficients and the correlations between the genetic variants. Genetic variant k makes a non-zero contribution to $\rho_V(X^*, Y^*)$ if it is directly associated with one or more Y_l (i.e., $\beta_{kl} \neq 0$ for some ls) *or* if it is correlated with one or more $X_{k'}$'s that are directly associated with one or more Y_l (i.e., there is a k' such that $\mathrm{Cor}(X_k, X_{k'}) \neq 0$ and an l such that $\beta_{k'l} \neq 0$).

2.3 Estimation, Sparse Correlation and Sum of Powered Correlations

We now turn to estimation of the contributions to the RV coefficient. The contribution from the kth marker is

$$\hat{\mathscr{C}}_k^* = \sum_{l=1}^{q} \mathrm{cor}^2(X_{\cdot k}^*, Y_{\cdot l}^*), \tag{2.11}$$

a sum of squared sample correlations. Our studies of simulated data suggest that when the correlation signal is sparse, in the sense that there are few truly non-zero correlations, and the sample size is modest compared to the number of phenotypes, sampling error in the many estimates of truly *zero* correlations can obscure the signal of the few truly non-zero correlations. A solution is to raise the squared correlations to a power, α; i.e., we consider the contributions

$$\hat{\mathscr{C}}_k^*(\alpha) = \sum_{l=1}^{q} \mathrm{cor}^{2\alpha}(X_{\cdot k}^*, Y_{\cdot l}^*) \tag{2.12}$$

to a modified RV coefficient

$$RV(\mathbf{X}^*, \mathbf{Y}^*|\alpha) \propto \sum_{k=1}^{p}\sum_{l=1}^{q} \text{cor}^{2\alpha}(X^*_{.k}, Y^*_{.l}) \tag{2.13}$$

for $\alpha \geq 1$. Raising correlations to powers larger than 2 has the effect of differentially shrinking all estimates toward zero, with estimates near zero shrunken more than those near one. Independently, Xu et al. (2017) arrived at the same modified RV coefficient in the context of testing the null hypothesis $H_0 : \rho_V(X^*, Y^*) = 0$ *versus* the alternative hypothesis $H_1 : \rho_V(X^*, Y^*) > 0$. They suggest the sum of powered correlation (SPC) test, in which $RV(\mathbf{X}^*, \mathbf{Y}^*|\alpha)$ is employed as a test statistic and its significance is assessed with a Monte Carlo permutation test. Xu et al. (2017) also suggest an adaptive sum of powered correlation (aSPC) test, in which the test statistic is a minimum p-value for the SPC test over a grid of powers. Though testing is not the focus of this project, we make use of their minimum-p-value idea to select the power α. In particular, our contribution plot is of contributions $\hat{\mathscr{C}}^*_k(\alpha)$ for the power α that minimizes the p-value of the test based on $RV(\mathbf{X}^*, \mathbf{Y}^*|\alpha)$, for values of α on a grid. In our study we chose $\alpha = 1, 2, 3$ or 4.

2.3.1 Example

In this subsection, we present contribution plots using standardized ($\mathbf{X}^*$ and $\mathbf{Y}^*$) data sets. The vertical axis of the contribution plot is $\hat{\mathscr{C}}^*(\alpha)$ where α is either 1 or the optimal value that minimizes the p-value of the $RV(\mathbf{X}^*, \mathbf{Y}^*|\alpha)$ test based on 5,000 permutations. The horizontal axis represents SNPs of the ADNI-1 genomic data sorted by chromosome number and base-pair location. A multivariate multiple regression model, $\mathbf{Y}= \mathbf{XB}+\mathbf{E}$, is used to simulate a matrix of responses. The description of each component is as follows.

- $\mathbf{X}_{n\times p}$ is the matrix of the ADNI-1 genomic data on 493 SNPs (p) in 179 CN subjects (n).
- $\mathbf{Y}_{n\times q}$ is a matrix of simulated response variables.
- $\mathbf{E}_{n\times q}$ is an error matrix generated from $MVN(0, I_{q\times q})$.
- $\mathbf{B}_{p\times q}$ is a coefficient matrix.

The choice of an identity matrix for the covariance of the error terms is for simplicity. The number of response variables (q) is set to be 56 to generate neuroimaging data of 56 brain regions for simulation. We set $B_{30,1} = 1$ and $B_{70,10} = 1$, to designate the 30th and 70th SNPs as causal markers on the 1st and 10th brain regions, respectively, and all other $B_{i,j} = 0$.

The results are given in Figure 2.1. The contribution plots at $\alpha = 1$ and $\alpha = 3$ are displayed in the upper and lower panels, respectively. The two spikes above the horizontal line in the each plot are the contributions corresponding to the 30th and 70th SNPs. The horizontal line is the estimated 95th percentile of the distribution of the maximum contributions, where the maximum is over all markers across the region. The estimate is based on an empirical null distribution from 5,000 data sets in which the rows of $\mathbf{X}$ are permuted. Individual contributions that exceed the 95% threshold are considered noteworthy. When α increases from 1 to 2, noise from non-causal variables in the background is reduced.

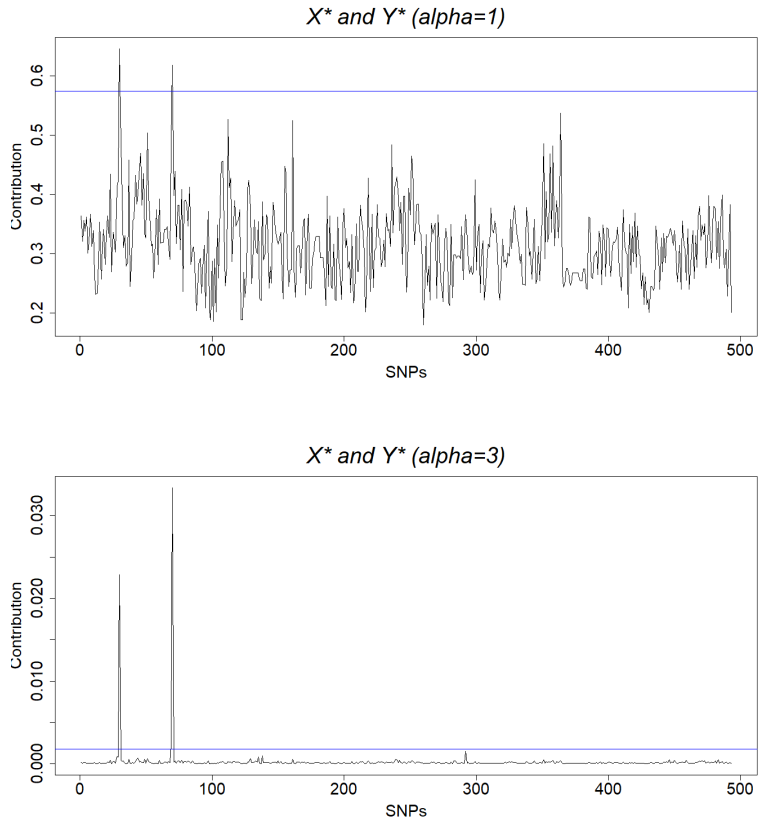

Figure 2.1: Example contribution plots of standardized genomic data of 493 SNPs and simulated neuroimaging data of 56 brain regions at $\alpha = 1$ (upper) and $\alpha = 3$ (lower). The horizontal line indicates the 95th percentile of the maximum contributions under the permutation null distribution.

Chapter 3

Application

In this chapter we apply the contribution plot to the ADNI-1 data set mentioned in the Introduction. We first describe how these data were obtained and what their features are.

3.1 Data Description

3.1.1 ADNI-1 Cohort Study

Both SNP and brain image data considered in this analysis were from the ADNI-1 study that was run from 2004 to 2009. One of the goals of the ADNI study is to identify biomarkers that predict AD. Our focus is on the 200 CN subjects collected in this study. The rationale for studying the CN subjects is that we are interested in genetic variation that predicts structural changes in the brain *before* subjects experience memory loss. Further details about the ADNI-1 study design is available on the ADNI website `http://adni.loni.usc.edu/study-design/`.

3.1.2 Genotype Data

Genotypes were measured as described in Saykin et al. (2010) and were subjected to quality control and imputation to fill in missing values as described in Szefer (2014). After data processing, 179 subjects with data on 493 SNPs in 33 genes remained for analysis. Table 3.1 gives a summary of gene names and the numbers of SNPs from each gene. SNP names are given in Appendix C.

3.1.3 Imaging Phenotype Data

The phenotypes were derived from baseline MRI scans taken for the ADNI-1 study for self-reported non-Hispanic white subjects. The MRI measurements were of volumes or cortical

Chromosome	Gene	No.	Chromosome	Gene	No.
1	CHRNB2	1	10	SORCS1	94
1	CR1	15	10	TFAM	6
1	ECE1	39	11	GAB2	19
1	MTHFR	10	11	PICALM	23
1	TF	3	11	SORL1	33
2	BIN1	12	15	ADAM10	19
2	IL1A	2	17	ACE	7
2	IL1B	1	17	GRN	1
6	NEDD9	69	17	THRA	3
6	PGBD1	6	17	TNK1	3
6	TNF	1	19	APOE	1
8	CLU	2	19	EXOC3L2	2
9	DAPK1	82	19	GAPDHS	3
9	IL33	14	19	LDLR	9
10	CALHM1	3	20	CST3	1
10	CH25H	1	20	PRNP	4
10	ENTPD7	4	Total		493

Table 3.1: Summary of the number of SNPs in analyzed genes.

thicknesses of 56 brain regions (Table 3.2), adjusted for covariates such as age, gender, education level, handedness and baseline intracranial volume (Wang et al., 2011).

Phenotype ID	Measurement	Cerebral region
AmygVol	Volume	Amygdala
CerebCtx	Volume	Cerebral cortex
CerebWM	Volume	Cerebral white matter
HippVol	Volume	Hippocampus
InfLatVent	Volume	Inferior lateral ventricle
LatVent	Volume	Lateral ventricle
EntCtx	Thickness	Entorhinal cortex
Fusiform	Thickness	Fusiform gyrus
InfParietal	Thickness	Inferior parietal gyrus
InfTemporal	Thickness	Inferior temporal gyrus
MidTemporal	Thickness	Niddle temporal gyrus
Parahipp	Thickness	Parahippocampal gyrus
PostCing	Thickness	Posterior cingulate
Postcentral	Thickness	Postcentral gyrus
Precentral	Thickness	Precentral gyurs
Precuneus	Thickness	Precuneus
SupFrontal	Thickness	Superior frontal gyrus
SupParietal	Thickness	Superior parietal gyurs
SupTemporal	Thickness	Superior temporal gyrus
Supramarg	Thickness	Supramarginal gyrus
TemporalPole	Thickness	Temporal pole
MeanCing	Mean thickness	Caudal anterior cingulate, isthmus cingulate, posterior cingulate, and rostral anterior cingulate
MeanFront	Mean thickness	Caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and frontal pole
MeanLatTemp	Mean thickness	Inferior temporal, middle temporal, and superior temporal gyri
MeanMedTemp	Mean thickness	Fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole
MeanPar	Mean thickness	Inferior and superior parietal gyri, supramarginal gyrus, and precuneus
MeanSensMotor	Mean thickness	Precentral and postcentral gyri
MeanTemp	Mean thickness	Inferior temporal, middle temporal, superior temporal, fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole

Table 3.2: Phenotype IDs and descriptions of 28 brain regions from a hemisphere, from Table 2.1 of Szefer (2014). Baseline structural MRI measurements of a total of 56 (= 28 × 2) regions from left and right hemispheres were estimated.

	SPC (α=1)	**SPC (α=2)**	**SPC (α=3)**	**SPC (α=4)**	**aSPC**
P-value	0.6834	0.3234	0.0624	0.0080	0.0154

Table 3.3: Summary of the p-values of SPCs with $\alpha = 1, 2, 3$, or 4, and the adaptive SPC test.

3.2 Contribution Plot for ADNI-1 Data

We initially standardize the genomic data of 493 SNPs and the neuroimaging data of 56 brain regions by subtracting column-wise means and dividing by column-wise SDs. The adaptive SPC test (Xu et al., 2017) of association between the genetic and phenotypic variables gives a p-value of 0.0154. The contribution plot may therefore be viewed as a *post hoc* investigation of the significant overall association. To select the power α for the contribution plot we calculate p-values for SPC tests with $\alpha = 1$, 2, 3 and 4 and find a minimum at $\alpha = 4$; see Table 3.3.

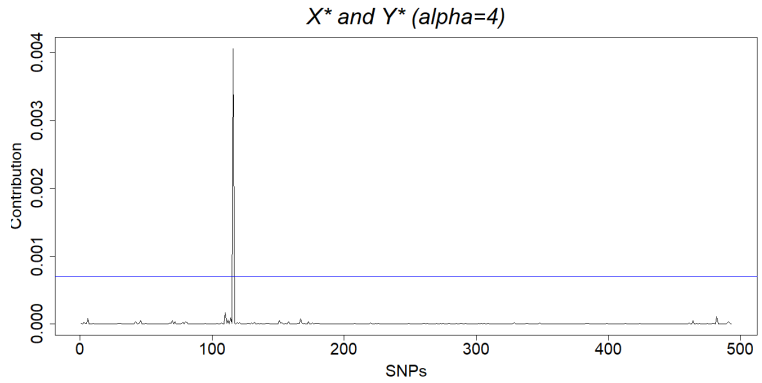

Figure 3.1: Contribution plot of standardized genomic data of 493 SNPs and 56 brain regions with $\alpha = 4$. The horizontal line indicates the 95th percentile of the maximum contributions under the permutation null distribution.

Figure 3.1 shows the contribution plot with $\alpha = 4$. SNPs on the x-axis are sorted by chromosome number and base-pair location. The spike above the permutation-based threshold is a strong signal of a linear association that comes from the SNP *rs16871157* within the *NEDD9* gene on chromosome 6.

We can further decompose the contribution of *rs16871157* by brain region. The results are shown in Figure 3.2 where the y-axis represents the individual sample correlation to the power of 8 between *rs16871157* and the 56 brain regions. Comparing the two panels of the figure, we can see that in general the correlations in the right hemisphere are stronger than those in the left hemisphere, but that the patterns of associations are very similar. Overall, it appears that *rs16871157* is associated with measures of cortical thickness, particularly in the temporal lobe of the brain (phenotype MeanTemp). The temporal lobe is involved in processing sensory input and memory.

Violin plots of the estimated MeanTemp and MeanLatTemp thickness by *rs16871157* genotypes are shown in Figure 3.3 for both the left and right hemisphere. In both hemi-

spheres the distribution of adjusted cortical thickness in CN subjects with the variant allele at *rs16871157* is shifted towards negative values compared to the distribution for CN subjects with two copies of the wild type allele, which is centred at zero. Thus, the presence of the variant allele at *rs16871157* is associated with reduced cortical thickness in CN subjects.

3.3 Discussion of *rs16871157* and *NEDD9*

rs16871157 is in an intron of the *NEDD9* gene and has no known function. Our analysis suggests that presence of the variant allele at *rs16871157* is associated with reduced cortical thickness in CN subjects. Reduced cortical thickness is associated with symptom severity in MCI and early AD patients, and has been observed in CN patients with amyloid binding (Dickerson et al., 2008).

NEDD9 stands for Neural Precursor Cell Expressed, Developmentally Down-Regulated 9. Much of the research to date on *NEDD9* has focussed on the association between variation in the gene and different cancers (e.g., Izumchenko et al., 2009), but, as the name suggests, the protein product of *NEDD9* is also involved in brain development. For example Vogel et al. (2009) found that the *NEDD9* protein plays a role in neuronal differentiation. In AD research, the SNP *rs760678* in *NEDD9* was found to be associated with late-onset AD (Wang et al., 2012). However, we note that the phenotypes associated with *rs760678* and *rs16871157* are quite different (late-onset AD *versus* baseline cortical thickness) and the two SNPs are in linkage equilibrium in Caucasian populations (estimated $R^2 < 0.01$ in Caucasian populations according to the online tool LDlink; Machiela and Chanock, 2015).

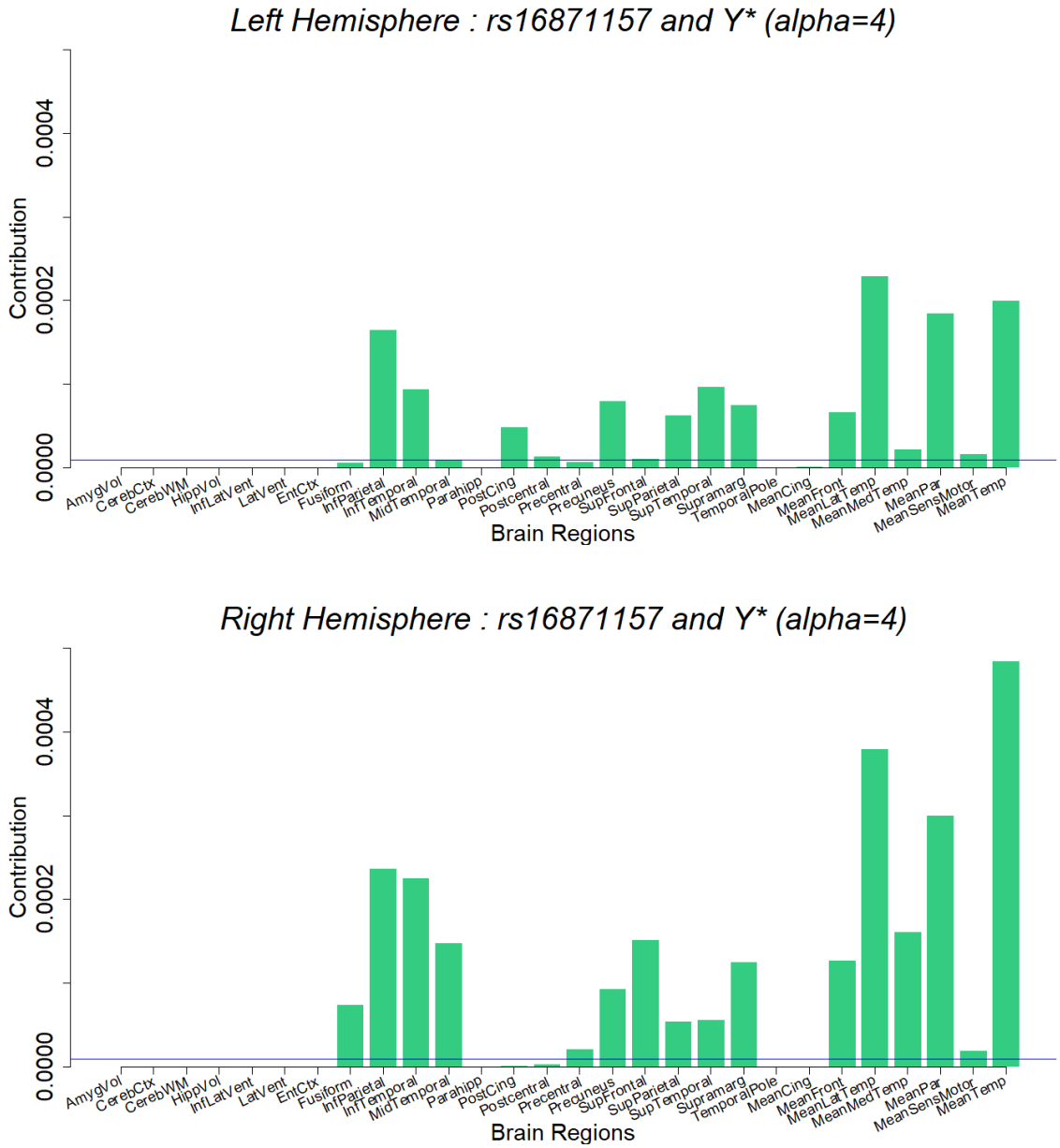

Figure 3.2: Contributions of *rs16871157* to brain regions in the left hemisphere (upper) and the right hemisphere (lower). The horizontal line indicates the 95th percentile of the maximum contributions under the permutation null distribution.

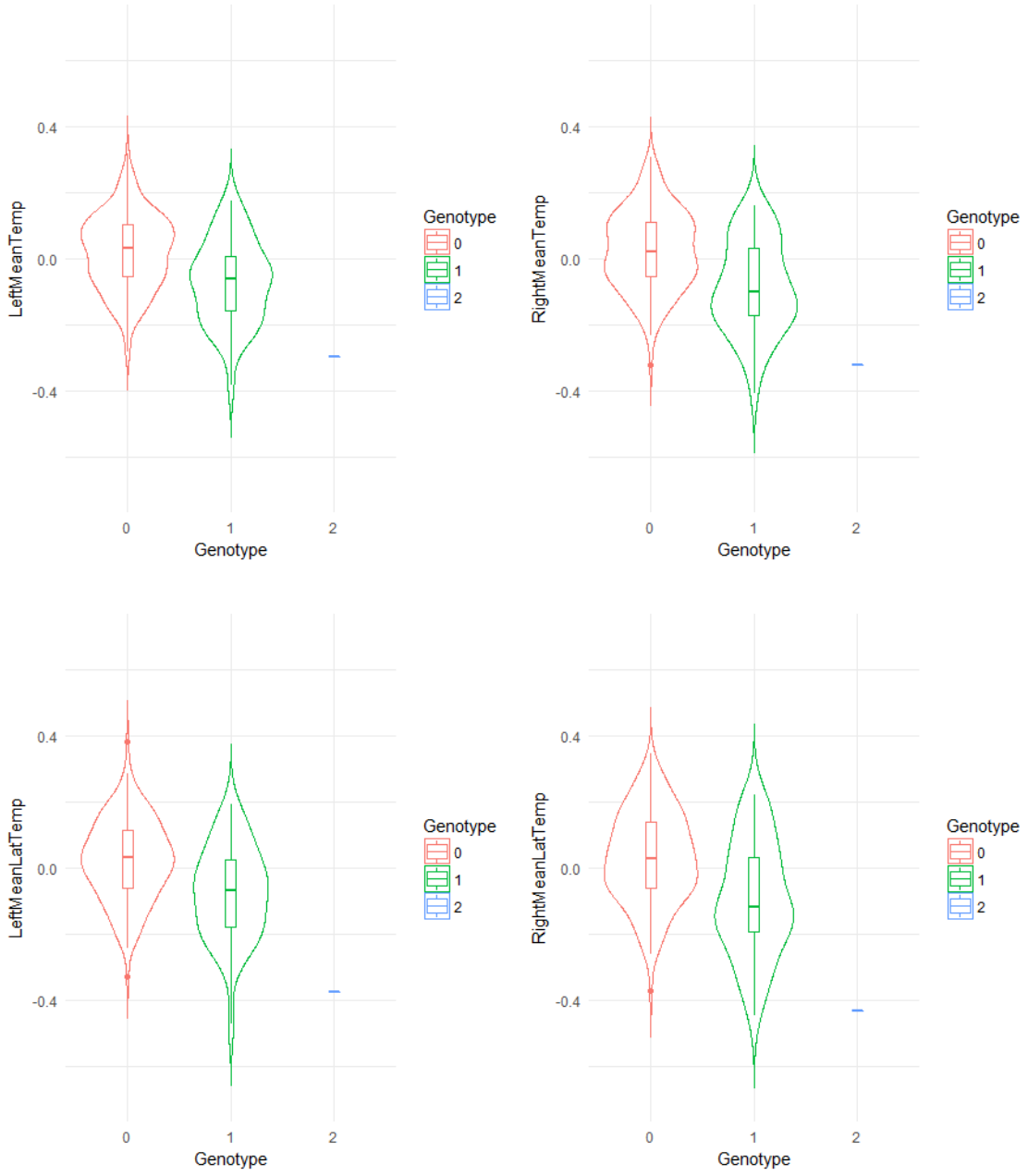

Figure 3.3: Violin plots of the distribution of the mean cortical thickness changes in MeanTemp (upper) and MeanLatTemp (lower) for each genotype of *rs16871157*. The left and right plots respectively represent the left and right hemispheres. The relative frequency of the minor allele in the CN subjects was 11.45%. Violin plots for genotype = 2 were not done because there is only one CN subject who is homozygous for the minor allele.

Chapter 4

Contribution Plots for Simulated Data Sets

In Chapter 2 we presented a simple example of the contribution plot using the ADNI-1 genetic data and simulated response variables. In this chapter, we display contribution plots for six additional simulated data sets. Our goal is to investigate the behaviour of the contribution plot under (i) different forms of dependence between explanatory variables and between response variables, and (ii) different numbers of explanatory and response variables.

4.1 Description of Simulation Configurations

This section describes the six different simulation settings. In all cases, data are simulated from the multivariate multiple regression model $\mathbf{Y}= \mathbf{XB}+\mathbf{E}$ in which:

- $\mathbf{X}_{n\times p}$ is a matrix of explanatory variables generated from $MVN(0, \Sigma_{p\times p})$ where $n = 100$ is the sample size,
- $\mathbf{Y}_{n\times q}$ is a matrix of response variables,
- $\mathbf{E}_{n\times q}$ is an error matrix generated from $MVN(0, \Sigma_{q\times q})$, and
- $\mathbf{B}_{p\times q}$ is a coefficient matrix.

The simulation parameters are p, q, $\Sigma_{p\times p}$, $\Sigma_{q\times q}$, and $\mathbf{B}$. We first give a brief overview of the different simulation settings, labelled setting 0, setting 1, ..., setting 5. In setting 0, data are simulated under the null hypothesis of no association between $\mathbf{X}$ and $\mathbf{Y}$; i.e., $B_{ij} = 0$ for all i and j. In settings 1 through 5 data are simulated under a sparse alternative, with $B_{30,1} = 1$, $B_{70,10} = 1$ and all other $B_{ij} = 0$. Simulation setting 1 specifies dependent explanatory

variables and setting 2 specifies dependent response variables. Simulation setting 3 specifies a large number of explanatory variables and setting 4 specifies a large number of response variables. Finally, setting 5 incorporates the most challenging features of settings 1 through 4: dependence among both explanatory and response variables, and large numbers of both explanatory and response variables. Further details about the simulation settings are as follows.

Setting 0 : No association.

- $n = 100$, $p = 130$, and $q = 25$
- $\Sigma_{p\times p} = I_{p\times p}$
- $\Sigma_{q\times q} = I_{q\times q}$
- $B = 0$

Setting 1 : The 25th to 35th X variables are correlated.

- $n = 100$, $p = 130$, and $q = 25$
- Letting $\Sigma_{p\times p}(i,j)$ be the $(i,j)^{th}$ entry of the matrix,

  $$\Sigma_{p\times p}(i,j)=0.9 \qquad (i \neq j,\ 25 \leq i,\ j \leq 35)$$

  All diagonal entries of $\Sigma_{p\times p}$ are 1, and the other entries are 0.
- $\Sigma_{q\times q} = I_{q\times q}$
- $B(30,1) = B(70,10) = 1$

Setting 2 : The 1st to 15th Y variables are correlated.

- $n = 100$, $p = 130$, and $q = 25$
- $\Sigma_{p\times p} = I_{p\times p}$
- $\Sigma_{q\times q}(i,j)=0.9 \qquad (i \neq j,\ 1 \leq i,\ j \leq 15)$

  All diagonal entries of $\Sigma_{q\times q}$ are 1, and the other entries are 0.
- $B(30,1) = B(70,10) = 1$

Setting 3 : There are 20 times more X variables.

- $n = 100$, $p = 2600$, and $q = 25$

- $\Sigma_{p\times p} = I_{p\times p}$
- $\Sigma_{q\times q} = I_{q\times q}$
- $B(30, 1) = B(70, 10) = 1$

Setting 4 : There are 20 times more Y variables.

- $n = 100$, $p = 130$, and $q = 500$
- $\Sigma_{p\times p} = I_{p\times p}$
- $\Sigma_{q\times q} = I_{q\times q}$
- $B(30, 1) = B(70, 10) = 1$

Setting 5 : All settings from 1 to 4 are adopted.

- $n = 100$, $p = 2600$, and $q = 500$
- $\Sigma_{p\times p}(i, j)$=0.9 $\quad (i \neq j,\ 25 \leq i,\ j \leq 35)$

  The other diagonal entries of $\Sigma_{p\times p}$ are 1, and the other entries are 0.
- $\Sigma_{q\times q}(i, j)$=0.9 $\quad (i \neq j,\ 1 \leq i,\ j \leq 15)$

  The other diagonal entries of $\Sigma_{q\times q}$ are 1, and the other entries are 0.
- $B(30, 1) = B(70, 10) = 1$

4.2 Analyses of Simulated Data Sets

Throughout, we standardize the explanatory variables and response variables and use $\mathbf{X}^*$ and $\mathbf{Y}^*$ to denote the standardized data matrices. The aSPC test is applied to each standardized data set and the p-value is reported. Contribution plots are of the contributions $\hat{\mathscr{C}}_k^*(\alpha) = \sum_{l=1}^{q} \text{cor}^{2\alpha}(X_{\cdot k}^*, Y_{\cdot l}^*)$ for explanatory variables $k = 1, \ldots, p$. Results are presented in figures with two panels; the top panel is for $\alpha = 1$ (i.e., contributions to the standard RV coefficient) and the bottom panel is for the α that minimizes the p-value of the test based on $RV(\mathbf{X}^*, \mathbf{Y}^*|\alpha)$ (see Chapter 2 for details). For settings 3 and 5 where a large number of explanatory variables are generated, plots are zoomed in a neighborhood of the causal variables to make it easier to see the correlation signal. A significance threshold is added to each plot to indicate the 95th percentile of the estimated distribution of maximum contributions, where the maximum is over all markers across the region. The estimate is based on an empirical null distribution from 5,000 data sets in which the rows of X are permuted. Individual contributions that exceed the 95% threshold are considered noteworthy.

4.3 Results

Setting 0 : None of X variables are associated with Y variables.

The p-value for the aSPC test on this simulated data set is 0.5055, correctly suggesting no association. Figure 4.1 displays the contribution plots. The significance threshold for the top panel is 0.5498 and the threshold for the bottom panel is 0.0059; both are outside the range of the vertical axes on the plots. In both panels there are no contributions that meet or exceed the significance thresholds. Thus, all contributions are considered true-negatives.

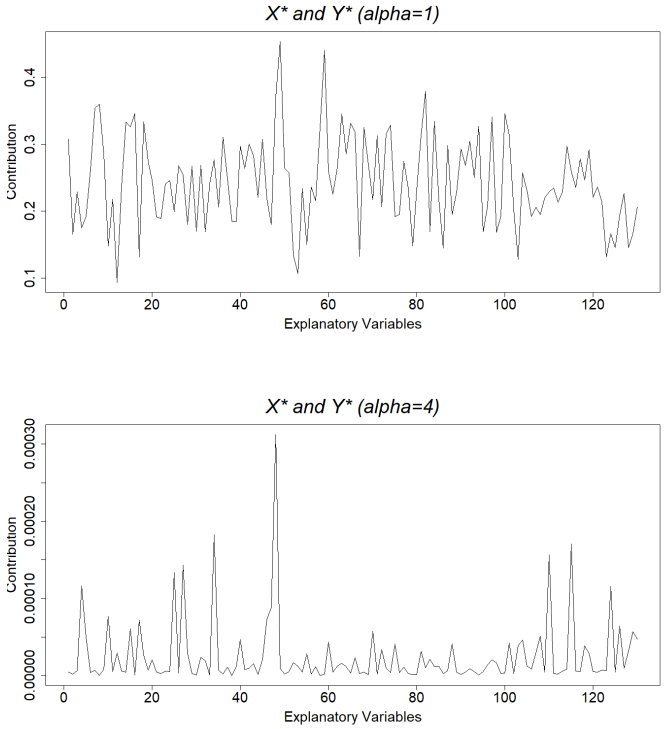

Figure 4.1: Simulation results of Setting 0 (B=0, p=130, q=25, $\Sigma_{p\times p}$=$I_{p\times p}$, $\Sigma_{q\times q}$=$I_{q\times q}$).

Setting 1 : The 25th to 35th X variables are correlated.

The p-value for the aSPC test on this simulated data set is 0.0006, reflecting the true association between the 30th explanatory variable, X_{30}, and the first response variable, Y_1, and between the 70th explanatory variable, X_{70}, and the 10th response variable, Y_{10}. The contribution plots are shown in Figure 4.2. The broad peak of signal toward the left end of the horizontal axes of the plots reflects the truly-associated X_{30}. In addition to a signal at X_{30}, other explanatory variables that are correlated with X_{30} have comparably-sized contributions, as predicted by equation (2.10). In particular, from equation (2.10),

$$\mathscr{C}_i^* = \sum_{l=1}^{25} \left\{ \beta_{il}^* + \sum_{k' \neq i} \beta_{k'l}^* \mathrm{Cor}(X_i^*, X_{k'}^*) \right\}^2 = \{\mathrm{Cor}(X_i^*, X_{30}^*)\}^2,$$

because $\beta_{il}^* = 0$ for $l = 1, \ldots, 25$ and $\beta_{k'l}^* = 0$ except when $k' = 30$ and $l = 1$, in which case $\beta_{30,1}^* = 1$. The contributions of X_i that are correlated with X_{30} should be roughly proportional to the squared correlation between X_i and X_{30} when $\alpha = 1$. Indeed, $\hat{\mathscr{C}}_{30}^*(\alpha = 1) = 0.7517$, $\hat{\mathscr{C}}_{28}^*(\alpha = 1) = 0.5619$, and $\mathrm{cor}(X_{\cdot 30}, X_{\cdot 28}) = 0.8837$, so that $0.5619 \approx 0.5870 (= 0.7517 \times (0.8837)^2)$. The narrow peak near the middle of the horizontal axes reflects the truly-associated X_{70}, which is not correlated with any of the other explanatory variables. There are two take-away messages here: (i) The contribution plots can identify the true signals, and (ii) correlation between explanatory variables can widen the peak signal.

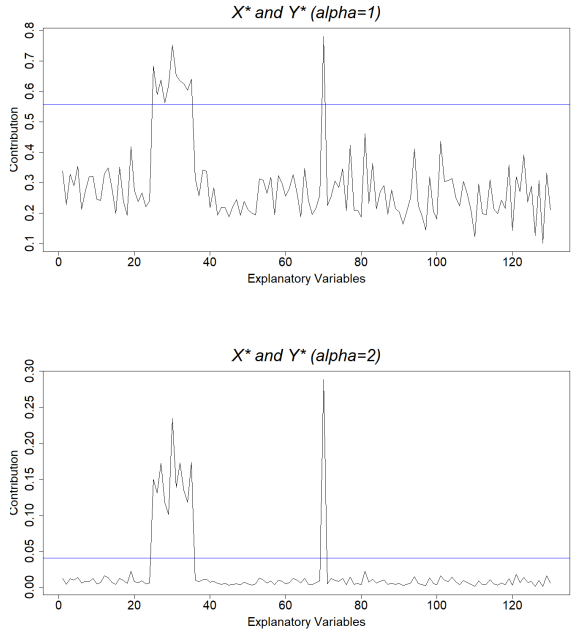

Figure 4.2: Simulation results of Setting 1 ($B_{30,1}$=$B_{70,10}$=1, p=130, q=25, $\Sigma_{p\times p} \neq I_{p\times p}$, $\Sigma_{q\times q}$=$I_{q\times q}$). The horizontal line indicates the 95th percentile of the maximum contributions under the permutation null distribution.

Setting 2 : The 1st to 15th Y variables are correlated.

The p-value for the aSPC test on this simulated data set is 0.0008, reflecting the true association between X_{30} and Y_1 and between X_{70} and Y_{10}. The contribution plots are shown in Figure 4.3. For contributions to the RV coefficient ($\alpha = 1$) the significance threshold is 1.7055. In the top panel we see that none of the contributions to the RV coefficient exceed this threshold. The increased threshold in setting 2 compared to setting 1 is a consequence of the increased variance in the contributions $\hat{\mathscr{C}}_k^*(\alpha) = \sum_{l=1}^{q} \mathrm{cor}^{2\alpha}(X_{.k}^*, Y_{.l}^*)$ resulting from positive dependence between response variables. In the top panel, the peak signal is at X_{100}, which is not truly associated with any of the response variables. By contrast, in the bottom panel the contributions of the two truly-associated variables do exceed the threshold.

The top panel in Figure 4.4 breaks down the signal at X_{100} into its squared sample-correlation components, $\mathrm{cor}^2(X_{.100}^*, Y_{.l}^*)$. X_{100} appears to be modestly associated with the first 15 (correlated) Y_i's, even though the true population correlations between X_{100} and these Y_i's are zero. What we have is essentially one modest sample correlation between X_{100} and the first-15 Y variables repeated by chance due to the population correlation in the error terms for the first-15 Y variables. The accumulation of these modest sample correlations with X_{100} across the first-15 response variables leads to the relatively large contribution for X_{100} in the top panel of Figure 4.3. The bottom panel of Figure 4.4 shows the squared sample correlations $\mathrm{cor}^2(X_{.30}^*, Y_{.l}^*)$, where $\mathrm{cor}^2(X_{.30}^*, Y_{.1}^*)$ reflects a true association. The main take-away message based on the result of setting 2 is that correlated response variables lead to more variable contributions for explanatory variables that can obscure the signal of truly-associated explanatory variables. However, we see that raising the squared sample correlations to a power reduces the variance of the contributions and may allow us to identify the truly-associated explanatory variables.

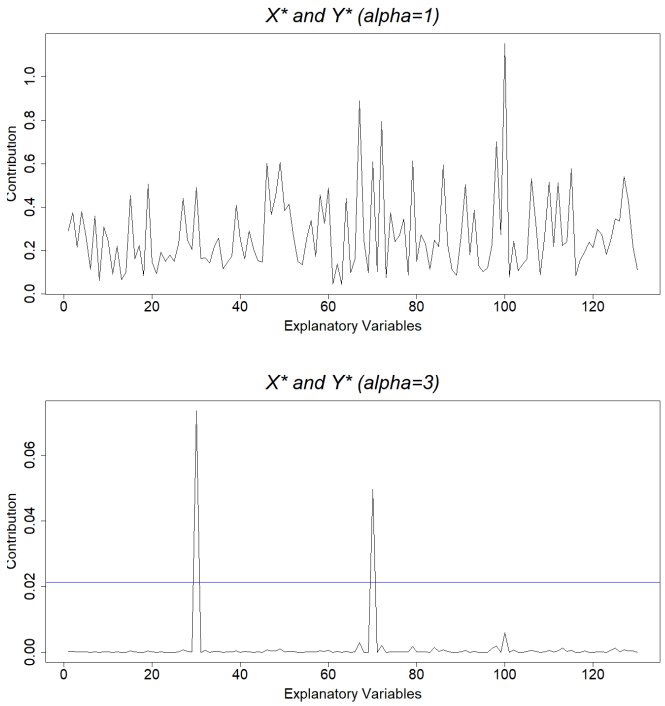

Figure 4.3: Simulation results of Setting 2 ($B_{30,1}$=$B_{70,10}$=1, p=130, q=25, $\Sigma_{p\times p}$=$I_{p\times p}$, $\Sigma_{q\times q}$≠$I_{q\times q}$). The horizontal line in the lower panel indicates the 95th percentile of the maximum contributions under the permutation null distribution.

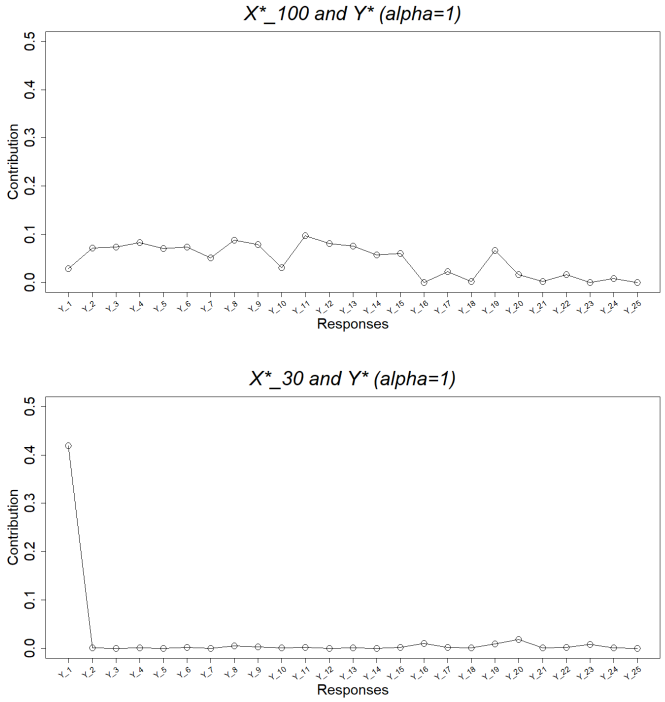

Figure 4.4: Squared correlations of $X^*_{.,100}$ and $Y^*_{.l}$ (upper) and $X^*_{.,30}$ and $Y^*_{.l}$ (lower).

Setting 3 : 20 times more X variables.

The p-value for the aSPC test on this simulated data set is 0.0008, reflecting the true association between X_{30} and Y_1 and between X_{70} and Y_{10}. The contribution plots are shown in Figure 4.5. The increase in the number of explanatory variables has little impact on the ability of the contribution plots to identify the source of the correlation signal.

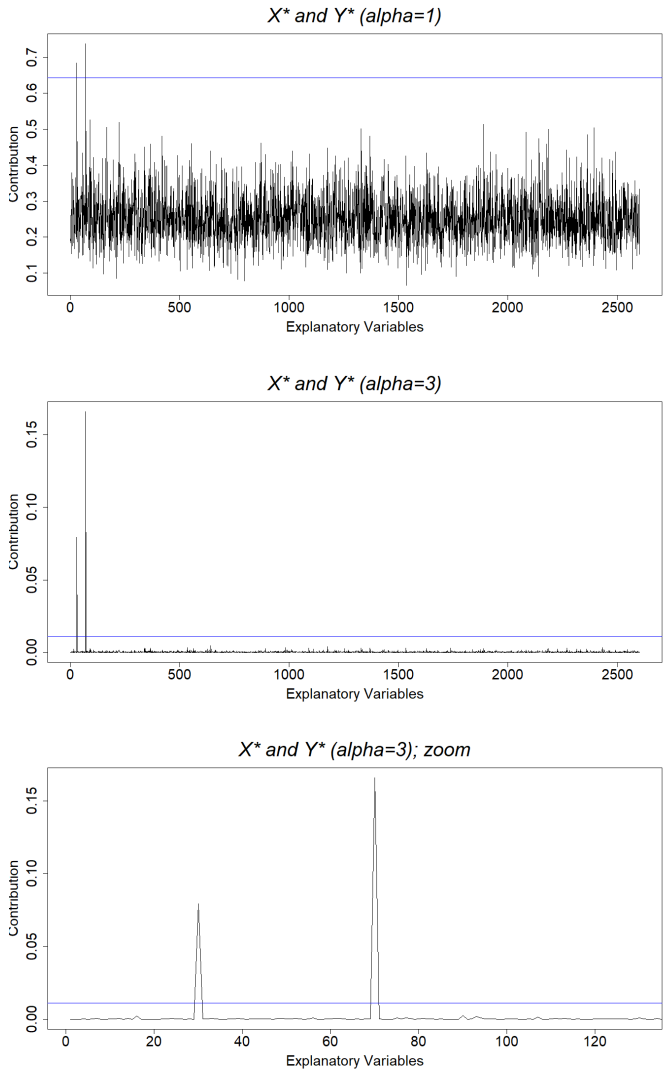

Figure 4.5: Simulation results of Setting 3 ($B_{30,1}$=$B_{70,10}$=1, p=2600, q=25, $\Sigma_{p\times p}$=$I_{p\times p}$, $\Sigma_{q\times q}$=$I_{q\times q}$). The horizontal line indicates the 95th percentile of the maximum contributions under the permutation null distribution.

Setting 4 : There are 20 times more Y variables.

The p-value for the aSPC test on this simulated data set is 0.0008, reflecting the true association between X_{30} and Y_1 and between X_{70} and Y_{10}. The contribution plots shown in Figure 4.6 are qualitatively similar to those for setting 2 (Figure 4.3), which involved correlated response variables. Increasing the number of response variables increases the variance of the contributions for the explanatory variables because the contributions are a sum of a large number of squared sample correlations between the targeted explanatory variable and responses. Thus, we have increased the variance of the contributions by adding more response variables. In setting 2, we also increased the variance of the contributions, but by adding correlated response variables. Here again, raising squared correlations to a power reduces variance and allows us to identify the source of the significant aSPC test. The inclusion of more response variables that are truly unassociated with any of the explanatory variables has no obvious effect on our ability to identify the truly-associated response variables.

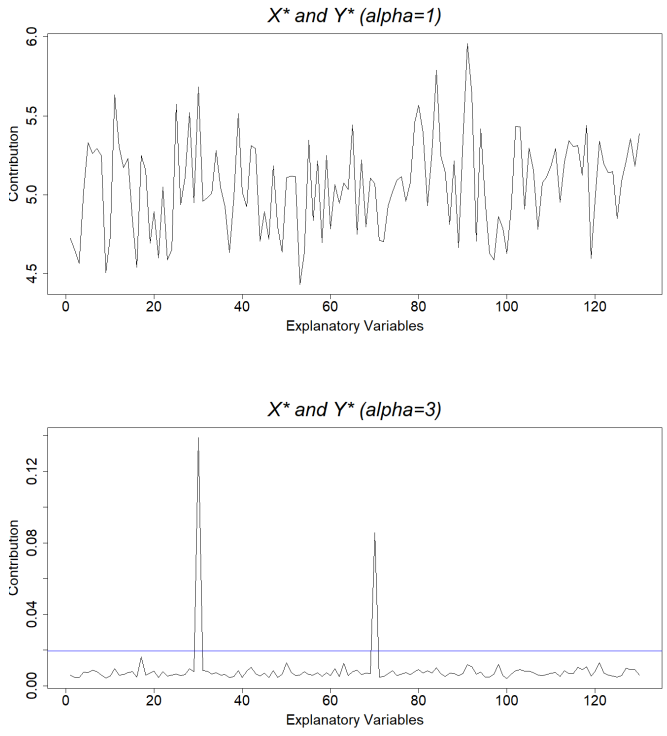

Figure 4.6: Simulation results of Setting 4 ($B_{30,1}$=$B_{70,10}$=1, p=130, q=500, $\Sigma_{p\times p}$=$I_{p\times p}$, $\Sigma_{q\times q}$=$I_{q\times q}$). The horizontal line in the lower panel indicates the 95th percentile of the maximum contributions under the permutation null distribution.

Setting 5 : All settings from 1 to 4 are adopted.

The p-value for the aSPC test on this simulated data set is 0.0008. The contribution plots, shown in Figure 4.7, illustrate all of the main features of the previous examples. Correlation between a truly-associated explanatory variable and other explanatory variables widens the peak signal around X_{30}. Correlation between response variables increases the variance of the contributions, which can obscure true associations, but this increased variance can be mitigated by raising squared correlations to higher powers.

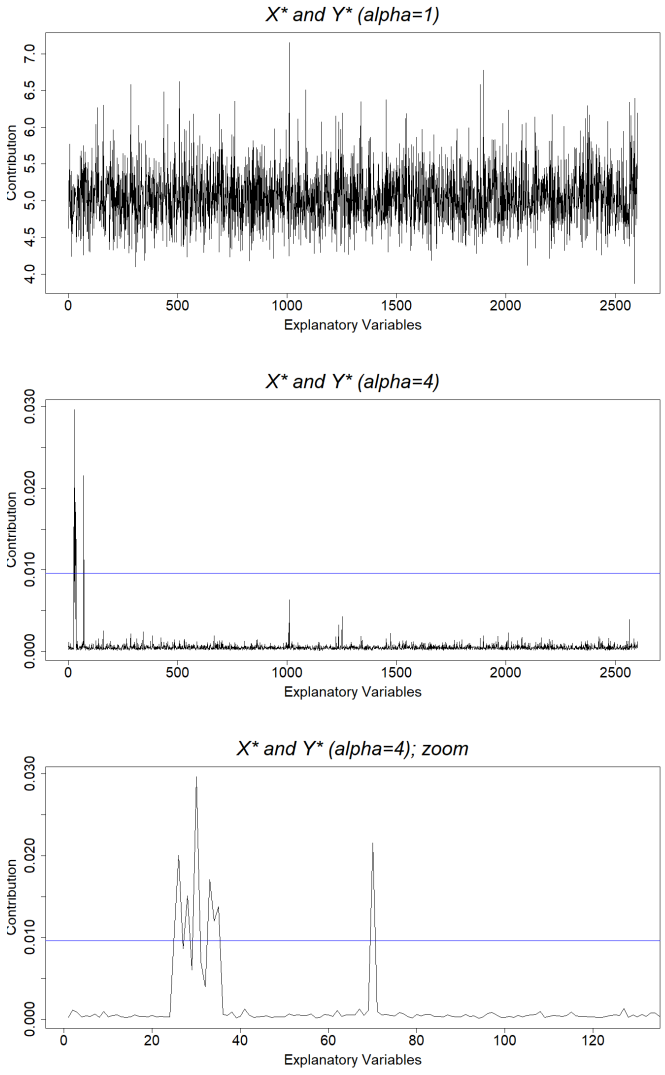

Figure 4.7: Simulation results of Setting 5 ($B_{30,1}$=$B_{70,10}$=1, p=2600, q=500, $\Sigma_{p\times p}\neq I_{p\times p}$, $\Sigma_{q\times q}\neq I_{q\times q}$). The horizontal line indicates the 95th percentile of the maximum contributions under the permutation null distribution.

4.3.1 Summary of Simulated Example Data Analyses

The contribution plot is intended as a *post hoc* investigation of an association between multiple explanatory variables and multiple response variables, to identify particular explanatory variables that may be responsible for the linear association with response variables. Our simulated data examples illustrate two main points about the contribution plot. First, correlation between explanatory variables can widen the peak of a signal, making it difficult to pin-point the particular variable(s) driving an association. Second, increasing the variance of the contributions, either through correlation between the responses or through increasing the number of responses, can obscure the signal. However, raising squared correlations to a power can counteract this increase in variance and may allow us to identify the explanatory variables that are responsible for an association.

Chapter 5

Conclusion

5.1 Project Summary

Measures of multivariate correlation are used in fields such as neurogenetics to find an association between a multivariate phenotype and a vector of explanatory variables. After an association is found, it may be of interest to identify the explanatory variables that are primarily responsible for the signal. In this project we have developed such a *post hoc* procedure and applied it to data from the ADNI-1 study. The contribution plot decomposes the RV coefficient into contributions from each explanatory variable and displays them graphically. A significance threshold for the maximum contribution under no association, determined by a permutation procedure, may be added to the plot. Signals above the threshold are considered noteworthy.

Chapter 2 introduced a population measure of correlation, ρ_V, and its estimator, the RV coefficient. Contributions to the population correlation were defined as sums of squared population covariances between individual explanatory variables (genotypes) and response variables (phenotypes). Formulas for these contributions were derived under a multivariate regression model and were seen to simplify if the response and explanatory variables are standardized, in which case covariances become correlations. We then discussed the estimation when the correlation signal is sparse and the idea of raising squared correlations to a power α. A method for selecting α was described, motivated by the adaptive sum of powered correlations (aSPC) test (Xu et al., 2017), and the approach was illustrated on a simulated data set.

In Chapter 3, we applied the methods of Chapter 2 to the ADNI-1 data. The aSPC test for correlation between SNP genotypes and phenotypes of brain regions of interest was significant (p=0.0154). The contribution plot suggested a sparse signal, driven by a single SNP, *rs16871157*, within the *NEDD9* gene on chromosome 6. Further investigation suggested that carriers of the variant allele at *rs16871157* had a tendency toward reduced

cortical thickness, whereas those with two copies of the wild type allele did not. Reduced cortical thickness has been observed to be associated with symptom severity in MCI and early AD patients.

Chapter 4 summarized the analyses of six data sets simulated under different scenarios. The two main conclusions from this investigation were that (i) correlation between explanatory variables can widen the peak of a signal, making it difficult to pin-point the particular variable(s) driving an association and (ii) correlation between response variables increases the variance of signal. Raising squared correlations to a power was found to reduce the variance of the contributions, allowing us to identify the explanatory variables responsible for the simulated associations.

5.2 Limitations and Future Work

Though the sample of CN subjects analyzed in this project were an ethnically homogeneous group of non-hispanic whites, it is still possible that the significant aSPC test was due to confounding by population stratification. Following Szefer et al. (2017), we intend to adjust both the phenotypes and genotypes for the genome-wide principal components and analyze the adjusted data to account for confounding by hidden ancestry. Further investigation of the role of *rs16871157* in the gene product of *NEDD9* is ongoing.

Finally, we note that the contribution plot can be extended to the case where study subjects are differentially weighted. The sample for our study was a population sample of CN subjects, and were all equally weighted. If we used the entire ADNI-1 sample instead, which is enriched for MCI and AD subjects, we would need to correct for the sampling bias by computing weighted covariances or correlations, where the weights are inversely proportional to the probability that each subject is included in the sample (Horvitz and Thompson, 1952). The contribution plot in terms of weighted covariance would be of the same form. See Appendix A for details. Investigating the properties of the contribution plot for unequally weighted subjects is an item for future work.

Bibliography

Adams, D. C. (2016). Evaluating modularity in morphometric data: challenges with the rv coefficient and a new test measure. *Methods in Ecology and Evolution*, 7(5):565–572.

Alzheimer's Association (2017). 2017 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 13(4):325–373.

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., et al. (2008). The cortical signature of alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild ad dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral cortex*, 19(3):497–510.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4):751–760.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M. M., Abraham, R., Hamshere, M. L., Pahwa, J. S., Moskvina, V., et al. (2011). Common variants at abca7, ms4a6a/ms4a4e, epha1, cd33 and cd2ap are associated with alzheimer's disease. *Nature genetics*, 43(5):429.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Izumchenko, E., Singh, M. K., Plotnikova, O. V., Tikhmyanova, N., Little, J. L., Serebriiskii, I. G., Seo, S., Kurokawa, M., Egleston, B. L., Klein-Szanto, A., et al. (2009). Ned9 promotes oncogenic signaling in mammary tumor development. *Cancer research*, 69(18):7198–7206.

Josse, J. and Holmes, S. (2014). Tests of independence and beyond. *arXiv preprint arXiv:1307.7383*.

Josse, J. and Holmes, S. (2016). Measuring multivariate association and beyond. *Statist. Surv.*, 10:132–167.

Josse, J., Pagès, J., and Husson, F. (2008). Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91.

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational statistics & data analysis*, 20(6):643–656.

Lee, E., Giovanello, K. S., Saykin, A. J., Xie, F., Kong, D., Wang, Y., Yang, L., Ibrahim, J. G., Doraiswamy, P. M., Zhu, H., et al. (2017). Single-nucleotide polymorphisms are associated with cognitive decline at alzheimer's disease conversion within mild cognitive impairment patients. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 8:86–95.

Machiela, M. J. and Chanock, S. J. (2015). Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21):3555–3557.

Mielke Jr, P. W., Berry, K. J., and Brier, G. W. (1981). Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. *Monthly Weather Review*, 109(1):120–126.

Naj, A. C., Jun, G., Beecham, G. W., Wang, L.-S., Vardarajan, B. N., Buros, J., Gallins, P. J., Buxbaum, J. D., Jarvik, G. P., Crane, P. K., et al. (2011). Common variants at ms4a4/ms4a6e, cd2ap, cd33 and epha1 are associated with late-onset alzheimer's disease. *Nature genetics*, 43(5):436.

Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., Risacher, S. L., Nho, K., Huentelman, M. J., Craig, D. W., et al. (2010). Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 6(3):265–273.

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C., and Van Erk, M. (2008). Matrix correlations for high-dimensional data: the modified rv-coefficient. *Bioinformatics*, 25(3):401–405.

Szefer, E., Lu, D., Nathoo, F., Beg, M. F., Graham, J., et al. (2017). Multivariate association between single-nucleotide polymorphisms in alzgene linkage regions and structural changes in the brain: discovery, refinement and validation. *Statistical applications in genetics and molecular biology*, 16(5-6):349–365.

Szefer, E. K. (2014). Joint analysis of imaging and genomic data to identify associations related to cognitive impairment.

Vogel, T., Ahrens, S., Büttner, N., and Krieglstein, K. (2009). Transforming growth factor β promotes neuronal cell fate of mouse cortical and hippocampal progenitors in vitro and in vivo: identification of nedd9 as an essential signaling component. *Cerebral Cortex*, 20(3):661–671.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., Shen, L., and Initiative, A. D. N. (2011). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237.

Wang, Y., Bi, L., Wang, H., Li, Y., Di, Q., Xu, W., and Qian, Y. (2012). Nedd9 rs760678 polymorphism and the risk of alzheimer's disease: a meta-analysis. *Neuroscience letters*, 527(2):121–125.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. (2013). The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(5):e111–e194.

Xu, Z., Xu, G., and Pan, W. (2017). Adaptive testing for association between two random vectors in moderate to high dimensions. *Genetic epidemiology*, 41(7):599–609.

Appendix A

Alternative RV Coefficient Forms

This appendix derives the connections between three different forms of the RV coefficient: (i) in terms of inner-product matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{Y}\mathbf{Y}^T$, (ii) in terms of squared covariances, as in equation (2.2) (repeated below as equation A.2), and (iii) in terms of Gower-centred distance matrices (defined below). We start with simple derivations that illustrate the connections between the forms of the RV coefficient, assuming all subjects are weighted equally. We then extend these results to unequally weighted subjects, using more abstract notation and results from linear algebra.

A.1 Equally-Weighted Subjects

Let $\mathbf{X}(n \times p)$ and $\mathbf{Y}(n \times q)$ denote data matrices in which each column has been centred by its ordinary arithmetic mean. Let $X_{i\cdot}$ and $Y_{i\cdot}$ be the ith rows of $\mathbf{X}$ and $\mathbf{Y}$, respectively and let x_{ij} and y_{ij} denote the (i,j)th elements of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Due to the column centring, $\sum_{i=1}^{n} x_{ij} = 0$.

A.1.1 Inner Products and Squared Covariances

The inner-product form of the RV coefficient is

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{tr(\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T)}{\sqrt{tr(\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T)tr(\mathbf{Y}\mathbf{Y}^T\mathbf{Y}\mathbf{Y}^T)}} \tag{A.1}$$

and the squared covariance form is

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q}\mathrm{cov}^2(X_{.k},Y_{.l})}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{cov}^2(X_{.k},X_{.l})\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{cov}^2(Y_{.k},Y_{.l})}}, \tag{A.2}$$

where $X_{.k}$ and $Y_{.k}$ are the kth columns of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

The key steps to show equality of the two forms are to (i) expand the traces in the inner-product formula into sums, (ii) reorder summations, and (iii) multiply numerator and denominator by an appropriate constant. To simplify writing traces as sums, let $\mathbf{S} = \mathbf{X}\mathbf{X}^T = \mathbf{S}^T$ and $\mathbf{T} = \mathbf{Y}\mathbf{Y}^T = \mathbf{T}^T$ be the inner-product matrices with elements s_{ij} and t_{ij}, respectively. We call these inner-product matrices because each matrix element can be shown to be an inner product between rows of the data matrices: $s_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk} = \langle X_{i\cdot}, X_{j\cdot}\rangle$; $y_{ij} = \sum_{k=1}^{p} y_{ik}y_{jk} = \langle Y_{i\cdot}, Y_{j\cdot}\rangle$. From the basic properties of the trace operator we can deduce that for matrices $\mathbf{A}$ and $\mathbf{B}$, $tr(\mathbf{AB}) = \sum_i \sum_j a_{ji}b_{ij}$. Thus the numerator of equation (A.1) is

$$\begin{aligned} tr(\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T) &= tr(\mathbf{S}\mathbf{T}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ji}t_{ij} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}\left\{\sum_{l=1}^{q} y_{il}y_{jl}\right\}. \end{aligned}$$

The denominator may be expanded into sums analogously and we obtain

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}\left\{\sum_{l=1}^{q} y_{il}y_{jl}\right\}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}^2 \times \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{q} y_{jk}y_{ik}\right\}^2}} \tag{A.3}$$

Now reorder the summations; e.g.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}\left\{\sum_{l=1}^{q} y_{il}y_{jl}\right\} = \sum_{k=1}^{p}\sum_{l=1}^{q}\left\{\sum_{i=1}^{n} x_{ik}y_{il}\right\}\left\{\sum_{j=1}^{n} x_{jk}y_{jl}\right\} = \sum_{k=1}^{p}\sum_{l=1}^{q}\left\{\sum_{i=1}^{n} x_{ik}y_{il}\right\}^2.$$

Since the data matrices have been column centred, the last expression is equal to $\sum_{k=1}^{p}\sum_{l=1}^{q}(n-1)^2\mathrm{cov}^2(X_{.k},Y_{.l})$. Similar rearrangements of the sums in the denominator give

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}^2 = \sum_{k=1}^{p}\sum_{l=1}^{p}(n-1)^2\mathrm{cov}^2(X_{.k},X_{.l})$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{q} y_{jk}y_{ik}\right\}^2 = \sum_{k=1}^{q}\sum_{l=1}^{q}(n-1)^2\mathrm{cov}^2(Y_{\cdot k}, Y_{\cdot l}).$$

Inserting these reordered sums into equation (A.3) and cancelling the constant $(n-1)^2$ terms gives the covariance form of the RV coefficient in equation (A.2).

A.1.2 Inner Products and Gower-Centred Distances

Define the Gower-centred distance matrices as follows. Let $d_{ij}^X = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$ be the squared Euclidean distance between $X_{i\cdot}$ and $X_{j\cdot}$, $\bar{d}_{i\cdot}^X = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^X$, $\bar{d}_{\cdot j}^X = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^X$ and $\bar{d}_{\cdot\cdot}^X = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^X$. Then the Gower-centred distance matrix for $\mathbf{X}$, Δ_X, is the symmetric matrix with (i,j) element

$$\delta_{ij}^X = d_{ij}^X - \bar{d}_{i\cdot}^X - \bar{d}_{\cdot j}^X + \bar{d}_{\cdot\cdot}^X.$$

The Gower-centred distance matrix for the $\mathbf{Y}$, Δ_Y, is defined analogously.

The form of the RV coefficient in terms of Gower-centred distances is

$$RV(\mathbf{X},\mathbf{Y}) = \frac{tr(\Delta_X\Delta_Y)}{\sqrt{tr(\Delta_X\Delta_X)tr(\Delta_Y\Delta_Y)}},$$

which can be expanded to sums as

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}^X\delta_{ij}^Y}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij}^X)^2 \times \sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij}^Y)^2}} \tag{A.4}$$

Equality of equations (A.3) and (A.4) follows from the equalities $\sum_{k=1}^{p} x_{ik}x_{jk} = -\delta_{ij}^X/2$, shown as follows. Briefly, we start by writing

$$d_{ij}^X = \sum_{k=1}^{p}(x_{ik}-x_{jk})^2 = \sum_{k=1}^{p} x_{ik}^2 - 2\sum_{k=1}^{p} x_{ik}x_{jk} + \sum_{k=1}^{p} x_{jk}^2. \tag{A.5}$$

From Section A.1.1 we recognize $\sum_{k=1}^{p} x_{ik}x_{jk}$ as the (i,j) element of $\mathbf{X}\mathbf{X}^T$. Expressions for $\sum_{k=1}^{p} x_{ik}^2$ and $\sum_{k=1}^{p} x_{jk}^2$ in terms of distances are derived as follows. By averaging equation (A.5) over index j, we obtain

$$\begin{aligned}
\bar{d}_{i\cdot}^X &= \frac{1}{n}\sum_{j=1}^{n}\left\{\sum_{k=1}^{p} x_{ik}^2 - 2\sum_{k=1}^{p} x_{ik}x_{jk} + \sum_{k=1}^{p} x_{jk}^2\right\} \\
&= \sum_{k=1}^{p} x_{ik}^2 - \frac{2}{n}\sum_{k=1}^{p} x_{ik}\sum_{j=1}^{n} x_{jk} + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{p} x_{jk}^2 \\
&= \sum_{k=1}^{p} x_{ik}^2 + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{p} x_{jk}^2.
\end{aligned}$$

Similarly, averaging over index i yields

$$\bar{d}^X_{\cdot j} = \sum_{k=1}^{p} x_{jk}^2 + \frac{1}{n}\sum_{i=1}^{p}\sum_{k=1}^{p} x_{ik}^2$$

and averaging over i and j yields

$$\bar{d}^X_{\cdot\cdot} = \frac{2}{n}\sum_{j=1}^{n}\sum_{k=1}^{p} x_{jk}^2.$$

Rearranging these expressions gives

$$\sum_{k=1}^{p} x_{ik}^2 = \bar{d}^X_{i\cdot} - \bar{d}^X_{\cdot\cdot}/2 \quad \text{and} \quad \sum_{k=1}^{p} x_{jk}^2 = \bar{d}^X_{\cdot j} - \bar{d}^X_{\cdot\cdot}/2.$$

Substituting into (A.5) and further rearranging leads to

$$\sum_{k=1}^{p} x_{ik}x_{jk} = -(d^X_{ij} - \bar{d}^X_{i\cdot} - \bar{d}^X_{\cdot j} + \bar{d}^X_{\cdot\cdot})/2 = -\delta^X_{ij}/2, \quad \text{as desired.}$$

A.2 Unequally-Weighted Subjects

We can generalize the RV coefficient for unequally weighted subjects with weights $w_1, \ldots, w_n$ that sum to one. Such an approach might be used to correct for sampling bias if sampling is stratified and some population subgroups are oversampled relative to others. Both $\mathbf{X}$ and $\mathbf{Y}$ are now column centred by weighted averages, so that, for example, $\sum_{i=1}^{n} w_i x_{ik} = 0$ for all $k = 1, \ldots, p$.

Of the three expressions for the RV coefficient, the squared covariance form is the most obvious for generalization. We assign weights to sample covariances and variances; e.g., $\text{cov}(X_{\cdot k}, Y_{\cdot l}) = \sum_{i=1}^{n} w_i x_{ik} y_{il}$. The RV coefficient becomes

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{\displaystyle\sum_{k=1}^{p}\sum_{l=1}^{q}\left\{\sum_{i=1}^{n} w_i x_{ik} y_{il}\right\}^2}{\displaystyle\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p}\left\{\sum_{i=1}^{n} w_i x_{ik} x_{il}\right\}^2 \sum_{k=1}^{q}\sum_{l=1}^{q}\left\{\sum_{i=1}^{n} w_i y_{ik} y_{il}\right\}^2}}$$

An implementation of this formula in R is:

```
RV.cov = function(X,Y,wts){
  S  = cov.wt(cbind(X,Y),wt=wts)$cov
  p = ncol(X); q = ncol(Y)
  SXX = S[1:p,1:p]
```

```
  SYY = S[(p+1):(p+q),(p+1):(p+q)]
  SXY = S[1:p,(p+1):(p+q)]
  return(sum(SXY^2)/sqrt(sum(SXX^2) * sum(SYY^2)))
}
```

Replacing the equal weights over subjects with unequal weights in (A.3) and (A.4) gives

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}\left\{\sum_{l=1}^{q} y_{il}y_{jl}\right\}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left\{\sum_{k=1}^{p} x_{jk}x_{ik}\right\}^2 \times \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left\{\sum_{k=1}^{q} y_{jk}y_{ik}\right\}^2}}$$

and

$$RV(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \delta_{ij}^X \delta_{ij}^Y}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j (\delta_{ij}^X)^2 \times \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j (\delta_{ij}^Y)^2}},$$

respectively. One can use the calculations in Section A.1 as a template to verify the equality of the above formulae.

The distance-based RV calculation can be implemented as

```
RV.dist = function(X,Y,wts){
  D.X = as.matrix(dist(X))^2
  D.Y = as.matrix(dist(Y))^2
  n = nrow(X)
  H = diag(rep(1,n)) - outer(rep(1,n),wts)
  Delta.X = H %*% D.X %*% t(H)
  Delta.Y = H %*% D.Y %*% t(H)
  WW = outer(wts,wts) # matrix whose i,j element is w_i w_j
  return(sum(WW*Delta.X*Delta.Y)/sqrt(sum(WW*Delta.X^2)*sum(WW*Delta.Y^2)))
}
```

A.2.1 Testing the Distance-Based Formula

For data sets with p and/or q larger than n, the distance-based formula may be faster to compute. This is illustrated below.

```
# Test RV stat by distances on a small problem with p>n
n = 60; p = 2200; q = 54
set.seed(123)
wts = (1:n) ; wts = wts/sum(wts) # Make up some wts; must sum to 1
X = matrix(rnorm(n*p),nrow=n)
Y = matrix(rnorm(n*q),nrow=n)
system.time({val1 = RV.cov(X,Y,wts)})
```

```
##    user  system elapsed
##    0.64    0.06    0.70
```

```
  val1
```

```
## [1] 0.7252929
```

```
  system.time({val2 = RV.dist(X,Y,wts)})
```

```
##    user  system elapsed
##    0.06    0.00    0.06
```

```
  val2
```

```
## [1] 0.7252929
```

Appendix B

Properties of the Multivariate Correlation and RV Coefficients

B.1 Properties of ρ_V

Recall the definition of the multivariate correlation coefficient

$$\rho_V(X,Y) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q}\mathrm{Cov}^2(X_k,Y_l)}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{Cov}^2(X_k,X_l)\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}^2(Y_k,Y_l)}}, \tag{B.1}$$

where Cov() denotes population covariance. ρ_V has the following properties (Josse and Holmes, 2016):

1. $0 \leq \rho_V(X,Y) \leq 1$. Geometrically, if X and Y are mean-zero random vectors, then $\rho_V(X,Y)$ is a cosine of the angle between them in an appropriately defined inner-product space.
2. Given $p = q = 1$, $\rho_V(X,Y) = \rho^2_{X,Y}$, where $\rho_{X,Y}$ is the population Pearson correlation coefficient.
3. $\rho_V(X,Y) = 0$ if and only if X and Y are uncorrelated.
4. $\rho_V(X,X) = 1$
5. $\rho_V(X,Y) = \rho_V(aX + \mathbf{c}, bY + \mathbf{d})$ where a and b are non-zero constants, $\mathbf{c}$ and $\mathbf{d}$ are constant vectors. In addition, $\rho_V(X,Y)$ is invariant to permutation of the elements of X and Y.
6. if $\mathbf{B}$ is an orthonormal matrix, $\rho_V(X,Y) = \rho_V(X, Y\mathbf{B})$.

B.2 Properties of RV

$RV(\mathbf{X}, \mathbf{Y})$ has the same properties as those listed above for ρ_V.

B.3 Dependence of the RV Coefficient on Sample Size and Dimension

Despite the reasonable geometrical interpretation, it has been pointed out that the RV coefficient has underlying problems triggered by two factors, sample size and dimensionality of data. Consequently, the RV coefficient should not be used directly as a measure of linear association between two data matrices.

B.3.1 Sensitivity to Sample Size

The main defect of the RV coefficient is its dependence on sample size. Such dependence is evident in the approximation (Smilde et al., 2008):

$$RV(\mathbf{X}, \mathbf{Y}) \approx \frac{pq}{\sqrt{\{p^2 + (n+1)p\}\{q^2 + (n+1)q\}}}, \tag{B.2}$$

and in the first moment of the RV coefficient under the the permutation distribution (Kazi-Aoual et al., 1995):

$$E(RV) = \frac{\sqrt{\beta_X \times \beta_Y}}{n-1} \quad \text{where } \beta_X = \frac{(tr(\mathbf{X}^T\mathbf{X}))^2}{tr(\mathbf{X}^T\mathbf{X})^2}. \tag{B.3}$$

Figure B.1 visualizes the change of the RV coefficient by sample size while the total number of variables remains 200. The left-hand plot shows the mean, 95% confidence interval, and the RV coefficient approximation of equation (B.2) under the null hypothesis of no association. As the sample size increases, the RV coefficient monotonically decreases. The right-hand plot illustrates that the vertical location of the approximation line varies at different proportions of the number of variables in $\mathbf{X}$ and $\mathbf{Y}$.

B.3.2 Sensitivity to Dimensionality

The RV coefficient is also affected by increases in the dimensions p and q, holding the sample size n fixed. The dependence on p and q is not obvious from (B.2) and (B.3). Adams (2016) showed an upward trend in the RV coefficient as the total number of variables increases.

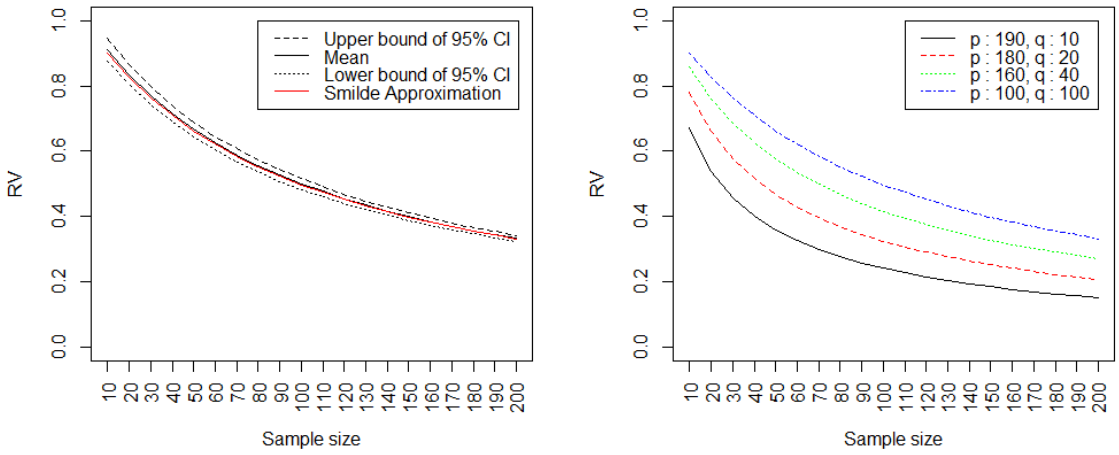

Figure B.1: RV coefficient with different sample sizes.

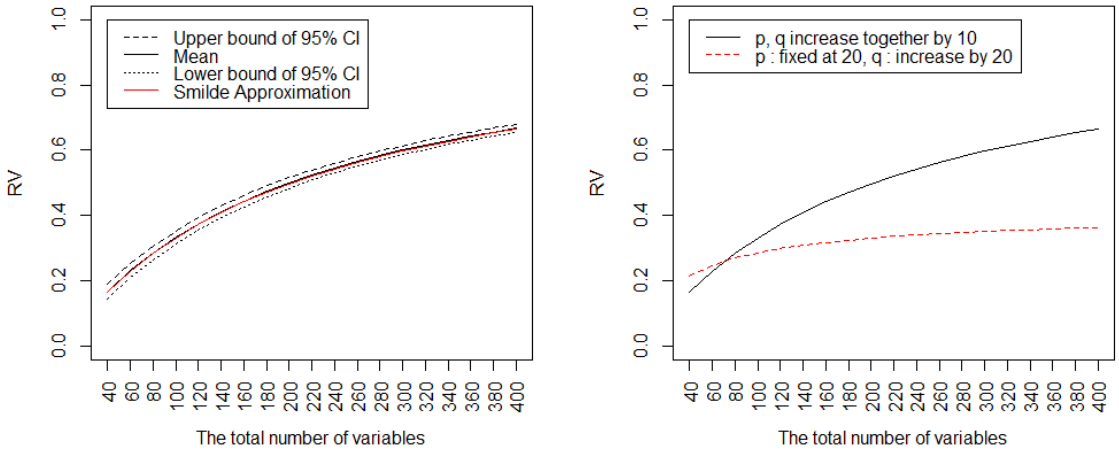

Figure B.2: RV coefficient with the different numbers of variables.

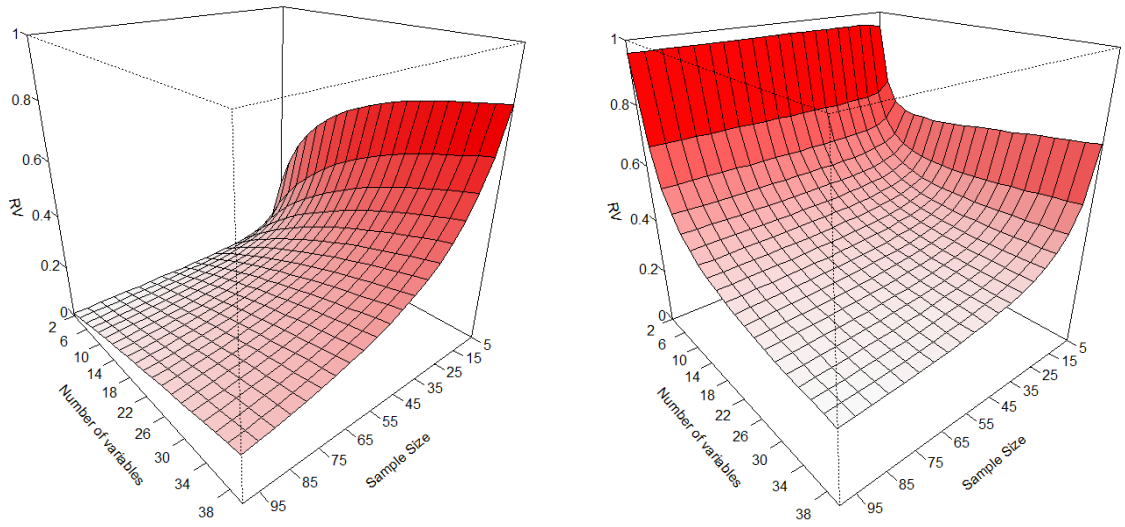

Figure B.3: RV coefficient change at different levels of the sample size and number of variables in two simulation sets without (left) and with (right) a linear association.

This is illustrated by the left-hand plot of Figure B.2. The right-hand plot illustrates that the rate of increase depends on how the dimensionality is increased (p and q together *versus* p fixed and q increasing). At $p+q = 400$, the RV cofficient is approximately 0.6 if p=q=200 (black line), whereas it is approximately 0.3 if $p = 20$ and $q = 380$ (red dashed line).

Additional notes on unexpected behaviour of the RV coefficient is as follows. First, we can see from Figure B.3 that the RV coefficient can either increase or decrease in the number of variables depending on whether there is or is not a linear association between data matrices. The left-hand plot in Figure B.3 (no linear association) shows an increase in RV for each sample size, while the right-hand plot shows a monotonic decrease in RV for some sample sizes and a curvelinear relationship in the number of variables for other sample sizes. Second, we note that the RV coefficient is not always affected by sample size and variable number. In particular, the RV coefficient remains constant in the exceptional case in which data matrices are artificially augmented by their duplications; for example, $RV(\mathbf{X}, \mathbf{Y}) = RV(rbind(\mathbf{X}, ..., \mathbf{X}), rbind(\mathbf{Y}, ..., \mathbf{Y})) = RV(cbind(\mathbf{X}, ..., \mathbf{X}), cbind(\mathbf{Y}, ..., \mathbf{Y}))$.

B.4 Hypothesis Test

Due to the dependence of the RV coefficient to both sample size and variable number, a hypothesis testing is usually conducted as a valid inferential method to test the significance of the association between two data matrices. As a test statistics, the standardized RV coefficient

$$Z_{RV} = \frac{RV(\mathbf{X}, \mathbf{Y}) - E(RV(\mathbf{X}, \mathbf{Y}))}{\sqrt{\text{var}(RV(\mathbf{X}, \mathbf{Y}))}} \tag{B.4}$$

is computed for a hypothesis testing under the null $H_0 : \rho_V(X, Y) = 0$ *versus* the alternative : $H_a : \rho_V(X, Y) > 0$. E() denotes the mean of $RV(\mathbf{X}, \mathbf{Y})$.

B.4.1 Permutation Distribution

The exact distribution of Z_{RV} is unknown, but approximations to the permutation distribution of Z_{RV} have been developed. The permutation distribution under the null hypothesis of no association is formulated by calculating all possible Z_{RV}s under rearrangment of the row labels of one of the data matrices. However, as the number of permutations increases factorially, it is extremely time costly and often even not feasible computationally to take into account all the possible cases. For example, if there are 30 units, then there are 30!=2.652529e+32 permutations and it is not feasible to enumerate them all.

B.4.2 Pearson Type III Distribution

The permutation distribution may be approximated by either Monte Carlo sampling, as discussed in the main text of this project, or by a continuous distribution, as follows. The approximating distribution is chosen from a parametric family to match the first three moments of (B.4). Kazi-Aoual et al. (1995) showed that the first moment is (B.3) and the second moment is

$$V(RV) = \frac{2\alpha_X \alpha_Y}{(n+1)(n-1)^2(n-2)} \left\{ 1 + \frac{n-3}{2n(n-1)} \Gamma_X \Gamma_Y \right\}$$

where $\alpha_X = (n - 1 - \beta_X)$

and $\Gamma_X = \dfrac{n-1}{(n-3)(n-1-\beta_X)} \left\{ n(n+1) \dfrac{\Sigma_i (\mathbf{X}^T\mathbf{X})^2_{ii}}{tr(\mathbf{X}^T\mathbf{X})^2} - (n-1)(\beta_X + 2) \right\}$.

The third moment can be found in Appendix of Kazi-Aoual et al. (1995).

The parametric family used for the approximation is the Pearson type III distribution, also known as a gamma distribution with the mean of 0 and the variance of 1. This approximating family allows for better approximation than the Normal distribution (Josse et al., 2008). The Pearson type III distribution is expressed by

$$f(z) = \frac{(2/\gamma)^{4/\gamma^2}}{\Gamma(4/\gamma^2)} \{(2 + z\gamma)/\gamma\}^{(4-\gamma^2)/\gamma^2} e^{-2(2+z\gamma)/\gamma^2}$$

where γ is the skewness of Z_{RV}. Correspondingly, $Z_{RV} + \beta \sim \Gamma(\alpha, \beta)$ where $\alpha = \frac{4}{\gamma^2}$ and $\beta = \frac{2}{\gamma}$. The distribution is always right-skewed with $\beta > 0$.

The p-value

$$P[RV(\mathbf{X}, \mathbf{Y}) \leq RV] = 1 - \int_{-\infty}^{Z_{RV}} f(z)dz$$

may be approximated using Simpson's rule (Mielke Jr et al., 1981).

Approximation of the permutation distribution by the Pearson type III and Normal distributions are illustrated in Figure B.4. We see that the shape of the sampling distribution does not significantly depend on the sample size, but does depend on the number of variables. The left plot in the first row shows a histogram of the standardized RV coefficients based on 5,000 resamples using the Monte Carlo method for $\mathbf{X}_{10\times2}$ and $\mathbf{Y}_{10\times2}$. The red line and the blue line represent the standard Normal approximation and the Pearson type III approximation respectively. The right-hand plot in the first row is for data matrices $\mathbf{X}_{10\times10}$ and $\mathbf{Y}_{10\times10}$. The left-hand plot in the second row is for $\mathbf{X}_{1000\times2}$ and $\mathbf{Y}_{1000\times2}$. Comparing the left-hand plots in the both rows we see similar skewness despite the difference in sample size. The bottom-left plot shows the Pearson type III approximation with different dimensionalities of data matrices.

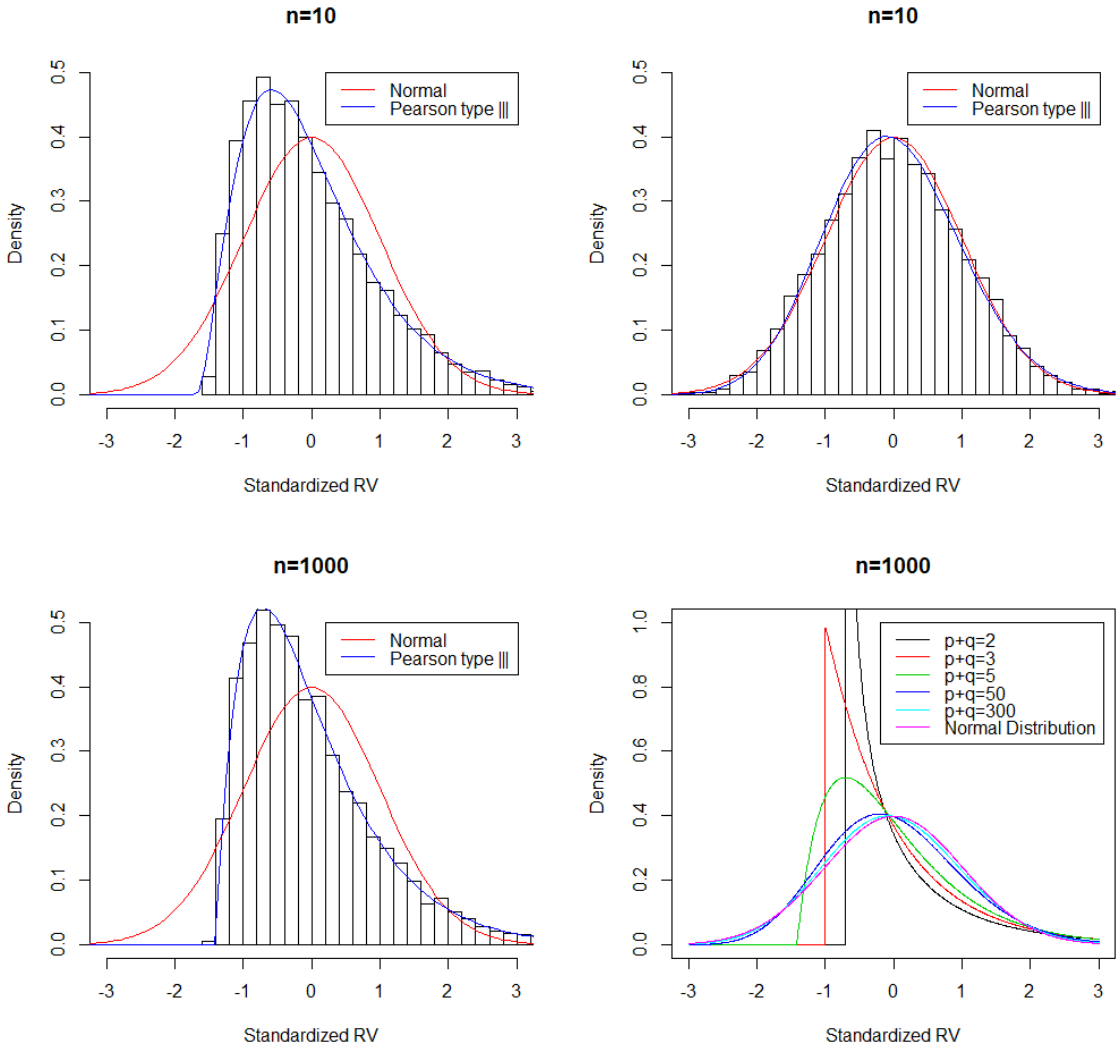

Figure B.4: Pearson type III approximation and Normal approximation of the standardized RV coefficient.

Appendix C

Names of SNPs in analyzed genes

Chr	Gene	SNPs
1	CHRNB2	rs3811450
	CR1	rs10127904, rs10779339, rs11117959, rs11118131, rs12734030, rs1408077, rs1571344, rs2025935, rs3737002, rs3818361, rs4310446, rs650877, rs6691117, rs6701713, rs677066
	ECE1	rs1076669, rs10916958, rs10916959, rs11590928, rs12562197, rs12756690, rs12758257, rs212515, rs212524, rs212525, rs212531, rs212534, rs212539, rs212540, rs212541, rs213010, rs213022, rs213023, rs213025, rs213028, rs213037, rs213039, rs213045, rs213052, rs213058, rs2282714, rs2282715, rs2745251, rs3026841, rs3026845, rs3026868, rs3026883, rs3026886, rs3026913, rs4654916, rs4654918, rs471359, rs84853, rs9426748
	MTHFR	rs1476413, rs1572151, rs17367504, rs1801131, rs1801133, rs2184226, rs3737964, rs4846048, rs6541003, rs9651118
	TF	rs696619, rs762484, rs762485
2	BIN1	rs10194375, rs10200967, rs11678252, rs13426725, rs13430599, rs17014873, rs17014923, rs2276575, rs6709337, rs749008, rs873270, rs880436
	IL1A	rs17561, rs3783526
	IL1B	rs1143634
6	NEDD9	rs1009667, rs1012503, rs1018374, rs10484451, rs10484453, rs10947009, rs10947021, rs11757904, rs11964334, rs11967989, rs12209631, rs1465131, rs1475345, rs16871072, rs16871157, rs16871166, rs16871236, rs16871247, rs16871253, rs17496723, rs1883235, rs1883238, rs2018334, rs2025676, rs2025677, rs2064111, rs2064112, rs2072834, rs2142739, rs2142741, rs2142742, rs2146342, rs2179179, rs2182335, rs2182337, rs2327389, rs2327394, rs2950, rs3734404, rs3798729, rs3798731, rs4713379, rs4713432, rs6457131, rs6457160, rs6457200, rs6905101, rs6908326, rs6912916, rs744970, rs760680, rs7738900, rs7741863, rs7748486, rs7769173, rs7775262, rs9295823, rs9295828, rs9296000, rs9348868, rs9368621, rs9380149, rs9393992, rs9393994, rs943008, rs9468690, rs9468793, rs967473, rs9791189
	PGBD1	rs1150724, rs13211507, rs1997660, rs2281043, rs2743554, rs9461448
	TNF	rs3093662
8	CLU	rs11136000, rs9314349

9	DAPK1	rs10125534, rs1014306, rs1015477, rs10512186, rs10512188, rs1056719, rs10746816, rs10780849, rs10868609, rs10868644, rs1105384, rs11141878, rs11141879, rs11141889, rs11141899, rs11141914, rs11141915, rs11141918, rs11141937, rs12001404, rs12378686, rs12685372, rs1316489, rs13283404, rs13288561, rs1329600, rs1421001, rs1473180, rs1475524, rs1475525, rs1554, rs1558889, rs1571515, rs17399090, rs17477673, rs17477827, rs1861828, rs1861832, rs1927976, rs1964911, rs1983973, rs2058882, rs2111554, rs2274606, rs3028, rs3095747, rs3095748, rs3118846, rs3118853, rs3118860, rs3118862, rs3124236, rs3124237, rs3124238, rs3128471, rs3128477, rs3128479, rs3128495, rs3128519, rs3128521, rs3739784, rs3793647, rs4877367, rs4877368, rs4878089, rs4878094, rs4878104, rs4878112, rs4878115, rs4878117, rs6560006, rs7025760, rs7027958, rs7036598, rs7036781, rs7046290, rs721936, rs7855635, rs913778, rs913782, rs943855, rs981292
	IL33	rs10815388, rs10975516, rs12551256, rs1330383, rs16924159, rs17498196, rs1891385, rs2066362, rs2210463, rs4740840, rs7025417, rs7033258, rs7037276, rs928413
10	CALHM1	rs11191692, rs2986018, rs729211
	CH25H	rs4933497
	ENTPD7	rs1057490, rs11190245, rs3740078, rs6584307
	SORCS1	rs1023024, rs10491052, rs10509823, rs10509825, rs10509826, rs10748924, rs10748932, rs10786972, rs10786978, rs10786998, rs10786999, rs10787010, rs10787011, rs10884339, rs10884374, rs10884381, rs10884387, rs10884399, rs10884402, rs10884409, rs11192997, rs11192998, rs11193007, rs11193130, rs11193190, rs11193198, rs11814111, rs11814145, rs11815967, rs12240854, rs12240947, rs12248379, rs12248564, rs1251753, rs1269918, rs12781860, rs1336978, rs1415020, rs1556758, rs17195022, rs17209374, rs1885352, rs1887635, rs1890457, rs2149196, rs2152676, rs2184796, rs2243454, rs2243581, rs2245123, rs2418811, rs2418828, rs2418834, rs2486154, rs2756251, rs4918255, rs4918274, rs4918282, rs596577, rs607437, rs610785, rs6584766, rs6584777, rs661319, rs669061, rs685316, rs7068978, rs7073924, rs7074484, rs7078098, rs7079264, rs7083707, rs7089127, rs7089234, rs7091546, rs7095427, rs7097380, rs717751, rs719965, rs7897726, rs7897974, rs7903481, rs7910584, rs7920985, rs821925, rs821927, rs821936, rs822094, rs822095, rs822097, rs822326, rs878183, rs911580, rs950809
	TFAM	rs1049432, rs11006130, rs11006132, rs11006133, rs12245545, rs2306604
11	GAB2	rs10501426, rs10899469, rs10899496, rs11237451, rs11601726, rs1318241, rs1893447, rs1981405, rs2292572, rs2450129, rs2511175, rs4944196, rs4945261, rs6592772, rs7107174, rs7112234, rs731600, rs7927923, rs7941639
	PICALM	rs10501602, rs10501604, rs10501608, rs10792820, rs10792821, rs10898427, rs11234495, rs11234532, rs17745273, rs1941375, rs2077815, rs475639, rs510566, rs527162, rs618679, rs642949, rs664629, rs666682, rs669556, rs677909, rs680119, rs713346, rs7938033
	SORL1	rs1010158, rs10502262, rs11218301, rs11218322, rs1133174, rs11600875, rs11601559, rs11605969, rs1503415, rs1614735, rs1620003, rs1699102, rs1699105, rs1790213, rs2070045, rs2276346, rs2298525, rs3781827, rs3781832, rs4420280, rs4631890, rs4935774, rs4935775, rs4936632, rs4936637, rs556349, rs661057, rs666004, rs676759, rs689021, rs7124060, rs726601, rs7945931
15	ADAM10	rs11854073, rs12439189, rs12441313, rs12906705, rs12908165, rs1427281, rs1427282, rs2081703, rs2305421, rs28455654, rs383902, rs4238331, rs4275799, rs6494029, rs653765, rs7174386, rs7182060, rs8027998, rs8043406
17	ACE	rs4305, rs4309, rs4311, rs4329, rs4343, rs4353, rs4362
	GRN	rs3785817
	THRA	rs1568400, rs3744805, rs7502966
	TNK1	rs2075760, rs6503018, rs7219773
19	APOE	rs405509
	EXOC3L2	rs10422797, rs346763
	GAPDHS	rs11882238, rs12151019, rs2239942
	LDLR	rs1433099, rs1799898, rs2228671, rs2569537, rs2569538, rs4508523, rs5930, rs6511720, rs688
20	CST3	rs2424577
	PRNP	rs12625444, rs2756271, rs6084833, rs6107516

Table C.1: Names of 493 SNPs in analyzed genes