Statistical Methods for Handling Missingness and Confounding in UK Biobank Population Brain Imaging



Lav Radosavljević

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2026

To Anja, Teodor, Vladana, Milan, Trpena, Branko, Vera and Miodrag.

Abstract

UK Biobank (UKB) brain imaging data is a one-of-a-kind resource for studying the links between the brain and demographic-, lifestyle- and genetic data. These links are often established via statistical modelling of the relation between Imaging Derived Phenotypes (IDPs) and non-IDPs (nIDPs). This type of analysis has been the source of multiple groundbreaking medical and epidemiological discoveries. There are, however, considerable challenges when establishing such links. This thesis focuses on solving two problems: missing data and confounding, for analysis linking IDPs and nIDPs. In the case of missing data, we identify the presence of Structured Missingness (SM) in the nIDPs, caused by the study design of UKB. We define a generative model for missing data in the nIDPs, and evaluate existing missing data methodology on synthetic data with access to the ground truth. Further, we develop methodology for performing Canonical Correlation Analysis (CCA) that is tailor-made to handle the type of missingness that we observed in the UKB brain imaging cohort. In the case of confounds, we create a fast and memory efficient pipeline for constructing UKB brain imaging confounds, generating a lower number of confounds compared to the previous, established pipeline (Alfaro-Almagro et. al. 2021). We also identify appropriate methodology for deconfounded inference when the number of confounds exceeds the number of subjects, enabling analysis of small subsets of subjects. The work presented in this thesis enables new types of analyses using the UKB brain imaging cohort, leading to deeper scientific understanding of the links between the human brain and everything around it.

Declarations

I declare that this thesis is entirely my own work and, except where stated, describes my own research under the supervision of professors Thomas E. Nichols and Stephen M. Smith.

- The work submitted in Chapter 1 has been published in the journal *BMC Medical Research Methodology* (URL: `https://link.springer.com/article/10.1186/s12874-025-02487-4`). It has also been presented at OHBM 2024 in Seoul, South Korea.
- The work submitted in Chapter 2 is under peer-review and is available as a preprint (URL: `https://www.medrxiv.org/content/10.1101/2025.10.09.25337581v1`).
- The work submitted in Chapter 3 is under peer-review and is available as a preprint (URL: `https://www.medrxiv.org/content/10.1101/2025.11.21.25340740v1`). The paper has a number of additional co-authors, whose names and contributions are listed in the preprint. It has also been presented at OHBM 2025 in Brisbane, Australia.
- The work submitted in Chapter 4 is soon to be submitted for peer-review. It has also been presented at OHBM 2025 in Brisbane, Australia.

Lav Radosavljević

January 2026

Acknowledgements

Firstly, I would like to thank both of my supervisors, Professors Thomas Nichols and Stephen Smith, for their help and commitment throughout the whole of this journey. Both of them have been truly marvellous supervisors.

I would also like to extend my thanks to the whole of my primary research group, NISOx, for their help, feedback, collaboration and friendship.

To my dear Anja, thank you for your support and kindness throughout my DPhil.

Equally, I thank my family; brother Teodor, mother Vladana, father Milan and grandmother Trpena, for their help and support.

I also thank the whole community of students from former Yugoslav republics at the University, for their friendship during my time as a student. And special thanks to my friend Kevin Wang for proof-reading this thesis.

Lastly, I would like to thank and acknowledge my funder, the EPSRC, for supporting my research (EP/S02428X/1).

Contents

1 Introduction **1**

1.1 UK Biobank 1

1.2 Brain Imaging 3

1.3 Missing Data 4

1.4 Confounding 6

1.5 Links Between Missingness and Confounding 9

2 A Generative Model For Evaluating Missing Data Methods in Large Epidemiological Cohorts **10**

2.1 Background 10

2.1.1 UK Biobank 12

2.2 Methods 12

2.2.1 Terminology 12

2.2.2 Parameters of the Generative Model 15

2.2.3 Estimating Parameters 16

2.2.4 Generating Synthetic Data 17

2.2.4.1 Calibrating the Predictability of Missingness 19

2.2.5 Simulation Study 19

2.2.6 Illustrative Example: Variable Selection for Total Grey Matter Volume on nIDP regression 23

2.3 Results 24

2.3.1 Analysis Pipeline 24

2.3.2 Simulation Study 30

2.3.3 Illustrative Example 30

2.4 Discussion . . . 33
2.5 Conclusions . . . 34

3 Accounting for Structured Missingness in Canonical Correlation Analysis 35
3.1 Introduction . . . 35
3.1.1 Canonical Correlation Analysis . . . 37
3.1.2 Confounding and Missingness . . . 37
3.1.3 Related Work . . . 39
3.2 Methods . . . 40
3.2.1 Other Methods . . . 43
3.2.2 Deconfounding . . . 44
3.2.3 Simulation Study . . . 44
3.2.4 Study on Real Data . . . 47
3.3 Results . . . 49
3.3.1 Simulated Data . . . 49
3.3.2 Real Data . . . 52
3.4 Discussion . . . 53

4 An Improved Pipeline for Constructing UK Biobank Brain Imaging Confounds 57
4.1 Introduction . . . 57
4.2 Background . . . 58
4.3 Methods . . . 59
4.3.1 Correcting R^2 after mean imputation . . . 60
4.3.2 Proposed Forward Selection Method . . . 60
4.3.3 Double Selection . . . 62
4.3.4 Proposed Pipeline . . . 62
4.3.5 Evaluating the New Pipeline . . . 63
4.4 Results . . . 66
4.5 Discussion . . . 69

5 Controlling for confounds in UK Biobank brain imaging data with small subsets of subjects **75**
5.1 Background . . . 75
5.1.1 UKB Imaging Confounds . . . 77
5.2 Methods . . . 77
5.2.0.1 Principal Component Regression . . . 77
5.2.0.2 Uncorrected Ridge Regression p-values . . . 78
5.2.0.3 Ridge Regression Permutation Testing . . . 79
5.2.0.4 Naive Feature Selection . . . 80
5.2.0.5 Double Selection . . . 81
5.2.0.6 De-Sparsified LASSO . . . 81
5.2.1 Simulation Method . . . 82
5.2.2 Real Data Demonstration . . . 84
5.3 Results . . . 84
5.3.1 Simulation Results . . . 84
5.3.2 Real Data Results . . . 93
5.4 Discussion . . . 96

6 Conclusions **98**

A A Generative Model For Evaluating Missing Data Methods in Large Epidemiological Cohorts **102**
A.1 Supplementary Proof . . . 102
A.2 Supplementary Figures . . . 107

B Accounting for Structured Missingness in Canonical Correlation Analysis **121**
B.1 Justification of Approximation . . . 121
B.2 Simulating Data . . . 123
B.3 CCA: Remaining Five Analyses . . . 124

C An Improved Pipeline for Constructing UK Biobank Brain Imaging Confounds **135**
C.1 Proof of consistency of $\hat{R}^2_{\text{cor}}$ . . . 135

C.2 Algorithm 139
C.3 Further Comparisons of Old and New Confounds 141
C.4 Comparing Double Selection and Single Selection 144

D Controlling for confounds in UK Biobank brain imaging data with small subsets of subjects **148**
D.1 Supplementary Figures 148

Bibliography **161**

Chapter 1

Introduction

1.1 UK Biobank

UK Biobank (UKB) is a large prospective cohort study involving just over half a million UK adults, aged 40-69 at the time of entry. The project started collecting data in 2006 and has since acquired demographic-, genetic-, metabolic-, health record-, and imaging data, along with many other smaller substudies and questionnaires. Table 1.1 shows the timeline of data collection for some of the most important UKB substudies, along with the envisioned initial sample size (Sudlow et al., 2015; Doherty et al., 2017; Littlejohns et al., 2020; UK Biobank, 2026; Foster et al., 2023).

Table 1.1: Timeline for data collection of major UK Biobank substudies, along with the approximate envisioned sample size at the start of collection. As we see, the number of participants in each substudy can vary greatly, and the collection of data spans across decades.

Substudy	**Start of Collection**	**Envisioned Sample Size**
Baseline Assessment	2006	∼500,000
Accelerometry	2013	∼100,000
Multimodal Imaging (initial imaging)	2014	∼100,000
Whole-Genome Sequencing	2018	∼ 500,000
Multimodal Imaging (repeat imaging)	2022	∼ 60,000

The goals of UKB are to be a state of the art resource for health data research

that deepens our understanding of population health, and leads to the development of new tools for diagnosis and early detection of disease. This type of research is particularly important as the UK and indeed World population is rapidly aging, with the world population over 65 expected to reach 1.5 billion by 2050 (Navaneetham and Arunachalam, 2025), leading to a great increase in healthcare needs, and making health research a more important global priority. Examples of groundbreaking medical and epidemiological discoveries made using UKB as a resource include: finding 44 new genetic risk variants for Major Depressive Disorder (Wray et al., 2018), identifying grip strength as a key biomarker for all-cause mortality, pulmonary- and cardiovascular disease (Celis-Morales et al., 2018), identifying the effect of COVID-19 infection on the brain (Douaud et al., 2022) and developing a Machine Learning (ML) tool for identifying Osteoarthritis based on bone density scans (Faber et al., 2022). As of August 2025, over 22,000 researchers in over 60 different countries have access to UKB data, and since the initial data release in 2012, there have been over 18,000 peer-reviewed articles published using UKB data (UK Biobank, 2025).

UKB is an invaluable resource for the global research community and it is therefore a key task to develop statistical methodology that is tailor-made for use on UKB data. This is the aim of this thesis. Our particular focus is the UKB brain imaging cohort, a substudy of UKB consisting of now roughly 20% of all participants. Examples of groundbreaking research using this data is the aforementioned study on the effect of COVID-19 on the human brain, an Artificial Intelligence (AI) tool used to identify early signs of Alzheimer's Disease (AD) in healthy UKB subjects (Azevedo et al., 2023) and a Genome Wide Association Study (GWAS) of brain measures, providing new important insights into the genetic architecture of the brain (Elliott et al., 2018). Along with brain imaging data, this thesis will include the use of non-imaging data sources, including: NMR-metabolomics data (Julkunen et al., 2023), which consists of 249 measures of lipids, fatty acids and small molecules measured from blood plasma samples, refractrometry data (Sudlow et al., 2015), which includes measurements of how light bends as it passes through the human eye, and food preferences data UK Biobank (2020), which detail the preferences of participants for different food items on a scale of 1 to 9. The reason for such a strong presence of non-imaging data alongside imaging data in this thesis is that our main aim is to develop statistical

methodology which enables the linking of imaging and non-imaging phenotypes in a way which deepens our understanding of their associations.

1.2 Brain Imaging

UKB started collecting Magnetic Resonance Imaging (MRI) based brain imaging data in 2014, with the aim of scanning a subset of around 100,000 volunteers from the study (Alfaro-Almagro et al., 2018). One of the main goals of this cohort is to provide neuroimaging researchers with a large sample of data, enabling the discovery of population level biomarkers for neurodegeneration, and overall brain health. Previous research to this end has been severely limited in terms of statistical power by the typically low sample sizes (tens or hundreds) in brain imaging studies (Miller et al., 2016). Additionally, since UKB contains a wide assortment of other data, the brain imaging cohort enables researchers to link the brain with non-brain variables, such as genetics and National Health Service (NHS) health records. The data consists of three different MRI imaging modalities: structural, diffusion (dMRI) and functional (fMRI), each with their own research purposes. Structural MRI captures information about the anatomical structure of tissues in the brain. dMRI measures the random diffusion of water molecules which reflects the microstructure of the brain tissue. fMRI measures brain activation, and is useful to see the brain response to tasks and stimuli, as well as measuring functional connectivity. The brain images themselves are useful for population health research, but usually require expert understanding. Additionally, voxel-wise analysis suffers from reduced statistical power resulting from the high number of voxels in a single image, because of the problem of correcting for multiple testing. Therefore, a pipeline has been developed to extract over 5,000 Imaging Derived Phenotypes (IDPs) from brain images. The IDPs are biologically much more easily interpretable than a voxel-level brain image, and greatly reduce the problem of multiple testing. An example of a structural MRI IDP is total brain volume, an example of a dMRI IDP is white-matter tract diffusivity, and an example of an fMRI IDP is the median Z-statistic for activation in a particular region of the brain.

A major research objective using these IDPs is to study them longitudinally as the cohort ages, and also to link them with non-imaging related variables (non-IDPs or nIDPs). Examples of nIDPs are NHS health records, food and lifestyle variables, physical measurements and genetic data. From a public health aspect, linking IDPs and nIDPs leads to deeper understanding of how to foster brain health on a population level. We have already mentioned groundbreaking studies linking the brain with COVID-19 and genetics, both using IDPs, and there have been numerous other studies that have given us important insights into how lifestyle affects brain health. Examples include: showing that even moderate alcohol consumption is associated with changes in brain health (Topiwala et al., 2017), showing association between obesity and gray matter (Dekkers et al., 2019), and finding a non-linear association between sleep duration and some structural IDPs, with brain health peaking at approximately 7 hours of sleep (Li et al., 2022). It is precisely this type of research that this thesis is aimed at; to develop methodology for robust inference linking IDPs and nIDPs. We focus on two challenges for users of the UKB brain imaging cohort: missing data and confounding. Our aim is to provide neuroimagers with a set of methods that they can use to handle these problems robustly, enabling further, more powerful research into the links between the brain and other health variables.

1.3 Missing Data

The problem of missing data is ubiquitous in the field of data science, and has been studied extensively since the dawn of statistical modelling from observational data. One of the earliest works to approach this problem is in a paper by Allan and Wishart (1930), which proposes a solution for performing Analysis of Variance (ANOVA) for testing treatment effects on crops, using data where one of the plots has been damaged or data has failed to be recorded. The authors treat the expected value of the missing observation as a variable and then perform ANOVA with one degree of freedom fewer. Yates (1933) extends this approach to handle multiple missing observations. In the decades following these initial examples, there are many other works in which specific missing data problems are handled via ad-hoc solutions (Finney, 1947; Healy and Westmacott, 1956; Anderson, 1957; Buck, 1960). However, generally little attention

was paid to the process that caused the missing data itself, and how this process could affect downstream analysis. In Rubin (1976), missing data was for the first time divided into groups depending on the process by which the missingness is generated. This initial taxonomy was amended in Marini et al. (1980) and finally formalised in Rubin (1987) by sorting missing data into three now famous categories: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR).

Looking at the modern day, with big datasets of ever growing size and complexity, we see that Rubin's taxonomy is insufficient to fully describe the different generating processes that can affect the analysis of incomplete data. Looking at tabular data from UKB in particular, we see a very distinct pattern to the missing data, resulting from its study design. The UKB has a set of baseline variables that are collected at the very first visit. In addition, it has many follow-up studies and questionnaires that not all subjects participate in. When a subject does not participate in a follow-up study, all variables from the follow-up study are recorded as missing for this subject. This creates large blocks of missing data for groups of subjects and variables. It is also possible that the characteristics of subjects that participate in follow-up studies are not the same as the cohort as a whole. Mitra et al. (2023) calls this type of missing data Structured Missingness (SM). There is very little work exploring how to best handle SM in the literature, so neuroimaging researchers using the UK Biobank have little guidance on how to handle this problem correctly.

The by far most common approach to handling missing data in large epidemiological cohorts is imputation of some variety, including Multiple Imputation (MI) and Single Imputation (SI) (Pedersen et al., 2017). Complete case analysis, i.e., using only samples with complete data, is also a fairly common approach, but has the disadvantages of leading to a reduced sample size, and also introducing bias in the data (Okpara et al., 2022). Indeed, for sufficiently extensive studies with many variables, it is easy to imagine a situation where there is not a single participant without at least some missing data.

We aim, in this thesis, to explore the way SM is generated in the UK Biobank, and identify methodology that is most suitable for handling SM in standard analysis pipelines for population neuroimaging data. This includes the development of

missing data methodology designed specifically to handle SM. Our specific contribution towards solving this problem is first creating a generative model which plausibly describes the process by which missingness is created in the nIDPs. This model is used for generating synthetic data to evaluate the performance of commonly used missing data methods on data with SM. Secondly, we develop a method for performing Canonical Correlation Analysis (CCA) in the presence of SM. CCA is a method commonly used to link multiple variables from two different substudies, making it a suitable place for developing methodology tailor-made for SM, as a number of subjects will have participated in both substudies, and a number of subjects will only have participated in one of them, leading to SM.

1.4 Confounding

When drawing conclusions about the relationship between an outcome variable y and an exposure variable x, it is possible to draw incorrect conclusions by merely examining the bivariate association between x and y. This is because the relationship between x and y can be induced by a third *confounding* variable z. A classical example of confounding in epidemiological research is a paper by Stark and Mantel (1966), which observed a statistically significant association between the risk of Down's Syndrome and sibling birth order, but the significance of this association disappeared after adjusting for maternal age. It is clear, in this simple case, that the bivariate association between birth order and the risk of Down's Syndrome is mostly a result of confounding by maternal age at birth, as higher birth order typically also implies a higher maternal age. To clarify, since different fields use different nomenclature, the terms "outcome variable" and "exposure variable" refer to two variables, one of which (exposure variable) we hypothesise has a predictive or causal relationship with the other (outcome variable). Equivalent or very similar terminology used in different fields are "target and feature", "dependent and independent", and "endogenous and exogenous".

Another phenomenon closely related to confounding is *Simpson's Paradox*, where the direction of an association is reversed after adjusting for a (usually discrete) covariate. A commonly used example is a paper by Bickel et al. (1975), which studies

the sex-bias in graduate admission rates for the University of California, Berkeley. The admissions rate is statistically significantly higher for men, but after adjusting for the department to which the application was made, a slight yet statistically significant bias in the opposite direction is observed, i.e., a bias in favour of female applicants. This is phenomenologically very similar to classical confounding, but the nuisance covariates here cause the true associations to flip direction, rather than inducing a spurious correlation. Indeed, in standard linear modelling, both problems are handled via simply including the confounds, along with the outcome and exposure variables in a multivariate Ordinary Least Squares (OLS) model.

Controlling for confounding effects is a crucial part of epidemiological research, as it prevents the drawing of incorrect conclusions from uncorrected association analysis. When choosing which confounds to control for, epidemiologists typically rely on subject matter expertise, for example from medical doctors, to define a set of variables which could plausibly act as confounders (Hernán et al., 2002). Using subject matter knowledge is crucial for understanding the causal relationships between variables, ensuring that for example *collider variables*, i.e., variables caused by both outcome and exposure, are not treated as confounders, as this can lead to spurious associations. The most typical example of this is *Berkson's Paradox*, where the collider variable is a participants' inclusion status in the study population. Elwert and Winship (2014) uses the example of studying the relationship between beauty and talent amongst Hollywood actors: while beauty and talent are likely independent, they both contribute to becoming a Hollywood actor, so conditioning upon Hollywood actor status, one would observe a spurious negative association between the two variables, due to selection bias. Another common challenge in controlling for confounds is imperfectly measured or noisy confounds, which are an unexpectedly challenging problem, sometimes causing a near 100% false positive rate (Westfall and Yarkoni, 2016). After selecting a viable candidate pool of confounds, it is also common to use statistical methods to further reduce confound sets, while still properly controlling for them (Belloni et al., 2014; Zhao et al., 2021).

In population neuroimaging in particular, the image acquisition protocol is likely the most important source of nuisance, as Quality Control (QC) variables have been shown to be associated with both IDPs and nIDPs (Alfaro-Almagro et al., 2021).

Previous work on UK Biobank brain imaging confounds has produced a pipeline that constructs ≈ 1000 confounds that are selected because they explain non-trivial variance in the IDPs. Deconfounded inference is performed by including these confounds in an OLS model along with the outcome and exposure of interest. There are, however, two significant limitations to this confound-generation pipeline. Firstly, it has a very long run-time and a high memory footprint. Secondly, there is redundancy in the confound set. Because the selection process is univariate, multiple highly correlated confounds can be selected, despite not having unique contribution in explaining variance in the IDPs. Therefore, the number of confounds is unnecessarily high. This creates a problem for analyses that involve a small subset of subjects, as the number of confounds can easily exceed the number of subjects, making standard de-confounded inference using OLS problematic. We therefore want to solve two problems in particular:

1. Improving the existing UKB imaging confounds pipeline to make it faster and more memory efficient.

2. Solving the problem of too many confounds. This includes improving the current pipeline to reduce redundancy in the confound set, and also exploring methods for deconfounded inference when the number of confounds exceeds the number of subjects.

Our first specific contribution to this is implementing a faster version of the confounds pipeline based on multivariate, rather than univariate feature selection. The new pipeline only requires 1 CPU and around 1 hour of runtime. It also greatly reduces the number of confounds (while explaining almost the same total variance space). Our second contribution is the evaluation of a large number of methods for deconfounded inference designed to handle a larger number of confounds. We provide a set of recommendations for neuroimaging researchers indicating which high-dimensional inference method is most appropriate to use at given sample sizes.

1.5 Links Between Missingness and Confounding

In this thesis, we consider both the problem of informative missingness and confounding in UK Biobank population neuroimaging. The reason for this joint focus is that the two problems are related in multiple ways, and it is therefore meaningful to consider them in the same body of work. The most obvious link between the two problems is that many typical confounders, such as baseline health and demographic characteristics (age, sex, socio-economic status etc.) are also associated with study participation, and therefore informative missingness (Fry et al., 2017). This makes it impossible to consider structured missingness without also considering confounds. Indeed, in Chapter 2 and 3 of this thesis, which are about missing data methodology, we assume that variables from the baseline assessment determine structured missingness. There are also other, less obvious links. For example, the faster pipeline for constructing imaging confounds presented in Chapter 4 is primarily an innovation in handling missingness in the IDPs, not in confound methodology itself. Throughout the thesis, it is obvious that there is a strong link between missingness and confounding which is inherent to UK Biobank data, as well as population health data more generally.

Chapter 2

A Generative Model For Evaluating Missing Data Methods in Large Epidemiological Cohorts

2.1 Background

Missing data is common in epidemiological and health data research and presents formidable challenges for many analytical approaches. The causes of missing data vary, from being inherent to the study design, to elective non-participation, or simply faults in measurement. Much work has therefore been devoted to evaluating the performance of methods for handling missing data. The most common approaches to comparing imputation methods include simulating data and inducing missingness using *a priori* chosen mechanisms (Le Morvan et al., 2020; van Loon et al., 2022). Alternatively, artificial missingness is induced on real, complete data (Jarrett et al., 2022; Ghalebikesabi et al., 2021; Beaulieu-Jones et al., 2018) or real missingness patterns are imposed on simulated data (Gentry et al., 2023). Simulation studies of this kind usually rely on quite restrictive assumptions that might not be reflective of large scale epidemiological cohorts such as UK Biobank (UKB). For example, while some studies induce missingness in a manner which lacks any clear structure (Le Morvan et al., 2020; Ghalebikesabi et al., 2021), in UKB missing data caused by non-participation

in a sub-study/questionnaire comes in “blocks". Specifically, if a subset of participants do not participate in an extension of the core study, then all of these subjects will have missing entries for the variables of this extension, and (when the rows and columns of the subjects-by-variables matrix are suitably reordered) this will form a solid block of missing data. In Mitra et al. (2023), this type of “blocky" missingness is categorised as *structured missingness* (SM). Expanding on this way of thinking, we refer to missing data lacking structure as *unstructured missingness* (UM). Since UKB and similar datasets consist of many different sub-studies and questionnaires, this is a crucial feature to consider when evaluating the performance of imputation methods.

While there has been work done on the evaluation of existing methods on data with structured missingness and the development of new methods for handling such data, it has been common to use a non data-driven method for inducing structured missingness (van Loon et al., 2022; Jackson et al., 2023; Little et al., 2014) or use real missingness patterns imposed on simulated data (Gentry et al., 2023). Assuming that SM is Missing Completely at Random (MCAR) is especially problematic in the case where it is caused by non-participation, since we know that participants often are disproportionately healthy (Fry et al., 2017), meaning that the data is not likely to be MCAR. Our aim is therefore to define a tool for generating synthetic data which has the same properties as a given data set. We want the pattern to satisfy the following three criteria:

1. There is structured missingness, blocks of missingness caused by non-participation in sub-studies, as well as unstructured missingness that is not in blocks and is attributable to any other cause.

2. Missingness is informative in the sense of MAR (Missing at Random), where there is a relationship between missingness in a given variable and the observed elements of other variables.

3. There is an association between inter-variable correlation and inter-variable missingness similarity, typically where tightly correlated variables are more likely to be jointly missing.

With our framework for simulating such synthetic data, we evaluate the performance of several imputation methods. While the general motivation of our work is developing

a tool for evaluating the accuracy of missing data methods for epidemiological data combined from different sub-studies, we are specifically interested in the problem of missing data when modelling relationships between brain imaging variables and health, demographic, behavioural and lifestyle variables in UK Biobank. We thus focus on the subset of 46 471 subjects with imaging derived phenotype (IDP) data, and a collection of 23 871 non-Imaging Derived Phenotypes (nIDPs) variables; these nIDP variables are a mixture of continuous and binary variables (some of the binary variables are 1-hot encodings of categorical variables). The illustrative study in subsection 2.2.6 is an example of tasks that involve imputing missing nIDP data with the final goal of creating a model describing the association between total grey matter volume (which is an IDP) and several nIDPs.

2.1.1 UK Biobank

UK Biobank is a biomedical database and epidemiological cohort study combining, among other things, demographic, genetic, medical, dietary, imaging, biomarker and questionaire data from around half a million subjects residing across the United Kingdom (UK Biobank, 2024; Bycroft et al., 2017; Miller et al., 2016; Allen et al., 2012; Julkunen et al., 2023; Collins, 2012). The brain imaging sub-cohort consists of roughly $\approx 50\,000$ subjects and is our object of interest in this work. IDPs for the brain imaging cohort are variables extracted from Magnetic Resonance Imaging (MRI) that inform various aspects of brain function and structure. Example IDPs include cortical volume, total grey matter volume and white matter hyperintensity volume. nIDPs on the other hand are demographic, health, cognition and lifestyle related variables from diverse sources that are deemed interesting in neuroimaging research.

2.2 Methods

2.2.1 Terminology

Let n and d be the number of subjects and variables respectively, $\mathbf{X}$ be our $n \times d$ dataset and $\mathbf{M}$ be the $n \times d$ missingness matrix where $M_{ij} = 1$ if variable j is missing

for subject i and $M_{ij} = 0$ if it is not missing. The following definitions and notation are central to our work:

1. **Variable-wise missingness pattern**. For any variable $j = 1, 2, ..., d$, the *variable-wise missingness pattern* for variable j is $\mathbf{m}_j^{\text{v}} = (M_{1j}, M_{2j}, ..., M_{nj}) \in \{0, 1\}^n$.

2. **Subject-wise missingness pattern**. For any subject $i = 1, 2, ..., n$, the *subject-wise missingness pattern* for subject i is $\mathbf{m}_i^{\text{s}} = (M_{i1}, M_{i2}, ..., M_{id}) \in \{0, 1\}^d$.

3. **Variable-wise missingness distance.** For any two variables j and j', the *variable-wise missingness distance* between them is the proportion of discordant missingness indicators

$$D_{jj'}^{\text{v}} = \frac{1}{n} \sum_{k=1}^{n} 1 \left\{ M_{kj} \neq M_{kj'} \right\},$$

   where $\mathbf{D}^{\text{v}} = D_{jj'}^{v}$ is the $d \times d$ variable-wise missingness distance matrix.

4. **Subject-wise missingness distance.** For any two subjects i and i' the *subject-wise missingness distance* between them is likewise

$$D_{ii'}^{\text{s}} = \frac{1}{d} \sum_{k=1}^{d} 1 \left\{ M_{ik} \neq M_{i'k} \right\},$$

   where $\mathbf{D}^{\text{s}}$ is the $n \times n$ subject-wise missingness distance matrix.

5. **Structured missingness (SM)**. We call missingness that is caused by non-participation in a sub-study/questionnaire *structured missingness*, resulting in a subset of subjects having missing data for a set of variables. This is also sometimes called block-wise missingness, as when subjects and variables are suitably reordered, this will result in solid blocks of missing data in the data matrix.

6. **Unstructured missingness (UM)**. We call missingness that is not caused by non-participation in a study/questionnaire *unstructured missingness*. This type of missingness does not induce any blocks of missing data.

We now define the stochastic mechanisms that can give rise to missing data. Let $\mathbf{x}$ be a d-dimensional random vector drawn from the same distribution as the data in our data set and $\mathbf{m}$ be its corresponding subject-wise missingness pattern. Let further $\mathbf{x}_{\text{obs}(\mathbf{m})}$ and $\mathbf{x}_{\text{miss}(\mathbf{m})}$ be the observed and unobserved parts of the random vector $\mathbf{x}$ respectively. We follow the terms in Rubin (1976) for different types of missingness:

- *Missing Completely at Random* (MCAR)

$$\mathbb{P}(\mathbf{m}|\mathbf{x}) = \mathbb{P}(\mathbf{m}).$$

  This means that the missingness mask is completely independent from underlying data.

- *Missing at Random* (MAR)

$$\mathbb{P}(\mathbf{m}|\mathbf{x}) = \mathbb{P}(\mathbf{m}|\mathbf{x}_{\text{obs}(\mathbf{m})}).$$

  This means that there exists some dependence between the missingness mask and the underlying data, but that this relationship can be described using only observed data, i.e., the relationship between $\mathbf{m}$ and $\mathbf{x}$ is determined exclusively by the observed part $\mathbf{x}_{\text{obs}(\mathbf{m})}$. For example, this can mean that there exists a group of variables with no missingness which determine the missingness mask $\mathbf{m}$.

- *Missing Not at Random* (MNAR)

$$\mathbb{P}(\mathbf{m}|\mathbf{x}) \neq \mathbb{P}(\mathbf{m}|\mathbf{x}_{\text{obs}(\mathbf{m})}).$$

  For this type of missingness, the relationship between the data and the missingness requires knowledge of data obscured by missingness. This is the most

difficult setting to handle, assuming no prior knowledge of the mechanism by which missingness is induced, since it has been shown that for any MNAR model explaining missing data in a given data set, there exists an MAR model with equal evidence (Molenberghs et al., 2008). In other words, there can be no theoretical guarantees of correctness for MNAR models explaining missing data barring direct knowledge of the missingness mechanism.

Characterising types of missingness is crucial to our work since many methods of handling missing data, most notably Multivariate Imputation by Chained Equations (MICE), have theoretical guarantees under MCAR and MAR (Azur et al., 2011), while MNAR requires additional assumptions (Van Buuren and Groothuis-Oudshoorn, 2011).

2.2.2 Parameters of the Generative Model

We assume that our data consists of C different sub-studies, where study $c = 0$ is assumed to be a baseline study with no missingness while studies $c = 1, ..., C-1$ are follow-up sub-studies with both structured and unstructured missingness. The following parameters define our generative model:

- $\{\rho\}_{c,c'}$, the distribution of between-variable correlations for all pairs of sub-studies c, c'. We assume a mixed data generative model (Demirtas and Doganay, 2012), where data arise from a multivariate normal distribution with zero means and unit variances: for continuous variables these values are directly observed, while for binary variables the normal variate is latent and the data is obtained by thresholding continuous variables to 0/1. Therefore, $\{\rho\}_{c,c'}$ represents the correlation distribution of the underlying data prior to thresholding.

- π_c the rate of structured missingness for sub-study c.

- (α_c, β_c), parameters governing the rate of unstructured missingness for each variable. We assume that the rate of unstructured missingness p_j^{um} is drawn from Beta(α_c, β_c) if feature j belongs to sub-study c.

- Σ_{core}. We assume that d_{core} variables from the baseline study $c = 0$ determine all structured missingness through a logistic model. Σ_{core} is the correlation matrix of these core variables. The core variables are assumed to all be continuous.
- AUC_c, the Area Under the Curve (AUC) score of the logistic model determining structured missingness for sub-study c.

2.2.3 Estimating Parameters

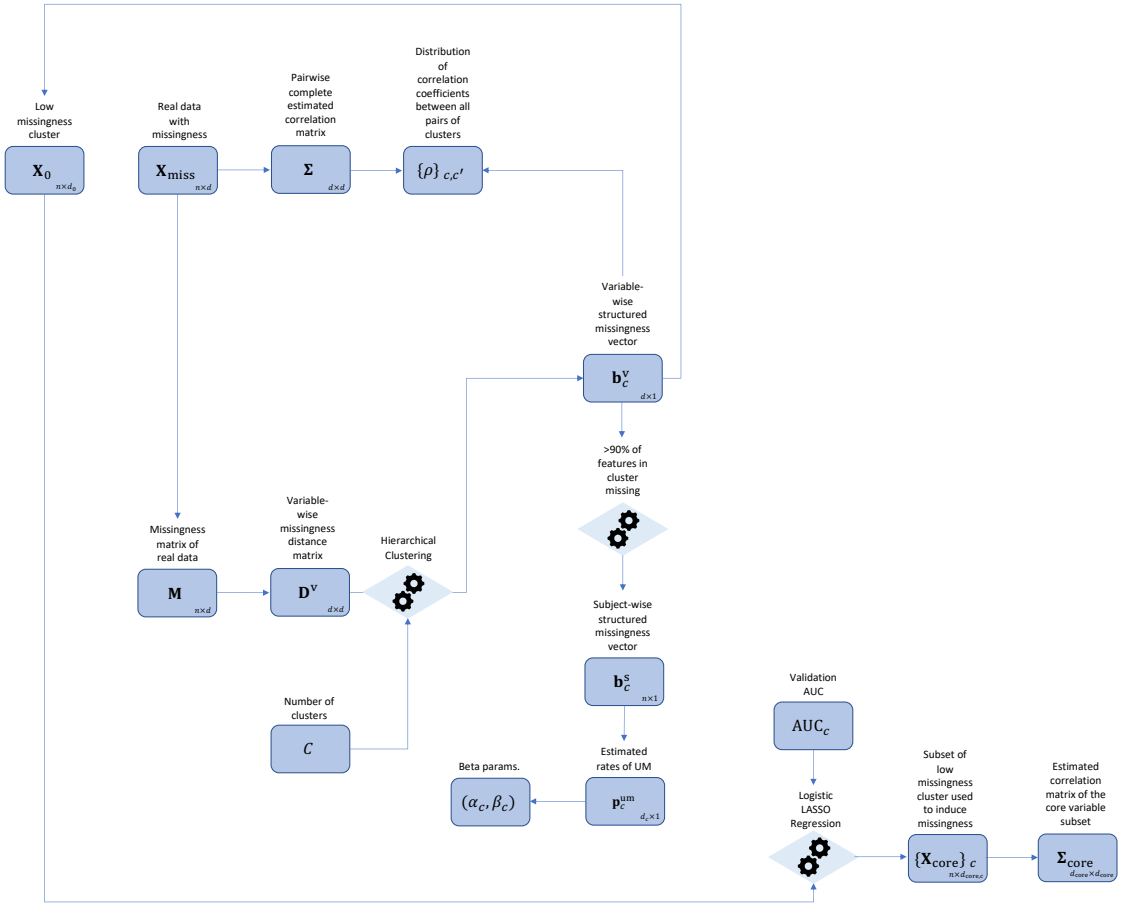


Figure 2.1: Flow chart of the data analysis pipeline. The initial input is the number of clusters C. Hierarchical clustering is used to then select C clusters of variables based on missingness patterns. After this, we identify structured missingness patterns for each cluster, and the level of informativeness of the same, as measured by the AUC of a prediction model for structured missingness as a function of baseline variables. We also identify the rates of structured and unstructured missingness for each cluster, and the distribution of variable correlations within and between clusters.

We estimate the parameters of the model using the following procedure, also detailed in the flowchart in Figure 2.1. To avoid confusion, we refer to $c = 0, 1, ..., C-1$ as "clusters" from here on out rather than "sub-studies", as a single cluster c may contain several similar sub-studies depending on the size of C that we choose.

1. C clusters are identified using hierarchical agglomerative complete linkage clustering (Sorensen, 1948).

2. The densities $\{\rho\}_{c,c'}$ are estimated using a histogram for each pair of clusters c, c'.

3. We define a subject i to be structurally missing for a cluster c if at least 90% of the variables from c are missing for subject i. This will give us the vectors $\mathbf{b}_c^{\text{s}} \in \{0,1\}^n$ where $\mathbf{b}_{c,i}^{\text{s}} = 1$ if subject i is structurally missing for cluster c. This result also directly gives us π_c, i.e., the probability of a subject having structured missingness for cluster c.

4. Having identified all structured missingness, we estimate (α_c, β_c) using the method of moments on the remaining, unstructured missingness.

5. We use LASSO Logistic Regression (LASSO-LR) (Tibshirani, 1996) to simultaneously identify the core variables that determine structured missingness and AUC_c, by fitting $C-1$ penalised logistic regression models that use the baseline study data $\mathbf{X}_0$ as predictors and the subject-wise structured missingness vectors $\mathbf{b}_c^{\text{s}}$ as outcomes. Specifically, AUC_c is estimated using 5-fold cross validation. Note that the core variables are cluster specific and may or may not overlap for different clusters. We then estimate Σ_{core}, the correlation matrix of the core variables for all clusters. The penalty term λ_c for each LASSO models can be chosen in multiple appropriate ways (see subsection 2.2.4.1).

2.2.4 Generating Synthetic Data

The data is generated using a step-wise procedure as seen in Figure 2.2. Since all continuous variables have unit variance correlation matrices and covariance matrices of continuous data are the same.

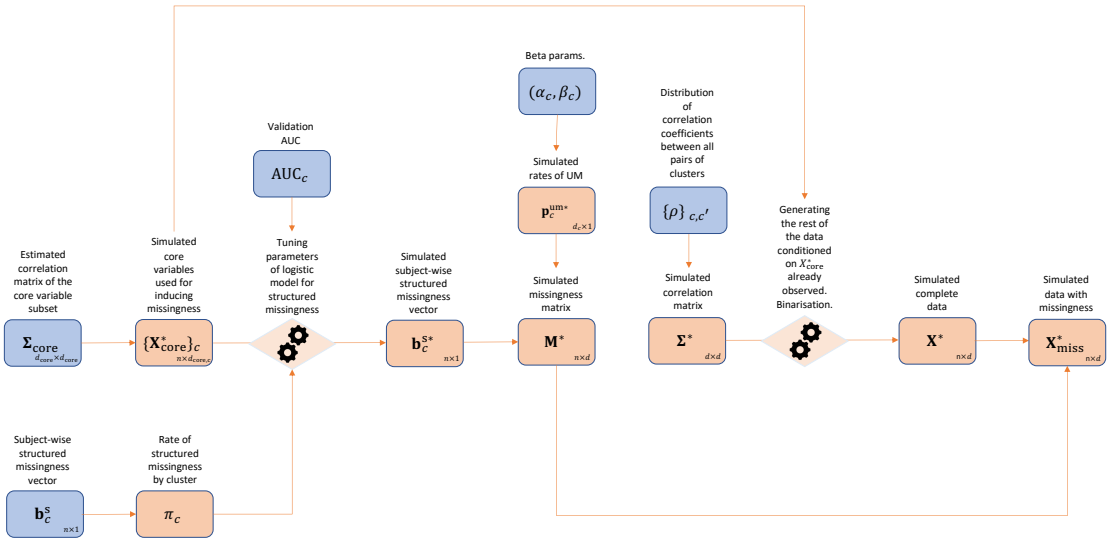


Figure 2.2: Flow chart of the synthetic data generation pipeline.

1. Using Σ_{core}, we simulate the cluster specific core variables $\{\mathbf{X}^*_{\text{core}}\}_c$ for clusters $c = 1, ..., C-1$, by drawing the full core variable data matrix $\mathbf{X}^*_{\text{core}}$ from $\mathcal{N}(\mathbf{0}, \Sigma_{\text{core}})$.

2. Using a binary search procedure, we determine intercepts and coefficients of $C-1$ logistic models determining structured missingness, with $\{\mathbf{X}^*_{\text{core}}\}_c$ as covariates, such that the model's AUC score and rate of positive cases will match AUC_c and π_c. All the coefficients of the logistic models are assumed to be equal. Using these models, we generate synthetic subject-wise structured missingness vectors $\mathbf{b}_c^{\text{S}*}$.

3. We generate the rates of unstructured missingness $p_j^{\text{um}*}$ by drawing them independently from $\text{Beta}(\alpha_c, \beta_c)$ where variable j is in cluster c. Unstructured missingness is assumed to be MCAR and is induced for each subject i and variable j with probability $p_j^{\text{um}*}$. By combining the generated structured and unstructured missingness, we obtain an $n \times d$ synthetic missingness indicator matrix $\mathbf{M}^*$.

4. We simulate the full $d \times d$ correlation matrix Σ^* by drawing its entries from $\{\rho\}_{c,c'}$ and, if necessary, projecting it to the nearest positive definite correlation matrix using Higham's algorithm (Higham, 2002).

5. Having the complete correlation matrix, we generate the rest of the data $\mathbf{X}^*$, conditioned on the already simulated core variable data $\mathbf{X}^*_{\text{core}}$. To allow for binary variables, we threshold a subset of variables to become binary, corresponding to the same number of binary variables in each cluster.

6. Finally the synthetic missingness mask $\mathbf{M}^*$ is imposed upon $\mathbf{X}^*$ to obtain the corresponding synthetic data set with missingness $\mathbf{X}^*_{\text{miss}}$.

This procedure generates synthetic datasets which satisfy the key criteria outlined in the introduction, including that the datasets fall under MAR. Crucially, we have access to the true mean vector and covariance matrix, as well as the underlying data obscured by missingness.

2.2.4.1 Calibrating the Predictability of Missingness

The choice of penalty term λ_c for each LASSO-LR model that predicts structured missingness can most easily be made by selecting the value of λ_c which minimises validation loss. This is, however, not always the option which is most faithful to the assumptions of our generative model. Since our generative model assumes that all core variables have equal importance in predicting structured missingness, we want to choose a value of λ_c which will minimise the number of core variables of low predictive importance, while having a validation loss that is close to that of the optimal value. This is an inevitably arbitrary feature of our generative model and we choose to select a reasonable value of λ_c through trial and error. Specifically, we fit the LASSO model for several penalty terms and compare the validation AUC between them. We then choose the strongest penalty which produced an AUC comparable to the maximum, as this greatly reduces the number of variables, while preserving informativeness. Tables 2.2 and 2.3 show that the number of selected variables can be greatly reduced while preserving a similar validation AUC.

2.2.5 Simulation Study

In order to demonstrate the use of our generative model, we conduct a simulation study on synthetic data mimicking the UK Biobank Brain nIDPs to evaluate the

performance of three commonly used imputation methods on this data set. Our study tests the accuracy of imputation for missing data entries, as measured by Mean Squared Error (MSE) for continuous variables and Balanced Accuracy (BA) for binary variables. Accuracy is recorded by individual variable across $B = 20$ synthetic data sets. Specifically, if we, for some variable j with n_{miss} missing values, have imputed values x_{imp} and corresponding true values x_{true}, MSE and BA are calculated according to the following formulae:

$$\text{MSE} = \frac{1}{n_{\text{miss}}} \sum_i (x_{\text{imp},i} - x_{\text{true},i})^2$$

$$\text{BA} = \frac{1}{2} \frac{\sum_{i|x_{\text{true},i}=0} 1\{x_{\text{imp},i} = x_{\text{true},i}\}}{\sum_{i|x_{\text{true},i}=0} 1} + \frac{1}{2} \frac{\sum_{i|x_{\text{true},i}=1} 1\{x_{\text{imp},i} = x_{\text{true},i}\}}{\sum_{i|x_{\text{true},i}=1} 1}.$$

Additionally, to illustrate the difficulty of imputing data with SM, we perform the same simulation study on copies of the synthetic datasets where missingness has been induced in an MCAR and completely unstructured manner. The data in each cluster is set to be missing using independent Bernoulli variables with probability equal to the total rate of missingness for the cluster.

The first imputation method is mean imputation, which serves as our benchmark method. The second is the matrix completion method SoftImpute (Mazumder et al., 2010), which assumes that there exists a low rank approximation of the data set. This method has a tuning parameter, the value of the low rank, which we vary as 5%, 15% and 30% of the full matrix rank. Both the mean imputation and SoftImpute methods are binarised to impute binary variables using 0.5 as the threshold, so imputed values greater than or equal to 0.5 are transformed to 1, while the rest are transformed to 0. The last method is called iterative imputation (Pedregosa et al., 2011) or ICE, Iterative Imputation by Chained Equations (Jarrett et al., 2022), which uses the same iterative procedure as MICE, but does not include randomness in the imputed values and only creates a single imputed data set. As we are interested in evaluating the imputation accuracy of missing values themselves, we evaluate this method rather than MICE, since it follows the same iterative procedure, but does not add noise to the imputed values like MICE does. MICE is also impractical to use in this

high dimensional setting with respect to memory use and computational time since it requires a large number of multiple imputed data sets. We choose to impute continuous values using Bayesian Ridge Regression (Pedregosa et al., 2011) and binary values using Logistic Regression with a Ridge penalty. In high dimensional setting, iterative imputation requires us to choose a subset of $k << d$ variables that are used to impute each variable j. These variables are normally set to be the k variables with highest absolute correlation with j (Van Buuren and Oudshoorn, 1999; Pedregosa et al., 2011; Graham, 2009), or the k variables with the most favorable missingness patterns (Van Buuren and Oudshoorn, 1999); all else being equal, we favour variables for imputing j which are observed the most often in rows where j is missing and therefore select variables using the rows of the matrix

$$\mathbf{V} = \mathbf{M}^T(\mathbf{1}_{n\times d} - \mathbf{M}), \tag{2.1}$$

where $V_{jj'}$ is the number of times j' is observed when j is missing.
We propose a third selection method which utilises correlation and missingness jointly, while being applicable to mixed data. It calculates a score $S_{jj'}$ which is proportional to the maximum expected reduction imputation error (MSE for continuous and misclassification rate for binary variables) under the assumption of MCAR and under the generative model described in Demirtas and Doganay (2012) for joint continuous and binary data, i.e., an underlying multivariate normal distribution with thresholding for binary variables.

- **j and j' are both continuous**

  $$S_{jj'} = V_{jj'}\rho^2,$$

  where ρ is the Pearson correlation between variables j and j'.

- **j is continuous and j' is binary**

  $$S_{jj'} = V_{jj'}\rho^2,$$

  where ρ is the Pearson correlation between variables j and j'.

- **j is binary and j' is continuous**

$$S_{jj'} = V_{jj'}\left[\int_{-\infty}^{D/\rho_b} \phi(x')\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)dx' + \int_{D/\rho_b}^{\infty} \phi(x')\left(1-\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)\right)dx' - \max\{p, 1-p\}\right]$$

if $\rho_b > 0$ and

$$S_{jj'} = V_{jj'}\left[\int_{-\infty}^{D/\rho_b} \phi(x')\left(1-\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)\right)dx' + \int_{D/\rho_b}^{\infty} \phi(x')\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)dx' - \max\{p, 1-p\}\right]$$

if $\rho_b < 0$, where ϕ and Φ are the probability density function and cumulative distribution function of the standard-normal distribution, ρ is the Pearson correlation between variables j and j', p is the rate of positive cases for variable j, $D = \Phi^{-1}(p)$ and

$$\rho_b = \rho\frac{\sqrt{p(1-p)}}{\phi(D)}.$$

D is the threshold of the underlying standard-normal variable that determines the binary value of j and ρ_b is the correlation between this underlying variable and j'. The reduction in misclassification loss can be calculated directly using these quantities by assuming that we predict 0/1 depending on whether the median of the latent variable conditioned on the value of j' is greater than D or not.

- **j and j' are both binary**

$$S_{jj'} = V_{jj'}\big[p' \max\{\mathbb{P}(x=1|x'=1), 1-\mathbb{P}(x=1|x'=1)\} + (1-p')\max\{\mathbb{P}(x=1|x'=0), 1-\mathbb{P}(x=1|x'=0)\} - \max\{p, 1-p\}\big],$$

where p and p' are the rates of positive cases for p and p' respectively. Here,

the reduction in misclassification loss is calculated directly from the 2×2 contingency table of j and j', since we know the most likely outcome of variable j given the value of j'. This contingency table is calculated using p, p' and ρ.

A formal proof of these results can be found in the supplementary material (see A.1). We vary the tuning parameter k to be 10, 50 and 150.

We therefore have considered three methods. Simple mean imputation replaces missing values in each variable with the mean of the observed values. SoftImpute uses matrix completion assuming a low rank approximation of the data exists, and iteratively completes the missing values in the matrix until they converge. Iterative imputation imputes missing data with a prediction model for each variable, using other variables as covariates, also iterating until the imputed values converge or the maximal number of iterations has been used.

2.2.6 Illustrative Example: Variable Selection for Total Grey Matter Volume on nIDP regression

In order to demonstrate the validity of the conclusions drawn from our simulation study, we apply our imputation methods to an exemplary analytical task on real data and see if there is agreement between the results of the analysis and the conclusions drawn from the simulation study. We choose the task of selecting 15 nIDPs for an Ordinary Least Squares (OLS) model, modelling the association between log-transformed normalised total grey matter volume and the selected nIDPs. The total pool to select from is $\approx 15\,000$ nIDPs (nIDPs with 0 variance or missingness above 40% are excluded). Not a single subject has all of these variables observed, so complete case analysis is not an option. Imputation is used here as a pre-processing step and LASSO-LR is used for variable selection. The outcome, i.e., the 15 variables that are selected varies depending on the imputation method. We compare four different approaches: using only complete variables (meaning only using variables that have no missingness), mean imputation, SoftImpute and iterative imputation. The tuning parameters for SoftImpute and iterative imputation are chosen based on their performance in the simulation study. The four approaches are evaluated by the relevance of the 15 selected variables, as measured by variance explained (R^2) of each OLS

model and the number of statistically significant variable coefficients. To ensure fair assessment of the R^2 scores and coefficient p-values irrespective of missingness in the selected variables, we return missing data entries post selection and use the `mice` package in `R` (Van Buuren and Groothuis-Oudshoorn, 2011) to create $m = 100$ multiple imputed data sets which are used to obtain robust p-values and also to quantify the uncertainty in the R^2 score. Pooled R^2 scores with corresponding standard errors are obtained according to "Rubin's rules" (Rubin, 1987). We use the following estimator for the standard error of R^2 (Olkin and Finn, 1995):

$$\text{se}(R^2) = \sqrt{\frac{4R^2(1-R^2)^2(n-p-1)^2}{(n^2-1)(n+3)}},$$

where n is the number of observations and p the number of variables.

We also ensure that the baseline variables age, age squared, sex and Townsend deprivation index are included in the OLS model as potential nuisance variables, to avoid detecting spurious associations. Pipelines similar to the one described here are commonly used for association modelling tasks on UKB data (Madakkatel et al., 2021; Liu et al., 2023; Zhang et al., 2023).

2.3 Results

2.3.1 Analysis Pipeline

As seen in subsection 2.2.3, we need to select a value for the number of clusters C as a parameter of our analysis pipeline. This choice is jointly driven by the data itself as well as the need to select a small number of clusters to allow us to clearly illustrate our methodology. Figure 2.3 shows the dendrogram of the 100 last agglomerations in the hierarchical clustering of nIDPs by variable-wise missingness pattern. It can be seen from this dendrogram that $C = 4$ will give us clusters which have a high between-cluster missingness distance relative to within-cluster distance. The choice of $C = 4$ clusters also gives us reasonably sized clusters for our analysis, as seen in Table 2.1.

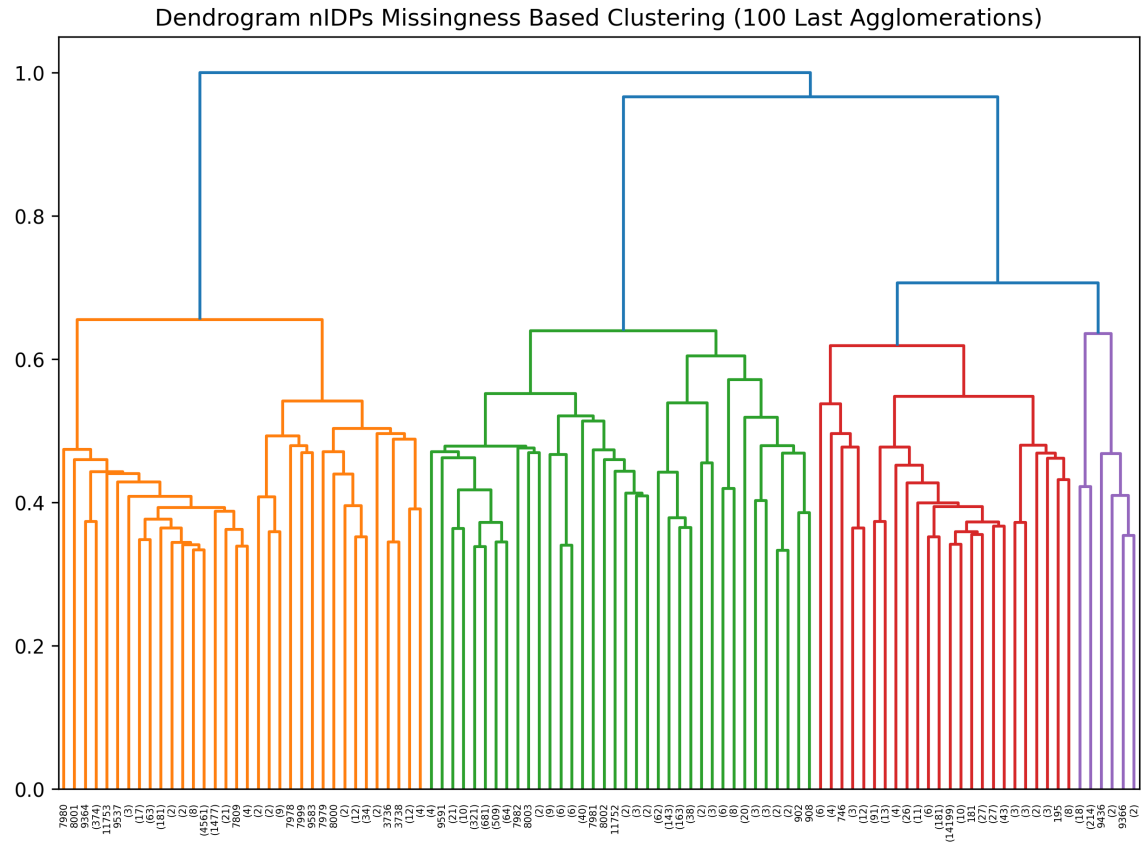


Figure 2.3: Dendrogram of the 100 last agglomerations in the hierarchical clustering of nIDPs by variable-wise missingness pattern, where distance between merged clusters (y-axis) is the maximal variable-wise missingness distance between agglomerated clusters. We determine that $C = 4$ clusters is an appropriate choice for illustrating the workings of our tool since it gives us reasonably sized clusters with a high between-cluster distance relative to within-cluster distance.

The share of each nIDP type by cluster is shown in Figure 2.4. Cluster $c = 0$ contains almost exclusively health and medical related nIDPs, cluster $c = 2$ contains mostly lifestyle and environment related variables, cluster $c = 3$ almost exclusively contains cognitive phenotype variables and cluster $c = 1$ contains a mix of the remaining types of variables. This results shows that nIDPs of the same type tend to have similar variable-wise missingness patterns.

Figure 2.5 plots the histograms of the proportions of variable-wise missing data, i.e., fraction of subjects missing for each variable in a cluster. As we can see, cluster $c = 0$, the cluster that contains mostly health- and medical related variables, has almost no missingness. This is an expected result, as health records usually either

Table 2.1: Table of cluster sizes.

Cluster	# of Binary Variables	# of Continuous Variables	Total # of Variables
$c = 0$	13458	1227	14685
$c = 1$	371	6434	6805
$c = 2$	131	2012	2143
$c = 3$	3	235	238

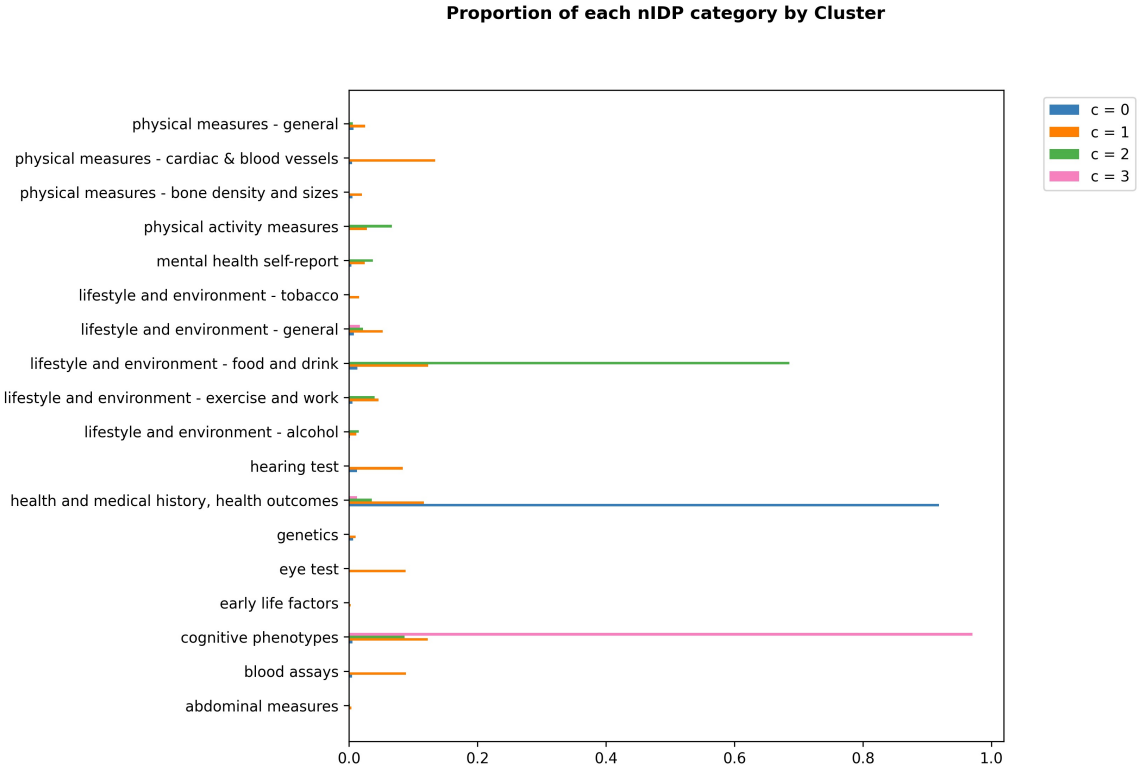


Figure 2.4: Bar plots detailing the proportion of nIDP variable types in each cluster. Cluster $c = 0$ contains almost exclusively health and medical related nIDPs, cluster $c = 2$ contains mostly lifestyle and environment related variables, cluster $c = 3$ almost exclusively contains cognitive phenotype variables and cluster $c = 1$ contains a mix of the remaining types of variables. This shows that nIDPs of the same type tend to have similar variable-wise missingness patterns.

contain too much missingness to be included in the first place or they have very little missingness as absence of data indicates absence of recorded disease or diagnosis. Clusters $c = 2, 3$ have intermediate rates of missingness with cluster $c = 3$ having lower rates of missingness as well as a lower variability of rates of missingness, while $c = 1$ has very high rates of missingness. For this reason, we exclude cluster $c = 1$ from our generative model, as its variables have too high rates of missingness to be

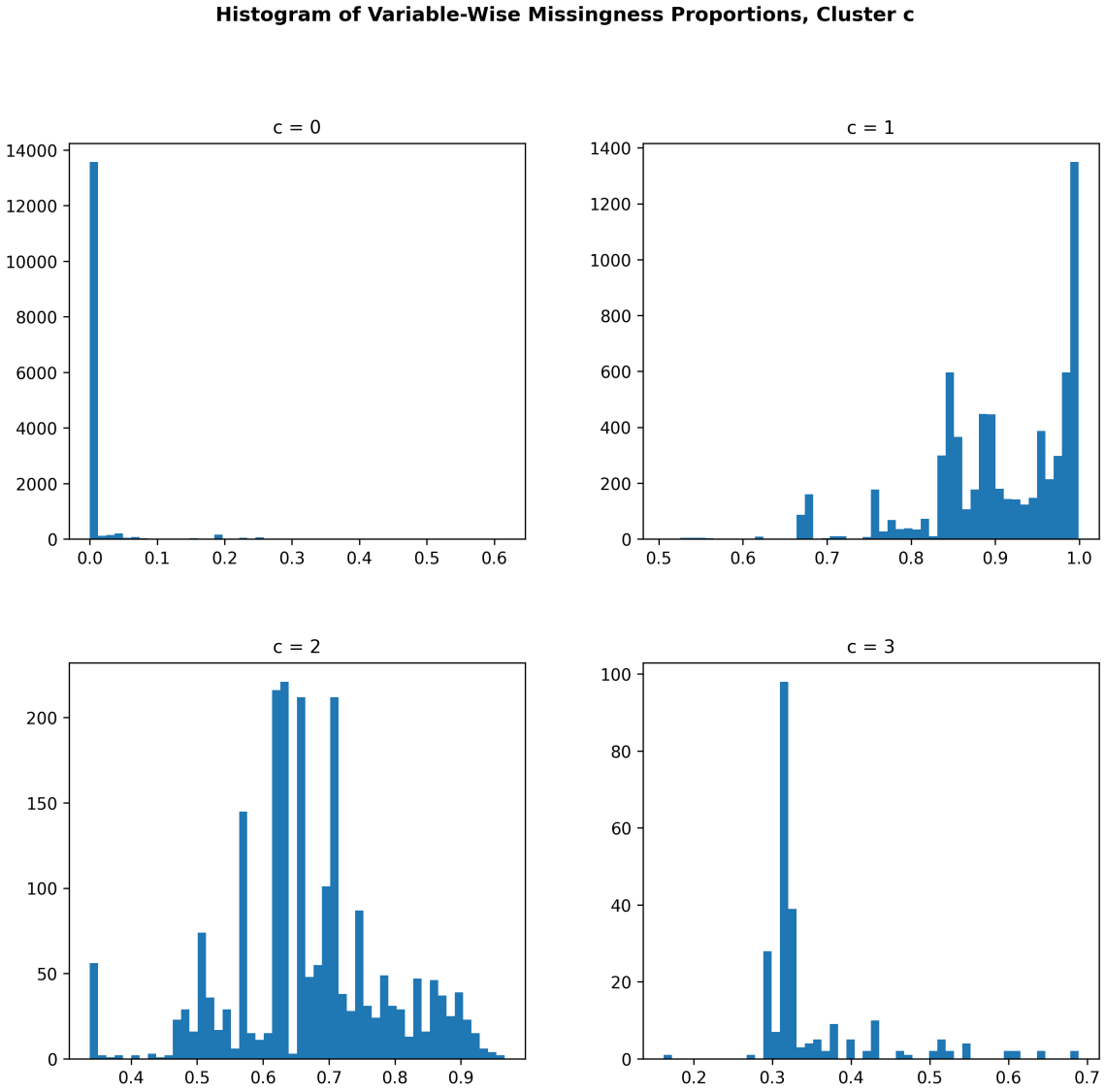


Figure 2.5: Histograms of the proportions of missing data for variables in each cluster. Each entry in the histogram for cluster c is the proportion of missing data for a single variable belonging to cluster c. Cluster $c = 0$, i.e., the cluster that contains mostly health- and medical related variables, has almost no missingness. Clusters $c = 2, 3$ have intermediate rates of missingness with cluster $c = 3$ having lower rates of missingness as well as a lower variability in the same, while $c = 1$ has very high rates of missingness.

interesting to use for imputation.

Figure 2.6 displays the subject-wise missingness histograms, the proportion of cluster-c variables missing for a given subject. The red line in each plot signifies the 90% threshold for structured missingness, meaning that subjects for which 90% or

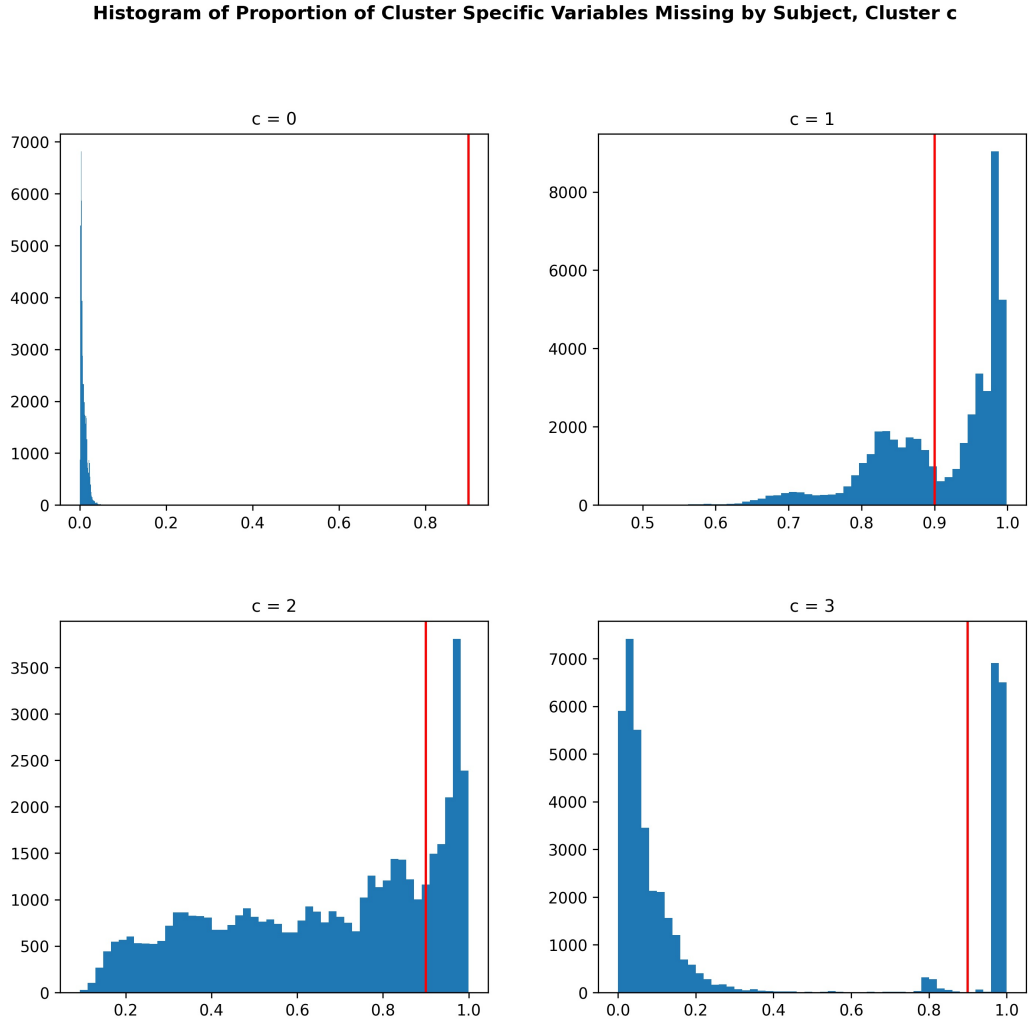


Figure 2.6: Histograms detailing the proportion of variables assigned to cluster c that are missing, by subject. The red line in each plot signifies the 90% threshold for structured missingness, meaning that subjects for which 90% or more of the features assigned to cluster c are missing are considered to have structured missingness for the variables in cluster c. We see that the cluster $c = 3$ has a much clearer separation between structured and unstructured missingness, whereas it is less clear for clusters $c = 1, 2$, likely due to higher rates of unstructured missingness as well as our approximation of $C = 4$ leading to different small sub-studies being assigned to the same cluster.

more of the features assigned to cluster c are missing are considered to have structured missingness for the variables in cluster c. We see that cluster $c = 3$ has a much clearer separation between structured and unstructured missingness, whereas it is less clear for clusters $c = 1, 2$, likely due to higher rates of unstructured missingness as well as our approximation of $C = 4$ leading to different small sub-studies being grouped

Table 2.2: Table detailing the validation AUC for predicting structured missingness using variables from cluster $c = 0$ and number of binary and continuous variables selected using LASSO Logistic Regression (LASSO-LR) for optimal values of λ_c.

Cluster	AUC_c	# of Binary Variables	# of Continuous Variables	$d_{\text{core},c}$
$c = 2$	0.72	48	575	623
$c = 3$	0.90	271	948	1219

Table 2.3: Table detailing the validation AUC for predicting structured missingness using variables from cluster $c = 0$ and number of binary and continuous variables selected using LASSO logistic regression using $\lambda_c = \exp(6)$ for both clusters $c = 2, 3$.

Cluster	AUC_c	# of Binary Variables	# of Continuous Variables	$d_{\text{core},c}$
$c = 2$	0.71	5	62	67
$c = 3$	0.86	5	78	83

Table 2.4: Table of results summarising the data analysis step of our tool. These results indicate that clusters $c = 2, 3$ have similar rates of structured missingness, while cluster $c = 2$ has a much higher rate of unstructured missingness, as indicated by the values of α_c and β_c, since the rate of unstructured missingness is $\alpha_c/(\alpha_c + \beta_c)$. It is also apparent that the variables in cluster $c = 3$ have a more informative type of structured missingness as we can see by the values of AUC_c.

Cluster	AUC_c	$d_{\text{core},c}$	π_c	α_c	β_c
$c = 2$	0.71	67	0.26	6.0	4.6
$c = 3$	0.86	83	0.29	0.44	4.4

together in the same cluster.

As discussed in subsection 2.2.4.1, the penalty terms λ_c need to be carefully chosen to not violate the assumptions of our generative model. Manual tuning arrived at a value of $\lambda_c = \exp(6)$ for both clusters $c = 2, 3$ which minimised both the total number of core variables as well as the proportion of binary variables, while having validation AUC scores close to those of the optimal values, as shown in Tables 2.2 and 2.3.

The final results of the data analysis are summarised in Table 2.4. These results indicate that clusters $c = 2, 3$ have similar rates of structured missingness, while cluster $c = 2$ has a much higher rate of unstructured missingness, as indicated by the values of α_c and β_c. It is also apparent that the variables in cluster $c = 3$ have a more informative type of structured missingness as we can see by the higher value of AUC_c.

2.3.2 Simulation Study

Figures 2.7 and 2.8 plot the imputation accuracy by variable for datasets using the generative model as well as the unstructured equivalent described in sub-section 2.2.5. The red line in each violin plot is the median of the best performing method in that comparison, i.e., the lowest median MSE and the highest median BA. To interpret the results, it should be noted that the continuous variables of the generative model all have zero mean and unit variance. Iterative imputation that uses Pearson correlation or the mixed score as its criterion for variable selection is the best performing method overall. It is also notable that the SoftImpute performs poorly for binary data and notably worse than the iterative imputation methods for continuous data. When comparing the performance between generative model versus completely unstructured missing data, we see that performance is better for the completely unstructured case, for both continuous and binary variables. This difference is particularly stark for cluster $c = 3$, where there is a lot of structured missingness and very little unstructured missingness, highlighting the difficulty of imputation in this setting. Finally, it should be noted that we are rarely able to explain more than 20% of variance in the missing values and that this could mean that the choice of imputation method will not greatly impact the outcome of many analytical tasks, as the modest accuracy of imputation may not be enough to greatly alter the final outcome.

2.3.3 Illustrative Example

Table 2.5 lists the variance explained for the OLS model using the selected variables along with the number of variables in the model that ended up being statistically significant. The results show a modest difference between the three imputation methods, with iterative imputation having the most statistically significant variables and the best R^2 score. The results when using only complete variables are worse with a considerably lower R^2 score as well as fewer variables ending up statistically significant. We also see that the complete variables method selected many more binary variables and this is because the complete variables are mostly health record data, i.e., data assigned to cluster $c = 0$, which is disproportionately binary as seen in Table 2.1. Meanwhile, the results for iterative imputation are the best, having the highest R^2

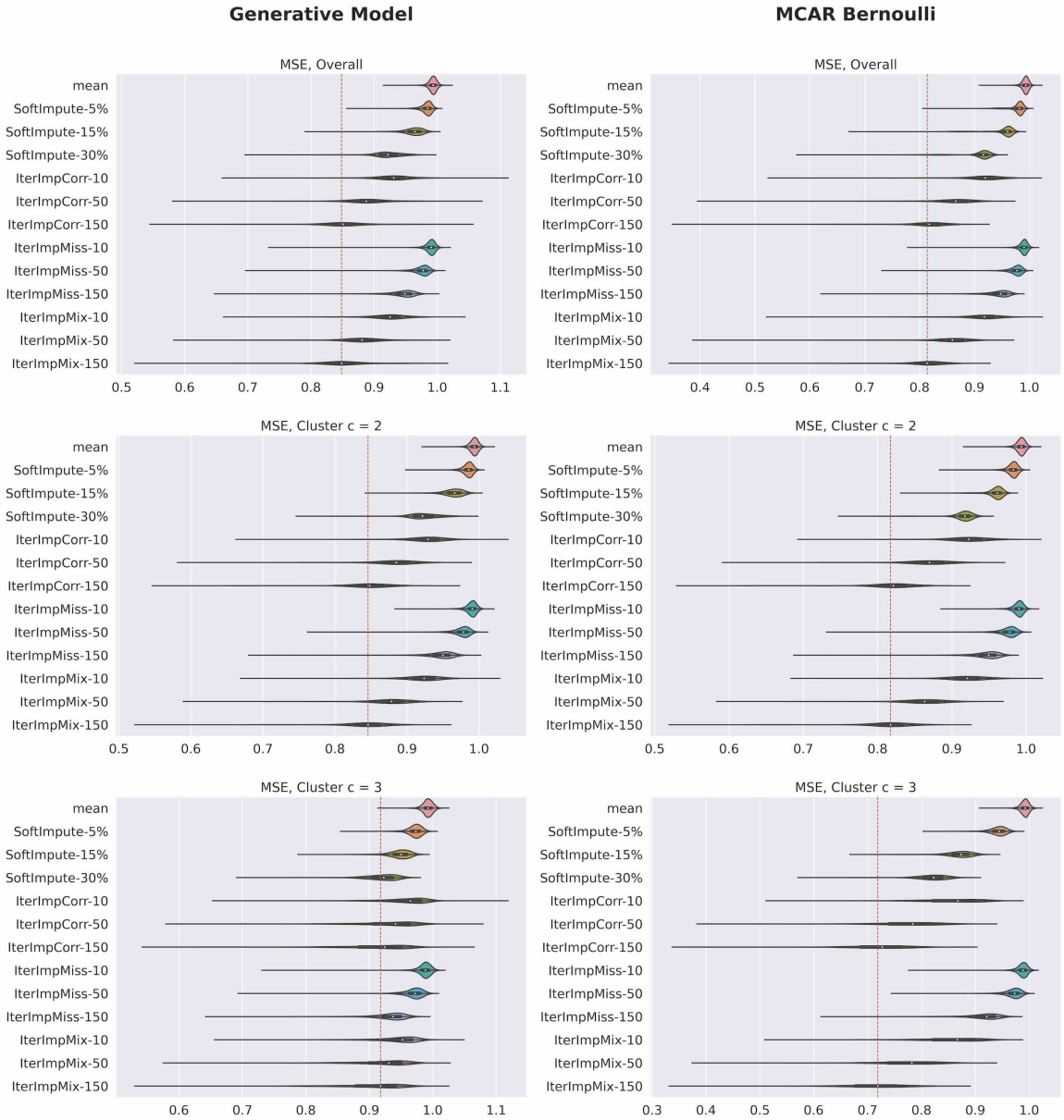


Figure 2.7: Violin plots of imputation accuracy for continuous variables, over $B = 20$ synthetic data sets. The red line in each violin plot is the median of the best performing method in that comparison, i.e., the lowest median MSE. All continuous variables have unit variance. It is clear that iterative imputation that uses Pearson correlation or the mixed score as its criterion for selecting k variables is the best performing method overall. Our tool estimates lower imputation accuracy than the unstructured approach, indicating that imputation is harder under SM.

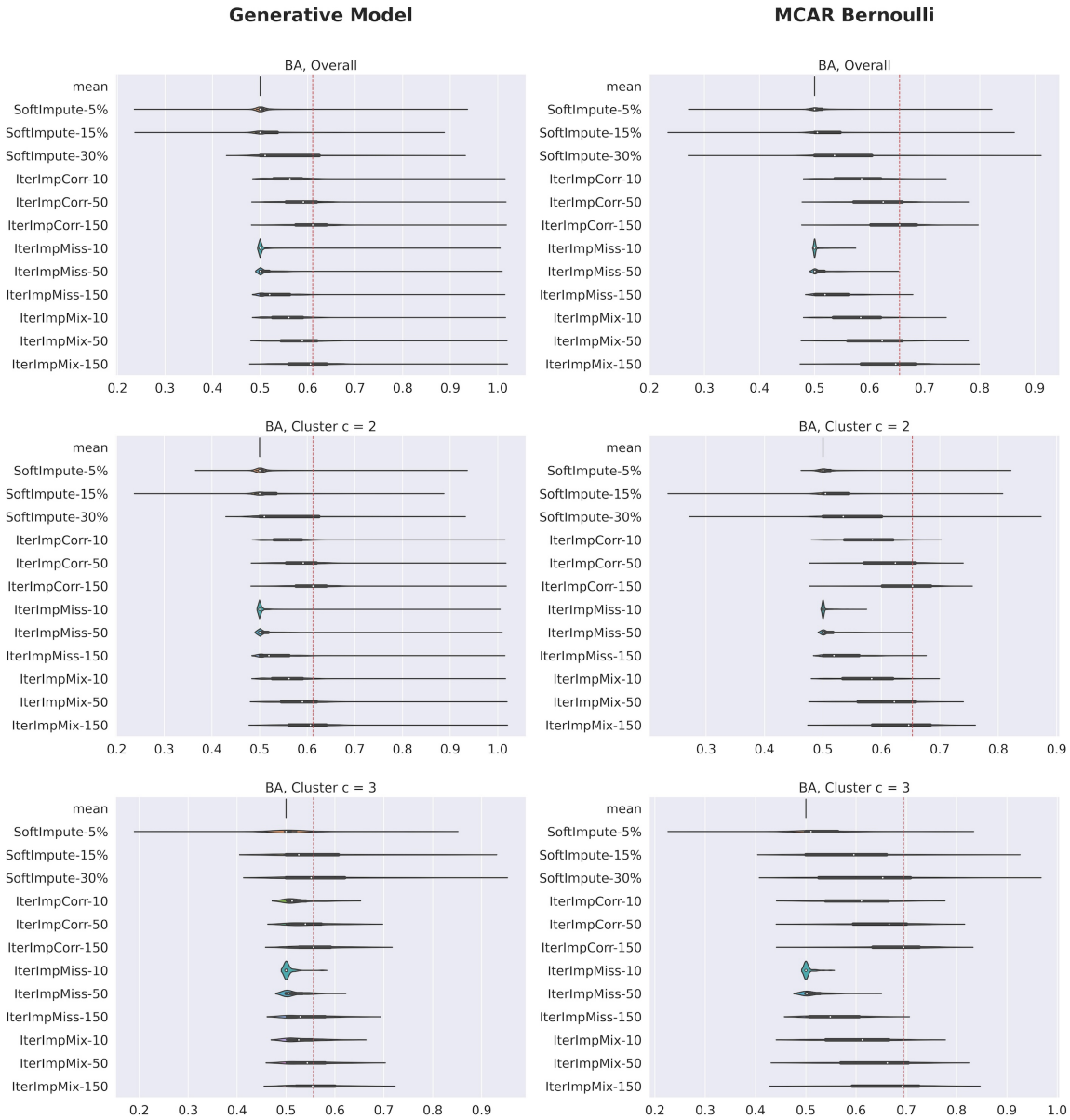


Figure 2.8: Violin plots of imputation accuracy for binary variables, over $B = 20$ synthetic data sets. The red line in each violin plot is the median of the best performing method in that comparison, i.e., the highest median BA. It is clear that iterative imputation that uses Pearson correlation or the mixed score as its criterion for selecting k variables is the best performing method overall. Our tool estimates lower imputation accuracy than the unstructured approach, indicating that imputation is harder under SM.

Table 2.5: R^2 scores for the OLS models using the selected variables along with the number of variables in the model that ended up being statistically significant. The results show a modest difference between the three imputation methods, with iterative imputation having the most statistically significant variables and the best R^2 score. The results when using only complete variables are worse with a considerably lower R^2 score as well as fewer variables ending up statistically significant.

Method	R^2 of OLS	se(R^2)	#Statistically Significant in OLS	# Binary vars. selected	Rate of missingness in selected vars.
Complete Variables	0.475	0.0034	11/15	12/15	0.00
Mean	0.510	0.0033	12/15	1/15	0.09
SoftImpute	0.504	0.0034	12/15	1/15	0.13
Iterative Imputation	0.516	0.0033	14/15	2/15	0.10

score as well as the highest number of statistically significant variables. These results align well with our simulation study; a small difference in the final outcome of the analytical task for different methods caused due to the difficulty of imputing structurally missing data, but with iterative imputation clearly being the best alternative.

2.4 Discussion

We have proposed a method for generating large-scale data with complex patterns of missing data that make imputation difficult. In particular, our data-driven simulation framework allows for highly informative missingness and joint missingness for variables that are strongly correlated. This ability to mimic the properties of large scale epidemiological datasets makes our tool useful for gaining insight into the performance of handling missing data for different analytical tasks. There are, however, limitations to our model which are important to note and represent potential future work on this topic. Our model assumes multivariate normality for all continuous features, which limits the generalisability of any conclusions drawn from simulation studies of analytical methods that are sensitive to non-gaussianity or strong outliers. In such scenarios, it is possible that conclusions drawn using our generative model would unduly favour linear methods over more complicated black-box methods that would fare better on real, non-Gaussian data. This could be solved by parameterising the model differently, allowing for more flexibility on the underlying multivariate

distribution, or by using non-parametric methods.

Another potential limitation of our generative model is that we assume that the correlation structure of the data closely follows the missingness structure. This is because we assume that for any pair of clusters c, c', the correlation between pairs of variables in c and c' are drawn independently from some distribution $\{\rho\}_{c,c'}$, i.e., we assume that there is no further covariance structure within or between clusters. We have found this to be approximately true for the nIDPs that we have been working with, but this might not be the case for other datasets. This could be solved by modelling missingness and correlation structure jointly in a way which allows for further complexity inside clusters.

In this project we chose to use $C = 4$ clusters of variables as an approximation of reality in order to be able to inspect the properties of these clusters separately. In all likelihood, the true number of sub-studies is higher, and more representative results could be obtained by choosing a higher figure. We deemed it necessary to use a lower number in order to demonstrate the inner workings of our tool. When allowing C to be higher, we found $C = 15$ clusters with 100 or more variables present in the data set. These clusters all had a very clear separation between the structured and unstructured missingness, which bolsters our hypothesis that missingness in UKB data can be effectively modelled as we have suggested.

2.5 Conclusions

The results from the simulation study combined with the illustrative example show that there is room for improvement in the current missing data methodology to accommodate for this specific type of missing data, and that the final result of many analytical tasks on data from the UKB Brain Imaging cohort will vary little depending which commonly used imputation method is chosen. This is due to the difficulty of imputation in this setting. Even so, we have shown some advantage in using iterative imputation over matrix completion methods. While we do not propose new missing data methods here, our results highlight the need for developing methods that specifically account for structured missingness.

Chapter 3

Accounting for Structured Missingness in Canonical Correlation Analysis

3.1 Introduction

There has recently been an increased interest in studying and characterising the structure of missingness patterns in large-scale data sets (Mitra et al., 2023; Jackson et al., 2023; Radosavljević et al., 2024). Structured Missingness (SM) can have diverse causes, such as combining sub-studies with overlapping sets of participants, batch failure or recording data from a questionnaire where some questions are recorded only if the subject answered 'Yes' or 'No' to a previous question, leading to blocks of missing data (Mitra et al., 2023). While missing data literature mostly ignores missing data structure, this might be problematic. It is not uncommon for SM to lead to a situation where highly correlated variables are mostly jointly missing, making it much more difficult to impute missing values using correlated variables, as we see in Chapter 2. Therefore, new missing data methodology is needed to address SM specifically.

In this work, we focus on handling SM for the task of Canonical Correlation Analysis (CCA) (Hotelling, 1992). Our work is inspired by tabular data from subjects in the

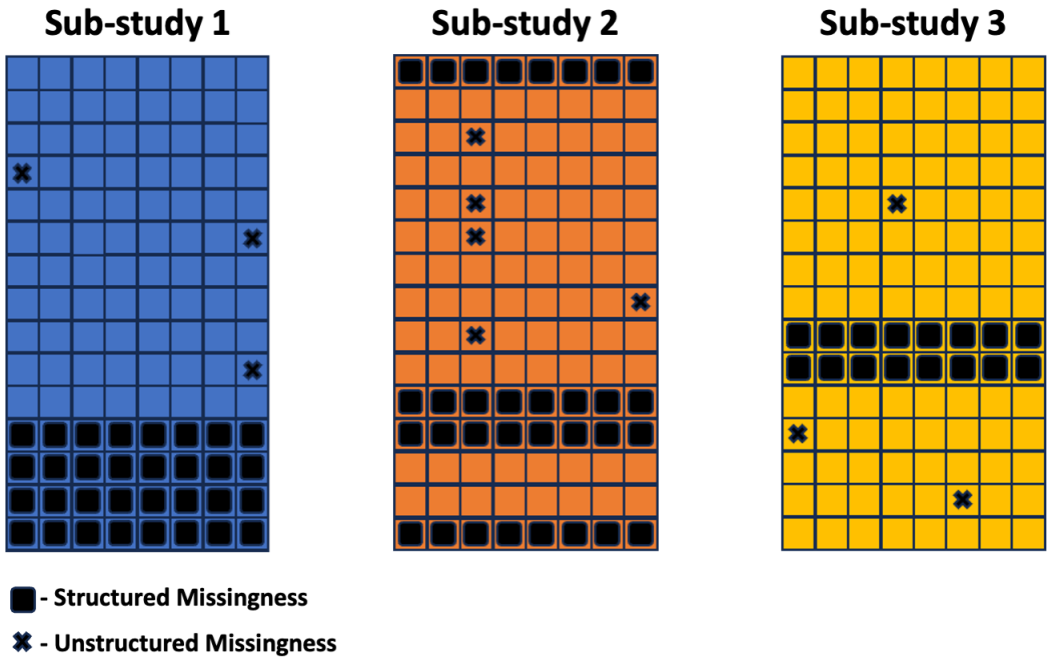


Figure 3.1: The missing data in a sub-study can be divided into structured missingness (SM), caused by non participation, and unstructured missingness (UM) attributable to any other cause.

UK Biobank (UKB) Brain Imaging Cohort. This data consists of imaging derived phenotypes (IDPs), which are derived from MRI images of the subjects' brains, and non-imaging derived phenotypes (nIDPs), which come from diverse sub-studies in UKB. SM in this type of data is caused by non-participation, as each subject has participated in a specific set of sub-studies, and non-participation in any sub-study will lead to missingness for all variable derived from the same. Additionally, a sub-study can also have unstructured missingness (UM) that is attributable to any other cause. Figure 3.1 illustrates this setting, which is similar to a multi-view learning problem with missing views (Xu et al., 2015; Zhang et al., 2018), and with UM added. CCA is commonly used to link IDPs and nIDPs, both of which can have SM, so it is therefore important to investigate how to best handle SM for this analytical task. We evaluate several existing methods on both real and synthetic data and propose a new method which handles SM and UM separately.

3.1.1 Canonical Correlation Analysis

CCA is a multivariate analysis method commonly used in e.g. neuroscience for tasks such as de-noising, detecting task fMRI activation or analysing and testing associations between different modalities (Smith et al., 2015; Farahibozorg et al., 2024; Zhuang et al., 2019, 2020; Winkler et al., 2020). It finds the maximal correlation between linear combinations of two variable sets. If $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$ and $\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}$ are random vectors representing the two variable sets, CCA seeks to find weight vectors $\beta_\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$, $\beta_\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}$ such that *canonical variables* $u = \mathbf{x}^\top \beta_\mathbf{x}$ and $v = \mathbf{y}^\top \beta_\mathbf{y}$ have maximal correlation. CCA can be performed via the generalised eigenvalue problem (GEP) in the following way (Ge et al., 2016):

$$\mathbf{w}_i \in \underset{\mathbf{w} \in \mathbb{R}^{d_\mathbf{x}+d_\mathbf{y}}}{\arg\max} |\mathbf{w}^\top \mathbf{A} \mathbf{w}| \text{ subject to } \begin{array}{l} \mathbf{w}^\top \mathbf{B} \mathbf{w} = 1 \\ \mathbf{w}^\top \mathbf{B} \mathbf{w}_j = 0 \ \forall j < i, \end{array} \tag{3.1}$$

where $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \mathbf{0} \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\mathbf{yy}} \end{pmatrix}$, $\mathbf{w} = \begin{pmatrix} \beta_\mathbf{x} \\ \beta_\mathbf{y} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}$ is the covariance matrix of $(\mathbf{x}, \mathbf{y})$.

In practice, since the joint covariance matrix of $\mathbf{x}$ and $\mathbf{y}$ is usually unknown, it is estimated from data; $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ and $\mathbf{Y}$ are $n \times d_\mathbf{x}$ and $n \times d_\mathbf{y}$ data matrices corresponding to n independent observations of $(\mathbf{x}, \mathbf{y})$.

3.1.2 Confounding and Missingness

When analysing associations between IDP and nIDP data, there are often sources of confounding which we want to adjust for prior to performing CCA. In epidemiological studies, variables such as age, sex, BMI, ethnicity, geographical region and variables related to socioeconomic status are examples of confounds (Alfaro-Almagro et al., 2021). These are the same types of variables that are also broadly predictive of study participation (Fry et al., 2017) and therefore also SM. Figure 3.2 shows the relationship between missingness and confounding variables that is assumed throughout the paper. Also, if we assume that our confound variable $\mathbf{b}$ is fully observed, we see that this missing data model falls under *Missing at Random* (MAR), since the relationship between data can be described using only observed variables. If $\mathbf{m}$ is the missingness

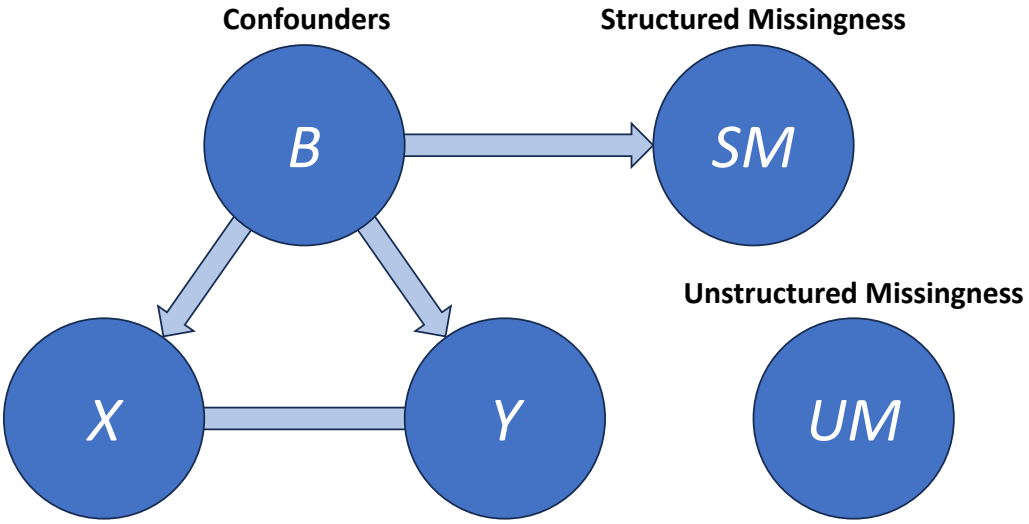


Figure 3.2: Structured missingness is assumed to be determined by the confounding variables, i.e., baseline characteristics such as age, sex, BMI, ethnicity, socioeconomic status etc., while unstructured missingness is assumed to be independent of all variables. This assumption results in data that falls under MAR, which our method relies on.

indicator of our joint random vector $\mathbf{d} = (\mathbf{b}, \mathbf{x}, \mathbf{y})$, which can be split into observed and unobserved components $\mathbf{d}_{\text{obs}}$ and $\mathbf{d}_{\text{mis}}$, then we have that

$$\mathbb{P}(\mathbf{m}|\mathbf{d}) = \mathbb{P}(\mathbf{m}|\mathbf{b}), \tag{3.2}$$

and since $\mathbf{b} \subseteq \mathbf{d}_{\text{obs}}$ (confounds being always fully observed), it follows that

$$\mathbb{P}(\mathbf{m}|\mathbf{d}) = \mathbb{P}(\mathbf{m}|\mathbf{b}) = \mathbb{P}(\mathbf{m}|\mathbf{d}_{\text{obs}}), \tag{3.3}$$

which by definition means that data is MAR.

3.1.3 Related Work

Since CCA requires estimation of the covariance matrix from observed data, the work most relevant to this paper concerns the estimation of covariance matrices from incomplete data. Under the assumption of multivariate Gaussianity and MAR, the Maximum Likelihood Estimate (MLE) of the covariance matrix Σ can be found via the Expectation-Maximisation (EM) algorithm (Schafer, 1997). This is, however, a highly computationally expensive procedure, as each iteration of the EM algorithm requires h inversions of $(d_{\mathbf{x}} + d_{\mathbf{y}}) \times (d_{\mathbf{x}} + d_{\mathbf{y}})$ matrices, where h is the number of missingness patterns observed in the n rows of $\mathcal{D}$. It is easy to see that any unstructured missingness will lead to $h \approx n$ in the high dimensional case, because there are $2^{d_{\mathbf{x}}+d_{\mathbf{y}}} >> n$ possible missingness patterns $\mathbf{m}$, and it is therefore unlikely that the same missingness pattern would appear more than once. This consequently leads to a very high time complexity for the EM algorithm. Therefore, the use of the MLE of the covariance matrix via EM has mostly been restricted to low-dimensional, small data settings (Enders and Peugh, 2004).

Several faster unbiased methods using inverse probability weighting (IPW) have been proposed for learning covariance matrices from incomplete data assuming *missing completely at random* (MCAR). The method proposed by Lounici (2014) assumes that missingness for each entry is determined by independent and identically distributed Bernoulli random variables with a known probability δ. However, this estimator is heavily biased in our structured setup where there is both strong dependence between missing data patterns and differing rates of missingness between variables. Park and Lim (2019) extend this approach to allow for dependence between missingness patterns and variable-specific missingness rates, but their proposed estimator requires a-priori knowledge of the joint probability distribution of the missingness over all variables. If these parameters are estimated from data, the Park and Lim estimator becomes the standard covariance matrix estimate using pairwise complete observations (PCO), which is not necessarily positive semi-definite (PSD).

Multivariate Imputation by Chained Equations (MICE) (Rubin, 1987; Van Buuren and Groothuis-Oudshoorn, 2011) is also commonly used to handle missing data and is applicable to the problem of covariance estimation from incomplete data. MICE

iteratively creates random synthetic replacements for the missing entries by fitting regression models for each variable using all other variables, which stochastically impute missing entries. Like the EM algorithm, this method also suffers from high computational cost, although this problem can be alleviated by using a carefully selected subset of variables to fit the prediction models, instead of using all available variables (Radosavljević et al., 2024; Pedregosa et al., 2011; Van Buuren and Oudshoorn, 1999; Graham, 2009).

Finally, it should be noted that in most cases, missing data is handled via simple, non-stochastic, Single Imputation methods such as mean imputation or matrix completion, prior to applying CCA. This is an inherently biased approach, unless imputation can be done with perfect accuracy, as non-stochastic imputation creates a downward bias in variance by not accounting for noise in the imputed values. And since the variance is biased post imputation, covariance is also generally biased.

3.2 Methods

Our method handles structured and unstructured missingness separately. When linking two modalities using CCA, we are often very close to a pure multi-view setting (i.e., a setting with no UM) where the EM algorithm can be practically used for covariance matrix estimation. However, the unstructured missingness makes this intractable, as discussed in subsection 3.1.3 and also in Le Morvan et al. (2020). Additionally, we assume that correlations within $\mathbf{X}$ and $\mathbf{Y}$ are strong, while correlations between $\mathbf{X}$ and $\mathbf{Y}$ are weak, making unstructured missingness easy to impute and structured missingness difficult to impute. For example, a study measuring the weights of different body parts will produce highly correlated variables, as the weights of individual body parts are strongly correlated with the subjects' height and BMI. Variables from different studies that measure completely different things tend to be more weakly correlated. We therefore propose to handle unstructured missingness prior to applying the EM algorithm, either via a single imputation method such as mean imputation or SoftImpute (SI) (Mazumder et al., 2010), or via MICE. After that, we use EM to obtain an approximate MLE of the covariance matrix from incomplete data. In Figure 3.3, we see an illustration that describes the estimation procedure. When using

MICE, we create r multiple imputed data sets, obtain MLE covariance matrices for each of them, and average them to obtain a single estimate.

There is mathematical intuition to support the validity of the proposed method using MICE. Let Σ be the covariance matrix described in equation 3.1. If $\mathcal{D}_{\text{obs}} = (\mathbf{X}_{\text{obs}}, \mathbf{Y}_{\text{obs}})$ is the observed data and $\mathcal{D}_{\text{UM}} = (\mathbf{X}_{\text{UM}}, \mathbf{Y}_{\text{UM}})$ is unobserved data obscured by unstructured missingness only, we know that

$$\mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}) = \int_{\mathcal{D}_{\text{UM}} \in \mathcal{X}} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}) \mathbb{P}(\mathcal{D}_{\text{UM}} | \mathcal{D}_{\text{obs}}) d\mathcal{D}_{\text{UM}}, \tag{3.4}$$

where $\mathcal{L}$ is the likelihood function. It can be argued further, using the logic of Monte-Carlo integration that

$$\mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}) \approx \frac{1}{r} \sum_{k=1}^{r} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}^{(k)}), \tag{3.5}$$

where $\mathcal{D}_{\text{UM}}^{(1)}, ..., \mathcal{D}_{\text{UM}}^{(r)}$ have been sampled independently from $\mathbb{P}(\mathcal{D}_{\text{UM}} | \mathcal{D}_{\text{obs}})$ using MICE. It follows that

$$\begin{aligned} \hat{\Sigma}_{\text{MLE}} = \arg\max_{\Sigma} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}) \approx \arg\max_{\Sigma} \frac{1}{r} \sum_{k=1}^{r} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}^{(k)}) \approx \\ \approx \frac{1}{r} \sum_{k=1}^{r} \arg\max_{\Sigma} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}^{(k)}), \end{aligned} \tag{3.6}$$

where the second approximation assumes that the functions $\mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}^{(k)}), k = 1, ..., r$ are well approximated by quadratic functions with the same Hessian (see B.1 in the supplementary material for a full justification). Since $\mathcal{D}_{\text{obs}}$ and $\mathcal{D}_{\text{UM}}^{(k)}$ together form an imputed data set with only structured missingness, it is computationally tractable to find $\arg\max_{\Sigma} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}^{(k)})$ via EM. Thus, our final estimate is

$$\hat{\Sigma} = \frac{1}{r} \sum_{k=1}^{r} \arg\max_{\Sigma} \mathcal{L}(\Sigma; \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{UM}}^{(k)}). \tag{3.7}$$

In practice, when finding the average of covariance matrices, they are often averaged in a reparametrised form and then returned to their original form after averaging (Pervaiz et al., 2020; Dryden et al., 2009). We average the covariance matrices in

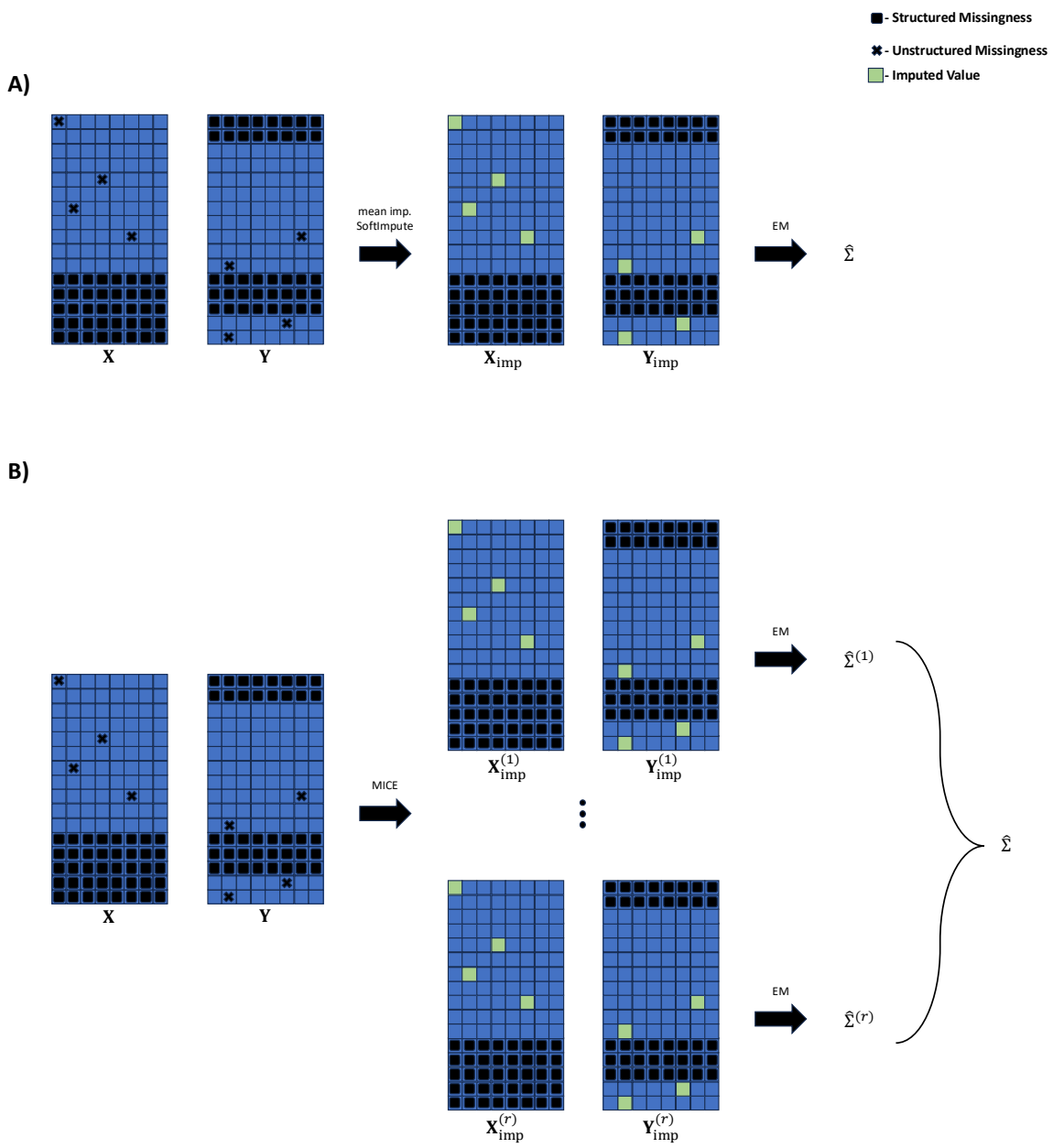


Figure 3.3: Illustration of our proposed method. A) We apply a Single Imputation method to impute the unstructured missingness in $\mathbf{X}$ and $\mathbf{Y}$. After this, the EM algorithm is applied to obtain the MLE covariance matrix. B) We use MICE to create r multiple imputed data sets where only unstructured missingness has been imputed. We apply EM to obtain r MLE covariance matrices, which we average to obtain a single estimate.

Cholesky space (Lin, 2019), because this is a simple and fast reparametrisation that has been shown to have good empirical performance for covariance matrices that are well conditioned (Dryden et al., 2009). This is a version of the standard method of pooling multiple imputed data, called Rubin's rules (Rubin, 1987), performed on covariance matrices in a reparametrised form.
In order to save computational time, we only use the variables in $\mathbf{X}$ to impute $\mathbf{X}_{\text{UM}}$ and only the variables in $\mathbf{Y}$ to impute $\mathbf{Y}_{\text{UM}}$, as variables from the same sub-study tend to be more strongly correlated than variables from different substudies, making them more suitable for imputation. For the version of our method that uses MICE to impute UM, we only use a selection of 5 imputation variables for each variable.

The main benefit of our proposed method is that the covariance matrix is estimated efficiently without imputation, after handling the small amount of UM. The EM MLE after imputing UM uses all data available, in contrast with MICE where we are forced to select a subset for each variable that we use for imputation. This both saves computational time and reduces bias by avoiding this type of truncation.

3.2.1 Other Methods

We compare several other methods to our approach. The simplest of these include Single Imputation methods, mean imputation and SoftImpute. For SoftImpute, we select the soft threshold for singular values heuristically using the elbow rule (Greenacre et al., 2022). Another method is MICE, where we create r multiple imputed data sets, calculate the sample covariance matrix on each set, and average them to obtain a single estimate (pooling is done in Cholesky space as described in section 3.2). Since we are working in a high dimensional setting, we use the proposed mixed variable selection procedure for iterative imputation models described in Subsection 2.2.5. Here, we use 20 imputation variables for each variable. Finally, we consider the method of estimating the covariance matrix using pairwise complete observations. This estimate is unbiased under MCAR, but is not necessarily PSD. We solve this by first calculating standard deviations for each variable using complete observation. We then calculate the correlation matrix using the PCO approach and project it to the closest PSD correlation matrix with Higham's Algorithm (Higham, 2002) using the `nearPD` function in R's `Matrix` package (Bates et al., 2010). Finally, we transform

this correlation matrix back to a PSD covariance matrix using the estimated standard deviations.

We acknowledge that there could be some benefit in excluding rows that have structured missingness in either $\mathbf{X}$ or $\mathbf{Y}$, especially for Single Imputation methods where bias is increased with the amount of missingness. Assuming that the causal structure shown in Figure 3.2 is correct, excluding rows with SM should yield correct inferential results, as fully observed confounds contain all information that is predictive of missingness. On the other hand, the loss of information by excluding whole rows with SM could be detrimental to inference, or lead to a situation where the number of rows and columns are similar, which is itself problematic for weight stability in CCA. We therefore perform all methods (excluding our proposed method) using both approaches, as shown in Figure 3.4.

3.2.2 Deconfounding

To adjust for confounding variables in our CCA, we first exclude the few rows where there is any missingness in the confounds and then perform standard Ordinary Least Squares (OLS) based deconfounding on $\mathbf{X}$ and $\mathbf{Y}$ as a preprocessing step. This is a standard approach to handling confounds in CCA (Wang et al., 2020). Interestingly, since we assume that structured missingness is determined in large part by confounds, our data will be approximately MCAR after deconfounding, as the data in $\mathbf{X}$ and $\mathbf{Y}$ will be orthogonal to the confounds, and thus by extension uncorrelated with the structured missingness patterns.

3.2.3 Simulation Study

To evaluate our method, we perform a simulation study mimicking the real data case. We simulate data sets $\mathcal{D} = (\mathbf{B}, \mathbf{X}, \mathbf{Y})$ where $\mathbf{B}$ are confounds and $(\mathbf{X}, \mathbf{Y})$ is the pair of sub-studies that we want to perform CCA on. The process for generating confounds is shown using pseudo-code in the supplementary material (see B.2). The data generation process is inspired by Singular Value Decomposition (SVD) and assumes the existence of a low rank approximation of data. $(\mathbf{X}, \mathbf{Y})$ are generated in a similar

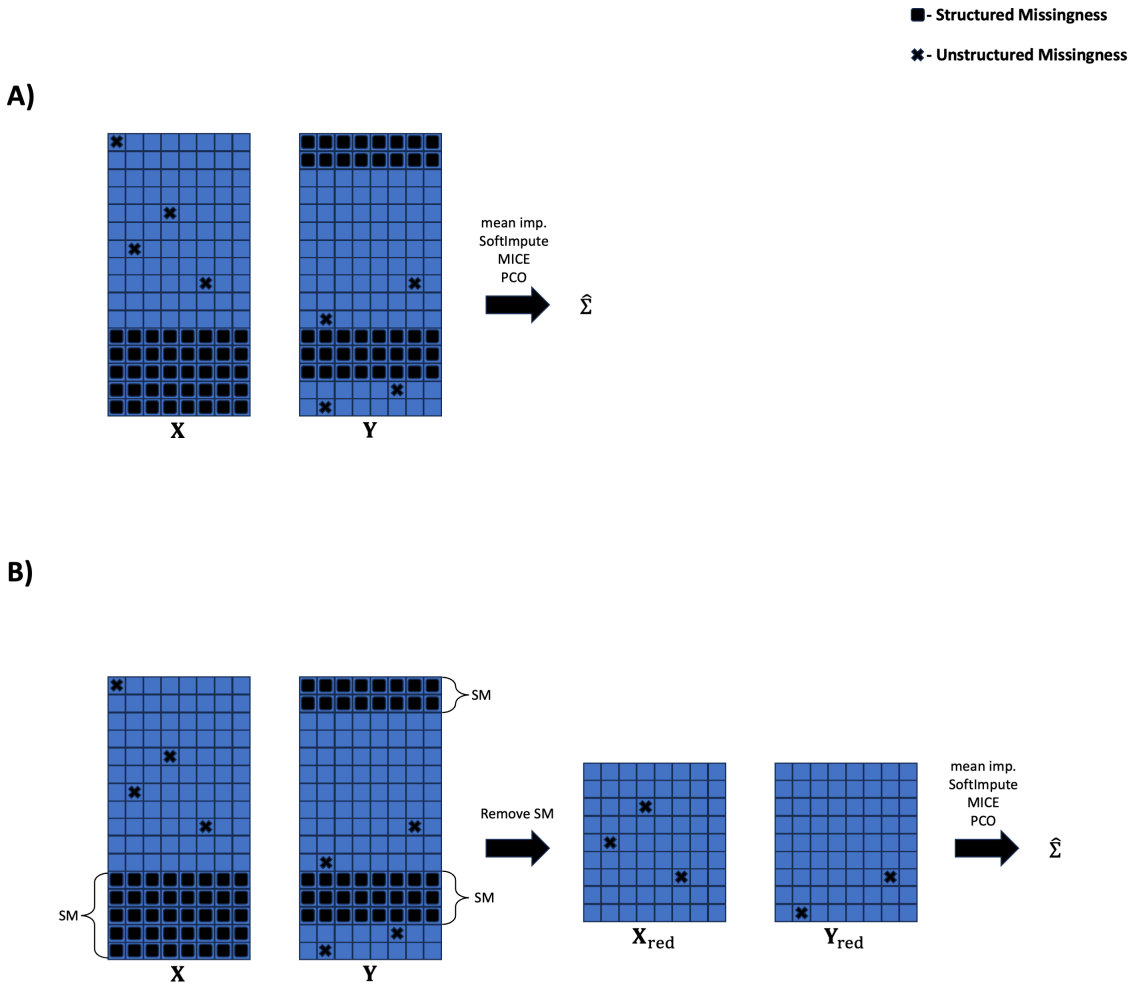


Figure 3.4: Illustration of the two approaches used for CCA. A) We use all rows in $\mathbf{X}$ and $\mathbf{Y}$, irrespective of structured missingness. B) All rows where there is SM in either $\mathbf{X}$ or $\mathbf{Y}$ are excluded prior to downstream analysis.

way, where $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ are two matrices with common variance generated using an SVD inspired procedure, and we set

$$(\mathbf{X}, \mathbf{Y}) = (\mathbf{B}\Gamma_{\mathbf{X}} + \tilde{\mathbf{X}}, \mathbf{B}\Gamma_{\mathbf{Y}} + \tilde{\mathbf{Y}}), \tag{3.8}$$

where $\Gamma_{\mathbf{X}}$ and $\Gamma_{\mathbf{Y}}$ are two matrices of confound effects.
We induce structured missingness in $(\mathbf{X}, \mathbf{Y})$ using our simulated $\mathbf{B}$, such that the probability of SM in $\mathbf{X}$ is p_{SM} and $1.5p_{\text{SM}}$ in $\mathbf{Y}$, while guaranteeing that SM can be predicted using $\mathbf{B}$ with an AUC score of $\text{AUC}_{\text{SM}} = 0.8$. The process we use for creating informative SM is described fully in subsection 2.2.4. SM is determined via a Logistic Regression model using the confounds as variables. The coefficients of the Logistic Regression model are scaled using a binary search procedure to match the AUC score that we have chosen (AUC_{SM}). We also induce unstructured missingness at a rate of p_{UM} in total, but such that the probability of UM is allowed to vary between variables. Specifically, the rates of UM for each variable are drawn from a Beta distribution with expectation p_{UM} and variance $p_{\text{UM}}^2/2$. Lastly, we induce 1% missingness in the confounds for a subset of variables that are not used to induce structured missingness.

We vary the rates of missingness to be $p_{\text{SM}} = 0.1, 0.2, 0.4$ and $p_{\text{UM}} = 0.01, 0.05, 0.15$. We also vary a parameter $\sigma_U = 0.02, 0.03, 0.05$, which controls the level of correlation between $\mathbf{X}$ and $\mathbf{Y}$ after removing confounding effects, with higher values of σ_U leading to stronger canonical correlation. σ_U is the standard deviation of the variance components that are shared between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, while the variance components that are unique to either $\tilde{\mathbf{X}}$ or $\tilde{\mathbf{Y}}$ all have standard deviation 1.
We set $n = 50000$ (roughly the number of subjects in the UKB Brain Imaging cohort) and $d_{\mathbf{x}} = d_{\mathbf{y}} = 200$.

We evaluate the performance of each missingness method on 100 synthetic data sets for each setting, by estimating canonical variable coefficients on a training set and then calculating the average correlation of the first ten canonical variables constructed on a non-confounded test set using our estimated coefficients. In other words, if $(\beta_X^{(1)}, \beta_Y^{(1)}), ..., (\beta_X^{(10)}, \beta_Y^{(10)})$ are our estimated canonical coefficients and $(\tilde{\mathbf{X}}_{\text{test}}, \tilde{\mathbf{Y}}_{\text{test}})$ is our non-confounded test set, the score is:

$$\text{score} = \frac{1}{10}\sum_{k=1}^{10} |\text{corr}(\tilde{\mathbf{X}}_{\text{test}}\beta_X^{(k)}, \tilde{\mathbf{Y}}_{\text{test}}\beta_Y^{(k)})|.$$

The goal is for this score to be as high as possible. We then obtain relative scores by dividing by the score obtained using the Gold Standard (GS), which in our case means analysis on data with no missingness. The relative scores ($\text{score}/\text{score}_{\text{GS}}$) are thus usually in the interval $(0, 1)$ and represent how close our missingness method gets us to replicating a scenario with no missingness. Given the assumptions of our method, we expect it to perform the best compared to other methods for cases where the rate of SM is high and the rate of UM is low, as UM has the lowest impact on covariance estimation in that scenario.

3.2.4 Study on Real Data

To evaluate our method on real data, we analyse four sub-studies from UK Biobank which have structured missingness. They include diffusion MRI (dMRI) weighted-mean tract diffusivity IDPs, food preferences, NMR metabolomics, and refractometry. We also include a set of confounds: sex, age, ethnic background (binary White British vs. not White British), income, Townsend deprivation index, BMI, frequency of alcohol consumption, smoking status and geographical north and east coordinates. Additionally, we include age-squared, age-by-sex and age-squared-by-sex, to account for non-linear age effects and interaction effects between sex and age. Variable numbers, rates of SM and UM, as well as informativeness of SM for the data sets are shown in table 3.1. A table showing the proportion of overlapping participants for each pair of studies is shown in table 3.2. The sets are chosen to represent typical analyses of this kind: The IDP data is high-dimensional and has low rates of SM and UM, the food preferences data has moderate rates of SM and almost no UM, NMR metabolomics has a high rate of SM and low rate of UM, and the Refractometry variables have high levels of SM and UM. Table 3.2 also shows the proportion of subjects that have participated in both studies for each pair of studies, showing diverse settings ranging from 77% for IDPs vs. Food Preferences, to 5.5% for NMR metabolomics vs. Refractometry.

Table 3.1: Number of variables, level of structured (p_{SM}) and unstructured (p_{UM}) missingness as well as informativeness of structured missingness (AUC_{SM}). AUC_{SM} is estimated by fitting a logistic regression model for predicting SM using confounds; AUC_{SM} is the AUC score for prediction calculated on test data.

Set	#Variables	p_{SM}	p_{UM}	AUC_{SM}
Confounds	21	0	0.002	NA
IDPs	300	0.053	0.012	0.61
Food Preferences	130	0.19	$2.4 \cdot 10^{-6}$	0.57
NMR metabolomics	73	0.76	0.0029	0.51
Refractometry	96	0.77	0.095	0.74

Table 3.2: Proportion of subjects (out of $n = 46471$), that participated in both studies. The table shows a wide range of settings for CCA from structured missing data; for IDPs vs. Food Preferences, a comfortable majority of subjects participated in both studies, while for NMR metabolomics vs. Refractometry, fewer than 6% of subjects participated in both studies, and is consequently a much more unstable setting.

Set 1	Set 2	Proportion Participated in Both
IDPs	Food Preferences	0.77
IDPs	NMR metabolomics	0.23
IDPs	Refractometry	0.22
Food Preferences	NMR metabolomics	0.19
Food Preferences	Refractometry	0.19
NMR metabolomics	Refractometry	0.055

We evaluate missing data methods for CCA on each of six possible pairs of sub-studies. For each pair of sub-studies, we make 100 train-test splits ($80\% - 20\%$). Deconfounded canonical coefficients are estimated on training data using each missingness method. On test data, the complete cases are kept, deconfounded and the canonical variables are calculated using CCA coefficients from each method. The correlation between these canonical variables constructed on test data is calculated and, with best methods having largest correlation.
We also show a real world example of how our method can be used to link two studies, food-preferences and NMR metabolomics, by applying our proposed method using EM combined with SoftImpute to find the first canonical modes of the two datasets. The canonical modes are then correlated with the variables in each respective study to find links between variables from different studies.

3.3 Results

3.3.1 Simulated Data

Table 3.3 shows the results of the simulation study described in subsection 3.2.3, detailing the performance of all methods relative to the Gold Standard under different simulation settings. We see that for most settings, the two versions of our proposed method that use SoftImpute and MICE (MICE+EM and SI+EM), have results comparable to those of MICE or better. This is particularly encouraging for the version that uses SoftImpute, as this is a very computationally cheap method. Our proposed methods give very similar performance to MICE's and dominate in a few settings. The PCO approach performs by far the worst overall. Figure 3.5 shows the computational time of each method on a single synthetic data set. Our method is much faster than MICE, especially when using mean imputation and SoftImpute to handle UM. MICE takes ≈ 2 hours in total, whereas MICE+EM takes ≈ 40 minutes, and mean+EM and SoftImpute+EM both take below 1 minute.

Table 3.3: Simulation results showing the performance of each method in different settings. The results in each entry show the average of the first 10 canonical correlations on held out data relative to the same average for the Gold Standard method, which is the case where we have no missing data (see subsection 3.2.3). The median result over 100 simulations is recorded. Bold font indicates the best result for a given setting. We see that the versions of our method that use SoftImpute or MICE to impute UM perform competitively compared to MICE.

Setting			**SM Excluded**				**SM Not Excluded**						
p_{UM}	p_{SM}	σ_U	mean	PCO	SI	MICE	mean	PCO	SI	MICE	mean+EM	SI+EM	MICE+EM
0.01	0.1	0.05	0.835	0.832	0.839	0.838	0.834	0.832	0.836	**0.844**	0.834	0.839	0.838
		0.03	0.678	0.672	0.683	0.687	0.679	0.670	0.676	**0.698**	0.673	0.686	0.677
		0.02	0.569	0.545	0.574	0.574	0.565	0.541	0.543	**0.596**	0.571	0.589	0.573
	0.2	0.05	0.807	0.804	0.809	0.809	0.805	0.800	0.801	**0.811**	0.807	0.809	0.809
		0.03	0.601	0.583	0.614	0.617	0.600	0.577	0.570	**0.636**	0.600	0.615	0.614
		0.02	0.490	0.466	0.507	0.508	0.496	0.471	0.467	**0.519**	0.483	0.506	0.500
	0.4	0.05	0.737	0.725	0.744	**0.745**	0.730	0.717	0.652	0.709	0.735	0.742	0.742
		0.03	0.435	0.400	0.432	0.443	0.423	0.406	0.284	0.437	0.410	0.439	**0.445**
		0.02	0.323	0.292	0.321	**0.337**	0.311	0.315	0.205	0.302	0.308	0.330	0.326
0.05	0.1	0.05	0.803	0.766	0.824	0.826	0.802	0.766	0.820	0.830	0.803	0.825	0.820
		0.03	0.639	0.496	0.670	0.679	0.643	0.495	0.665	**0.678**	0.639	0.672	0.675
		0.02	0.502	0.344	0.533	**0.554**	0.494	0.355	0.510	0.547	0.498	0.544	0.547
	0.2	0.05	0.793	0.731	0.808	0.810	0.791	0.729	0.802	0.806	0.792	**0.811**	0.808
		0.03	0.576	0.432	0.613	0.621	0.575	0.436	0.598	**0.631**	0.577	0.615	0.613
		0.02	0.455	0.320	0.481	0.487	0.445	0.320	0.431	**0.496**	0.442	0.488	0.482
	0.4	0.05	0.659	0.493	0.681	**0.687**	0.648	0.504	0.562	0.672	0.659	0.682	0.681
		0.03	0.358	0.222	0.377	0.389	0.356	0.236	0.235	**0.423**	0.349	0.372	0.377
		0.02	0.301	0.217	0.332	0.342	0.298	0.236	0.217	**0.348**	0.309	0.335	0.338
0.15	0.1	0.05	0.777	0.494	0.811	0.811	0.776	0.499	0.809	0.810	0.776	**0.817**	0.800
		0.03	0.549	0.203	0.620	**0.629**	0.549	0.210	0.608	0.629	0.549	0.627	0.608
		0.02	0.419	0.205	0.478	**0.516**	0.419	0.209	0.453	0.501	0.427	0.487	0.502
	0.2	0.05	0.714	0.326	0.753	0.765	0.709	0.342	0.741	0.757	0.714	0.762	0.752
		0.03	0.467	0.157	0.528	**0.545**	0.484	0.162	0.469	0.533	0.467	0.540	0.518
		0.02	0.379	0.164	0.418	**0.453**	0.378	0.167	0.378	0.430	0.376	0.442	0.418
	0.4	0.05	0.595	0.129	0.644	**0.669**	0.588	0.166	0.492	0.634	0.594	0.655	0.657
		0.03	0.283	0.079	0.304	**0.351**	0.284	0.091	0.187	0.364	0.274	0.320	0.337
		0.02	0.251	0.132	0.269	**0.295**	0.240	0.134	0.206	0.284	0.237	0.261	0.272

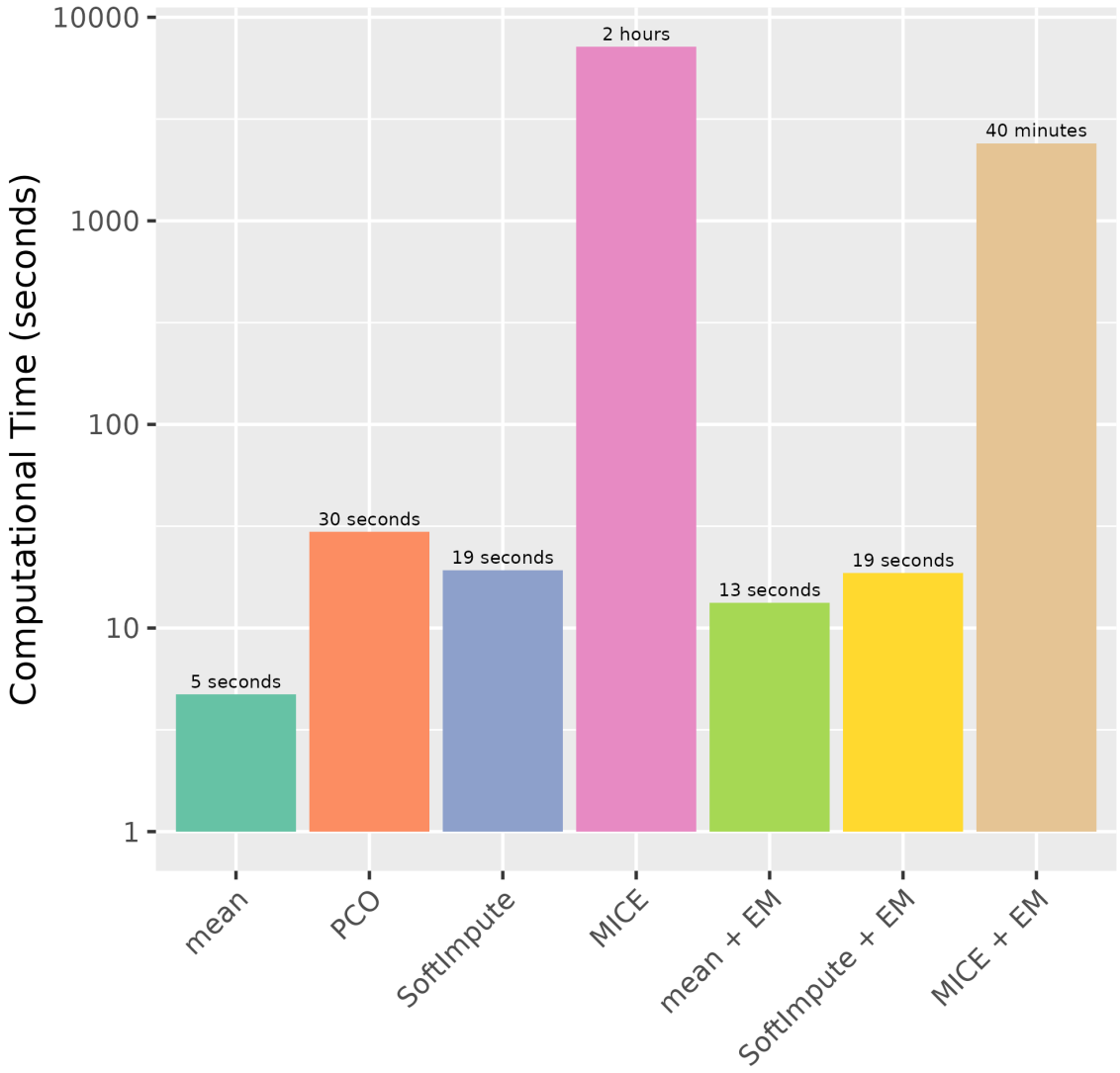


Figure 3.5: Computational time for each method on a single synthetic data set ($p_{\text{SM}} = 0.2$, $p_{\text{UM}} = 0.05$, $\sigma_U = 0.03$). The y-axis is shown on a logarithmic scale. We can see that our method is much faster than MICE, especially when using mean imputation and SoftImpute to impute UM.

3.3.2 Real Data

Table 3.4 shows the results on real data. As we can see, the three versions of our proposed method perform the best relative to other methods for pairs of data sets that include the IDPs. A reasonable explanation for this is that the IDPs have by far the most number of variables, and our proposed method is designed for high dimensional settings. Conversely, all three versions of our method have low scores for the NMR metabolomics and Refractometry pair. This can possibly be explained by the low level of correlation between the data sets, and low percentage of participants of both studies, leading to very unstable coefficients. For the two remaining comparisons, our methods perform similar to MICE. A possible explanation for this is that all other sets have ≈ 100 variables, so the dimensionality is sufficiently low for MICE to work almost optimally (see subsection 3.2.1 for details on "truncating" MICE in the high-dimensional setting). As with simulated data, we also see that the PCO approach performs the worst.

Table 3.4: Results for evaluation on real data. The scores are the observed correlations on held out, deconfounded data, between canonical variables constructed using fitted weights. A higher score is considered better. Bold font indicates the best result for a given pair of sub-studies. As we can see, MICE combined with EM has the best performance for all pairs of studies that involve the IDPs. A reasonable explanation for this is that the IDPs have by far the most number of variables, and our proposed method is designed for high dimensional settings. All three versions of our method have low scores for the NMR metabolomics and Refractometry pair. This can possibly be explained by the low level of correlation between the data sets, and low percentage of participants of both studies, leading to very unstable coefficients.

Sub-studies		**SM Excluded**				**SM Not Excluded**						
Set 1	Set 2	mean	PCO	SI	MICE	mean	PCO	SI	MICE	mean+EM	SI+EM	MICE+EM
IDPs	Food Pref.	0.215	−0.003	0.217	0.217	0.214	0.000	0.217	0.217	0.215	0.216	**0.218**
IDPs	NMR metabol.	0.113	0.012	0.113	**0.116**	0.115	0.004	0.068	0.057	0.103	0.104	**0.116**
IDPs	Refrac.	0.068	0.013	0.077	0.100	0.054	0.003	0.036	0.065	0.062	0.070	**0.113**
Food Pref.	NMR metabol.	0.303	0.300	0.303	0.303	0.298	0.293	**0.311**	0.301	0.304	0.304	0.304
Food Pref.	Refrac.	0.012	0.000	0.014	0.025	0.011	−0.007	**0.044**	0.033	0.004	0.000	0.015
NMR metabol.	Refrac.	0.215	−0.013	0.214	0.230	**0.240**	0.041	0.056	−0.001	0.111	0.120	0.177

In Figures 3.6 and 3.7, we see the variables most strongly correlated or anti-correlated with the first canonical modes of the food preference and NMR metabolomics datasets. The first canonical correlation between the two datasets is $r = 0.38$. As we

can see, the main link between the two datasets seems to be the positive correlation between a preference for fatty fish, and Omega-3 and Docosahexaenoic fatty acids. The correlation between consuming fatty fish and Omega-3 is a well established results in nutrition science (Harris et al., 2013; Milte et al., 2008). The equivalent plots for the five remaining CCA comparisons are found in the supplementary material (see B.3).

3.4 Discussion

We have proposed a new method for handling missing data for CCA. As seen from both real and simulated data, our proposed method is most useful relative to other methods in cases where the rate of UM is low, and we have also shown that it is computationally much faster than the best method, MICE. Conversely, we have shown that the proposed method performs poorly when there is a high level of UM. For those cases, SoftImpute or MICE on data where SM is excluded are better alternatives. For example, in a situation where we have two data sets with > 300 variables each, and the rate of UM is $< 3\%$ in both sets, while the rate of SM is $> 20\%$ in either of the sets, we strongly recommend the use of our method as compared to a more standard approach such as MICE, because it is much faster than MICE and at least as accurate.
In this work, we have limited ourselves to cases where CCA is performed on pairs of sub-studies. It is however, completely viable to use our proposed method for cases where we are dealing with multiple sub-studies, either as parts of $\mathbf{X}$ and $\mathbf{Y}$ or in the case of multi-set CCA (MCCA) (Kettenring, 1971; Parra, 2018), i.e., an extension of CCA with multiple data sets instead of merely two. In fact, it is entirely possible that our proposed method is more useful in these settings, as it becomes, in these cases, impossible to remove all rows with SM without reducing the number of rows to be either zero or lower than the total number of variables.
Lastly, we would like to note that our work shows that incorporating knowledge of SM in missing data methodology can yield better and faster results than standard approaches that are agnostic to structure. It is therefore necessary to study the effect of SM on analysis and develop suitable methodology for SM more broadly.
Our proposed method of covariance estimation and CCA has been released as an `R`

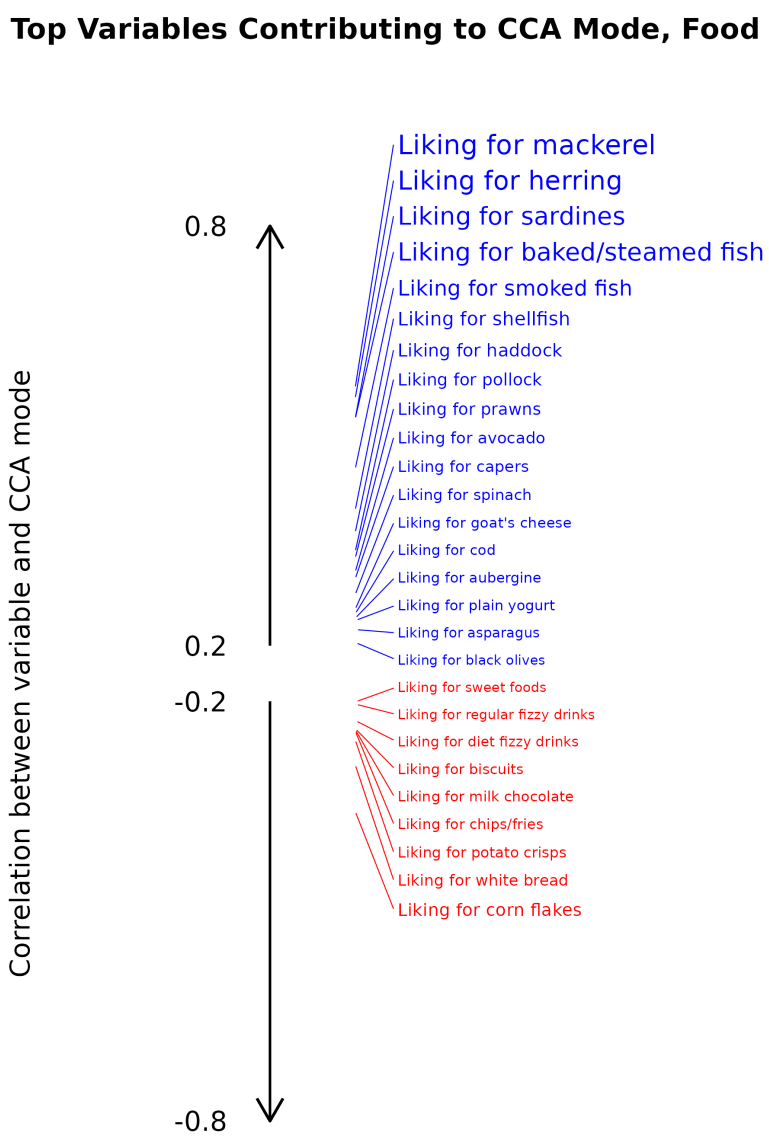


Figure 3.6: Food preferences: top variables in terms of association with the first canonical mode. We see a clear pattern; the canonical mode is positively correlated with a liking for fatty fish and other seafood, and negatively associated with a liking for a number of carbo-hydrate rich foods. The first canonical correlation is $r = 0.38$.

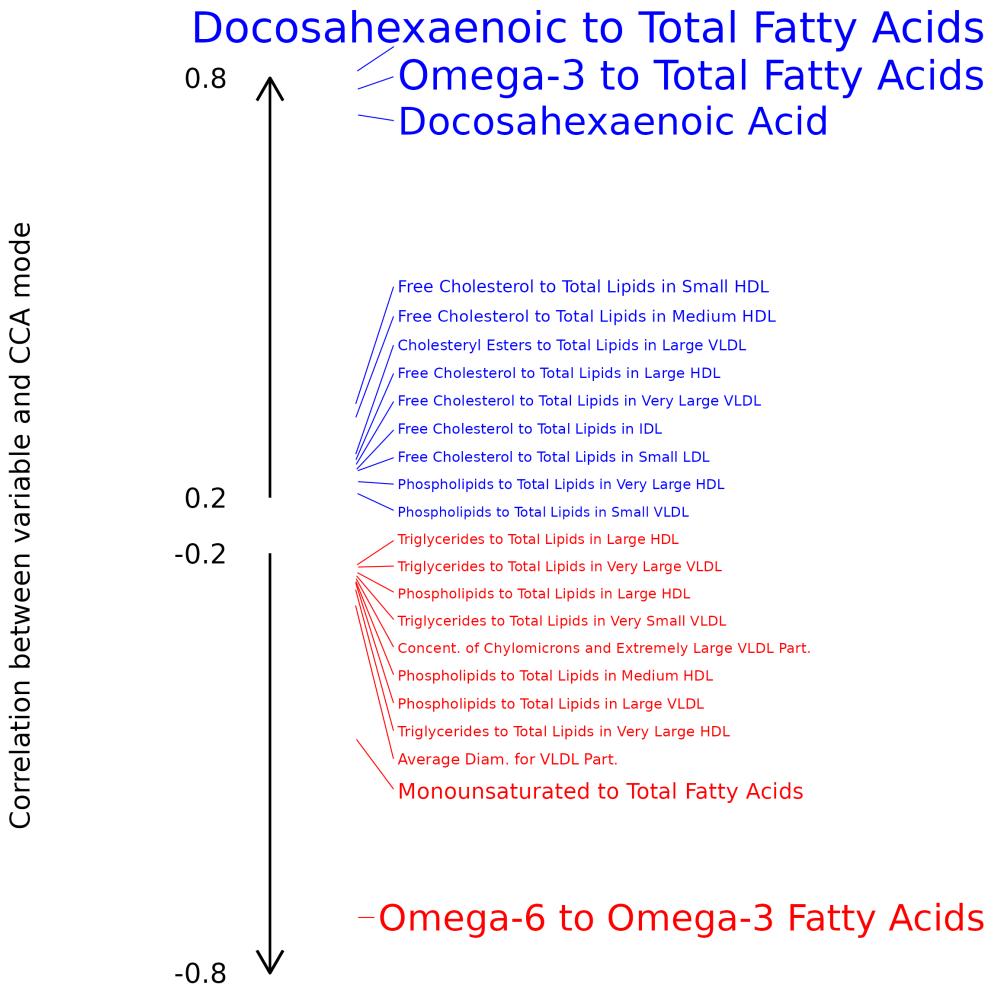


Figure 3.7: NMR metabolomics: top variables in terms of association with the first canonical mode. We see a very clear pattern; the canonical mode is strongly positively correlated with the proportion of Omega-3 and Docosahexaenonic fatty acids out of total fatty acids. The first canonical correlation is $r = 0.38$.

package and can be found in the following GitHub repository: `https://github.com/lavrad99/SM_CCA`

Chapter 4

An Improved Pipeline for Constructing UK Biobank Brain Imaging Confounds

4.1 Introduction

The UK Biobank (UKB) brain imaging analysis pipeline (Alfaro-Almagro et al., 2018) produces around 6,000 Imaging Derived Phenotypes (IDPs), and also over $10,000$ Quality Control (QC) and diagnostic variables that may be valuable as confounds when estimating associations between IDPs and other outcome variables (nIDPs, non-imaging-derived phenotypes), such as lifestyle variables, health outcomes and genetics. In previous work (Alfaro-Almagro et al., 2021) we systematically identified around 1,000 confounds that explained non-trivial variance in IDP on nIDP associations. Our previous approach to identifying UKB imaging confounds selects confounds, one-by-one, that explain a high proportion of variance across all IDPs and/or a very high proportion of variance in individual IDPs. As this is a serial and not joint process, it has the disadvantage of selecting several highly correlated confounds, leading to redundancy in the confound set. To avoid redundancy in the selected confounds we propose using a forward stepwise selection procedure, at each step selecting the next confound variable that increases R^2 (proportion of variance explained) the most over

IDPs. However, with varying missingness (different IDPs have different patterns of missing or outlier values), stepwise regression over thousands of IDPs is computationally burdensome, due to the need to fit a unique linear model for each IDP. We solve this via mean imputation of the IDPs and post-imputation correction of all R^2-estimates. We show that this does not degrade the choice of confounds, and that we can estimate a largely equivalent but much more compact set of brain imaging confounds.

4.2 Background

Handling confounding effects is a ubiquitous problem in epidemiological and biostatistical research (Brenner and Blettner, 1997; Greenland and Morgenstern, 2001; Morabia, 2011; Robins and Morgenstern, 1987; Basto et al., 2023). In classical epidemiological modelling, sex, age, socio-economic status and smoking status are typical examples of confounding variables that modify the relationship between exposure and outcome (McNamee, 2003; Papageorgiou, 2019; Hajat et al., 2021). In the case of imaging variables, the situation is much more complicated, as the imaging process itself causes nuisance effects that are associated with brain IDPs and nIDPs. UKB brain imaging confounds are constructed by initially considering a relatively small set of imaging confounds such as sex, age, site, head-motion, scanner-protocol, table position, etc., and treating their effect on IDPs as linear and additive. To fully capture the complicated relationships between confounds and IDPs, non-linear transformations of these confounds along with interaction terms and time trends are also considered. However, since the number of possible confounds is in the tens of thousands, a subset of the most relevant potential confounds is selected. In particular, a subset of non-linear confounds are first selected based on the variance explained in the IDPs (after first regressing out linear confounds); then a subset of variable interactions are selected in the same way. Finally, a set of confounds is constructed to capture trends caused by scanner drift over time or distribution shifts in the population, as well as Time of Day (ToD) trends, using the following procedure: All confounds constructed so far are regressed out of the IDPs, and the residuals are ordered by ToD and Date, separately, are smoothed, and the Principal Components (PCs) of these two smoothed

matrices are selected such that they explain 90% of the variance of the smoothed data. These PCs form the basis of the time trends confounds in the IDPs. Currently, the whole pipeline yields 966 brain imaging confounds.

Previously, because the evaluation of potential confounds was computationally expensive, each potential confound was considered for inclusion in parallel. As this can lead to mass redundancy, it would be preferable to use stepwise forward selection, building up a set of the most important confounds, and only adding additional potential confounds to this set if they explain IDP variance that is unique compared with the existing working confound set. The main obstacle to forward selection in this pipeline is heterogeneous, i.e., variable-specific missingness in the IDPs. Specifically, if we insist on using Complete Cases (CC), each IDP has a unique design matrix associated with it, and therefore we require a unique linear model for each IDP. Since there are around 6,000 IDPs, it is very computationally burdensome to perform forward selection over all of them simultaneously. A classical solution to this problem is imputation of missing data, but this biases the selection process, because the effect of imputation on R^2, i.e., the variance explained, varies between IDPs depending on their Signal to Noise Ratio (SNR) and their rate of missingness. For example, if we use an imputation method that shrinks R^2 the higher the rate of missingness, the IDPs with low missingness consequently disproportionally influence the selection process, thus biasing all downstream analysis using confounds.

4.3 Methods

For a scalar random variable y following a linear model of the form

$$y = \mathbf{x}^\top \beta + \epsilon, \tag{4.1}$$

where $\mathbf{x}$ is a multivariate random vector, β a fixed set of coefficients and ϵ a mean zero random variable independent of $\mathbf{x}$, R^2 is defined as

$$R^2 = 1 - \frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(y)}, \tag{4.2}$$

i.e., the proportion of variance in y that can be explained by $\mathbf{x}$ (Fisher, 1928; Olkin and Pratt, 1958). For a linear model fitted on data from independent draws of $\mathbf{x}$ and y, the standard point estimate of R^2 is

$$\hat{R}^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{4.3}$$

where RSS is the residual sum of squares and TSS is the total sum of squares. This is a consistent estimator of R^2 (Koerts and Abrahamse, 1969). Using the change in R^2 as the criterion for variable selection is a common modelling choice in forward selection pipelines (Lindsey and Sheather, 2010; Blanchet et al., 2008; James et al., 2023; Ghani and Ahmad, 2010). However, we cannot use the standard estimator of R^2 for this purpose if the missing data of the dependent variable y has been imputed, as this estimator is no longer consistent.

4.3.1 Correcting R^2 after mean imputation

To solve the problem of R^2 bias under imputation, we propose a simple correction of the R^2 estimate following mean imputation. If $\hat{R}^2_{\text{imp}}$ is the standard R^2-estimate calculated on mean imputed data, our proposed corrected estimate is

$$\hat{R}^2_{\text{cor}} = \frac{\hat{R}^2_{\text{imp}}}{1 - \hat{p}_{\text{miss}}}, \tag{4.4}$$

where $\hat{p}_{\text{miss}}$ is the rate of missingness in our dependent variable, i.e., our IDP. What this means intuitively is that if the rate of missingness in the dependent variable is p_{miss}, mean imputation leads to a change in R^2 by a factor of $1 - p_{\text{miss}}$. In the supplementary material, we prove that Eqn. (4.4) is a consistent estimator of the true R^2 under Missing Completely at Random (MCAR) (see Section C.1).

4.3.2 Proposed Forward Selection Method

Our forward selection method assumes the following linear model for each IDP $\mathbf{y}$:

$$\mathbf{y} = \mathbf{C}_{\text{fix}}\beta_{\text{fix}} + \mathbf{C}_{\text{pot}}\beta_{\text{pot}} + \epsilon, \tag{4.5}$$

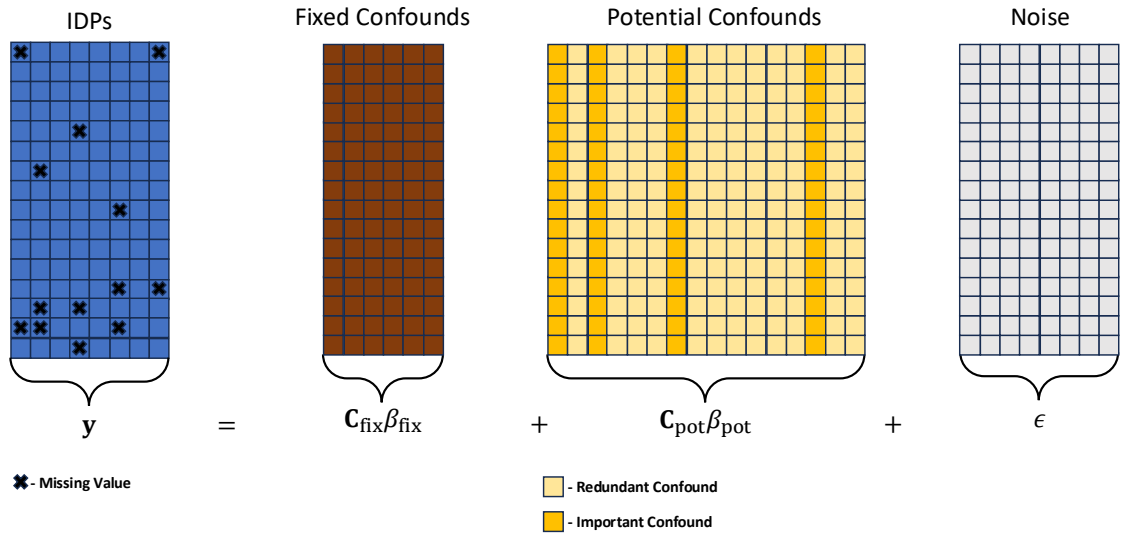


Figure 4.1: Illustration of the regression equation that is the focus of this work. We assume that only a very small subset of potential confounds are relevant to the IDPs and that there is heterogeneous missingness in the IDPs.

where $\mathbf{C}_{\text{fix}}$ is the matrix of fixed confounds (e.g. age, sex, or previously selected confounds), and $\mathbf{C}_{\text{pot}}$ is the matrix of potential confounds from which we want to select. We assume that many of the potential confounds are redundant, that missing data in the IDPs occurs under MCAR, and that there is no missing data in the confounds. Figure 4.1 illustrates this selection problem. Using our corrected estimate, we select a maximum of two variables to add to our confound set in each iteration: the variable that has the maximal change in corrected $\hat{R}^2_{\text{cor}}$ for a single variable, and the one that has the maximal average change in $\hat{R}^2_{\text{cor}}$ over all variables. We set two pre-determined thresholds τ_{max} and τ_{mean}, determining when to not include variables based on the two criteria. The algorithm stops when the maximal change in $\hat{R}^2_{\text{cor}}$ is lower than τ_{max} and the maximal average change is lower than τ_{mean}. The default values we use for these thresholds are $\tau_{\text{max}} = 0.0025$ and $\tau_{\text{mean}} = 0.0005$. If the two selected variables are too highly correlated with each other (having an absolute Pearson correlation exceeding 0.95), we only include the variable selected by the mean criterion. An essential feature of using mean imputation is that our algorithm can be implemented in a vectorised fashion, such that each iteration only consists of matrix manipulations; pseudocode can be found in the supplementary material (Section C.2), and python code is openly available at `https://git.fmrib.ox.ac.`

`uk/fsl/fastconfounds/`.

4.3.3 Double Selection

In our previous pipeline for constructing imaging confounds, we only select confounds in the context of their relationship to IDPs. However, since confounds can relate to both IDPs and nIDPs, it makes sense to also consider nIDPs when constructing imaging confounds. So-called *double selection*, i.e., selecting all variables that are predictive of either outcome or exposure, has been shown to have desirable properties in terms of false positive control and power when controlling for confounding variables in high-dimensional settings (Belloni et al., 2014). We therefore include 979 nIDPs in the selection process. The nIDPs were chosen by excluding all nIDPs with rates of missingness higher than 20% and all binary nIDPs where the most common value represents over 90% of cases. We also consider a version of the pipeline with *single selection*, i.e., where only IDPs are used for selection, and compare it to the version using double selection.

4.3.4 Proposed Pipeline

Using the corrected R^2 estimate, we construct a forward selection pipeline for constructing confounds that closely follows the previous pipeline (Alfaro-Almagro et al., 2021), while applying forward selection instead of univariate selection. The pipeline follows this procedure:

1. We construct "linear confounds", which are the raw continuous confounds and categorical confounds encoded as dummy variables.

2. We construct "non-linear confounds", three sets of transformation of all continuous linear confounds. There is a squaring of each variable (after centering); a quantile-transform to a standard normal distribution (Peterson and Cavanaugh, 2019); and a squaring of the normal-transformed variables.

3. We construct a baseline set of confounds consisting of age, sex and site, and select a subset of the remaining linear and non-linear variables using forward selection.

4. We construct "crossed terms", considering all possible variable interactions for the selected variables (i.e., second-order multiplicative interactions between all the above confounds) and again perform forward selection to select a subset.

5. To construct time trends, we mean-impute the IDPs and regress out the selected confounds.

   (a) For each site, we perform Gaussian kernel smoothing for all the IDPs' residuals over time of day. We perform PCA on the smoothed residuals and include enough PCs to explain 99% of the variance.

   (b) Using all extracted PCs, we again perform forward selection to obtain a subset of PCs.

   (c) We regress the newly selected variables out of the IDPs and perform smoothing over date. We extract PCs again and perform forward selection.

Flowcharts illustrating the old and new pipelines are shown in Figures 4.2 and 4.3. Confounds are constructed using the full $\approx 63,000$ subjects in the UK Biobank brain imaging cohort. Almost all IDPs are used in the pipeline, but 50 ASL IDPs are currently excluded for having too high rates of missingness, with some of them having over 90% missing data. For the new pipeline, double-selection is performed by adding the nIDPs to the IDP data, with the only exception being the construction of time trends, as nIDPs and IDPs do not have the same acquisition times.

4.3.5 Evaluating the New Pipeline

We evaluate the new pipeline relative to the old pipeline in a number of ways. First we evaluate the corrected estimator $\hat{R}_{\text{cor}}$ by comparing it to $\hat{R}_{\text{CC}}$, the standard estimator of R^2 using CC (i.e., the gold standard under MCAR using complete cases data only). We make this comparison over all IDPs and our subset of nIDPs as part of double-selection (see Section 4.3.3). To assess whether the two confound sets explain the same amount of variance in each variable, we compare the variance explained by the old and new confound sets in a Bland-Altman plot. Additionally, we compare the actual variance explained, not just the amount. We assess this by measuring the increase in variance explained by concatenating the two confound sets; if the same confound

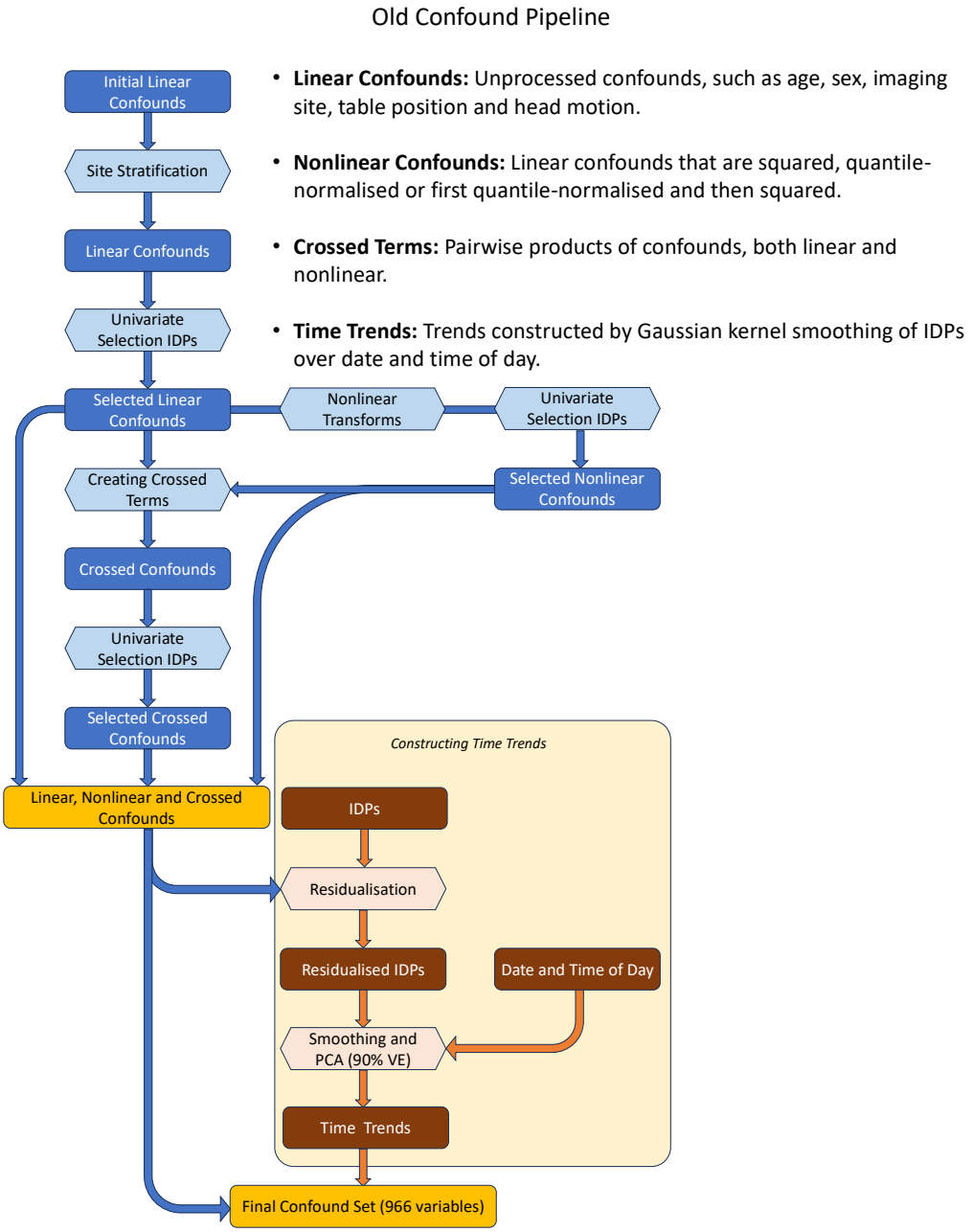


Figure 4.2: Flow chart illustrating the old confound pipeline. Unlike the new pipeline, the old pipeline uses univariate selection to include confounds, instead of multivariate forward selection. Another difference is in constructing time trends; the old pipeline includes all PCs up to 90% variance explained, regardless of their ability to explain IDPs, whereas the new pipeline performs forward selection to choose PCs. Only IDPs are used for selection, whereas the new pipeline uses nIDPs as well.

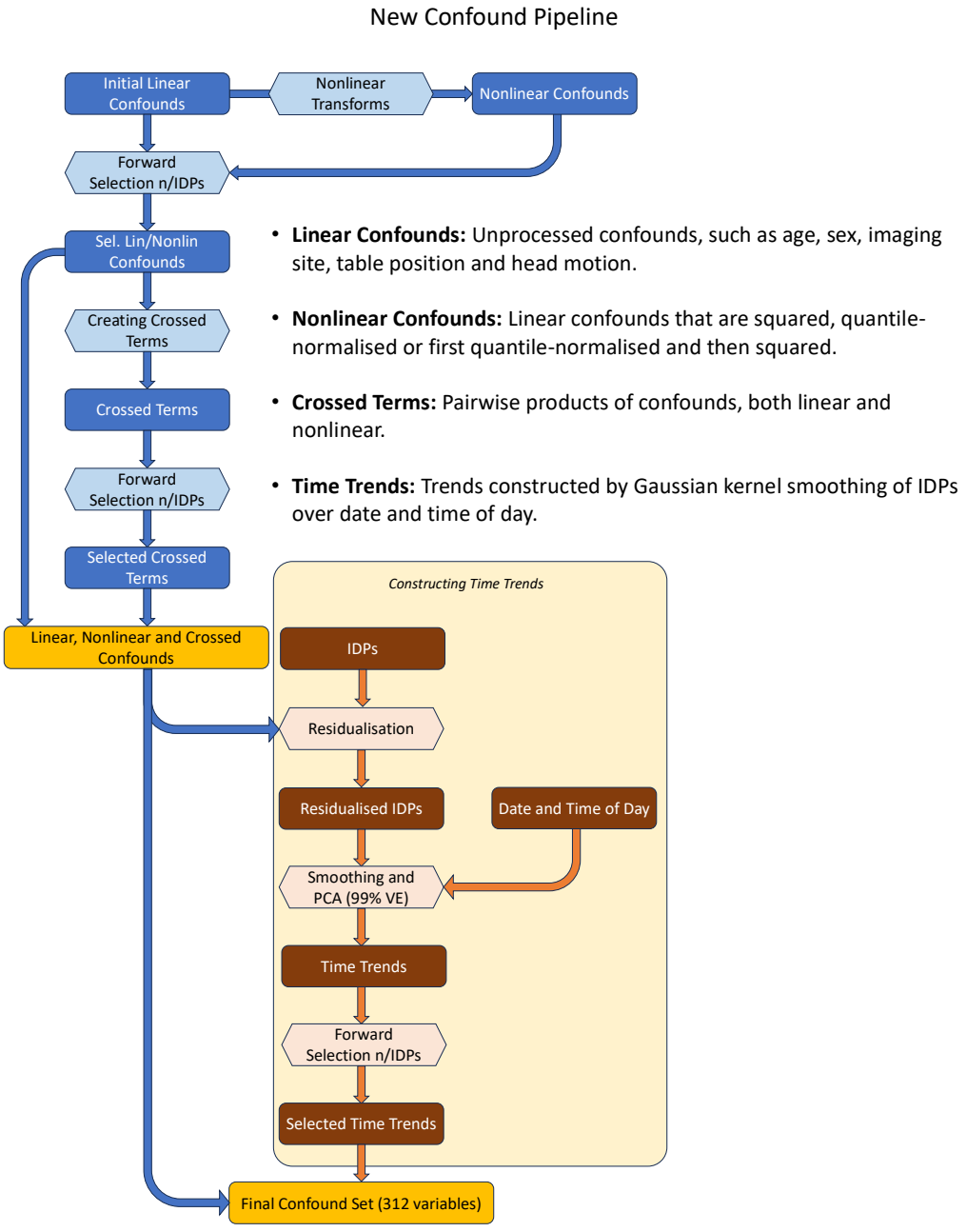


Figure 4.3: Flow chart illustrating the new confound pipeline. Unlike the old pipeline, the new pipeline uses multivariate forward selection to include confounds, instead of univariate selection. Another difference is in constructing time trends; the old pipeline includes all PCs up to 90% variance explained, regardless of their ability to explain IDPs, whereas the new pipeline performs forward selection to choose PCs. The new pipeline uses nIDPs in addition to IDPs for selection.

variance is explained, there will be no increase. We also examine the correlation of each IDPs/nIDPs after deconfounding using old vs. new confounds; a high correlation indicates that deconfounding using old and new confounds has a roughly equivalent effect.

Lastly, we compare the p-values for univariate associations between IDPs and nIDPs, when using old vs. new confounds for deconfounding. Here we also construct a Bland-Altman plot showing the agreement in $-\log$ p-values when using old and new confounds for deconfounding.

When comparing the variance explained in the IDPs/nIDPs between old and new confound sets, we use adjusted R^2 (Ezekiel, 1930) instead of the regular estimate, to account for the fact that there are different numbers of confounds in the two sets:

$$\hat{R}^2_{\text{adj}} = 1 - \frac{n-1}{n-d-1}(1-\hat{R}^2), \tag{4.6}$$

where n is the number of subjects, d the number of confounds and $\hat{R}^2$ the standard R^2 estimator.

4.4 Results

Our new pipeline yields 312 confounds, compared to 966 confounds for the old pipeline. The lower number of confounds is a result of forward selection in the new pipeline, which avoids the joint selection of highly correlated variables caused by greedy univariate selection in the old pipeline. Table 4.1 shows the number of confounds in each confound group for the old and new pipelines. The most stark difference is in the time trend confounds, where the new pipeline only constructs two confounds compared to 186 for the old pipeline. The reason for this is that in the old pipeline, after smoothing the residualised IDPs all PCs up to 90% total variance explained are included, whereas in the new pipeline, the same forward selection procedure is applied to the extracted PCs and only PCs that pass the thresholds for variance explained are included.

Figure 4.4 shows the relative difference between $\hat{R}^2_{\text{CC}}$ and $\hat{R}^2_{\text{cor}}$ plotted against the rate of missingness, for IDPs and nIDPs separately. The percentage of relative

	Old Pipeline	New Pipeline
Linear	378	64
Non-Linear	178	52
Crossed Terms	224	194
Time Trends	186	2
Total	**966**	**312**

Table 4.1: Number of confounds in each confound group, for old and new confounds. We see that the new pipeline constructs significantly fewer confounds. The difference is particularly stark for the time trend confounds. This is because the new pipeline further reduces them by applying forward selection to the PCs that are extracted from the smoothed IDP residuals.

difference here means:

$$100 \cdot \frac{\hat{R}_{\text{CC}} - \hat{R}_{\text{cor}}}{\hat{R}_{\text{CC}}.} \tag{4.7}$$

Each point in the plot represents a single IDP/nIDP and compares the estimated proportion of variance explained in the variable by the old confound set, as calculated using $\hat{R}^2_{\text{CC}}$ and $\hat{R}^2_{\text{cor}}$, plotted as percent relative difference. There is strong agreement between the two estimators at different levels of missingness, with the vast majority of estimates being within 5% relative difference of each other (99.2% for IDPs and 98.6% for nIDPs). No estimates differ more than 25% from each other.

Figure 4.5 shows the average proportion of variance explained in the IDPs and nIDPs by the old and new confound sets, as well as the proportion explained by the union of both confound sets. The two confound sets explain similar proportions of variance on average, but the variance explained is distributed differently between confound groups. In particular, the non-linear confounds explain a larger portion of variance in the new confound sets. This is expected, as the non-linear and linear confounds are selected simultaneously in the new confound pipeline, while the linear are selected before the non-linear in the old pipeline, thereby leading to a larger proportion of variance initially accounted for by linear confounds. Additionally, we see that the crossed terms explain relatively more variance in the new confound set, and the time trends explain relatively less variance for the new confound set. Importantly, there is no notable increase in variance explained when combining new and old confounds, meaning that almost all important mutual variance between confounds and

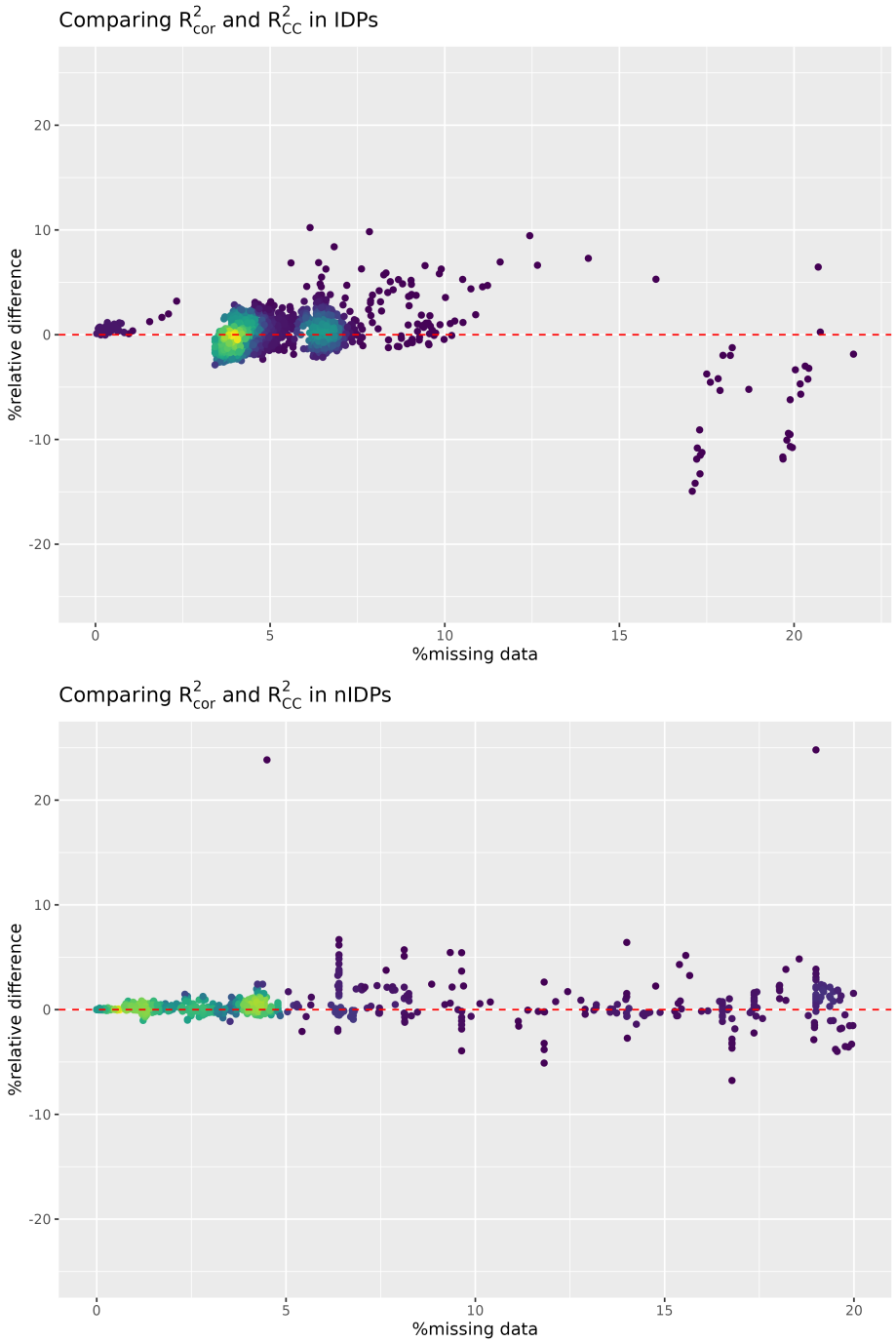


Figure 4.4: Scatter plots of missingness rate against the percentage difference between $\hat{R}^2_{\text{cor}}$ and $\hat{R}^2_{\text{CC}}$. For the vast majority of IDPs and nIDPs, the two estimates are within 5% relative difference of each other (99.2% of IDPs and 98.6% of nIDPs), which is sufficiently close for our purposes.

IDPs/nIDPs is captured by the new confound set.

Figure 4.6 illustrates the proportion of variance explained in individual IDPs and nIDPs by old and new confound sets. We see very good alignment between the two confound sets across IDPs and nIDPs, with generally very small differences in the proportion of variance explained. To further verify that the two confound sets have the same effect, we also look at the correlation between IDPs/nIDPs that have been deconfounded using old and new confounds. The correlation is generally very high with the vast majority having correlations above 0.95 (98.4% for IDPs and 99.9% for nIDPs). Lastly, Figure 4.7 shows a Bland-Altman plot of $-\log_{10}$ p-values of univariate IDP–nIDP associations after deconfounding using new and old confounds. We see very strong agreement between p-values, especially when looking at low p-values. For 98% of IDP on nIDP associations, the two p-values are within one order of magnitude of each other. Of those p-values which are below the Bonferroni threshold controlling the Family Wise Error Rate (FWER) at 5% ($p \approx 10^{-8}$) for either old or new confounds, 54% are within one order of magnitude of each other, and 89.0% are within three orders of magnitude of each other. Two additional plots comparing the two confound sets are found in Section C.3 of the Supplementary Material.

We compared versions of the new pipeline in which single- and double-selection are used, respectively. We found very little difference between the two pipelines, with the version using double selection having a slight but noticeable advantage in explaining variance in the nIDPs. Plots comparing the two pipelines are found in the Supplementary Material, Section C.4.

4.5 Discussion

We have proposed an improved method for confound selection on large-scale epidemiological datasets. The new confound set constructed using our pipeline consists of 312 confounds compared the previous 966 confounds, while having a very similar effect as the old set. Our proposed method is sufficiently fast to be used for forward selection (incremental forward selection of confounds, one at a time) using UK Biobank brain imaging data (tens of thousands of subjects, tens of thousands of potential confounds,

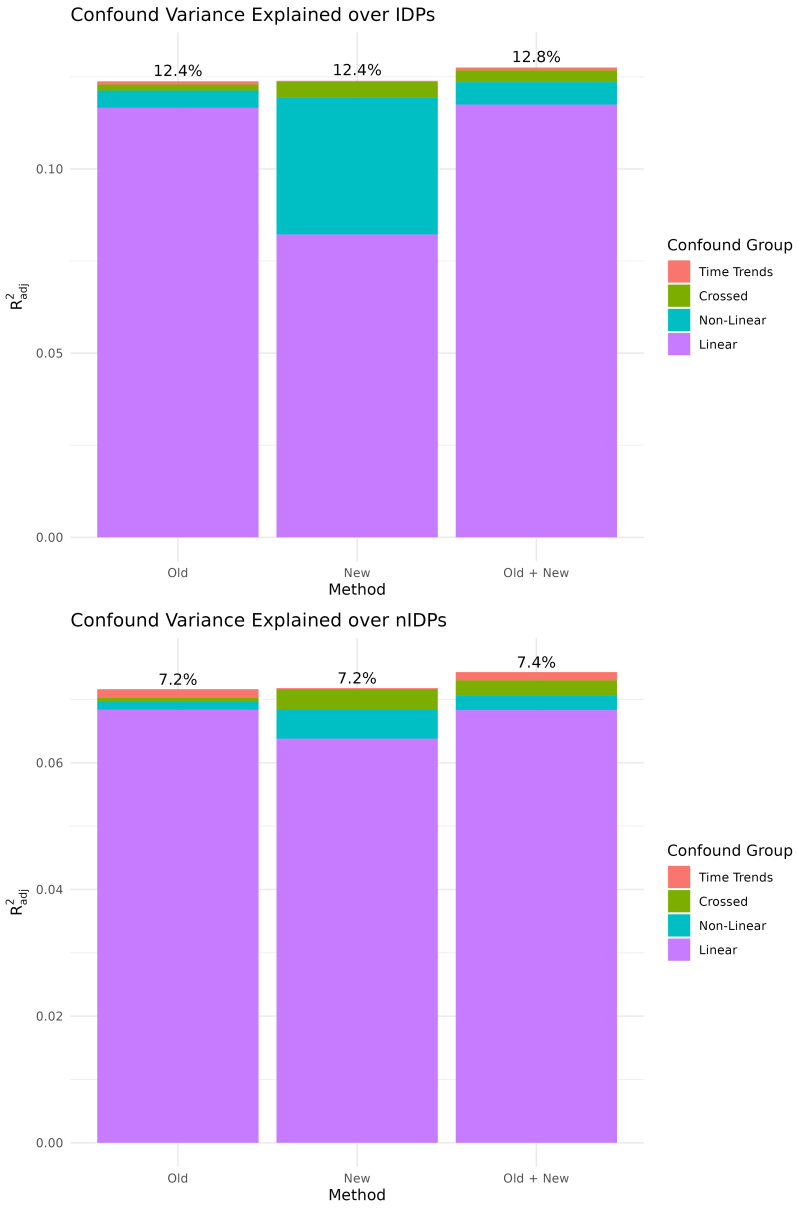


Figure 4.5: Bar plots showing the adjusted R^2 for the two different confound sets and for the combined set of confounds. The variance is distributed differently between different confound groups for the two sets, but the proportion explained is very similar. Importantly, there is no notable increase in variance explained when combining the confound groups, meaning that almost all important mutual variance between confounds and IDPs/nIDPs is captured by the new confound set.



Figure 4.6: Bland-Altman plots illustrating the difference in variance explained in each IDP/nIDP by the old and new confound sets. We see that the two confound sets generally explain very similar proportions of variance in individual IDPs and nIDPs.

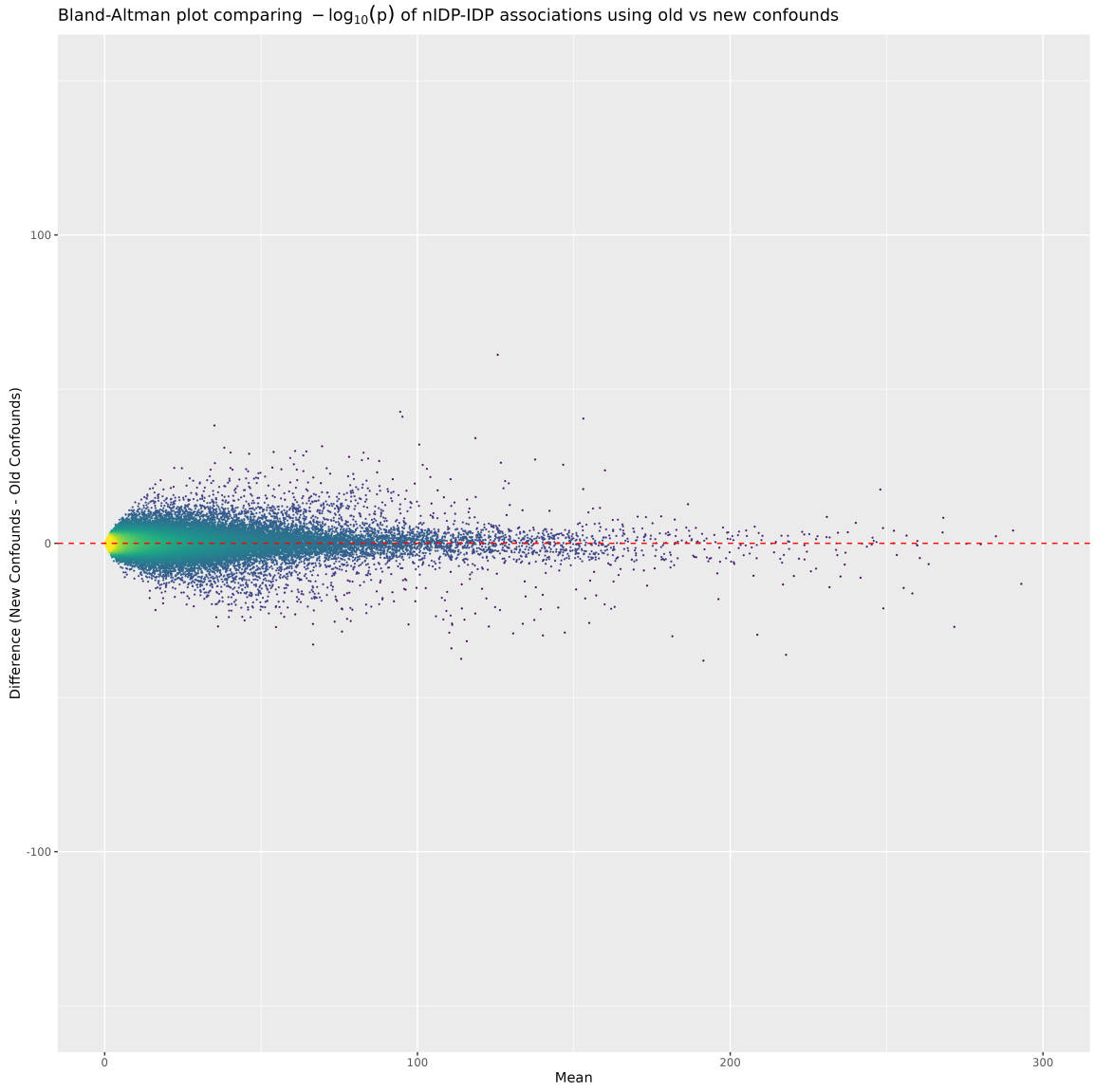


Figure 4.7: Bland-Altman plot of $-\log_{10}(p)$ for univariate IDP on nIDP associations using old and new confounds. The plot shows very strong agreement between p-values obtained after deconfounding with either confound set.

and thousands of IDPs). It has the advantage of being neutral to the rate of missingness in IDPs and nIDPs, meaning that it does not systematically inflate/deflate the estimates of R^2 for variables with high/low missingness. This is not the case when using uncorrected estimates of R^2 on data that has been imputed, where (depending on the imputation method) higher missingness leads to inflation/deflation of the R^2 estimate.

The main limitation of this work is reliance on the assumption of MCAR, which has been shown to not hold for nIDPs in general, as seen in Chapter 2. Future work includes developing similar methodology that only assumes Missing at Random (MAR), a weaker missingness criterion, and is therefore applicable to a larger scope of feature selection problems. Another limitation of our pipeline is that it uses classical forward selection, which has been shown to yield inflated false positive rates. There have been solutions proposed to handle this problem (Blanchet et al., 2008; Tibshirani et al., 2016; Benjamini and Gavrilov, 2009), and future work includes exploration of such options. However, in large sample settings, such as that considered by the current work, it has been shown that this problem is largely mitigated (Li et al., 2007).

The confound set produced by our pipeline is designed to be used for mass-univariate analysis of associations between IDPs and nIDPs (although the resulting confounds should be relevant for multivariate analyses too, for example, canonical correlation analysis). It therefore includes variables that explain meaningful variance in any individual IDP/nIDP. Hence, when considering a smaller problem, for example mass-univariate analysis of all IDPs on one nIDP, many of the confounds might be redundant and, consequently, inefficient power-wise. This is especially true when considering a smaller subsets of subjects. Future work therefore also includes efficient deconfounded inference when considering subsets of subjects and variables; however, as we would recommend regenerating the confound set when using a highly reduced set of subjects, that should ameliorate this concern, and the fact that the new pipeline is computationally efficient makes this easily feasible. When regenerating confounds on a reduced subset of subjects, the choice of value for $\tau_{\max}$ and τ_{mean} is an important design detail, as these thresholds prevent the inclusion of noisy confounds. We recommend the use of our thresholds for the full data set $\tau_{\max} = 0.0025$ and $\tau_{\text{mean}} = 0.0005$ as a default starting point, but if the subset is much smaller than

$n \approx 50000$, larger thresholds might be necessary. In particular, using a smaller subset of subjects should not result in a significantly larger number of confounds, so if the pipeline returns much more than ≈ 300 confounds, it should be rerun with higher thresholds. More optimized ways of choosing $\tau_{\max}$ and τ_{mean} is also an area of future inquiry.

Chapter 5

Controlling for confounds in UK Biobank brain imaging data with small subsets of subjects

5.1 Background

The UK Biobank brain imaging cohort has recently scanned its 100,000-th subject and researchers now have access to data well over $85,00$ subjects. This resource has, among other things, yielded invaluable understanding of the links between imaging and non-imaging variables (Douaud et al., 2022; Dutt et al., 2022; Elliott et al., 2017; Topiwala et al., 2021, 2023). When establishing such links, it is crucial to control for confounding factors to avoid the detection of spurious associations. In Alfaro-Almagro et al. (2021), a set of imaging confounds are identified that are predictive of Imaging Derived Phenotypes (IDPs) and/or non-Imaging Derived Phenotypes (nIDPs). The following regression equation describes the assumed relationship between some univariate exposure $\mathbf{x}$, univariate outcome $\mathbf{y}$ and confound matrix $\mathbf{C}$:

$$\mathbf{y} = \mathbf{C}\beta_C + \mathbf{x}\beta_x + \epsilon, \tag{5.1}$$

where ϵ is assumed to be homoscedastic, uncorrelated white noise with $\text{Var}(\epsilon_i) = \sigma^2$. In Neuroimaging research, $\mathbf{x}$ is usually an nIDP and $\mathbf{y}$ an IDP or vice versa. We are interested in testing null-hypotheses on the form $H_0 : \beta_x = 0$ against $H_1 : \beta_x \neq 0$. The most common approach to performing this test is by way of a standard t-test on the coefficient $\hat{\beta}_x$ obtained by Ordinary Least Squares (OLS) regression. This simple approach is appropriate when the sample size n is equal to the full number of subjects in the UKB Brain imaging cohort, but when working on a subset of all subjects, which is common in Neuroimaging research, this approach can be problematic in several ways. Firstly, this approach is not at all possible when $n < p$, where p is the number of confounds. Secondly, if n is not much larger than p, i.e., if the design matrix of the OLS regression is close to square-shaped, the standard approach can be severely underpowered. This is because the variance of $\hat{\beta}_x$ is

$$\text{Var}(\hat{\beta}_x) = \frac{\sigma^2}{\left(1 - \hat{R}^2_{\text{OLS}}(\mathbf{C}, \mathbf{x})\right) \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{5.2}$$

where $\hat{R}^2_{\text{OLS}}(\mathbf{C}, \mathbf{x})$ is the proportion of variance explained when regressing $\mathbf{x}$ on $\mathbf{C}$. In other words, the variance of $\hat{\beta}_x$ is proportional to the Variance Inflation Factor (VIF) of $\mathbf{x}$

$$\text{VIF}_x = \frac{1}{1 - \hat{R}^2_{\text{OLS}}(\mathbf{C}, \mathbf{x})}. \tag{5.3}$$

Since the proportion of variance explained is severely inflated when n is not much larger than p, this will lead to high standard errors for $\hat{\beta}_x$, which in turn leads to low power. Several approaches have been formulated to tackle these problems. These methods typically assume some level of sparsity in the coefficients β_C. In this work, we describe a number of these methods and evaluate them using simulated exposure and outcome variables, and real confound data for different sample sizes, levels of sparsity of β_C and levels of partial correlation between outcome and exposure. The goal is to find the method with the highest statistical power, while having a controlled False Positive Rate (FPR). We also show an illustrative example of the use of the most suitable method from the simulation study on real data, by performing mass univariate testing of the association between IDPs and Alzheimer's Disease (AD).

5.1.1 UKB Imaging Confounds

UK Biobank Brain Imaging confounds are constructed to control for nuisance effects in the IDPs caused by the participants physical and demographic attributes, the technology and protocol used for imaging and miscellaneous drift effects over time. These confounds include, among other things, sex, age, site, head-motion and scanner-protocol. Non-linear transformations of confounds and interaction terms are also considered to ensure that as much nuisance variance as possible is captured. These confounds are also correlated with nIDPs (Alfaro-Almagro et al., 2021), which means that they are more than just nuisance variables which cause regression dilution; if not controlled for, they cause detection of spurious associations between IDPs and nIDPs.

5.2 Methods

In this sub-section, we describe a range of methods that are used for high-dimensional inference that we consider as candidates for use on UKB Brain Imaging Data.

5.2.0.1 Principal Component Regression

Using Principal Component Analysis (PCA) to reduce the dimensionality of confound data is an approach commonly used in Genome Wide Association Studies (GWAS), where principal components of the kinship matrix are included in multivariate regression models to control for confounding effects caused by relatedness of the participants (Thomson and McWhirter, 2017). We include two versions of this approach, `PCA-OLS1` and `PCA-OLS2`, in our simulation study. `PCA-OLS1` uses the following procedure for high-dimensional inference:

1. Perform PCA on confound matrix $\mathbf{C}$.
2. Include Principal Components (PCs) by the size of corresponding singular values in an OLS model with outcome $\mathbf{y}$ as dependent variable and exposure $\mathbf{x}$ as one of the independent variables. Stop including PCs once the model has only 20 error degrees of freedom remaining, or there are no more PCs remaining to include.

3. Use the standard OLS estimator on this variable set to test the hypothesis $H_0 : \beta_x = 0$

This is our baseline high-dimensional inference approach, as it is as close to the default OLS approach as possible. Indeed, for settings where $n - p - 2 \geq 20$, this approach is exactly equivalent to OLS. `PCA-OLS2` is a less conservative approach and uses the following procedure:

1. Perform PCA on confound matrix $\mathbf{C}$.

2. Include Principal Components (PCs) by the size of corresponding singular values in an OLS model with outcome $\mathbf{y}$ as dependent variable and exposure $\mathbf{x}$ as one of the independent variables. Stop including PCs once the PCs explain 95% of the variance in $\mathbf{C}$, or when the model reaches 20 remaining error degrees of freedom.

3. Use the standard OLS estimator on this variable set to test the hypothesis $H_0 : \beta_x = 0$

`PCA-OLS2` is less conservative than `PCA-OLS1`, as it includes a lower number of PCs in the multivariate regression. Both approaches rely to different extents on the *manifold hypothesis*, i.e., that our high-dimensional confound data has a latent low dimensional structure that we can discover using PCA (Whiteley et al., 2022).

5.2.0.2 Uncorrected Ridge Regression p-values

Coefficient estimates of the Ridge Regression (RR) estimator are biased and can therefore not be directly used for exact parametric inference. However, under certain conditions, the bias in the estimator becomes negligible compared to its variance under the null hypothesis, giving us coefficient t-tests that are close to unbiased (Halawa and El Bassiouni, 2000). We use approximate t-test detailed in Cule et al. (2011). Assuming we have a linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{5.4}$$

where ϵ is white noise with variance σ^2, the null hypothesis $H_0 : \beta_j = 0$ is tested using the following approximate t-statistic:

$$t^{(\lambda)} = \frac{\hat{\beta}_j^{(\lambda)}}{\text{se}(\hat{\beta}_j^{(\lambda)})}, \tag{5.5}$$

where $\hat{\beta}^{(\lambda)} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$ is the RR estimate of β using penalty λ, and $\text{se}(\hat{\beta}_j^{(\lambda)})$ is the square root of the jth diagonal element of the matrix

$$\text{Var}(\hat{\beta}^{(\lambda)}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}. \tag{5.6}$$

$\hat{\sigma}^2$ is estimated in the following way:

$$\hat{\sigma}^2 = \frac{(y - \mathbf{X}\hat{\beta}^{(\lambda)})^\top(y - \mathbf{X}\hat{\beta}^{(\lambda)})}{\nu}, \tag{5.7}$$

where the error degrees of freedom ν is set to be $\nu = n - \text{trace}(2\mathbf{H} - \mathbf{H}\mathbf{H}^\top)$ for "hat matrix" $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top$. $t^{(\lambda)}$ is assumed to approximately follow Student's t-distribution with degrees of freedom $n - \text{trace}(\mathbf{H})$.

The most appropriate method of choosing penalty parameter λ is still a hotly debated subject in the scientific community (Ding et al., 2018). We allow for two selection methods, one proposed in Cule and De Iorio (2013) and a standard approach by selecting λ through 5-fold Cross-Validation. The two methods are referenced as `Ridge.Cule-DeIorio` and `Ridge.CV` respectively.

5.2.0.3 Ridge Regression Permutation Testing

Hemerik et al. (2021) outline three different RR-based permutation tests for confound adjusted association. The first two tests are based on the Freedman-Lane permutation testing procedure (Freedman and Lane, 1983) and the third test is based on a permutation test by Kennedy where both outcome and exposure are residualised with respect to confounds (Winkler et al., 2014). Each method is described in detail in the original article, but we will only go through the first Freedman-Lane based method `Ridge.PT.FLH1`. Assuming the model described in equation 5.1, we select penalty term λ_x for regression of $\mathbf{x}$ on $\mathbf{C}$ and λ_y for regression of $\mathbf{y}$ on $\mathbf{C}$ using cross-validation. The absolute sample correlation between the residuals of both regressions is estimated:

$$T_0 = |\rho\left(\mathbf{R}_{\lambda_x}\mathbf{x}, \mathbf{R}_{\lambda_y}\mathbf{y}\right)|, \tag{5.8}$$

where $\mathbf{R}_{\lambda_x} = \mathbf{I}_n - \mathbf{H}_{\lambda_x} = \mathbf{I}_n - \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \lambda_x\mathbf{I}_p)^{-1}\mathbf{C}^\top$ and $\mathbf{R}_{\lambda_y} = \mathbf{I}_n - \mathbf{H}_{\lambda_y} = \mathbf{I}_n - \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \lambda_y\mathbf{I}_p)^{-1}\mathbf{C}^\top$. For b subsequent permutations $\mathbf{P}_1, ..., \mathbf{P}_b$, the following statistic is calculated

$$T_j = |\rho\left(\mathbf{R}_{\lambda_x}\mathbf{x}, (\mathbf{H}_{\lambda_y} + \mathbf{P}_j\mathbf{R}_{\lambda_y})\mathbf{y}\right)|, \tag{5.9}$$

for $j = 1, ..., b$. The estimated p-value is

$$\hat{p}_{\text{PT}} = \frac{1}{b+1}\sum_{j=0}^{b} 1\{T_0 \leq T_j\}. \tag{5.10}$$

The two methods based on the Freedman-Lane procedure are referred to as `Ridge.PT.FLH1` and `Ridge.PT.FLH2`, and the one based on double residualisation is referred to as `Ridge.PT.DR`.

5.2.0.4 Naive Feature Selection

Feature selection is a common pre-processing step to reduce the dimensionality of inference tasks. The post-selection estimator using the Least Absolute Selection and Shrinkage Operator (LASSO) has been proposed as a viable option for high dimensional inference tasks (Zhao et al., 2021). It works in the following way:

1. Fit a LASSO regression model predicting $\mathbf{y}$ using $\mathbf{C}$.
2. Create a reduced set of confounds $\tilde{\mathbf{C}}$, containing only variables that have non-zero coefficient estimates in the LASSO regression.
3. Fit an OLS model regressing $\mathbf{y}$ on $\tilde{\mathbf{C}}$ and $\mathbf{x}$. If the model has insufficient degrees of freedom, i.e., lower than 20, we use PCA on $\tilde{\mathbf{C}}$ as in sub-section 5.2.0.1.

Since LASSO is sometimes not appropriate for strongly correlated data, and there are confounds in our data set that are highly correlated, we also use ElasticNet to perform selection instead of LASSO. We call the two methods `Naive.LASSO` and `Naive.ElasticNet`. All tuning parameters are selected via cross-validation.

5.2.0.5 Double Selection

Belloni et al. (2014) propose a post-selection estimator where variables that are predictive of $\mathbf{x}$ and/or $\mathbf{y}$ are selected, likewise using LASSO. The selection process works in the following way:

1. Fit a LASSO regression model predicting $\mathbf{x}$ using $\mathbf{C}$.
2. Fit a LASSO regression model predicting $\mathbf{y}$ using $\mathbf{C}$.
3. Create a reduced set of confounds $\tilde{\mathbf{C}}$, containing only variables that have non-zero coefficient estimates in either LASSO regression.
4. Fit an OLS model regressing $\mathbf{y}$ on $\tilde{\mathbf{C}}$ and $\mathbf{x}$. If the model has insufficient degrees of freedom, i.e., lower than 20, we use PCA on $\tilde{\mathbf{C}}$ as in sub-section 5.2.0.1.

It is very similar to the naive selection estimator, but with two key advantages. Firstly, it increases power by more strongly orthogonalising $\mathbf{x}$ with respect to confounds as we include more confounds that are highly predictive of $\mathbf{x}$. Secondly, FPR control is improved by including variables that are strongly predictive of $\mathbf{x}$ and weakly predictive of $\mathbf{y}$. Failure to include such variables in the single selection estimator (see sub-section 5.2.0.4) leads to inflation in the FPR.

We refer to this method as LASSO Double Selection (`LADS`). As with the single selection estimator, we might expect to see some benefit in using ElasticNet over LASSO, and therefore also include an estimator which we refer to as ElasticNet Double Selection (`ENDS`).

5.2.0.6 De-Sparsified LASSO

De-Sparsified or de-biased LASSO is a process by which bias in the LASSO estimator is corrected so that inference can be performed on the parameter estimates (Chevalier, 2020). Assuming a standard linear model as in equation 5.4, Javanmard and Montanari (2014) propose the following estimator:

$$\beta_{\text{DESLA}} = \beta^{(\lambda)} - \frac{1}{n}\mathbf{M}\mathbf{X}^{\top}(\mathbf{y} - \mathbf{X}\beta^{(\lambda)}), \tag{5.11}$$

where $\beta^{(\lambda)}$ is a LASSO coefficient estimate and $\mathbf{M}$ is an estimate of the inverse of the covariance matrix of $\mathbf{X}$. It is easy to see intuitively why this correction generally reduces the bias in the LASSO estimate $\beta^{(\lambda)}$. If $\mathbb{E}[\beta^{(\lambda)}] = \beta + \gamma$, i.e., $\beta^{(\lambda)}$, has bias γ, then $\mathbb{E}[\beta_{\text{DESLA}}] = \beta + \left(\mathbf{I}_{p+1} - \mathbf{M}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\right)\gamma$. Since $\mathbf{M}$ is an approximation of the precision matrix of $\mathbf{X}$, it follows that $\mathbf{I}_{p+1} - \mathbf{M}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)$ will be close to the null matrix, thus reducing the bias. The estimator has approximately the following covariance matrix:

$$\text{Cov}(\beta_{\text{DESLA}}) \approx \frac{1}{n^2}\hat{\sigma}^2\mathbf{M}\mathbf{X}^\top\mathbf{X}\mathbf{M}^\top, \tag{5.12}$$

where $\hat{\sigma}^2$ is an estimate of the residual variance.

The choice of method for finding λ and $\mathbf{M}$ are crucial details in constructing this estimator. Javanmard and Montanari propose a convex program for determining $\mathbf{M}$ and a heuristic for determining λ which are the default settings in their `R` function `SSLasso`. A much faster and more conservative version of this estimator has been proposed in Boot and Nibbering (2017) that uses the Moore-Penrose pseudo-inverse of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ as $\mathbf{M}$. For this version of the de-sparsified LASSO estimator, we use cross-validation to select λ. The two methods are referred to as `DESLA-JM` and `DESLA-MP`. De-sparsified LASSO is generally not used when $n > p$, so we limit the use of these methods to scenarios where $n < p$, in the simulation study.

5.2.1 Simulation Method

We are interested in the FPR control and power of all estimators under different levels of sparsity, sample size and strength of confounding. We design a simulation framework that uses real, standardised UKB imaging confound data $\mathbf{C}$ with simulated outcomes $\mathbf{y}$ and exposures $\mathbf{x}$. Specifically, we assume the following system of regression equations for describing $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x} = \mathbf{C}\beta_x + \mathbf{x}_0 \tag{5.13}$$

$$\mathbf{y} = \mathbf{C}\beta_y + \rho_{xy|C}\mathbf{x}_0 + \sqrt{1 - \rho^2_{xy|C}}\mathbf{y}_0, \tag{5.14}$$

where $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and $\mathbf{y}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ are independent from each other and all other variables. Here, $\mathbf{C}\beta x$ and $\mathbf{C}\beta y$ is the confounding signal that goes into $\mathbf{x}$ and $\mathbf{y}$ respectively, $\mathbf{x}_0$ is the variance component in $\mathbf{x}$ that is independent from confounds, and $\mathbf{y}_0$ is the variance component in $\mathbf{y}$ that is independent from both confounds and $\mathbf{x}$. Since $\mathbf{x}_0$ and $\mathbf{y}_0$ are simulated to have unit variance, $\rho_{xy|C}$ is the partial correlation between $\mathbf{x}$ and $\mathbf{y}$ after adjusting for confounds, which is precisely the association that we wish to test.

β_x and β_y are also simulated to have varying degrees of approximate sparsity and correlation between each other. The reason we are also interested in sparsity is that the imaging confounds are unlikely to all be equally relevant to each variable of interest, and this can be leveraged to increase inference power. We induce correlation between β_x and β_y, because some variables, such as age, are likely to be strongly associated with both outcome and exposure, which makes it unrealistic to simulate them independently. Specifically, for each confound j, we simulate

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{uv} \\ \rho_{uv} & 1 \end{pmatrix} \right). \tag{5.15}$$

We then set $\beta_{xj} = \text{sign}(u_j)|u_j|^\delta$ and $\beta_{yj} = \text{sign}(v_j)|v_j|^\delta$, where δ is a positive integer. Depending on the size of δ, this introduces some controlled level of approximate sparsity in β_x and β_y, where a higher δ means higher sparsity. The correlation ρ_{uv} is mathematically calibrated so that $\rho(\beta_{xj}, \beta_{yj}) = \rho_\beta$, for some pre-chosen ρ_β. A higher value of ρ_β is more likely to cause false positives, since it means that more confounds are strongly predictive of both outcome and exposure. β_x and β_y are also normalised so that the confounds explain 25% of variance in both $\mathbf{x}$ and $\mathbf{y}$.

The number of imaging confounds that we use is $p = 754$ and the sample size n is varied to be $n = 50, 100, 200, 500, 1000, 2000$. The sparsity parameter δ is varied to be $\delta = 1, 2, 3, 5, 7, 10$, and ρ_β is varied to be $\rho_\beta = 0, 0.25, 0.5$, representing weak, intermediate and strong confounding. Finally if $n \leq 1000$, we vary $\rho_{xy|C}$ to be $\rho_{xy|C} = 0, 0.05, 0.1, 0.2, 0.4$, and if $n = 2000$, we let $\rho_{xy|C} = 0, 0.01, 0.02, 0.05, 0.1$. This is because smaller effect sizes are more relevant to the setting of $n = 2000$. $\rho_{xy|C} = 0$ means that there is no association between $\mathbf{x}$ and $\mathbf{y}$ after adjusting for confounds and

is the setting that is used to evaluate the FPR control of our methods. Each setting is evaluated over 1000 random simulations to calculate power and FPR.

Our simulation framework differs from previous work on the topic, as it produces approximate sparsity in confounding effects β_C, rather than exact sparsity (Zhao et al., 2021; Hemerik et al., 2021). We see this as a more realistic setting than exact sparsity, as confounding effect sizes are unlikely to be truly 0 in reality, but may vary heavily in magnitude. Most previous simulation studies also treat the confounding effects as block homogeneous, i.e., that large blocks of confounds have the exact same effect. This differs significantly from real data, where different confounds have very different effect sizes.

5.2.2 Real Data Demonstration

To demonstrate the use of high-dimensional inference methodology on real data, we perform mass univariate testing between 3939 IDPs and one nIDP, adjusting for all imaging confounds. Here, the IDPs are modelled as outcome and the nIDP as exposure. The nIDP that we choose is AD and the data consists of 64 AD subjects and 128 age- and sex-matched controls, giving us a sample size of $n = 192$ in total. The mass univariate analyses are also performed using the baseline method, `PCA-OLS1`, for comparison.

5.3 Results

5.3.1 Simulation Results

The simulation results are too extensive to present in the main text of this thesis, considering that we evaluate $6 \cdot 6 \cdot 3 \cdot 5 = 540$ simulation settings. We therefore focus on a relatively small subset of simulation settings. Being conservative, we fix $\rho_\beta = 0.5$, which is the setting most prone to cause false positives for all methods. We vary the sample size to be $n = 50, 100, 200, 500, 1000, 2000$, covering all sample sizes of interest, and we vary the sparsity parameter δ to be $\delta = 1, 3, 7$, representing a dense, medium and sparse setting. The results for the same simulation settings but with $\rho_\beta = 0, 0.25$ are in the supplementary material (see D.1).

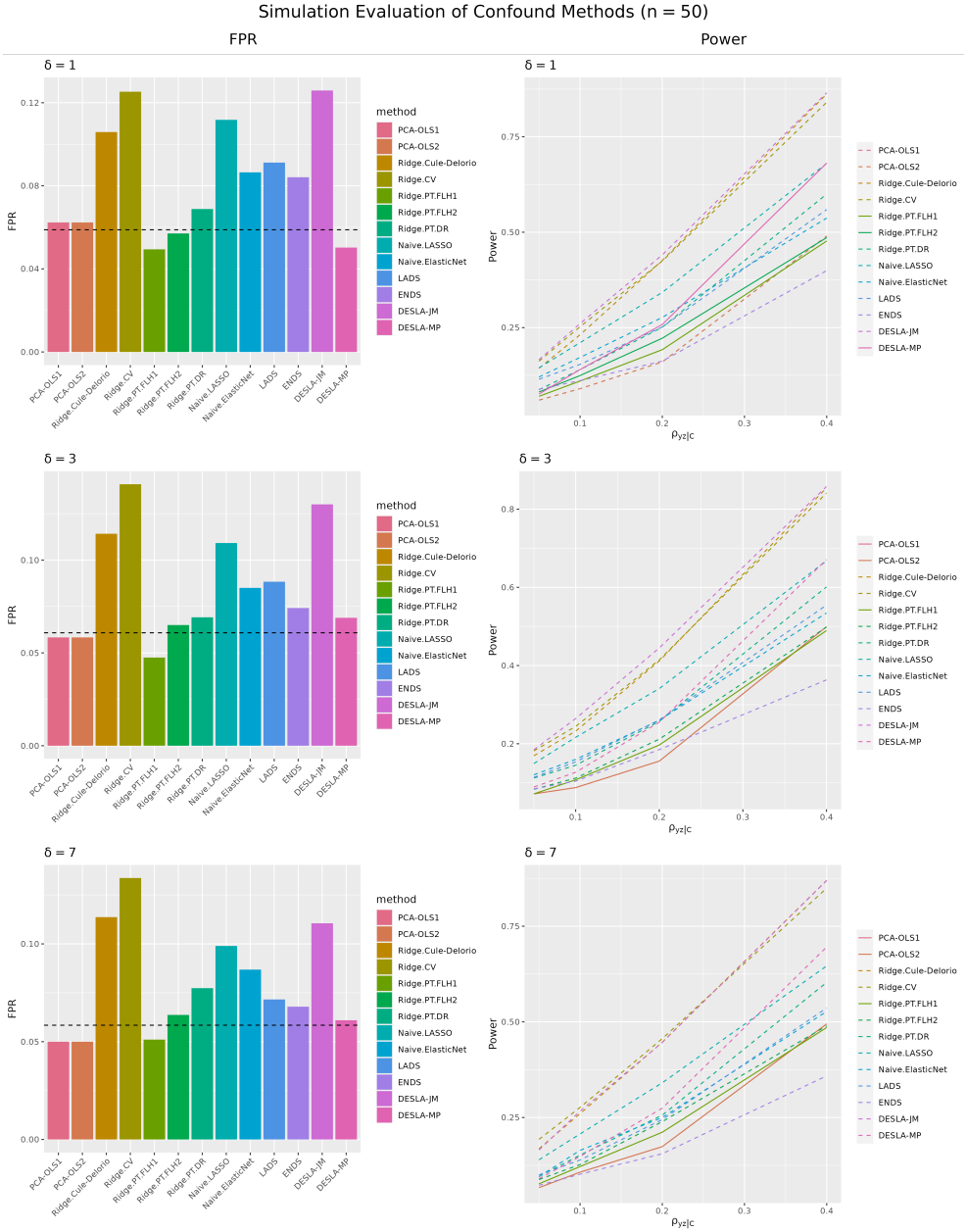


Figure 5.1: Simulation results for $n = 50$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot. Here, we see that methods `PCA-OLS1`, `PCA-OLS2` and `Ridge.PT.FLH1` have reasonably controlled FPRs, with `DESLA-MP` having the highest power.

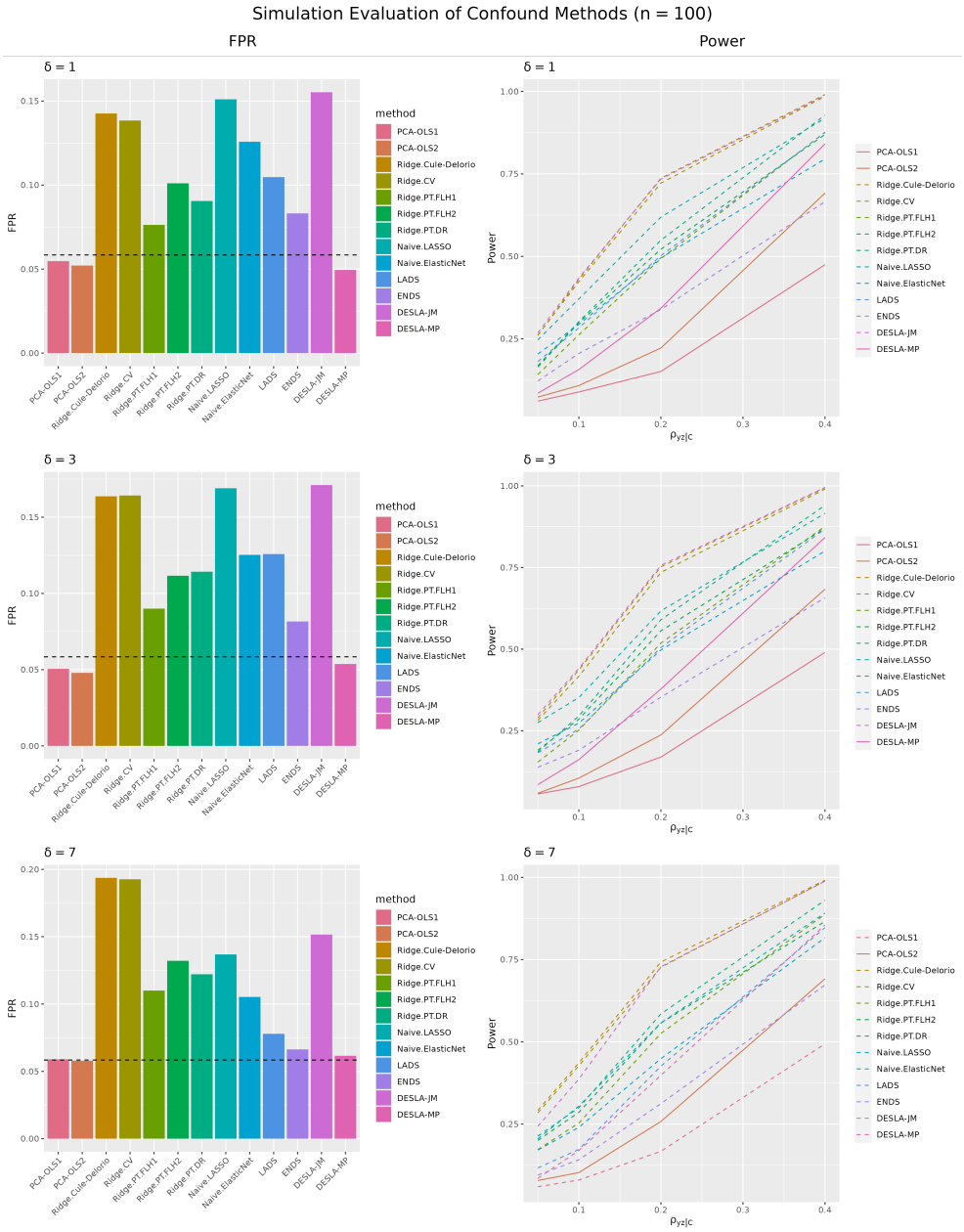


Figure 5.2: Simulation results for $n = 100$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot. Here, we see that methods `PCA-OLS1`, `PCA-OLS2` and `DESLA-MP` have reasonably controlled FPRs, with `DESLA-MP` having the highest power.

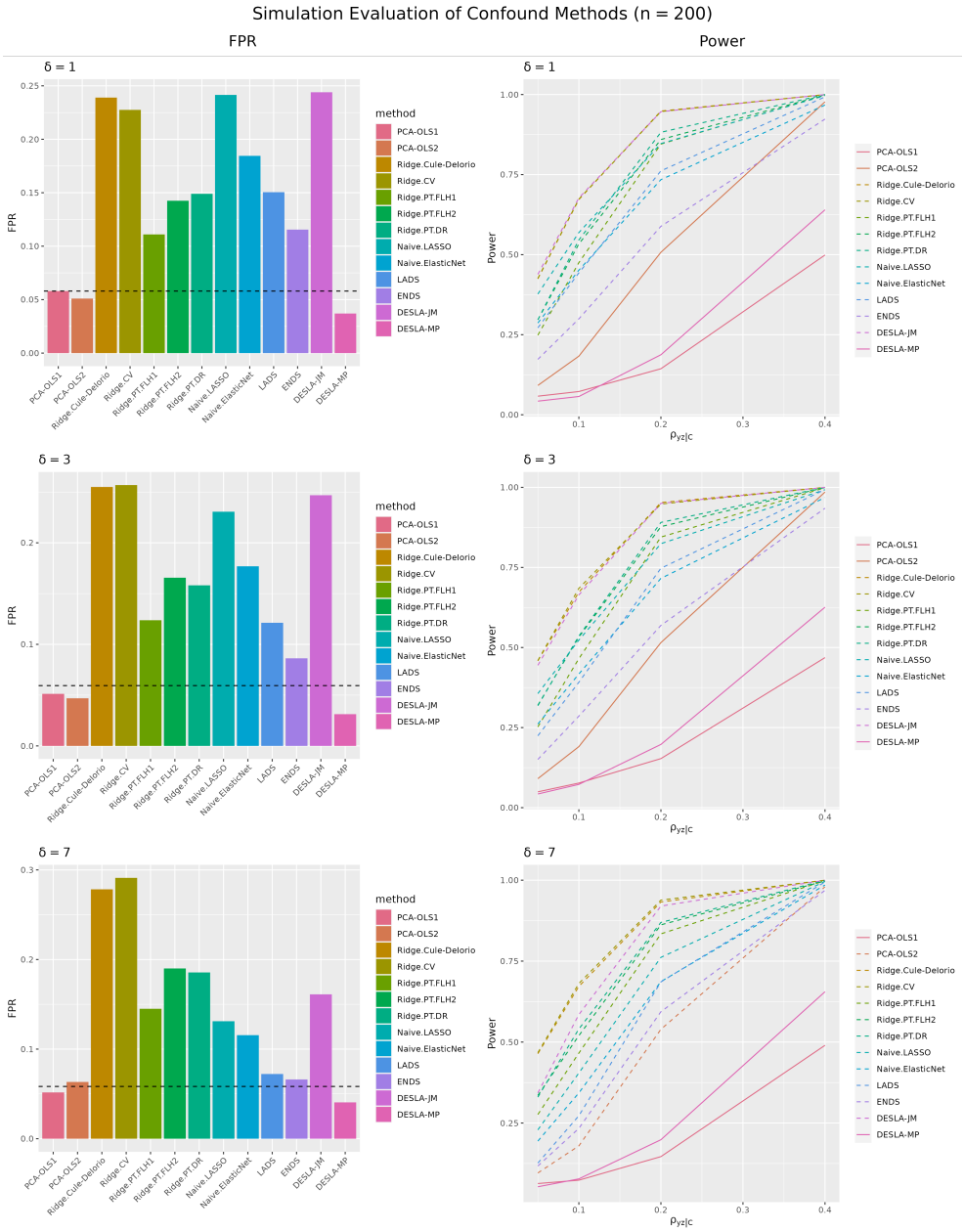


Figure 5.3: Simulation results for $n = 200$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot. Here, we see that methods `PCA-OLS1`, `PCA-OLS2` and `DESLA-MP` have reasonably controlled FPRs, with `PCA-OLS2` having the highest power. `DESLA-MP` is shown to have a conservative FPR.

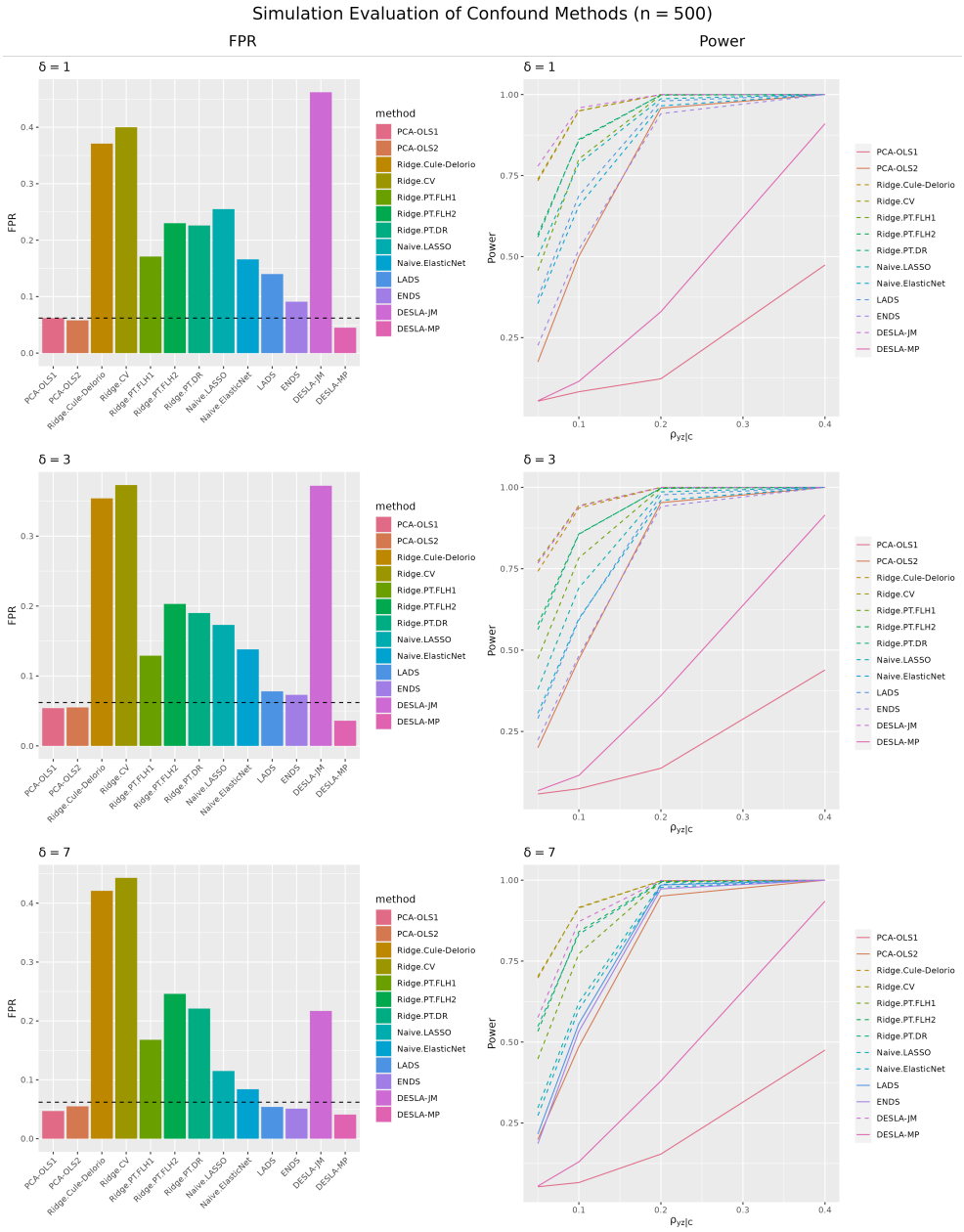


Figure 5.4: Simulation results for $n = 500$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot. Here, we see that methods `PCA-OLS1`, `PCA-OLS2` and `DESLA-MP` have reasonably controlled FPRs, with `PCA-OLS2` having the highest power. `DESLA-MP` is shown to have a conservative FPR.

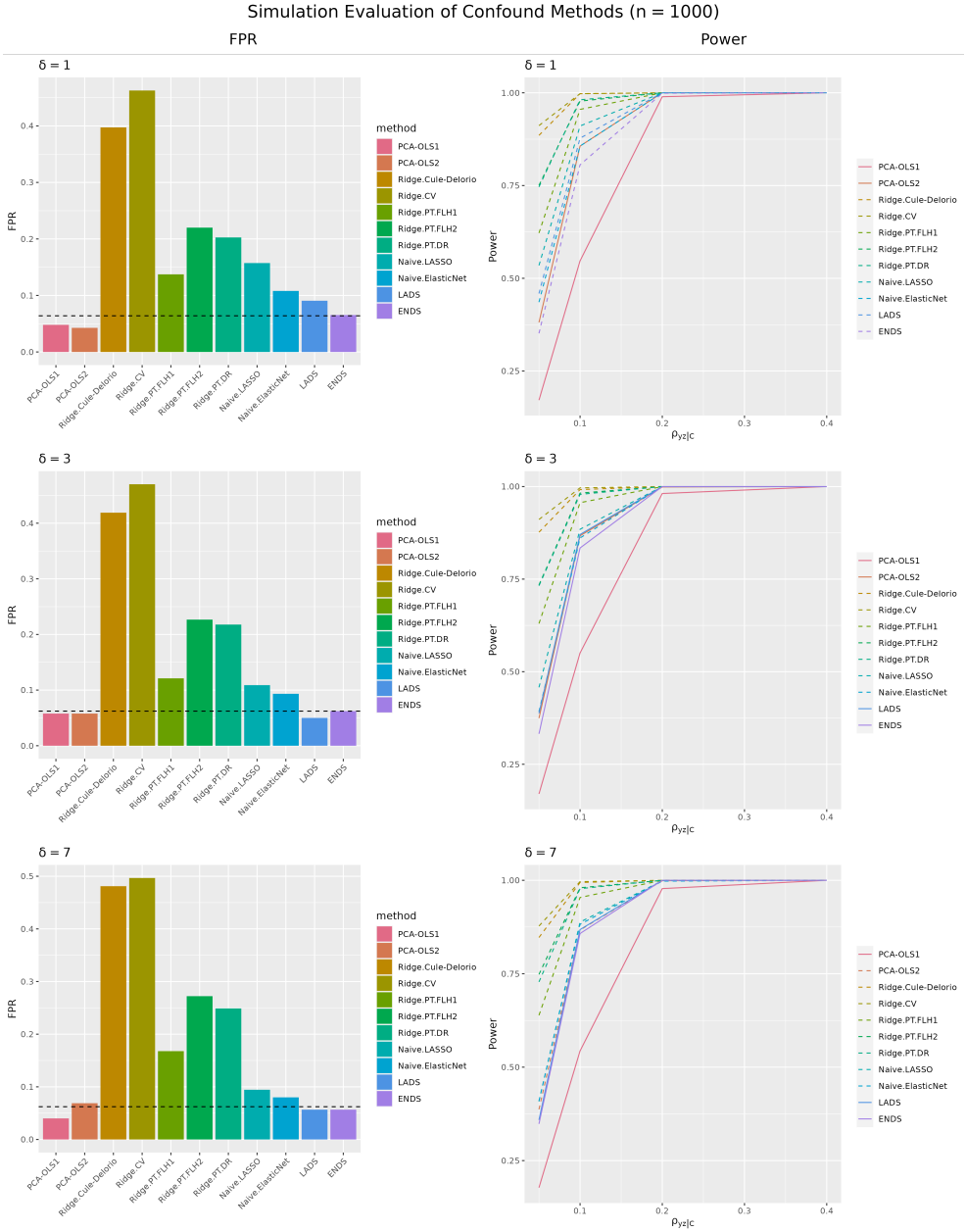


Figure 5.5: Simulation results for $n = 1000$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot. Here, we see that only `PCA-OLS1` has a reasonably controlled FPRs, with `PCA-OLS2` and `ENDS` having borderline controlled FPRs and higher power than `PCA-OLS1`. `LADS` has a controlled FPR for $\delta > 1$.

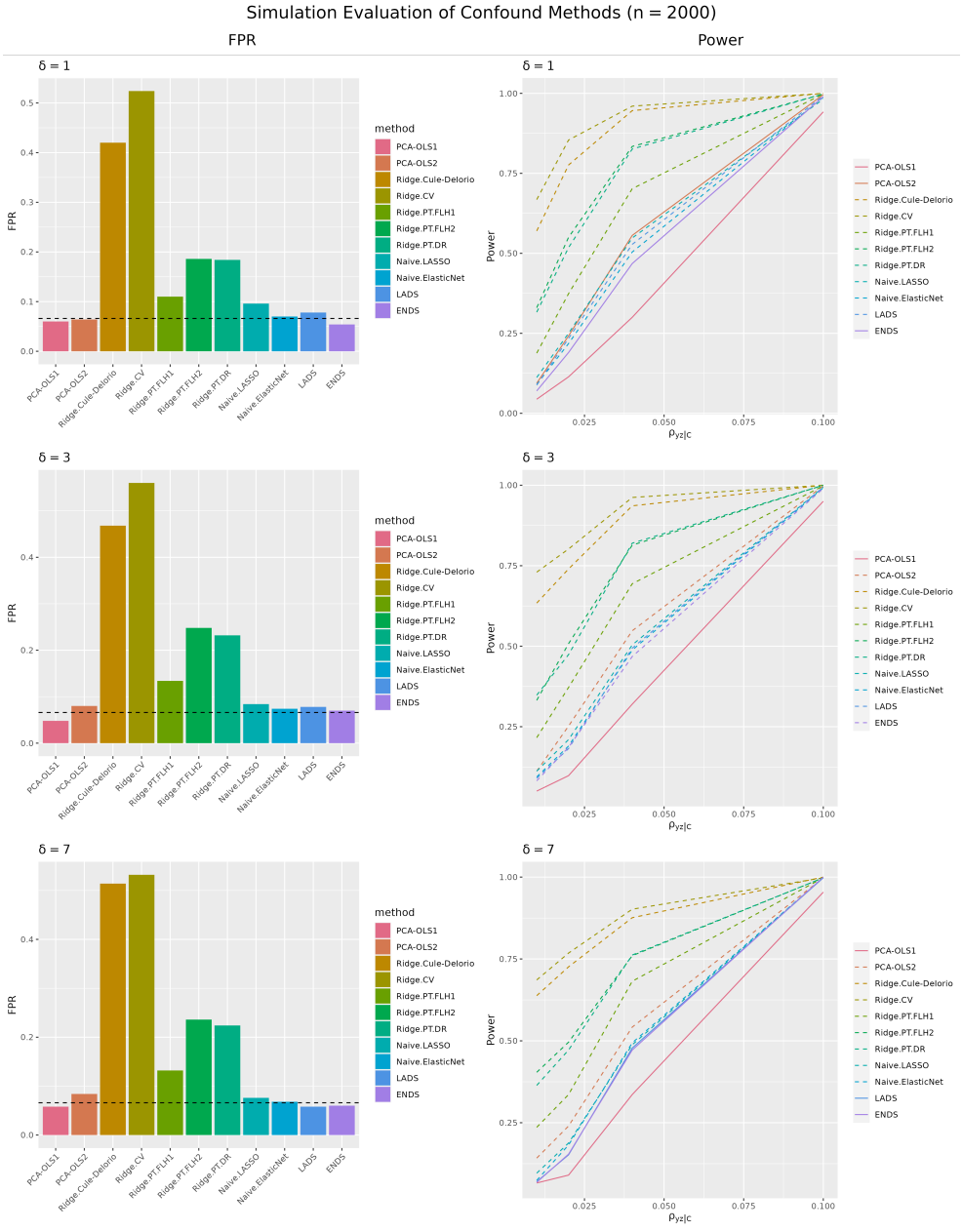


Figure 5.6: Simulation results for $n = 2000$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot. Here, we see that only `PCA-OLS1` and `ENDS` have controlled FPRs, but with quite similar statistical power.

The results of the simulation study are shown in Figures 5.1 through 5.6, each showing the results for a given number of subjects n, with varying sparsity $\delta = 1, 3, 7$. For $n = 50$ subjects, we see that `PCA-OLS1`, `PCA-OLS2` and `Ridge.PT.FLH1` seem to have reasonably controlled FPRs, with `Ridge.PT.FLH1` having a somewhat conservative FPR, and also the highest power. For $n = 100$ subjects, we see that `PCA-OLS1`, `PCA-OLS2` and `DESLA-MP` are the only methods that have controlled FPR. Of these three, `DESLA-MP` has the highest power. For $n = 200$, the same three methods are the only ones with controlled FPR, but `PCA-OLS2` now has the highest power, followed by `DESLA-MP`. Also, we see that `DESLA-MP` has a conservative FPR compared to `PCA-OLS1` and `PCA-OLS2`. Additionally, we see that the methods based on double selection, `LADS` and `ENDS`, are close to having a controlled FPR for the most sparse setting $\delta = 7$. The results follow a very similar pattern for $n = 500$, where the same three methods have a controlled FPR and `PCA-OLS2` has by far the highest power, followed by `DESLA-MP`. Here, we also clearly see a conservative FPR for `DESLA-MP` compared to `PCA-OLS1` and `PCA-OLS2`. Also for the highest sparsity, we now have a fully controlled FPR for the two methods based on double selection. For $n = 1000$, `OLS-PCA1` is the only method with a fully controlled FPR in all scenarios (for $n = 1000$, it is equivalent to regular OLS), with `OLS-PCA2` and `ENDS` showing borderline controlled FPRs and high power. `LADS` also seems to have a controlled FPR, except for in the most dense scenario, $\delta = 1$. For $n = 2000$, we now see that only `OLS-PCA1` and `ENDS` have controlled FPRs, but with fairly similar power.

From these results, it is clear that there are suitable methods for high-dimensional inference in UK Biobank Neuroimaging research at every level of sample size. Our real data task of testing the association between IDPs and AD, has sample size $n = 192$, which is closest to $n = 200$. In this setting, we see that `PCA-OLS2` and `DESLA-MP` both have higher power than `PCA-OLS1`, but that `PCA-OLS2` has a slightly elevated empirical FPR for $\delta = 7$, while `DESLA-MP` has a conservative FPR. We therefore choose to use `DESLA-MP` for the exemplar study, comparing the performance of this method with the baseline approach, `PCA-OLS1`.

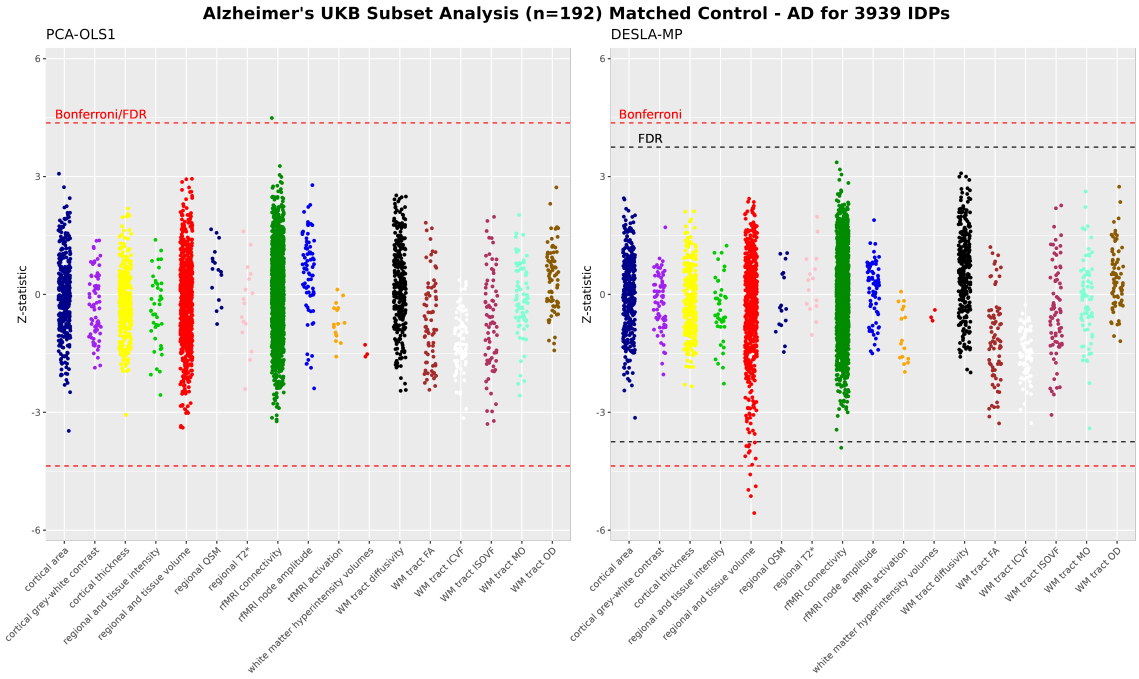


Figure 5.7: Z-statistic plots of associations between 3939 IDPs and AD, using `PCA-OLS1` (left) and `DESLA-MP` (right). `DESLA-MP` detects a number of negative associations between regional/tissue volumes and AD that `PCA-OLS1` fails to detect. The 5 IDPs that pass the Bonferroni threshold for `DESLA-MP` measure volumes in the hippocampus, which is severely affected by AD.

5.3.2 Real Data Results

Figure 5.7 shows two Manhattan plots of Z-scores for the association between 3939 IDPs and AD, controlling for 754 imaging confounds, using `PCA-OLS1` and `DESLA-MP`. For `DESLA-MP`, 22 IDP associations pass the Benjamini-Hochberg False Discovery Rate (FDR) multiple testing threshold (Benjamini and Hochberg, 1995), controlling the FDR at $\alpha = 0.05$. 5 IDP associations pass the Bonferroni threshold for controlling the Family Wise Error Rate (FWER) at $\alpha = 0.05$. For `PCA-OLS1`, the FDR and Bonferroni threshold are the same, and only one IDP association passes both thresholds. All 5 IDP associations that pass the Bonferroni threshold using `DESLA-MP` are negative associations between AD and hippocampal volumes. It is well known and established that the hippocampus is severely affected by AD, and that AD is associated with atrophy of the hippocampus (Halliday, 2017). This gives us further confidence in the suitability of `DESLA-MP` as a high-dimensional inference method in population health neuroimaging, as it has replicated a well known result that the baseline approach `PCA-OLS1`, fails to detect.

Figure 5.8 shows a Bland-Altman plot comparing the $-\log_{10}$ transformed p-values for all IDP on AD associations using `PCA-OLS1` and `DESLA-MP`. The plot does not show a clear, systematic tendency for `DESLA-MP` to have lower p-values than `PCA-OLS1`, but shows in the top right corner that there is a small number of associations for which `DESLA-MP` has significantly lower p-values. These are the regional and tissue volume IDPs that we found to have negative association with AD. This further raises our confidence in `DESLA-MP`, as it shows that the previously detected associations between IDPs and AD are unlikely to be the result of a systematic downward bias in p-values for `DESLA-MP`.

Lastly, Figure 5.9 shows thw QQ-plots for the Z-scores of IDP on AD associations using `PCA-OLS1` and `DESLA-MP`. For `PCA-OLS1`, the QQ-plots show that the Z-scores are close to having a Gaussian distribution, which is indeed their true distribution under the null hypothesis of no association. For `DESLA-MP`, the highly negative Z-scores are more negative than expected under the null hypothesis. These Z-scores likely correspond to the IDPs that were found to have a negative association with AD. On the other hand, the highly positive Z-scores are less positive than expected

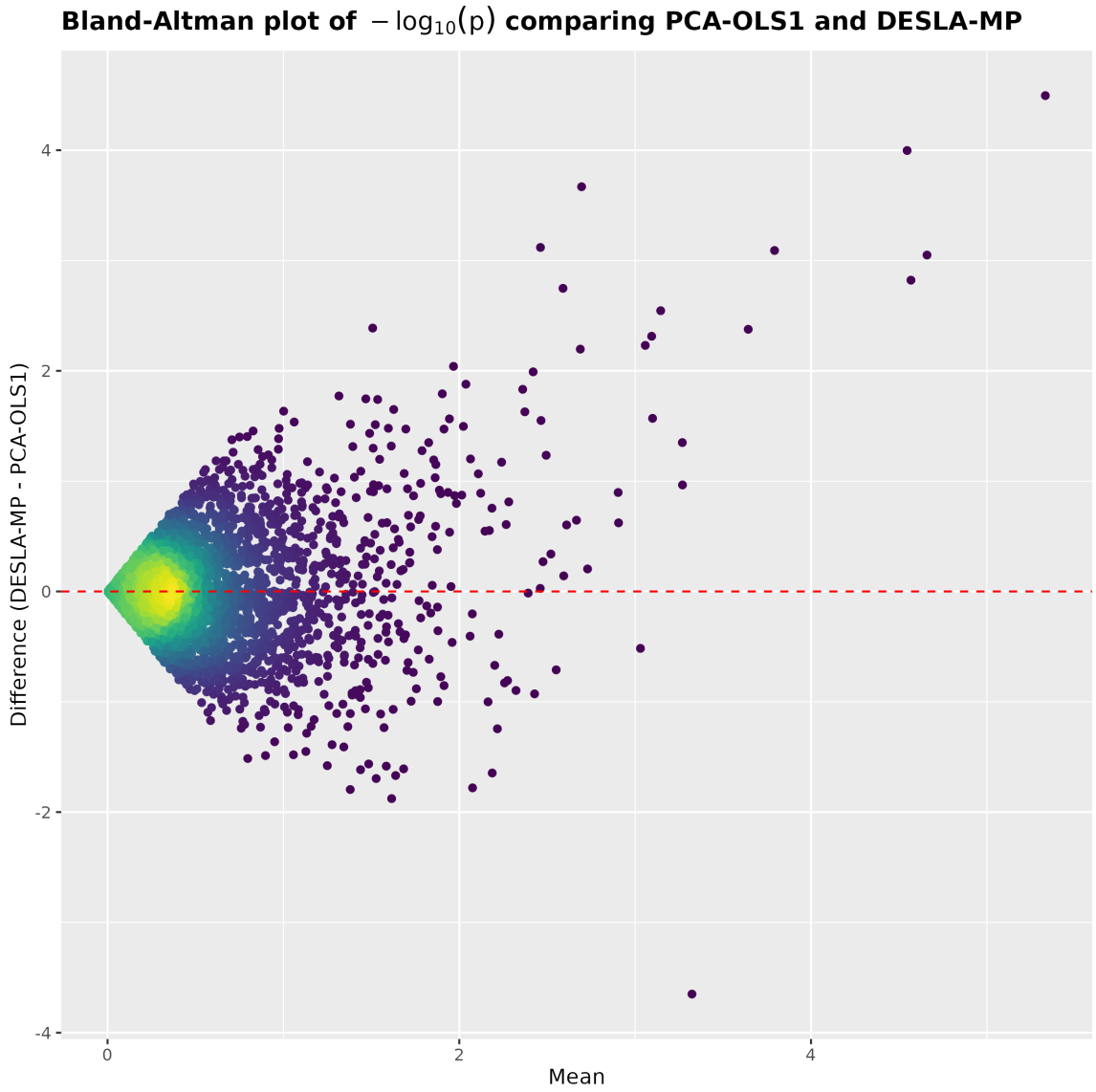


Figure 5.8: Bland-Altman plot comparing the $-\log_{10}$ transformed p-values of all IDP on AD associations using `PCA-OLS1` and `DESLA-MP`. We see no systematic bias in the p-values when comparing the two methods, but we see that there are a number of low p-values for `DESLA-MP` that are much higher for `PCA-OLS1`. These are the IDPs measuring hippocampal volumes, also seen in Figure 5.7.

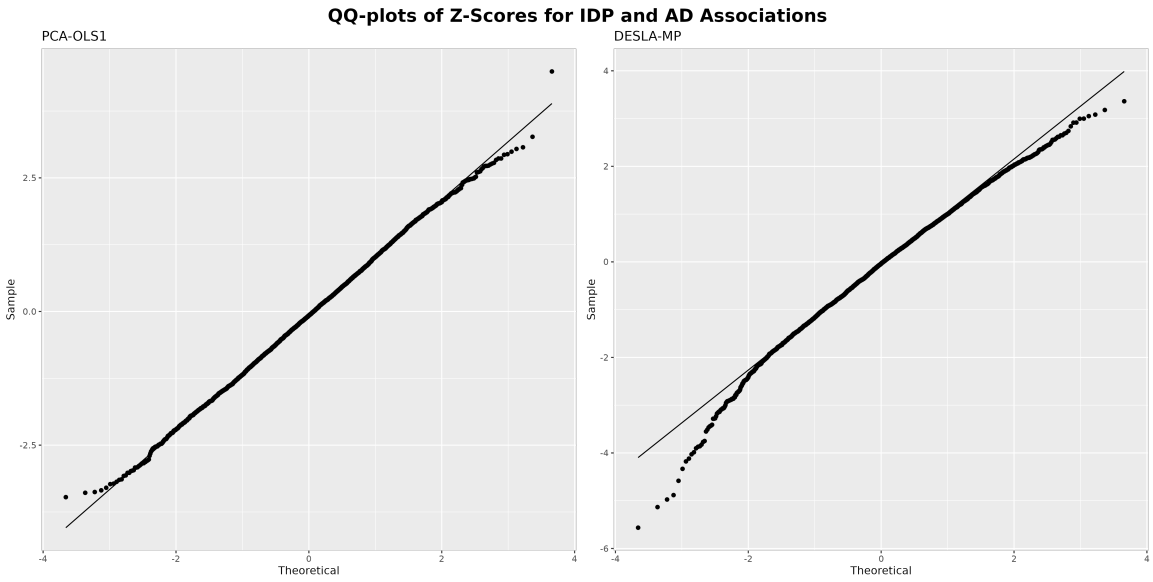


Figure 5.9: QQ-plots of the Z-scores corresponding to IDP on AD associations, for `PCA-OLS1` and `DESLA-MP`. We see that the Z-scores for `PCA-OLS1` look fairly Gaussian, which is their distribution under the null-hypothesis of no association. Conversely, for `DESLA-MP`, we see that the highly negative Z-scores are more negative than expected under the null-hypothesis, which corresponds to the hippocampal IDPs with strong negative associations with AD, and the highly positive Z-scores are less positive than expected under the null-hypothesis, which corresponds to the conservative FPR for `DESLA-MP` seen in the simulation study.

under the null hypothesis, which is in agreement with our simulation results which show that DESLA-MP is conservative under the null hypothesis.

5.4 Discussion

We have evaluated 13 high-dimensional inference methods extensively in a simulation study, varying the sample size, strength of confounding and approximate sparsity of confounding effects. Unlike most previous studies on this topic, our simulation framework has assumed approximate sparsity, rather than strict sparsity, to give a more accurate picture of the performance on real data, which is not guaranteed to fall under exact sparsity. Overall, we have found that there is a number of methods that are reliable for the purposes of high-dimensional inference in UK Biobank population Neuroimaging, even under conditions that are the most prone to induce false positives. Table 5.1 shows our recommendation for the method to use under different sample size scenarios, assuming that we are taking a conservative approach of choosing the highest powered option among methods that we are highly confident have a controlled FPR. For example, we recommend DESLA-MP over PCA-OLS2 for $n = 500$, even though both methods seem to have reasonably controlled FPRs at this sample size, and PCA-OLS2 has higher power. This is because PCA-OLS2 is shown to have an elevated FPR for higher sample sizes than $n = 500$, while DESLA-MP shows conservative FPRs overall.

The main limitation of our work is that we have only evaluated the FPR and power of our methods using confidence level $\alpha = 0.05$, but the confidence level $\alpha = 0.05$ is a somewhat arbitrary threshold, and much lower thresholds are often required, especially when correcting for multiple testing. A simulation study evaluating the FPR and power of all methods for some confidence level $\alpha << 0.05$ will necessarily require a substantially larger number of simulations, and will due to computational constraints therefore lead to a simulation framework that is more focussed on a few specific methods and settings.

We have created a user-friendly R package, HDConf, which includes all of the methods included in this work. The package is freely available for download on GitHub: https://github.com/lavrad99/HDConf

Sample Size	Recommended Method
$n = 50$	`Ridge.PT.FLH1`
$n = 100$	`DESLA-MP`
$n = 200$	`DESLA-MP`
$n = 500$	`DESLA-MP`
$n = 1000$	`PCA-OLS1`
$n = 2000$	`PCA-OLS1`

Table 5.1: The recommended methods for high-dimensional inference in the UK Biobank Brain Imaging Cohort at different sample sizes. At the lowest sample size, $n = 50$, we recommend a ridge regression permutation test based on the Freedman-Lane procedure. For sample sizes ranging between $n = 100$ and $n = 500$, we recommend a de-sparsified LASSO approach using the Moore-Penrose inverse for approximating the precision matrix of our data. For $n \geq 1000$, we recommend `PCA-OLS1`, which is equivalent to vanilla OLS at this sample size. The recommendations are based on the results of our simulation study.

Chapter 6

Conclusions

In this thesis, we have focused on two problems faced by neuroimaging researchers that use the UKB brain imaging cohort: missingness and confounding. In Chapter 2, we explored the structure of missing data in the nIDPs, and proposed a generative model which plausibly describes the mechanism by which structured missing data is created in the nIDPs, and the correlation structure present in the nIDPs. Our simulation study finds only small differences between established missing data methods when imputing missing nIDPs, with a bespoke version of iterative imputation having the best performance. This conclusion from the simulation study was replicated in an exemplary task of identifying 15 nIDPs that explain a high degree of variance in total grey matter volume; very similar levels of variance explained, but with iterative imputation clearly having the better performance. In essence, the results validate previous research results using simpler missing data methods such as mean imputation, while showing that there is indeed some benefit to using iterative imputation. Future research directions on this topic involve incorporating non-Gaussianity and non-linearity in the generative model. Departing from our linear and Gaussian assumption for continuous data is both more representative of the real world, and will likely lead to different simulation results that are more representative of the true performance of missing data methods.

In Chapter 3, we departed from established missing data methodology, and developed a method tailored for data with a mix of SM and UM. Specifically, we focused on CCA linking pairs of UKB sub-studies, a common type of analysis in UKB neuroimag-

ing research. The key to our method is handling SM and UM separately. This design choice is based on results from Chapter 2, which showed that variables with similar missingness patterns tend to be more highly correlated, likely due to the fact that variables from the same sub-study tend to be more highly correlated. This makes it possible to impute UM with very high accuracy compared to SM, which has to rely on more loosely correlated variables for imputation. Our bespoke method showed good performance compared to standard missing data methodology in evaluations using both simulated and real data, and when applied to a real UKB data task, it managed to replicate a well-known medical result via CCA, linking blood levels of Omega-3 with a preference for consuming fatty fish. The most important future direction for this work is extending it to multi-set CCA, where it is an even more useful approach, as the number of subjects with complete study participation rapidly decreases with the number of studies, and our method, notably, does not rely on complete cases.

Moving on to the problem of confounding, Chapter 4 presents our new version of the UKB brain imaging confounds pipeline. The main innovation is the departure from the univariate selection approach for constructing confounds, to a multivariate, fast approach. Multivariate selection leads to a reduced set of confounds, as we avoid the problem of redundant confounds being selected because they are highly correlated with important confounds. Previously, multivariate approaches were too computationally expensive, due to the heterogeneous missing data patterns in the IDPs, requiring a unique design matrix for each of thousands of IDPs. We made it possible by proposing a new estimator for R^2, which is calculated on imputed data, thereby avoiding the problem of missing data altogether. The new confounds pipeline is very computationally fast and memory efficient, and has been made a part of the UKB brain imaging pipeline so that researchers can easily regenerate imaging confounds for new data releases and for idiosyncratic subsets of subjects. The main work that remains to be done on this project is finding and testing a systematic way of selecting threshold parameters for subsets of varying size, as we only recommend a rule of thumb for choosing these parameters. Another, more methodological, area of future inquiry is adjusting the new pipeline to weaker missingness assumptions, as it currently relies on MCAR.

Lastly, Chapter 5 evaluates different methods for deconfounded inference when

the number of confounds exceeds the number of data points. This is a problem in neuroimaging research when dealing with a small subset of subjects, for example a subset of UKB brain imaging subjects with a particularly rare phenotype. Using real UKB imaging confound data, we evaluated 13 methods for deconfounded inference, by simulating outcome and exposure variables, and recording the False Positive Rate (FPR) and statistical power of each method. This study led to a set of recommendations for neuroimaging researchers, detailing which methods to use at given sample sizes. To demonstrate the use of this set of recommendations, we conducted an association study linking Alzheimer's Disease (AD) with IDPs, using only 192 subjects. This exemplar study identified a number of negative associations between AD and regional volumes, specifically in the hippocampus, replicating a well-known result about AD, which is that it is associated with hippocampal atrophy. While the simulation study tests a wide array of sparsity settings, it only evaluates FPR and power for the significance level $\alpha = 0.05$. However, since we often perform multiple testing in epidemiological research, power and FPR control at significance levels much smaller than $\alpha = 0.05$ are relevant, and have not been directly evaluated in our study. This will, however, require many more independent simulation runs to be reliable, so in all likelihood such a study requires either a more focussed set of methods and simulation settings or expanded computational resources.

In summary our work has produced a number of insights, pipelines and methods to aid neuroimaging researchers in handling the problems of confounding and missingness when using UKB brain imaging data. We believe that these results will empower UKB researchers, but likewise believe that most of them are, to varying extents, applicable to other large epidemiological cohorts, such as the China kadoorie biobank (Chen et al., 2011). The study in Chapter 2 is specific to the missingness structure in UKB, so for it to become applicable to a different study it would require a rerun of the analysis pipeline on a new dataset. The method for CCA with SM presented in Chapter 3 is useful to any epidemiological study with multiple substudies that have different subsets of participants. The new method in Chapter 4 solves a forward selection problem under missing data, so it has a broad use case within large population health data sets, beyond constructing UKB imaging confounds. The simulation study in Chapter 5 uses real UKB confounds and simulates exposure and outcome variables.

Because of this, the conclusions from the simulation study are not strictly applicable to other epidemiological studies. However, the simulation framework can easily be rerun with confound data from another study and lead to the same type of insights for that specific study.

Appendix A

A Generative Model For Evaluating Missing Data Methods in Large Epidemiological Cohorts

A.1 Supplementary Proof

Here is a supplementary proof for the proposed variable selection method for iterative imputation. The results assume the joint continuous and binary model described in Demirtas and Doganay (2012), meaning that we assume that the data is drawn from a underlying multivariate normal distribution, where binary data is obtained through thresholding. For the sake of simplicity, we will assume that the latent normal variables that have been thresholded have unit variance and zero mean.

Let j and j' be two variables in our data set $\mathbf{X}$ with missing data and let (x, x') be a generic random draw of the pair j and j' where x is missing and x' is observed. We want to use the variable j' to impute the values in j. It is assumed that we know the Pearson correlation ρ between them, the missingness matrix $\mathbf{M}$ and the proportion of positive cases if either of them are binary. We also assume that the missingness falls under MCAR, so the missingness and the underlying values are completely independent. We want to know how much better than mean/mode imputation it can possibly be to use variable j' to impute variable j, using square- or misclassification loss de-

pending on the type of variable j. We want to preface by observing that the expected reduction in MSE/increase in accuracy will be proportional to $V_{jj'}$, the number of times variable j is missing while j' is observed, multiplied by the expected reduction in loss using the single generic pair (x, x'), because using j' to impute j is only useful in those instances where j is missing and j' is not. It is also clear that

$$\mathbf{V} = (\mathbf{1}_{n\times d} - \mathbf{M})^T\mathbf{M}.$$

x and x' are both continuous

If x and x' are both continuous normal variables with correlation ρ and $\mathbb{E}[x] = \mu$ and $\text{Var}[x] = \sigma^2$, then it is obvious that

$$\mathbb{E}[(x-\mu)^2] - \mathbb{E}[(x - f_{\text{Bayes}}(x'))^2] = \sigma^2 - (1-\rho^2)\sigma^2 = \rho^2\sigma^2,$$

because we know that the Bayes predictor under the square loss for x is the conditional expectation of x given x' and $\text{Var}[x|x'] = (1-\rho^2)\sigma^2$. Therefore, the decrease in MSE for using j' to impute j is proportional to $V_{jj'}\rho^2$.

x is continuous and x' is binary

Let $f_{\text{Bayes}}(x') = \theta_0 + \theta_1 x'$ be the Bayes predictor under the square loss when x' is a binary r.v. This form can be assumed without loss of generality (WLOG), since f_{Bayes} is only defined on the set $\{0, 1\}$. Also, let p' be the proportion of positive cases for j'. We know that

$$\mathbb{E}[xx'] = \rho\sqrt{\text{Var}[x]\text{Var}[x']} + \mathbb{E}[x]\mathbb{E}[x'] = \rho\sigma\sqrt{p'(1-p')} + p'\mu.$$

We also know that f_{Bayes} minimises the expression

$$\mathbb{E}[(x - f_{\text{Bayes}}(x'))^2] =$$

$$= \mathbb{E}[x^2] - 2\mathbb{E}[x f_{\text{Bayes}}(x')] + \mathbb{E}[f_{\text{Bayes}}(x')^2] =$$

$$\mu^2 + \sigma^2 - 2\mu\theta_0 - 2(\rho\sigma\sqrt{p'(1-p')} + p'\mu)\theta_1 + (\theta_0 + \theta_1 p')^2 + \theta_1^2 p'(1-p').$$

We find the minimum with respect to $\theta_{0,1}$ by finding the point at which the gradient vector is 0. We have that:

$$\frac{\partial}{\partial\theta_0}\mathbb{E}[(x - f_{\text{Bayes}}(x'))^2] = -2\mu + 2(\theta_0 + \theta_1 p') = 0 \iff \theta_0 + \theta_1 p' = \mu.$$

And also:

$$\frac{\partial}{\partial\theta_1}\mathbb{E}[(x - f_{\text{Bayes}}(x'))^2] = -2(\rho\sigma\sqrt{p'(1-p')} + p'\mu) + 2p'(\theta_0 + \theta_1 p') + 2p'(1-p')\theta_1 = 0$$

$$\iff$$

$$-\rho\sigma\sqrt{\frac{1-p'}{p'}} - \mu + \theta_0 + \theta_1 p' + (1-p')\theta_1 = 0$$

$$\iff$$

$$\theta_0 + \theta_1 = \mu + \rho\sigma\sqrt{\frac{1-p'}{p'}}.$$

This gives us the following system of equations:

$$\begin{cases} \theta_0 + p'\theta_1 = \mu \\ \theta_0 + \theta_1 = \mu + \rho\sigma\sqrt{\frac{1-p'}{p'}}, \end{cases}$$

which evaluates to

$$\begin{cases} \theta_0 = \mu - \rho\sigma\sqrt{\frac{p'}{1-p'}} \\ \theta_1 = \frac{\rho\sigma}{\sqrt{p'(1-p')}}. \end{cases}$$

Using these values for $f_{\text{Bayes}}(x') = \theta_0 + \theta_1 x'$, we get that

$$\mathbb{E}[(x - f_{\text{Bayes}}(x'))^2] =$$

$$= \mu^2 + \sigma^2 - 2\mu\left(\mu - \rho\sigma\sqrt{\frac{p'}{1-p'}}\right) - 2(\rho\sigma\sqrt{p'(1-p')} + p'\mu)\left(\frac{\rho\sigma}{\sqrt{p'(1-p')}}\right) + (\mu)^2$$

$$+ \left(\frac{\rho\sigma}{\sqrt{p'(1-p')}}\right)^2 p'(1-p') =$$

$$= \mu^2 + \sigma^2 - 2\mu^2 + 2\mu\rho\sigma\sqrt{\frac{p'}{1-p'}} - 2\rho^2\sigma^2 - 2\mu\rho\sigma\sqrt{\frac{p'}{1-p'}} + \mu^2 + \rho^2\sigma^2 =$$

$$= (1-\rho^2)\sigma^2.$$

This is the same result as in the previous case, which means that the decrease in MSE for using j' to impute j is proportional to $V_{jj'}\rho^2$.

x is binary and x' is continuous

Here we assume that $\tilde{x}$ is the underlying normal variable variable that is binarised using threshold $D \in \mathbb{R}$, i.e., $x = 1\{\tilde{x} > D\}$. $(\tilde{x}, x')$ is multivariate normal distributed. We can further assume WLOG that $\tilde{x}$ and x' have zero mean, unit variance and correlation ρ_b. It is obvious that $f_{\text{Bayes}}(x') = 1\{\mathbb{E}[\tilde{x}|x'] > D\}$ is the Bayes predictor of x using x' under the misclassification loss, since $\tilde{x}|x'$ is normally distributed with expectation $\mathbb{E}[\tilde{x}|x']$ and the expected value and median of normally distributed variables are the same. We also know that $\mathbb{E}[\tilde{x}|x'] = \rho_b x'$, $\text{Var}[\tilde{x}|x'] = 1 - \rho_b^2$. This means that

$$\begin{cases} f_{\text{Bayes}}(x') = 1\{x' > D/\rho_b\} \text{ if } \rho_b > 0 \\ f_{\text{Bayes}}(x') = 1\{x' < D/\rho_b\} \text{ if } \rho_b < 0. \end{cases}$$

The expected decrease in misclassification loss by imputing x using x' is

$$\mathbb{E}[1\{x = f_{\text{Bayes}}(x')\}] - \mathbb{E}[1\{x = \arg\max(1-p, p)\}] = \mathbb{E}[1\{x = f_{\text{Bayes}}(x')\}] - \max\{p, 1-p\}.$$

Finally, we get that the expected decrease in expected loss is proportional to

$$V_{jj'}\left[\int_{-\infty}^{D/\rho_b}\phi(x')\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)dx'+\int_{D/\rho_b}^{\infty}\phi(x')\left(1-\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)\right)dx'-\max\{p,1-p\}\right]$$

if $\rho_b > 0$ and

$$V_{jj'}\left[\int_{-\infty}^{D/\rho_b}\phi(x')\left(1-\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)\right)dx'+\int_{D/\rho_b}^{\infty}\phi(x')\Phi\left(\frac{D-\rho_b x'}{\sqrt{1-\rho_b^2}}\right)dx'-\max\{p,1-p\}\right]$$

if $\rho_b < 0$, where ϕ and Φ are the the probability density function and cumulative distribution function of the standard-normal distribution. It should be noted that if p is the proportion of positive cases for variable j, then $D = \Phi^{-1}(p)$ and that

$$\rho_b = \rho\frac{\sqrt{p(1-p)}}{\phi(D)},$$

where ρ is the correlation between x and x'.

x and x' are both binary

Let p and p' be the proportion of positive cases in j and j' respectively. Let further $p_{11} = \mathbb{P}(x = 1 \cap x' = 1)$. We know in this case that

$$f_{\text{Bayes}}(x') = \underset{k\in\{0,1\}}{\arg\max}\,\mathbb{P}(x = k|x').$$

This means that it is enough for us to know the contingency table of x and x' to know the Bayes predictor and we know the whole contingency table if we know p, p' and p_{11}. The maximal expected decrease in the misclassification loss is proportional to

$$V_{jj'}[p'\max\{\mathbb{P}(x = 1|x' = 1), 1-\mathbb{P}(x = 1|x' = 1)\}$$
$$+(1-p')\max\{\mathbb{P}(x = 1|x' = 0), 1-\mathbb{P}(x = 1|x' = 0)\}-\max\{p,1-p\}].$$

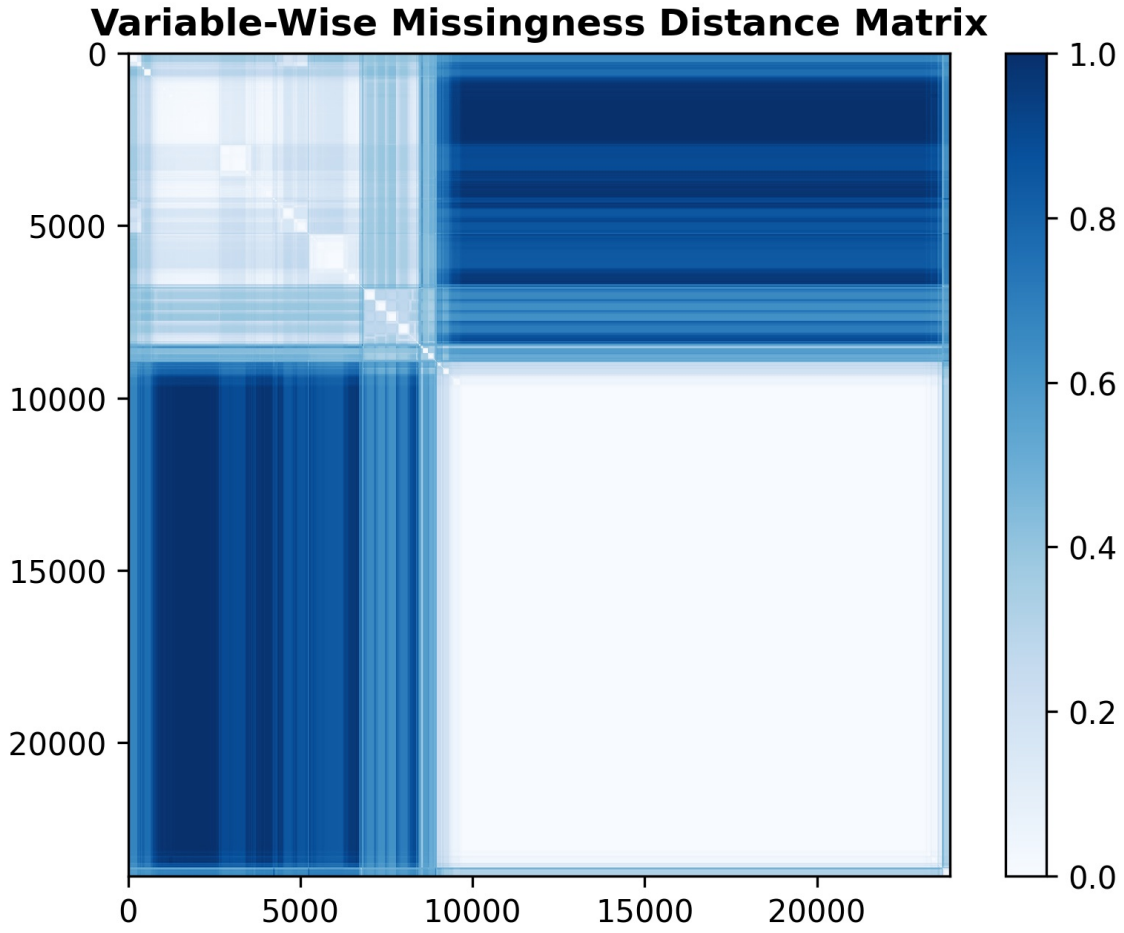


Figure A.1: Full variable-wise missingness distance matrix. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

The quantity p_{11} is calculated in the following way:

$$p_{11} = \mathbb{E}[xx'] = \mathrm{Cov}[x, x'] + \mathbb{E}[x]\mathbb{E}[x'] = \rho\sqrt{\mathrm{Var}[x]\mathrm{Var}[x']} + pp' =$$

$$= \rho\sqrt{p(1-p)p'(1-p')} + pp'.$$

A.2 Supplementary Figures

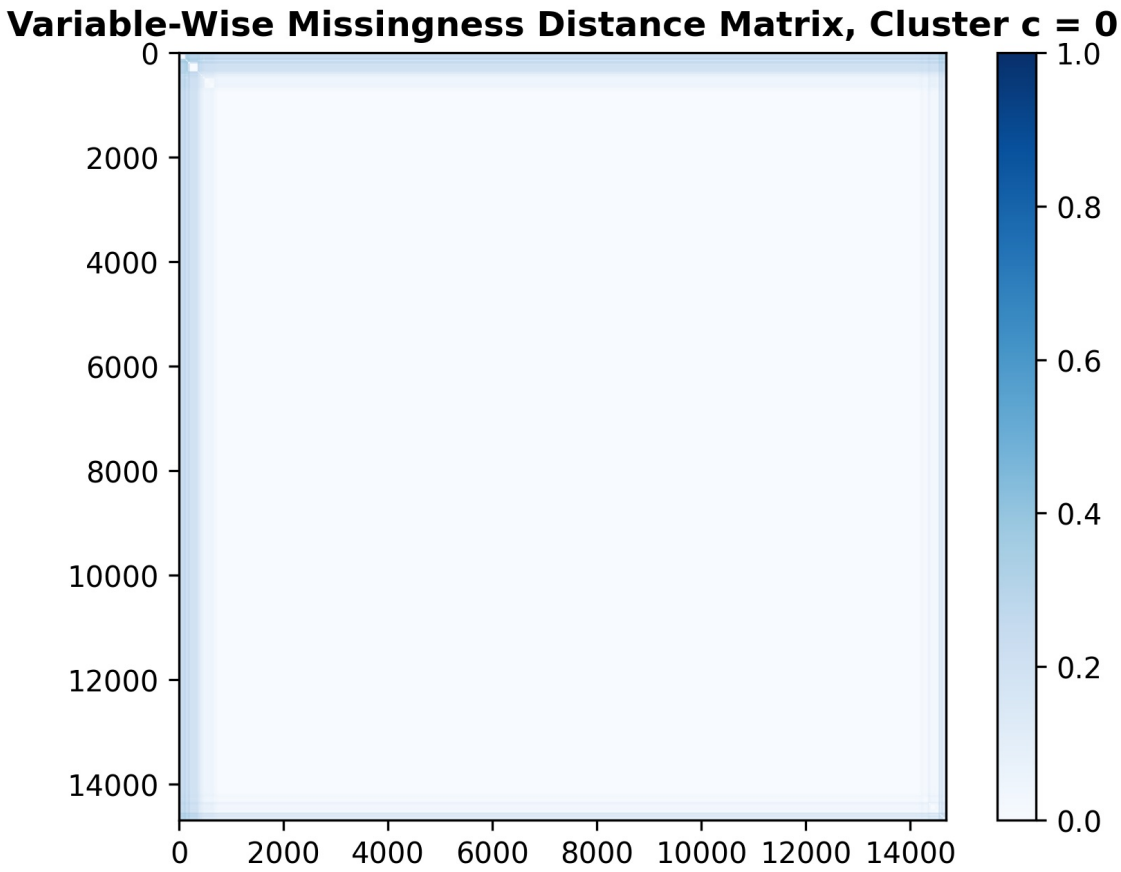


Figure A.2: Variable-wise missingness distance matrix for cluster $c = 0$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.



Figure A.3: Variable-wise missingness distance matrix for cluster $c = 1$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

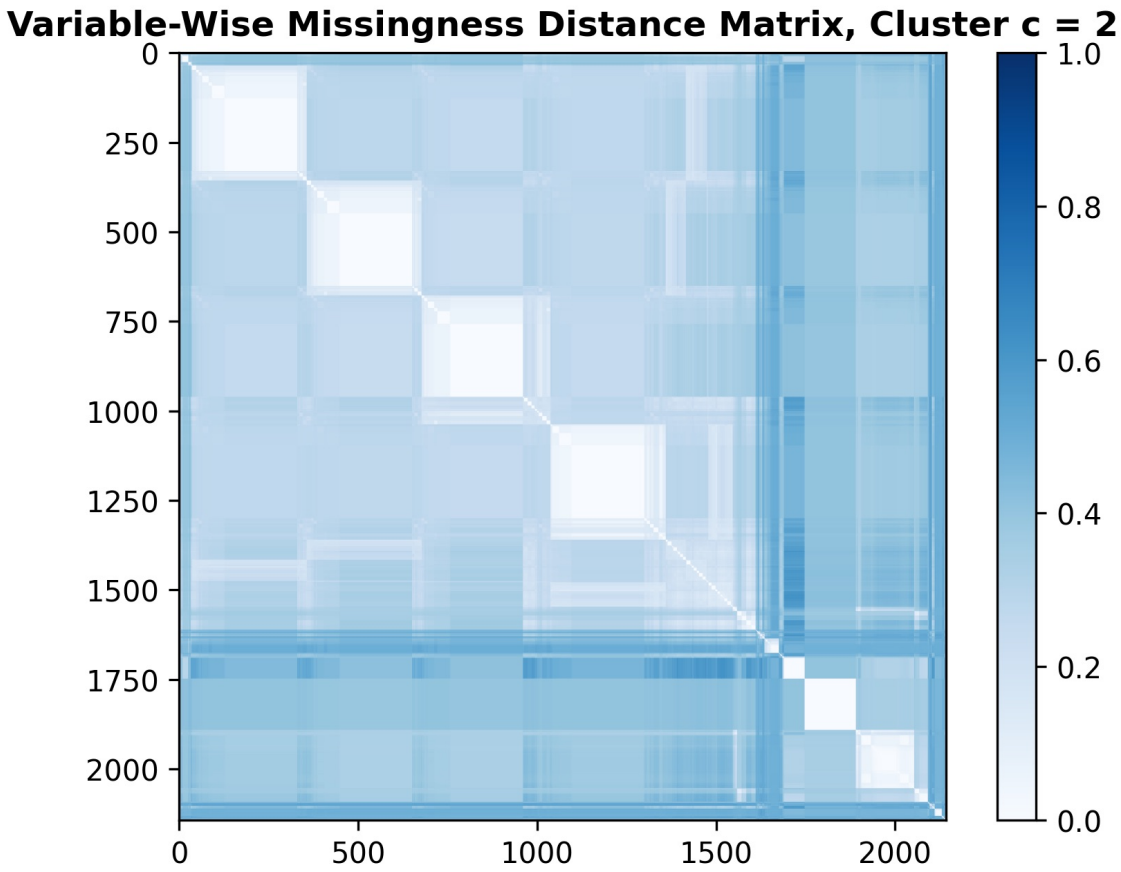


Figure A.4: Variable-wise missingness distance matrix for cluster $c = 2$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

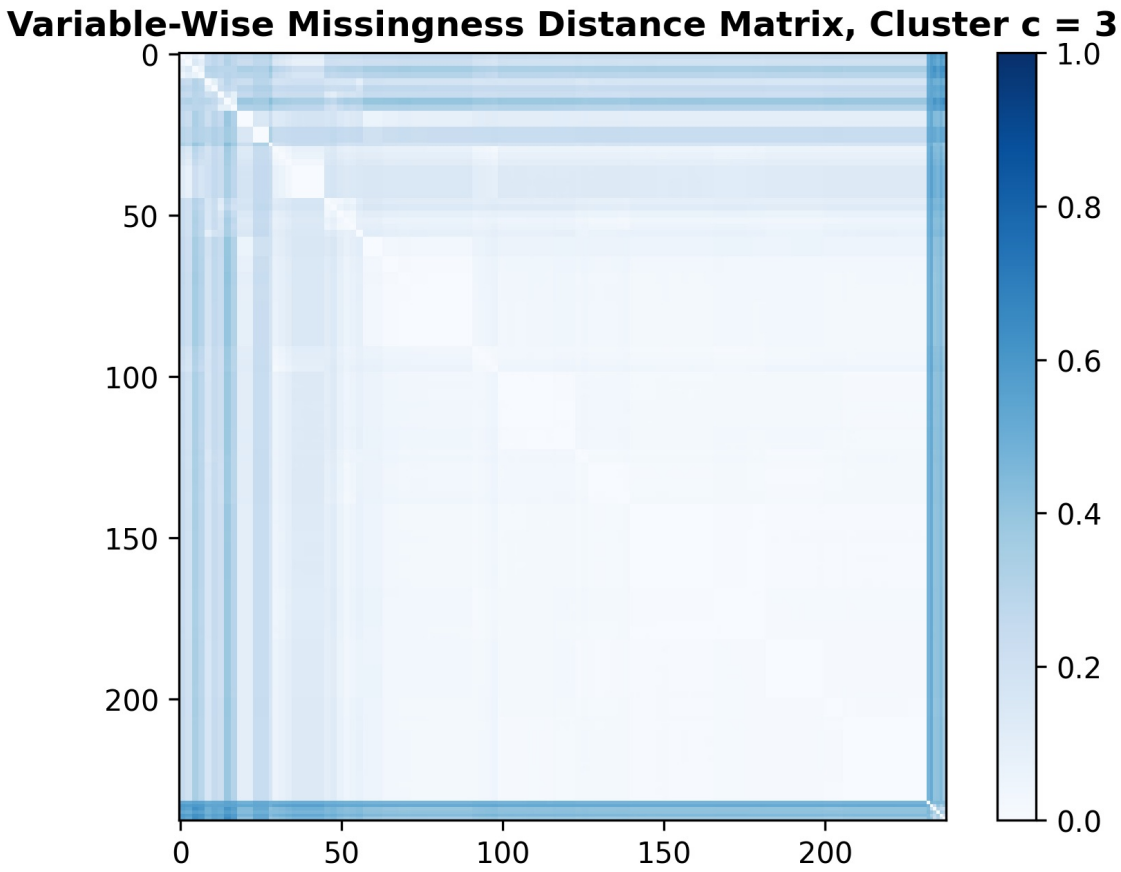


Figure A.5: Variable-wise missingness distance matrix for cluster $c = 3$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

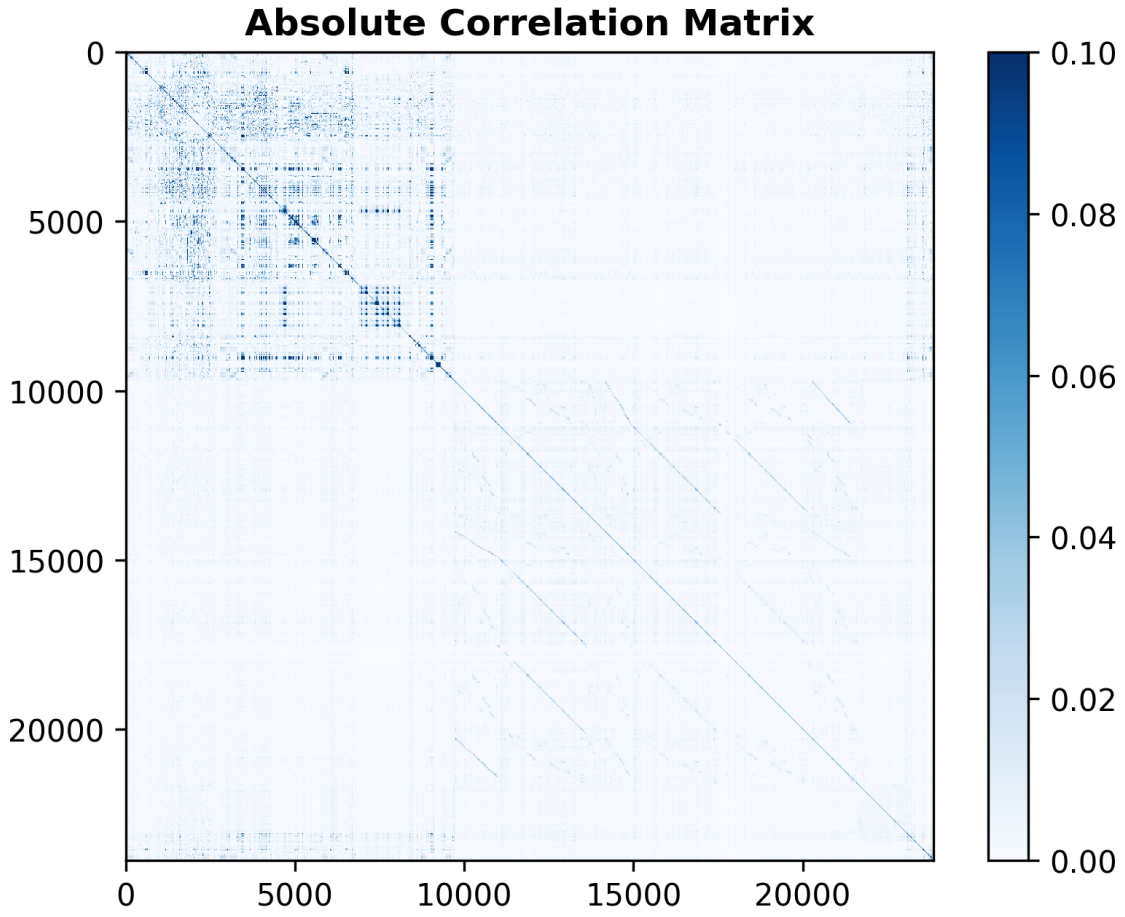


Figure A.6: Full phenotype correlation matrix. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

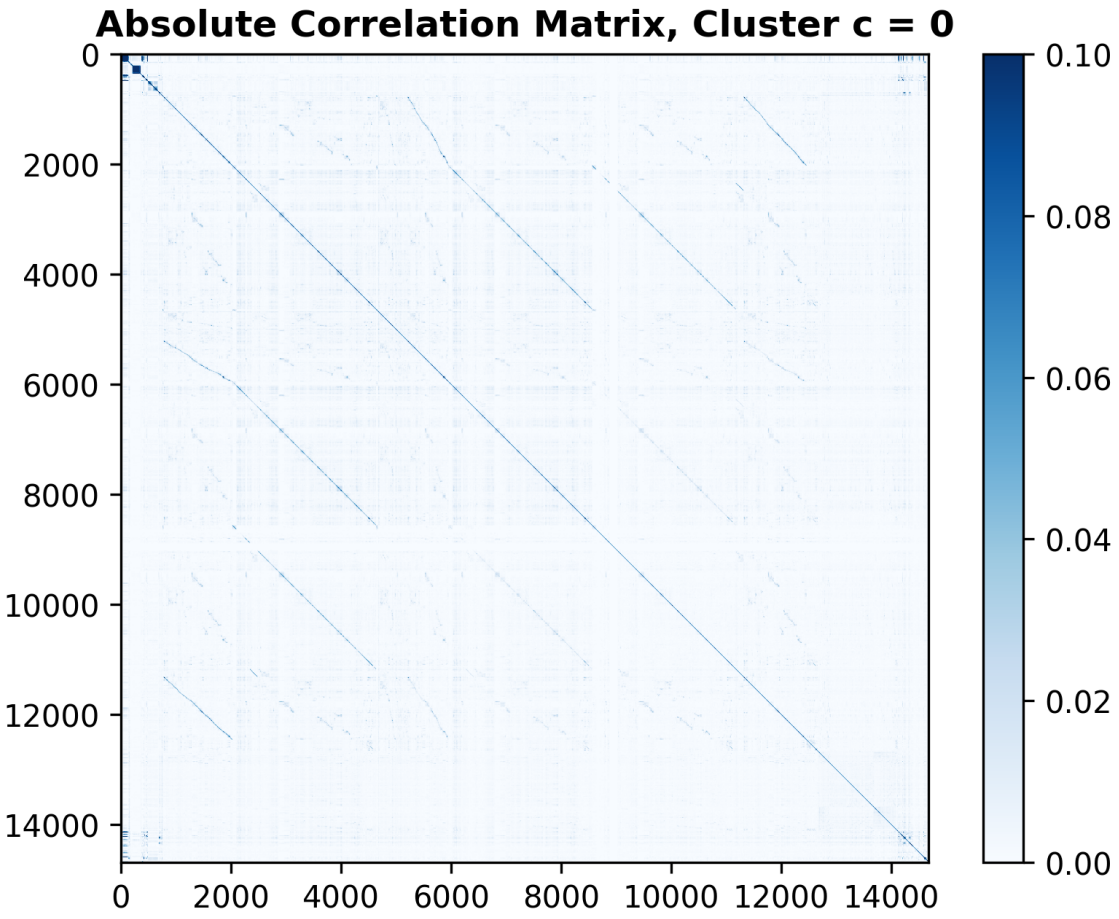


Figure A.7: Phenotype correlation matrix for cluster $c = 0$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

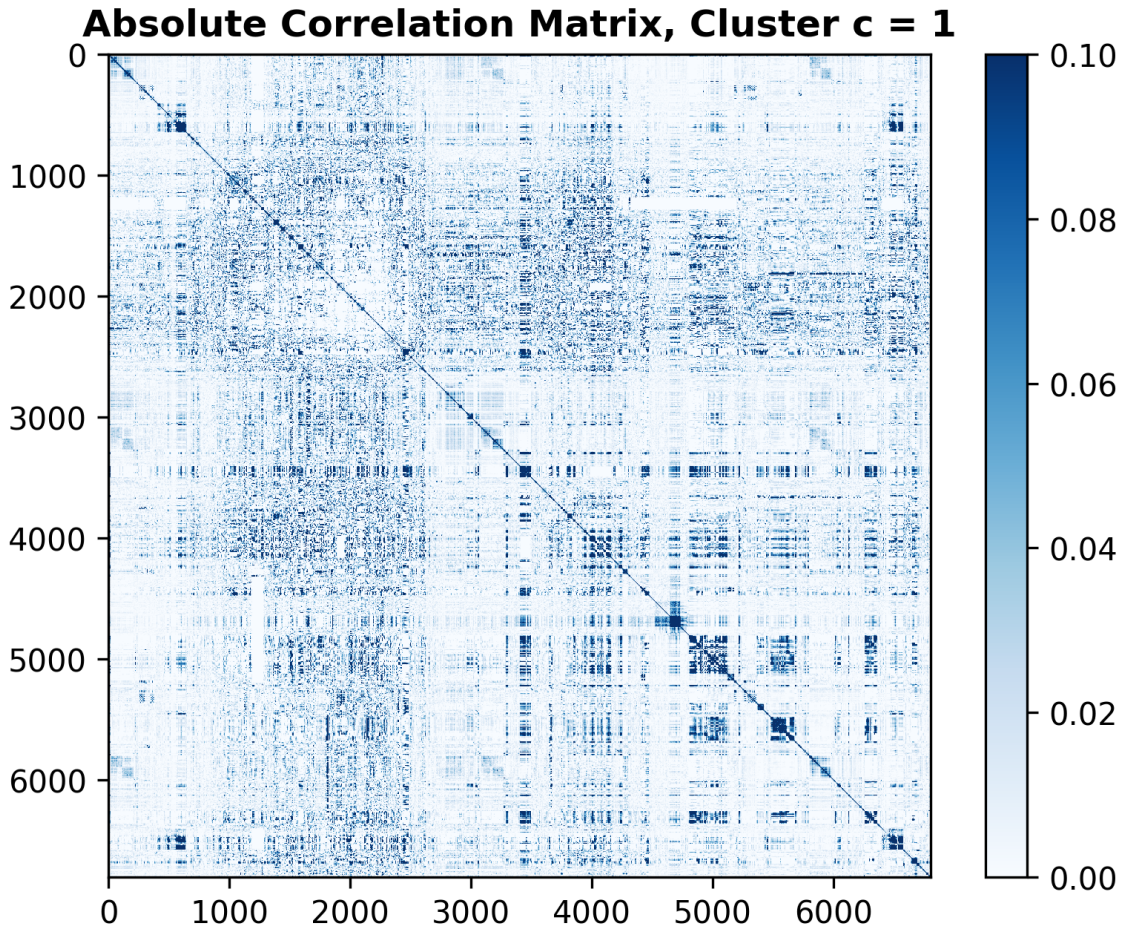


Figure A.8: Phenotype correlation matrix for cluster $c = 1$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

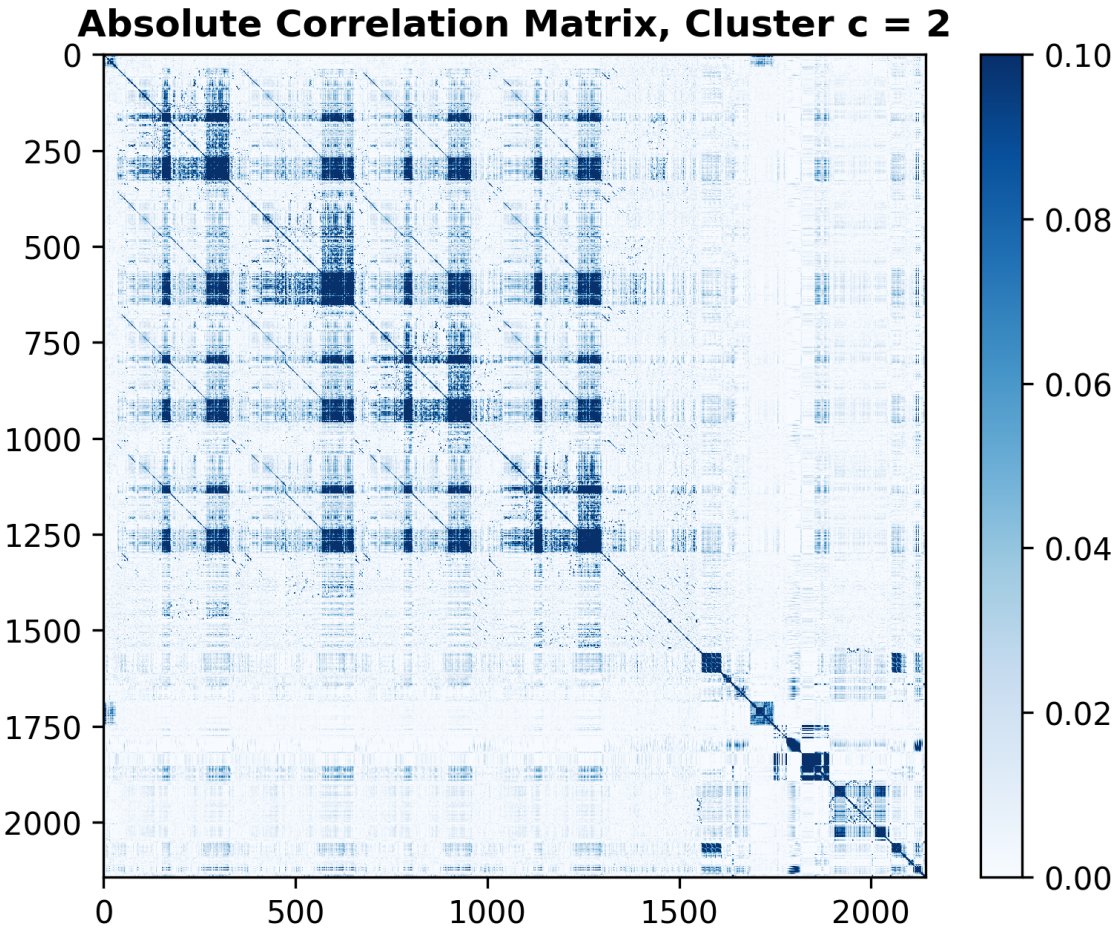


Figure A.9: Phenotype correlation matrix for cluster $c = 2$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

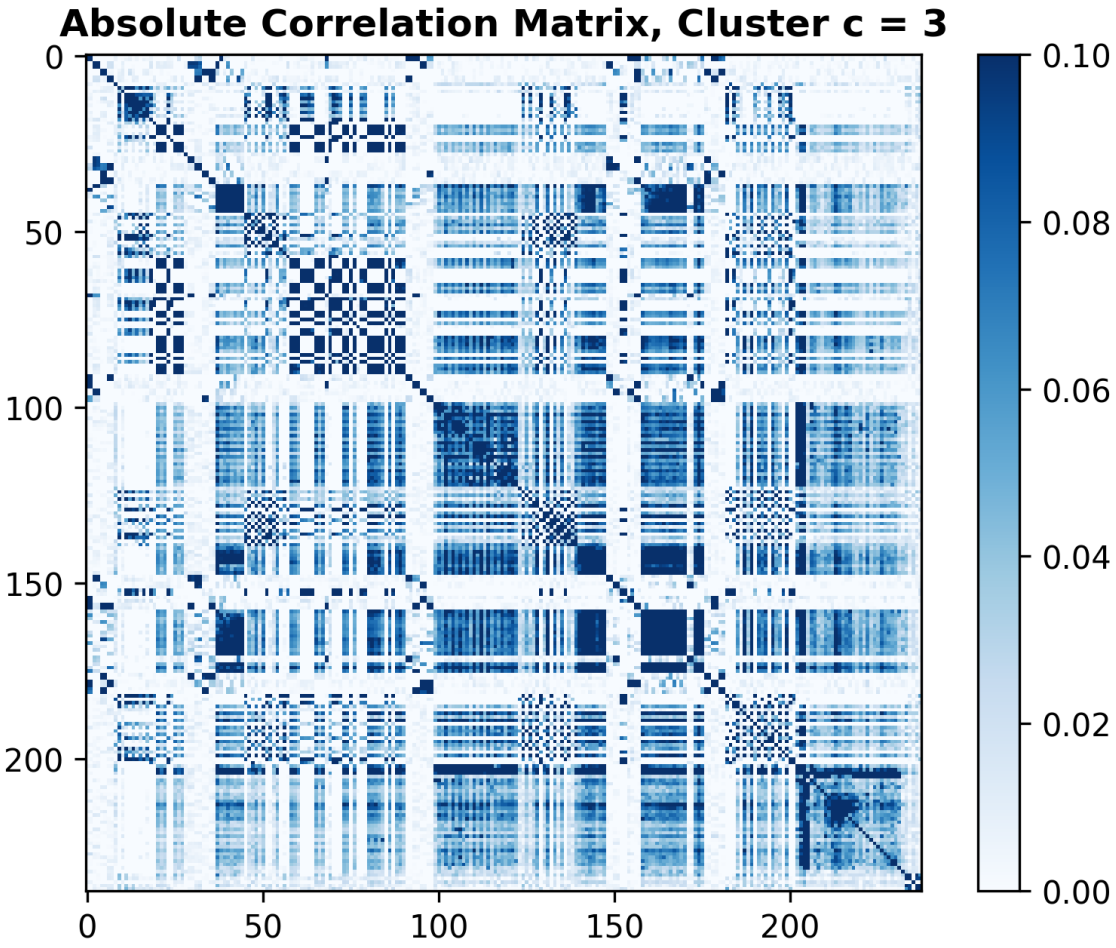


Figure A.10: Phenotype correlation matrix for cluster $c = 3$. Rows and columns have been resorted by order of agglomeration in the hierarchical clustering by variable-wise missingness distance.

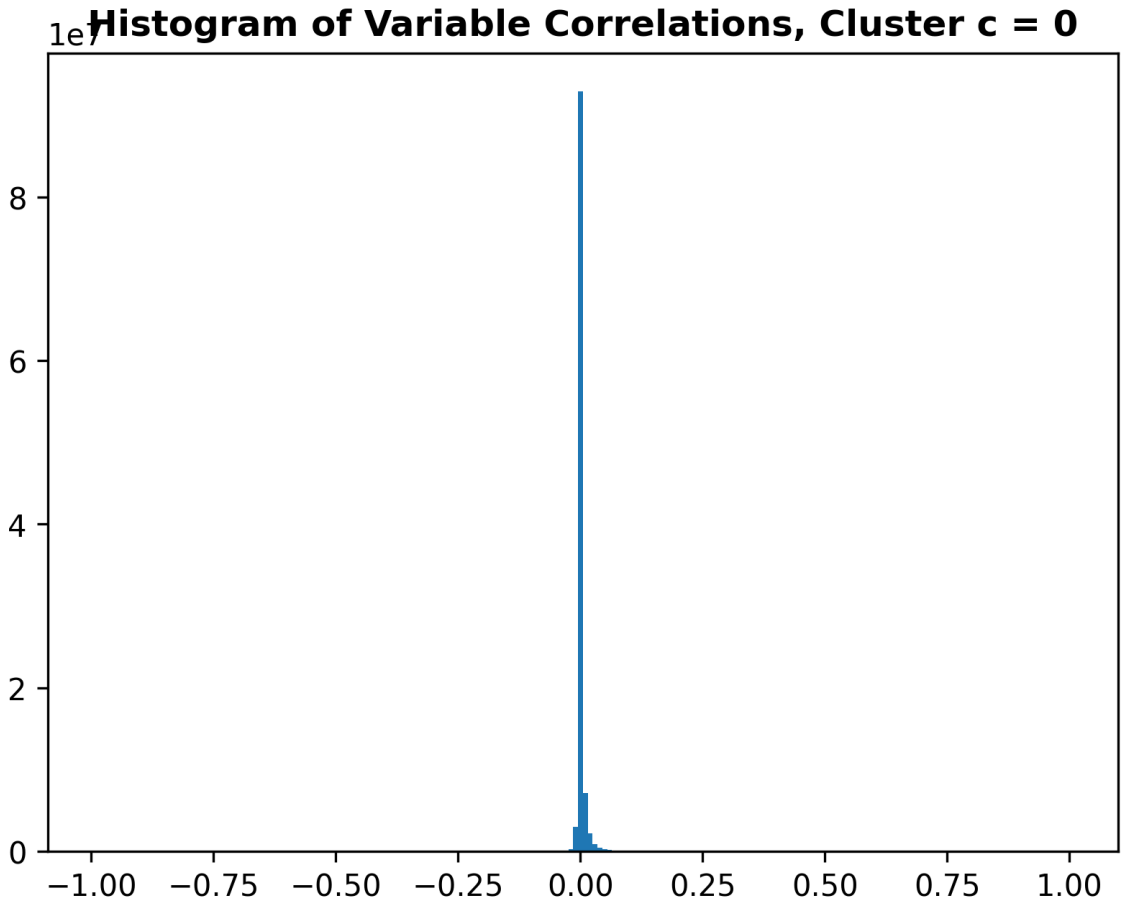


Figure A.11: Histogram of correlations, cluster $c = 0$.

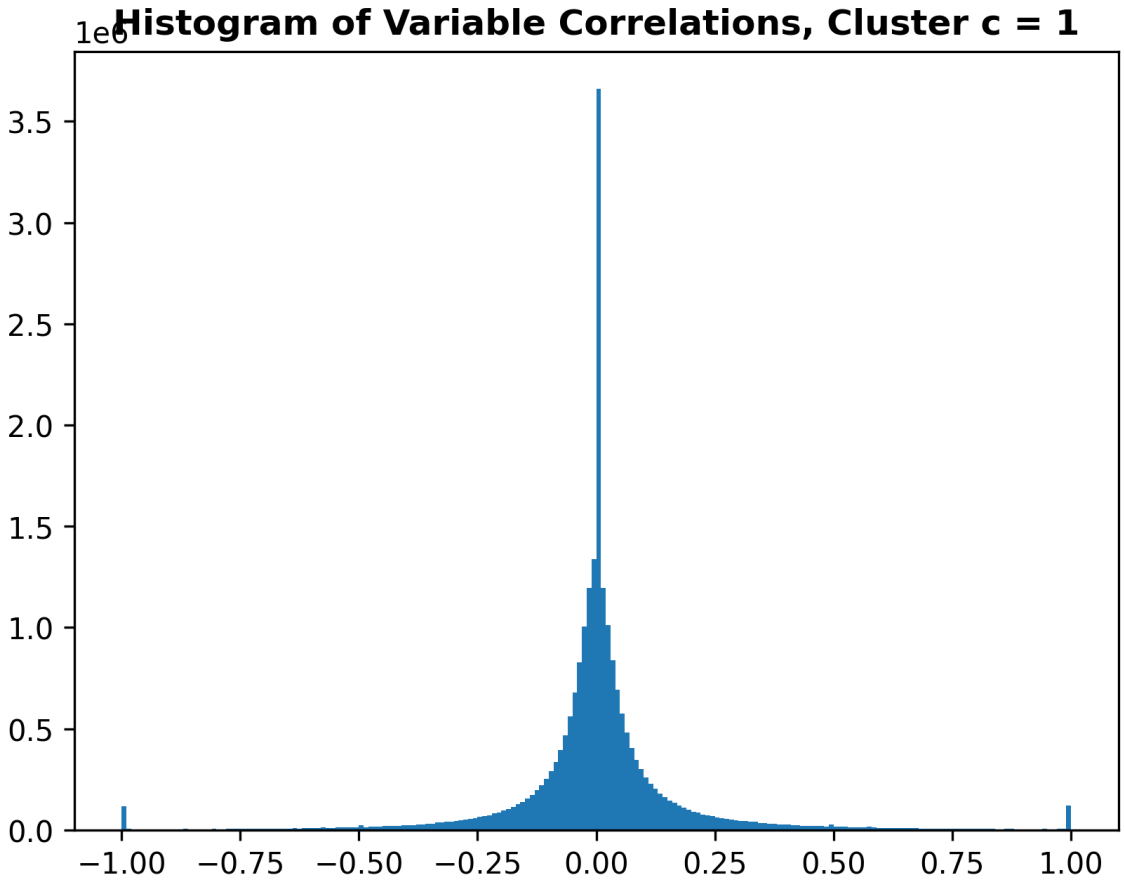


Figure A.12: Histogram of correlations, cluster $c = 1$.

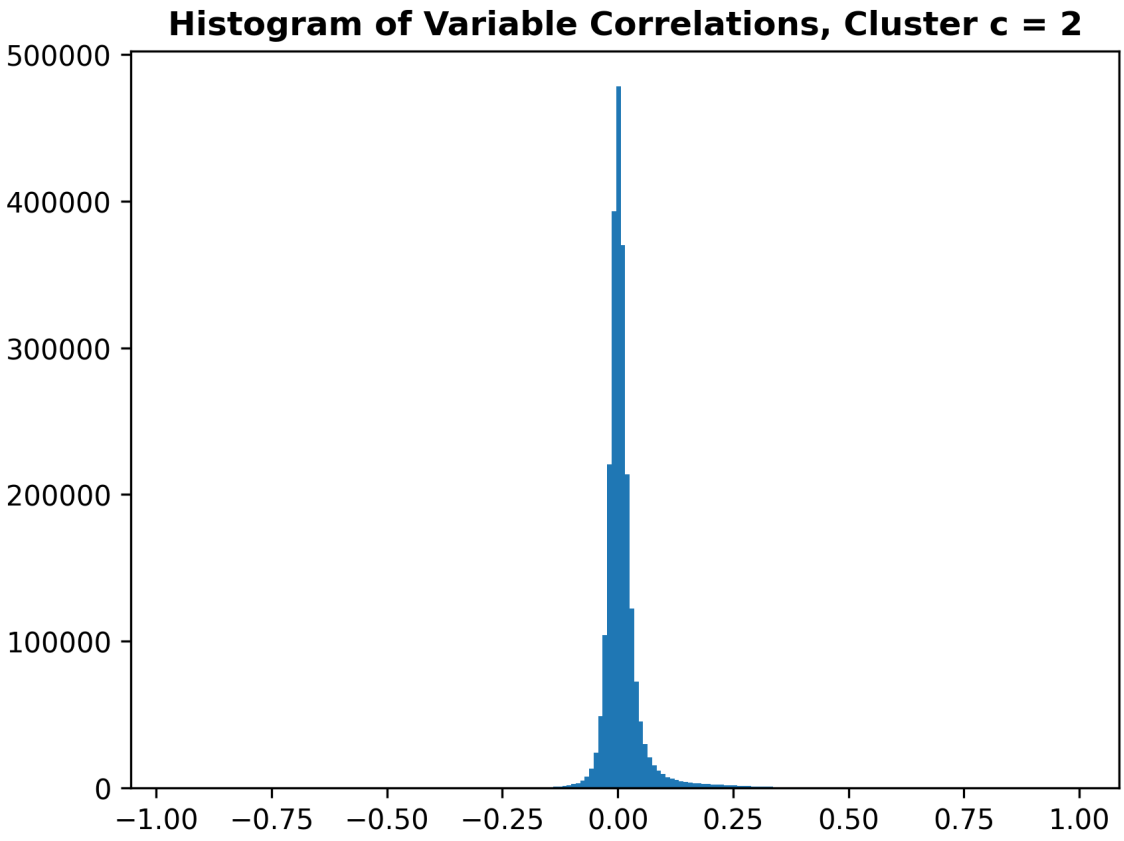


Figure A.13: Histogram of correlations, cluster $c = 2$.

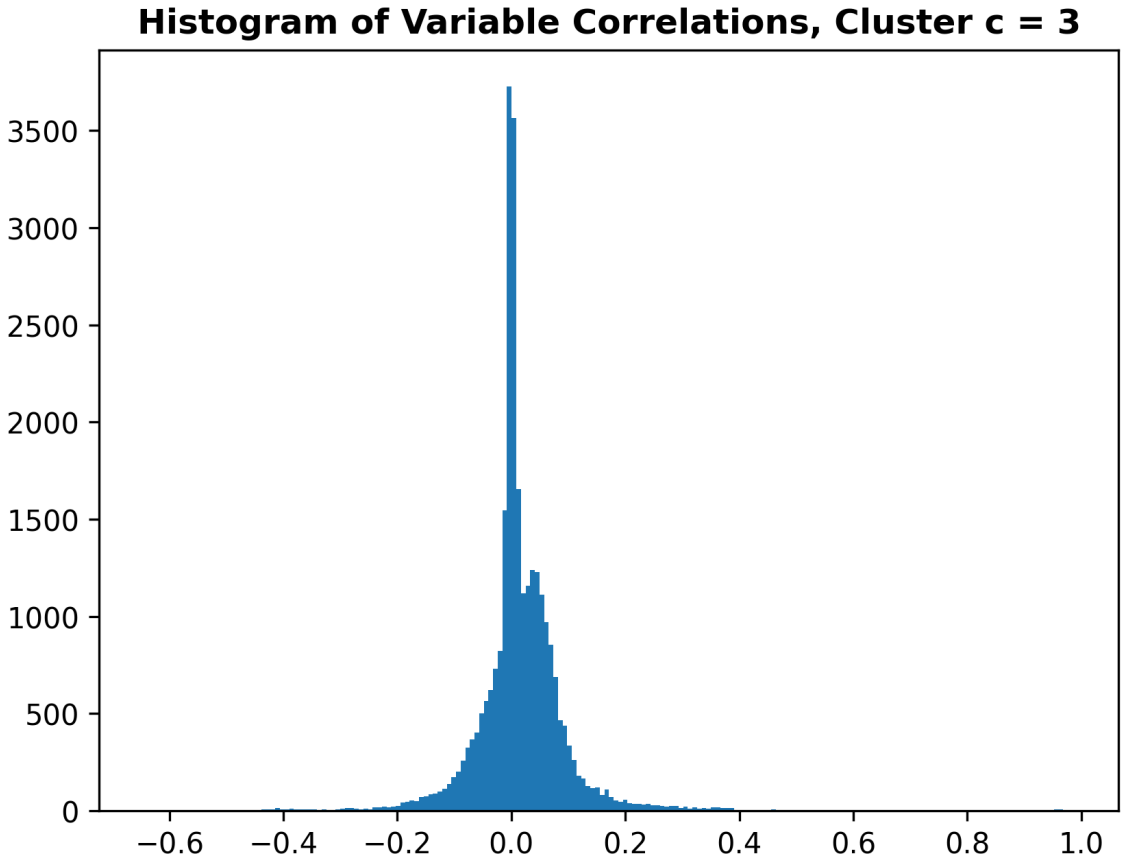


Figure A.14: Histogram of correlations, cluster $c = 3$.

Appendix B

Accounting for Structured Missingness in Canonical Correlation Analysis

B.1 Justification of Approximation

Assume that $f_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}_i^\top \mathbf{x} + \mathbf{c}_i$, $f_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, ...n$, is a series of quadratic functions with the same Hessian $\mathbf{A}$. Assume further that A is negative definite, so that f_i all have global maxima. We want to show that

$$\arg\max_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \arg\max_{\mathbf{x}\in\mathbb{R}^d} f_i(\mathbf{x}).$$

It is easy to derive that

$$\arg\max_{\mathbf{x}\in\mathbb{R}^d} f_i(\mathbf{x}) = -\mathbf{A}^{-1}\mathbf{b}_i,$$

by setting the gradient of f_i to $\mathbf{0}$. It therefore follows that

$$\frac{1}{n}\sum_{i=1}^{n} \arg\max_{\mathbf{x}\in\mathbb{R}^d} f_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} -\mathbf{A}^{-1}\mathbf{b}_i = -\mathbf{A}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{b}_i\right).$$

Let

$$\tilde{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{b}_i\right)^\top \mathbf{x} + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{c}_i\right).$$

$\tilde{f}$ is therefore also a quadratic function with Hessian $\mathbf{A}$, and it follows that

$$\arg\max_{\mathbf{x}\in\mathbb{R}^d} \tilde{f}(\mathbf{x}) = \arg\max_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = -\mathbf{A}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{b}_i\right).$$

$$\therefore \arg\max_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \arg\max_{\mathbf{x}\in\mathbb{R}^d} f_i(\mathbf{x}).$$

Q.E.D

This lemma shows that our pooling of covariance matrices after MICE approximately corresponds to finding the MLE, if the likelihood functions for the covariance matrix of the multiply imputed datasets with only SM are well approximated by quadratic functions that have the same Hessian.

B.2 Simulating Data

Algorithm 1: Generating Confounds $\mathbf{B}$

$r \leftarrow \text{round}(0.3d_B)$ **for** $i \leftarrow 1$ ***to*** r **do**
end
$s_i \sim \chi^2(1)$ $s_{\max} \leftarrow \max\{s_1, s_2, \ldots, s_r\}$ **for** $i \leftarrow 1$ ***to*** r **do**
end
$s_i \leftarrow 0.05 + s_i / s_{\max}$ **for** $i \leftarrow 1$ ***to*** n **do**
end
$j \leftarrow 1$ **to** r $U_{ij} \sim \mathcal{N}(0,1)$ **for** $i \leftarrow 1$ ***to*** d **do**
end
$j \leftarrow 1$ **to** r $V_{ij} \sim \mathcal{N}(0,1)$ $V_{ij} \leftarrow V_{ij}^3$ $\mathbf{B} \leftarrow \mathbf{U}\text{diag}\{s_1, s_2, \ldots, s_r\}\mathbf{V}^\top$ $\Sigma_{\text{NONOISE}} \leftarrow \mathbf{V}\text{diag}\{s_1^2, s_2^2, \ldots, s_r^2\}\mathbf{V}^\top$ **for** $i \leftarrow 1$ ***to*** n **do**
end
$j \leftarrow 1$ **to** d $t \sim \mathcal{N}(0, 1.5\{\Sigma_{\text{NONOISE}}\}_{jj})$ $B_{ij} \leftarrow B_{ij} + t$

B.3 CCA: Remaining Five Analyses

The plots for the remaining five pairs of CCA analyses are presented below in the following order:

1. IDPs vs. Food Preferences
2. IDPs vs. NMR Metabolomics
3. IDPs vs. Refractometry
4. Food Preferences vs Refractometry
5. NMR Metabolomics vs Refractometry

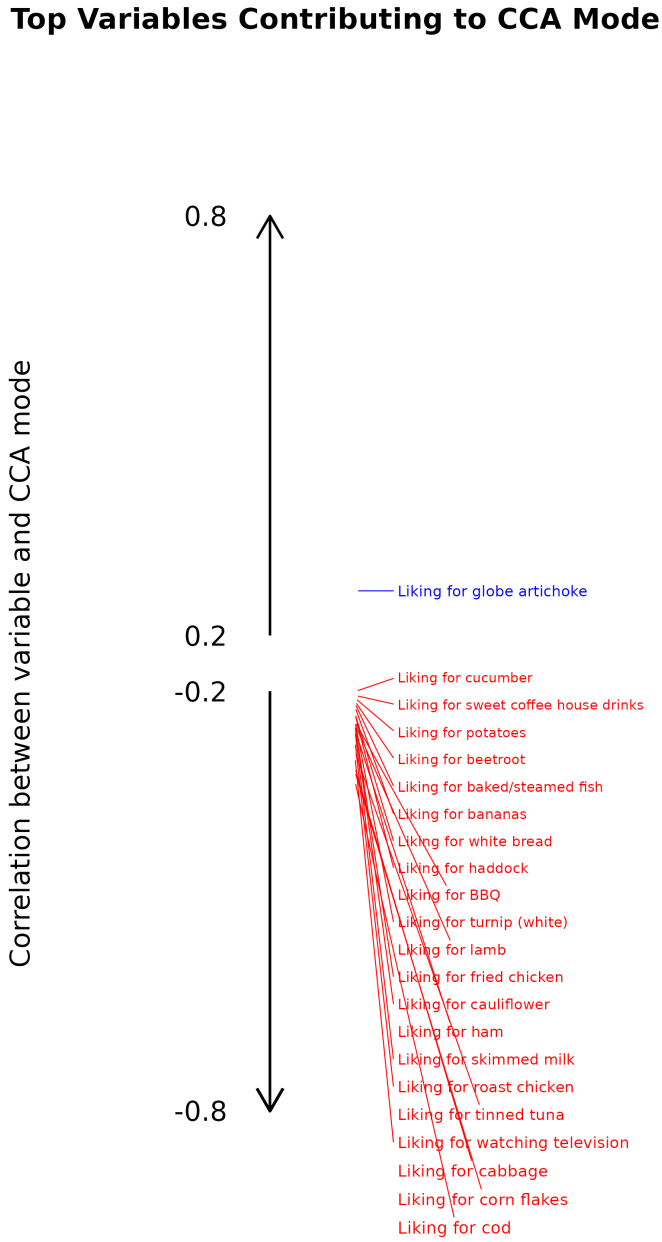


Figure B.1: Food preferences: top variables in terms of association with the first canonical mode (IDPs vs Food Preferences). The first canonical correlation is $r = 0.25$.

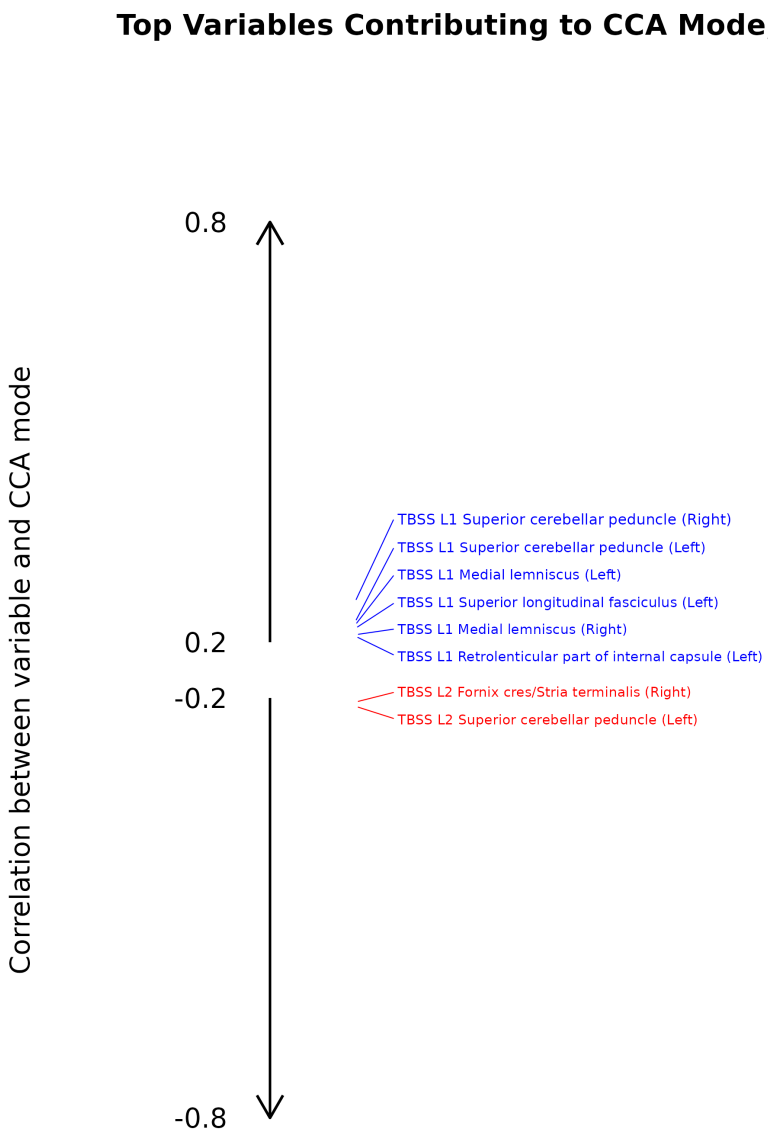


Figure B.2: IDPs: top variables in terms of association with the first canonical mode (IDPs vs Food Preferences). The first canonical correlation is $r = 0.25$.

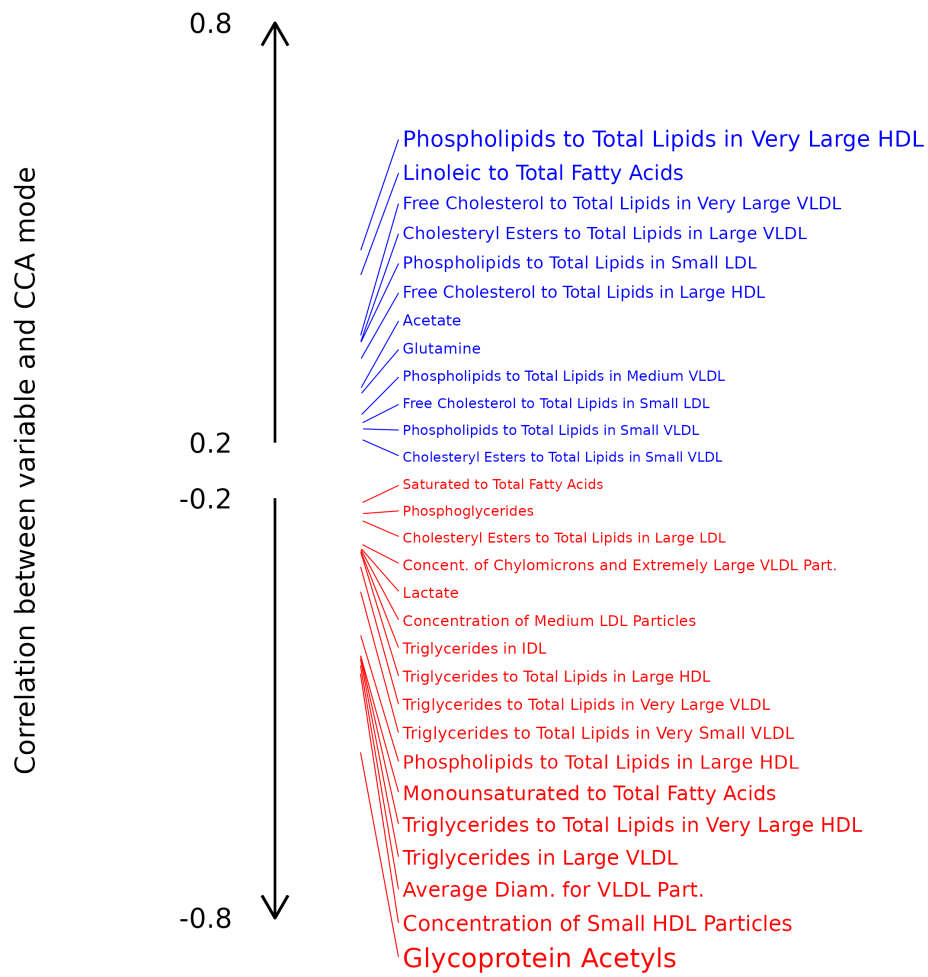


Figure B.3: NMR Metabolomics: top variables in terms of association with the first canonical mode (IDPs vs NMR Metabolomics). The first canonical correlation is $r = 0.29$.

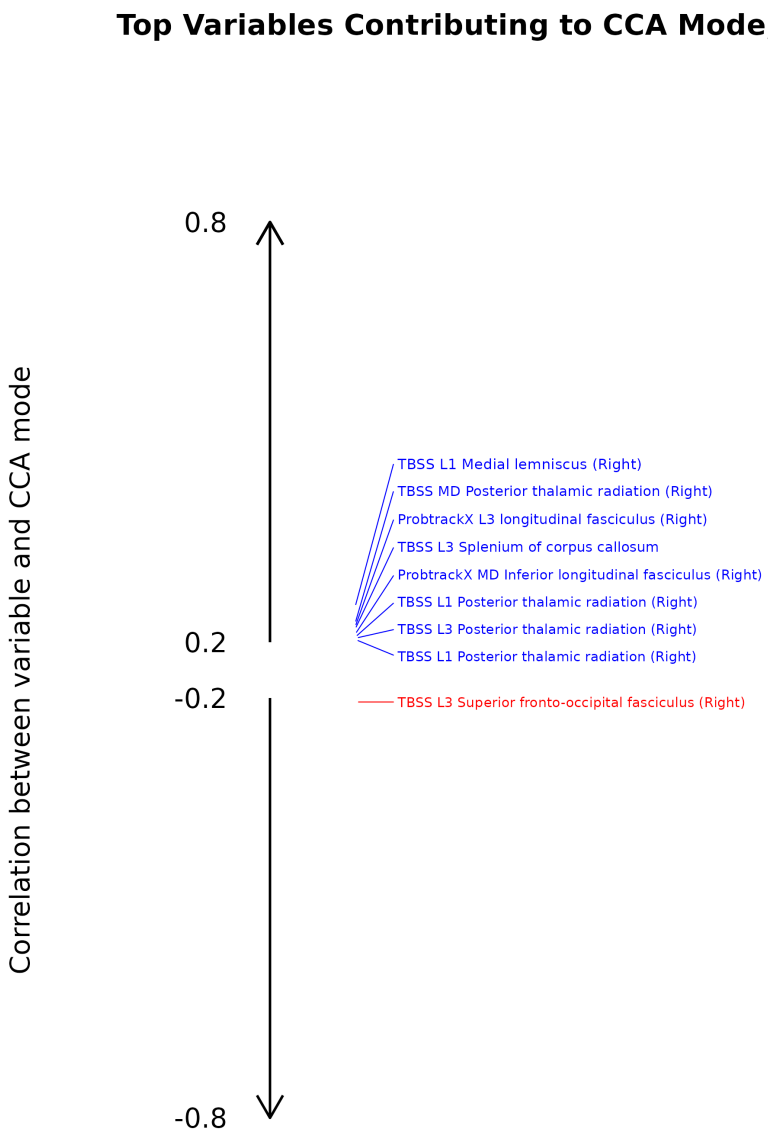


Figure B.4: IDPs: top variables in terms of association with the first canonical mode (IDPs vs NMR Metabolomics). The first canonical correlation is $r = 0.29$.

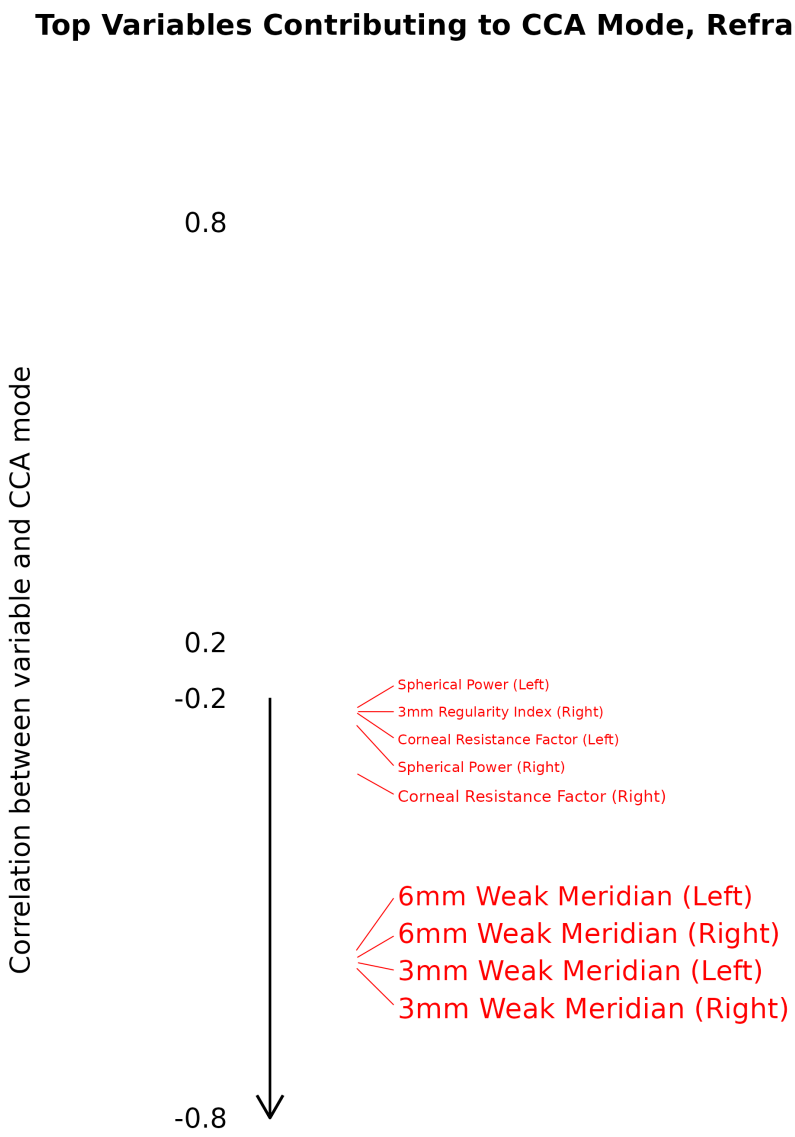


Figure B.5: Refractometry: top variables in terms of association with the first canonical mode (IDPs vs Refractometry). The first canonical correlation is $r = 0.28$.

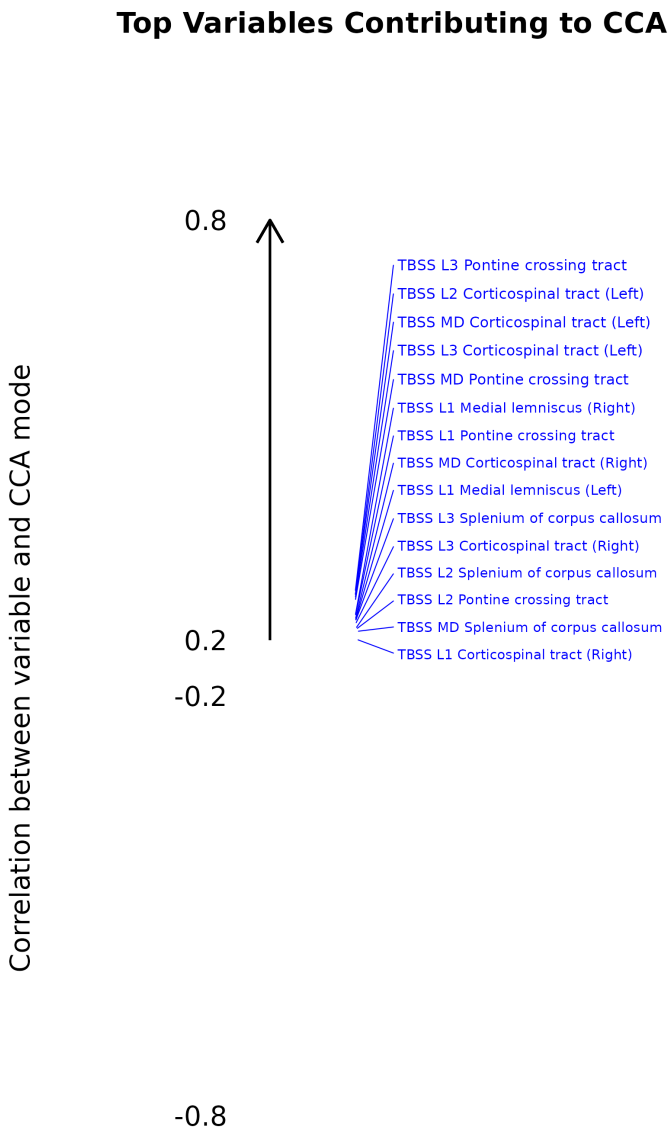


Figure B.6: IDPs: top variables in terms of association with the first canonical mode (IDPs vs Refractometry). The first canonical correlation is $r = 0.28$.



Figure B.7: Food Preferences: top variables in terms of association with the first canonical mode (Food Preferences vs Refractometry). The first canonical correlation is $r = 0.23$.

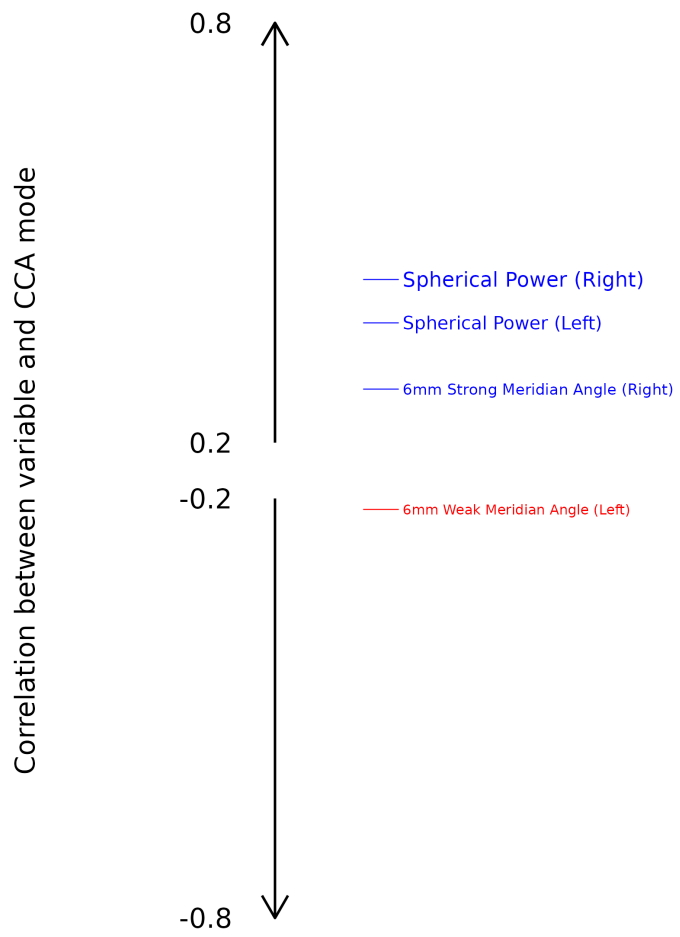


Figure B.8: Refractometry: top variables in terms of association with the first canonical mode (Food Preferences vs Refractometry). The first canonical correlation is $r = 0.23$.

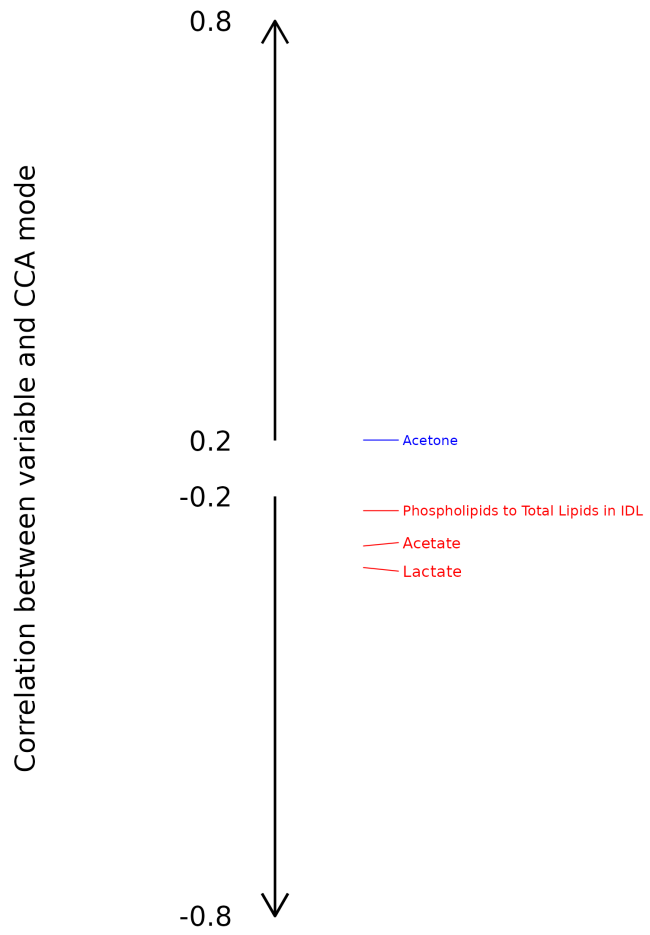


Figure B.9: NMR Metabolomics: top variables in terms of association with the first canonical mode (NMR Metabolomics vs Refractometry). The first canonical correlation is $r = 0.42$.

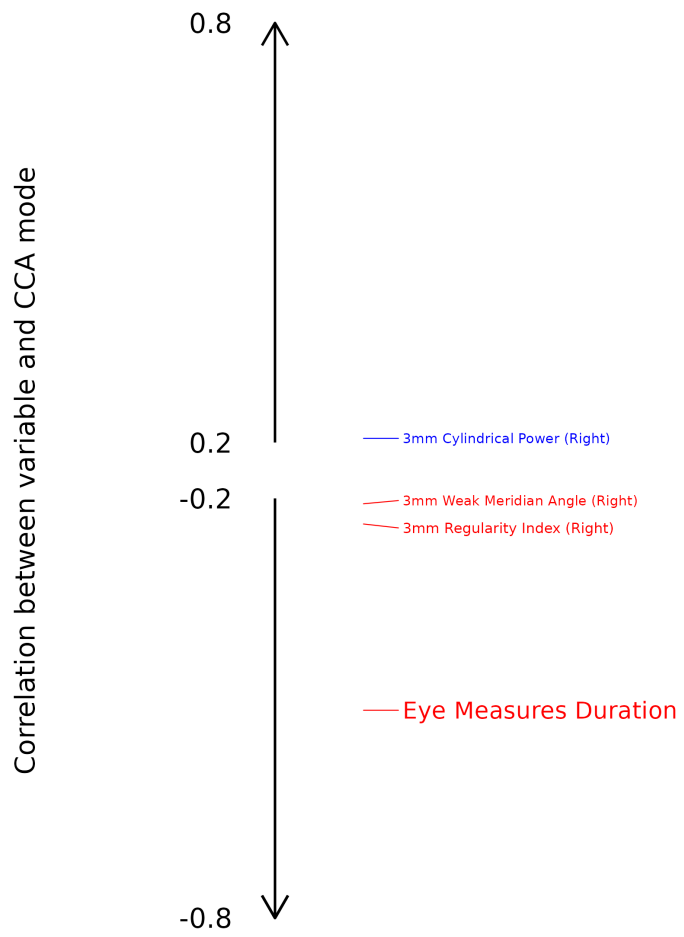


Figure B.10: Refractometry: top variables in terms of association with the first canonical mode (NMR Metabolomics vs Refractometry). The first canonical correlation is $r = 0.42$.

Appendix C

An Improved Pipeline for Constructing UK Biobank Brain Imaging Confounds

C.1 Proof of consistency of $\hat{R}^2_{\text{cor}}$

Assume the following linear model:

$$y = \mathbf{x}^\top \beta + \epsilon,$$

where y is a scalar random variable, $\mathbf{x}$ is a d-dimensional explanatory variable, a random vector with expectation $\mathbf{0}$ and covariance matrix Σ, β is a fixed d-dimensional vector of coefficients and ϵ is white noise with variance σ^2_ϵ. While the standard setting assumes a fixed design matrix, we consider behavior for ever-growing n samples of $(\mathbf{x}, y)$. For this linear model, the proportion of variance explained is:

$$R^2 = \frac{\text{Var}(\mathbf{x}^\top \beta)}{\text{Var}(y)} = \frac{\beta^\top \Sigma \beta}{\beta^\top \Sigma \beta + \sigma^2_\epsilon} = 1 - \frac{\sigma^2_\epsilon}{\beta^\top \Sigma \beta + \sigma^2_\epsilon}.$$

Let us assume that we have $n = n_{\text{obs}} + n_{\text{mis}}$ samples of $(\mathbf{x}, y)$, where n_{mis} samples

of $\mathbf{y} = (y_1, y_2, ..., y_n)$ have been randomly censored in an MCAR fashion at the rate of p_{miss}. We thus construct $\mathbf{y}_{\text{imp}}$, where the missing values of $\mathbf{y}$ are imputed with $\bar{y}_{\text{obs}}$, the arithmetic mean of the observed values. We now have a design matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top$ and response $\mathbf{y}_{\text{imp}} = (y_{\text{imp},1}, .., y_{\text{imp},n})^\top$, and we want to show that

$$\hat{R}^2_{\text{cor}} = \frac{\hat{R}^2_{\text{imp}}}{1 - \hat{p}_{\text{miss}}} = \frac{1}{1 - \hat{p}_{\text{miss}}} \left(1 - \frac{\text{RSS}_{\text{imp}}}{\text{TSS}_{\text{imp}}}\right),$$

where $\hat{p}_{\text{miss}}$ is the estimated rate of missingness, is a consistent estimator of R^2.

We start by showing that

$$\hat{\sigma}^2_{y,\text{imp}} = \frac{1}{n} \sum_{i=1}^{n} (y_{\text{imp},i} - \bar{y}_{\text{imp}})^2$$

is a consistent estimator of $(1 - p_{\text{miss}})\sigma^2_y$. We first observe that $\bar{y}_{\text{imp}} = \bar{y}_{\text{obs}}$. This means that $y_{\text{imp},i} - \bar{y}_{\text{imp}} = 0$ for all indices i where y is missing.

$$\therefore \hat{\sigma}^2_{y,\text{imp}} = \frac{1}{n} \sum_{i=1}^{n} (y_{\text{imp},i} - \bar{y}_{\text{imp}})^2 = \frac{n_{\text{obs}}}{n} \left(\frac{1}{n_{\text{obs}}} \sum_{i=1}^{n_{\text{obs}}} (y_{\text{obs},i} - \bar{y}_{\text{obs}})^2 \right).$$

Since $\frac{n_{\text{obs}}}{n}$ is a consistent estimator of $1 - p_{\text{miss}}$ and $\frac{1}{n_{\text{obs}}} \sum_{i=1}^{n_{\text{obs}}} (y_{\text{obs},i} - \bar{y}_{\text{obs}})^2$ is a consistent estimator of σ^2_y under MCAR, it follows from the continuous mapping theorem that $\hat{\sigma}^2_{y,\text{imp}}$ is a consistent estimator of $(1 - p_{\text{miss}})\sigma^2_y$. We continue by proving that

$$\frac{1}{n} \mathbf{X}^\top \mathbf{y}_{\text{imp}}$$

is a consistent estimator of $(1 - p_{\text{miss}})\text{Cov}(\mathbf{x}, y) = (1 - p_{\text{miss}})\Sigma\beta$. We know that

$$\frac{1}{n} \mathbf{X}^\top \mathbf{y}_{\text{imp}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \cdot y_{\text{imp},i} = \frac{n_{\text{obs}}}{n} \left(\frac{1}{n_{\text{obs}}} \sum_{i=1}^{n_{\text{obs}}} \mathbf{x}_{\text{obs},i} \cdot y_{\text{obs},i} \right) + \frac{n_{\text{miss}}}{n} \left(\frac{1}{n_{\text{miss}}} \sum_{i=1}^{n_{\text{miss}}} \mathbf{x}_{\text{miss},i} \cdot \bar{y}_{\text{obs}} \right),$$

and since $\frac{n_{\text{obs}}}{n}$ is a consistent estimator of $1 - p_{\text{miss}}$, $\frac{1}{n_{\text{obs}}} \sum_{i=1}^{n_{\text{obs}}} \mathbf{x}_{\text{obs},i} \cdot y_{\text{obs},i}$ is a consistent estimator of $\text{Cov}(\mathbf{x}, y)$ and $\frac{1}{n_{\text{miss}}} \sum_{i=1}^{n_{\text{miss}}} \mathbf{x}_{\text{miss},i}$ is a consistent estimator of $\mathbb{E}(\mathbf{x}) = \mathbf{0}$, it follows from the continuous mapping theorem that $\frac{1}{n} \mathbf{X}^\top \mathbf{y}_{\text{imp}}$ is a consistent estimator of $(1 - p_{\text{miss}})\text{Cov}(\mathbf{x}, y) = (1 - p_{\text{miss}})\Sigma\beta$.

Now we want to prove that

$$\hat{\sigma}^2_{\epsilon,\text{imp}} = \frac{1}{n}\left((\mathbf{I}-\mathbf{H})\mathbf{y}_{\text{imp}}\right)^\top(\mathbf{I}-\mathbf{H})\mathbf{y}_{\text{imp}} = \frac{1}{n}\mathbf{y}_{\text{imp}}^\top(\mathbf{I}-\mathbf{H})\mathbf{y}_{\text{imp}} =$$

$$= \frac{1}{n}\mathbf{y}_{\text{imp}}^\top\mathbf{y}_{\text{imp}} - \frac{1}{n}\mathbf{y}_{\text{imp}}^\top\mathbf{H}\mathbf{y}_{\text{imp}},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a consistent estimator of $(1-p_{\text{miss}})\sigma_y^2 - (1-p_{\text{miss}})^2\beta^\top\Sigma\beta$. We know that

$$\frac{1}{n}\mathbf{y}_{\text{imp}}^\top\mathbf{y}_{\text{imp}} = \frac{1}{n}\sum_{i=1}^{n} y_{\text{imp},i}^2 = \frac{n_{\text{obs}}}{n}\left(\frac{1}{n_{\text{obs}}}\sum_{i=1}^{n_{\text{obs}}} y_{\text{obs},i}^2\right) + \frac{n_{\text{miss}}}{n}\bar{y}_{\text{obs}}^2,$$

and since $\frac{n_{\text{obs}}}{n}$ is a consistent estimator of $1-p_{\text{miss}}$, $\bar{y}_{\text{obs}}$ is a consistant estimator of $\mathbb{E}(y)=0$ and $\frac{1}{n_{\text{obs}}}\sum_{i=1}^{n_{\text{obs}}} y_{\text{obs},i}^2$ is a consistent estimator of $\mathbb{E}(y)^2+\text{Var}(y) = 0^2+\sigma_y^2 = \sigma_y^2$, it follows from the continuous mapping theorem that $\frac{1}{n}\mathbf{y}_{\text{imp}}^\top\mathbf{y}_{\text{imp}}$ is a consistent estimator of $(1-p_{\text{miss}})\sigma_y^2$. We continue by observing that

$$\frac{1}{n}\mathbf{y}_{\text{imp}}^\top\mathbf{H}\mathbf{y}_{\text{imp}} = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{y}_{\text{imp}}\right)^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y}_{\text{imp}}\right).$$

Here we use our previous result that $\frac{1}{n}\mathbf{X}^\top\mathbf{y}_{\text{imp}}$ is a consistent estimator of $(1-p_{\text{miss}})\Sigma\beta$, as well as the fact that $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is a consistent estimator of Σ, and get from the continuous mapping theorem that $\frac{1}{n}\mathbf{y}_{\text{imp}}^\top\mathbf{H}\mathbf{y}_{\text{imp}}$ is a consistent estimator of

$$(1-p_{\text{miss}})\beta^\top\Sigma^\top\Sigma^{-1}(1-p_{\text{miss}})\Sigma\beta = (1-p_{\text{miss}})^2\beta^\top\Sigma\beta.$$

From this, it finally follows that $\hat{\sigma}^2_{\epsilon,\text{imp}}$ is a consistent estimator of

$$(1-p_{\text{miss}})\sigma_y^2 - (1-p_{\text{miss}})^2\beta^\top\Sigma\beta.$$

We now have all results necessary to prove the consistency of $\hat{R}^2_{\text{cor}}$. We note that

$$\frac{\mathrm{RSS}_{\mathrm{imp}}}{\mathrm{TSS}_{\mathrm{imp}}} = \frac{\mathbf{y}_{\mathrm{imp}}^{\top}\mathbf{y}_{\mathrm{imp}} - \mathbf{y}_{\mathrm{imp}}^{\top}\mathbf{H}\mathbf{y}_{\mathrm{imp}}}{\sum_{i=1}^{n}(y_{\mathrm{imp},i} - \bar{y}_{\mathrm{imp}})^2} = \frac{\frac{1}{n}\mathbf{y}_{\mathrm{imp}}^{\top}\mathbf{y}_{\mathrm{imp}} - \frac{1}{n}\mathbf{y}_{\mathrm{imp}}^{\top}\mathbf{H}\mathbf{y}_{\mathrm{imp}}}{\frac{1}{n}\sum_{i=1}^{n}(y_{\mathrm{imp},i} - \bar{y}_{\mathrm{imp}})^2} =$$

$$= \frac{\hat{\sigma}^2_{\epsilon,\mathrm{imp}}}{\hat{\sigma}^2_{y,\mathrm{imp}}}.$$

It thus follows that

$$\hat{R}^2_{\mathrm{cor}} = \frac{1}{1 - \hat{p}_{\mathrm{miss}}}\left(1 - \frac{\hat{\sigma}^2_{\epsilon,\mathrm{imp}}}{\hat{\sigma}^2_{y,\mathrm{imp}}}\right).$$

Using our results and a combination of the continuous mapping theorem and Slutsky's theorem, we finally get that $\hat{R}^2_{\mathrm{cor}}$ is a consistent estimator of

$$\frac{1}{1 - p_{\mathrm{miss}}}\left(1 - \frac{(1 - p_{\mathrm{miss}})\sigma_y^2 - (1 - p_{\mathrm{miss}})^2\beta^{\top}\Sigma\beta}{(1 - p_{\mathrm{miss}})\sigma_y^2}\right) = \frac{1}{1 - p_{\mathrm{miss}}}\left(\frac{(1 - p_{\mathrm{miss}})^2\beta^{\top}\Sigma\beta}{(1 - p_{\mathrm{miss}})\sigma_y^2}\right) =$$

$$= \frac{\beta^{\top}\Sigma\beta}{\sigma_y^2} = R^2.$$

Q.E.D

C.2 Algorithm

The proposed forward selection algorithm has two pre-determined thresholds τ_{mean} and τ_{max} which determine when no remaining potential confounds are important enough to be included in the final confound set. The algorithm is described in detail below using pseudocode. The functions `P_MISS`, `PINV` and `MEAN_IMPUTER` perform the calculation of column-wise missingness rates, the Moore-Penrose pseudo-inverse and mean imputation respectively. All other function names are assumed to be self-explanatory. The operator $*$ is used for matrix multiplication, all other operations are assumed to be applied element-wise. All data is standardised before running the algorithm.

Algorithm 1: Algorithm: Selecting Potential Confounds

IDPs $\leftarrow$ `MEAN_IMPUTER`(IDPs) Mean imputation in the IDPs $p_{\text{miss}} \leftarrow$ `P_MISS`(IDPs) Calculate missingness rates for each IDP
TSS $\leftarrow \mathbf{1}_n^\top * \text{IDPs}^2$ Total Sum of Squares for each IDP
$\mathbf{C}_{\text{fix}}^+ \leftarrow$ `PINV`($\mathbf{C}_{\text{fix}}$) Pseudo-inverse of $\mathbf{C}_{\text{fix}}$ $\mathbf{C}_{\text{pot}} \leftarrow \mathbf{C}_{\text{pot}} - \mathbf{C}_{\text{fix}} * (\mathbf{C}_{\text{fix}}^+ * \mathbf{C}_{\text{pot}})$ Regress $\mathbf{C}_{\text{fix}}$ out of $\mathbf{C}_{\text{pot}}$
IDPs $\leftarrow$ IDPs $- \mathbf{C}_{\text{fix}} * (\mathbf{C}_{\text{fix}}^+ * \text{IDPs})$ Regress $\mathbf{C}_{\text{fix}}$ out of IDPs RSS $\leftarrow \mathbf{1}_n^\top * \text{IDPs}^2$ Residual Sum of Squares for each IDP after regressing out $\mathbf{C}_{\text{fix}}$
$R_0^2 \leftarrow (1 - \text{RSS/TSS})/(1 - p_{\text{miss}})$ %Variance Explained by $\mathbf{C}_{\text{fix}}$ in each IDP. **while** *TRUE* **do**
end
`YY` $\leftarrow \mathbf{C}_{\text{pot}}^\top * \text{IDPs}$ Inner product matrix between $\mathbf{C}_{\text{pot}}$ and IDPs `YY` $\leftarrow \mathbf{1}_n^\top * \text{IDPs}^2$ TSS for all IDPs `XX` $\leftarrow \mathbf{1}_n^\top * \mathbf{C}_{\text{pot}}^2$ TSS for all potential confounds RSS $\leftarrow$ `YY` $-$ `XY`$^2/$(`XX`)$^\top$ Matrix of new RSS after regressing each potential confound out of each IDP $R_1^2 \leftarrow (1 - \text{RSS/TSS})/(1 - p_{\text{miss}})$ Matrix of new R^2 after regressing each potential confound out of each IDP $\Delta R^2 \leftarrow R_1^2 - R_0^2$ Matrix of change in R^2 after regressing each potential confound out of each IDP $\Delta_{\text{mean}} = \{$`MEAN`$(\{\Delta R^2\}_{\cdot i}) | i = 1, ..., \#\text{IDPs}\}$ Find mean change in R^2 for each IDP $\Delta_{\text{max}} = \{$`MAX`$(\{\Delta R^2\}_{\cdot i}) | i = 1, ..., \#\text{IDPs}\}$ Find maximal change in R^2 for each IDP **if** *MAX*$(\Delta_{\text{mean}}) < \tau_{\text{mean}}$ *and* *MAX*$(\Delta_{\text{max}}) < \tau_{\text{max}}$ **then**
end
`BREAK` If no IDPs have changes in R^2 above the thresholds, the algorithm ends
if *MAX*$(\Delta_{\text{mean}}) < \tau_{\text{mean}}$ **then**
end
$i_{\text{mean}} \leftarrow$ `ARGMAX`(Δ_{mean}) Find potential confound with maximal max change in R^2 **else**
end
$i_{\text{mean}} \leftarrow \emptyset$ **if** *MAX*$(\Delta_{\text{max}}) < \tau_{\text{max}}$ **then**
end
$i_{\text{max}} \leftarrow$ `ARGMAX`(Δ_{max}) Find potential confound with maximal mean change in R^2 **else**
end
$i_{\text{max}} \leftarrow \emptyset$
if $|$*CORR*$(\{\mathbf{C}_{\text{pot}}\}_{\cdot i_{\text{mean}}}, \{\mathbf{C}_{\text{pot}}\}_{\cdot i_{\text{max}}})| < 0.95$ **then**
end
$i_{\text{max}} \leftarrow \emptyset$ If the two selected variables are too correlated, only select one $\mathbf{C}_{\text{sel}} \leftarrow \{\mathbf{C}_{\text{pot}}\}_{\cdot \{i_{\text{mean}}, i_{\text{max}}\}}$
Matrix of selected confounds $\mathbf{C}_{\text{sel}}^+ \leftarrow$ `PINV`($\mathbf{C}_{\text{sel}}$) Pseudo-inverse of selected confounds matrix
IDPs $\leftarrow$ IDPs $- \mathbf{C}_{\text{sel}} * (\mathbf{C}_{\text{sel}}^+ * \text{IDPs})$ Regress selected confounds out of IDPs
$\mathbf{C}_{\text{pot}} \leftarrow \mathbf{C}_{\text{pot}} - \mathbf{C}_{\text{sel}} * (\mathbf{C}_{\text{sel}}^+ * \mathbf{C}_{\text{pot}})$ Regress selected confounds out of $\mathbf{C}_{\text{pot}}$
$\mathbf{C}_{\text{fix}} \leftarrow$ `CONCATENATE`($\mathbf{C}_{\text{fix}}, \mathbf{C}_{\text{sel}}$) Add selected confounds to fixed confounds
$\mathbf{C}_{\text{pot}} \leftarrow$ `REMOVE_COLUMNS`($\mathbf{C}_{\text{pot}}, \{i_{\text{mean}}, i_{\text{max}}\}$) Remove selected confounds from potential confounds
$R_0^2 \leftarrow (1 - \mathbf{1}_n^\top * \text{IDPs}^2/\text{TSS})/(1 - p_{\text{miss}})$ New %variance explained of fixed confounds

We can see from the algorithm that the most computationally costly operations

is initially calculating the pseudo-inverse of $\mathbf{C}_{\text{fix}}$, and also performing the matrix multiplication $\mathbf{C}_{\text{pot}}^{\top} * \text{IDPs}$ in each iteration. If d_{fix} is the number of fixed confounds, d_{pot} the number of potential confounds, d_{IDPs} the number of IDPs, n the number of subjects, and $s_{\text{pot}} << d_{\text{pot}}$ the number of relevant potential IDPs, then the time complexity of our algorithm is:

$$\mathcal{O}(nd_{\text{fix}}^2 + nd_{\text{pot}}d_{\text{IDPs}}s_{\text{pot}}).$$

It should also be noted that all computations are vectorised as we only use matrix multiplication and element-wise matrix/vector manipulation in the algorithm.

C.3 Further Comparisons of Old and New Confounds

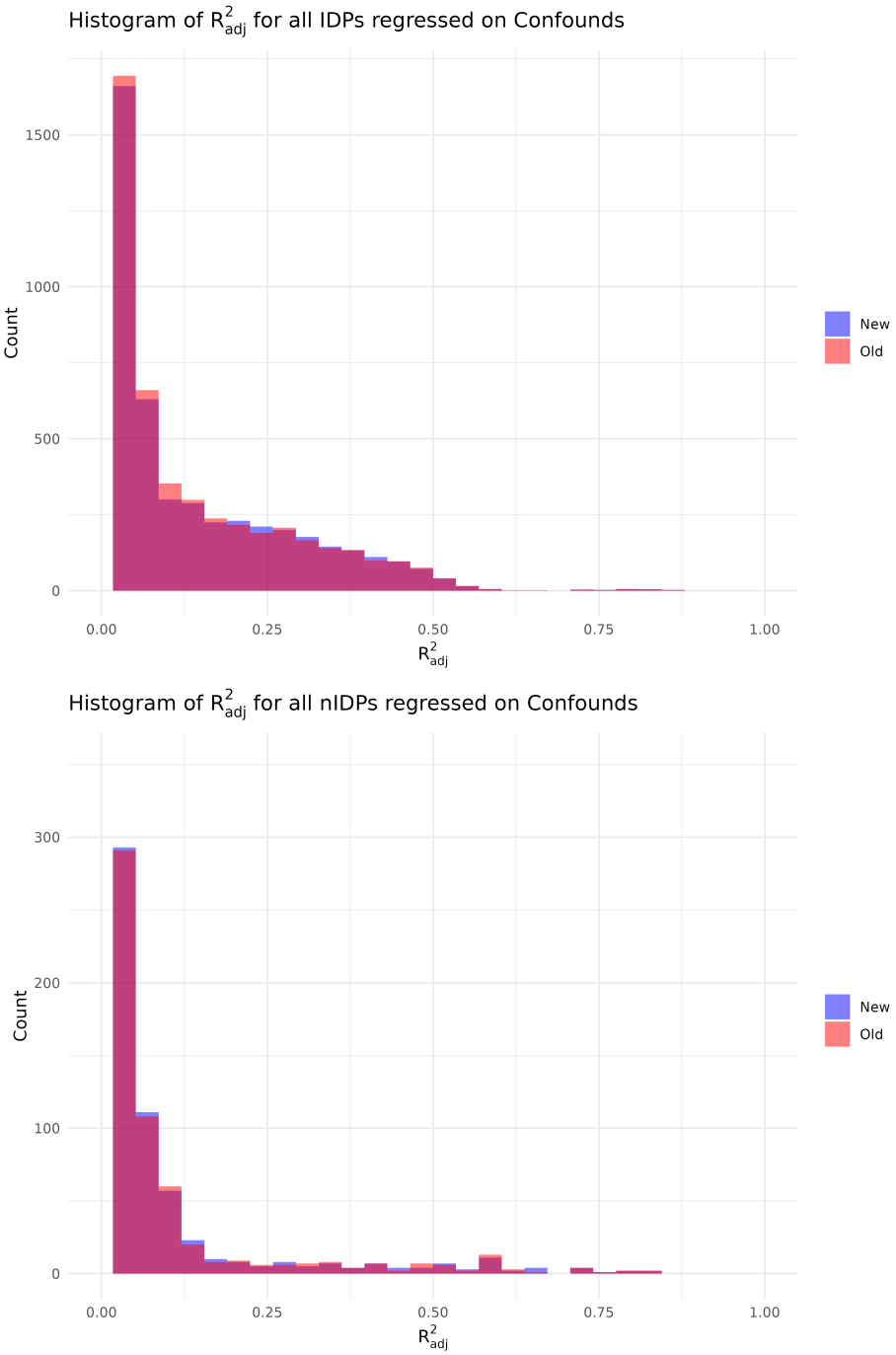


Figure C.1: Overlapping histograms of the proportion of variance explained in the IDPs and nIDPs by old and new confounds. We see that there is great overlap in distribution of variance explained for the old and new confounds.

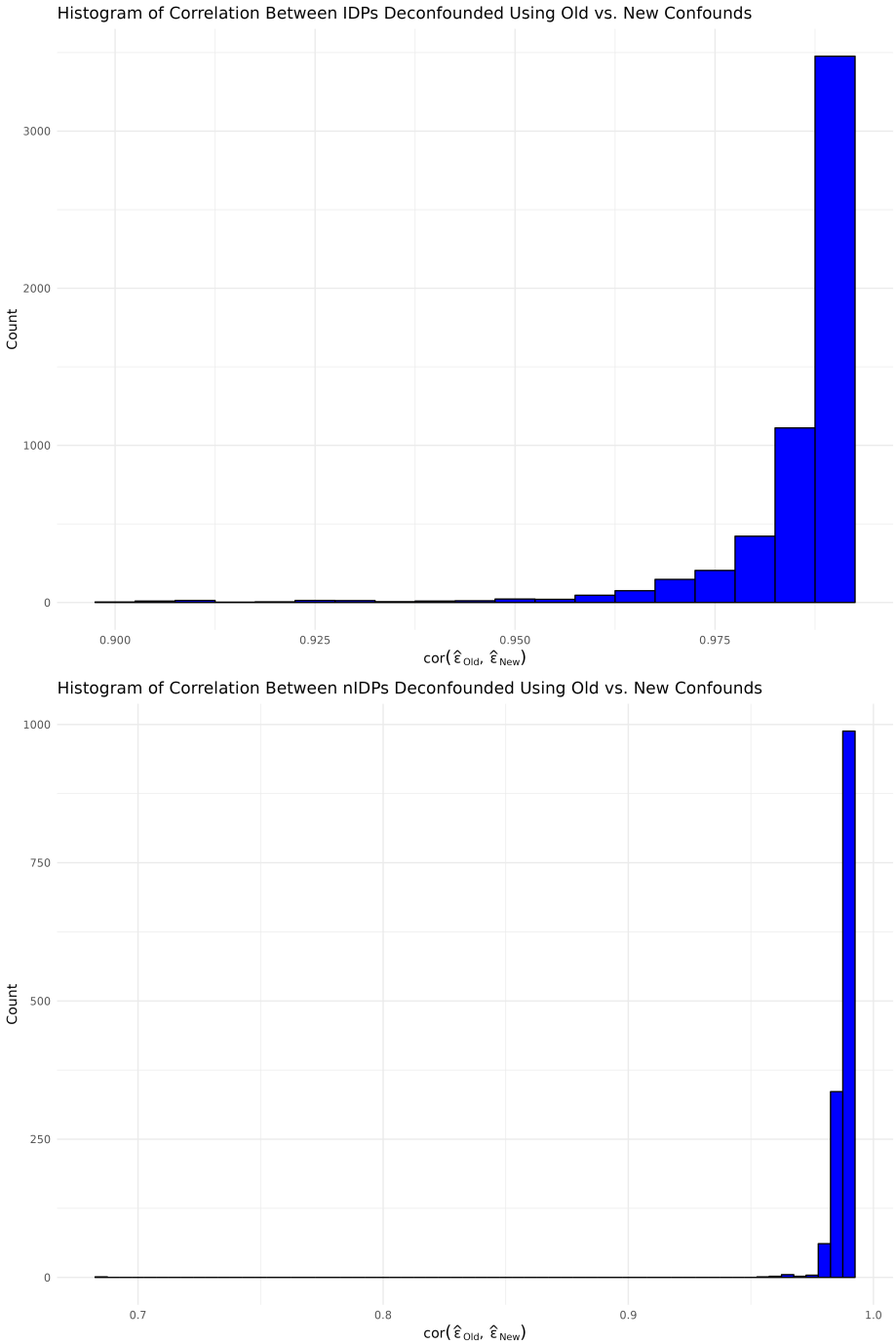


Figure C.2: Histograms of the correlation between IDPs and nIDPs deconfounded using old and new confounds. We see a very high correlation between IDPs and nIDPs deconfounded using different confound sets, meaning that deconfounding has a very similar effect on IDPs/nIDPs when using either confound set.

C.4 Comparing Double Selection and Single Selection

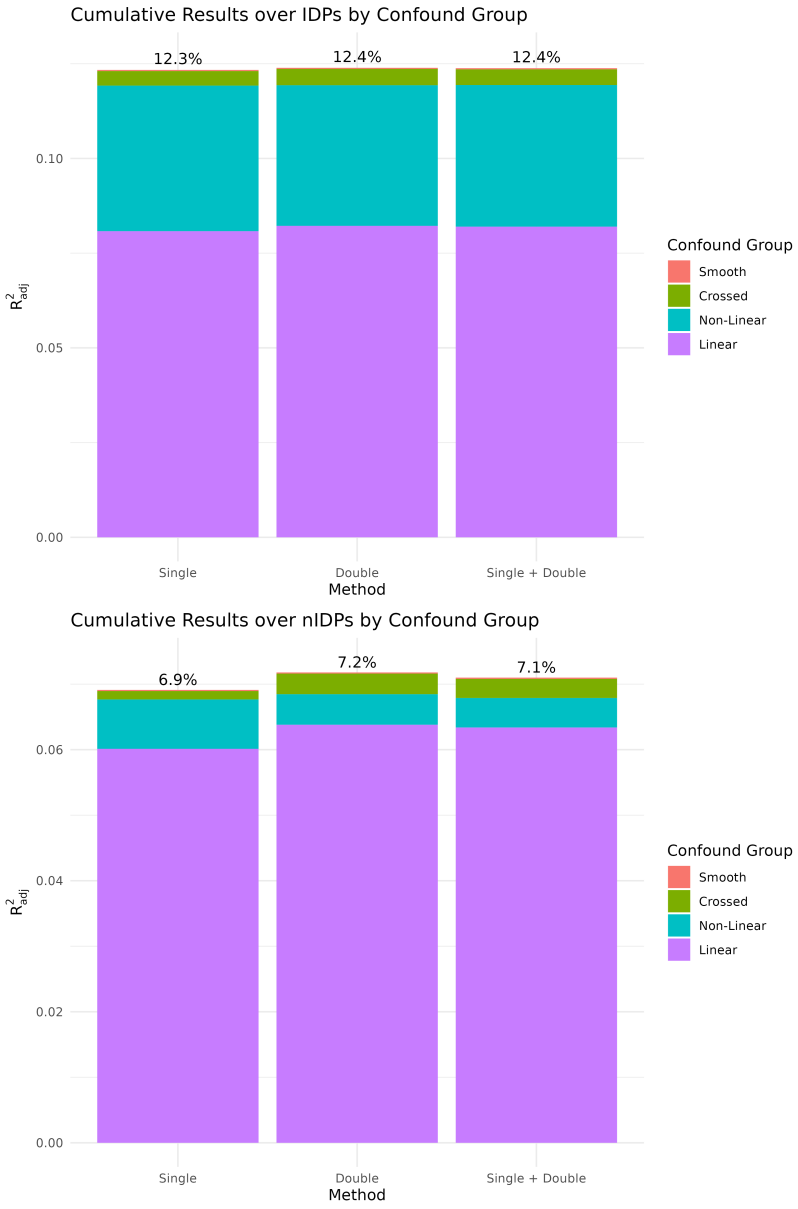


Figure C.3: Bar plots comparing the average variance explained using double- vs. using single-selection. For IDPs, the proportions are almost identical. For nIDPs, double selection leads to a small but noticeable rise in variance explained.

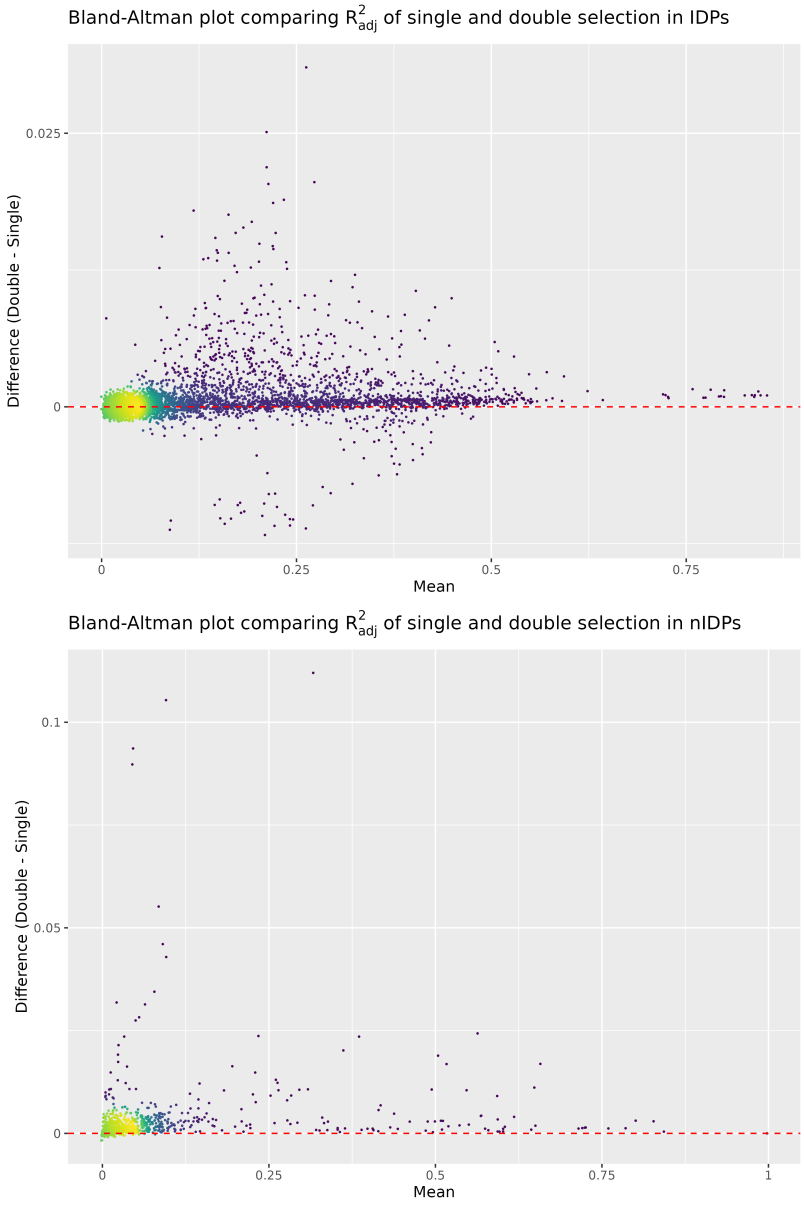


Figure C.4: Bland-Altman plots illustrating the difference in variance explained in each IDP/nIDP by confound sets constructed using double- vs. single selection. For IDPs, we see very strong agreement between the two different methods. For nIDPs, we see that double-selection leads to consistently higher variance explained but generally not drastically higher.

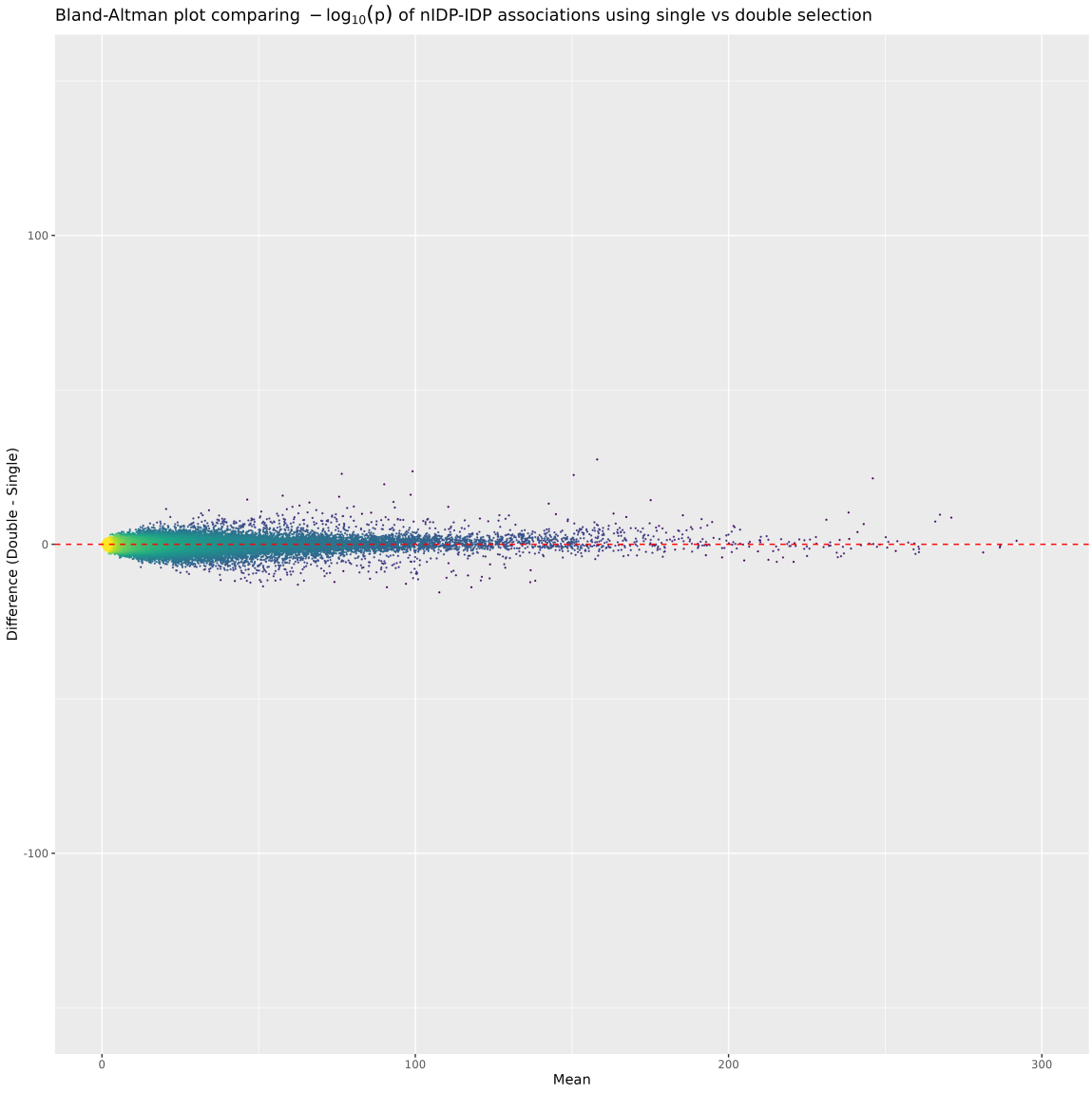


Figure C.5: Bland-Altman plot of $-\log_{10}(p)$ for univariate IDP on nIDP associations using confounds constructed with single- or double-selection. The plot shows that p-values obtained after deconfounding with either confound set are extremely close.

Appendix D

Controlling for confounds in UK Biobank brain imaging data with small subsets of subjects

D.1 Supplementary Figures

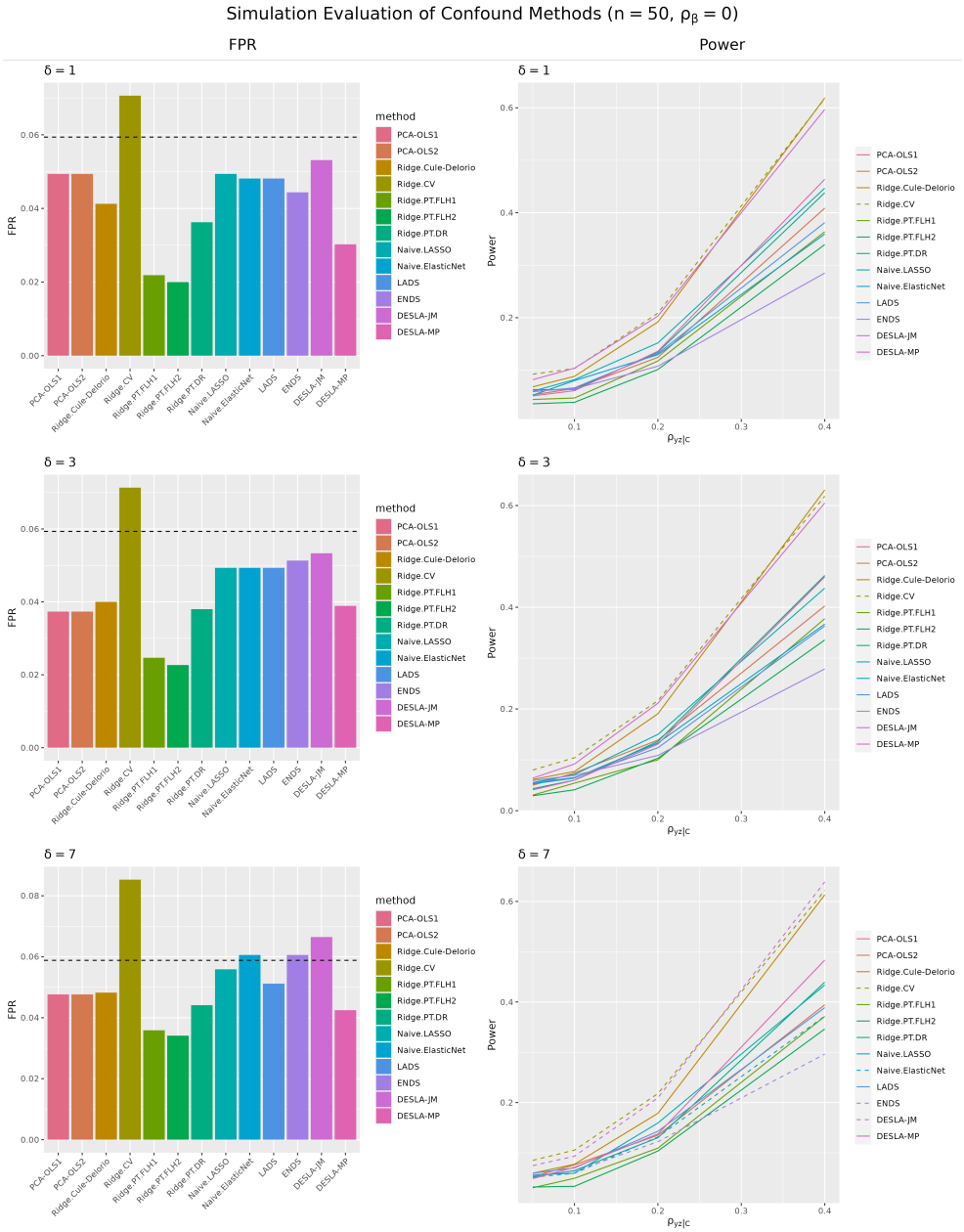


Figure D.1: Simulation results for $n = 50$, $\rho_\beta = 0$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

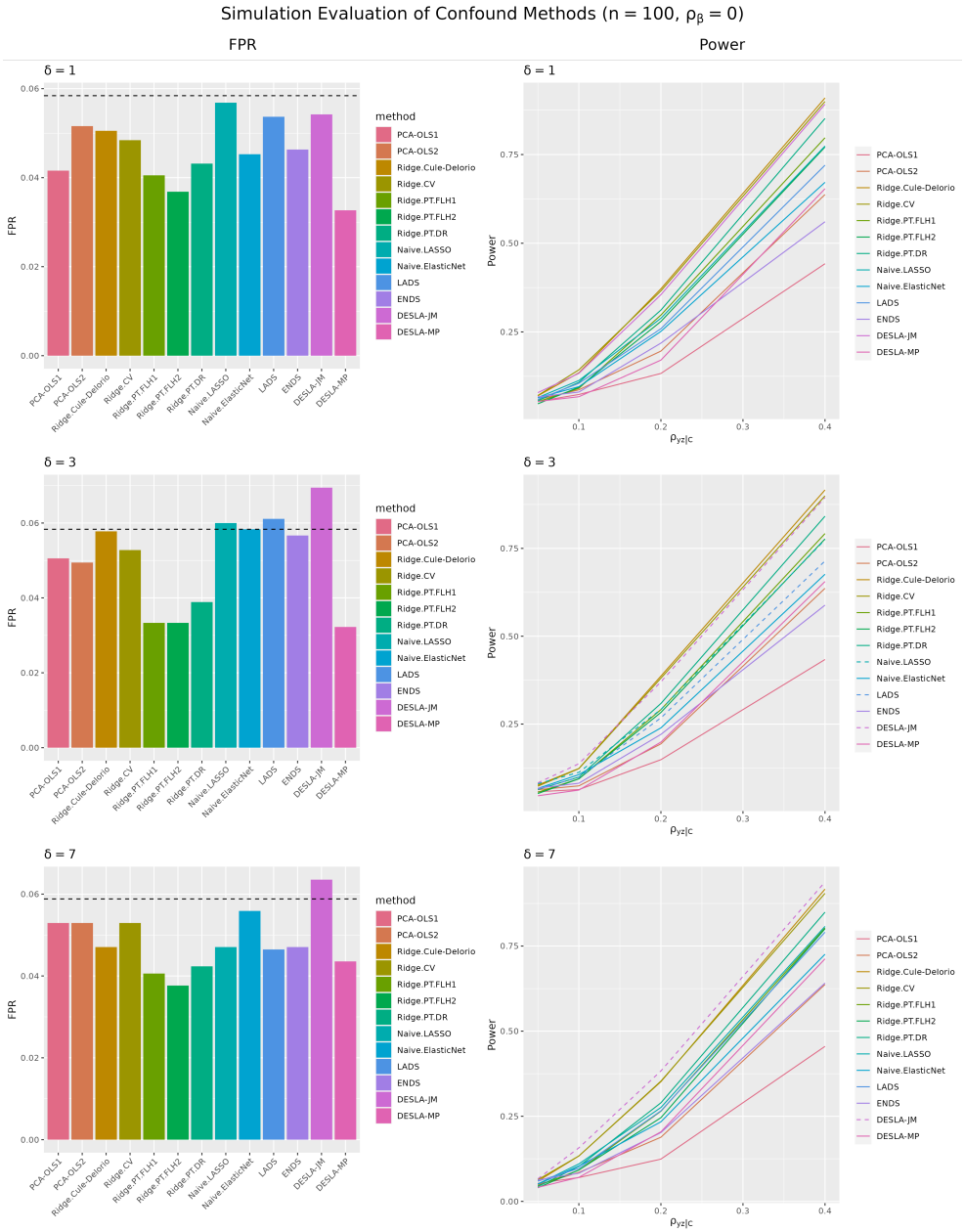


Figure D.2: Simulation results for $n = 100$, $\rho_\beta = 0$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

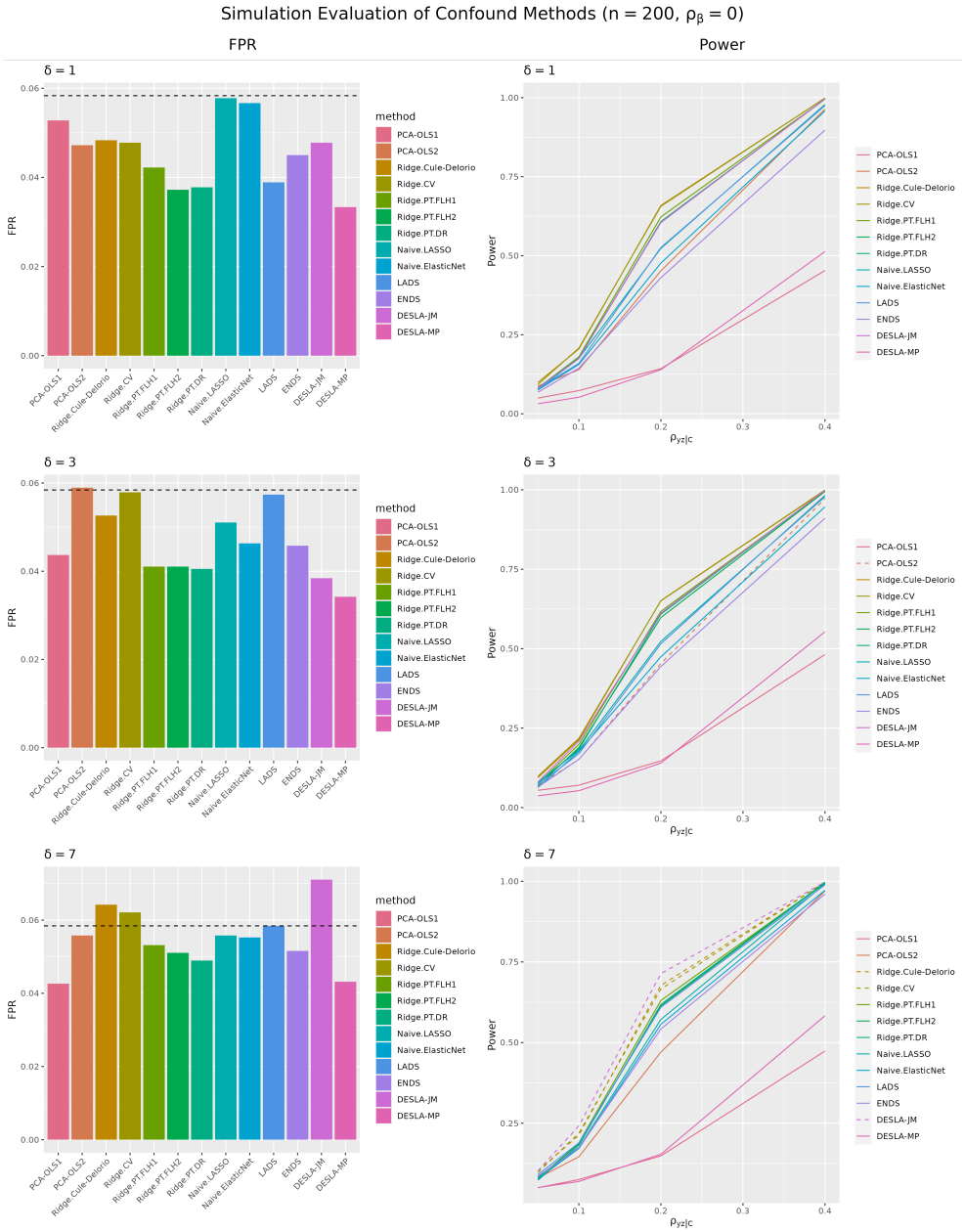


Figure D.3: Simulation results for $n = 200$, $\rho_\beta = 0$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

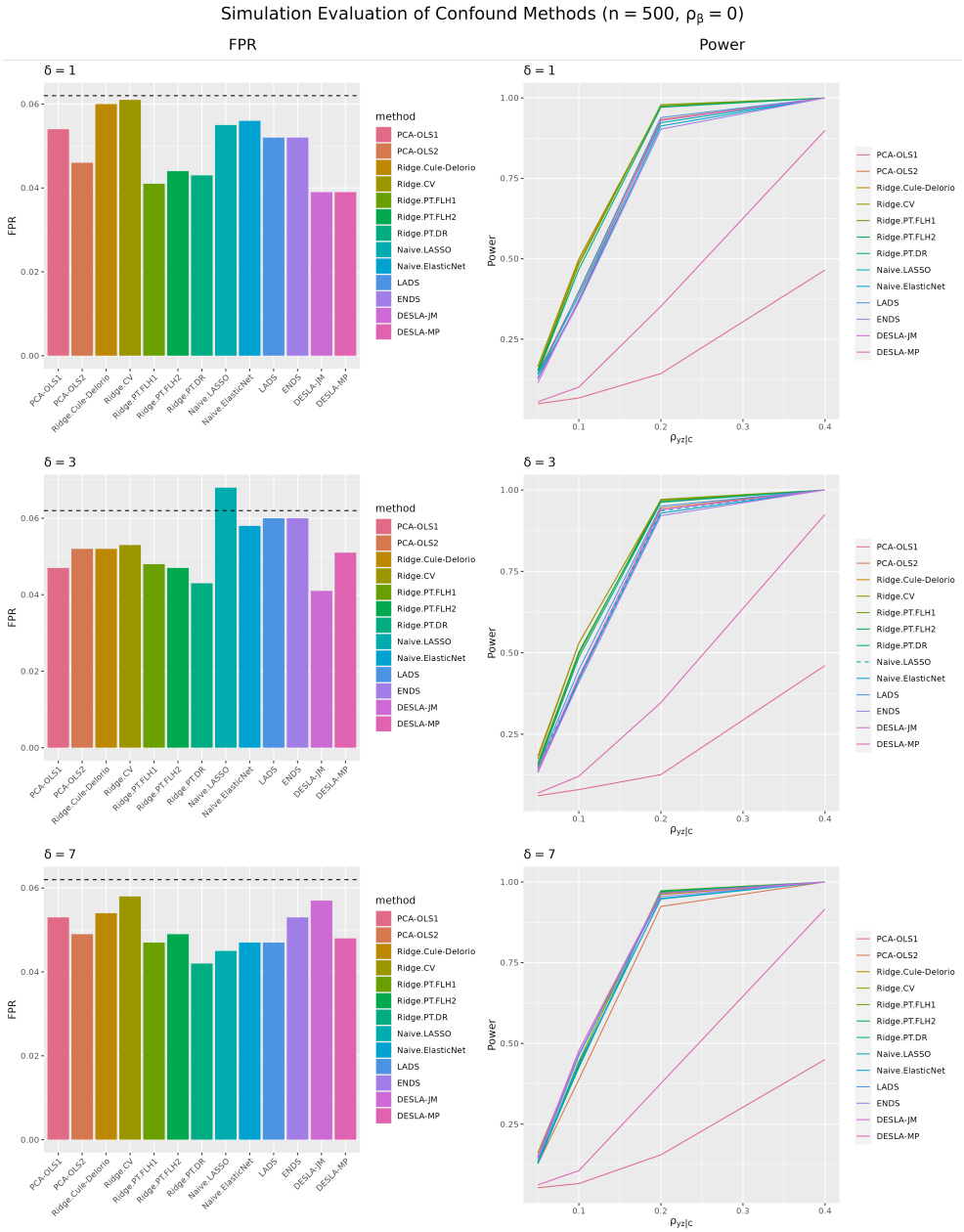


Figure D.4: Simulation results for $n = 500$, $\rho_\beta = 0$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

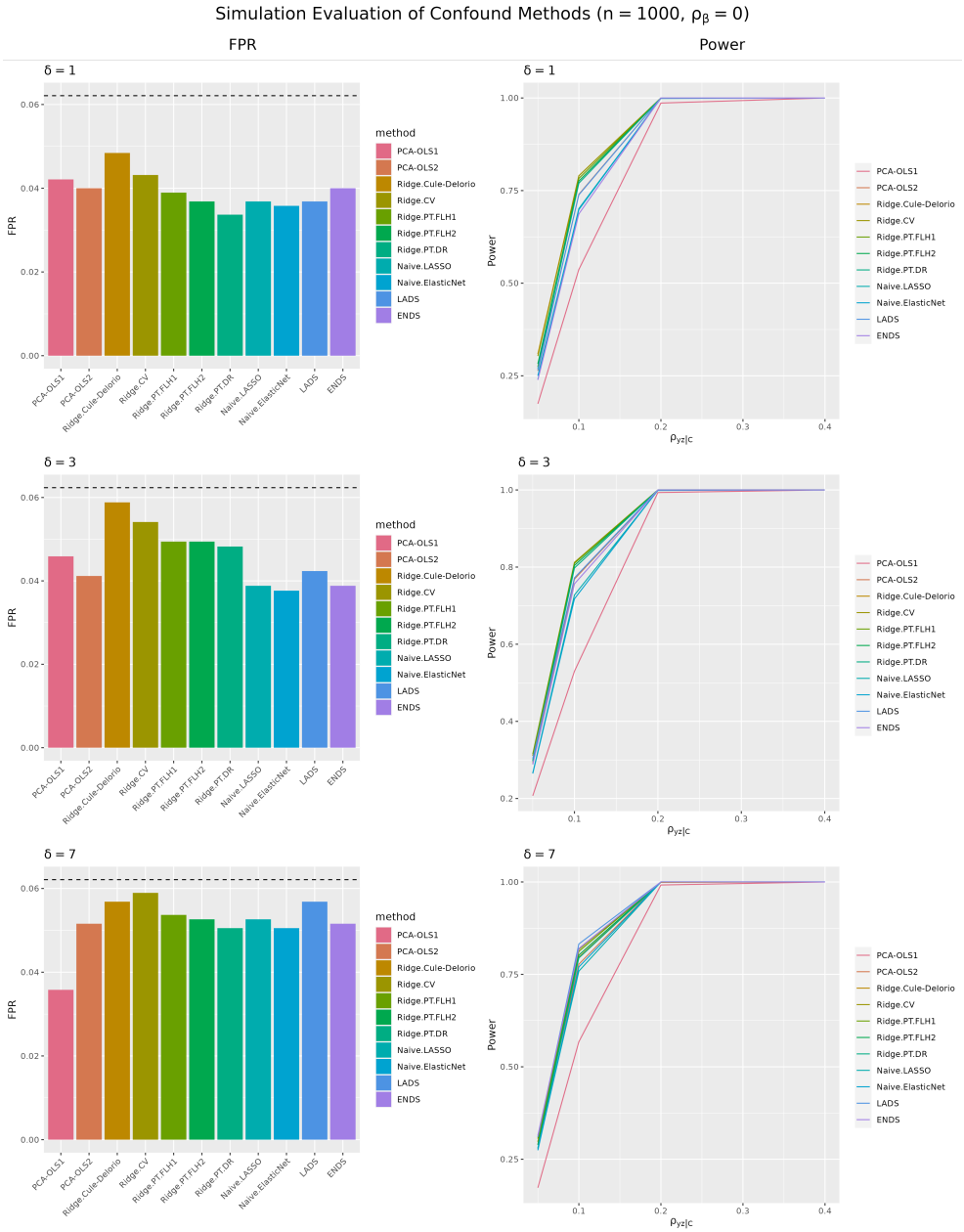


Figure D.5: Simulation results for $n = 1000$, $\rho_\beta = 0$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

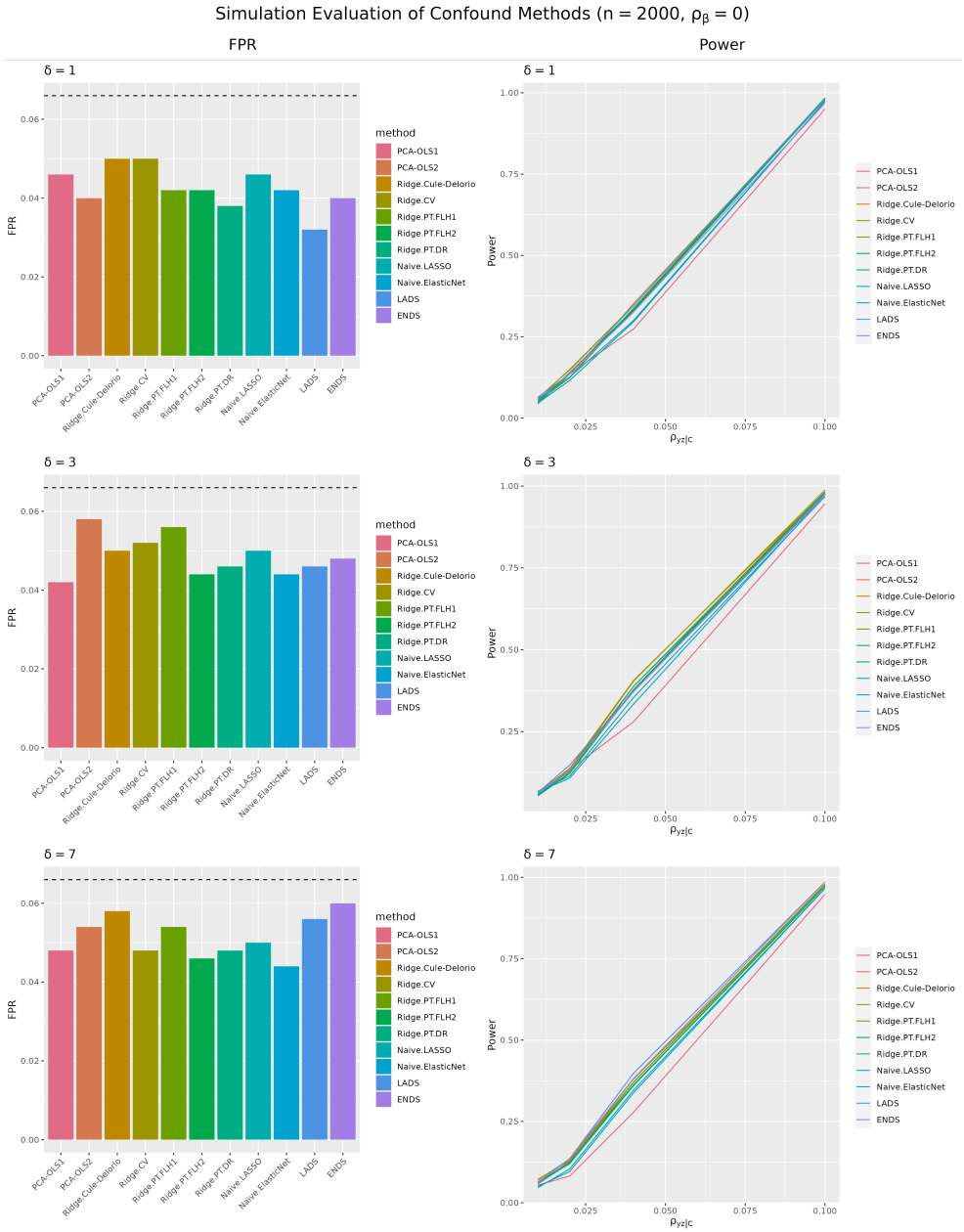


Figure D.6: Simulation results for $n = 2000$, $\rho_\beta = 0$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

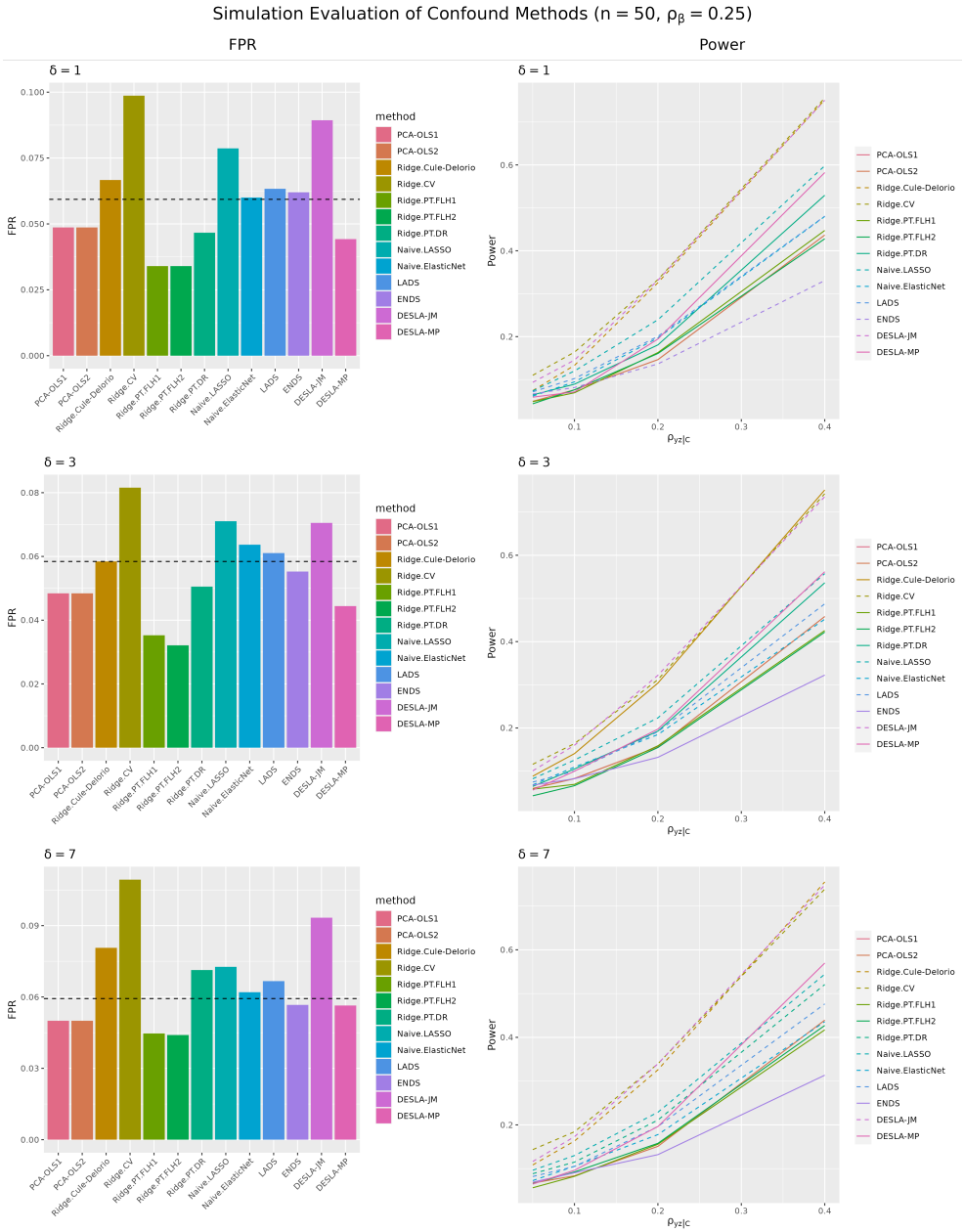


Figure D.7: Simulation results for $n = 50$, $\rho_\beta = 0.25$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

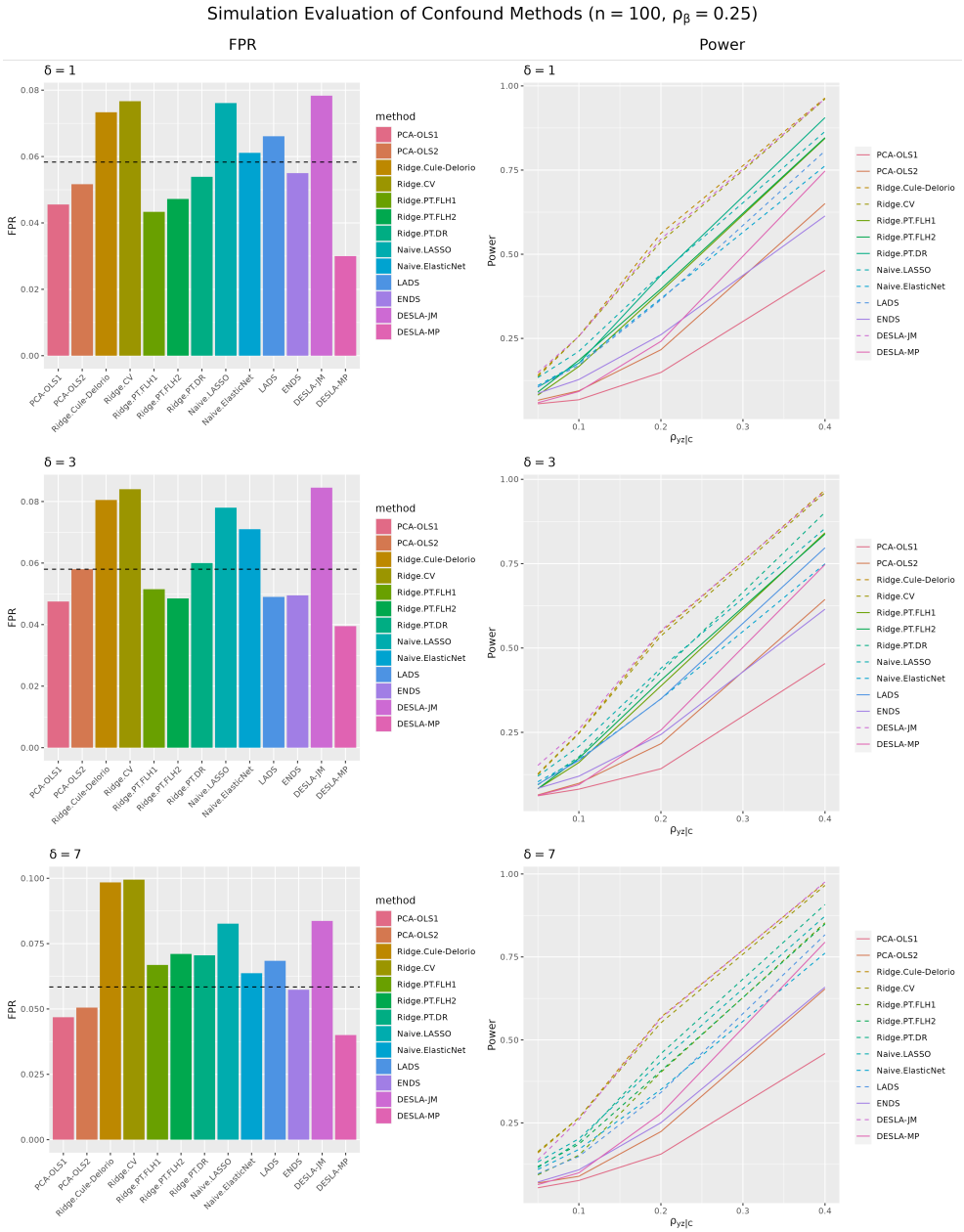


Figure D.8: Simulation results for $n = 100$, $\rho_\beta = 0.25$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

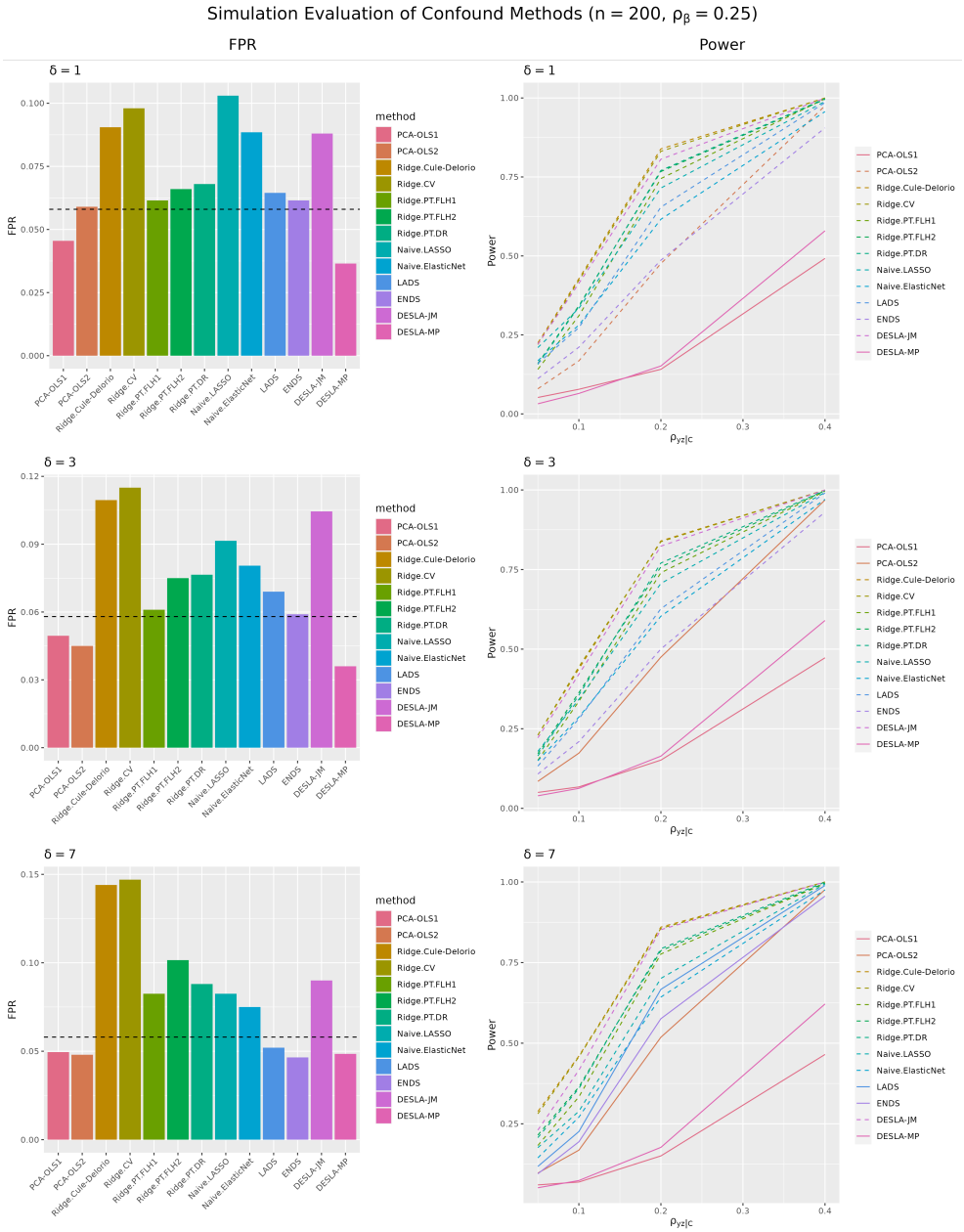


Figure D.9: Simulation results for $n = 200$, $\rho_\beta = 0.25$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

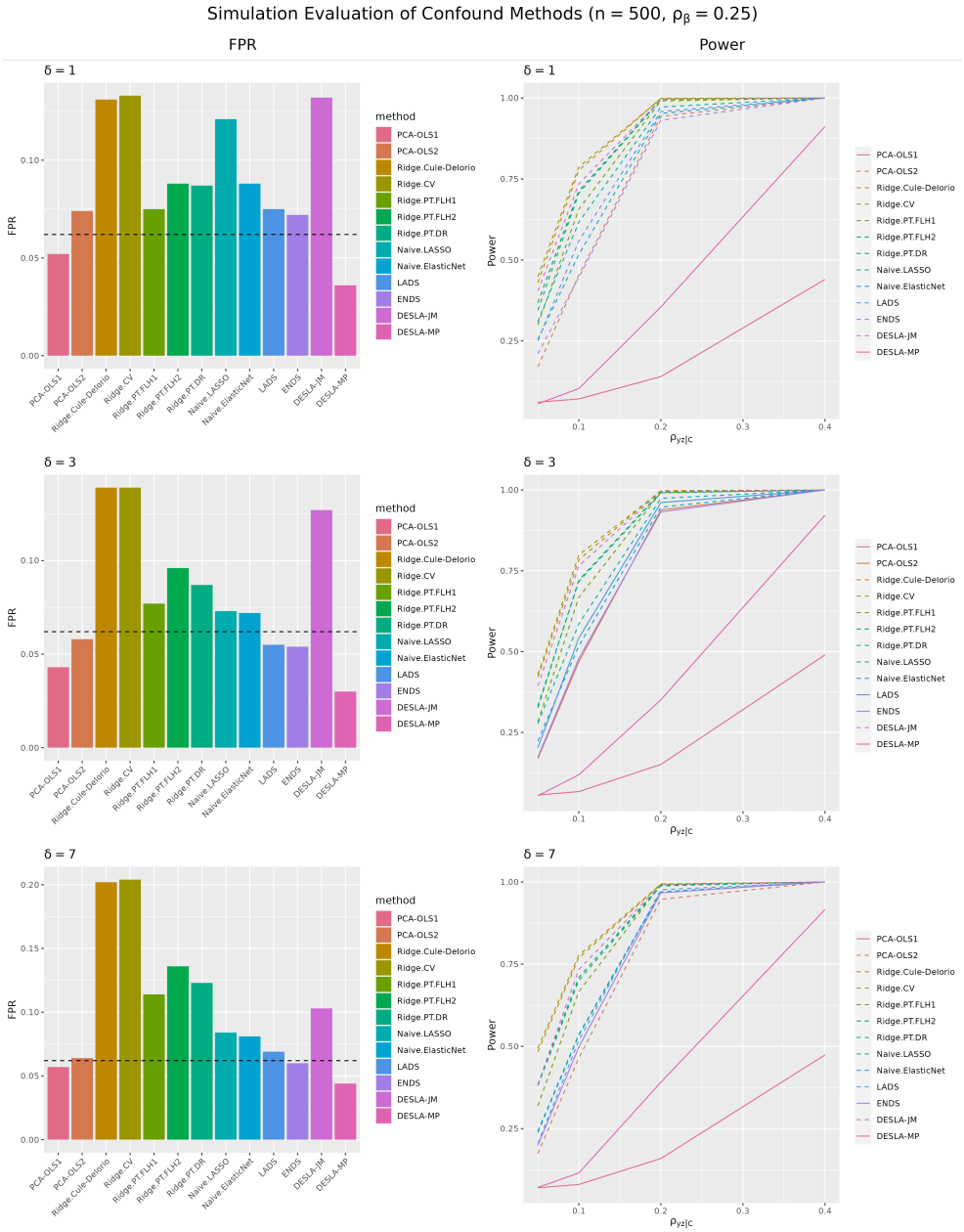


Figure D.10: Simulation results for $n = 500$, $\rho_\beta = 0.25$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

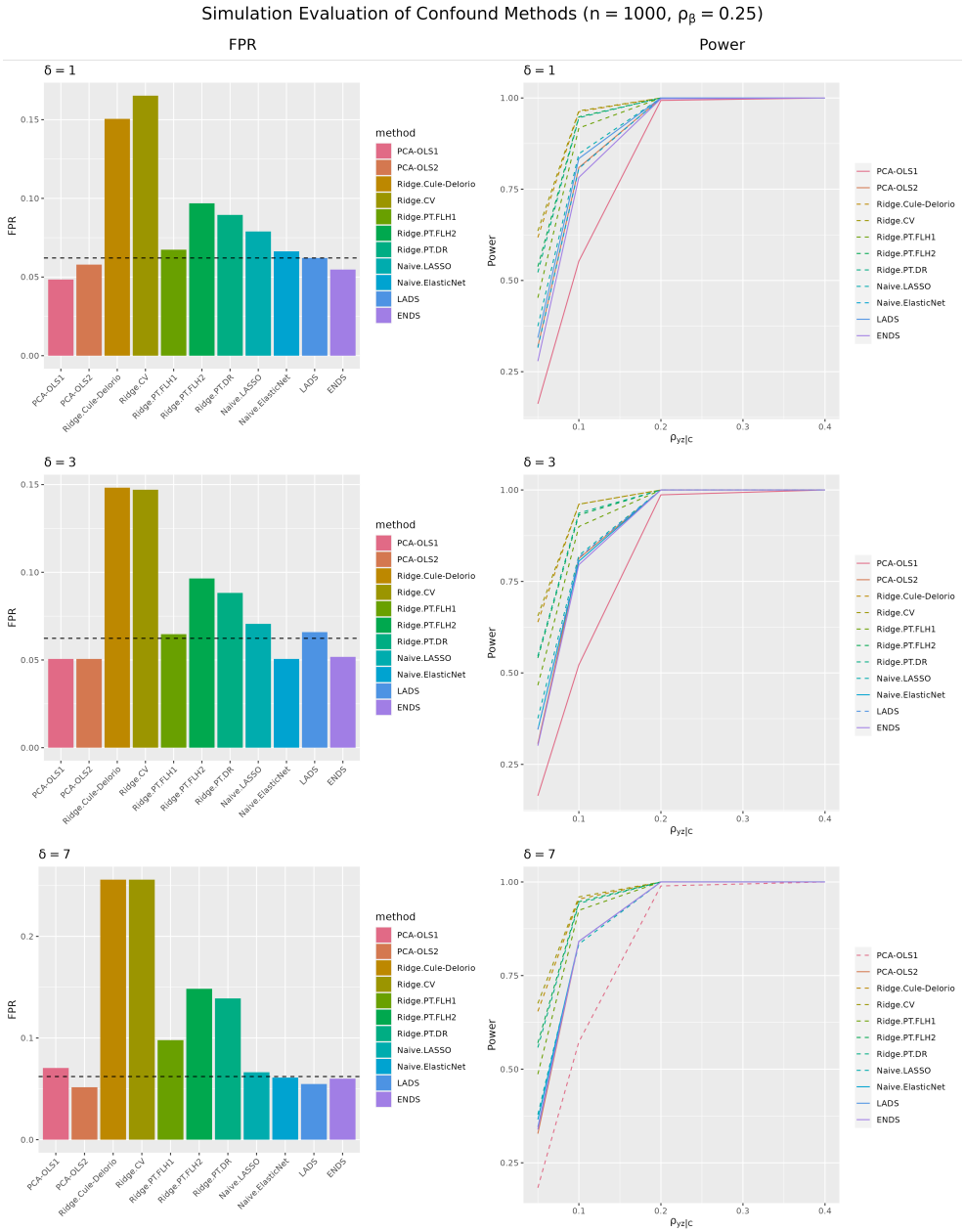


Figure D.11: Simulation results for $n = 1000$, $\rho_\beta = 0.25$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

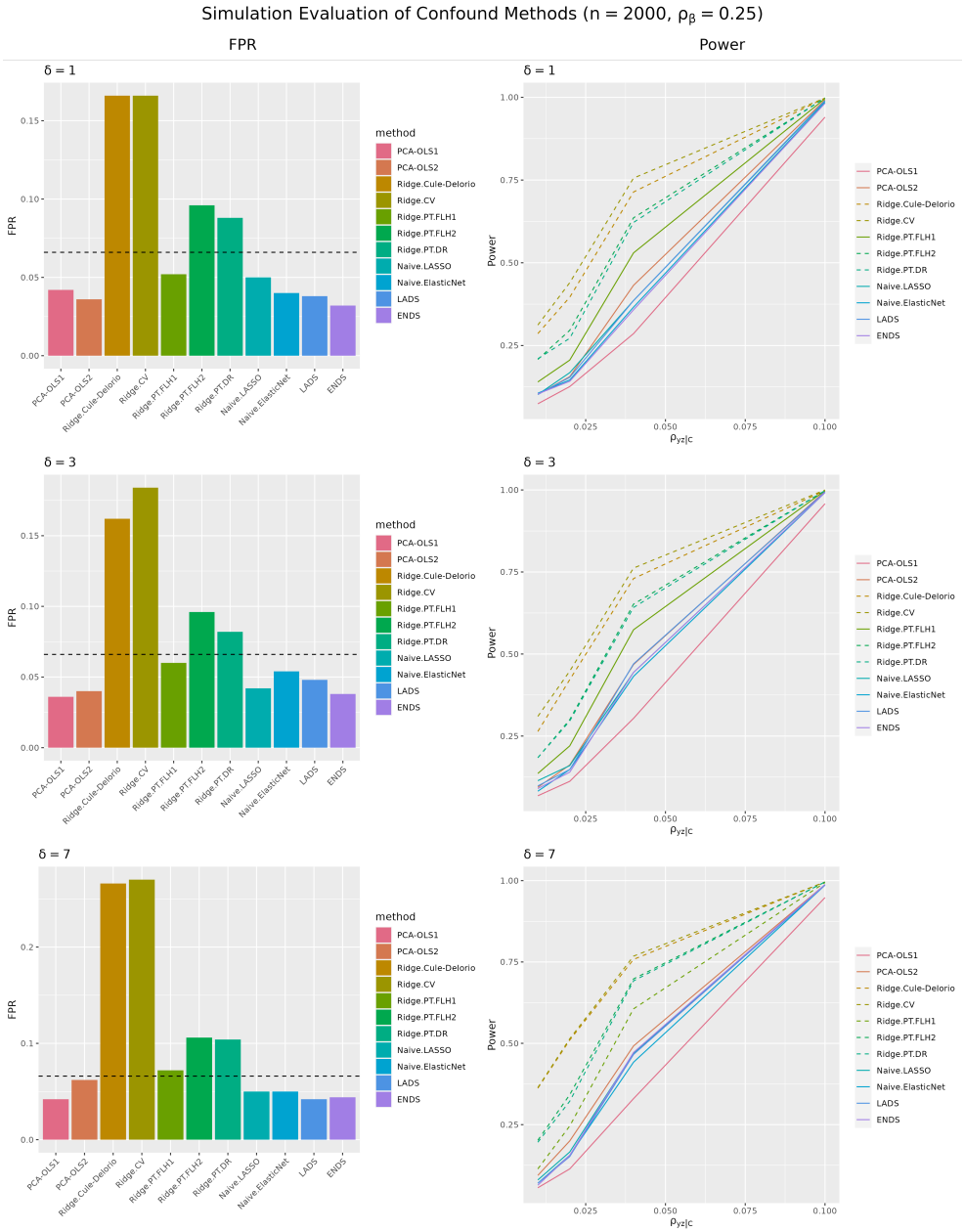


Figure D.12: Simulation results for $n = 2000$, $\rho_\beta = 0.25$. The bar plots to the left show the FPR and lineplots the statistical power of different methods. The dashed line in the bar plots is the 95% upper confidence limit for a FPR of 0.05. Methods above this limit have dashed lines in the power plot.

References

Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JL, Griffanti L, Douaud G, Sotiropoulos SN, Jbabdi S, Hernandez-Fernandez M, Vallee E, et al. (2018) Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage 166:400–424

Alfaro-Almagro F, McCarthy P, Afyouni S, Andersson JL, Bastiani M, Miller KL, Nichols TE, Smith SM (2021) Confound modelling in uk biobank brain imaging. NeuroImage 224:117002

Allan F, Wishart J (1930) A method of estimating the yield of a missing plot in field experimental work. The Journal of Agricultural Science 20(3):399–406

Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, Gallacher J, Green J, Matthews P, Pell J, et al. (2012) Uk biobank: Current status and what it means for epidemiology. Health Policy and Technology 1(3):123–126

Anderson TW (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the american Statistical Association 52(278):200–203

Azevedo T, Bethlehem RA, Whiteside DJ, Swaddiwudhipong N, Rowe JB, Lió P, Rittman T (2023) Identifying healthy individuals with alzheimer's disease neuroimaging phenotypes in the uk biobank. Communications medicine 3(1):100

Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research 20(1):40–49

Basto M, Abreu T, Gonçalves R, Pereira JM (2023) Confounding and effect modification in biostatistics: Na overview. UJAM 11(1):8–15

Bates D, Maechler M, Jagan M, et al. (2010) Matrix: Sparse and dense matrix classes and methods. R package version 0999375-43, URL http://cran r-project org/package= Matrix

Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR (2018) Characterizing and managing missing structured data in electronic health records: data analysis. JMIR medical informatics 6(1):e8960

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies 81(2):608–650

Benjamini Y, Gavrilov Y (2009) A simple forward selection procedure based on false discovery rate control. The Annals of Applied Statistics pp 179–198

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57(1):289–300

Bickel PJ, Hammel EA, O'Connell JW (1975) Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. Science 187(4175):398–404

Blanchet FG, Legendre P, Borcard D (2008) Forward selection of explanatory variables. Ecology 89(9):2623–2632

Boot T, Nibbering D (2017) Scalable simultaneous inference in high-dimensional linear regression models. arXiv preprint arXiv:170303282

Brenner H, Blettner M (1997) Controlling for continuous confounders in epidemiologic research. Epidemiology 8(4):429–434

Buck SF (1960) A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society: Series B (Methodological) 22(2):302–306

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'ŹConnell J, et al. (2017) Genome-wide genetic data on 500,000 uk biobank participants. BioRxiv p 166298

Celis-Morales CA, Welsh P, Lyall DM, Steell L, Petermann F, Anderson J, Iliodromiti S, Sillars A, Graham N, Mackay DF, et al. (2018) Associations of grip strength with cardiovascular, respiratory, and cancer outcomes and all cause mortality: prospective cohort study of half a million uk biobank participants. bmj 361

Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L (2011) China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. International journal of epidemiology 40(6):1652–1666

Chevalier JA (2020) Statistical control of sparse models in high dimension. PhD thesis, Université Paris-Saclay

Collins R (2012) What makes uk biobank special? The Lancet 379(9822):1173–1174

Cule E, De Iorio M (2013) Ridge regression in prediction problems: automatic choice of the ridge parameter. Genetic epidemiology 37(7):704–714

Cule E, Vineis P, De Iorio M (2011) Significance testing in ridge regression for genetic data. BMC bioinformatics 12:1–15

Dekkers IA, Jansen PR, Lamb HJ (2019) Obesity, brain volume, and white matter microstructure at mri: a cross-sectional uk biobank study. Radiology 291(3):763–771

Demirtas H, Doganay B (2012) Simultaneous generation of binary and normal data with specified marginal and association structures. Journal of Biopharmaceutical Statistics 22(2):223–236

Ding J, Tarokh V, Yang Y (2018) Model selection techniques: An overview. IEEE Signal Processing Magazine 35(6):16–34

Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, White T, Van Hees VT, Trenell MI, Owen CG, et al. (2017) Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. PloS one 12(2):e0169649

Douaud G, Lee S, Alfaro-Almagro F, Arthofer C, Wang C, McCarthy P, Lange F, Andersson JL, Griffanti L, Duff E, et al. (2022) Sars-cov-2 is associated with changes in brain structure in uk biobank. Nature 604(7907):697–707

Dryden IL, Koloydenko A, Zhou D (2009) Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics pp 1102–1123

Dutt RK, Hannon K, Easley TO, Griffis JC, Zhang W, Bijsterbosch JD (2022) Mental health in the uk biobank: A roadmap to self-report measures and neuroimaging correlates. Human Brain Mapping 43(2):816–832

Elliott LT, Sharp K, Alfaro-Almagro F, Douaud G, Miller K, Marchini J, Smith S (2017) The genetic basis of human brain structure and function: 1,262 genome-wide associations found from 3,144 gwas of multimodal brain imaging phenotypes from 9,707 uk biobank participants. BioRxiv p 178806

Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, Marchini J, Smith SM (2018) Genome-wide association studies of brain imaging phenotypes in uk biobank. Nature 562(7726):210–216

Elwert F, Winship C (2014) Endogenous selection bias: The problem of conditioning on a collider variable. Annual review of sociology 40(1):31–53

Enders CK, Peugh JL (2004) Using an em covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. Structural Equation Modeling 11(1):1–19

Ezekiel M (1930) Methods of correlation analysis.

Faber BG, Ebsim R, Saunders FR, Frysz M, Lindner C, Gregory JS, Aspden RM, Harvey NC, Davey Smith G, Cootes T, et al. (2022) A novel semi-automated classifier of hip osteoarthritis on dxa images shows expected relationships with clinical outcomes in uk biobank. Rheumatology 61(9):3586–3595

Farahibozorg SR, Harrison SJ, Bijsterbosch JD, Woolrich MW, Smith SM (2024) Multiscale modes of functional brain connectivity. bioRxiv pp 2024–05

Finney D (1947) The truncated binomial distribution. Annals of Eugenics 14(1):319–328

Fisher RA (1928) The general sampling distribution of the multiple correlation coefficient. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character 121(788):654–673

Foster PJ, Atan D, Khawaja A, Lotery A, MacGillivray T, Owen CG, Patel PJ, Petzold A, Rudnicka A, Sun Z, et al. (2023) Cohort profile: rationale and methods of uk biobank repeat imaging study eye measures to study dementia. BMJ open 13(6):e069258

Freedman D, Lane D (1983) A nonstochastic interpretation of reported significance levels. Journal of Business & Economic Statistics 1(4):292–298

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE (2017) Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. American journal of epidemiology 186(9):1026–1034

Ge R, Jin C, Netrapalli P, Sidford A, et al. (2016) Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In: International Conference on Machine Learning, PMLR, pp 2741–2750

Gentry AE, Kirkpatrick RM, Peterson RE, Webb BT (2023) Missingness adapted group informed clustered (magic)-lasso: a novel paradigm for phenotype prediction to improve power for genetic loci discovery. Frontiers in Genetics 14

Ghalebikesabi S, Cornish R, Holmes C, Kelly L (2021) Deep generative missingness pattern-set mixture models. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 3727–3735

Ghani IMM, Ahmad S (2010) Stepwise multiple regression method to forecast fish landing. Procedia-Social and Behavioral Sciences 8:549–554

Graham JW (2009) Missing data analysis: Making it work in the real world. Annual review of psychology 60:549–576

Greenacre M, Groenen PJ, Hastie T, dEnza AI, Markos A, Tuzhilina E (2022) Principal component analysis. Nature Reviews Methods Primers 2(1):100

Greenland S, Morgenstern H (2001) Confounding in health research. Annual review of public health 22(1):189–212

Hajat A, MacLehose RF, Rosofsky A, Walker KD, Clougherty JE (2021) Confounding by socioeconomic status in epidemiological studies of air pollution and health: challenges and opportunities. Environmental health perspectives 129(6):065001

Halawa A, El Bassiouni M (2000) Tests of regression coefficients under ridge regression models. Journal of Statistical Computation and Simulation 65(1-4):341–356

Halliday G (2017) Pathology and hippocampal atrophy in alzheimer's disease. The Lancet Neurology 16(11):862–864

Harris WS, Varvel SA, Pottala JV, Warnick GR, McConnell JP (2013) Comparative effects of an acute dose of fish oil on omega-3 fatty acid levels in red blood cells versus plasma: implications for clinical utility. Journal of clinical lipidology 7(5):433–440

Healy M, Westmacott M (1956) Missing values in experiments analysed on automatic computers. Journal of the Royal Statistical Society: Series C (Applied Statistics) 5(3):203–206

Hemerik J, Thoresen M, Finos L (2021) Permutation testing in high-dimensional linear models: an empirical investigation. Journal of Statistical Computation and Simulation 91(5):897–914

Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American journal of epidemiology 155(2):176–184

Higham NJ (2002) Computing the nearest correlation matrixâĂŤa problem from finance. IMA journal of Numerical Analysis 22(3):329–343

Hotelling H (1992) Relations between two sets of variates. In: Breakthroughs in statistics: methodology and distribution, Springer, pp 162–190

Jackson J, Mitra R, Hagenbuch N, McGough S, Harbron C (2023) A complete characterisation of structured missingness. arXiv preprint arXiv:230702650

James G, Witten D, Hastie T, Tibshirani R, Taylor J (2023) Linear model selection and regularization. In: An Introduction to Statistical Learning: with Applications in Python, Springer, pp 229–288

Jarrett D, Cebere BC, Liu T, Curth A, van der Schaar M (2022) Hyperimpute: Generalized iterative imputation with automatic model selection. In: International Conference on Machine Learning, PMLR, pp 9916–9937

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. The Journal of Machine Learning Research 15(1):2869–2909

Julkunen H, Cichońska A, Tiainen M, Koskela H, Nybo K, Mäkelä V, Nokso-Koivisto J, Kristiansson K, Perola M, Salomaa V, et al. (2023) Atlas of plasma nmr biomarkers for health and disease in 118,461 individuals from the uk biobank. Nature communications 14(1):604

Kettenring JR (1971) Canonical analysis of several sets of variables. Biometrika 58(3):433–451

Koerts J, Abrahamse APJ (1969) The correlation coefficient in the general linear model

Le Morvan M, Josse J, Moreau T, Scornet E, Varoquaux G (2020) Neumiss networks: differentiable programming for supervised learning with missing values. Advances in Neural Information Processing Systems 33:5980–5990

Li L, Hui S, Pennello G, Desta Z, Todd S, Nguyen A, Flockhart D (2007) Estimating a positive false discovery rate for variable selection in pharmacogenetic studies. Journal of Biopharmaceutical Statistics 17(5):883–902

Li Y, Sahakian BJ, Kang J, Langley C, Zhang W, Xie C, Xiang S, Yu J, Cheng W, Feng J (2022) The brain structure and genetic mechanisms underlying the nonlinear association between sleep duration, cognition and mental health. Nature Aging 2(5):425–437

Lin Z (2019) Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. SIAM Journal on Matrix Analysis and Applications 40(4):1353–1370

Lindsey C, Sheather S (2010) Variable selection in linear regression. The Stata Journal 10(4):650–669

Little TD, Jorgensen TD, Lang KM, Moore EWG (2014) On the joys of missing data. Journal of pediatric psychology 39(2):151–162

Littlejohns TJ, Holliday J, Gibson LM, Garratt S, Oesingmann N, Alfaro-Almagro F, Bell JD, Boultwood C, Collins R, Conroy MC, et al. (2020) The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature communications 11(1):2624

Liu X, Morelli D, Littlejohns TJ, Clifton DA, Clifton L (2023) Combining machine learning with cox models to identify predictors for incident post-menopausal breast cancer in the uk biobank. Scientific Reports 13(1):9221

van Loon W, Fokkema M, de Rooij M (2022) Imputation of missing values in multi-view data. arXiv preprint arXiv:221014484

Lounici K (2014) High-dimensional covariance matrix estimation with missing observations

Madakkatel I, Zhou A, McDonnell M, Hyppönen E (2021) Can we use machine learning to discover risk factors? testing the proof of principle using data on 11,000 predictors and mortality in the uk biobank. medRxiv pp 2021–05

Marini MM, Olsen AR, Rubin DB (1980) Maximum-likelihood estimation in panel studies with missing data. Sociological methodology 11:314–357

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. The Journal of Machine Learning Research 11:2287–2322

McNamee R (2003) Confounding and confounders. Occupational and environmental medicine 60(3):227–234

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JL, et al. (2016) Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience 19(11):1523–1536

Milte CM, Coates AM, Buckley JD, Hill AM, Howe PR (2008) Dose-dependent effects of docosahexaenoic acid-rich fish oil on erythrocyte docosahexaenoic acid and blood lipid levels. British Journal of Nutrition 99(5):1083–1088

Mitra R, McGough SF, Chakraborti T, Holmes C, Copping R, Hagenbuch N, Biedermann S, Noonan J, Lehmann B, Shenvi A, et al. (2023) Learning from data with structured missingness. Nature Machine Intelligence 5(1):13–23

Molenberghs G, Beunckens C, Sotto C, Kenward MG (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society Series B: Statistical Methodology 70(2):371–388

Morabia A (2011) History of the modern epidemiological concept of confounding. Journal of Epidemiology & Community Health 65(4):297–300

Navaneetham K, Arunachalam D (2025) Global population aging, 1950–2050. In: Handbook of aging, health and public policy: perspectives from Asia, Springer, pp 99–116

Okpara C, Edokwe C, Ioannidis G, Papaioannou A, Adachi JD, Thabane L (2022) The reporting and handling of missing data in longitudinal studies of older adults is suboptimal: a methodological survey of geriatric journals. BMC medical research methodology 22(1):122

Olkin I, Finn JD (1995) Correlations redux. Psychological Bulletin 118(1):155

Olkin I, Pratt JW (1958) Unbiased estimation of certain correlation coefficients. The annals of mathematical statistics pp 201–211

Papageorgiou SN (2019) Confounders in epidemiological associations. Journal of orthodontics 46(3):267–269

Park S, Lim J (2019) Non-asymptotic rate for high-dimensional covariance estimation with non-independent missing observations. Statistics & Probability Letters 153:113–123

Parra LC (2018) Multi-set canonical correlation analysis simply explained. arXiv preprint arXiv:180203759

Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I (2017) Missing data and multiple imputation in clinical epidemiological research. Clinical epidemiology pp 157–166

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Pervaiz U, Vidaurre D, Woolrich MW, Smith SM (2020) Optimising network modelling methods for fmri. NeuroImage 211:116604

Peterson RA, Cavanaugh JE (2019) Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. Journal of Applied Statistics 47(13-15):2312–2327

Radosavljević L, Smith SM, Nichols TE (2024) A generative model for evaluating missing data methods in large epidemiological cohorts. medRxiv pp 2024–04

Robins JM, Morgenstern H (1987) The foundations of confounding in epidemiology. Computers & Mathematics with Applications 14(9-12):869–916

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley Series in Probability and Statistics, Wiley

Schafer JL (1997) Analysis of incomplete multivariate data. CRC press

Smith SM, Nichols TE, Vidaurre D, Winkler AM, Behrens TE, Glasser MF, Ugurbil K, Barch DM, Van Essen DC, Miller KL (2015) A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nature neuroscience 18(11):1565–1567

Sorensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biol Skar 5:1–34

Stark CR, Mantel N (1966) Effects of maternal age and birth order on the risk of mongolism and leukemia. Journal of the National Cancer Institute 37(5):687–698

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. (2015) Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12(3):e1001779

Thomson R, McWhirter R (2017) Adjusting for familial relatedness in the analysis of gwas data. Bioinformatics: Volume II: Structure, Function, and Applications pp 175–190

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58(1):267–288

Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R (2016) Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association 111(514):600–620

Topiwala A, Allan CL, Valkanova V, Zsoldos E, Filippini N, Sexton C, Mahmood A, Fooks P, Singh-Manoux A, Mackay CE, et al. (2017) Moderate alcohol consumption as risk factor for adverse brain outcomes and cognitive decline: longitudinal cohort study. bmj 357

Topiwala A, Ebmeier KP, Maullin-Sapey T, Nichols TE (2021) No safe level of alcohol consumption for brain health: observational cohort study of 25,378 uk biobank participants. MedRxiv pp 2021–05

Topiwala A, Nichols TE, Williams LZ, Robinson EC, Alfaro-Almagro F, Taschler B, Wang C, Nelson CP, Miller KL, Codd V, et al. (2023) Telomere length and brain imaging phenotypes in uk biobank. PLoS One 18(3):e0282363

UK Biobank (2020) Uk biobank food preferences web questionnaire (version 1.1). Tech. rep., UK Biobank, URL `https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/foodpref.pdf`

UK Biobank (2024) UK Biobank. `https://www.ukbiobank.ac.uk`, accessed: 2024-08-21

UK Biobank (2025) Uk biobank: Health research data for the world. `https://www.ukbiobank.ac.uk/`, accessed: December 19, 2025

UK Biobank (2026) Major achievements. URL `https://www.ukbiobank.ac.uk/discoveries-and-impact/major-achievements/`, accessed: 2026-07-02

Van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in r. Journal of statistical software 45:1–67

Van Buuren S, Oudshoorn K (1999) Flexible multivariate imputation by MICE. Leiden: TNO

Wang HT, Smallwood J, Mourao-Miranda J, Xia CH, Satterthwaite TD, Bassett DS, Bzdok D (2020) Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. NeuroImage 216:116745

Westfall J, Yarkoni T (2016) Statistically controlling for confounding constructs is harder than you think. PloS one 11(3):e0152719

Whiteley N, Gray A, Rubin-Delanchy P (2022) Statistical exploration of the manifold hypothesis

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014) Permutation inference for the general linear model. Neuroimage 92:381–397

Winkler AM, Renaud O, Smith SM, Nichols TE (2020) Permutation inference for canonical correlation analysis. Neuroimage 220:117065

Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TM, et al. (2018) Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature genetics 50(5):668–681

Xu C, Tao D, Xu C (2015) Multi-view learning with incomplete views. IEEE Transactions on Image Processing 24(12):5812–5825

Yates F (1933) The analysis of replicated experiments when the field results are incomplete.

Zhang L, Zhao Y, Zhu Z, Shen D, Ji S (2018) Multi-view missing data completion. IEEE Transactions on Knowledge and Data Engineering 30(7):1296–1309

Zhang Y, Chen SD, Deng YT, You J, He XY, Wu XR, Wu BS, Yang L, Zhang YR, Kuo K, et al. (2023) Identifying modifiable factors and their joint effect on dementia risk in the uk biobank. Nature human behaviour 7(7):1185–1195

Zhao S, Witten D, Shojaie A (2021) In defense of the indefensible: A very naive approach to high-dimensional inference. Statistical science: a review journal of the Institute of Mathematical Statistics 36(4):562

Zhuang X, Yang Z, Sreenivasan KR, Mishra VR, Curran T, Nandy R, Cordes D (2019) Multivariate group-level analysis for task fmri data with canonical correlation analysis. NeuroImage 194:25–41

Zhuang X, Yang Z, Cordes D (2020) A technical review of canonical correlation analysis for neuroscience applications. Human brain mapping 41(13):3807–3833