PARTIAL LEAST SQUARES ON DATA WITH MISSING COVARIATES: A COMPARISON APPROACH

by

Dana L. Tudorascu

Diploma de licenta, Universitatea din Craiova, Romania, 1999

M.S. in Computational Mathematics, Duquesne University, 2003

Submitted to the Graduate Faculty of

Graduate School of Public Health

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

This dissertation was presented

by

Dana L. Tudorascu

It was defended on

June 4, 2009

and approved by

Lisa A. Weissfeld, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Julie C. Price, PhD
Associate Professor
School of Medicine
Department of Radiology
University of Pittsburgh

Stewart J. Anderson, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Lan Kong,PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Dissertation Director:
Lisa A. Weissfeld, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

PARTIAL LEAST SQUARES ON DATA WITH MISSING COVARIATES: A COMPARISON APPROACH

Dana L. Tudorascu, PhD

University of Pittsburgh, 2009

The correlation between any two random variables can be estimated using a variety of techniques including parametric methods based on the Pearson correlation coefficient, nonparametric methods, and regression analysis. While these estimators have been widely used, the computation of these estimates in the presence of missing data has not been as widely studied. There has been some work addressing the estimation of parameters in the presence of missing data for regression analysis; including imputation, inverse probability weighted regression and weighted estimating equations. However, there has been little work focused on the estimation of the correlation coefficient. To assess the usefulness of these methods in a practical setting, we present simulation studies comparing imputation, inverse probability weighting and complete cases and provide recommendations on the basis of these results. Furthermore, computation of Partial Least Squares (PLS) scores with the correlation matrix computed using the above mentioned techniques are also presented. We apply these results in a positron emission tomography data set consisting of several different brain regions as response variables and cognitive tasks as covariates of interest. Alzheimer's disease is a progressive and fatal health disease. The application presented in this work is significant for public health since it provides us with a better understanding of variability in different brain regions as it relates to neuropsychological tests that are helpful in diagnosis of progressive brain disease (i.e Alzheimer's disease).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my advisor, Dr. Lisa Weissfeld for all her support and guidance that she has provided for me. Dr. Lisa Weissfeld has shown me the path to independent research and guided me throughout the years. I was very fortunate to have such an incredible advisor. I would have never come so far without Dr. Weissfeld's help.

I would also like to acknowledge my committee members, Dr. Julie Price, Dr. Lan Kong and Dr. Stewart Anderson for all their suggestions and their time. Dr. Lan Kong has provided great help with her simulations/summary suggestions. Dr. Julie Price has provided me with very useful information regarding the data example presented in this dissertation. Furthermore, over the years I have spent with the PET group, Dr. Price has always been very helpful and understanding.

I have to mention the incredible support of my friends Gina, Bedda, Kira, Amy and all my romanian friends from Pittsburgh during all these years. They have always been there for me. I am very grateful to my husband, Adi. He has encouraged me and supported me during all these years. My son, Andrei has motivated me and gave me hope since the day he was born.

I am very grateful to the Physics and Computing group I worked with at Emory University, especially Dr. John Votaw that taught me useful things about PET imaging.

Last but not least, I would like to thank my parents, Mioara and Benone for all their love and support. They taught me to be ambitious, strong and always be persistent in accomplishing my goals.

1.0 INTRODUCTION

Missing values in a dataset, an issue that statisticians often have to overcome, is the main motivation of our research. Statistical inference from data with missing covariates has been a heavily researched topic over the last two decades. Since most statistical methods were derived for fully observed data, the impact of missing values is an issue. Missing values can occur on independent variables (predictors) and on dependent variables (outcomes). The goal of a statistician still remains the same with or without missing values, namely, to draw valid and efficient inferences about the population of interest. This work is an attempt to address the problem of missing covariates when an estimate of the covariance matrix is desired. Furthermore, the estimated covariance matrix will further be employed in the computation of partial least squares scores. Partial least squares (PLS) is a technique that operates on the covariance among two or more blocks of variables and uses this information to obtain a new set of variables, called scores, that relate the blocks using fewer dimensions.

In 1976 Rubin and Little [20] developed a terminology for different missing values processes. The three missing data mechanisms defined by Little and Rubin are: MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random). MCAR assumes that the probability that an observation z_i is missing is not related to the value of z_i or to the value any other variables in the study. MAR assumes that the probability that an observation z_i is missing is not related to the value of z_i but could be related to the value of other variables in the study. MNAR assumes that the missingness could be related to the value as well as to the other variables. Throughout our work the missing at random mechanism will be assumed.

1.1 SUMMARY

The layout of this dissertation will be as follows. Chapter 2, entitled Background, will provide a literature review with respect to handling data with missing covariates along with important definitions that will be use throughout our work. Partial Least Squares is also introduced as well in this chapter and discussed in detail. Chapter 3, entitled "Covariance matrix" will present the methodology that was used for estimation of the covariance matrix in the presence of missing covariates, along with simulation studies, a positron emission tomography example and a short discussion. Chapter 4, entitled "Partial Least Squares", will present the methodology developed for computing PLS scores in the presence of missing data. Simulation studies will be presented, along with a data example followed by a short discussion section. Chapter 5 will conclude this work with a discussion section and further recommendations.

2.0 BACKGROUND

Positron Emission Tomography (PET) measures emissions from radioactively labelled metabolically active chemicals that have been injected into the bloodstream. The emission data are computer-processed to produce 2- or 3-dimensional images of the distribution of the chemicals throughout the brain [16]. The labelled compound, called a radiotracer (molecule labelled with a positron emitting isotope), is injected into the bloodstream and accumulates in various regions throughout the body/brain. Sensors in the PET scanner detect the radioactivity as the compound accumulates in various regions of the brain.

There are four important radionuclides that are used in PET to label compounds: Carbon (^{11}C), Nitrogen (^{13}N), Oxygen (^{15}O) and Fluorine (^{18}F). They emit radiation that will pass through the body and can be detected externally. The amount of tracer is small and it does not interfere or influence in any way the activity of the compound. Based on what they measure, PET tracers can be divided into three broad classes: metabolic, blood flow and receptors and transporters. The first category includes tracers that provide metabolic data such as glucose uptake and protein synthesis obtained from tracers $18-fluorine$ (^{18}F), $Carbon-11$ (^{11}C). These biomolecules leave the bloodstream and enter cells. The second category of compounds measure physiological activities (e.g., blood flow) and uses the tracer Oxygen (^{15}O). The compound remains in the bloodstream over the established study duration. The third broad category includes compounds that quantify molecular target (receptors and transporters) and a tracer that could be used with these is $Carbon-11$ (^{11}C).

A mathematical model (kinetic model) is used to describe the kinetics of the tracer during the biological process. The final result is a three-dimensional image of the anatomical distribution of the biological process under the investigation. A sequence of images of tracer activity distribution is recorded for a specified amount of time. Using a kinetic model, useful

information regarding the tracer uptake during the time period of interest is extracted at the end of the process. Predefined regions of interest are placed over the organ tissue under investigation on a specified number of slices (images), so that the activity in that region (tracer concentration) can be tracked over time. Data obtained through the regions of interested process of the images will produce the time activity curves for the tissue under investigation.

PET scanning is used for diagnosis of brain disease, most notably because brain tumors, strokes, and neuron-damaging diseases which cause dementia (such as Alzheimer's disease) all cause changes in brain metabolism, which in turn results in easily detectable changes in PET scans. In PET imaging, $[^{18}F]$-2-deoxy-2-fluoro-D-glucose (FDG) can be used for the assessment of glucose metabolism in the heart, lungs, and the brain. The fluorine in the FDG molecule is chosen to be the positron-emitting radioactive isotope $fluorine - 18$, to produce $[^{18}F]$-FDG. The isotope $fluorine - 18$ is clinically attached to the chemical compound. The tracer reflects metabolic activity and most of the FDG uptake occurs within 30 minutes after it has been injected. As a glucose analog, FDG is taken up by high-glucose-using cells such as brain, kidney, and cancer cells, where phosphorylation (the addition of a phosphate (PO_4) group to a protein or other organic molecule) prevents the glucose from being released intact [17]. This tracer is also widely used for imaging different types of tumors. As a result FDG-PET can be used for the diagnosis, staging, and monitoring of cancer treatment. Some type of diseases for which PET-FDG is used include: Hodgkin's disease, non-Hodgkin's lymphoma, colorectal cancer, breast cancer, melanoma, and lung cancer as well as in diagnosing Alzheimer's disease.

In the present study FDG is used as a tracer for the exchange of glucose between plasma and the brain. Positron Emission Tomography with $[^{18}F]$ flouorodeoxyglucose (PET-FDG) is thought to aid diagnosis of dementing disorders. Recognizing Alzheimer's disease is a promising application for FDG-PET because of the sharp contrast in its pattern of glucose hypometabolism [14].

It is well known in the neuroimaging community that the data that comes from imaging studies is highly variable. This variability can make it very difficult to analyze data. In order to measure the variation, we want to know how much two variables are changing together

or in opposition to each other. As a visualization of this concept we could refer to geometry, thinking about each variable as a vector with a focus on the length and direction of these vectors. Geometrically, the correlation of two variables is defined as being the cosine of the angle formed by the two vectors. A perfect correlation of two variables will be either equal to 1, a perfect positive correlation, or -1, a perfectly negative association.

The covariance matrix is a very useful tool in many different areas forming the basis for a large number of statistical techniques. In this setting it is extensively used since data are generally continuous and highly variable. When it comes to reducing dimensionality and explaining variability in large data sets, the covariance matrix plays an extremely important role since it forms the basis for many dimension reduction procedures.

Principal components analysis (PCA) is a technique used to describe the variation in a set of correlated variables, $x_1, x_2, ..., x_q$, in terms of a new set of uncorrelated variables $y_1, y_2, ..., y_q$ each of which is a linear combination of x variables. Principal components analysis is interpreted in terms of the correlations or covariances between the original variables and derived components. The derived components of the PCA's are calculated from the covariance matrix. Factor analysis is another multivariate method, mostly used in studies in the social sciences, that uses the covariance matrix. Correlations/covariances of the manifest (measurable) variables are central to factor analysis.

Partial Least Squares, first introduced in 1975 by Herman Wold, [25] is another method that is used in the multivariate setting for which the covariance matrix plays a very important role. Partial Least Squares has been used to extract new information from imaging data that is not accessible through other univariate and multivariate image analysis tools. Brain images are very rich data sets containing a tremendous number of temporal, spatial and statistical signals. The general idea of PLS is to extract the latent factors accounting for as much of the factor variation as possible while modelling the responses well. The first step of PLS is the singular value decomposition (SVD) of the cross-block covariance matrix Σ_{YX}. (Our data contain missing values for some of the covariates and therefore the covariance matrix based only on available cases will not make use of all of the data. Some of the methods that make best use of all of the data are being investigated here.) Incorporating missing values into the modelling process requires knowledge of the nature of the missing data mechanisms

as well as its implications for the statistical inference.

2.1 DEFINITIONS

Let $Y_{n\times q}$ be a vector of response (or dependent) variables and $X_{n\times p}$ be a vector of independent variables or covariates, completely observed and Z_i be a covariate for which some of the observations are missing. Let R_i, the indicator for missing status, be defined as:

$$R_i = \begin{cases} 1, & \text{if } z_i = observed, \\ 0, & \text{if } z_i = missing. \end{cases} \tag{2.1}$$

Throughout this work a missing at random mechanism (MAR) is assumed. We denote the data as (X, Y, Z), where X, Y is the complete part of the data and Z is the part containing missing values. Rubin [20] defined missing data to be MAR if the distribution of missingness does not depend on Z, that is, if

$$P(R|Z) = P(R|X, Y).$$

The basis of estimation discussed here comes from a multivariate normal distribution which is defined as:

Definition 1. *A random vector $X \in \mathbb{R}^{p\times 1}$ has a multivariate normal distribution with a nonsingular covariance matrix Σ if $\Sigma \in \mathbb{R}^{p\times p}$ is a positive definite matrix and the probability density function of X is:*

$$f(x) = [const.]|\Sigma|^{-1/2} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)), \tag{2.2}$$

where $\mu \in \mathbb{R}^{p\times 1}$ is the expected value.

The parameters we shall estimate are (μ, Σ), the population mean and population covariance, respectively. The likelihood function of the multivariate normal distribution is given by the following expression:

$$L(\mu, \Sigma) = [const.] \prod_{i=1}^{n} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)), \tag{2.3}$$

or equivalently:

$$L(\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1}(x_i - \mu)). \tag{2.4}$$

The unbiased parameter estimates for (μ, Σ) are given by the sample mean $\bar{x}$ and the sample covariance matrix S:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{2.5}$$

respectively,

$$S = \frac{1}{n-1} \sum (x_i - \bar{x})(x_i - \bar{x})^T. \tag{2.6}$$

Definition 2. *The population covariance of two random variables X and Y with $E(X) = \mu_X$ and $E(Y) = \mu_Y$ is defined by:*

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)),$$

where E is the expected value.

Let us consider the *bivariate* normal as a special case of the multivariate normal distribution. Let (X, Y) have a bivariate normal distribution with density function given by:

$$f(x, y) = \frac{1}{2\sigma_x \sigma_y (1 - \rho^2)^{1/2}} exp\{-\frac{1}{2(1 - \rho^2)}$$
$$(((\frac{x - \mu_x}{\sigma_x})^2 + (\frac{y - \mu_y}{\sigma_y})^2 - 2\rho(\frac{x - \mu_x}{\sigma_x})(\frac{y - \mu_y}{\sigma_y}))\}$$

and mean vector:

$$E \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}. \tag{2.7}$$

The covariance matrix is written as:

$$\Sigma = E\begin{bmatrix} (X-\mu_X)^2 & (X-\mu_X)(Y-\mu_Y) \\ (X-\mu_X)(Y-\mu_Y) & (Y-\mu_Y)^2 \end{bmatrix}. \tag{2.8}$$

Therefore, the estimate of the Σ matrix will be:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{XX} & \hat{\sigma}_{XY} \\ \hat{\sigma}_{YX} & \hat{\sigma}_{YY} \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_X^2 & \hat{\rho}\hat{\sigma}_X\hat{\sigma}_Y \\ \hat{\rho}\hat{\sigma}_Y\hat{\sigma}_X & \hat{\sigma}_Y^2 \end{bmatrix}, \tag{2.9}$$

where

$$\hat{\sigma}_{XX} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \hat{Var}(X) = \hat{\sigma}_X^2$$

$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \hat{\sigma}_{YX} = \hat{\rho}\hat{\sigma}_X\hat{\sigma}_Y$$

and

$$\hat{\sigma}_{YY} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \hat{Var}(Y) = \hat{\sigma}_Y^2.$$

Definition 3. *The correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:*

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}.$$

The Pearson correlation coefficient, also known as the "sample correlation coefficient" is the best estimate of the correlation of X and Y when both, X and Y are normally distributed.

Definition 4. *The Pearson correlation coefficient r, is given by*

$$r = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where $\bar{x}$, $\bar{y}$ are the sample means of X and Y.

An equivalent formula for r that illustrates its mathematical relationship to the least squares estimate of the slope of the fitted regression line is

$$r = \frac{S_X}{S_Y}\hat{\beta}_1 \tag{2.10}$$

where S_X, S_Y are the sample standard deviations of X and Y and, as mentioned above, $\hat{\beta}_1$ is the slope of the line fitted to the data.

Also, this is equivalent to

$$r_{xy}^2 = 1 - \frac{S_{y|x}^2}{S_y^2}, \tag{2.11}$$

where $S_{y|x}^2$ is the square of the error of the linear regression of x_i on y_i determined by the equation $y = \beta_0 + \beta_1 x_i$ and

$$S_{y|x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Note that S_y^2 is the variance of y:

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

2.2 MISSING DATA

In 1976 Rubin developed a framework of inference from data with missing values that remains in use today. In 1988 Little [12] provided a review of methods that can handle missing covariates. Little classifies the methods that can handle missing covariates into six classes. The first class would be the complete-case (CC) analysis. In a complete case analysis all of the observations that have missing values are eliminated from the analysis. The second class would be the available-case (AC) methods. Available-case analysis methods uses the largest sets of available cases for individual parameters [11]. The problem with the AC analysis is that the estimated covariance matrix of the $X's$ is not necessarily positive definite, which leads to inferior results compared to CC analysis for highly correlated data [5]. The third class of methods discussed by Little consists of least squares (LS) on imputed data methods.

In this setting the missing $Z's$ are imputed and a regression of $Y's$ on $Z's$ is performed on the filled in data by ordinary least squares or weighted least squares regression. The imputation methods included were: unconditional and conditional mean imputation. The inferences based on these methods (tests and confidence intervals) appear to be biased and imprecise. The fourth class is the class of maximum likelihood (ML) methods. In this class a classical ML estimate for a model for the joint distribution of Y and (X, Z) would be one approach, where the joint distribution would be multivariate normal with mean μ and covariance matrix Σ. Another method mentioned by Little [12] would be the Expectation Maximization algorithm (EM).

The Expectation-Maximization algorithm (Dempster, Laird and Rubin, 1977; [1]) is a general iterative algorithm for maximum likelihood estimation in data with missing values. The EM algorithm consists of two major steps: (1) Expectation (E-step), (2) Maximization (M-step) [11]. The E-step consists of computing the conditional expectation of the complete data log-likelihood given the observed data and the current estimated parameters and then substitutes these expectations for the missing data. The M-step consists of computing the maximum likelihood estimates of the parameters as if there were no missing data. Specifically, the M-step finds the parameter estimates that maximize the complete-data log-likelihood from the E-step. Specifically, in our incomplete data setting, the distribution of the complete data X, Y, Z could be factored as [22]:

$$f(X, Y, Z|\theta) = f(X, Y|\theta)f(Z|X, Y, \theta)$$

and the likelihood function

$$L(\theta|X, Y, Z) = L(\theta|X, Y) + logf(Z|X, Y, \theta) + constant,$$

where $L(\theta|X, Y, Z) = logf(X, Y, Z|\theta)$ denotes the complete data likelihood and $L(\theta|X, Y)$ the observed data likelihood. The term $f(Z|X, Y, \theta)$ cannot be calculated because Z is unknown, so an average over the predictive distribution $f(Z|X, Y, \theta^{(t)})$ is calculated (θ^t is a preliminary estimate of the unknown parameter). This averaging will give [22]:

$$Q(\theta|\theta^{(t)}) = L(\theta|X, Y) + H(\theta|\theta^{(t)}) + const,$$

where

$$H(\theta|\theta^{(t)}) = \int log f(Z|X,Y,\theta) f(Z|X,Y,\theta^{(t)}) dZ.$$

The two steps of the EM algorithm can then be specified as follows [22]:

1. The Expectation or E-step, in which the function $Q(\theta|\theta^{(t)})$ is calculated by averaging the complete data likelihood $L(\theta|X,Y,Z)$ over $f(Z|X,Y,\theta^t)$ and;
2. The Maximization or M-step, in which $\theta^{(t+1)}$ is found by maximizing $Q(\theta|\theta^{(t+1)})$.

The ML class tends to perform better than the previously described classes, yielding consistent estimates and being more efficient. The fifth class mentioned in Little's paper was the Bayesian methods class. The Bayesian methods seem satisfactory for small sample inferences but they were mostly explored in the case of missing values for dependent variables and not for independent variables. With the likelihood function being very complex, the marginal posterior distributions have to be approximated by numerical integration or simulations which makes this method complicated to use. The last class described by Little includes the multiple imputation (MI) methods. Rubin (1987) introduced the idea of multiple imputation (MI) in which each missing value is replaced with $m > 1$ simulated values prior to analysis. This will produce m possible complete data sets that are analyzed in the same manner as a complete data set. The results are then combined using simple computation steps to obtain the overall estimates along with their standard errors that reflect missing data uncertainty and the sample variation. The maximum likelihood approach assumes a specific distribution model:

$$f(X,Y,Z) = \int f(X,Y,Z|\theta) f(\theta) d\theta,$$

while in the multiple imputation approach we are interested in drawing missing values from the posterior distribution of $Z|(X,Y)$:

$$f(Z|X,Y) = \int f(Z|X,Y,\theta) f(\theta|X,Y) d\theta,$$

where θ is the parameter of interest. The multiple imputation procedure assumes that the data are from a continuous multivariate distribution and that it contains missing values that can occur on any of the variables. Each value is a conditional draw from the conditional

distribution of the missing observations given the observed data. This is done in such a way that the set of imputations properly represents the information about the missing value that is contained in the observed data for the chosen model. Additionally when a Markov Chain Monte Carlo (MCMC) or regression method is used, MI assumes a multivariate normal distribution for the data.

MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution. In simple words a Markov chain is a stochastic process that has the Markov property, that is, the process is such that given the present state, the future states are independent of past states. Yang C. Yuan [26] gives a very good overview of the multiple imputation method implemented in SAS software. The MCMC method has its origins in physics where it was used as tool for exploring the equilibrium distributions of interacting molecules. Bayesian inference uses MCMC as a method for exploring posterior distributions. Using MCMC one could simulate the joint distribution of the unknown quantities and obtain simulation-based estimates of the parameters of interest.

The MCMC method has two important steps [26]: the imputation I-step and the posterior P-step. The imputation (I-step) simulates the missing values for each observation independently, with previously estimated mean and covariance matrix. Let the missing values be denoted z_i and the variables with observed values by (x_i, y_i), then the I-step draws values for z_i from the conditional distribution of $Z|(X, Y)$. The P-step simulates the posterior mean and covariance matrix from the complete sample and uses these new estimates in the imputation step. The iterates have to converge to their stationary distribution and then to simulate an approximately independent draw of the missing values. Therefore, with the interest estimate $\theta^{(t)}$ at the t^{th} iteration, the imputation step will draw Z^{t+1} from $f(Z|(X, Y), \theta^{(t)})$ and the posterior step will draw $\theta^{(t+1)}$ from $f(\theta|(X, Y), Z^{(t+1)})$. This will create a Markov chain long enough $(Z^{(1)}, \theta^{(1)}), (Z^{(2)}, \theta^{(2)}), ...$, to converge in distribution to $f(Z, \theta|(X, Y))$. The MI procedure uses the means and standard deviations from available cases as the initial estimates for the EM algorithm. The MI procedure (default) uses the MCMC method with a single chain to create five imputations, with initial estimates from the EM and correlations set to zero. The highest observed data posterior density with a

noninformative prior is computed from the EM algorithm and it is used as the starting value for the chain. The MI procedure takes 200 burn-in iterations before the first imputation and 100 iterations between imputations. The MI imputation model does not make any distinctions between the response (dependent) or predictor (independent) variables but treats all as a multivariate response. The imputation model does not provide a parsimonious description of the data nor does it represent any type of relationship between variables. Distinctions between dependent and independent variables should be left to post imputation analysis. The multiple imputation technique has many attractive features. It allows the user to proceed with familiar complete data analysis methods. Rubin has also shown that there is no need for many repetitions in order to find precise estimates. The efficiency of an estimate based on m imputations relative to one based on an infinite number is calculated using the following formula:

$$(1 + \frac{\lambda}{m})^{-1}, \tag{2.12}$$

where λ is the rate of missing information and m is the number of imputations. Also missing values for each variable are predicted from its own observed values with random noise added to preserve a correct amount of variability in the imputed data. Rubin (1987; [21]) has also provided rules for combining the results of an analysis from all completed datasets resulting from imputations. Multiple imputation is very useful for database construction, because once the imputations are created, the data can be analyzed using complete data methods.

2.3 INVERSE PROBABILITY WEIGHTING

The estimation of regression coefficients from data with missing covariates has been widely studied and a variety of methods are available. Robins et al. [18] and Zhao et al. [27] have proposed weighted estimating equations that lead to consistent and asymptotically normal estimators of the $\beta's$. With weighted estimating equations, the contribution to the estimating equation from a complete observation is weighted by the inverse of the probability of being observed. This method has been mostly used in regression analysis, two-stage sample survey analysis and in generalized estimating equations. The roots of inverse probability weighting

are to be found in survey analysis [7]. In the past ten years several authors [19]; [18]; [23] have proposed improved IPW estimates that performed theoretically more efficiently under the MAR assumption [9]. The idea that lies behind the inverse probability weighting method is based on weighting each observation by the inverse probability of being observed. Carpenter and Kenward [9] present an overview of this method along with its recent developments. In the context of linear regression, the inverse probability weighting technique would apply as described in the following paragraph. Consider the general linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where ϵ are independent and $\epsilon_i \sim N(0, \sigma^2)$, or

$$Y = X\beta + \epsilon,$$

where $Y_{n\times 1}$, $X_{n\times 2}$, $\beta_{2\times 1}$ and $\epsilon_{n\times 1}$. The residuals are $\epsilon = Y - X\beta$ and we need to minimize the residuals sums of squares: $\epsilon^T\epsilon = (Y - X\beta)^T(Y - X\beta)$. Therefore, we take the derivatives with respect to β from $\frac{d}{d\beta}(Y - X\beta)^T(Y - X\beta)$ which gives the score, or the estimating equations: $\sum_{i=1}^{n} X_i(Y_i - X_i\beta)$. In the case of a complete data set these equations are set to zero and the parameter estimates are easily calculated. In the case of missing values the estimates of these parameters will change. Then the observed data estimating equations can be written [9]:

$$\sum_{i=1}^{n} R_i X_i(Y_i - X_i\beta).$$

Therefore, weighting by $\frac{1}{\Pi_i}$ will give [9]:

$$\sum_{i=1}^{n} \frac{R_i}{\Pi_i} X_i(Y_i - X_i\beta) = 0.$$

The IPW estimates are not as efficient as likelihood-based estimates and they are very sensitive to the precise form of model for the probability of response.

2.4 PARTIAL LEAST SQUARES

Partial least squares is a technique that extends multiple linear regression without imposing the restrictions that discriminant analysis, principal components regression, and canonical correlation do. Principal components, discriminant analysis and canonical correlation allow factors to be extracted from either X^TX or Y^TY as opposed to partial least squares which would allow the factors to be extracted from the cross-product matrices, (i.e. X^TY or Y^TX). In partial least squares, prediction functions are represented by factors extracted from the Y^TXX^TY matrix, where Y is the matrix of dependent variables and X is the matrix of independent variables. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of Y and X variables. The use of traditional multivariate methods is limited, especially when there are fewer observations than there are predictor variables which makes the use of the partial least square a desirable technique in this situation. The goal of PLS is to predict Y from X and, at the same time, to describe the common structure [13]. PLS searches for a set of components called latent vectors that performs a simultaneous decomposition of X and Y and explain, at the same time, as much as possible of the covariance between X and Y. PLS decomposes $X_{n\times p}$ and $Y_{n\times q}$ as a product of a orthogonal factors and a set of specific loadings. The matrix of independent variables, X is decomposed as $X = TP^T$ with $T^TT = I$, with I being the identity matrix. $T_{n\times r}$ is called the X-score matrix and $P_{p\times r}$ is called the loading matrix. Similarly, $Y_{n\times q}$ is estimated as $Y = BQ^T$ where $B_{n\times r}$ is the Y-scores matrix and $Q_{q\times r}$ is the loading Y matrix.

Each extracted x-score is a linear combination of X. The first extracted x-score t of X is of the form $t = Xv$, where v is the eigenvector corresponding to the first eigenvalue of X^TYY^TX. Similarly, the first y-score of Y, b, is of the form $b = Yu$ where u is the eigenvector corresponding to the first eigenvalue of Y^TXX^TY(i.e note that X^TY is the $Cov(X, Y)$). The first step of the partial least squares is to compute the singular value decomposition of the cross-block covariance matrix, $Cov(X, Y) = \rho\sigma_X\sigma_Y$. Let us denote this covariance matrix by Σ. In order to perform a singular value decomposition on Σ we need to find the matrices

U, S and V such that

$$\Sigma_{n\times p} = U_{n\times n} S_{n\times p} V^t_{p\times p}, \tag{2.13}$$

where $U^tU = I_{n\times n}$ and $V^tV = I_{p\times p}$.The columns of V are actually the eigenvectors of $\Sigma^t\Sigma$ and the columns of U are the eigenvectors of $\Sigma\Sigma^t$. To compute the eigenvectors of a matrix we use the following definition:

Definition 5. *Let us denote $\Sigma\Sigma^t = W$. For an $n \times n$ matrix W, a nonzero vector x is the eigenvector of W if:*

$$(W - \lambda I)x = 0,$$

for some scalar λ and I being the $n \times n$ identity matrix. Then the scalar λ is called an eigenvalue of Σ, and x is said to be an eigenvector of Σ corresponding to λ.

More specifically, the columns of the matrix U: $u_1, u_2, ..., u_n$, $n \in \mathbb{N}$, from the SVD represent the eigenvectors of the decomposition of the product of the covariance matrix and its transpose, $\Sigma_{xy}\Sigma^T_{xy}$ and will be used to calculate the $X - scores$. The columns of V, with columns $v_1, v_2, ..., v_p$, $p \in \mathbb{N}$, represent the eigenvectors corresponding to the eigenvalues of the product $\Sigma^T_{xy}\Sigma_{xy}$ and will be used to calculate the $Y - scores$. The matrix of $X - scores$, T, is composed of columns, $t_1, ..., t_i$, where i represents the number of scores one would like to calculate. Each of these columns is computed as:

$$t_{1j} = Xu_{1j}, j = 1, ..., i.$$

The same process is employed to calculate the matrix of $Y - scores$, B, $b_1, ..., b_j$, where j represents the number of $Y - scores$ to be extracted. The first computed $Y - scores$ will be:

$$b_1 = Yv_1.$$

The diagonal matrix, D from the SVD algorithm is the matrix of singular values. The sum of the squared singular values is equal to the sum of the squared cross covariances. More specifically, the diagonal matrix D is given by: $D = diag(d_1, d_2, d_3, ..., d_n)$ and the sum of squared singular values is given by: $\sum_{i=1}^{n} d_i^2$. The ratio, $(\frac{d_i^2}{\sum_{i=1}^{n} d_i^2})$, is "the proportion of the

summed squared cross-block covariance accounted for by the singular vectors". This matrix will give the percentages of variability that are explained by the PLS scores after performing the described calculations.

2.5 PARTIAL LEAST SQUARES SCORES FROM DATA WITH MISSING VALUES

Calculation of PLS scores from data with missing values has not received much attention in the literature. PLS has been extensively used in the chemometrics field, but until recently, the missing data problem has not been investigated. Some work has been done with respect to missing values in PCA. P.Ho and Silva [6] used multiple imputation to create m data sets and performed PCA on each data set and the scores were averaged over the m data sets to produce the overall score estimates. Walczak and Massart [24] use the iterative algorithm (IA) to deal with the missing data problem in PLS and PCA. Each iteration of the IA algorithm consists of two steps. In the first step, estimation of the model parameters is performed as if there were no missing data. The second step consists of computation of the conditional expectation of the missing elements given the observed data and the current estimated parameters. The elements are replaced with the expected values calculated as the mean of the corresponding row and column mean. Another way to calculate initial estimates that is described in the paper is by use of the matching procedure. In the proposed matching procedure, the missing elements for the $i-th$ object are replaced with the corresponding values observed for the most similar object from the studied data set. The so called IA (iterative algorithm) that Walczak and Massart [24] describe is basically the EM algorithm. The IA is summarized as follows [24]: (1) fill in missing elements with their initial estimates for factors (1:A); (2) calculate the mean of X and of Y; (3) center X and Y; (4) calculate weights, scores and loadings; (5) substract the predicted X and Y from the original X and Y and go to step (4); (6) reconstruct X with the actual set of scores and loadings(and A factors); (7) fill in missing elements in X with their estimates and go to step (2) until convergence.

Nelson, Taylor and McGregor [15] provide a comparison of the single component projection method derived from the NIPALS (nonlinear iterative partial least squares) algorithm and conditional mean replacement for dealing with the missing problem in PLS. NIPALS is the algorithm developed by H. Wold [25] in the late sixties for the econometrics field and adopted later in chemistry. Various versions of the NIPALS algorithm have been developed and used in different areas since then. The NIPALS algorithm was first used for PCA and later for PLS. The NIPALS algorithm is an alternative way to compute the singular value decomposition of a covariance matrix for calculating eigenvectors, and can be applied to PLS. The PLS-NIPALS algorithm with missing values can be associated with a simple imputation method where the missing data are estimated using simple regression. In every iteration for calculation of principal components or latent variables, the residuals for the missing elements in the least square function are set to zero or the missing values are replaced by their minimum distance projection onto the current estimate of the loading and score vector [15]. Nelson et. al have shown that this works well when the percentage of missing data is no higher than 20%. Furthermore, they have shown that data replacement by conditional mean, single component projection (derived from the NIPALS algorithm) and a method of simultaneous projection to the model plane performed well with no more than 20% missing data. The NIPALS algorithm is usually recommended only when the missing data pattern is random rather than structured. The single projection method consists of applying the NIPALS missing data model building algorithm to each dimension separately. It treats the missing data separately in the calculations of each latent dimension. The projection method based on the projection to the model plane is a complicated and computationally intensive algorithm. This method consists of calculating all of the scores at once by projecting onto the hyperplane formed by the loading vectors ($p_i's$) [15]. The missing data replacement using conditional mean replacement is computed in the expectation step of the EM algorithm. This method is used after the construction of the PLS model, so the interest lies only in the handling of the missing data in future multivariate observations since the estimates of the mean and covariance matrix are already available from the modelling step. All of these methods performed well for a small amount of missing data (20%) although there were issues with all of them. These issues include: (1) collinearity of the loading vectors after loadings

for the missing measurements; (2) similarities of the magnitudes of the scores (in single component projection); (3) underestimating the dimensionality of the data and (4) noise (with the simultaneous projection method). The expected mean replacement method was superior when compared to the other two [15].

3.0 COVARIANCE MATRIX

3.1 INTRODUCTION

Most of the datasets that statisticians use in different areas of research contain missing values. Depending on the goal of the study the researcher chooses to replace the missing values using different techniques or to drop the observations from the analysis and use only complete cases for further investigation. Completely deleting cases with missing values leads to potential bias in the estimates and affects inferences drawn from the data. Our research questions have arisen from a dataset with missing values. The goal of our study was the estimation of the correlation/covariance matrix when some covariates are not completely observed. The estimation of correlation/covariance matrix in the presence of missing data has not received a lot of attention in the statistical literature. In general, maximum likelihood methods have been used to estimate the covariance/correlation.

Dempster, Laird and Rubin [1] proposed a general approach to iterative computation of maximum-likelihood estimates for data with missing values, the expectation maximization algorithm (EM). Dixon (1983) mentioned four methods that could be used to calculate the correlation matrix: listwise, pairwise, allvalue and samemean methods which were discussed in further detail by Javaid Kaiser in 1994, [8]. Rubin (1976) developed a framework of inference from data with missing values that remains in use today. Little (1992) provided a review of methods that can address missing covariates. Little classifies the methods that can handle missing covariates into six classes (complete-case (CC), available-case (AC), Least Squares (LS) on imputed data methods, maximum likelihood (ML) methods (i.e. EM algorithm), Bayesian methods and multiple imputation methods.

Multiple imputation was first introduced by Rubin in 1987 and has become very popular

and widely used in recent years. The idea that lies behind multiple imputation (MI) is the replacement of each missing value with m simulated values, $(m > 1)$, prior to analysis. This will produce m possible data sets with no missing values that are analyzed in the same manner as a complete data set. The results are then combined to obtain overall estimates along with their standard errors that reflect missing data uncertainty as well as the sample variation. The multiple imputation technique has many attractive feature since it allows the user to proceed with familiar complete data analysis methods. Rubin has also shown that there is no need for many repetitions in order to find precise estimates. Multiple imputation is very useful for database construction since once the imputations are created, the data can be analyzed using complete data methods.

In recent years weighted statistical methods, (i.e. inverse probability weighting, weighted estimating equations) have become of greater interest in the statistical community facing datasets with missing observations. The roots of inverse probability weighting (IPW) method can be found in survey sampling techniques, [7] and over the last decade several authors [19]; [18]; [23] have proposed improved inverse probability weighting (IPW) estimates. With weighted estimating equations, the contribution to the estimating equation from a complete observation is weighted by the inverse of the probability of being observed. This method has been mostly used in regression analysis, two stage sample surveys and generalized estimating equations. The inverse probability weighting method could be easily implemented since many statistical routines include an option for weighting, which makes it very attractive.

The inverse probability weighting method is a new approach used in this specific context for reconstructing a covariance matrix from data with missing covariates. The technique consists of calculating the probability of being observed for each record in the dataset for the variable with missing values and then performing a weighted regression analysis with the weights given by the inverse probability of being observed. The covariance matrix is then reconstructed from estimated pairwise correlations between two variables by weighted linear regression. A comparison of the weighted method with the covariance calculated using multiple imputation and complete cases is presented. This new approach provides a useful technique without the burden of exhaustive computation or technicalities. The robustness of the method is somewhat limited by the missing data model specification, since the missing

data model needs to be correctly specified for unbiased estimates.

In the present work two different techniques for calculating the covariance matrix from data with missing covariates; namely multiple imputation and inverse probability weighting are compared to covariance matrix calculation based on complete cases. Complete case analysis consists of deleting all observations with missing values and performing the calculations only on those cases that are completely observed. In the multiple imputation case the missing values are imputed based on a specified multiple imputation method, in this case a Markov Chain Monte Carlo and then computing the covariance matrix as if there were no missing data.

In our context, the implementation is performed under the assumption of data missing at random (MAR) as defined by [20]. In addition, we assume that missing values occur only in the covariate data. A detailed description of these methods and their implementation in our context is presented in section two of this chapter. Simulation studies have been conducted for assessing the theoretical properties of the covariance matrix estimates obtained from the above mentioned techniques. A description of simulation studies along with their results are given in the third section of present chapter. An application of these methods is illustrated in a Positron Emission Tomography (PET) study conducted at University of Pittsburgh, PET Center. A discussion, along with related results is presented in section four of this chapter. The conclusions of our findings and further recommendations are discussed in the fifth section entitled "Discussion".

3.2 METHODS

Consider estimating the correlation/covariance matrix from a dataset (X, Y) where X represents the vector of covariates, x_i, that are always observed, a covariate z_i, that is missing for some subjects, and Y, the vector of responses, y_i, in this case completely observed. An indicator variable, R_i is created and it is defined as follows:

$$R_i = \begin{cases} 1, & \text{if } z_i = observed, \\ 0, & \text{if } z_i = missing, \end{cases}$$

where z_i is the covariate with missing values previously introduced. The complete case analysis method uses available cases for each available pair of observations with all other observations dropped from analysis. Each entry of the estimated covariance matrix, $S_{n\times n}$, is calculated using the classic formula:

$$s_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ij} - \overline{x}_i)(y_{jk} - \overline{y}_j), \tag{3.1}$$

where $i = 1, ..., n,\ j = 1, ..., n$. The advantage of this method is its simplicity, since standard statistical methods can be applied for statistical inference. The main disadvantage of complete cases is the loss of information that results from discarding all cases with missing values. Complete-case analysis makes no use of cases with missing values when estimating the covariance between the given variable and the other variable of interest.

The method of multiple imputation, introduced first in 1976 by Rubin, [20] is a very useful technique that is often used when missing observations are present in datasets under investigation. Implementation of multiple imputation in standard statistical computer package routines (i.e. SAS PROC MI) makes it easy to apply, further increasing use in a data analysis. In the multiple imputation case, each missing value is replaced by m simulated values, $(m > 1)$, prior to performing the data analysis, resulting in the creation of m complete data sets. The analysis results are then combined to obtain overall estimates along with their standard errors that reflect missing data uncertainty as well as the sample variation. In the multiple imputation approach, we are interested in drawing missing values from the posterior distribution of $Z|(X, Y)$:

$$f(Z|(X,Y)) = \int f(Z|(X,Y),\theta) f(\theta|X,Y) d\theta,$$

where θ is the parameter of interest.

The main attractive feature of multiple imputation is the fact that after performing the imputations, the complete data sets are analyzed by standard statistical techniques.

Rubin [21] has also shown that there is no need for many repetitions in order to find precise estimates. The efficiency of an estimate based on m imputations relative to one based on an infinite number is calculated using the following formula:

$$(1 + \frac{\lambda}{m})^{-1}, \tag{3.2}$$

where λ is the rate of missing information and m is the number of imputations (i.e. 20% missing information, $m = 5$ efficiency is $\frac{1}{1.04} = 96\%$). Also missing values for each variable are predicted from its own observed values with random noise added to preserve a correct amount of variability in the imputed data. Rubin (1987), [21] has also provided rules for combining the results of the m complete datasets resulting from imputations. In our context, let Σ denote the population parameter that we would like to estimate, in this case the ij entry of the sample covariance matrix, S, and let se denote the standard error of this estimator. Since $m = 5$ data sets are created using MI and for each data set we will have one estimate of our parameter of interest, $\hat{s}_{ij}$, the overall estimate will be simply: $\overline{s_{ij}} = (\frac{1}{5}) \sum_{m=1}^{5} \hat{s}_{kl}^{(m)}$. Also, the uncertainty in $\overline{s}_{ij}$ contains the average of the within imputation variance, $\overline{se} = (\frac{1}{5}) \sum_{m=1}^{5} se^{(p)}$ and the between-imputation variance $B = (\frac{1}{4}) \sum_{m=1}^{5} [\hat{s}_{ij}^{(m)} - \overline{s}_{ij}]^2$. The total variance will be a modified sum of the two variance components, $T = \overline{se} + (1 + \frac{1}{5})B$, and the square root of T is the overall standard error. Multiple imputation is useful for dataset reconstruction, since once the imputations are created, the data can be analyzed using complete data methods.

The inverse probability weighting method was also used to reconstruct the correlation/covariance matrix. The inverse probability weighting method has been applied primarily to linear and logistic regression estimates with one continuous covariate missing or more than one covariate missing for the categorical case. Carpenter and Kenward (2006) present a nice overview of this method along with its recent developments as they apply to the estimation of coefficients in a linear regression problem with one missing covariate and compare it with multiple imputation. In our context, the inverse probability weighting method is used for calculating the pairwise correlations of any two variables by using a linear regression with weights given by the inverse probability of being observed. The probabilities

of being observed (Π) are estimated from a logistic regression model with R_i, previously defined, as the dependent variable and all $y_i's$ and $x_i's$ are introduced into the model along with interaction terms of interest. Note that incorrect specification of the logistic model might result in inconsistent estimates of the probabilities of being observed. Once the estimated $\hat{\Pi}_i$ have been calculated from:

$$\Pi_i = Pr(z_i = observed|x_i, y_i) = \frac{exp(-\beta_0 - \beta_i x_i - \beta_j y_i)}{1 + exp(-\beta_0 - \beta_i x_i - \beta_j y_i)}, \tag{3.3}$$

the weights are then calculated as $\frac{1}{\hat{\Pi}_i}$ and used in further analysis. In the present case these were used as weights in a linear regression model, where each dependent variable was regressed on each independent variable and the pairwise correlation was calculated from $\sqrt{R^2}$ and used to reconstruct the corresponding entries of the correlation/covariance matrix when the covariates involved had missing values. In this case, given that weighting is performed using the inverse probability of being observed, the sum of squares will incorporate weights as well. Therefore the sum of squares will be,

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (x_i - \bar{x})^2, \tag{3.4}$$

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (y_i - \bar{y})^2, \tag{3.5}$$

and the estimate of the slope, $\hat{\beta}$, will be

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \frac{1}{\hat{\Pi} i} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (x_i - \bar{x})^2}. \tag{3.6}$$

Substituting in $r_{xy} = \frac{S_x}{S_y} \hat{\beta}$ we will have the following equation for the estimate of r_{xy}:

$$\hat{r}_{xy} = \frac{\sqrt{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (x_i - \bar{x})^2 \sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (y_i - \bar{y}_i)^2}}{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (y_i - \bar{y}_i)^2} \frac{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i} (x_i - \bar{x})^2}. \tag{3.7}$$

In the next section we present our simulation studies and results along with a discussion section and future recommendations.

3.3 SIMULATIONS

In this work, several simulation studies have been conducted to examine the performance of the methods in practice. Covariance matrices were created under different structures and each covariance structure was used as the starting matrix for the simulations. Several datasets were created with different percentages of missing values (i.e. $20\%, 30\%, 50\%$). Also, a variety of missing data scenarios were explored as well. The missing values were created so that they depend on either the $x_i's$ or the $y_i's$ or neither x_i or y_i. Model misspecifications were explored for some of the simulations for both multiple imputation and inverse probability weighting.

Each dataset, (X, Y), X, Y vectors, was simulated from a multivariate normal distribution, $MVN(\mu, \Sigma)$. The first set of simulations consisted of 1000 simulated data sets of 100 observations each. The simulations were conducted using R software. We have simulated 20%, 30% and 50% missing values for each data set. We have simulated datasets with 50 and 200 observations and examined the performance of our employed methods. In the simulated datasets, the outcome variable, Y had no missing values. Also, some of the $X's$ were fully observed covariates while some had missing observations.

Let us denote our observed data set by X and partition it into (X, Z), where X contains the covariates with fully observed values while Z contains the covariates for which some of the values are missing. For any data set with missing values one could define an indicator variable, R_i, that will define the missingness process in the following way:

$$R_i = \begin{cases} 1, & \text{if } z_i = observed, \\ 0, & \text{if } z_i = missing, \end{cases} \text{ where, } R_i \sim Bernoulli(\Pi).$$

In our simulations the missing values were created based on a logistic regression model. We have calculated the probability of each record in the data set being observed based on the logistic regression model:

$$\Pi_i = Pr(z_i = observed | x_i, y_i) = \frac{exp(-\beta_0 - \beta_i x_i - \beta_j y_i)}{1 + exp(-\beta_0 - \beta_i x_i - \beta_j y_i)}. \tag{3.8}$$

We fixed the logistic regression coefficients so that we could create roughly 20%, 30% and 50% of the data with missing values. The indicator variable R_i has been simulated

from a binomial distribution with probability of success ($R_i = 1$) being equal to Π_i. If the missingness is ignorable then Π_i does not depend on the z_i. We assumed a MAR missing data mechanism in which Π_i does not depend on the z_i. We have first calculated correlation coefficients using only complete cases, therefore all of the missing observations were excluded from the analysis (each pairwise correlation was calculated based on the available cases). For each data set the correlation of any pair of two variables (X_{ik}, X_{jk}) is calculated and denoted by r_{ijk}, (where k corresponds to the data set). The overall correlation coefficient for any pair for all of the simulated data sets was calculated as $\bar{r} = \sum_{k=1}^{p} r_{ijk}$, where $k = 1, ..., p$ datasets are used. Also, the bias and mean square error were calculated.

The second method performed was multiple imputation where values where imputed for the missing covariates and the correlation for each data set was calculated as if the dataset were complete. In the multiple imputation model, all variables that were used to generate the data as MAR as well as variables that were further employed in the analysis were included in the model. PROC MI in SAS was used for the computations. By default this procedure uses the MCMC method with a single chain to create $m = 5$ imputations. The EM algorithm is used to compute the starting value for the chain. The MI procedure performs 200 burn-in iterations before the first imputation and 100 iterations between imputations. PROC CORR was used to calculate the pairwise correlations. Since 5 data sets were created using MI and for each data set one estimate for each pairwise correlation of our parameter of interest was calculated denoted as $\hat{r}^i$, the overall estimate was: $\overline{r} = (\frac{1}{5}) \sum_{i=1}^{5} \hat{r}^{(i)}$. Also, the uncertainty in $\overline{r}$ contains the average of the within imputation variance, $\overline{S} = (\frac{1}{5}) \sum_{i=1}^{5} S^{(i)}$ and the between-imputation variance $B = (\frac{1}{4}) \sum_{i=1}^{5} [\hat{r}^{(i)} - \overline{r}]^2$. The total variance will be a modified sum of the two variance components, $T = \overline{S} + (1 + \frac{1}{5})B$, and the square root of T is the overall standard error.

The third method used was the inverse probability weighted method. The first step was a linear regression ($y_i = \beta_0 + \beta_1 x_i + \epsilon$, $\epsilon \sim N(0, \sigma^2)$) with weights $\frac{1}{\Pi_i}$. The estimated probability of being observed (Π_i) was calculated for each record using a logistic regression. Each outcome was regressed on each covariate and the correlation was calculated as $r = \sqrt{R^2}$

or $r = \frac{S_X}{S_Y}\hat{\beta}_1$ given by equation 3.7.

3.3.1 Results with correct models

For the results that are discussed in this section we have used correct models for both the multiple imputation and inverse probability weighting methods. More specifically, when the missing observations were dependent on any of the $x's$ or the $y's$, the variables were kept in the model when estimating the probability $P(R_i = 1)$, i.e., the probability of observing the $i'th$ subject's covariate in that specific case. In the first set of simulations, we performed 1000 simulations and each dataset had 100 observations. The probability of being observed did not depend on any of the $x's$ or the $y's$ and the coefficients of the logistic model were $(\beta_0, \beta_1, \beta_2) = (\ln 4, 0, 0)$. Even though the missing values were not dependent on any of the $y's$ or the $x's$, the probability of being observed was still estimated using a stepwise logistic model with the constraint that a variable was kept in the model if $p = .20$. The weights were calculated based on the estimated model in all cases. The estimated logistic regression coefficients were almost the same or very close in value to the true logistic coefficients for all models.

The results for the first set of simulations are presented in Table 1 along with the associated bias and mean square error. Note that only one row of the correlation matrix is presented due to the fact that missing data affect only one covariate. The estimated logistic regression coefficients for estimating the probability of being observed for the covariate with missing values were very close to the true coefficients with the true values being:$(\beta_0, \beta_1, \beta_2) = (\ln 4, 0, 0)$, and the estimated values being:$(\beta_0, \beta_1, \beta_2) = (\ln 4, 0.007, 0)$. The percentage of missing values in Table 1 is 20%.

The crude bias was calculated as $E(\hat{r}) - r$ and the mean square error as $Var(\hat{r}) + (Bias(\hat{r}))^2$. Different percentages of missing values were also investigated and the results are reviewed in the Discussion section. We also used a different random correlation matrix and new simulations were performed under different scenarios of missingness. In this case the missing values were created based on a logistic model with the probability of being observed

Table 1: Correlations from MI, IPW and CC when the missing data is independent of both x and y

Name	y_1	y_2	y_3	x_2	x_3
true_value	0.337	-0.228	-0.411	-0.260	-0.591
method					
CC	0.333	-0.225	-0.410	-0.259	-0.590
MI	0.334	-0.226	-0.408	-0.261	-0.590
IPW	0.333	-0.227	-0.410	-0.260	-0.590
Bias					
CC	-0.004	0.003	0.001	0.001	0.001
MI	-0.003	0.002	0.003	-0.001	0.001
IPW	-0.004	0.001	0.001	-0.000	0.001
MSE					
CC	0.010	0.011	0.009	0.011	0.006
MI	0.008	0.010	0.008	0.010	0.005
IPW	0.009	0.010	0.009	0.010	0.005

depending on none or depending on the $y's$. The coefficients of the logistic regression models used to create the missing observations were $(\beta_0, \beta_1, \beta_2, \beta_3) = (ln(4), -.7, 1, -.3)$ when the missing values were dependent on the values of $y's$. Also, the multiple imputation model contains all of the variables that were used to create the missing at random data (y_1, y_2, y_3). The results for 20% missing data, when the probability of missing is dependent on $y's$ and for a sample size of 100 are presented in table 2. Also, for the second covariance structure, the results for the case where missing values were not dependent on $x's$ or $y's$ are presented in the discussion section.

The results for the third covariance structure with 20% of the data missing, where missing data values were created dependent on some of the $x's$, are presented in Table 3.

Table 2: Correlations from MI, IPW and CC, when the missing data depend on y

Name	y_1	y_2	y_3	x_2	x_3
true_value	-0.548	-0.390	0.216	-0.578	0.140
method					
CC	-0.555	-0.375	0.275	-0.568	0.107
MI	-0.542	-0.386	0.214	-0.573	0.136
IPW	-0.548	-0.385	-0.245	-0.574	0.142
Bias					
CC	-0.007	0.015	0.059	0.010	-0.033
MI	0.006	0.004	-0.002	0.005	-0.004
IPW	0.000	0.005	0.029	0.004	0.002
MSE					
CC	0.007	0.011	0.016	0.007	0.014
MI	0.008	0.010	0.007	0.006	0.012
IPW	0.008	0.009	0.010	0.006	0.008

Additional results for 30% percent of missing values for this scenario are presented in the Appendix section. The coefficients of the logistic regression models used to create the missing observations were $(\beta_0, \beta_1, \beta_2, \beta_3) = (ln(4), -1, 1)$ when the missing values were dependent on the values of $x's$.

Table 3: Correlations from MI, IPW and CC when the missing data is dependent on x

Name	y_1	y_2	y_3	x_2	x_3
true_value	-0.436	0.583	0.294	0.518	0.538
method					
CC	-0.447	0.578	0.302	0.538	0.570
MI	-0.434	0.576	0.288	0.512	0.536
IPW	-0.442	0.578	0.298	0.515	0.542
Bias					
CC	-0.011	-0.005	0.008	0.020	0.032
MI	-0.002	-0.007	0.005	-0.006	-0.002
IPW	-0.006	-0.005	0.005	-0.003	0.004
MSE					
CC	0.010	0.013	0.012	0.008	0.009
MI	0.008	0.011	0.007	0.007	0.012
IPW	0.008	0.011	0.006	0.006	0.008

3.3.2 Results with incorrect models

In this section, using some of the simulated data sets we explored the possibility of estimating the covariance matrix when the models for the probability of being observed were not correctly specified (some of the key variables were not included in the model) and also when the imputation models do not contain all of the variables. For the second covariance structure, with results presented in Table 4, the missing values were dependent on $y's$ and the correct model for estimating probability of missingness had the coefficients: $(\beta_0, \beta_1, \beta_2, \beta_3) = (ln(4), -.7, 1, -.3)$. In this case we have run simulation studies without including the variable y_2 in the multiple imputation model or in the inverse probability weighted model. The results are presented in Table 4.

Table 4: Correlations from MI, IPW and CC when missing data is dependent on y and the model is misspecified

Name	y_1	y_2	y_3	x_2	x_3
true_value	-0.548	-0.390	0.216	-0.578	0.140
method					
MI	-0.576	-0.322	0.242	-0.556	0.115
IPW	-0.551	-0.365	0.285	-0.561	0.128
Bias					
MI	0.011	0.069	0.026	0.023	-0.025
IPW	-0.027	-0.025	0.069	0.017	0.012
MSE					
MI	0.009	0.021	0.018	0.010	0.017
IPW	0.009	0.012	0.018	0.008	0.009

Under the incorrect model, with estimates computed using multiple imputation and inverse probability weighting, the bias is obvious for both methods. As mentioned before, both the multiple imputation and inverse probability weighting are sensitive to the misspecification of the model of missingness. The bias is larger when compared to the bias obtained with correct models and also, the mean square errors are larger under misspecification. The mean square errors for the inverse probability weighting method, seem to be either equal to the multiple imputation ones or smaller, indicating that inverse probability weighting performs better than multiple imputation under this misspecification. The bias of the estimates is comparable between the two; some variables have smaller bias for inverse probability weighting and some have a little smaller bias for the multiple imputation.

3.4 DISCUSSION

The most important thing we would like to point out is that in the case where the missing values are not dependent on either $x's$ or $y's$, all three methods give very similar results. The percentages of missingness were varied for some of the simulations, i.e. 30%, 50% missing values were also considered for some cases, and some of the results are presented below. Only one table is included due to the fact that a similar trend was observed across the simulations. More specifically, the conclusions are very similar regarding the percentage of missing values (see Table number 5). Also, as expected, not much difference has been observed between estimates obtained with 20% missing data as compared to those obtained when 30% percent of the data are missing (i.e Table 5 and Table 6). When the percentage of missing is roughly around 50%, the estimates are more biased as compared to previously discussed cases (20%, 30%).

Across all of the simulation studies, a similar trend was observed, with an increase in the percentage of missingness (i.e. 50%), correlation coefficient estimates tend to be more biased and have higher mean square errors. The complete case analysis results are similar to the other two methods, when missingness does not depend on any other variable in the dataset. In the case where the missing values are dependent on either of the $x's$ or of the $y's$, the complete case analysis tends to be more biased when compared to inverse probability weighting and multiple imputation. The mean square errors of the inverse probability weighting were generally of equal or smaller values as compared to those of multiple imputation across simulations. The estimates of the correlation for complete cases are close to those computed using multiple imputation and inverse probability weighting, especially when the missing data does not depend on any other variable. In the case where the missing values are dependent on either some of the $x's$ or some of the $y's$ the complete case estimates are more biased when compared to those obtained from multiple imputation and inverse probability weighting.

One advantage of the inverse probability weighting method is that is less computational than multiple imputation. In addition, inverse probability weighting is easier to implement in any software, as opposed to multiple imputation which is easiest to implement in SAS.

Furthermore, the inverse probability weighting method relies on fewer assumptions as compared to multiple imputation. This work shows that the specification of the missing model plays a trivial role for both multiple imputation and inverse probability weighting. A downside of multiple imputation is that the data should be missing at random. If the data is not missing at random, this will impact the estimates. When a combination of discrete and ordinal variables are included in the model multiple imputation can be difficult to implement. In addition, the joint distribution, $(f(Z, X, Y))$ should be correctly specified or the estimates will be biased and inconsistent. The inverse probability weighting method is preferred since it does not require all of these assumptions and also performs as well as multiple imputation. Furthermore, the inverse probability method is not imputing any value in the dataset, but it is computing the best estimates making use of the available information at hand. The multiple imputation method creates values that are based on several assumptions established prior to analysis, and this method is not always well received because of the scepticism that the imputed values are just a " sophisticated guess".

3.5 POSITRON EMISSION TOMOGRAPHY EXAMPLE

The Mini Mental State Examination is a tool that can be used to asses mental status. The MMSE is effective as a screening instrument to separate patients with cognitive impairment from those without it. The MMSE is a 30 point questionnaire that is used to screen for cognitive impairment. An MMSE score above 27 is considered to be normal; between 20 and 26 indicates mild cognitive impairment; between 10 and 19 indicates moderate cognitive impairment, and below 10 indicates severe cognitive impairment (Folstein et al., 1975). Some experiments have found that individuals with Alzheimer's disease (AD) show increased activity in the prefrontal cortex compared to control subjects during some cognitive tasks [10].

A Positron Emission Tomography research study was conducted at the University of Pittsburgh Medical Center and one of the questions of interest was to explore the connec-

tion that exists between the increased activity in some of the brain regions and cognitive tasks. Each subject underwent a PET scan where FDG ($[^{18}F]$-2-deoxy-2-fluoro-D-glucose) was used as a imaging tracer. The FDG tracer (a radioactive form of sugar) is used to assess glucose metabolism in the brain. The regions-of-interest (ROI) measurements or normalized functional image data under investigation are: Frontal Cortex (FRC), Lateral Temporal Cortex (LTC), Mesial Temporal Cortex (MTC), Dorsal Frontal Cortex (DFC), Parietal (PAR), Occipital (OCC), Anterior Cingulate Gyrus (ACG). The ROI's are predefined prior to analysis and the accumulation of the tracer is evaluated using Positron Emission Tomography imaging. The accumulation of the FDG tracer is calculated by averaging across all voxels within that specific region, resulting in a continuous measurement for each of the brain regions under investigation. In subjects in whom some mild or severe cognitive impairment is suspected, low levels of accumulations of the compound are expected to be seen in certain parts of the brain.

The research study included 111 subjects classified into three different groups. The three groups were healthy subjects (Control or non-disease group), the Alzheimer disease (AD) and the mild cognitive impairement (MCI) group. The control group consists of 67 subjects, 22 males (age 74 ± 10) and 45 females (age 71 ± 9). In the MCI group the number of subjects was 25 with 19 males (age 71 ± 8) and 6 females (age 67 ± 7) and the AD group consists of 19 subjects, 14 males (age 70 ± 10) and 5 females (age 72 ± 8.3). Additionally, a set of covariates were measured for each of the subjects including: age, gender and cognitive status (i.e. Mini Mental State Examination, abbreviated MMSE). Several MMSE tests were performed. Some of the MMSE values are missing. The MMSE score measurements for each group divided by gender class is presented in table 5.

The correlation matrix calculated using only complete cases was performed in SAS using PROC CORR and results are provided in Table 6. The pairwise correlation coefficients were calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Table 5: Mini Mental State examination

Group	Gender	MMSE	$MMSE_7$	$MMSE_W$
AD(19)	M(14)	22.41±3.82	22.42±3.32	23.92±3.51
	F(5)	21.33±3.78	21.4±3.84	24.4±2.7
MCI(25)	M(19)	26.75±2.8	27.05±2.8	27.15±2.5
	F(6)	27.8±0.44	28.16±0.75	28.6±1.03
Controls(67)	M(22)	27.8±2.27	27.8±1.9	27.8±1.73
	F(45)	28.9±1.27	28.86±1.48	29.2±0.83

where $\bar{x}$, $\bar{y}$ are the sample means of X and Y.

Table 6: Correlations for complete cases

Name	DFC	FRC	LTC	MTC	OCC	PAR	ACG
AGE	.261	.279	.187	.399	.219	.297	.175
$MMSE_7$	.131	.151	.484	.068	-.132	.245	-.036
$MMSE_W$	.170	.183	.477	.043	-.154	.247	-.062
MMSE	.220	.237	.486	.148	-.23	.254	-.046

We applied multiple imputation to this data set. The covariate that had missing values was the MMSE score. A Markov Chain Monte Carlo (MCMC) method was used for the multiple imputation. The MCMC method assumes that data come from a multivariate distribution and that the values are missing at random (MAR) in the sense defined by Little and Rubin[11]. The multiple imputation model contains all of the variables in the data set since all of them will be used in further analysis. The multiple imputation model should contain all variables that will be further employed in the analysis. The maximum and

minimum value for imputed variables was specified in the multiple imputation procedure. Those values were previously calculated from the available cases in the data. The average, maximum, minimum and standard deviation for each variable with missing values were calculated before and after each imputation by group status. The results are discussed here and additional tables are provided in the Appendix section.

In the AD group the minimum value of the MMSE before imputation was 16 and maximum before imputation was 29 with an average of 22.2 and a standard deviation of 3.74. The average value of MMSE for the first imputation was 22.87 with a standard deviation of 3.87, for the second imputation it was 22.30 with a standard deviation of 3.54, for the third imputation 22.63, standard deviation 3.63, fourth imputation 22.72, standard deviation 4.07, and fifth imputation 22.70 with a standard deviation of 3.89. For each imputation the minimum and maximum values were 16.00 and 29.00 respectivelly. Similarly, for the control group, the minimum value of MMSE was 24.00 while maximum value was 30.00, with an average of 28.6 and a standard deviation equal to 1.68. For each imputation, the average values of MMSE were: $28.53, 28.56, 28.58, 28.65, 28.51$ and the standard deviations were $1.54, 1.49, 1.54, 1.56, 1.60$ respectively for the first, second, third, fourth and fifth imputation. In the same manner, for the MCI group the MMSE minimum was 20.00 and maximum was 30.00. The average value before imputation was 27.00 with a standard deviation equal to 2.56. The average values for each imputation were $26.85, 27.21, 27.24, 27.04, 27.03$ with the following standard deviations $2.52, 2.39, 2.39, 2.31, 2.31$ for imputation one, two, three, four and five.

In the multiple imputation case, all of the correlations will stay the same, the only one that is changing is the correlation between MMSE and all of the other variables since MMSE is the variable with missing values. The correlation matrix between covariates and brain regions obtained from the multiple imputation method is presented in Table 7.

Table 7: Correlations for all covariates with brain regions-MI

Name	DFC	FRC	LTC	MTC	OCC	PAR	ACG
AGE	.261	.279	.187	.399	.219	.297	.175
$MMSE_7$	.131	.151	.484	.068	-.132	.245	-.036
$MMSE_W$	.170	.183	.477	.043	-.154	.247	-.062
MMSE	.200	.214	.454	.150	-.171	.220	-.013

3.5.1 Inverse probability weighting

The inverse probability weighting method has also been investigated for this example. The estimated probability of being observed has to be calculated for each of the covariates with missing values. The regression models were constructed to extract the correlation between the covariate with missing values and the response variables. These models use the inverse estimated probabilities as weights. An indicator variable, R is created where R defines the missing process as follows:

$$R = \begin{cases} 1, & \text{if } MMSE = observed, \\ 0, & \text{if } MMSE = missing. \end{cases}$$

A logistic regression model with R as the response variable and all other variables as predictors is fit to the data. The predictor variables include all seven brain regions (ACG, DFC,FRC, LTC, MTC, OCC, PAR), group, gender, and the other two MMSE scores ($MMSE_7$ and $MMSE_W$). A stepwise logistic regression model was performed with a selection entry equal to selection stay with a value of .25. The best fitting model had only two of the predictors based on the criteria mentioned above namely: MTC and DFC. The estimates of the logistic regression are presented in Table 8. The correlation between each covariate and region-of-interest using inverse probability weighting is also computed. The correlation values are presented in Table 9.

Table 8: Logistic regression parameter estimates

Name	Parameter Estimate	St. error	p-value
Intercept	2.27	3.04	0.454
DFC	4.45	2.76	0.106
MTC	-8.00	4.18	0.056

Table 9: Correlations for all covariates with brain regions-IPW

Name	DFC	FRC	LTC	MTC	OCC	PAR	ACG
AGE	.261	.279	.187	.399	.219	.297	.175
$MMSE_7$	.131	.151	.484	.068	-.132	.245	-.036
$MMSE_W$	.170	.183	.477	.043	-.154	.247	-.062
MMSE	.225	.242	.490	.151	-.228	.250	-.045

3.6 ADDITIONAL TABLES FOR THE DATA EXAMPLE

Additional tables with summary statistics for the covariate with missing values (MMSE) in this case are provided in the first part of this Appendix. Table 10 is for the AD group, Table 11 is for the Control group and Table 12 is for the MCI group, along with the statistics for the data set prior to imputations. Table 13 provides the correlation estimates for each multiple imputation between MMSE and each brain region.

Table 10: Summary statistics for MMSE before and after imputations for AD group

Variable	Imputation status	Minimum	Maximum	Average	Standard Deviation
MMSE	Before Imputation	16.00	29.00	22.2	3.74
MMSE	Imputation_**1**	16.00	29.00	22.87	3.87
	Imputation_**2**	16.00	29.00	22.30	3.54
	Imputation_**3**	16.00	29.00	22.63	3.63
	Imputation_**4**	16.00	29.00	22.72	4.07
	Imputation_**5**	16.00	29.00	22.70	3.89

Table 11: Summary statistics for MMSE before and after imputations for Control group

Variable	Imputation status	Minimum	Maximum	Average	Standard Deviation
MMSE	Before Imputation	24.00	30.00	28.6	1.68
MMSE	Imputation_**1**	24.00	30.00	28.53	1.54
	Imputation_**2**	24.00	30.00	28.56	1.59
	Imputation_**3**	24.00	30.00	28.58	1.54
	Imputation_**4**	24.00	30.00	28.65	1.56
	Imputation_**5**	24.00	30.00	28.51	1.60

Table 12: Summary statistics for MMSE before and after imputations for MCI group

Variable	Imputation status	Minimum	Maximum	Average	Standard Deviation
MMSE	Before Imputation	20.00	30.00	27.00	2.56
MMSE	Imputation_**1**	20.00	30.00	26.85	2.52
	Imputation_**2**	20.00	30.00	27.21	2.39
	Imputation_**3**	20.00	30.00	27.24	2.39
	Imputation_**4**	20.00	30.00	27.04	2.31
	Imputation_**5**	20.00	30.00	27.03	2.31

Table 13: Correlations for MMSE with each brain region for each imputation

Name	Imputation	DFC	FRC	LTC	MTC	OCC	PAR	ACG
MMSE	Imputation_**1**	.217	.230	.441	.173	-.158	.218	-.017
MMSE	Imputation_**2**	.194	.210	.475	.127	-.176	.241	-.010
MMSE	Imputation_**3**	.209	.223	.469	.143	-.175	.229	-.011
MMSE	Imputation_**4**	.194	.208	.430	.162	-.156	.202	-.006
MMSE	Imputation_**5**	.185	.200	.453	.145	-.191	.212	-.020
MMSE	Average	.200	.214	.454	.150	-.171	.220	-.013

4.0 PARTIAL LEAST SQUARES

4.1 INTRODUCTION

Partial least squares (PLS) was first introduced in 1960's by Herman Wold [25] and used in the Economics field. PLS has been widely used in chemometrics, econometrics and more recently in neuroimaging due to its dimension reduction capability. It was first introduced to neuroimaging in 1996 by A.R. McIntosh et. al, [13] and has received a lot of attention lately in the neuroimaging community. Partial least squares (PLS) is a technique which is based on a singular value decomposition of the covariance matrix between two blocks of variables and uses the extracted information to obtain a new set of variables that optimally relate the blocks using the fewest dimensions. The neuroimage datasets are very rich databases containing a substantial amount of temporal, spatial and statistical signals. PLS is a technique that provides the researcher with a comprehensive tool capable of explaining the factor variation and models the responses well at the same time.

A first step in PLS is the computation of the covariance matrix for which a Singular Value Decomposition (SVD) is performed. The matrices obtained from the SVD will then be used in calculation of the PLS scores which are the final product of the analysis. The final product of the technique is one, or more, PLS scores for each individual which are computed using one of the resulting matrices from SVD and the value of each variable from the dataset. Thus, missing data in any of these variables is a serious limitation in achieving the final goal of the PLS; computing the scores. The present work will provide a method for computing the PLS scores in the presence of missing data. The complete case analysis technique will result in the elimination of subjects from the analysis and the PLS scores will not be available for subjects with missing data due to elimination of these subjects in the

first place. This is a serious limitation of the complete cases analysis since one will never be able to compute the actual values of the scores due to missing data. To date, there is almost no existing published literature involving calculation of PLS scores resulting from data with missing values. Also, the inverse probability weigthing method cannot be used due to a similar limitation, there will be no data available on subjects with missing observations in order to compute PLS scores. Therefore, only the multiple imputation technique will be able to provide scores for all subjects due to the imputation of the missing values, which will result in complete observations for all subjects. The multiple imputation procedure will provide all of the information needed to arrive at the final product of PLS and it is the only approach to handling missing data that addresses this issue.

An interesting approach that we present here is a combination of inverse probability weighting and multiple imputation. Explicitly, the covariance matrix used as input for the singular value decomposition is calculated using the reconstruction proposed in chapter 3, from the inverse probability weights. We will then use a complete dataset from one of the datasets created using multiple imputation, so that all of the variables will have complete observations. This matrix and the dataset are then used to compute the scores.

A methods section with detailed information regarding the mentioned techniques is presented in section two. Section three consists of a description of simulation studies, section four presents an example using PLS scores obtained from the above mentioned methods and section five presents a short discussion along with further recommendations.

4.2 METHODS

The main step of partial least squares is the computation of a singular value decomposition of the covariance matrix. The singular value decomposition algorithm takes a rectangular matrix of data, $\Sigma_{(n\times p)}$, and finds U, D and V such that:

$$\Sigma_{n\times p} = U_{n\times n} D_{n\times p} V^t_{p\times p}, \tag{4.1}$$

where $U^tU = I_{n\times n}$ and $V^tV = I_{p\times p}$.The columns of V are actually the eigenvectors of $\Sigma^t\Sigma$ and the columns of U are the eigenvectors of $\Sigma\Sigma^t$. To find the eigenvectors of a matrix we use the following definition:

Definition 6. *Let us denote $\Sigma\Sigma^t = W$. For an $n \times n$ matrix W, a nonzero vector x is the eigenvector of W if:*

$$(W - \lambda I)x = 0,$$

for some scalar λ and I being the $n \times n$ identity matrix. Then the scalar λ is called an eigenvalue of Σ, and x is said to be an eigenvector of Σ corresponding to λ.

The singular value decomposition computed for each of the covariance matrices derived by complete cases, multiple imputation and the inverse probability weighted method. The dataset is composed of two matrices, one matrix consisting of independent variables, denoted X, and one matrix consisting of dependent variables, denoted Y. The covariance matrix of X and Y, Σ_{xy}, is computed to be used in the analysis. The singular value decomposition is then performed on $\Sigma_{xy} = Cov(X, Y)$. The SVD will return three different matrices, U, D, V such that $\Sigma = UDV^T$. The goal of the PLS is to find the matrix of $X-scores$, T, and $Y-scores$, B, which are computed using the original dataset combined with the matrices calculated by the SVD algorithm. More specifically, the columns of the matrix U, $u_{11}, u_{12}, ..., u_{1n}$, from the SVD represent the eigenvectors from the decomposition of the product of the covariance matrix and its transpose, $\Sigma_{xy}\Sigma_{xy}^T$, and will be used to calculate the $X-scores$. The columns of V, denoted $v_{11}, v_{12}, ..., v_{1p}$, represent the eigenvectors corresponding to the eigenvalues of the product $\Sigma_{xy}^T\Sigma_{xy}$ and will be used to calculate the $Y-scores$. The matrix of $X-scores$, T, with columns, $t_1, ..., t_i$, where i represents the number of scores one would like to calculate, is computed as follows:

$$t_1 = Xu_{11},$$

and the process is continued until all of the desired scores are calculated. The same process is employed to calculate the matrix of $Y-scores$, B, $b_1, ..., b_j$, where j represents the number of $Y-scores$ to be extracted. The first computed $Y-score$ will be:

$$b_1 = Yv_{11}.$$

The diagonal matrix, D from the SVD algorithm is the matrix of singular values. The diagonal matrix D is given by: $D = diag(d_{11}, d_{22}, d_{33}, ..., d_{nn})$ and the sum of squared singular values is given by: $\sum_{i=1}^{n} d_{ii}^2$. The ratio of a squared singular value, d_{ii}^2 divided by the above sum, $(\frac{d_{ii}^2}{\sum_{i=1}^{n} d_{ii}^2})$, is "the proportion of summed squared cross-block covariance accounted for by the singular vectors". This matrix will give the percentages of variability that are explained by the PLS scores after performing the described calculations.

4.2.1 PLS using complete cases

In the complete cases scenario, only cases where all the variables are observed are used in calculations. More specifically, each entry of the cross-correlation matrix is calculated using:

$$s_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ij} - \overline{x}_i)(x_{jk} - \overline{x}_j), \tag{4.2}$$

where $i = 1, ..., n$, $j = 1, ..., n$. then the SVD algorithm is applied to the specified covariance matrix and the U, D and V matrices are obtained from the SVD. The $X - scores$ are obtained by multiplying the matrix of independent variables with the U matrix obtained from SVD. The number of scores to be calculated is given by the percentage of variability explained by the singular vectors. For each subject, one PLS score will be calculated as follows: $X\_score_i = u_i X$, where u_i represents the first column of the matrix U. The subjects for which some of the observations are missing (x_{ij}) will result in a missing PLS score. Therefore, in this case only subjects with complete observations will have a corresponding PLS score.

4.2.2 PLS using multiple imputation

Partial least squares is performed on the covariance matrix calculated from data with missing values which were filled in using the multiple imputation method previously described. In this case, a number of $m = 5$ imputations were performed for each covariate with missing values. Therefore, instead of one dataset there will be 5 datasets and the covariance matrix, Σ^m, $m = 1, ..., 5$, will be calculated separately for each dataset. Note that the upper subscript denotes the imputation number, m. The SVD will be performed on each covariance

matrix, Σ^m, and 5 different sets of singular values will be obtained. Each covariance matrix will be decomposed into $\Sigma_{n\times p} = U_{n\times n}D_{n\times p}V^t_{p\times p}$ using the SVD algorithm, and the resulting matrices, U,D,V will be used in further calculations of the scores.

In this case, for each complete dataset, a set of scores will be calculated using the corresponding matrix from the SVD decomposition. More specifically, for the first imputation, $m = 1$, with the corresponding covariance matrix Σ^1 with the SVD decomposition, $\Sigma^1 = U^1D^1V^{1t}$, the first corresponding $X - score$ will be $t^1_1 = u^1_{11}X$, where X represents the data matrix with all covariates and u^1_{11} is the first column of matrix U^1 (the columns of the matrix U^1, u_{ij}, represent the eigenvectors corresponding to the eigenvalues of the covariance matrix, Σ^1). In the case where more than one score will be needed, similar calculations will be performed, the only change being the column of U_1 that will be used in the specified calculation (i.e. $t^1_2 = u^1_{12}X$).

Similarly, the covariance matrix corresponding to the second imputation, $m = 2$, Σ^2, is calculated and the SVD is applied to this matrix as well, resulting in U^2, D^2, V^2 and the corresponding first $X - score$, $t^2_1 = u^2_{11}X$. Also, if more than one score is desired, through similar calculations, the second $X - score$ is computed, $t^2_2 = u^2_{12}$. This algorithm is repeated for all $m = 5$ completed datasets. Thus a total of 5 sets of one $X - score$, t^1_1, t^2_1, t^3_1, t^4_1, t^5_1 will be calculated, one for each imputation. An average of all scores for all $m = 5$ imputations could also be computed for the $X - scores$ resulting in only one set of averaged scores for X. Specifically, $t_{mean} = \frac{1}{m}\sum_{m=1}^{5}(t^m_1)$ will denote the average of the resulting scores using multiple imputation.

4.2.3 PLS using inverse probability weighting

Inverse probability weighting is a new approach used in this context for calculation of PLS scores in the presence of missing data. The covariance matrix is calculated using inverse probability weighting and it is further used in singular value decomposition algorithm. Let us consider the cross-covariance matrix of X and Y where X represents the data matrix consisting of all independent variables and Y is the matrix consisting of all dependent variables,

and let X_1 be the covariate with missing values:

$$\Sigma = \begin{bmatrix} Cov(X_1, Y_1) & Cov(X_2, Y_1) & Cov(X_3, Y_3) \\ Cov(X_1, Y_2) & Cov(X_2, Y_2) & Cov(X_3, Y_3) \\ Cov(X_1, Y_3) & Cov(X_2, Y_3) & Cov(X_3, Y_3) \end{bmatrix}. \tag{4.3}$$

The entries of the covariance matrix Σ will be calculated using the inverse probability weighting technique since the variable X_1 is not fully observed. Therefore, each entry will be calculated using the estimated correlations described in Chapter 3:

$$\hat{r}_{xy} = \frac{\sqrt{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i}(x_i - \bar{x})^2 \sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i}(y_i - \bar{y}_i)^2}}{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i}(y_i - \bar{y}_i)^2} \frac{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} \frac{1}{\hat{\Pi}_i}(x_i - \bar{x})^2}. \tag{4.4}$$

The SVD is performed on this covariance matrix, Σ, resulting in the U, D and V matrices that will further be used in calculations of the scores. The first $X - score$ will be: $X - score = u_i X$, where u_i is the first column of matrix U (the columns of matrix U, are the eigenvectors corresponding to the eigenvalues of the decomposition of the covariance matrix, Σ). Similarly, the first $Y - score$ could also be calculated as: $Y - score = v_i X$, where v_i is the first column of the matrix V (the columns of matrix V, are the eigenvectors corresponding to the eigenvalues of the decomposition of the covariance matrix, Σ). The PLS scores computed here in this dissertation were only the $X - scores$ since those were the scores involving the covariate with missing values. As already described in the context for the complete cases and multiple imputation, the ratio of a squared singular value, d_{ii}^2 divided by the above sum,$(\frac{d_{ii}^2}{\sum_{i=1}^{n} d_{ii}^2})$, is "the proportion of summed squared cross-block covariance accounted for by the singular vectors" is valid here as well. This matrix will give the percentages of variability that is explained by the PLS scores after performing the described calculations.

An interesting combination that was performed here was the combination of the cross-correlation matrix calculated using the inverse probability method and then computation of the scores, using this matrix for the SVD and using one dataset with all of the observations completed resulting from the multiple imputation. The SVD is performed on the cross-correlation computed using inverse probability weighting. Since the inverse probability

weighting will not substitute any value in the dataset, in order to have complete observations for all subjects, one dataset with imputed values obtained by performing multiple imputation was used in the computation of scores.

The cross-correlation is computed using formula 4.4 from this chapter and then the SVD is performed on this matrix. The resulting U, D and V matrices are computed along with the percentages of variability explained by the singular vectors that results from $(\frac{d_{ii}^2}{\sum_{i=1}^{n} d_{ii}^2})$. The first $X - scores$ will be then calculated as: $t_i = u_i X^j$, where u_i represents the first column of the matrix U and X^j represents the vector of covariates with missing observations substituted by imputation, (j represents the imputation index). In this work, even though 5 imputations were performed, the IPWMI scores were calculated using only one of the datasets from the imputation.

In order to quantify the agreement between all of the methods (complete cases, multiple imputation, inverse probability weigthing and inverse probability weighting combined with multiple imputation) used for computing the scores, Kendall's concordance coefficient, denoted by W was computed. The mathematical formula used for computing Kendall's concordance coefficient (W) is given by the following expression [2]:

$$W = \frac{\sum_{i=1}^{n}(R_i - \frac{k(n+1)}{2})^2}{\sum_{i=1}^{n}(ik - \frac{k(n+1)}{2})^2},$$

where k represents the technique (i.e. CC, MI, IPW, IPWMI), n represents the objects that are ranked, (i.e. PLS scores), and R_i represents the sum of the ranks given to the specific objects by the k techniques.

4.3 SIMULATIONS

In this case, partial least squares was performed using datasets previously created to compute covariance matrices. Some of the covariance matrices simulated under scenarios described in chapter three, were used in the singular value decomposition. For the first

set of simulations, each dataset had a sample size equal to 100 and we simulated 1000 datasets. The covariance matrices previously calculated were based on different percentages of missingness (i.e. 20%, 30%, 50%), therefore the same percentages of misingness will be used in this context. Calculations of the PLS scores were performed using the methods previously described: complete cases, multiple imputation, inverse probability weighting and inverse probability weighting combined with multiple imputation. For the complete case method, for each dataset the covariance matrix was calculated and then SVD was performed on each covariance matrix of the 1000 simulations, resulting in one set of scores for each dataset. Therefore, a total of 100 scores were calculated for each dataset and the estimated PLS score was the average of the 1000 datasets. Therefore for each subject, the average PLS score will be:

$$X\_score_{avg-cc} = \frac{1}{1000} \sum_{i,k=1}^{1000,100} X\_score_{ik}, \tag{4.5}$$

where i represents the dataset and k represents the subject index, $i = 1, ..., 1000$, $k = 1, ..., 100$. In the case of multiple imputation, for each imputation one set of scores will be calculated for each dataset, that is one set of averaged scores for each imputation. Mathematically, this could be written as:

$$X\_score^{j}_{mi} = \frac{1}{1000} \sum_{i,k=1}^{1000,100} X\_score^{j}_{ik}, \tag{4.6}$$

where the subscript i represents the dataset, k the subject index and j represents the imputation, here $j = 1, ..., 5$. If an averaged score across imputations was needed it would be calculated using:

$$X\_score_{avg-mi} = \frac{1}{5} \sum_{j=1}^{5} X\_score^{j}_{avg-mi}. \tag{4.7}$$

The averaged scores across imputations were not computed here. The scores computed in this dissertation were obtained from each imputation separately and were not combined in an average score across the multiple imputations. The third method, inverse probability weighting is similar to complete cases when calculating the PLS scores. The only difference between the two is the calculation of the covariance matrix. In this case the covariance matrix

is calculated by weighting each observation by the inverse probability of being observed, as previously described. Therefore, the scores will be calculated using the following formula:

$$X\_score_{avg-ipw} = \frac{1}{1000} \sum_{i,k=1}^{1000,100} X\_score_{ik}. \tag{4.8}$$

The last method used to calculate scores was inverse probability weighting combined with multiple imputation. The only difference between inverse probability weighting and this method is the use of one of the datasets with missing values substituted using imputation. Mathematically, scores will be calculated using the following formula:

$$X\_score_{avg-ipwmi} = \frac{1}{1000} \sum_{i,k=1}^{1000,100} X^j\_score_{ik}, \tag{4.9}$$

In this case, the subscript j denotes the imputation dataset that was used.

4.3.1 Simulation results

The SVD was computed for the each of the cross-correlation matrices obtained using complete cases, multiple imputation and inverse probability weighting. As previously described, the diagonal matrix, D from the SVD algorithm is the matrix of singular values. The ratio of a squared singular value, d_{ii}^2 divided by sum,$(V_i = \frac{d_{ii}^2}{\sum_{i=1}^{n} d_{ii}^2})$, gives the percentages of variability that are explained by the singular values. For the first set of simulations, the averaged percentages of variability computed by each method are presented in Table 14. In this case scenario, 20% of the data was missing for one covariate.

The most important thing that we would like to mention in this case, is the fact that the resulting scores are the averaged scores for each subject across all simulations. When only one dataset is available, such computations are not possible for complete cases and inverse probability weighting, and in that case (as we will further describe using the dataset example) not all of the scores for all subjects can be computed. The missing percentage of scores for each subject was roughly 20%. Also, since only 64% of variability is explained by the first set of x-scores, it is necessary to compute the second set of x-scores. The PLS scores for the situation when the percentages of missingness were 30% and 50% percent

Table 14: Variability accounted for by the singular values, 20% missing

Method/Percent	V_1	V_2	V_3
Percent (%) true	64.32	26.61	9.07
Percent (%) CC	64.11	26.71	9.18
Percent (%) MI1	64.08	26.72	9.18
Percent (%) MI2	64.09	26.72	9.18
Percent (%) MI3	64.09	26.71	9.19
Percent (%) MI4	64.04	26.77	9.18
Percent (%) MI5	64.10	26.70	9.19
Percent(%) IPW	64.06	26.75	9.17

respectively were computed for this scenario and the percentages of variability explained by singular values for each method are displayed in Tables 15 and 16.

Kendall's concordance coefficient was calculated to measure the agreement between the first set of x-scores obtained using complete cases, multiple imputation, inverse probability weighting, inverse probability weighting combined with multiple imputation and the true scores. The computed value of Kendall's concordance coefficients are presented in Tables $17 - 19$. All of the corresponding p-values were $< .01$. We can conclude that there is strong agreement between the scores calculated using the above mentioned techniques. Also, the agreement between all of the imputations was measured using Kendall's concordance coefficient (W) and the value was .99. This was expected since the variability explained by the singular vectors using multiple imputation are very close to each other. Furthermore, the agreement between multiple imputation and the true values of the PLS scores was also investigated. The Kendall's computed values were above .80 for all percentages of missing values (20%, 30%, 50%).

Another scenario that we have considered was the computation of all PLS scores for all datasets together. More specifically, the computed score for each subject, $X - score_{ikj}$ (where i represents the dataset, k represents the subject index and j represents number of

Table 15: Variability accounted for by the singular values, 30% missing

Method/Percent	V_1	V_2	V_3
Percent (%) true	64.32	26.61	9.07
Percent (%) CC	63.91	26.92	9.15
Percent (%) MI1	63.88	26.94	9.16
Percent (%) MI2	63.89	26.95	9.16
Percent (%) MI3	63.91	26.92	9.16
Percent (%) MI4	63.88	26.94	9.16
Percent (%) MI5	63.89	26.94	9.16
Percent(%) IPW	63.83	27.02	9.13

Table 16: Variability accounted for by the singular values, 50% missing

Method/Percent	V_1	V_2	V_3
Percent (%) true	64.32	26.61	9.07
Percent (%) CC	64.27	26.48	9.23
Percent (%) MI1	63.91	26.91	9.16
Percent (%) MI2	63.92	26.91	9.17
Percent (%) MI3	63.90	26.92	9.16
Percent (%) MI4	63.94	26.87	9.17
Percent (%) MI5	63.90	26.92	9.16
Percent(%) IPW	64.10	26.69	9.19

the imputation, $i = 1, ..., 1000$, $k = 1, ..., 100$, $j = 1, ..., 5$), using the multiple imputation method was compared with the $X - score_{ik}$ computed from the dataset before creating the missing values.

In this case, the inverse probability weighting method and complete cases could not

Table 17: Kendall's W for averaged scores across simulations for MI and truth

missing percentage	Kendall's W first $x - score$	Kendall's W second $x - score$
20%	.97	.98
30%	.92	.90
50%	.81	.80

Table 18: Kendall's W for averaged scores across simulations for CC and truth

missing percentage	Kendall's W first $x - score$	Kendall's W second $x - score$
20%	.70	.70
30%	.62	.60
50%	.47	.48

Table 19: Kendall's W for averaged scores across simulations for IPW and truth

missing percentage	Kendall's W first $x - score$	Kendall's W second $x - score$
20%	.70	.71
30%	.64	.63
50%	.52	.48

be evaluated due to observations with missing values in the computed scores from each of these two methods. As described before, a combination of inverse probability and multiple imputation is also evaluated. Since multiple imputation provides a dataset with no missing values, we have combined one multiple imputation dataset and singular value decomposition

performed on the correlation matrix computed using inverse probability weighting.

To evaluate the agreement between multiple imputation, inverse probability weighting combined with multiple imputation and a dataset without missing observations, Kendall's concordance coefficient (W) was calculated. The Kendall's concordance correlation coefficients, $W's$, and their associated p-values for this case scenario are presented in Table 22. Table 20 provides the results associated with this scenario, with 20% missing data. Additional tables with percentages of missingness equal to 30% and 50% percent are presented in Table 21.

Table 20: Variability explained by the singular vectors, second scenario, 20% missing

Method/Percent	V_1	V_2	V_3
Percent (%) no missing	61.13	38.00	.84
Percent (%) MI1	60.96	38.18	.80
Percent (%) MI2	60.92	38.23	.84
Percent (%) MI3	60.91	38.23	.80
Percent (%) MI4	60.98	38.13	.85
Percent (%) MI5	61.10	38.00	.80
Percent (%) IPWMI1	61.20	38.00	.80

Table 21: Variability explained by the singular vectors, second scenario, 30%, 50% missing

Method/Percent	V_1	V_2	V_3
Percent (%) no missing	61.13	38.00	.84
Percent (30%) MI1	61.05	38.10	.84
Percent (50%) MI1	61.05	38.20	.86
Percent (30%) MI2	61.09	38.04	.85
Percent (50%) MI2	61.07	38.05	.86
Percent (30%) MI3	60.94	38.20	.84
Percent (50%) MI3	61.09	38.05	.85
Percent (30%) MI4	60.98	38.14	.85
Percent (50%) MI4	61.07	38.06	.86
Percent (30%) MI5	61.10	38.21	.84
Percent (50%) MI5	61.03	38.10	.86
Percent (30%) IPWMI1	61.26	37.90	.82
Percent (50%) IPWMI1	61.60	37.60	.80

Table 22: Kendall's W for multiple imputation and data with no missing values

missing percentage	Kendall's W first $x - score$	Kendall's W second $x - score$
20%	.955	.918
30%	.958	.901
50%	.948	.878

4.4 DISCUSSION

The most important observation that we can make here is the fact that when 20% of the data is missing, the percentages of variability explained by the singular values computed using complete cases, multiple imputation and inverse probability weighting are very similar. Each method gives the same percentage of variability, therefore in this case we could certainly conclude, as expected, that there is not much difference between the three methods. When the percentages of missingness increase a similar trend is observed as well, the agreement is a little lower when 50% of the values in the dataset are missing as opposed to when only 20% or 30% are missing. Since the cross-correlation matrices were very similar, we expected to see these results.

In the first case scenario when the average score is computed across simulations for complete cases, inverse probability weighting and true value, the Kendall's concordance coefficient is smaller (.72, respectively .74) when compared to the situation where no average is performed. This is explained by the fact that we are comparing the averaged scores across simulations, and in the case of complete cases and inverse probability weighting there are 20%, 30% and 50% missing values for the averaged score. These values are expected given the presence of missing values in our scores.

The second case scenario where the scores across all simulated datasets were computed using multiple imputation and compared with the scores obtained before deleting the data, we observed a very good agreement, above .90, which was also expected. In this case, complete cases and inverse probability weighting was not investigated since the missing values would have been present. As an alternative, the inverse probability weighting combined with multiple imputation was examined. Good agreement was also observed when this was included in the calculation of Kendall's concordance coefficients. This did not come as a surprise either, since one of the multiple imputation datasets was used to compute the scores and the singular values from the SVD performed on the cross-correlation were obtained from inverse probability weighting.

As a conclusion to our research, our recommendation would be to use the multiple imputation method to calculate PLS scores, since it is the only available option. The inverse

probability method is a nice alternative if no more than variability percentages are required by the researcher. The combination of the two, is an interesting thing that was explored here, but it does not provide any additional benefits since it uses both multiple imputation and inverse probability weighting. While multiple imputation is an intensive computational technique, it is still the only candidate for handling missing data in this setting.

Histogram plots for each subject across simulations are attached in the appendix. Only 25 subject plots were included for illustration, since the plots are similar across simulations. The percentages of missing values were roughly 20% for each subject for the plots presented in the Appendix section of this chapter.

4.5 POSITRON EMISSION TOMOGRAPHY EXAMPLE; TRAILS

The relationship between brain metabolism and performance on the Trail Making Test (TMT) was examined in our studied population in order to detect an indication of cognitive dysfunction. The TMT test explores visual-conceptual and visual-motor tracking [3]. The TMT test was first developed in 1944 for the War Department of the US Army and has been widely used as a neuropsychological test consisting of two parts. Part A consists of connecting consecutive numbered circles by the subject going under examination. Part B of the test consists of alternatively connecting numbered and lettered circles. The score of the test is recorded as the number of seconds it takes to the subject to perform the required task. The block of dependent variable consists of several brain region-of-interest, all of them completely observed. A region-of-interest is a selected subset of samples within a dataset. In this example a region-of-interest represents a specific region of the brain. The regions are identified prior to analysis. The regions-of-interest (ROI) under investigation are: Frontal Cortex (FRC), Lateral Temporal Cortex (LTC), Dorsal Frontal Cortex (DFC), Parietal (PAR), Occipital (OCC), Anterior Cingulate Gyrus (ACG).

4.5.1 Methods

The research study under investigation included 111 subjects classified into three different groups. Some of these subjects had values missing for the Trail A test. It has been shown that there is a relationship between cerebral impairment and Trail A test results [4]. The three groups were the healthy control group (Control or non-disease group), the Alzheimer disease (AD) group and the mild cognitive impairement (MCI) group. The control group consists of 67 subjects, 22 males (age 74 ± 10) and 45 females (age 71 ± 9). In the MCI group the number of subjects was 25 with 19 males (age 71 ± 8) and 6 females (age 67 ± 7) and the AD group consists of 19 subjects, 14 males (age 70 ± 10) and 5 females (age 72 ± 8.3). A summary of the Trail A measurements divided by subject groups is provided in Table 23.

The cross-correlation matrix was computed using the method of multiple imputation, inverse probability weighting and complete cases and the results are presented in Table 24.

Table 23: Trail Making Test A summary by group and gender

Group	Gender	Age	Trail A	missing
AD(15)	M(12)	69.9±10.61	48.72±16.58	
	F(3)	71.8±8.31	42.33±10.06	4
MCI(23)	M(17)	70.9±8.70	37.35±16.24	
	F(6)	67.16±6.36	26.83±4.62	2
Controls(61)	M(19)	73.5±10.01	34.78±11.24	
	F(42)	70.53±9.19	28.04±9.47	6

The cross-correlation will be used in PLS. To calculate the probability of being observed a logistic model was fit to the data and variables with a p-value less than .20 were kept in the model. In this case, three variables were selected, age, FRC (Frontal Cortex) and LTC (Lateral Temporal Cortex) in the final model $Pr(R = 1) = -3.05 + .05 * Age - 5.52 * FRC + 7.9 * LTC$. No other variable met the selection criteria.

Table 24: Correlations of Trail A with each region-of-interest

Name	ACG	DFC	FRC	LTC	OCC	PAR
CC_Trail A	.115	−.108	−.121	−.245	.067	−.142
MI_Trail A	.110	−.108	−.120	−.243	.067	−.144
IPW_Trail A	.112	−.110	−.122	−.246	.066	−.144

A summary of the multiple imputation for the Trails A given in Table 25. The minimum and maximum values for the Trail A variable remained the same before and after the imputation for each group. The minimum value for the Trail A variable for the AD group was equal to 26 before imputation and after imputation and maximum value was equal to

Table 25: Average values for multiple imputation for Trail A variable

Imputation by Group	MI1	MI2	MI3	MI4	MI5	mean before
AD	45.47	45.54	47.02	47.68	47.68	47.35
Control	30.68	30.67	30.27	30.46	30.11	30.14
MCI	34.09	34.58	34.59	34.10	35.00	34.60

83 before and after imputation. The minimum value for the Trail A variable for the Control group was equal to 13.85 before and after imputation and a maximum equal to 60.06 before and after imputation. Also, for the MCI group, the minimum of 21 and maximum of 72 before imputation remained unchanged after imputation.

The singular value decomposition was applied to the cross-correlation matrix between age, Trail A and all regions-of interest. The partial least squares scores were computed using the cross-correlation calculated using complete cases, multiple imputation and inverse probability weighting. The scores calculated using complete cases and the inverse probability weighting method were compared using the Kendall's correlation ($\tau = .98$). The Kendall's concordance coefficient value was computed to verify agreement between all multiple imputation and multiple imputation combined with inverse probability weighting and the value was W=.99 with a p-value $< .00001$.

The inverse probability method, is then combined with one of the multiple imputation datasets and therefore the scores will no longer have missing values. Table number 26 provides the percentages of variability explained by the singular vectors.

The variability is very similar across all methods. Since the proportion of the summed squared cross-correlation accounted for by the first singular vector is higher than 83%, only one $x - score$ is needed. The matrix of singular vectors related to the covariates that is computed from the singular value decomposition is very similar across the methods. The singular vectors are the columns of the matrix V from the singular value decomposition

Table 26: Variability explained by singular vectors for Trail A example; all methods

Method	V1	V2
CC	83.88%	16.11%
IPW	84.15%	15.84%
MI1	83.05%	16.9%
MI2	83.15%	16.84%
MI3	83.85%	16.14%
MI4	84.37%	15.62%
MI5	84.20%	15.79%

($SVD = UDV^T$). These vectors indicate the covariates that are most related to the singular values. In this case, the covariate age (average value of the singular value across methods equal to $-.97$), is more related to the first singular value and Trail A is most related to the second singular value (average value of singular vector across methods equal to .98).

Similarly, the singular vectors that will give an indication of which region-of-interest is most related to the singular values are the columns of the matrix U from the singular value decomposition ($SVD = UDV^T$). The higher the value (referred also as weight) in that specific column of the matrix, the stronger the relationship between that specific region and the singular values. In this example, the variable that was most related to the the first singular value was the parietal (average value of singular vectors across methods equal to $-.52$), followed by frontal cortex ($-.48$). Since this singular value is most related to age as well, the interpretation would be that these are the regions that differ most with age.

The region-of-interest with the highest value (weight) in the second column of the matrix U computed using SVD was anterior cingulate gyrus (.62) followed by occipital (.62). Therefore these regions differ most for the Trail A. The graphs that are included here represent the scores computed using complete cases and inverse probability weighting plotted versus Trail A test values for each group (figure 1, 3, 5). Similarly, figures with plots of the

scores computed using multiple imputation and inverse probability weighting and multiple imputation versus Trail A test values are provided as well (figure 2, 4, 6).

4.6 ADDITIONAL TABLES FOR TRAILS A DATA EXAMPLE

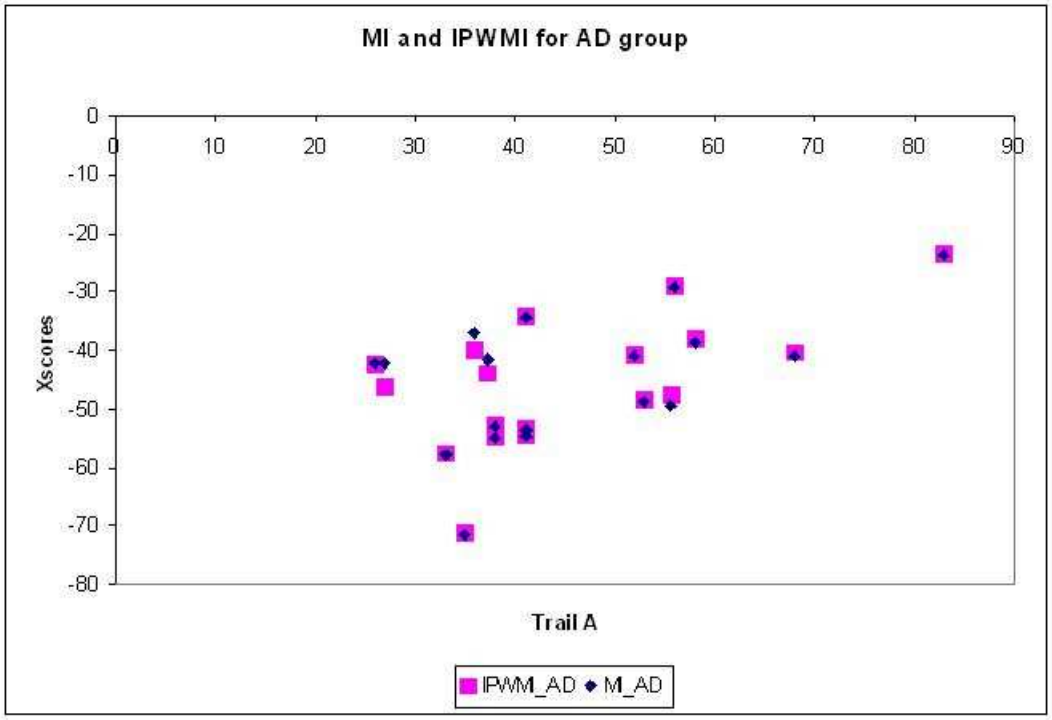

Figure 1: X-scores for AD group using MI and IPWMI method

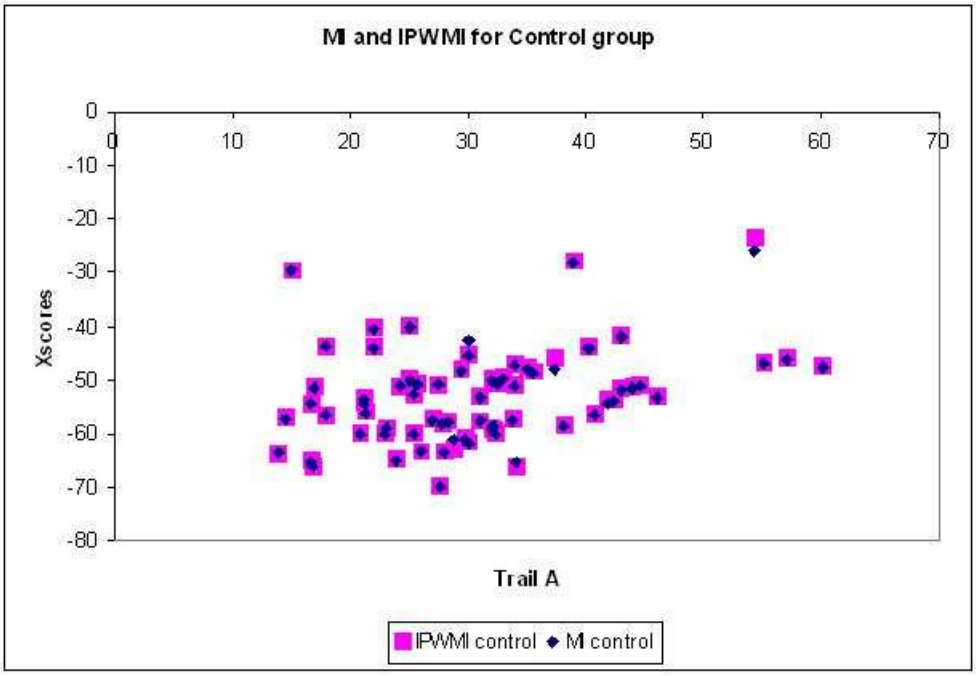

Figure 2: X-scores for Control group using MI and IPWMI method

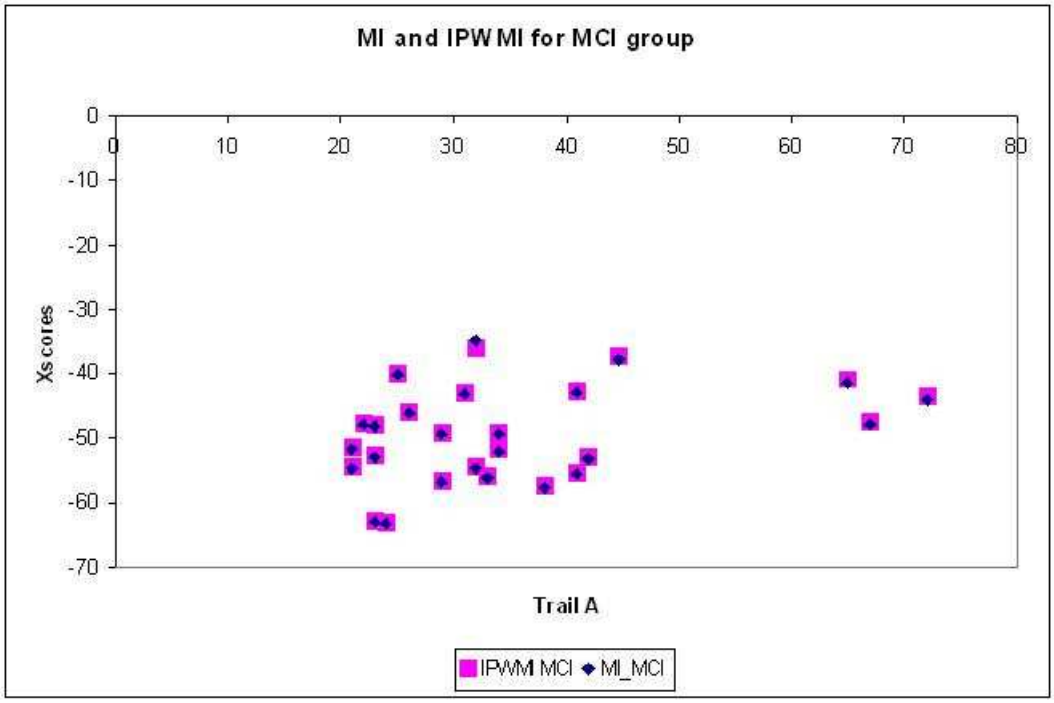

Figure 3: X-scores for MCI group using MI and IPWMI method

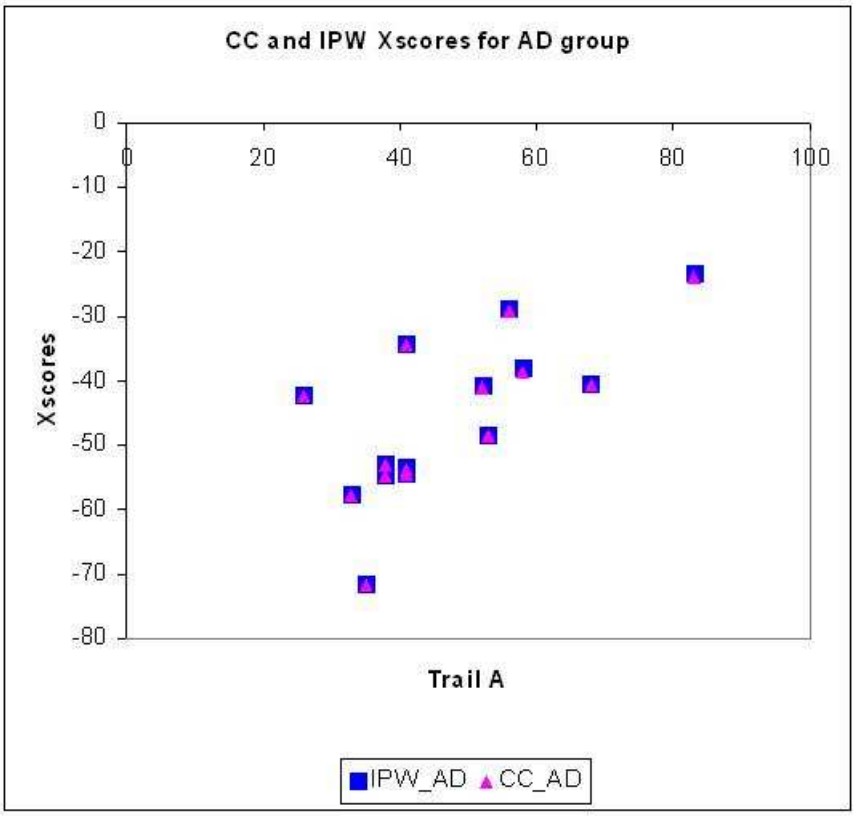

Figure 4: X-scores for AD group using CC and IPW method

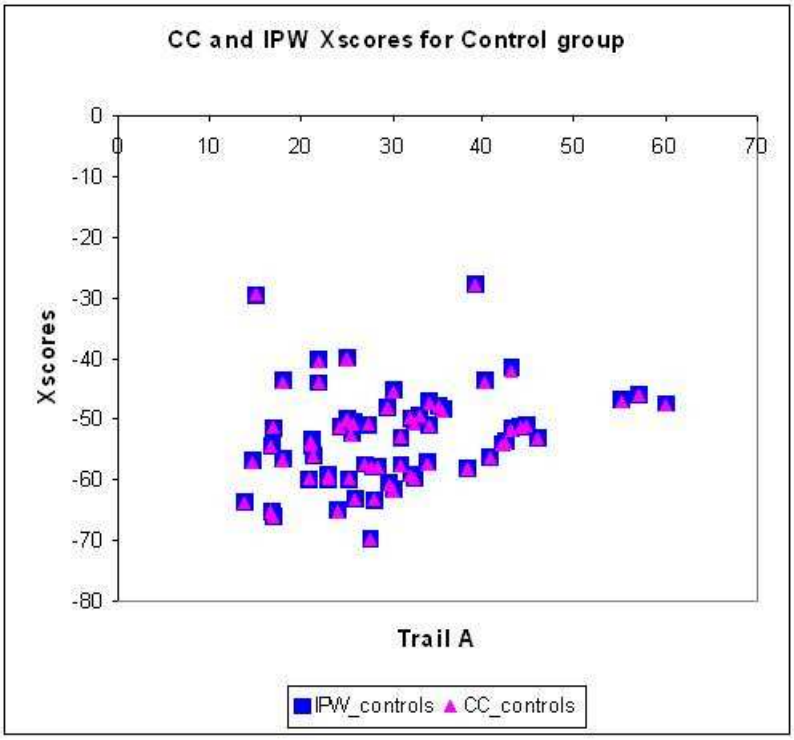

Figure 5: X-scores for Control group using CC and IPW method

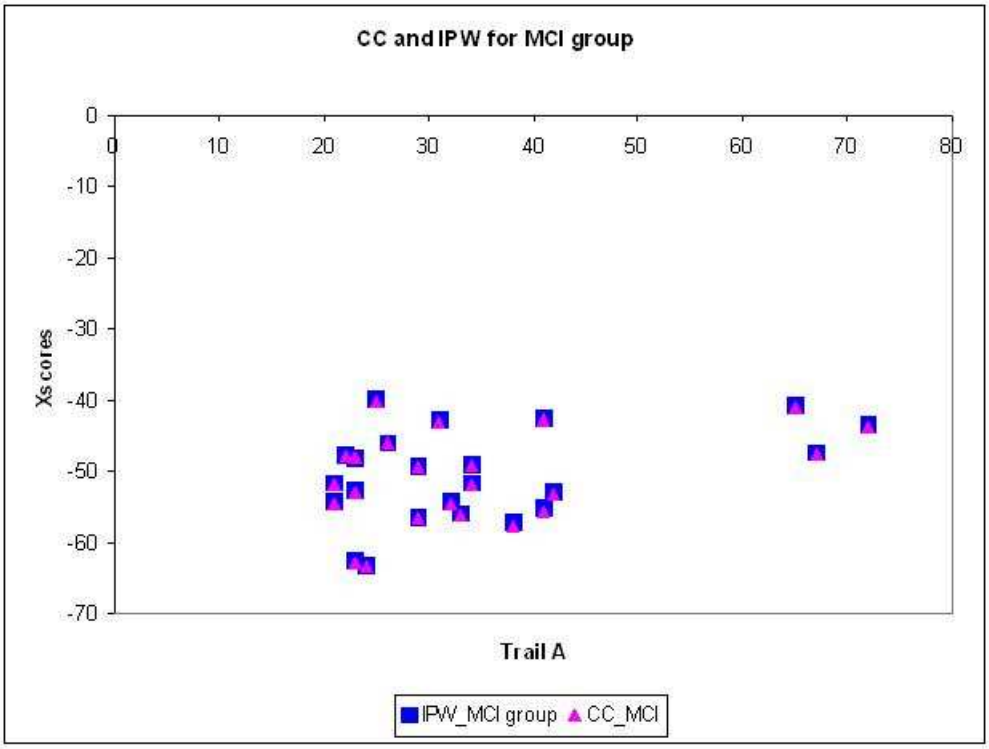

Figure 6: X-scores for MCI group using CC and IPW method

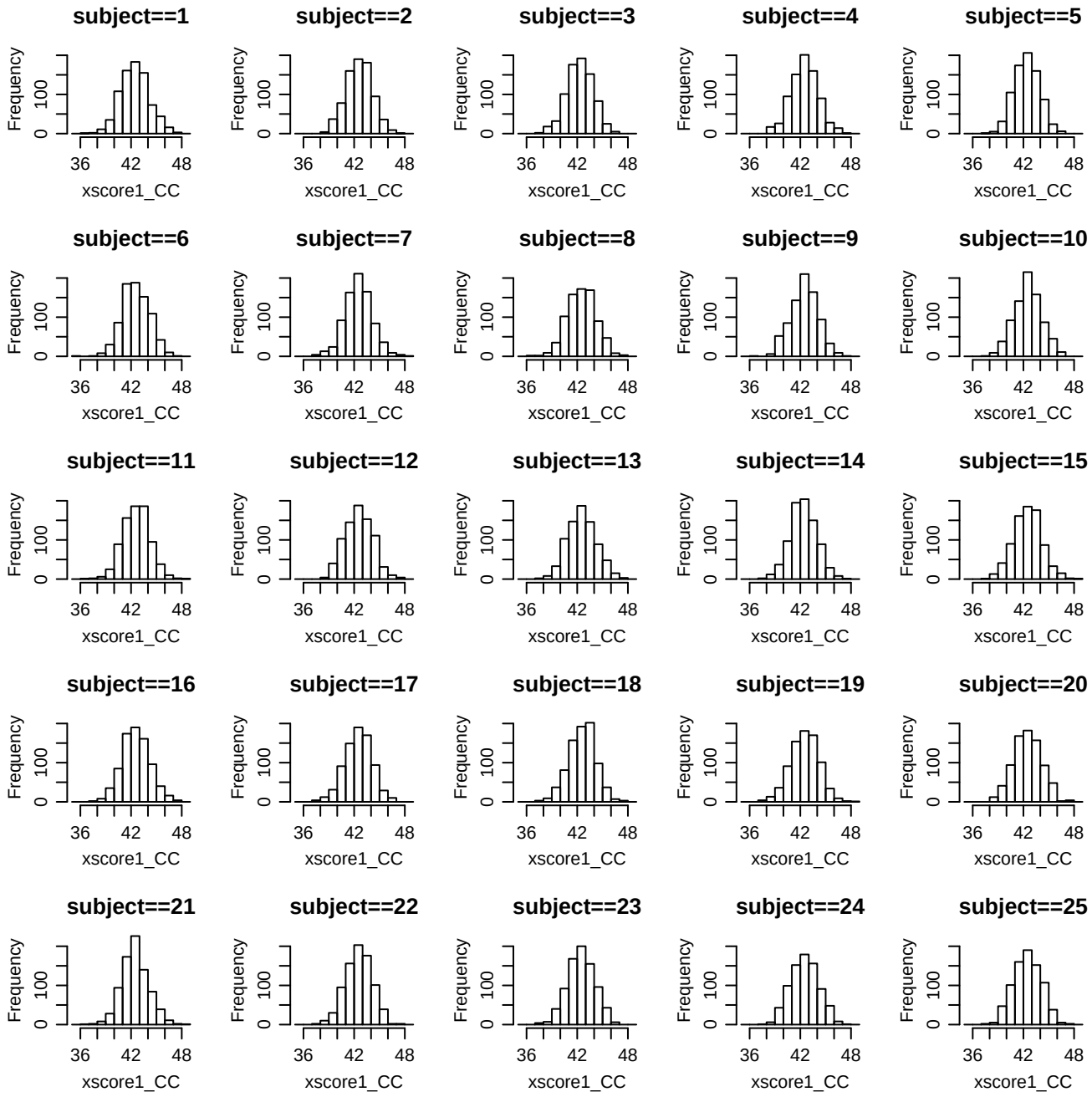

Figure 7: First X-score set: 25 subjects

5.0 DISCUSSION AND FUTURE WORK

The methods for the estimation of the cross-correlation matrix and partial least squares (PLS) scores presented in this dissertation were developed in the presence of one covariate with missing values. The problem is partially solved at the level of one covariate with missing values, in the continous case under the assumption of multivariate normality of the dataset under investigation. Extensions of this methodology, inverse probability weighting, in the case when more than one covariate with continuous measurements is missing would be the next step. The development of such an extension has came to our attention while working on these problems. To extend the present technique to more than one covariate with missing values requires further investigation.

The first step considered would be the estimation of the probability of missingness and then the computation of cross-correlation from a linear regression with weights based on the estimated probabilities of being observed, when these probabilities can no longer be computed from a Bernoulli distribution. To estimate the probability of missing values for only one missing covariate in the continuous case, one establishes an indicator function, as defined previously, R_i, that can only take two values, either a 0 or a 1, corresponding to either observed (i.e 0) or missing (i.e 1). Thus, via classical logistic regression with the indicator variable R_i as response variable, the estimated probabilities of being observed are computed. This will have to be adapted for the case of more than one missing covariate. More specifically, let us briefly introduce the missing scenario for two continuous covariates. Let us consider $X = (X_{complete}, Z)$, where $X_{complete}$ consists of variables that are completely observed and $Z = (Z_1, Z_2)$ that are not always observed. Obviously, different possibilities ought to be considered; one possibility would be for (Z_1, Z_2) to be either missing or observed at the same time; another possibility would be for (Z_1, Z_2) to have missing values alternatively, with no

obvious pattern or for (Z_1, Z_2) to have missing values based on a specific pattern.

Let R_j be the indicator variable, which in this case will no longer have a Bernoulli distribution but a multinomial distribution with four parameters that need to be estimated, p_1, p_2, p_3, p_4. The indicator variable would be defined as follows:

$$R_i = \begin{cases} 1, & \text{if } Z_1 \& Z_2 = observed, \\ 2, & \text{if } Z_1 = missing \& Z_2 = observed, \\ 3, & \text{if } Z_2 = missing \& Z_1 = observed, \\ 0, & \text{if } Z_1 \& Z_2 = missing. \end{cases}$$

In this case, a multinomial logistic regression would be used to estimate each possible probability. The probability of being observed for each of the $Z_i's$ is estimated based on a conditional probability function. The probability of being observed will be of this form, $Pr(Z_1 = 1|Z_2, X) = Pr(Z_1|Z_2 = 1, Z2 = 3, X)$ as opposed to the case of only one covariate missing when $Pr(Z_1) = Pr(Z_1|X)$. Once the probabilities of being observed are calculated, the weights would be computed as well and the methodology presented should be adapted accordingly to the statistical analysis. Also, another possibility will arise from the mechanism of the missing data. Throughout this work, the missingness mechanism, (Π_i) considered was an ignorable missing mechanism. When an ignorable missing mechanism is considered, the missing values for that specific variable are independent of the values that the variable could take. In the case of a non-ignorable missing mechanism, a specific mathematical model for the missing data mechanism has to be considered while performing the statistical analysis.

This work should also be extended to the longitudinal case scenario. Many of the neuroimaging studies are performed over an extended period of time and covariates as well as response variables are measured over time. It would be interesting to adapt the method to the longitudinal case and to study its behavior.

APPENDIX

ADDITIONAL SIMULATION TABLE

The cross-correlation matrices are calculated using complete cases, inverse probability weighting and multiple imputation. These matrices are for the first case scenario of simulations where missingness is independent of x and y, 20% of the data points are missing. The difference between the true correlation matrix entries and the entries of the matrices computed using complete cases, inverse probability weighting and multiple imputation for the covariates that did not have any missing values is small. Their values are identical for almost all of the entries in the matrix. For the entries where there is some difference, the magnitude of the difference is very small (i.e. .001).

Table 27: Correlations from MI, IPW and CC when the missing data is independent of x and y, first correlation structure

Name	y_1	y_2	y_3	x_2	x_3
true_value	0.337	-0.228	-0.411	-0.260	-0.591
method					
CC 30%	0.335	-0.231	-0.408	-0.259	-0.585
MI 30%	0.338	-0.230	-0.407	-0.260	-0.589
IPW 30%	0.336	-0.227	-0.409	-0.261	-0.586
CC 50%	0.335	-0.221	-0.403	-0.258	-0.585
MI 50%	0.338	-0.229	-0.406	-0.260	-0.589
IPW 50%	0.336	-0.229	-0.403	-0.261	-0.587
Bias					
CC 30%	-0.002	-0.003	0.003	0.001	0.006
MI 30%	-0.003	0.002	0.004	-0.000	0.002
IPW 30%	-0.001	0.001	0.002	-0.001	0.005
CC 50%	-0.002	0.007	0.009	0.002	0.006
MI 50%	-0.001	0.001	0.006	0.001	0.002
IPW 50%	-0.001	0.001	0.008	-0.001	0.004
MSE					
CC 30%	0.012	0.013	0.010	0.013	0.007
MI 30%	0.011	0.011	0.006	0.010	0.006
IPW 30%	0.012	0.012	0.007	0.010	0.007
CC 50%	0.016	0.019	0.013	0.018	0.010
MI 50%	0.010	0.012	0.012	0.013	0.006
IPW 50%	0.012	0.012	0.009	0.013	0.007

Table 28: Correlations from MI, IPW and CC, when the missing data is independent of x and y, second correlation structure

Name	y_1	y_2	y_3	x_2	x_3
true_value	-0.548	-0.390	0.216	-0.578	0.140
method					
CC 20%	-0.546	-0.391	0.214	-0.572	0.141
MI20%	-0.546	-0.390	0.213	-0.577	0.140
IPW20%	-0.548	-0.391	0.216	-0.579	0.141
CC 30%	-0.549	-0.391	0.213	-0.579	0.142
MI 30%	-0.547	-0.390	0.212	-0.576	0.141
IPW30%	-0.549	-0.391	0.216	-0.579	0.143
CC 50%	-0.544	-0.380	0.222	-0.573	0.132
MI 50%	-0.543	-0.376	0.217	-0.572	0.132
IPW 50%	-0.544	-0.382	0.220	-0.573	0.133

Table 29: Bias, MSE from MI, IPW and CC, when the data is independent of x and y, second correlation structure

Name	y_1	y_2	y_3	x_2	x_3
Bias					
CC 20%	0.002	-0.001	-0.002	0.006	0.001
MI 20%	0.002	0.000	-0.003	0.001	0.000
IPW 20%	0.000	-0.001	0.000	-0.001	0.001
CC 30%	-0.001	-0.001	-0.003	0.001	0.002
MI 30%	0.001	0.000	-0.004	-0.000	0.001
IPW 30%	-0.001	0.001	0.000	0.001	0.000
CC 50%	0.004	0.010	0.006	0.005	-0.008
MI 50%	0.005	0.014	0.001	0.006	-0.008
IPW 50%	0.004	0.008	0.004	0.005	-0.007
MSE					
CC 20%	0.007	0.010	0.011	0.006	0.012
MI 20%	0.006	0.010	0.009	0.006	0.012
IPW 20%	0.006	0.009	0.010	0.006	0.008
CC 30%	0.007	0.010	0.013	0.006	0.014
MI 30%	0.008	0.011	0.013	0.007	0.014
IPW 30%	0.007	0.010	0.011	0.006	0.010
CC 50%	0.010	0.017	0.020	0.010	0.020
MI 50%	0.009	0.012	0.014	0.010	0.014
IPW 50%	0.010	0.016	0.016	0.010	0.011

Table 30: Correlations from MI, IPW and CC when the missing data is dependent on x, 30% missing values

Name	y_1	y_2	y_3	x_2	x_3
Bias					
CC (30%)	-0.010	-0.006	0.011	0.023	0.034
MI (30%)	0.000	-0.007	-0.005	-0.006	-0.003
IPW (30%)	-0.001	-0.005	0.003	0.001	0.006
MSE (30%)					
CC (30%)	0.009	0.008	0.013	0.008	0.009
MI (30%)	0.008	0.007	0.011	0.007	0.007
IPW (30%)	0.007	0.007	0.010	0.006	0.005

BIBLIOGRAPHY

[1] N.M Rubin D.B Dempster, A.P Laird. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society*, 39:1—38, 1977.

[2] Jean D. Gibbons. *Nonparametric Measures of Association*. SAGE Publications Inc., Newbury Park, CA, 1993.

[3] M. Mascheroni S. Simoncelli M. Laiacona M. Capitani E. Giovagnoli, A.R. Del Pesce. Trail making test: normative values from 287 normal adult controls. *Italian Journal of Neorological Science*, 17:305—309, 1996.

[4] Charles J. Osmon Moses Berg Golden. *Interpretation of the Halstead-Reitan Neuropsychological Test Battery: A Casebook Approach*. Grune & Stratton, NY, 1981.

[5] Y. Haitovsky. Missing data in regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30:67—82, 1968.

[6] M.C.M Hogg T.A. Ho, P. Silva. Multiple imputation and maximum likelihood principal components analysis of incomplete multivariate data from a study of the ageing of port. *Chemometrics and Intelligent Laboratory Systems*, 55:1—11, 2001.

[7] Thompson D. J. Horvitz, D.G. A generalization of sampling without replacement from a finite univers. *American Statistical Association Journal*, 00:663—685, 1951.

[8] Kaiser Javaid. The estimation of correlation matrix from data having missing values. *Research-Report Islamic Countries Conference on Statistical Sciences*, 1994.

[9] M.G. Kenward J.R. Carpenter. A comparison of multiple imputation and inverse probability weighting for analysis with missing data. *Journal of the Royal Statistical Society*, A, 2006.

[10] M. Kullowicz, L. Wallace. The mini mental state examination (mmse). *Best Practices in Nursing Care to Older Adults*, 3:1, 1999.

[11] R.J.A Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, NY, 1987.

[12] Roderick J.A Little. Robust estimation of the mean and covariance matrix from data with missing values. *Journal of Applied Statistics*, 37:22—38, 1988.

[13] F.L Haxby J.V. McIntosh, R.A. Brookstein and C.L. Grady. Spatial pattern analysis of functional brain images using partial least squares. *Neuroimaging*, 3:143—157, 1996.

[14] S. et al. Minoshima. Fdg-pet improves accuracy in distinguishing frototemporal dementia and alzheimer's disease. *Brain*, 130:2616—2635, 2007.

[15] Taylor Paul A. MacGregor John F. Nelson, R.C. Philip Paul. Missing data methods in pca and pls: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, 35:45—65, 1996.

[16] D.W Townsend P.E Valk, D.L Bailey and M.N Maisey. *Positron Emission Tomography: Basic Science and Clinical Practice.* Springer,, New York, NY, 2003.

[17] M. et al. Reivich. The fluorodeoxyglucose method for the measurement of local cerebral glucose utilization in man. *Circulation Research*, 44:127—137, 1979.

[18] A. Robins, J.L. Rotnitzsky and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistcal Association*, 89:846—866, 1994.

[19] A. Robins, J.L. Rotnitzsky and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106—121, 1995.

[20] D.B Rubin. Inference and missing data. *Biometrika*, 65:581—592, 1976.

[21] D.B Rubin. *Multiple Imputation for Nonresponse in Surveys.* J. Wiley & Sons,, New York, NY, 1987.

[22] J. L. Schaffer. *Analysis of incomplete multivariate data.* CRC Press, Cambridge, MA, 1997.

[23] J.L. Rotnitzsky A. Scharfstein, D.O Robins. Adjusting for non-ignorable drop-out using semi-parametric nonresponse model. *Journal of the American Statistcal Association*, 94:1096—1146, 1999.

[24] D.L Walczak, B. Massart. Dealing with missing data part i,ii. *Chemometrics and Intelligent Laboratory Systems*, 58:15—27, 2001.

[25] H. Wold. Soft modelling by latent variables: the nonlinear iterative partial least squares(nipals) approach. *Perspectives in Probability and Statistics*, In Honor of M.S. Bartlett:117—144, 1975.

[26] C. Yang Yuan. Multiple imputations for missing data: Concepts and new development. *SAS proceedings.*

[27] S. Zhao, L.P. Lipsitz. Design and analysis of two-stage studies. *Statistics in Medicine*, 11:769—782, 1992.