Unweighted versus Weighted Estimation for the Multivariate Growth Curve Model Useful in the Analysis of Longitudinal Data

by

Cody Halden

A thesis
presented to the University of Ottawa
in partial fulfillment of the
thesis requirement for the degree of
Master of Science
in
Statistics

Department of Mathematics and Statistics, Faculty of Science

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor(s):	Jemila Hamid Professor, Dept. of Mathematics and Statistics, University of Ottawa
Internal Member:	Kelly Burkett A. Professor, Dept. of Mathematics and Statistics, University of Ottawa
External Member:	Sanjoy Sinha Professor, School of Mathematics and Statistics, Carleton University

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Growth Curve Models (GCMs) are useful in analyzing longitudinal data or studies involving response curves. GCMs, a type of bilinear regression model, include within- and between-individual design matrices. Assuming multivariate normality, explicit likelihood solutions exist, often optimal even for small samples. This thesis examines the estimation of mean parameters in GCMs, using simulations to compare unweighted and weighted estimators in various scenarios: large sample ($n > p$), near singularity ($n \approx p$), and high-dimensional ($n < p$). Longitudinal data with different levels of within-individual correlations were considered, and estimators were compared based on bias and mean squared error (MSE). Matrix bias and matrix MSE aggregation methods were explored for comparative evaluations. Illustrations using three real datasets were provided.

Simulation results showed unweighted estimators performed well in most scenarios, except with unstructured variance-covariance matrices. Unweighted estimators were more optimal for structured variance-covariance matrices, regardless of within-individual correlation strength. Further investigation is needed to evaluate weighted estimators in other covariance structures and assess robustness against model misfits and non-normality.

Acknowledgements

I would like to thank Dr. Hamid for her dedication to my learning throughout my entire thesis. I want to thank her for the many hours of her time that I used during the correction period of this thesis. Without her continued support and dedication, this thesis would not have been possible.

I would like to thank my parents for their continuous support throughout my education and being there for me when I needed to talk about anything.

Finally, I want to thank my friends for understanding the time commitment of a thesis. Without their continuous understanding and support, I would not have finished this thesis.

Dedication

This is dedicated to all of the people in my life who have supported my education and have pushed me to be the best that I can be. Without them, I wouldn't have pursued higher education and wouldn't be in the position that I am in today.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Objectives

1.1 Introduction and Rationale

The multivariate growth curve model (GCM) is a useful alternative (to the generalized linear mixed model (GLMM)) for analyzing longitudinal data. It can also be used to model other response curves, for instance blood pressure or blood glucose measurements modeled as a function of doses of a medication, tensile strength of a material modeled as a function of temperature or pressure, etc. (Chen and Gupta, 2005; Pan and Fang, 2002, Kollo and von Rosen, 2005, Hamid et al., 2011).

The GCM is a natural extension to the classical multivariate analysis of variance (MANOVA) model, in the sense that MANOVA with linear restrictions (i.e. with a linearly structured mean) becomes the GCM (Potthoff and Roy, 1964; Khatri, 1966). For this reason, the GCM is also referred to as the generalised multivariate analysis of variance (GMANOVA) model. The MANOVA model is, therefore, a special case of the GMANOVA model, where the mean does not have a structured functional form (as a function of continuous predictors such as time). The term GMANOVA is used interchangeably to mean the GCM or extended growth curve model (EGCM) as well as to encompass any generalisations or extensions of such models. However, the term GCM or EGCM is often reserved for response variables modelled as a function of time; for example, in studies involving longitudinal data.

We would like to note here that MANOVA can be used to model and compare multiple outcomes (repeated measurements) taken from the same individuals, as well as repeated measurements taken from the same outcome at different time points or at different doses of a treatment or temperature, etc. On the other hand, the GCM in particular, and

GMANOVA models in general, are used in situations where the same outcome is measured repeatedly. In this thesis, we focus on longitudinal data, hence the repeated measurements are assumed to be taken across different time points. Nevertheless, the statistical methods presented and the analysis strategies provided in the simulations are valid for situations other than longitudinal data (e.g., dose-response relationships). We would also like to note here that we will use GCM and GMANOVA interchangably.

Suppose that repeated measurements of a continuous outcome were taken from individuals in k groups and from p time points. Suppose also that the mean is structured (has a functional form with respect to time) and can be represented as a polynomial function of degree $q-1$. That is, for the j^{th} group, the mean can be represented as

$$\mu_j = \beta_{0j} + \beta_{1j}t + \beta_{2j}t^2 + ... + \beta_{(q-1)j}t^{q-1} \tag{1.1}$$

Data with such a mean can be modeled using the GCM, which, using matrix formulations, can be written as

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

where $\mathbf{Y}$ is a $p \times n$ matrix consisting of the p repeated measurements of the continuous outcome (which is also the dependent variable in the model), $\mathbf{Z}$: $p \times q$ and $\mathbf{X}$: $k \times n$ are known within-individual (across time) and between-individual design matrices respectively. The within-individual design matrix represents the functional relationship between the response variable and time. On the other hand, the between-individual design matrix consists of dummy variables (i.e. zeros and ones) representing group membership for each individual, which is exactly the same as the design matrix in the MANOVA model. The parameter matrix $\mathbf{B}$: $q \times k$ consists of the unknown regression coefficients for the k groups, as described in the polynomials presented in Equation (1.1) above. The matrix $\mathbf{E}$ is the error matrix, where the columns of $\mathbf{E}$ are often assumed to be distributed as the p-variate multivariate normal distribution with mean vector $\mathbf{0}$ and unknown variance-covariance matrix $\mathbf{\Sigma}$. (Potthoff and Roy, 1964; Khatri, 1966; Hamid et al., 2011; Hamid and Beyene, 2009; Jana et al., 2016; von Rosen, 1995; Kollo and von Rosen, 2005; Woolson and Leeper, 1980). More details on the GCM, including representations of the various matrices involved in the model, are provided in Chapter 2.

As we can see in the above model, the GCM consists of two design matrices (the within- and between- individual design matrices), and the conditional mean of the dependent variable (given the two design matrices) is modelled using a bilinear multivariate regression framework, often modelling the mean of the response variable with respect to time (von Rosen, 1991; Pan and Fang, 2002; Hamid et al., 2011; Jana, 2013; Jana et al., 2017). For this and other reasons discussed in Chapter 2, the GCM model is also referred to as the

multivariate bilinear regression model (BRM). Under the assumption of multivariate normality, explicit likelihood solutions for the parameters of the GCM are available. Maximum likelihood estimates (MLEs) for both $\mathbf{B}$ (the mean parameters) and the unknown variance-covariance matrix $\mathbf{\Sigma}$ are available in the literature, both under full-rank assumptions and in general conditions (Potthoff and Roy, 1964; Khatri, 1966; von Rosen, 1989).

The GCM model has been extended to the EGCM which is also a GMANOVA model and is useful in the analysis of clustered longitudinal data, where the mean trajectory for different groups is allowed to have different shapes (represented by different degrees of polynomials) unlike the GCM model, where the mean for all groups are assumed to follow polynomials of the same degree, but can have different parameters (Verbyla and Venables, 1988; Hamid, 2001; Hamid and von Rosen, 2006; Zellner, 1962-1963). Under multivariate normality, explicit likelihood solutions for the parameters of the EGCM also exist. Several more studies investigating the various aspects of the GCM and EGCM, including estimation, hypothesis testing and model diagnostics, are also available (Verbyla and Venables, 1988; Hamid et al., 2011; Hamid and von Rosen, 2006; Jana et al., 2020).

The explicit nature of the solutions for both the GCM and EGCM provides a better understanding of residuals from correlated outcomes in general and specifically in the analysis of longitudinal data. As such, more optimal model diagnostics and goodness-of-fit assessments can be done through these bilinear regression models (Srivastava and Carter, 1983; von Rosen, 1995; Hamid and Von Rosen, 2006). The normality assumption has also been relaxed to provide robust estimators for both the GCM and EGCM models, where the multivariate skewed normal distribution was considered (Jana et al., 2018; 2020). Extensions that allow analysis of high-dimensional longitudinal data have also been proposed in the literature (Srivastava, 2007; Hamid & Beyene, 2009; Ahmad et al., 2013; Jana et al., 2013; 2017).

There is considerable literature on GMANOVA models with structured covariance matrices, including aspects of residuals and model diagnostics (Pan and Fang, 2002; Ohlsen and von Rosen, 2010; Nzabanita et al., 2012). However, more extensive research still needs to be done to examine the performance of the covariance estimators (unstructured or structured) and their impact on estimators of the mean parameters and hypothesis tests involving the mean parameters, which are often of interest in practical applications.

In this thesis, we focused on the estimation of the mean parameters of the GCM, represented by the matrix $\mathbf{B}$. For traditional regression models, both univariate and multivariate regression, the maximum likelihood procedure (under the normality assumption) and the least squares method lead to linear projections with respect to the design matrix (or the matrix of covariates). On the other hand, inference for the mean in GMANOVA

models in general, and in particular for the GCM and EGCM, leads to bilinear projections corresponding to the two design matrices (i.e. the within-individual design matrix $\mathbf{Z}$ and the between-individual design matrix $\mathbf{X}$) (Potthoff and Roy, 1964; Khatri, 1966; Kollo and von Rosen, 2005).

The projection with respect to the between-individual design matrix, $\mathbf{X}$, remains the same as it is in the traditional univariate and multivariate linear regression models, including the MANOVA model. The GMANOVA model also involves an additional projection with respect to the within-individual design matrix $\mathbf{Z}$, which involves weighted inner products stemming from dependency in the response (i.e. correlation between measurements over time). This is similar to traditional weighted regression approaches used to overcome limitations due to the presence of heteroskedasticity and / or correlated errors (Kaufman, 2013; Berry and Feldman, 1985).

Potthoff and Roy (1964), in their original paper introducing the GCM, suggested a transformation of the GCM to a MANOVA model, where the transformation involved an arbitrary, nonrandom weight. The arbitrariness of the transformation has been highlighted by the authors themselves. In subsequent literature, arbitrariness is also described as a limitation to their approach. To overcome this limitation, the use of the identity matrix was suggested by the authors, which basically meant using unweighted bilinear projections, ultimately ignoring the dependency in the data. However, the authors argued that ignoring dependency (using the identity matrix in the transformation) might lead to sub-optimal inference, and they suggested that choosing a matrix close to the true variance-covariance matrix might lead to better inference. However, the true variance-covariance matrix is often unknown in practical applications.

Later in 1966, Khatri provided MLEs for the parameters of the GCM under some full-rank conditions. The MLEs also led to bilinear projections with respect to the two design matrices. However, unlike the Potthoff and Roy approach, the likelihood procedure led to projections involving a random weight. The weight, in fact, turned out to be the pooled sample variance-covariance matrix, which is an unbiased estimator of the unknown variance-covariance matrix $\mathbf{\Sigma}$. MLEs under general conditions (without assuming full rank) were also derived for both the GCM and EGCM models (von Rosen, 1989), which also involve the sample variance-covariance matrix as a weight.

Several studies investigating optimality and robustness of the estimators, levels, and powers of hypothesis tests, as well as residual and goodness-of-fit analysis for the GMAVOVA models have been considered in the literature (Pan and Fang, 2002; Hamid and von Rosen, 2006; Ohlsen and von Rosen, 2010; Nzabanita et al., 2012; Jana, 2013; Jana et al., 2017, 2018, 2019, 2020). However, to our knowledge, there are no studies involving formal in-

vestigations in terms of the loss of information (if any) incurred as a result of ignoring the dependency (i.e., unweighted bilinear projections) or relative efficiency gained by using weighted projections. Moreover, there are no studies investigating comparative performance of the different weights. This is particularly important considering that the unknown variance-covariance matrix is poorly estimated in some cases, including situations where we have a small sample size, as well as in high-dimensional scenarios.

1.2 Objectives

The overall objective of this thesis was to fill the methodological gaps identified and presented above. Specifically, our objective was to perform extensive simulations to investigate comparative optimality of weighted versus unweighted bilinear projections for inference involving the mean parameters of the GCM. Our extensive simulations included small and large sample scenarios, high-dimensional scenarios, as well as scenarios with weak, moderate, and strong correlations (dependency) between outcome measurements at different time points (within-individual correlations). We also introduced a novel weighting approach using the MLE of the unknown variance-covariance matrix $\mathbf{\Sigma}$, which might be useful in some situations. We hypothesised that the weighted methods should perform better (than the unweighted method) in all situations given their overall behaviour, and because of the fact that they should simulate a closer value to the true $\mathbf{\Sigma}$, as indicated in the original GCM paper by Potthoff and Roy. However, we also hypothesize that in situations where $\mathbf{\Sigma}$ is poorly estimated (eg. small sample size), the weighted estimators might perform poorly.

1.3 Thesis Organization

Following this introduction chapter, the results of our extensive literature review will be presented in Chapter 2, where we provide detailed historical and theoretical background on the GCM, which also include some new derivations that we believe provide a clear connection between existing theory and the rationale for this thesis. In Chapter 2, we will also discuss bias and mean square error (MSE) for matrix parameters, and present different approaches and strategies we used (in the simulations) to facilitate comparisons between estimators obtained using different methods. In Chapter 3, we will present descriptions of the extensive simulations we performed followed by the simulation results. In Chapter 4, we will present results from analysis of multiple real data sets, including high-dimensional data, and provide interpretations of the findings in light of the simulation results. Summary, discussions and future directions are provided in Chapter 5.

Chapter 2

Background

2.1 Univariate and Multivariate Analysis of Variance Models

Consider an outcome measured from n individuals in k groups , where n_j of the n individuals belong to group j. Suppose we are interested in estimating and comparing the means across the different groups. The structure of the data for such a study can be seen as presented in Table 2.1. Assuming normality, one-way analysis of variance (ANOVA) using the F test can be used to compare the means across the different groups (Srivastava et al., 1983; Rosner, 2006).

Group 1	Group 2	.	.	.	Group k
y_{11}	y_{12}	.	.	.	y_{1k}
y_{21}	y_{22}	.	.	.	y_{2k}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$y_{n_1 1}$	$y_{n_2 2}$	.	.	.	$y_{n_k k}$

Table 2.1: Data structure corresponding to an ANOVA model, where a continuous response variable (outcome) is measured from n individuals across k groups.

Let y_{ij} represent the measurement for the individual i in group j, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., k$. Using linear regression formulations, the ANOVA model can be written as

$$y_{ij} = \mu_j + e_{ij}, \qquad e_{ij} \overset{iid}{\sim} N(0, \sigma^2) \tag{2.1}$$

where μ_j represents the mean for group j, e_{ij} denote the error terms that are assumed to be normally distributed and $\overset{iid}{\sim}$ represents identically and independently distributed. Using matrix formulations, the above regression model can be re-written as

$$\mathbf{y} = \mathbf{BX} + \epsilon, \qquad \epsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.2}$$

where $\mathbf{y}$ is a row vector of length n representing the outcome (dependent) variable, $\mathbf{X}$ is a $k \times n$ matrix of zeros and ones, representing group membership, and is referred to as the between-individual design matrix, $\mathbf{B}$ is a row vector of length k consisting of unknown parameters representing the differences between the means across the k groups and ϵ is a row vector of length n representing the error terms, which is assumed to follow a multivariate normal distribution with variance-covariance matrix of $\sigma^2\mathbf{I}$, where $\mathbf{I}$ represents an $n \times n$ identity matrix.

Consider now the data presented in Table 2.2, where p repeated measurements are taken from each individual across the k groups. Suppose the l^{th} observation for individual i in group j is denoted by Y_{ilj}.

Group 1	Group 2	...	Group k
$y_{111}, y_{121}, ..., y_{1p1}$	$y_{112}, y_{122}, ..., y_{1p2}$	...	$y_{11k}, y_{12k}, ..., y_{1pk}$
$y_{211}, y_{221}, ..., y_{2p1}$	$y_{212}, y_{222}, ..., y_{2p2}$	...	$y_{21k}, y_{22k}, ..., y_{2pk}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$y_{n_1 11}, y_{n_1 21}, ..., y_{n_1 p1}$	$y_{n_2 12}, y_{n_2 22}, ..., y_{n_2 p2}$	...	$y_{n_k 1k}, y_{n_k 2k}, ..., y_{n_k pk}$

Table 2.2: Data structure corresponding to a MANOVA or a GMANOVA model, where p repeated measurements from the same continuous outcome or single measurement from multiple continuous outcomes are taken from individuals across k groups

For a study generating such multivariate data, we are often interested in estimating and comparing the mean vectors among the k groups. Comparison of the mean vectors, for instance, can be done using the MANOVA model and the likelihood ratio test, which

is also referred to as the Wilk's Lambda test (Srivastava et al., 1979, 1983; Casella and Berger, 2002; Srivastava, 2002). There are also other tests alternative to Wilk's Lambda, which include the Lawley-Hotelling Trace test and Roy's maximum root test (Srivastava et al., 1979, 1983; Casella and Berger, 2002; Srivastava, 2002). Using matrix notations, the MANOVA model can be written as

$$\mathbf{Y} = \mathbf{BX} + \mathbf{E}, \qquad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}) \tag{2.3}$$

where $\mathbf{Y}$: $p \times n$ is the matrix of responses, $\mathbf{B}$: $p \times k$ is the matrix of unknown regression coefficients where column j of $\mathbf{B}$ represents the mean vectors of group j, and $\mathbf{X}$: $k \times n$ is the between-individual design matrix representing group membership, the same as what we have in the ANOVA model presented in equation (2.2) (Anderson, 1958; Everitt et al., 2001; Casella and Berger, 2002). The columns of the error matrix $\mathbf{E}$ are identically and independently distributed as a p-variate normal distribution, often written as $N_p(0, \mathbf{\Sigma})$, with mean vector $\mathbf{0}$ and an unknown positive definite variance-covariance matrix $\mathbf{\Sigma}$. Alternatively, we can write the distribution of the error matrix as $\mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I})$ to indicate that the columns of $\mathbf{E}$ are independent and the rows are dependent, and the dependency is represented by $\mathbf{\Sigma}$.

Note that in the MANOVA model presented in Equation (2.3), the mean is unstructured. That is, there is no functional relationship between the elements of the mean vector. For instance, suppose the p measurements taken from each individual represent different cardiovascular related outcomes (say, blood pressure, cholesterol, blood sugar etc). Then, the mean vector for group j represents the mean blood pressure, mean cholesterol and mean blood sugar level, and does not have any functional form with respect to the elements of the mean vector. This is what is referred to as having an unstructured mean, unlike the structured mean for longitudinal data, which we will discuss in the next section.

2.2 The Growth Curve Model

Consider again the data structure presented in Table 2.2, and suppose that the repeated measurements are taken across p time points, p dose levels, or with respect to any other continuous variables (eg. p temperatures or pressure levels, etc). Moving forward, we will, without loss of generality, assume that the repeated measurements are taken at different time points (which leads to longitudinal data). However, we would like to note that the methods are applicable to any multivariate data examining response curves (eg. dose-response curves).

Suppose now that the mean is structured (has a functional form with respect to time) and can be represented as a polynomial function (of time) of degree $q-1$, as described in Equation (1.1). For instance, consider a classical textbook data set from a dental study first analyzed by Potthoff and Roy (1964). The data consists of four repeated dental measurements (at ages 8, 10, 12 and 14) taken from 11 girls and 16 boys ($n = 27$, $k = 2$, $p = 4$). The individual and mean profiles as a function of time for the two groups are provided in Figure 2.1.

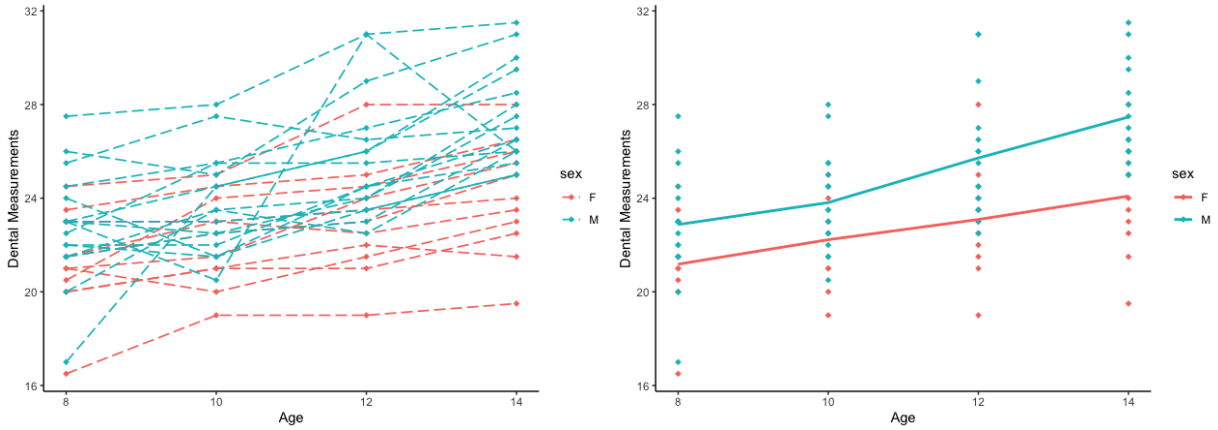

Figure 2.1: Individual Profile and Mean Profile Plots for Potthoff and Roy's dental data consisting of repeated measurements from 16 boys (group 1) and 11 girls (group 2).

From the above figure, we can see that the group means (for both boys and girls) can be represented by linear functions of time (i.e. a structured mean with a polynomial of degree $q-1=1$). The corresponding GCM, using matrix formulations, can be written as

$$\mathbf{Y} = \mathbf{ZBX} + \mathbf{E}, \qquad \mathbf{E} \sim N_{4,27}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}) \tag{2.4}$$

where $\mathbf{Y}$: 4×27 is a matrix of observed outcomes (p repeated measurements of the dependent variable taken from n individuals, where $p = 4$ and $n = 27$), $\mathbf{Z}$: 4×2 and $\mathbf{X}$: 2×27 are the within-individual (across time) and between-individual design matrices, respectively. Here, $\mathbf{B}$: 2×2 is the matrix consisting of the unknown intercepts and slopes corresponding to the 2 groups (boys and girls), and $\mathbf{E}$ is the error matrix which we will assume follows the multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$. The estimates of the mean and variance-covariance parameters of the dental data are used in our simulations to generate data with dependency similar to a real life scenario. The data will also be analyzed in Chapter 3 and the results presented as an illustration.

To provide an additional motivating example, consider another dataset, which we will also analyse in Chapter 3. The data is obtained from a glucose tolerance test and consists of 8 repeated measurements (at 0 (baseline) and 0.5, 1, 1.5, 2, 3, 4, and 5 hours post administration of oral glucose) taken from 13 control (non-obese) and 20 obese individuals ($n = 33$, $k = 2$, $p = 8$) (Zerbe, 1979). The individual and mean profiles for the two groups are provided in Figure 2.2. The estimates obtained from this data set were also used in our simulations to generate a real life scenario with $p = 8$.

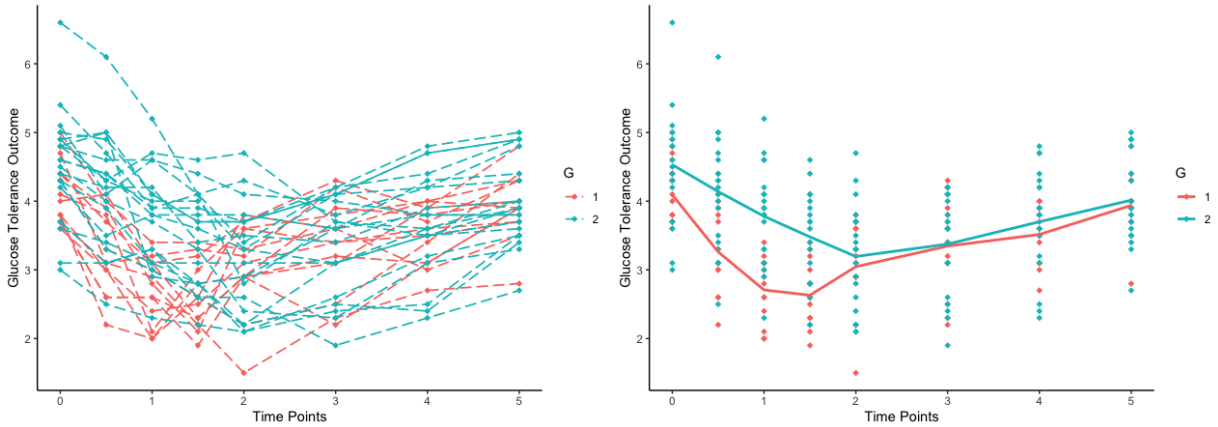

Figure 2.2: Individual Profile and Mean Profile Plots for Zerbe's Glucose data consisting of repeated measurements from 13 control (group 1) and 20 obese individuals (group 2).

From Figure 2.2, we can see that the group means (for both groups) can be represented as quadratic functions of time (i.e., a structured mean with a polynomial of degree $q-1 = 2$). The corresponding GCM, using matrix formulations, can be written as

$$\mathbf{Y} = \mathbf{ZBX} + \mathbf{E}, \qquad \mathbf{E} \sim N_{8,33}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}) \tag{2.5}$$

where $\mathbf{Y}$: 8×33 is a matrix of observed outcomes (p repeated measurements of the dependent variable taken from n individuals, where $p = 8$ and $n = 33$), $\mathbf{Z}$: 8×3 and $\mathbf{X}$: 2×33 are the within-individual (across time) and between-individual design matrices, respectively. Here, $\mathbf{B}$: 3×2 is the matrix consisting of the unknown intercepts and slopes corresponding to the 2 groups (control and obese) and results on the estimation of this matrix are presented in Chapter 4. $\mathbf{E}$ is the error matrix which is often assumed to follow the multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$. If the trajectory for the two groups were different, we would have fit the EGCM (Jana, 2018). That is, if one of the groups (say the control group as an example) presented with a more linear mean trajectory, we could have fit the control group with a linear mean trajectory and the obese group with the original quadratic mean trajectory.

In general, the within-individual design matrix $\mathbf{Z}$ (described below) specifies the time at which measurements of the response (dependent) variable $\mathbf{Y}_{ij}$ are taken, and represents the functional relationship between the mean and time. The first column of the matrix $\mathbf{Z}$ is a vector of 1's corresponding to the intercept term in the polynomial function that represents the mean trajectory over time.

$$\mathbf{Z} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{q-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{q-1} \\ 1 & t_3 & t_3^2 & \cdots & t_3^{q-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^{q-1} \end{pmatrix}$$

For the dental data previously introduced in equation (2.4), the within-individual design matrix, assuming linear mean structures over time for both the girls and the boys, is given by

$$\mathbf{Z} = \begin{pmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{pmatrix}$$

On the other hand, the between-individual design matrix $\mathbf{X}$ is the same as the design matrix involved in ANOVA and MANOVA models, and represents the group membership. Suppose n_1 individuals belong to group 1, n_2 individuals belong to group 2 and n_k belong to group k, the between-individual design matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_2} & \mathbf{0}_{n_3} & \cdots & \mathbf{0}_{n_k} \\ \mathbf{0}_{n_1} & \mathbf{1}_{n_2} & \mathbf{0}_{n_3} & \cdots & \mathbf{0}_{n_k} \\ \mathbf{0}_{n_1} & \mathbf{0}_{n_2} & \mathbf{1}_{n_3} & \cdots & \mathbf{0}_{n_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_1} & \mathbf{0}_{n_2} & \mathbf{0}_{n_3} & \cdots & \mathbf{1}_{n_k} \end{pmatrix}$$

where $\mathbf{1}_{n_k}$ and $\mathbf{0}_{n_k}$ represent row vectors of $1's$ and $0's$ of length n_k. For the dental data, for instance, the between-individual design matrix is given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{11} & \mathbf{0}_{16} \\ \mathbf{0}_{11} & \mathbf{1}_{16} \end{pmatrix}$$

Finally, the unknown parameter matrix $\mathbf{B}$ corresponding to the coefficients of the polynomials representing the mean structure given by:

$$\mathbf{B} = \begin{pmatrix} b_{01} & b_{02} & b_{03} & \cdots & b_{0k} \\ b_{11} & b_{12} & b_{13} & \cdots & b_{1k} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{(q-1)1} & b_{(q-1)2} & b_{(q-1)3} & \cdots & b_{(q-1)k} \end{pmatrix}$$

For the dental data, it consists of the intercepts and the coefficients of the linear terms for the girls and boys, and is given by

$$\mathbf{B} = \begin{pmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \end{pmatrix}$$

That is, the mean for the girls and the boys, respectively, can be written (as a function of time) as

$$\mu_g = b_{01} + b_{11}t \quad \text{and} \quad \mu_b = b_{02} + b_{12}t$$

The GCM, presented in equation (2.4) in its matrix form, was formally introduced by Potthoff and Roy (1964). The model assumes that the mean trajectories for all k groups have the same shape, hence requiring the same degree of polynomials to be fitted for all the groups, although the coefficients of the polynomials for the different groups can vary. However, in practical applications, there may be scenarios whereby the mean trajectories over time for the different groups have different shapes. The Extended Growth Curve Model (EGCM) can be used to overcome this limitation. The EGCM was independently introduced by Srivastava and Khatri (1979) and Verbyla and Venables (1988). The MLEs for the model parameters are provided in von Rosen (1989). Hypothesis testing, residual analysis, and model diagnostics for the EGCM are also studied (Hamid and von Rosen, 2006; Jana et al., 2020). The model allows analysis of clustered longitudinal data, which highlights one of the benefits of using multivariate bilinear models instead of, for instance, using Generalized Linear Mixed Models (GLMMs).

The EGCM, under special scenarios, was also considered by Verbyla and Venables and is presented as a special case of Zellner's multivariate seemingly unrelated regression (SUR) model (Verbyla and Venables, 1988; Zellner, 1962,1963). This 1962 model is mentioned very frequently in econometrics literature (Zellner, 1962,1963; Stanek and Koch, 1985; Mendoza et al., 1995). However, there exist many other scenario based papers, allowing for the GCM to be used in less than optimal settings. Jana's Skewed Normal GCM and EGCM account for data that do not follow a normal distribution and provides robust estimators (Jana

et al., 2018; 2020). This is particularly important when performing inference involving nonnormal longitudinal data, as the normal-based estimators are shown to be biased and associated with large variance, especially when data are highly skewed.

2.3 Inference for the Growth Curve Model

There is considerable literature on both estimation and hypothesis testing involving the GMANOVA models. However, only estimation is considered in this thesis. As such we only present a summary of existing literature related to estimation of the model parameters. The main focus in most practical applications is the mean parameter $\mathbf{B}$, however, the variance-covariance matrix, $\mathbf{\Sigma}$, often appears as a weight in the estimators of the mean parameters. This could be an additional topic of discussion for a future paper.

Consider first the MANOVA model, presented in equation (2.3), which also includes the ANOVA model as its special case (i.e. when $p = 1$). The MLE under the normality assumption (which is also the least squares estimator) for $\mathbf{B}$ is given by (Srivastava et al., 1979, 1983; Casella and Berger, 2002; Srivastava, 2002)

$$\hat{\mathbf{B}} = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1},$$

which leads to the estimated (predicted) mean and residuals, which, respectively as:

$$\hat{\mathbf{Y}} = \hat{\mathbf{B}}\mathbf{X} = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}$$
$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y}(\mathbf{I} - \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X})$$

The MLE for $\mathbf{\Sigma}$ is given by

$$\hat{\mathbf{\Sigma}} = \frac{\mathbf{R}\mathbf{R}'}{n} = \frac{(\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X})(\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X})'}{n} = \frac{\mathbf{Y}(\mathbf{I} - \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X})\mathbf{Y}'}{n}$$

In practice, we often use the unbiased estimator given in Equation 2.6, which can also be shown to be equivalent to the pooled sample variance-covariance matrix (von Rosen, 1989; Hamid and von Rosen, 2006; Jana et al., 2020).

$$\hat{\mathbf{\Sigma}}^* = \frac{(\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X})(\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X})'}{n - k} = \frac{\mathbf{R}\mathbf{R}'}{n - k} \tag{2.6}$$

Formal inference involving the GMANOVA model first appeared in Potthoff and Roy's paper, where they used transformations to reduce the GMANOVA model to a MANOVA model. Once the model was transformed, methods developed for the MANOVA model were used to provide estimators and perform hypothesis testing. Let us consider the GMANOVA model again

$$\mathbf{Y} = \mathbf{ZBX} + \mathbf{E}$$

Suppose $\mathbf{\Sigma}$ is known, we can multiply the above model by $\mathbf{P} = (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}$, to get

$$\begin{aligned} \mathbf{PY} &= \mathbf{PZBX} + \mathbf{PE} \\ &= (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{ZBX} + \mathbf{PE} \\ &= \mathbf{BX} + \mathbf{PE} \\ \Leftrightarrow \mathbf{Y}^* &= \mathbf{BX} + \mathbf{E}^*, \end{aligned}$$

which is a MANOVA model, whereby

$$\begin{aligned} \mathbf{Y}^* = \mathbf{PY} = (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Y} \\ \mathbf{E}^* = \mathbf{PE} = (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{E} \end{aligned}$$

Let us now check the distributional assumptions for the transformed model. Recall that the columns of the error matrix $\mathbf{E}$ are independent, and hence the columns of $\mathbf{E}^* = \mathbf{PE}$ are also independent, since $\mathbf{P}$ is a known, fixed (non-random), matrix. Now, let $\mathbf{E}_i$ and $\mathbf{E}_i^*$ represent the i^{th} columns of $\mathbf{E}$ and $\mathbf{E}^*$, respectively. Hence, we have

$$\begin{aligned} cov(\mathbf{E}_i^*) &= cov(\mathbf{PE}_i\mathbf{P}') = \mathbf{P\Sigma P}' \\ &= (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{\Sigma\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1} \\ &= (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1} \\ &= (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1} \end{aligned}$$

The transformed model (from GMANOVA to MANOVA) can be used in inference involving the parameter matrix $\mathbf{B}$. Hence, the MLE for $\mathbf{B}$ in the GCM, assuming $\mathbf{\Sigma}$ is known, is given by:

$$\begin{aligned} \hat{\mathbf{B}} &= \mathbf{Y}^*\mathbf{X}(\mathbf{XX}')^{-1} \\ &= (\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{YX}(\mathbf{XX}')^{-1}, \end{aligned}$$

leading to the estimated mean (predicted values):

$$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\mathbf{B}}\mathbf{X} = \mathbf{Z}(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X},$$

which is a bilinear projection with respect to the two design matrices, hence generating the design space for the GCM (Hamid and von Rosen, 2006; Jana et al., 2020). This is one of the reasons the GCM and GMANOVA models are often referred to as multivariate bilinear regression models. We also see that the projection involving $\mathbf{Z}$ is weighted, to account for the within-individual (across time) dependency, where in this case the weight $\boldsymbol{\Sigma}^{-1}$ is assumed to be known. In our simulations, this estimator of $\mathbf{B}$ is referred to as the $\boldsymbol{\Sigma}$-weighted estimator.

In practical applications, $\boldsymbol{\Sigma}$ is unknown. To overcome this practical limitation, Potthoff and Roy suggested using the identity matrix, that is a transformation using $\mathbf{P} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, which also leads to the MANOVA model but with

$$\mathbf{E}_i^* \sim N_p\left(\mathbf{0}, (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\right)$$

In this case, the estimators for the parameter $\mathbf{B}$ and the mean, respectively, are given by

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1} \\ \\ \hat{\mathbf{Y}} &= \mathbf{Z}\hat{\mathbf{B}}\mathbf{X} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X},\end{aligned}$$

which is still a bilinear projection with respect to $\mathbf{Z}$ and $\mathbf{X}$. However, unlike the $\boldsymbol{\Sigma}$-weighted estimator, both of the projections involved in the estimator are unweighted, and hence this approach ignores the dependency in the data. This provides simplifications, especially in hypothesis testing involving the GMANOVA models (Potthoff and Roy, 1964), which could be useful in certain scenarios. However, no investigation was performed to evaluate the loss of precision and accuracy as a result of ignoring the within-individual (across time) dependency, which is one of the motivations for this thesis.

In theory, the transformation $\mathbf{P} = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}^{-1}$, with an arbitrary positive-definite matrix $\mathbf{V}$, can be used to transform the GMANOVA model to the MANOVA model. Potthoff and Roy indicated that the closer the weight is to the true, but unknown, variance-covariance matrix $\boldsymbol{\Sigma}$, the more optimal the estimator of $\mathbf{B}$ would be. Nevertheless, no formal evaluation of this was done to confirm if this is indeed the case. Hence, one of the objectives of this thesis is to formally evaluate this and other issues related to weighted and unweighted estimation.

It is important to highlight that the projections above are all known, and hence there is no randomness involved in these projections. Therefore, once the data is observed, both

projection matrices in the bilinear projections are completely specified. This is not the case for the MLEs derived by Khatri (1966). The MLEs for $\mathbf{B}$ and the mean trajectory over time, respectively, are given by

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1} \\ \hat{\mathbf{Y}} &= \mathbf{Z}\hat{\mathbf{B}}\mathbf{X} = \mathbf{Z}(\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X},\end{aligned}$$

where $\mathbf{S} = \mathbf{Y}(\mathbf{I} - \mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X})\mathbf{Y}'$, which represents the squared deviation from the mean across all groups. Note also that the pooled sample variance-covariance matrix is given by (Kollo and von Rosen, 2005; Hamid and von Rosen, 2006)

$$\mathbf{S}_p = \frac{\mathbf{Y}(\mathbf{I} - \mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X})\mathbf{Y}'}{n-k}.$$

The weighting using the pooled variance-covariance matrix leads to the same estimators as the MLEs presented above. That is

$$\hat{\mathbf{B}} = (\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1} = (\mathbf{Z}'\mathbf{S}_p^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}_p^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}$$

and

$$\hat{\mathbf{Y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X} = \mathbf{Z}(\mathbf{Z}'\mathbf{S}_p^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}_p^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}$$

Therefore, in our simulations, we will refer to the MLE estimator of $\mathbf{B}$ as $\mathbf{S}$-weighted or covariance-weighted. Next, we can also produce the MLE of the true variance-covariance matrix $\mathbf{\Sigma}$ now that we know the MLE for $\hat{\mathbf{B}}$ and $\hat{\mathbf{Y}}$.

Similar to the MANOVA model (which is also the case for univariate and other multivariate models), the MLE for $\mathbf{\Sigma}$ for the GMANOVA models is given by:

$$\hat{\mathbf{\Sigma}}_{MLE} = \frac{\mathbf{R}\mathbf{R}'}{n} = \frac{(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})'}{n} = \frac{(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}\mathbf{X})(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}\mathbf{X})'}{n} \tag{2.7}$$

Additionally, in the GMANOVA model, the above estimator can be re-written as

$$n\hat{\mathbf{\Sigma}} = \mathbf{R}\mathbf{R}' = \mathbf{R_1}\mathbf{R_1}' + \mathbf{R_2}\mathbf{R_2}' + \mathbf{R_3}\mathbf{R_3}', \tag{2.8}$$

where we define

$$\begin{aligned}
\mathbf{P}_{Z,S} &= \mathbf{Z}(\mathbf{Z}'\mathbf{S^{-1}}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{S}^{-1} \\
\mathbf{P}_X &= \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X} \\
\mathbf{R_1} &= (\mathbf{I} - \mathbf{P_{Z,S}})\mathbf{Y}(\mathbf{I} - \mathbf{P_X}) \\
\mathbf{R_2} &= \mathbf{P_{Z,S}}\mathbf{Y}(\mathbf{I} - \mathbf{P_X}) \\
\mathbf{R_3} &= (\mathbf{I} - \mathbf{P_{Z,S}})\mathbf{Y}\mathbf{P_X} \\
\mathbf{S} &= \mathbf{Y}(\mathbf{I} - \mathbf{P_X})\mathbf{Y}'
\end{aligned}$$

Moreover, note that

$$\begin{aligned}
\mathbf{R_1}\mathbf{R_1}' + \mathbf{R_2}\mathbf{R_2}' &= (\mathbf{I} - \mathbf{P}_{Z,S})\mathbf{Y}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}'(\mathbf{I} - \mathbf{P}_{Z,S})' + \mathbf{P}_{Z,S}\mathbf{Y}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}'\mathbf{P}'_{Z,S} \\
&= \mathbf{Y}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}' \\
&= (n-k)\mathbf{S}_p
\end{aligned}$$

So, we have that

$$\hat{\mathbf{\Sigma}}_{MLE} = \left(\frac{n-k}{n}\right)\mathbf{S}_p + \frac{\mathbf{R}_3\mathbf{R}_3'}{n},$$

where $\mathbf{R}_3$ comes from the residual decomposition done by von Rosen and Hamid (von Rosen, 1996; Hamid & von Rosen, 2006; Hamid et al., 2011), and is given by

$$\mathbf{R}_3 = (\mathbf{I} - \mathbf{P_{Z,S}})\mathbf{Y}\mathbf{P_X}$$

Alternatively, we have

$$n\hat{\mathbf{\Sigma}}_{MLE} = \mathbf{S} + \mathbf{R}_3\mathbf{R}_3'$$

As we can see above, the MLE estimator of $\mathbf{\Sigma}$ has a shrinkage structure, hence may be a better estimator in situations involving small sample sizes, and in particular might

overcome limitations when $\mathbf{S}$ is singular. For this reason, we will consider an additional estimator for $\mathbf{B}$, where the inverse of the MLE of $\mathbf{\Sigma}$ is used as the weight. That is,

$$\hat{\mathbf{B}} = (\mathbf{Z}'\mathbf{\Sigma}_{MLE}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}_{MLE}^{-1}\mathbf{Y}\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}$$

In our simulations, we will refer to this estimator as the MLE-weighted estimator. We would like to highlight here that weighting by $\mathbf{\Sigma}_{MLE}^{-1}$ and $(n\mathbf{\Sigma}_{MLE})^{-1} = (\mathbf{S} + \mathbf{R}_3\mathbf{R}_3')^{-1}$ lead to the same estimator, since n will cancel out.

2.4 Comparing Optimality of Matrix Estimators

When comparing multiple estimators using simulations, bias and mean squared error (MSE) are the two most commonly used optimality criteria. In studies involving matrix parameters, bias and MSE are also matrices, hence making the comparisons quite challenging. Calculation of bias for a matrix estimator is straightforward, where bias, similar to univariate and vector estimators, is defined as the difference between the expected value of the estimator and the true parameter. That is, for a parameter an unknown matrix $\mathbf{B}$ of interest and its estimator $\hat{\mathbf{B}}$,

$$Matrix\, Bias(\hat{\mathbf{B}}) = E[\hat{\mathbf{B}}] - \mathbf{B}.$$

In our simulations, we used element-wise bias to show the overall distribution of bias of the different estimators, and to evaluate how the distribution compares under the many scenarios we considered. We also considered the Euclidean distance of the $Matrix\, Bias(\hat{\mathbf{B}})$ from the zero matrix, which is given by

$$||Bias(\hat{\mathbf{B}})||^2 = \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{k}\left(bias(\hat{b}_{ij})\right)^2} = \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{k}(\hat{b}_{ij} - b_{ij})^2}, \tag{2.9}$$

where $\hat{b}_{ij}$ and b_{ij} are the ij^{th} elements of the matrix $\hat{\mathbf{B}}$ and $\mathbf{B}$, respectively.

The MSE for vector and matrix estimators is not well defined in literature, and a simple generalization of univariate MSE can lead to misleading interpretations, since this definition for matrix estimators involves covariance between the element-wise estimators and the cross-products of biases of the elements of the estimator matrix. Nevertheless, this definition of MSE exists in the literature despite this limitation, especially in its utility

involving comparative analysis. Consider a univariate estimator $\hat{\theta}$ of θ, the MSE is defined as an expectation of the squared difference between the estimator and the true parameter. That is,

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2], \tag{2.10}$$

which can easily be shown to be equivalent to the sum of variance and square of bias:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \Big(Bias(\hat{\theta})\Big)^2$$

For a vector estimator $\hat{\mathbf{\Theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)'$ of $\mathbf{\Theta} = (\theta_1, \theta_2, ..., \theta_k)$', the MSE is defined as (Casella and Berger, 2002):

$$MSE(\hat{\mathbf{\Theta}}) = E[(\hat{\mathbf{\Theta}} - \mathbf{\Theta})(\hat{\mathbf{\Theta}} - \mathbf{\Theta})'],$$

which involves an outer product of two vectors, and hence is a matrix, which we will refer to as $Matrix\,MSE$ to highlight that it is a matrix and involves more than the element-wise MSEs. We show below that the MSE of a vector of estimators can be represented as the sum of the variance-covariance matrix of the estimator and the outer product between the bias vector.

$$\begin{aligned}
Matrix\,MSE(\hat{\mathbf{\Theta}}) &= MSE(\hat{\mathbf{\Theta}}) \\
&= E[(\hat{\mathbf{\Theta}} - \mathbf{\Theta})(\hat{\mathbf{\Theta}} - \mathbf{\Theta})'] \\
&= E[(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}) + E(\hat{\mathbf{\Theta}}) - \mathbf{\Theta})(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}) + E(\hat{\mathbf{\Theta}}) - \mathbf{\Theta})'] \\
&= E[(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}))(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}))'] - E\left[(\mathbf{\Theta} - E(\hat{\mathbf{\Theta}}))(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}))'\right] \\
&- E\left[(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}))(\mathbf{\Theta} - E(\hat{\mathbf{\Theta}}))'\right] + (E(\hat{\mathbf{\Theta}}) - \mathbf{\Theta})(E(\hat{\mathbf{\Theta}}) - \mathbf{\Theta})') \\
&= \mathbf{\Omega} + Bias(\hat{\mathbf{\Theta}})Bias(\hat{\mathbf{\Theta}})'
\end{aligned}$$

Note that the cross-product term vanishes since we have

$$E\left[(\mathbf{\Theta} - E(\hat{\mathbf{\Theta}}))(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}))'\right] = E\left[(\hat{\mathbf{\Theta}} - E(\hat{\mathbf{\Theta}}))(\mathbf{\Theta} - E(\hat{\mathbf{\Theta}}))'\right] = \mathbf{0}.$$

Here, $Matrix\,MSE(\hat{\mathbf{\Theta}})$ is a matrix of dimensions $k \times k$ and $\mathbf{\Omega}$ represents the true variance-covariance matrix for the vector of estimators, which is also referred to as the dispersion matrix of the estimators. In simulations, $\mathbf{\Omega}$ is estimated empirically, which we will denote by $\hat{\mathbf{\Omega}}$. Note also that $Bias(\hat{\mathbf{\Theta}})Bias(\hat{\mathbf{\Theta}})'$ is a matrix, where the diagonal elements of which are the square of the element-wise biases and the off diagonal elements correspond

to the product between the element-wise biases. As such, comparing two estimators, say $\hat{\mathbf{\Theta}}_1$ and $\hat{\mathbf{\Theta}}_2$ of $\mathbf{\Theta}$, based on the matrix MSE stemming from the above definition is quite a challenge - not only that comparing two matrices is difficult in itself, but also that the MSE involving covariances and cross-product of biases makes it even more complicated to interpret.

To overcome this challenge, we can use the trace of $Matrix\,MSE(\hat{\mathbf{\Theta}})$, which is equivalent to the sum of the element-wise MSEs. Alternatively, one can also use the average of the element-wise MSEs, which is equivalent to $trace(Matrix\,MSE(\hat{\mathbf{\Theta}}))/k$. The Euclidean distance of $Matrix\,MSE(\hat{\mathbf{\Theta}})$ from the zero matrix, is also another way of dimension reduction for the purpose of comparative evaluation. However, because of the involvement of the variance-covariances of the estimators and cross-products of biases of the estimators in $Matrix\,MSE(\hat{\mathbf{\Theta}})$, which makes interpretation complicated, we use a different strategy in this thesis, whereby we simply use the matrix consisting of element-wise MSEs and its distance from the zero matrix.

Consider now matrix estimator $\hat{\mathbf{B}}$ of $\mathbf{B}$. Following the definition of MSE for the univariate and vector estimators, the MSE for $\hat{\mathbf{B}}$ is given by (Jana, 2013)

$$Matrix\,MSE(\hat{\mathbf{B}}) = MSE(\hat{\mathbf{B}}) = E[(\hat{\mathbf{B}} - \mathbf{B})(\hat{\mathbf{B}} - \mathbf{B})'], \tag{2.11}$$

which can also be shown to be $\mathbf{\Omega} + Matrix\,Bias(\hat{\mathbf{B}})Matrix\,Bias(\hat{\mathbf{B}})'$, where $\mathbf{\Omega}$ is a matrix of variances and covariances for the estimator $\hat{\mathbf{B}}$ and can be empirically estimated. Similarly to what we discussed for vector estimators, $Matrix\,MSE(\hat{\mathbf{B}})$ consists of variance covariances and cross products of biases, which leads to the same challenges of feasibility and interpretability. To overcome this challenge, Jana et al. (2013) used the trace of the the $Matrix\,MSE(\hat{\mathbf{B}})$, which represents the sum of element-wise MSEs. This may lead to larger MSEs , which can be misleading.

In our simulations, we will use element-wise MSEs as well as the matrix consisting of element-wise MSEs. For comparison purposes, involving all elements of the matrix estimator, we used the Euclidean distance of the element-wise MSEs from the zero matrix. That is,

$$||MSE^*(\hat{\mathbf{B}})||^2 = \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{k}\left(MSE(\hat{b}_{ij})\right)^2}, \tag{2.12}$$

where $MSE^*(\hat{\mathbf{B}})$ is a matrix consisting of element-wise MSE. Note that $MSE^*(\hat{\mathbf{B}})$ is different from $Matrix\,MSE(\hat{\mathbf{B}})$, which consists of covariances and cross-products of element-wise biases.

Chapter 3

Simulations

3.1 Simulation Design and Settings

We performed an extensive simulation in R consisting of 665 scenarios with 100 replications of 1,000 samples for each bias scenario, to allow visualisation of the distribution graphically. Data within each scenario of 1,000 samples was replicated 100 times and element-wise empirical bias and empirical MSE were calculated from the 1,000 samples to generate an average of these measurements for each of the 100 replications. The simulation scenarios involved various sample sizes, dimensions, mean structures, and overall nature of within-individual dependency (Table 3.1). Without loss of generality, we considered two groups ($k = 2$). However, the methods, statistical analysis, and results are valid for one group as well as more than two groups.

Parameters	Values
Mean Trajectory	Linear ($q = 2$), Quadratic ($q = 3$)
Range of Sample Sizes	$n = 6 - 1000$
Dimension	$p = 4, 8, 40$
Covariance	Weak, Moderate, Strong
Groups	$k = 2$
Distribution	Multivariate Normal

For each scenario, longitudinal data with p repeated measurements from n individuals (n_1 from the first group and n_2 from the second group) was generated using the GCM:

$$\mathbf{Y} = \mathbf{ZBX} + \mathbf{E}, \qquad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}) \tag{3.1}$$

where design matrices $\mathbf{Z}$ and $\mathbf{X}$ were first generated, with $k = 2$ (without loss of generality) and the desired dimension (i.e. the number of time points p). Without loss of generality, the value of time is set at $t = 1, 2, ..., p$. We also used real-life scenarios, where time is specified based on a real data set analysed in Chapter 4.

In the next step, we specified the degree of the polynomial for the mean structure and the value of the mean parameter $\mathbf{B}$ (the coefficients of the polynomials). We considered linear and quadratic mean trajectory over time, nevertheless, the results can be generalized for any polynomial of degree $q - 1$. The elements of $\mathbf{B}$ were initially taken from real data sets, to allow some of the scenarios to align with those of the real world. We have also considered other scenarios by modifying the values we obtained from real data sets. The only unknown parameter we still need to describe is the variance-covariance matrix $\mathbf{\Sigma}$, which is specified with the desired level of dependency (correlation) between the measurements over time (the strength of within-individual correlation). Once all the parameters of the model are specified, the dependent variable (response) is generated according to the model in Equation 3.1, where we first generated $\mathbf{E}$ from the multivariate normal distribution (*mvrnorm*) with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$. Detailed specifications for $\mathbf{B}$ and $\mathbf{\Sigma}$ are presented in subsequent sections.

Since the objective is to evaluate comparative performance of unweighted and weighted approaches for estimating the mean parameters of the GCM, we considered the following four estimators, discussed in Chapter 2:

- Unweighted: $\hat{\mathbf{B}} = (\mathbf{Z}{\mathbf{I_p}}^{-1}\mathbf{Z}')^{-1}\mathbf{Z}'{\mathbf{I_p}}^{-1}\mathbf{YX}'(\mathbf{XX}')^{-1}$
- $\mathbf{S}$-Weighted: $\hat{\mathbf{B}} = (\mathbf{ZS}^{-1}\mathbf{Z}')^{-1}\mathbf{Z}'\mathbf{S}^{-1}\mathbf{YX}'(\mathbf{XX}')^{-1}$
- MLE-Weighted: $\hat{\mathbf{B}} = (\mathbf{Z}\hat{\mathbf{\Sigma}}_{MLE}^{-1}\mathbf{Z}')^{-1}\mathbf{Z}'\hat{\mathbf{\Sigma}}_{MLE}^{-1}\mathbf{YX}'(\mathbf{XX}')^{-1}$
- $\mathbf{\Sigma}$-Weighted: $\hat{\mathbf{B}} = (\mathbf{Z\Sigma}^{-1}\mathbf{Z}')^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{YX}'(\mathbf{XX}')^{-1}$

We used both element-wise biases and MSEs as well as aggregated biases and MSEs in our comparison, where aggregation of the element-wise biases and MSEs were done using the Euclidean distance, as discussed in Chapter 2. In addition to the traditional sample size scenarios ($n > p$ and $n >> p$) , we also considered high-dimensional scenarios ($n < p$) and scenarios near singularity ($n = p$ and $n \approx p$) .

In our initial simulations, which are presented in the next section as motivating scenarios, we mimicked real-life datasets, where we used the parameter estimates from real data sets as parameters for our simulations. For these scenarios, the dental and glucose data (briefly presented in Chapter 2) were used. We observed change in the behaviour of bias and MSE when sample size scenarios approached singularity. To further investigate this, we expanded the simulations to include a larger number of time points to be able to truly assess the behaviour in the high-dimensional and near singularity scenarios, notably $p = 40$. Further exploration into even larger p may lead to different results and should be explored.

To evaluate the effect of covariance (dependency across time) on comparative performance evaluations, we generated data such that observations across time are associated with weak, moderate, and strong correlation. The analysis is in the final section of this chapter.

All simulations were performed using Version 4.2.3 of the R statistical software (R Core Team, 2023).

3.2 Simulation Results

3.2.1 Motivating Scenarios

For the scenarios in this section, we used real data sets that involved longitudinal measurements at 8 (using glucose data) and 4 (using dental data) time points. Both data sets have two groups. We first focus on the glucose scenario with 8 time points and 33 individuals (13 cases and 20 controls), where the variance-covariance matrix for the simulation was generated from the sample variance-covariance matrix of the glucose data, described in Chapter 2. We calculated the MLE estimator for $\mathbf{B}$ assuming a quadratic mean profile over time for the two groups (given below) and calculated the pooled sample variance-covariance matrix (denoted by $\mathbf{S}_p$), which is also provided below. These matrices were then used to generate a multivariate outcome matrix represented by $\mathbf{Y}$.

$$\hat{\mathbf{B}} = \begin{bmatrix} 5.00 & 5.12 \\ -1.04 & -0.66 \\ 0.11 & 0.07 \end{bmatrix}$$

$$\mathbf{S}_p = \begin{bmatrix} 0.47 & 0.44 & 0.33 & 0.3 & 0.26 & 0.28 & 0.24 & 0.23 \\ 0.44 & 0.57 & 0.31 & 0.27 & 0.26 & 0.26 & 0.29 & 0.22 \\ 0.33 & 0.31 & 0.4 & 0.29 & 0.26 & 0.21 & 0.2 & 0.19 \\ 0.3 & 0.27 & 0.29 & 0.37 & 0.32 & 0.3 & 0.22 & 0.19 \\ 0.26 & 0.26 & 0.26 & 0.32 & 0.52 & 0.37 & 0.28 & 0.23 \\ 0.28 & 0.26 & 0.21 & 0.3 & 0.37 & 0.49 & 0.37 & 0.3 \\ 0.24 & 0.29 & 0.2 & 0.22 & 0.28 & 0.37 & 0.41 & 0.32 \\ 0.23 & 0.22 & 0.19 & 0.19 & 0.23 & 0.3 & 0.32 & 0.34 \end{bmatrix}$$

The variance-covariance matrix is unstructured and dependency ranges from moderate (correlation of 0.47) to relatively strong (correlation of 0.87). We can also see that the variability between the different time points is not consistent. For the mean parameters in our simulations, to be able to examine if the results are sensitive to potential changes in the estimators, we also scaled some of the elements of the $\hat{\mathbf{B}}$ matrix:

$$\begin{bmatrix} 5.09 & 4.92 \\ -10.41 & -6.61 \\ 11.05 & 7.54 \end{bmatrix}$$

Next, we focus on the dental scenario with 4 time points and 27 individuals (11 boys and 17 girls), where the variance-covariance matrix for the simulation was generated from the sample variance-covariance matrix of the dental data, described in Chapter 2. Additionally, the mean parameter values and the variance-covariance matrix (obtained from the dental data) used for the scenarios involving $p = 4$ are provided below.

$$\hat{\mathbf{B}} = \begin{bmatrix} 20.28 & 20.80 \\ 0.95 & 1.65 \end{bmatrix}$$

,

$$\mathbf{S}_p = \begin{bmatrix} 5.42 & 2.72 & 3.91 & 2.71 \\ 2.72 & 4.18 & 2.93 & 3.32 \\ 3.91 & 2.93 & 6.46 & 4.13 \\ 2.71 & 3.32 & 4.13 & 4.99 \end{bmatrix}$$

.

The element-wise bias for one of the motivating scenarios is presented in Figure 3.1, where the solid black line represents zero bias and the broken lines represent the empirical 95% confidence intervals (CI) corresponding to the bias obtained from each of the three methods.

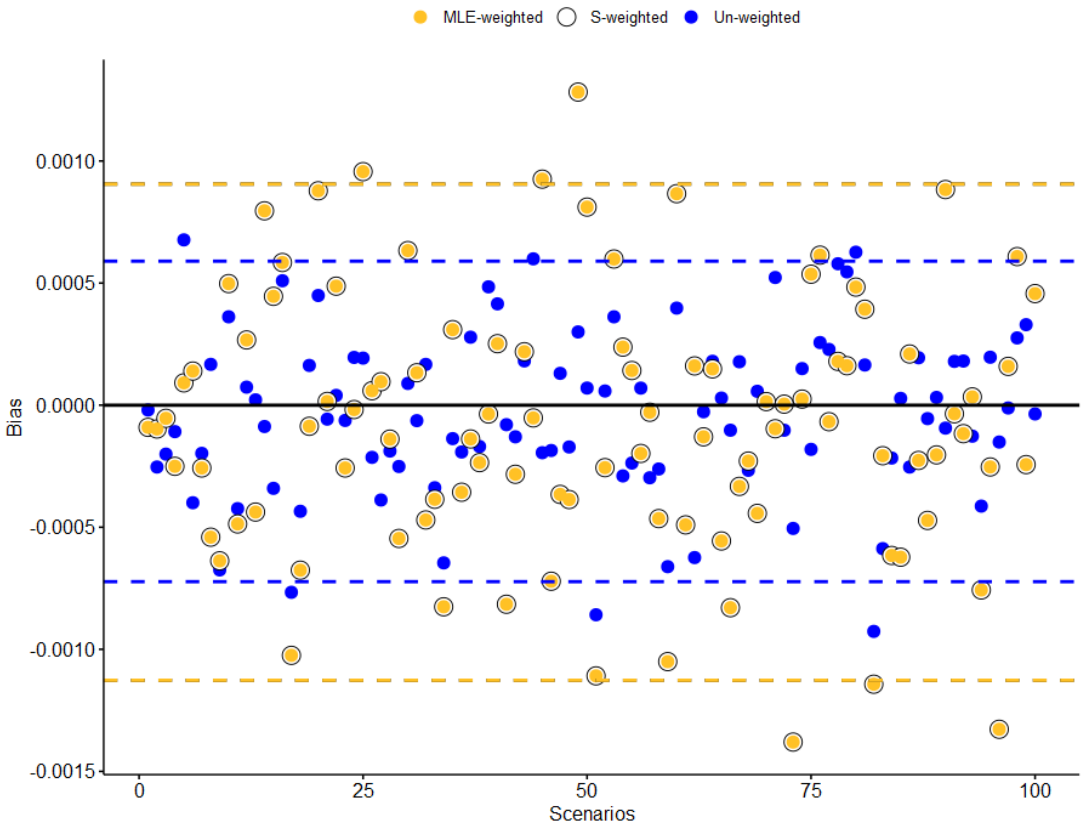

Figure 3.1: Element-wise bias of $b_{1,2}$ calculated for 100 replications of the same scenario. Each replication is generated 1,000 times to calculate the element-wise bias corresponding to each method. $n = 50$ and $p = 8$ is used.

The results in Figure 3.1 clearly show that bias is randomly distributed around zero, for all methods, indicating that the methods are unbiased. The results also show that the distribution of bias corresponding to the unweighted approach is slightly less variable than the weighted approaches, indicating that the unweighted estimators are less variable and more favourable. The results are consistent for all of the elements of the matrix $\hat{\mathbf{B}}$, different

degrees of polynomials but are not consistent between different sample sizes and different values of p (Figure 3.2 and 3.3, Table 3.1). We will explore these relationships in more detail in Table 3.2 when we explore the behaviour of the MSE as well as in subsequent sections. For large sample sizes, all four methods perform similarly with respect to Bias. Bias from additional scenarios are included in Table 3.1.

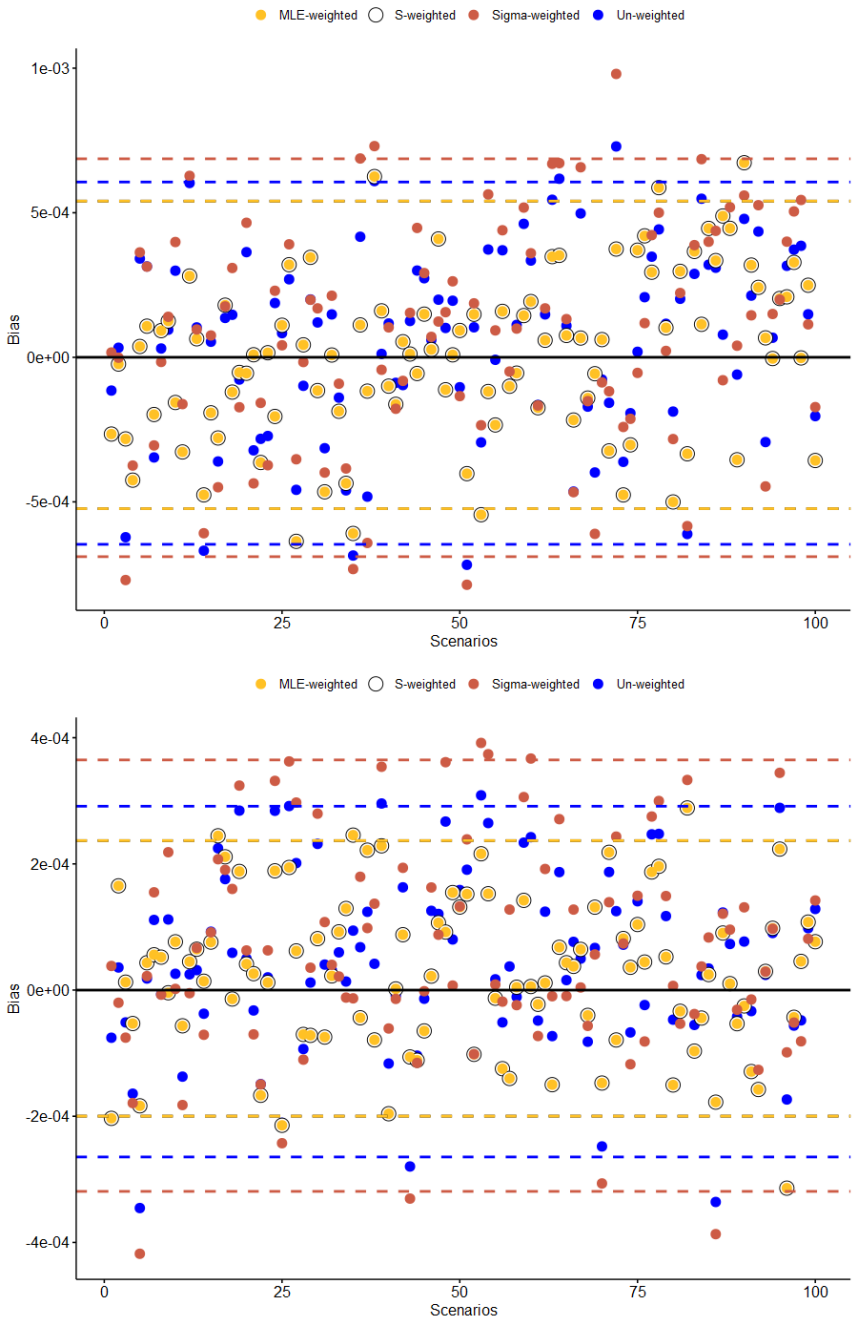

Figure 3.2: Element-wise bias of $b_{1,2}$ calculated for 100 replications of the same scenario. Each replication is generated 1,000 times to calculate element-wise bias corresponding to each method. $n = 20, 100$ and $p = 8$ are used.

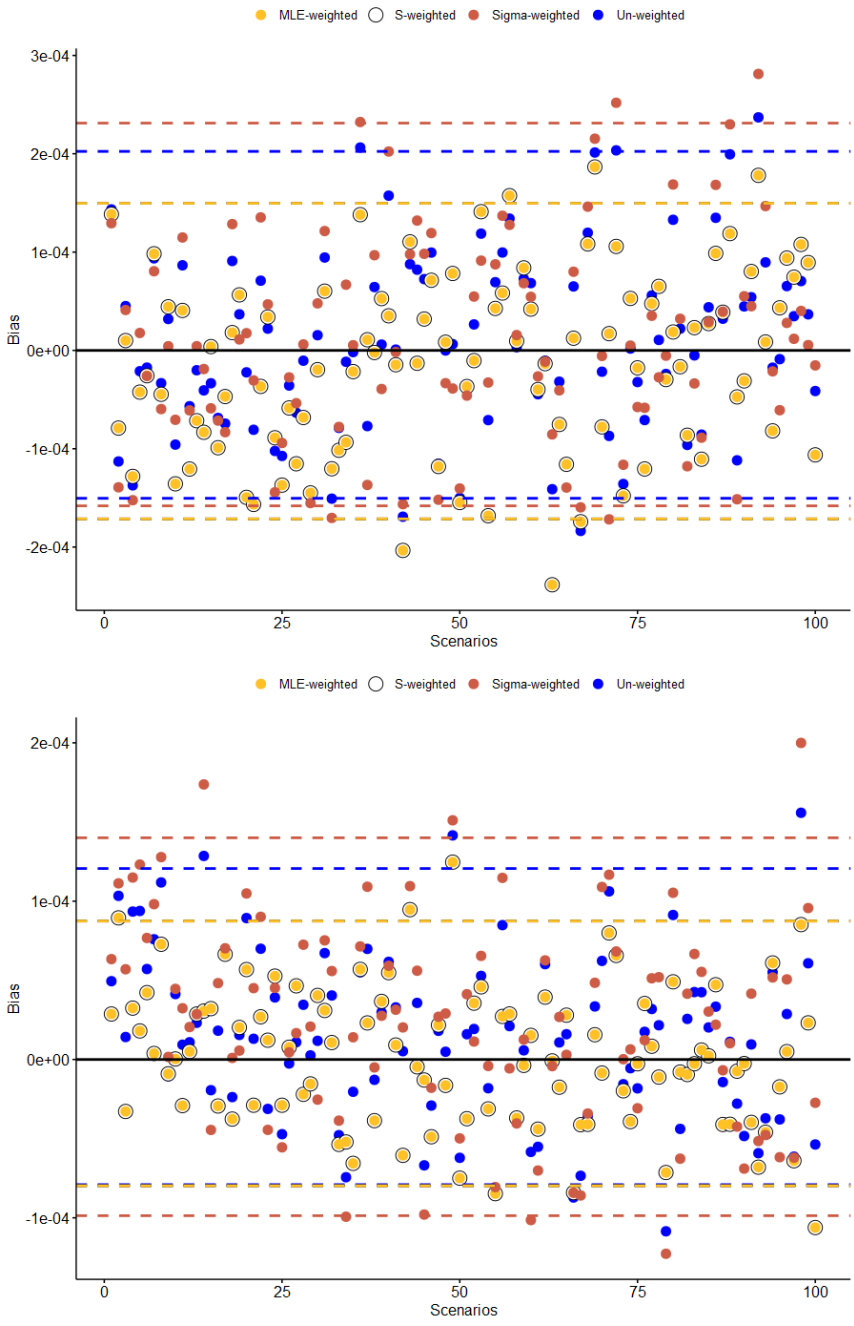

Figure 3.3: Element-wise bias of $b_{1,2}$ calculated for 100 replications of the same scenario. Each replication is generated 1,000 times to calculate element-wise bias corresponding to each method. $n = 200, 500$ and $p = 8$ are used.

In all scenarios in Figure 3.2 and Figure 3.3, it is clear that element-wise bias is very small, with the weighted methods (using the MLE and $\mathbf{S}$) presenting with slightly lower variability of the bias. Given the within-individual variability, we can hypothesize that this is the reason why the weighting methods perform better. Additionally, from these figures, it seems all of the methods are unbiased, meaning the results are centered around zero. To facilitate comparison, we considered Euclidean distance of the matrix of element-wise biases, and the results are presented in Table 3.1.

Overall, the results in Figure 3.2 and 3.3 show that there is no clear "best" method to use. It is evident, from Table 3.1 that the performance of the unweighted method between $n = 20$ and $n = 50$ improves drastically for $p = 8$. It is also evident that that performance is not uniform as we increase sample size. For instance, for Figure 3.2 and 3.3, we see that the weighted methods perform uniformly better, however in Figure 3.1, which was for $n = 50$ the unweighted method is slightly less variable than the weighted method and these results are consistent with the results in Table 3.2 where the Euclidean bias is large. Overall, there is no uniformly better method as the performance changes across sample sizes (see Table 3.1).

p	n	Unweighted	$\mathbf{S}$-Weighted	MLE-Weighted	$\mathbf{\Sigma}$-Weighted
4	10	3.042	9.993	9.993	7.521
4	20	0.639	0.651	0.651	0.505
4	30	1.025	0.713	0.713	0.397
4	40	0.094	0.102	0.102	0.140
4	50	0.278	1.087	1.087	0.525
4	60	0.364	0.496	0.496	0.471
4	70	0.057	0.061	0.061	0.040
4	80	0.087	0.040	0.040	0.043
4	90	0.058	0.038	0.038	0.033
4	100	0.058	0.010	0.010	0.017
8	10	0.765	1.318	1.318	0.351
8	20	0.022	0.220	0.220	0.124
8	30	0.069	0.036	0.036	0.024
8	40	0.096	0.007	0.007	0.015
8	50	0.012	0.097	0.097	0.079
8	60	0.072	0.010	0.010	0.022
8	70	0.003	0.009	0.009	0.008
8	80	0.023	0.013	0.013	0.015
8	90	0.065	0.001	0.001	0.003
8	100	0.020	0.019	0.019	0.012
8	200	0.004	0.007	0.007	0.004
8	500	0.009	0.004	0.004	0.007
8	1000	0.003	0.003	0.003	0.003

Table 3.1: Comparison of Euclidean Distances of the Bias Matrix of the Various Weighting Methods using Covariance Matrices Generated from the Dental ($p = 4$) and Glucose data ($p = 8$).

It is clear from the results that close to singularity for $p = 8$, the $\mathbf{\Sigma}$ method performs best. However, deviating from singularity, we have that different methods perform better at different sample sizes, with sporadic behaviour. It is clear from the results in Table 3.1 that close to singularity, the results are also inconsistent between different values of p. We have that the unweighted method performs better near singularity for $p = 4$ but the $\mathbf{\Sigma}$ method performs better near singularity for $p = 8$; however, as we increase the sample size, the weighting methods begin to perform better. As previously mentioned in the discussion of Figure 3.2 and 3.3, we have that the bias of the MLE- and **S**- weighted methods begin performing better than the unweighted method asymptotically. The high-variability of these methods and somewhat sporadic behaviour present in Table 3.1 should be explored further in subsequent papers given the nature of the covariance structure of the real data.

In Figure 3.4, we present element-wise MSE for all six elements of the estimator, for a scenario involving $p = 8$ and sample size ranging from $n = 6$ to $n = 100$ The results show that the weighted methods worked better for $n > p$ scenarios. Near singularity ($n \approx p$), we see some erratic behaviour corresponding to the **S**-weighted and MLE-weighted estimators, which we will explore in subsequent sections. We can also clearly see from Figure 3.4 that the unweighted method performed better in high-dimensional scenarios ($n < p$). Additional sample sizes are considered in Table A.3.

Figure 3.4 shows that the MSE corresponding to all methods decreases as the sample size increases, indicating that the estimators are consistent. In order to see the difference between the methods clearly, we focus on selected sample sizes, and these modified MSE plots are presented in Figure 3.5. The figure clearly shows that, despite the small differences, the weighted methods are uniformly better than the unweighted method for non high-dimensional scenarios. Of course, $\mathbf{\Sigma}$-weighted, being the gold standard, is uniformly better than all other methods. This confirms Potthoff and Roy's (1964) assertion that the estimators used are more efficient if the weights are chosen close to the true variance-covariance matrix, as is evident for this specific scenario.

We observed that both **S**- and MLE-weighted estimators performed the best asymptotically up to near singularity, more specifically beginning around $n = 20$ (note: $\frac{p}{n} = \frac{8}{20} = \frac{2}{5}$) . Clearly, additional simulations are required to assess performance under high-dimensional scenarios, since $p = 8$ does not provide us enough room to be able to assess more sample size options for high-dimensional scenarios.

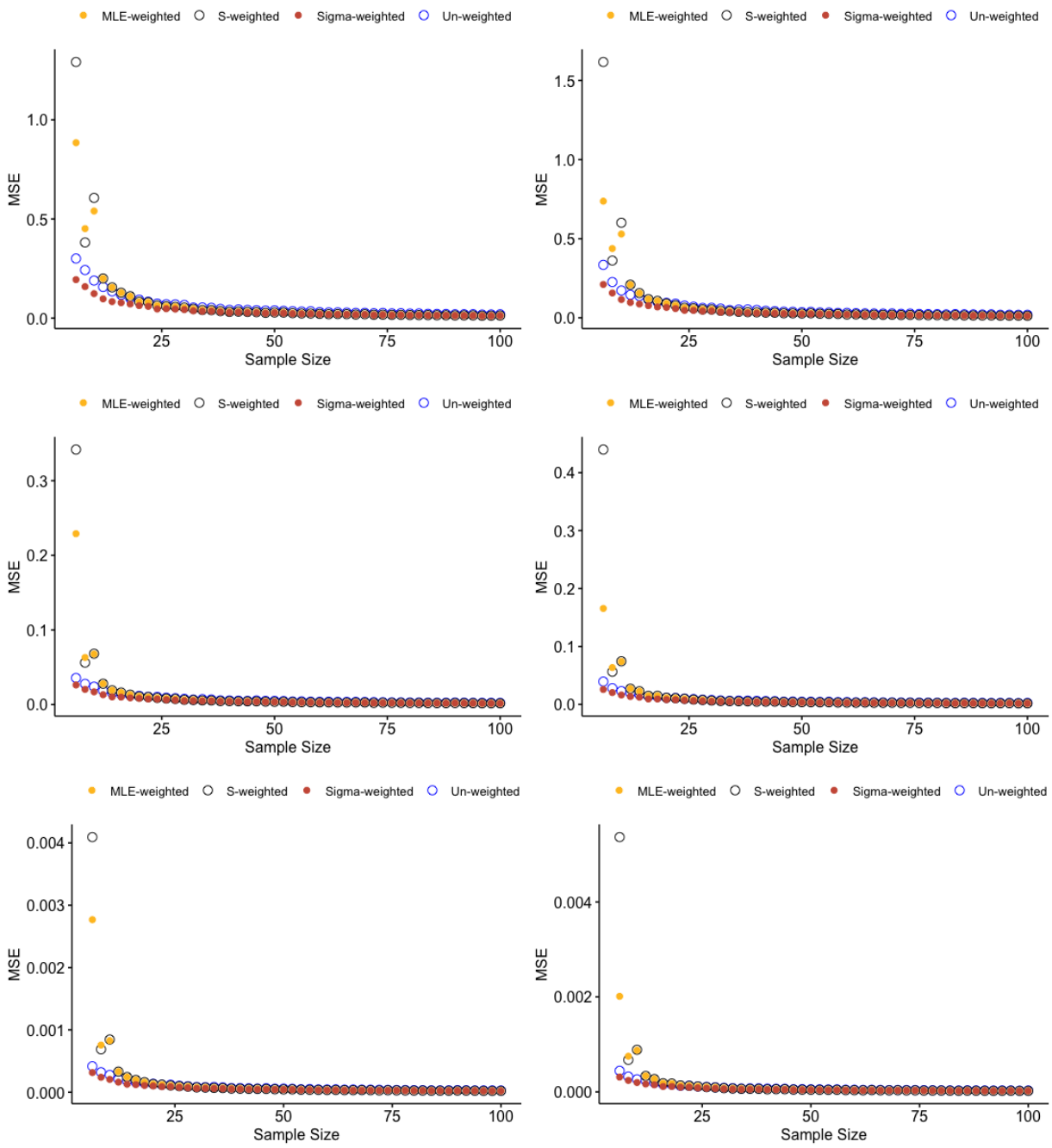

Figure 3.4: Element-wise MSE of $\hat{\mathbf{B}}$ calculated for 1,000 replications of the same scenarios. An average was generated to calculate the element-wise MSE corresponding to each element. $n = 14 - 100$ and $p = 8$ are used.

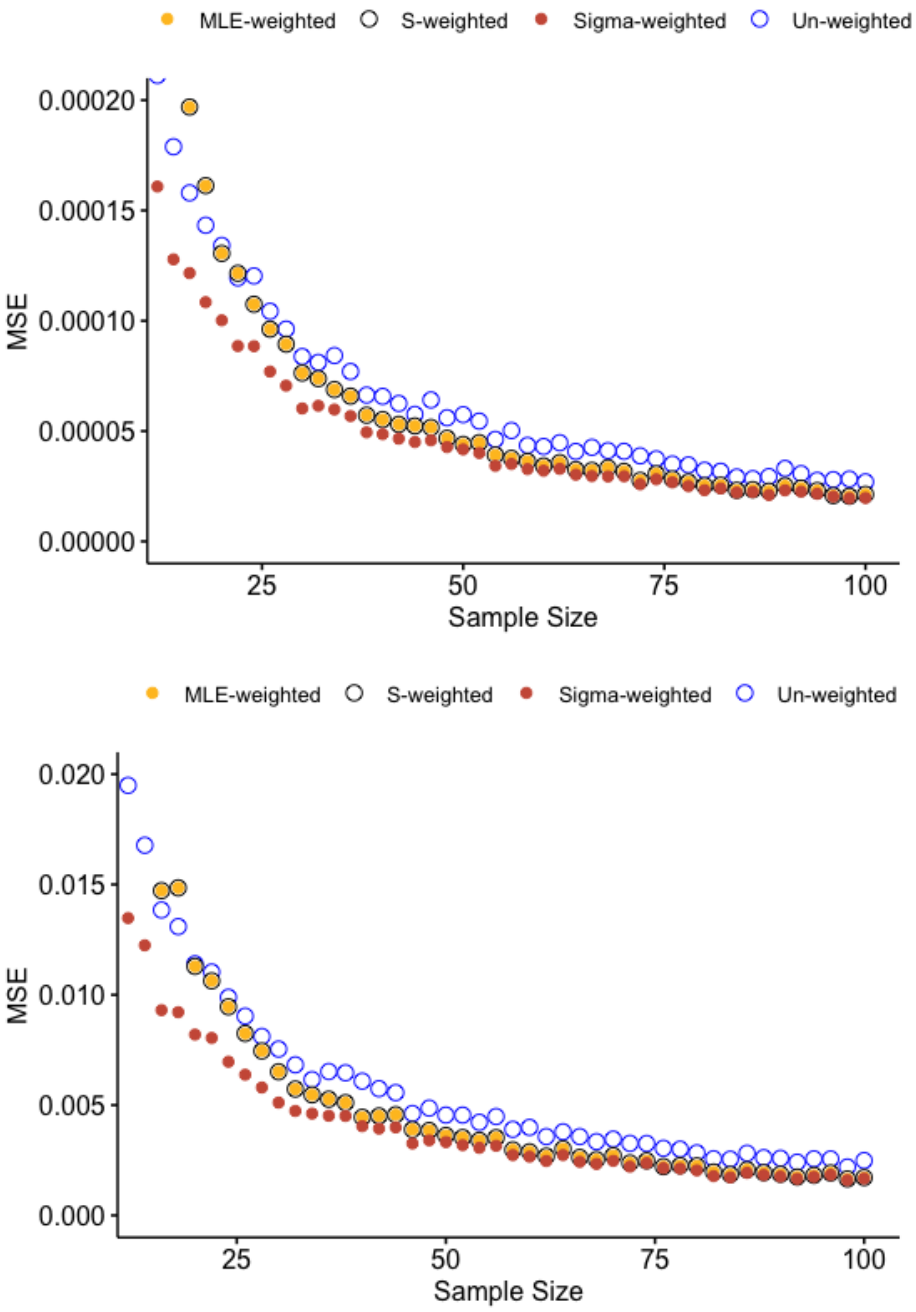

Figure 3.5: Element-wise MSE for $b_{1,2}$ and $b_{2,1}$ calculated for 1,000 replications of the same scenarios. An average was generated to calculate the element-wise MSE corresponding to each element. $n = 10 - 100$ and $p = 8$ are used.

Similarly to bias, we calculated the Euclidean distance of the matrix consisting of element-wise MSEs, and the results are presented in Table 3.2 for near singularity and otherwise, results are in Table A.3. Although we are presenting the results for $p = 4$ and $p = 8$, the results are consistent for other values of p and will be discussed in the next sections.

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
4	6	7.578	43.030	42.298	6.996
4	8	5.885	8.759	8.759	5.419
4	10	4.562	5.817	5.817	4.216
8	6	0.453	2.143	1.186	0.289
8	8	0.334	0.532	0.635	0.225
8	10	0.258	0.859	0.763	0.170

Table 3.2: Comparison of Euclidean Distances of the MSE Matrix of the Various Weighting Methods using Covariance Matrices Generated from the Dental ($p = 4$) and Glucose data ($p = 8$) showing behaviour near singularity.

It is clear from Table 3.2 that near singularity, all methods perform poorly. Additionally, it can be seen that close to singularity, weighting with the true variance-covariance matrix does not perform well either. This is perhaps due to other factors related to the variability (or instability) of the weighted estimators, rather than the weighting mechanism itself, and hence requires further investigation. On the other hand, in high-dimensional scenarios, this is not be the case and will be discussed in the next section. In the figures and tables presented in this section, we can clearly see that the $\mathbf{S}$-weighted and MLE-weighted methods are equivalent to the performance of the true variance-covariance matrix asymptotically, suggesting that in real life scenarios, where the true variance-covariance matrix is unknown, weighting using the estimators of $\mathbf{\Sigma}$ provides the best alternative, especially asymptotically. Nevertheless, we will show later that this might not be the case for all covariance structures. We would also like to highlight that the $\mathbf{S}$-weighted and MLE-weighted estimators performed almost exactly the same, save for near singularity. This leads us to believe that in a well-fitted model, the impact of the R_3 residual is next to none. Exploring this would be beneficial for similar studies to this paper.

3.2.2 Singularity and Near Singularity

High-dimensional scenarios can happen when the dimension of the data (e.g. number of repeated measurements, the number of covariances etc.) is close to ($n \approx p$) or larger than

the number of observations in question ($p > n$). Singularity and near singularity brings additional problems to statistical inference, which, as we saw in our simulation results, was the case for our weighted estimators involving $\mathbf{S}$ and the MLE estimator of $\boldsymbol{\Sigma}$. More specifically, we saw in the previous section that these weighted estimators became unstable near singularity and they performed worse than the unweighted estimator, which ignores the dependency in the data. In this section, we will explore this further using $p = 40$, which will allow us to include additional sample size scenarios in near singularity and high-dimensional settings.

We generated the true variance-covariance matrix $\boldsymbol{\Sigma}$ from the moderate covariance structure which we further study in Section 3.4 and considered sample sizes of $n = 12; 24; 36$ for the bias plots presented in Figures 3.6 and $n = 30 - 50$ and $n = 12 - 38$ (increasing by 2) for the MSE plots presented in Figures 3.7 and 3.8. Note that when $\mathbf{S}$ is singular, the Moore-Penrose generalised inverse is used. Additional simulations involving high-dimensional scenarios are included in Table 3.3 and Tables A.4 and A.5 for different covariance structures. In those scenarios, $\boldsymbol{\Sigma}$ was generated following the paper by Jana and colleagues (Jana et al., 2017), where we were able to control the level of dependency (correlation). These variance-covariance matrices are defined as weak (0.2-0.4), moderate (0.5-0.65) and strong (0.7-0.9) based on the off-diagonal covariance terms. We also considered different levels of dependency for cases where $n > p$, where the results are presented in Section 3.2.3.

The bias results in Figure 3.6 clearly show that bias is randomly distributed around zero for most methods, with slightly higher variability when using the $\mathbf{S}$-weighted and MLE-weighted methods. The results also show that the distribution of bias corresponding to the unweighted and $\boldsymbol{\Sigma}$-weighted appear to be minimally better with respect to their distribution around zero, and are less variable than the weighted approaches using covariance estimators. For $n = 12$ and $n = 36$, all of these methods are unbiased. The results are consistent across the elements of the matrix $\hat{\mathbf{B}}$, the various different values of p we considered, as well as the different degrees of polynomials and different sample sizes included in our simulation scenarios (Table 3.3).

In Figure 3.6, with all sample sizes considered, the unweighted and $\boldsymbol{\Sigma}$-weighted methods perform uniformly better. Given the nature of the model and the fact that its performance under singularity conditions is poor; to see that the unweighted method is still unbiased for sample sizes $n = 12$ and $n = 36$ and outperforms the weighted methods is an important note to make. Especially considering that the unweighted method ignores the dependency of the data across time, and this performance is contrary to what has been indicated in the literature. What is consistently interesting is the fact that the $\boldsymbol{\Sigma}$ method performs well for all scenarios considered thus far.

It is clear to see that in all these cases, there is slight skewness with respect to the bias, as previously mentioned, save for the results presented in the previous paragraph about $n = 12$ and $n = 36$. The variability in the distribution of bias observed for the weighted examples (both **S**-weighted and MLE-weighted) could be due to the fact that there is erratic behaviour in the weighted methods in high-dimensional and near singular scenarios, which can be seen in Section 3.2.1 as well as further shown in the MSE plots presented in Figure 3.7. This is even more evident in the Euclidean distances of bias and MSE provided in Table 3.3 and Table 3.4 in the next section. Additionally, given that the true $\mathbf{\Sigma}$ is never known in real-world situations, we can safely say that in high-dimensional scenarios, the unweighted method performs the best uniformly.

The MSE plot provided in Figure 3.7 highlights that the **S**-weighted and MLE-weighted estimators are unstable near singularity. This might have affected the distribution of bias that we saw in Figure 3.6. Using several additional scenarios, we have indeed confirmed that for scenarios involving $n > p$, biases for all the methods are randomly distributed around zero, indicating unbiasedness. The unstable behaviour of the weighted estimators is partially derived by the stochastic behaviour of the Moore-Penrose inverse, in particular when near singularity situations (Imori and von Rosen, 2019) which we had previously mentioned.

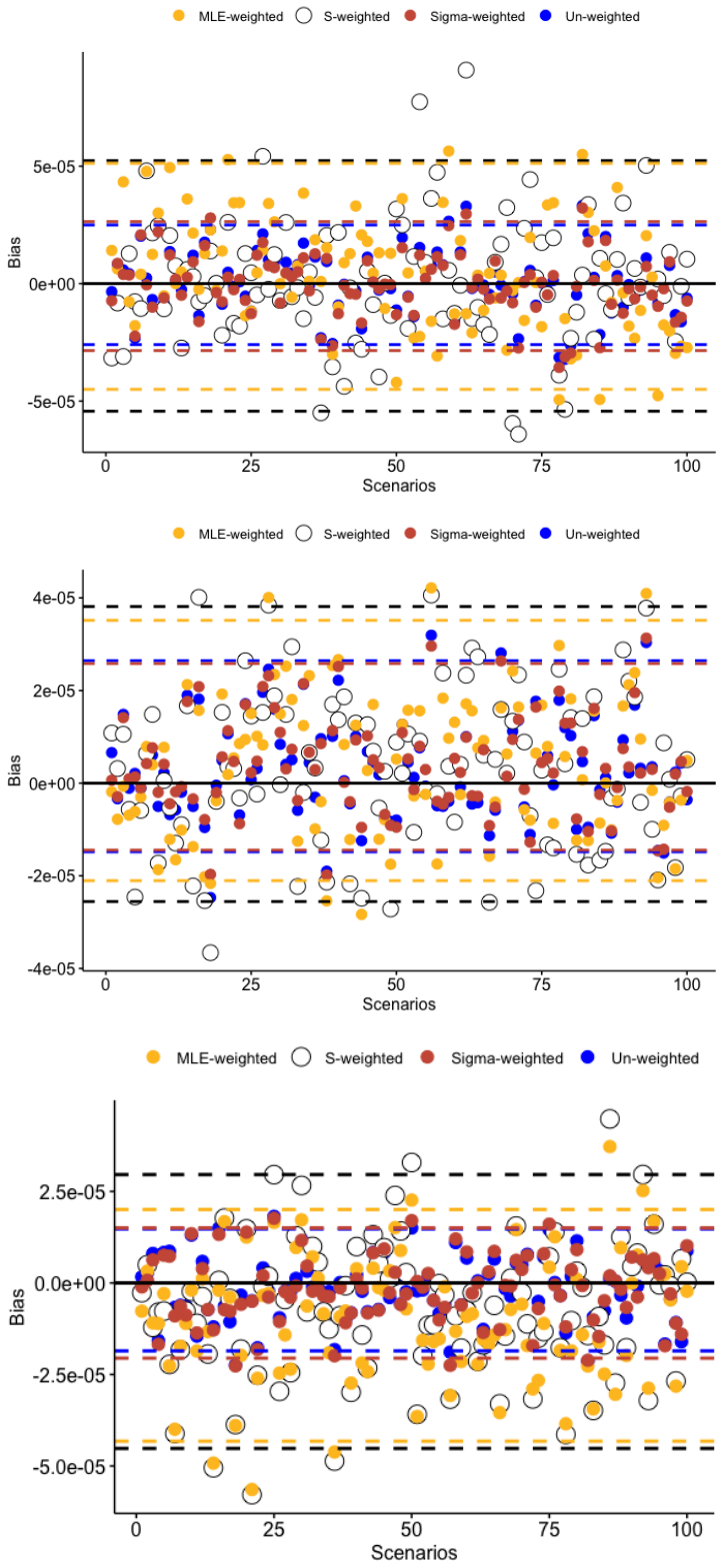

Figure 3.6: Element-wise bias of $b_{1,2}$ calculated for 100 replications of the same high-dimensional scenarios. Each replication is generated 1,000 times to calculate element-wise bias corresponding to each method. $n = 12, 24, 36$ (from left to right) and $p = 40$ are used.

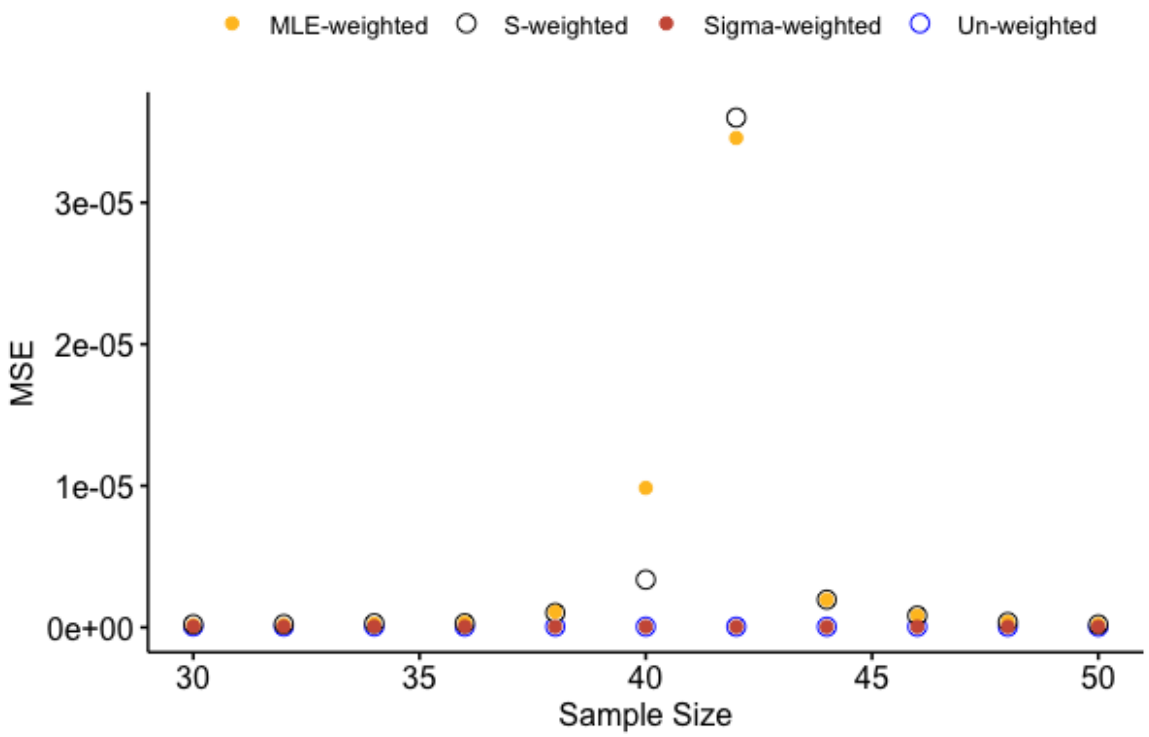

Figure 3.7: Element-wise MSE for $b_{1,2}$ calculated for 1,000 replications of the same high-dimensional scenarios. An average was generated to calculate the element-wise MSE corresponding to this element. $n = 30 - 50$ and $p = 40$ are used.

In Section 3.2.1, we saw from the MSE plots that the weighted estimators ($\mathbf{S}$ and MLE) and the $\mathbf{\Sigma}$-weighted estimator performed better when $n > p$. We also mentioned that this appears to be reversed in high-dimensional scenarios, where we observed uniformly better MSE corresponding to the unweighted method. This can be clearly seen when we use large values of p, such as $p = 40$, where the results in this high-dimensional scenario are presented in Figure 3.8. The figure also clearly shows where the MSE starts to increase near singularity for the weighted methods, which we saw behaviour of in Figure 3.7.

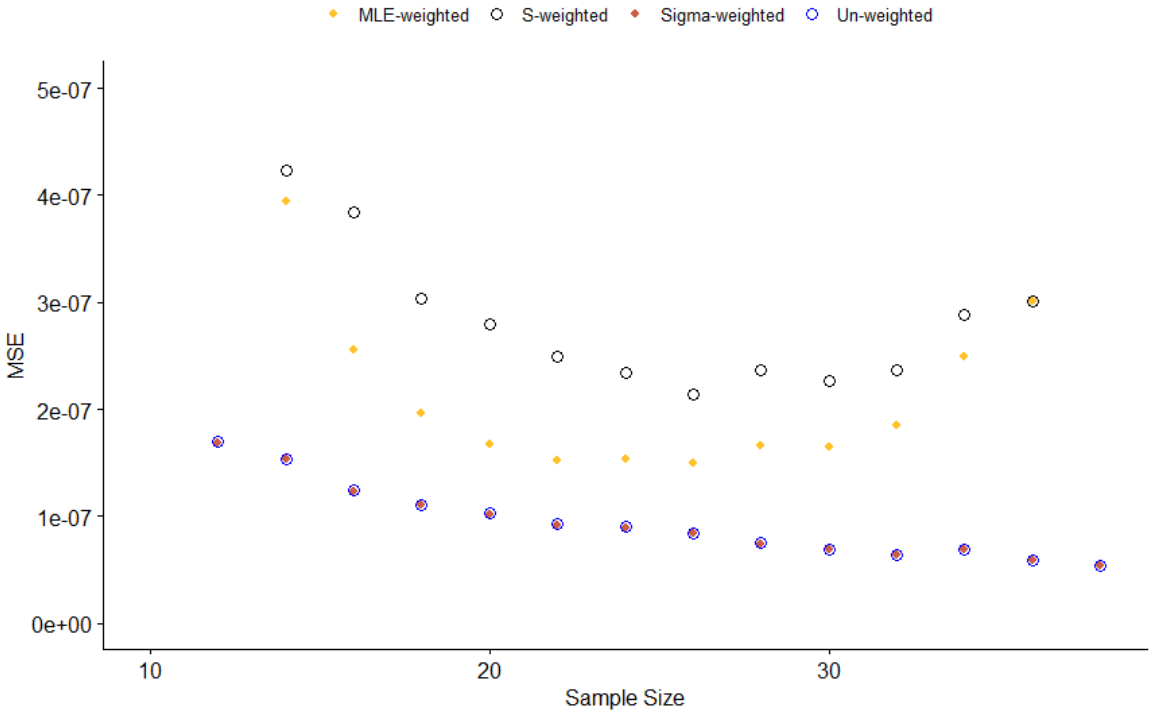

Figure 3.8: Element-wise MSE for $b_{1,2}$ calculated for 1,000 replications of the same high-dimensional scenarios. An average was generated to calculate the element-wise MSE corresponding to this element. $n = 12 - 38$ and $p = 40$ are used.

3.2.3 The Effect of Covariance

In the motivating example (Section 3.2.1), we used variance-covariance matrices (for the error matrix) generated from real data sets. Both covariance matrices used appeared to be unstructured. In the subsequent scenario with $p = 40$ (Section 3.2.2), we opted for a structured variance-covariance matrix, as we had seen that no method performed uniformly well with an unstructured covariance matrix and our objective was to simply have more range in sample sizes to allow us to see the high-dimensional and near singularity scenarios. In this section, we want to explore different levels of dependency in the data, for different structured covariances. Since it is difficult to create a positive-definite, unstructured, variance-covariance matrix with the desired levels of dependency, we resorted to using covariance matrices with the Toeplitz structure (code in Appendix B). We followed the approach used by Jana and colleagues (Jana et al., 2017) in creating variance-covariance

matrices with weak (0.2-0.4), moderate (0.5-0.65) and strong (0.7-0.9) correlations.

For the three different dependency levels, we considered $p = 40$ and $n = 50; 100$ to generate the bias plots and $n = 10 - 300$ for the MSE plots, a sufficiently wide range of sample size values to allow us to see large sample behaviours of the estimators as well as see behaviour in the singularity and near singularity scenarios. The results are presented in Figures 3.9 and 3.10. Euclidean distances of bias and MSE corresponding to these scenarios are provided in Tables 3.3, 3.4, A.1, A.2, A.4 and A.5, respectively.

Evidently, the unweighted and $\mathbf{\Sigma}$ methods perform uniformly better in terms of element-wise bias and are unbiased no matter the covariance structure used. Further to this section, we will focus solely on the moderate covariance structure, with other structures results available in Tables A.1 - A.5. The overall spread of these weighting methods and their performance is less clear when we move from $n = 50$ to $n = 100$. That is, the asymptotic behaviour of the weighted method allows them to perform better as the sample size increases. For $n = 50$, the unweighted and $\mathbf{\Sigma}$ methods are unbiased. However, the weighted methods are unbiased but slightly more variable for each sample size and method considered.

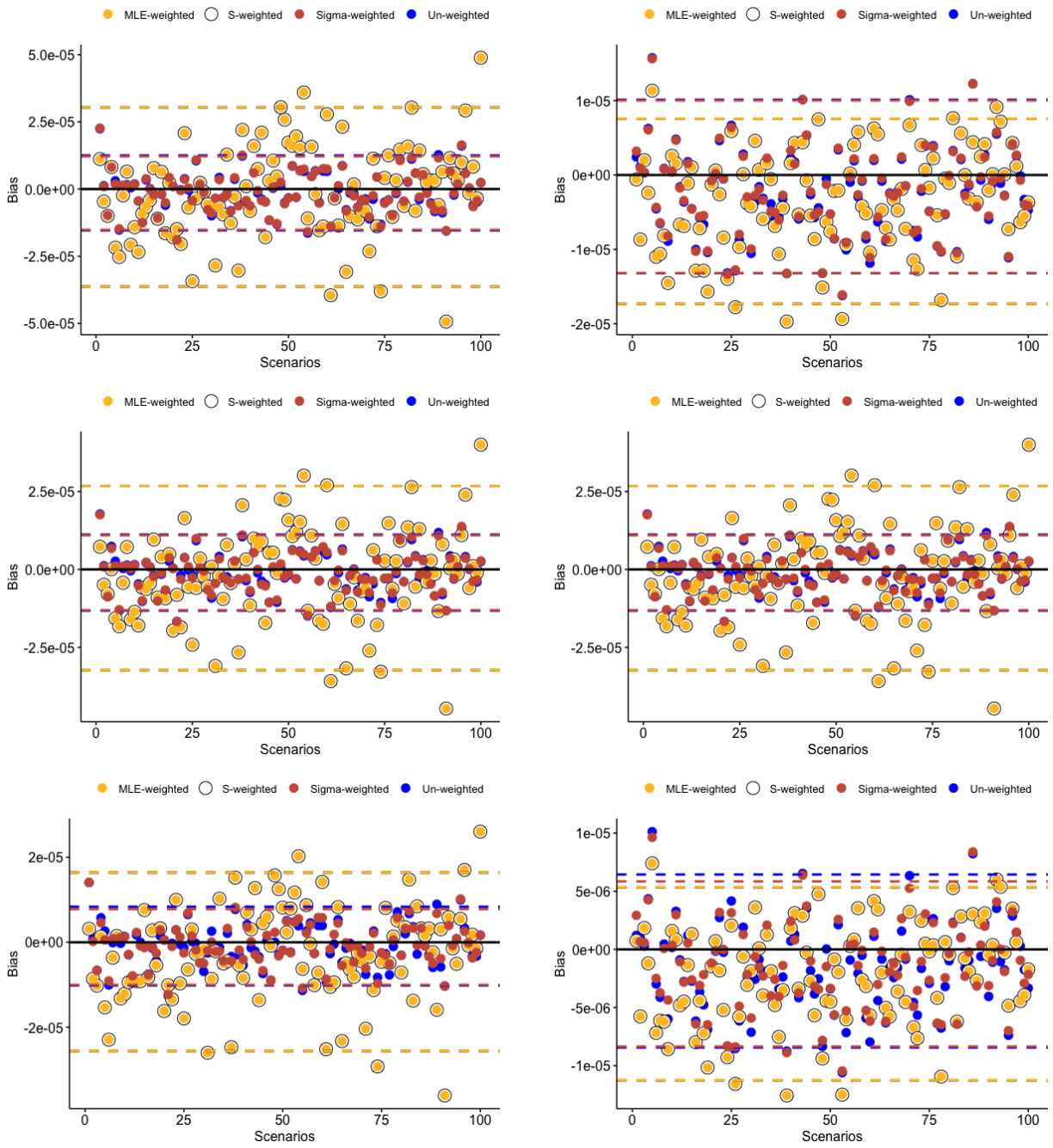

Figure 3.9: Element-wise bias of $b_{1,2}$ calculated for 100 replications of the same high-dimensional scenarios. Each replication is generated 1,000 times to calculate element-wise bias corresponding to each method. First row corresponds to weak covariance, second to moderate, and final row to strong covariance. $n = 50, 100$ (from left to right) and $p = 40$ are used.

We can see from Figure 3.9 that the bias is uniformly better for the unweighted and $\boldsymbol{\Sigma}$ methods, as previously discussed. Additionally, there seems to be less variability with respect to the bias across scenarios corresponding to these methods. This is also reflected in the MSE plots presented in Figure 3.10, where the unweighted method is uniformly better than both the **S**-weighted and MLE-weighted estimators, across all sample sizes considered. This is also further confirmed in Figure 3.11 with the asymptotic behaviour of the MSE of the moderate covariance method observed. This is true whether we have weak, moderate or strong correlations across the different time points, and whether the variances are small or large. Table 3.3 provides additional samples for the bias results.

p	n	Unweighted	**S**-weighted	MLE-weighted	$\boldsymbol{\Sigma}$-weighted
40	10	0.547	0.260	0.700	0.594
40	20	0.272	0.501	0.733	0.280
40	30	0.115	0.195	0.163	0.099
40	40	0.020	489.508	489.787	0.023
40	50	0.189	0.108	0.108	0.192
40	100	0.001	0.019	0.019	0.002
40	150	0.007	0.040	0.040	0.006
40	200	0.004	0.006	0.006	0.005
40	250	0.006	0.013	0.013	0.006
40	300	0.019	0.013	0.013	0.020
40	350	0.009	0.004	0.004	0.009
40	400	0.003	0.002	0.002	0.003
40	450	0.010	0.010	0.010	0.010
40	500	0.008	0.004	0.004	0.007

Table 3.3: Comparison of Euclidean Distances of the Bias Matrix of the Various Weighting Methods using a Covariance Matrix with Moderate Dependency

The results from Table 3.3 are inconclusive. At $n = 10$, the weighted methods outperform the unweighted methods and then again around $n = 50$, which is most likely due to the sporadic behaviour of these methods in high-dimensionality, close to singularity and not their true accuracy being shown. However, the $\boldsymbol{\Sigma}$-weighted method outperforms all of the other methods from $n = 20$ to $n = 40$ and then the unweighted method begins to perform consistent with the $\boldsymbol{\Sigma}$ method with respect to overall bias around $n = 100$. However, at $n = 300$ and beyond, the overall bias of the weighted methods is superior. From the MSE figures presented below however (Figures 3.10 and 3.11), the MSE is consistently better for the unweighted method. This means that the overall variability of the weighted methods, which contribute to the MSE calculation, in the scenarios where they

have weaker bias is superior to that of the unweighted method as well as the $\mathbf{\Sigma}$-weighted method. We had previously noted the bias variability for the weighted methods which contributes to the overall MSE calculation. This result is different from what we observed in Sections 3.2.1, where we used unstructured covariance matrices and found that the MSE of the weighted methods outperformed the unweighted method very quickly after reaching non high-dimensional scenarios, and stayed consistently better.

In Section 3.2.1, we saw from the MSE plots that the $\mathbf{S}$-weighted and MLE-weighted estimators performed better when $n > p$. We also mentioned that this appears to be reversed in high-dimensional scenarios, where we observed uniformly better MSE corresponding to the unweighted and $\mathbf{\Sigma}$-weighted methods. Similarly to those scenarios, from Figure 3.10, we can see again that the unweighted method (along with the $\mathbf{\Sigma}$-weighted method) performs uniformly better for all sample sizes across the different covariance structures. This leads to the generalisation that all of the covariance structures perform the same for MSE, meaning that even with different levels of covariances, the performance of the estimation techniques is consistent and are shown in Tables A.4 and A.5. As such, to provide a final picture of the true behaviour of the MSE, we show Figure 3.11, noting the behaviour of the MSE in high-dimensional, non high-dimensional as well as asymptotic scenarios, for the moderate covariance matrix. It is absolutely evident that the unweighted method performs uniformly better for the structured covariances we chose, consistent with Section 3.2.2, which also used a structured covariance matrix, but was inconsistent with the motivating scenario, which used an unstructured covariance matrix, which led to different, more inconclusive results. Additional results highlighting this consistent behaviour are available in Appendix A for the weak and strong covariance matrices respectively. Figure 3.11 is shown below.

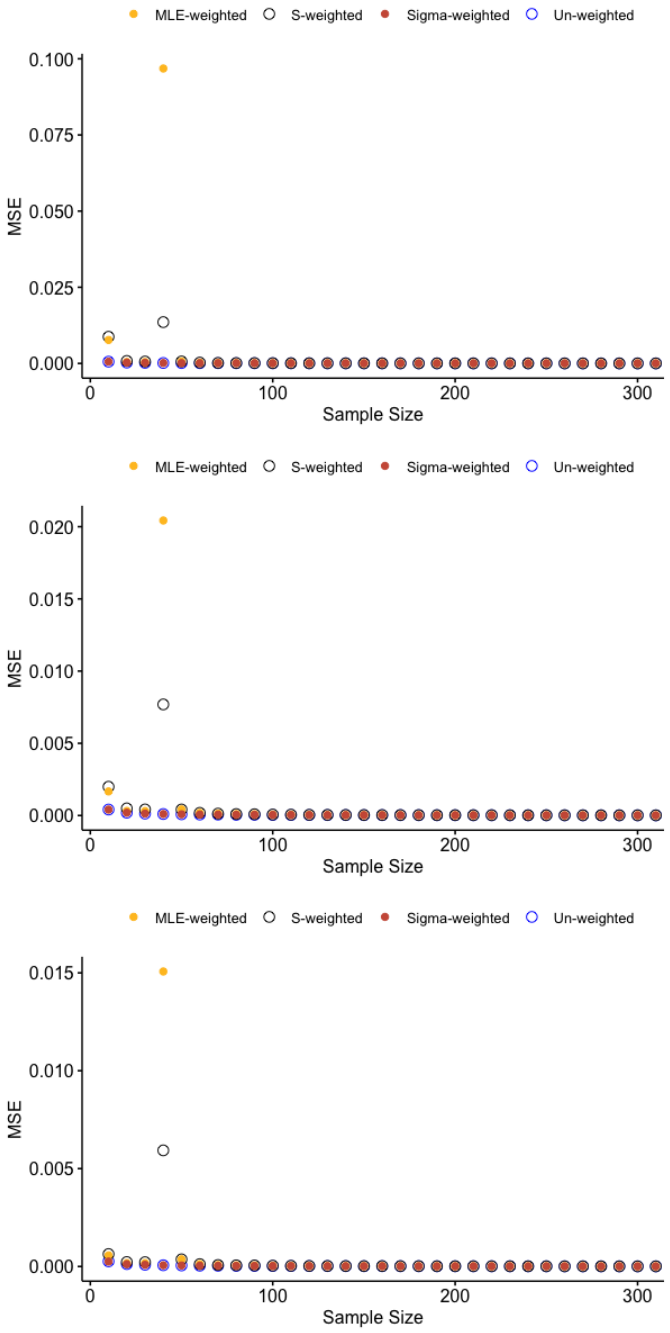

Figure 3.10: Element-wise MSE for $b_{1,2}$ calculated for 1,000 replications of the same scenarios. An average was generated to calculate the element-wise MSE corresponding to this element for various covariance matrices of weak, moderate and strong dependency. $n = 10 - 300$ and $p = 40$ are used.

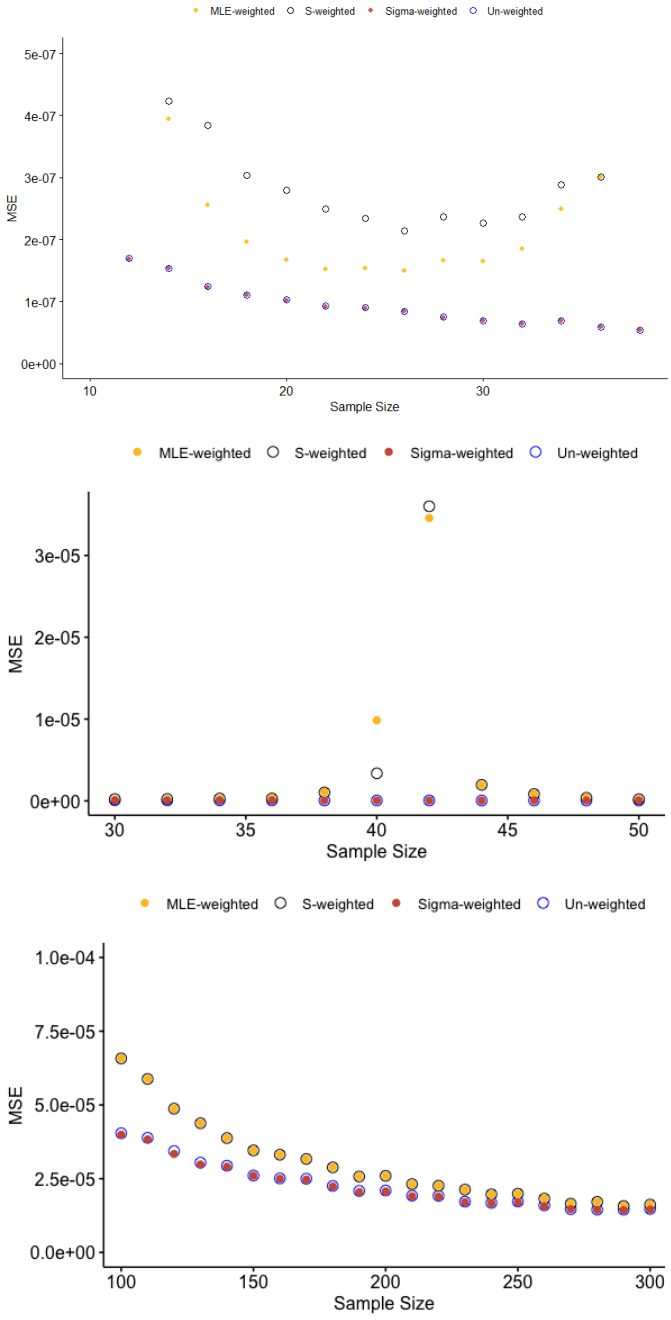

Figure 3.11: Element-wise MSE for $b_{1,2}$ calculated for 1,000 replications of the same scenarios. An average was generated to calculate the element-wise MSE corresponding to this element for a covariance matrix of moderate dependency across high-dimensional ($n = 12 - 38$), non high-dimensional ($n = 100 - 300$) as well as asymptotic ($n = 30 - 50$) cases. $p = 40$ is used.

Figure 3.11 confirms that, even asymptotically, the unweighted method performs better with these covariance matrices. These results are consistent with Figure 3.7 as well wherein the MSE starts to increase near singularity for the weighted methods. These results are also seen through Table 3.4 below and are further confirmed by other choices of the covariance matrices defined in Tables A.4 and A.5.

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
40	10	0.2120	0.3576	0.3295	0.2103
40	20	0.1038	0.1476	0.1368	0.1029
40	30	0.0668	0.1606	0.1400	0.0666
40	40	0.0499	32.1456	83.3184	0.0496
40	50	0.0409	0.9956	0.9956	0.0407
40	100	0.0207	0.0329	0.0329	0.0207
40	150	0.0134	0.0182	0.0182	0.0134
40	200	0.0105	0.0130	0.0130	0.0105
40	250	0.0084	0.0095	0.0095	0.0084
40	300	0.0071	0.0084	0.0084	0.0071
40	350	0.0063	0.0071	0.0071	0.0063
40	400	0.0055	0.0055	0.0055	0.0055
40	450	0.0045	0.0055	0.0055	0.0045
40	500	0.0045	0.0045	0.0045	0.0045

Table 3.4: Comparison of Euclidean Distances of the MSE Matrix of the Various Weighting Methods using a Covariance Matrix with Moderate Dependency

Table 3.4 further confirms our assumptions that the MSE of the unweighted method is consistently better than the weighted methods and that further to Potthoff and Roy's initial assumptions, the true variance-covariance weight $\mathbf{\Sigma}$ provides for a uniformly better estimate with respect to the MSE. These results are further shown for different variance-covariance matrices in Table A.4 and Table A.5 and are consistent with their results. Overall, the unweighted method seems to perform better with structured covariance methods, but does not have a consistently better bias and performs worse than the weighted methods with unstructured covariances. The unstructured behaviour of the methods should be explored further in subsequent papers to determine whether this result is consistent across covariance matrices.

Chapter 4

Real Data Examples

4.1 Motivating Scenario - Glucose Data

As described previously in Chapter 2, the glucose data was first analysed by Zerbe in 1979 and consists of 8 repeated measurements (at times 0, 0.5, 1, 1.5, 2, 3, 4, and 5 hours post administration of oral glucose) taken from 13 control (non-obese) and 20 obese ($n = 33$, $k = 2$, $p = 8$) individuals. The individual and mean profile plots are provided in Figure 2.2, Chapter 2. This will be the first data set to be analysed using the unweighted and weighted methods to assess estimations of these methods. The within- and between-individual design matrices, assuming a quadratic mean profile over time, are given by:

$$\mathbf{Z}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 1 & 1.5 & 2 & 3 & 4 & 5 \\ 0 & 0.25 & 1 & 2.25 & 4 & 9 & 16 & 25 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{1_{13}} & \mathbf{0_{20}} \\ \mathbf{0_{13}} & \mathbf{1_{20}} \end{pmatrix}$$

We present the element-wise estimators in Table 4.1, where confidence intervals for the element-wise estimators were calculated using bootstrapping. We took random samples (with replacement) of the 13 individuals from the obese groups and 20 individuals were sampled (again with replacement) from the control group. Re-sampling was done separately for the two groups so that we keep the within- and between- individual design matrices consistent across each method. We used 1000 bootstrap samples, from which we calculated the mean and 95% CI for the elements of $\hat{\mathbf{B}}$, which provides an empirical measure for precision. The results from this section were interpreted in light of the simulation scenarios presented in Chapter 3.

Method	Unweighted	**S**-weighted	MLE-weighted
$b_{1,0}$	3.705 (3.70, 3.71)	3.97 (3.96, 3.99)	3.97 (3.96, 3.99)
$b_{1,1}$	-0.74 (-0.745, -0.735)	-1.11 (-1.12, -1.10)	-1.11 (-1.12, -1.10)
$b_{1,2}$	0.17 (0.16, 0.17)	0.22 (0.21, 0.22)	0.22 (0.21, 0.22)
$b_{2,0}$	3.87 (3.86, 3.88)	4.02 (4.01, 4.04)	4.02 (4.01, 4.04)
$b_{2,1}$	-0.80 (-0.815, -0.795)	-1.04 (-1.06, -1.03)	-1.04 (-1.06, -1.03)
$b_{2,2}$	0.17 (0.17, 0.17)	0.20 (0.20, 0.21)	0.20 (0.20, 0.21)

Table 4.1: Element-wise estimates of $\hat{\mathbf{B}}$ and 95% CIs for Glucose Data

We can see from Table 4.1 that the weighted methods provided exactly the same estimators, indicating that the assumed quadratic mean profile is quite accurate. In other words, it means that $\mathbf{R}_3$, which represents the difference between the observed and estimated mean, is close to zero. We also see from Table 4.1 that the unweighted method provided slightly different estimators, but the differences in most cases were not big. From Section 3.2.1, in the simulation scenarios that mimicked the glucose data, we saw that the weighted estimators performed better in general, except in near singularity and high-dimensional situations. As such, we would be inclined to choose the weighted methods over the unweighted method as the weighted methods performed better at $n = 30$. However, it is important to note that the overall CIs are lower for the unweighted method. This behaviour will be assessed throughout the analysis of these sections. Our chosen estimator $\hat{\mathbf{B}}$ in this case would be:

$$\hat{\mathbf{B}} = \begin{pmatrix} 3.97 & 4.02 \\ -1.11 & -1.04 \\ 0.22 & 0.20 \end{pmatrix}$$

We can clearly see that the quadratic terms between the groups are different, which is consistent with the profile plots of Figure 2.2. Clearly, from the mean profile plot in Chapter 2 (Figure 2.2), we can see that the quadratic term for the obese group should be larger, meaning that the quadratic curve is more pronounced which is what we see in our estimator. Additionally, the intercept seems to be clearly defined accurately through this estimator, as seen in Figure 2.2.

All analysis was performed using Version 4.2.3 of the R statistical software (R Core Team, 2023).

4.2 Asymptotic Scenario - Cardiovascular Data

As an additional example, we considered a sample from the Framingham study. The original study aimed to evaluate the cardiovascular health of individuals living in Framingham, Massachusetts, over a long period of time (Lee et al., 2007). Although the data included many cardiovascular indicators, we focused on cholesterol measurements for two groups of individuals, those with diabetes and those who were not diabetic. Data consisted of $n = 2555$ individuals, of whom 261 had diabetes, but some involved missing data. We removed the missing data, since the objective was for illustration purposes only. Without the missing data, we still had a large sample size $n = 2025$ with 227 having diabetes. However, the results were clearly going to be asymptotic in this case, given the large value of n, and the true behaviour of the methods would be very similar. As such, we took a random sample of 150 individuals within the study and ended up with a sample of $n = 150$ with $n_1 = 19$ and $n_2 = 131$ representing the number of people with diabetes and those without, respectively. We used this random sample in the estimation of the $\hat{\mathbf{B}}$ estimator. Unbalanced sample sizes are not a topic of this thesis but could be considered in future papers.

Next, we plotted the individual and mean trajectory plots for this random sample, to ensure that both follow the same mean structure, to examine the overall trajectory of the mean and to determine an appropriate functional form (i.e. decide on the degree of polynomial $q-1$ that should be fitted). Although the mean structure might differ, as long as they still follow the same polynomial in time, we can continue with the GCM and not require the EGCM.

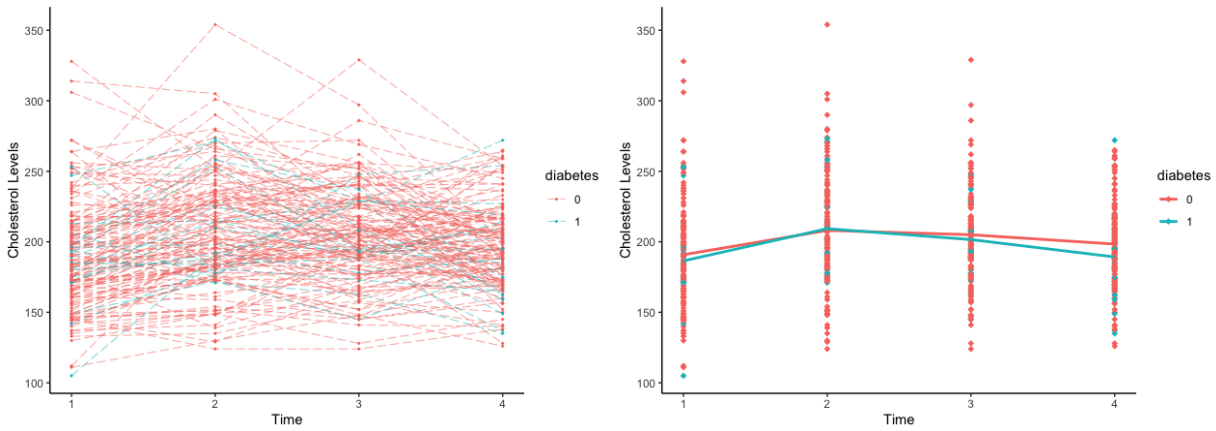

Figure 4.1: Individual Profile and Mean Profile Plots for the Framingham data from 19 diabetic (1) and 131 non-diabetic (2) individuals.

Indeed, it appears in Figure 4.1 that the means for both groups follow quadratic trajectories over time. We will look for this in the estimators derived from the analysis. The within- and between- individual matrices corresponding to the data are given by:

$$\mathbf{Z}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{1_{19}} & \mathbf{0_{131}} \\ \mathbf{0_{19}} & \mathbf{1_{131}} \end{pmatrix}$$

We expect all the methods to perform similarly given the asymptotic behaviour we saw in the simulations. However, we also saw in the simulations that in such scenarios (small p, large n) and unstructured covariance matrix, the weighted methods performed better. The covariance matrix of this method is:

$$\mathbf{\Sigma} = \begin{pmatrix} 0.66 & 0.49 & 0.34 & 0.18 \\ 0.49 & 0.79 & 0.47 & 0.25 \\ 0.34 & 0.47 & 0.62 & 0.31 \\ 0.18 & 0.25 & 0.31 & 0.63 \end{pmatrix}$$

Note also that the two weighted estimators provided exactly the same estimators, indicating that $\mathbf{R}_3$ was very small, hence indirectly confirming that the quadratic trajectory

over time fitted the data well. In this case, the CIs of the weighted method are better than the unweighted method, save for the CI for $b_{1,2}$. This result is interesting and the overall element-wise estimates are presented in Table 4.2

Method	Unweighted	**S**-weighted	MLE-weighted
$b_{1,0}$	176.96 (175.88, 178.04)	174.68 (173.63, 175.73)	174.68 (173.63, 175.73)
$b_{1,1}$	18.91 (17.82, 20.00)	16.83 (15.72, 17.93)	16.83 (15.72, 17.93)
$b_{1,2}$	-3.32 (-3.54, -3.10)	-2.63 (-2.86, -2.41)	-2.63 (-2.86, -2.41)
$b_{2,0}$	165.75 (165.33, 166.17)	164.86 (164.49, 165.24)	164.86 (164.49, 165.24)
$b_{2,1}$	31.39 (30.98, 31.81)	29.70 (29.34, 30.06)	29.70 (29.34, 30.06)
$b_{2,2}$	-5.83 (-5.92, -5.75)	-5.33 (-5.41, -5.26)	-5.33 (-5.41, -5.26)

Table 4.2: Element-wise estimates of $\hat{\mathbf{B}}$ and 95% CIs for Cardiovascular Data

Our chosen estimator $\hat{\mathbf{B}}$ in this case would be:

$$\hat{\mathbf{B}} = \begin{pmatrix} 174.68 & 164.86 \\ 16.83 & 29.70 \\ -2.63 & -5.33 \end{pmatrix}$$

We can clearly see that the quadratic terms between the groups are negative, which we expected from the mean plot presented in Figure 4.2. We can also see that the first row, the intercept, is defined accurately in the estimator.

4.3 Near Singularity Scenario - Lung Cancer Data

In this section, we considered gene expression measurements taken from a study examining human lung tissue, taken from a known and publicly available data set. The objective was to investigate if exposure to certain chemicals resulted in lung cancer. This data was also further analysed in the context of the GCM by Jana (2013). The data consists of $n = 9$ individuals, with three equal groups of 3 individuals. The groups include one control group, not exposed to any treatment plan as well as two different treatment groups. The study

was conducted to identify gene expression profiles among genes exposed to V_2O_5 as well as H_2O_2 (the treatments) and compare their expression with healthy (normal) lung tissue. These individuals are measured at $p = 5$ different time points, notably after 1 hour, 4 hours, 8 hours, 12 hours, and 24 hours of treatment. Thus, we have $n = 9$ and $p = 5$. For illustration purposes, we focus on the top 2 genes, from a gene filtering analysis done by Jana and colleagues (Jana et al., 2013). The individual and mean profile plots for the first gene, ribosomal protein S12 is given in figure 4.2.

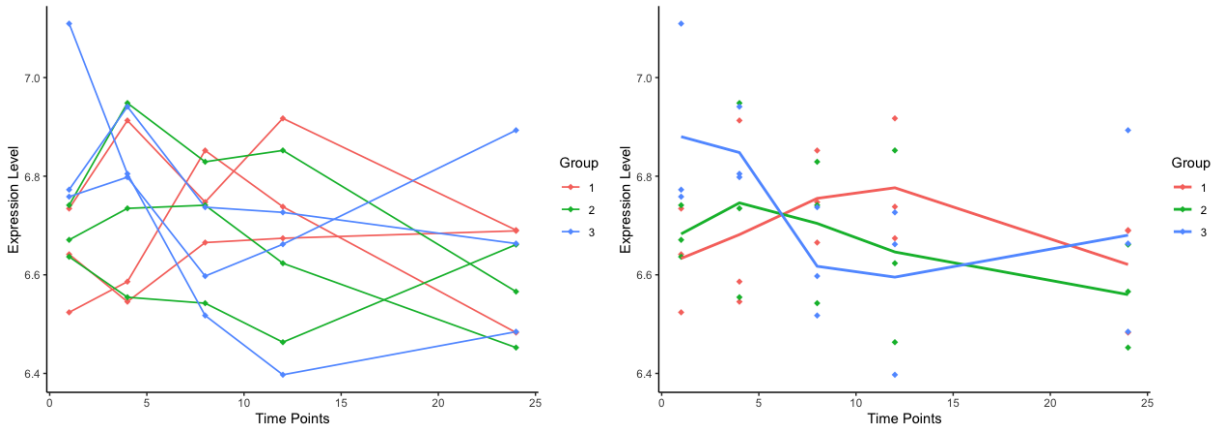

Figure 4.2: Individual Profile and Mean Profile Plots for Gene expression measurements of Ribosomal Protein S12 from control (1), V_2O_5 Treatment (2) and H_2O_2 Treatment (3)

We can see in Figure 4.2 that there is high variability within each group. However, the sample size is very limited to make any conclusive statements in that regard. Nevertheless, it appears that the means across the three groups may be modelled using quadratic mean trajectories over time, which is the same choice made by Jana (2013). As such, we used $q - 1 = 2$ and the within- and between-individual design matrices are given below:

$$\mathbf{Z}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 4 & 9 & 12 & 24 \\ 1 & 16 & 81 & 144 & 576 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{1_3} & \mathbf{0_3} & \mathbf{0_3} \\ \mathbf{0_3} & \mathbf{1_3} & \mathbf{0_3} \\ \mathbf{0_3} & \mathbf{0_3} & \mathbf{1_3} \end{pmatrix}$$

For the second gene we considered from the same study, using the same within- and between- individual design matrices, profile plots are given in Figure 4.3.

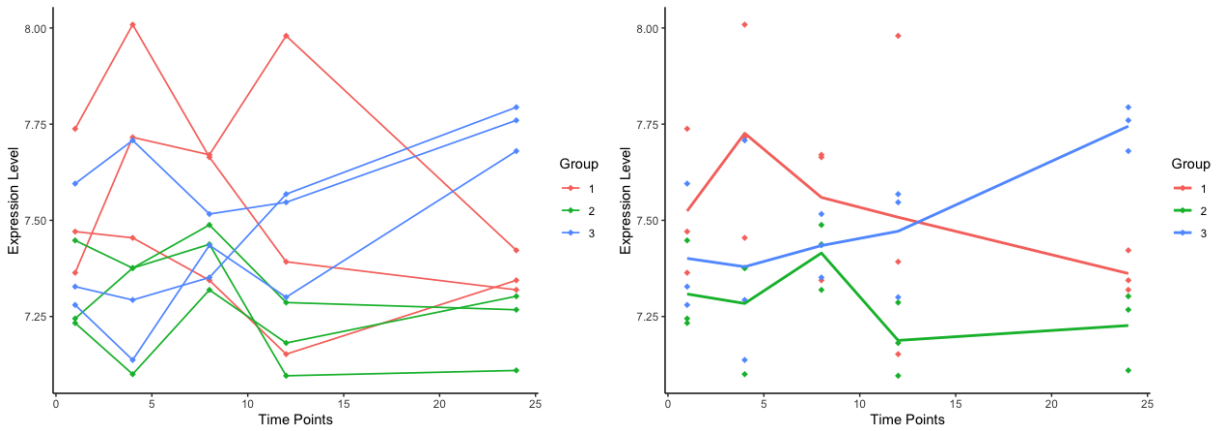

Figure 4.3: Individual Profile and Mean Profile Plots for Gene expression measurements for Proteasome (Prosome, Macropain) Subunit, Beta Type 7 from control (1), V_2O_5 Treatment (2) and H_2O_2 Treatment (3)

Figure 4.3 appears to be quite different from Figure 4.2, where in this case the 3 mean trajectories seems to follow different shapes. Although the EGCM might be more appropriate for such data, we still considered analysis using the GCM, mainly to illustrate how the estimators behave in such a scenario. In both of these cases, we expect different results than in the previous sections as we are in the near singularity scenario, where the weighted methods perform poorly compared to the unweighted method. Moreover, we expect the two weighted estimators to provide different results, in particular for the second gene. This is because the quadratic mean profile for all the 3 groups might not represent the true mean trajectory accurately, hence leading to large residuals, which will be reflected in the difference between the two weighted estimators.

Indeed, the results presented in Tables 4.3 and 4.4 show that the weighted estimators in general led to different estimators, and that the difference was more pronounced in Table 4.4. Given that we were in near singularity scenarios, it was evident from our simulation results (both using unstructured and structured covariance) that the unweighted method performed better. This was also confirmed from the width of the 95% bootstrap CIs, where the unweighted method led to narrower CIs, indicating more precision in the estimators.

Method	Unweighted	**S**-weighted	MLE-weighted
$b_{1,0}$	6.6015 (6.5306, 6.6725)	5.03 (2.2882, 7.7718)	5.0449 (2.3184, 7.7714)
$b_{1,1}$	0.0271 (0.0153, 0.0388)	0.2204 (-0.3731, 0.8138)	0.2292 (-0.3539, 0.8122)
$b_{1,2}$	-0.0011 (-0.0016, -0.0006)	-0.0055 (-0.035, 0.024)	-0.0064 (-0.034, 0.0212)
$b_{2,0}$	6.7092 (6.6805, 6.7379)	5.1128 (2.3246, 7.901)	5.1555 (2.3657, 7.9453)
$b_{2,1}$	0.0007 (-0.0117, 0.013)	0.1927 (-0.4117, 0.797)	0.1862 (-0.4213, 0.7937)
$b_{2,2}$	-0.0003 (-0.0009, 0.0003)	-0.0046 (-0.0346, 0.0253)	-0.0048 (-0.0334, 0.0239)
$b_{3,0}$	6.9629 (6.8648, 7.0611)	5.3117 (2.4109, 8.2124)	5.3332 (2.4465, 8.2199)
$b_{3,1}$	-0.0493 (-0.0731, -0.0255)	0.1516 (-0.4786, 0.7817)	0.1485 (-0.4712, 0.7683)
$b_{3,2}$	0.0016 (0.0008, 0.0023)	-0.003 (-0.0341, 0.028)	-0.0033 (-0.0326, 0.026)

Table 4.3: Element-wise estimates of $\hat{\mathbf{B}}$ and 95% CIs for Ribosomal Gene Data

Therefore, we will use the $\hat{\mathbf{B}}$ obtained from the unweighted analysis, which is given by:

$$\hat{\mathbf{B}} = \begin{pmatrix} 6.60 & 6.71 & 6.96 \\ 0.0271 & 0.0007 & -0.0493 \\ -0.0011 & -0.0003 & 0.0016 \end{pmatrix}$$

Similarly, it is clear from Table 4.4 we can see that the unweighted approach leads to a better estimator., where the 95% CIs are in general narrower, compared to the weighted methods. As mentioned above, we also see larger difference between the **S** -weighted and MLE-weighted estimators, indicating that the systematic component of the model might not be adequate to represent the mean trajectories over time. Nevertheless, given that the data scenario matches the near singularity scenarios we considered in our simulations, it is evident that the unweighted approach provides a better estimator than the weighted estimators.

Method	Unweighted	**S**-weighted	MLE-weighted
$b_{1,0}$	7.5958 (7.4517, 7.7399)	6.112 (3.2605, 8.9634)	6.0986 (3.3221, 8.8751)
$b_{1,1}$	0.0023 (-0.0329, 0.0376)	0.3084 (-0.4583, 1.0751)	0.2943 (-0.4389, 1.0275)
$b_{1,2}$	-0.0005 (-0.0018, 0.0007)	-0.0162 (-0.0962, 0.0638)	-0.0142 (-0.075, 0.0465)
$b_{2,0}$	7.319 (7.2437, 7.3943)	5.8851 (3.143, 8.6273)	5.9121 (3.218, 8.6062)
$b_{2,1}$	-0.0014 (-0.0133, 0.0105)	0.3028 (-0.4383, 1.0439)	0.2988 (-0.4229, 1.0205)
$b_{2,2}$	-0.0001 (-0.0005, 0.0002)	-0.0157 (-0.0928, 0.0614)	-0.0141 (-0.0697, 0.0415)
$b_{3,0}$	7.3914 (7.2749, 7.5079)	5.9838 (3.1935, 8.7741)	6.0851 (3.3519, 8.8184)
$b_{3,1}$	-0.0021 (-0.0108, 0.0066)	0.2949 (-0.4563, 1.0461)	0.2606 (-0.4455, 0.9667)
$b_{3,2}$	0.0007 (0.0005, 0.0009)	-0.0147 (-0.0931, 0.0637)	-0.0098 (-0.0476, 0.028)

Table 4.4: Element-wise estimates of $\hat{\mathbf{B}}$ and 95% CIs for Proteasome Gene Data

Thus, if the GCM is to be fitted for the model, we would prefer to use the unweighted estimator, which is given by :

$$\hat{\mathbf{B}} = \begin{pmatrix} 7.60 & 7.32 & 7.39 \\ 0.0023 & -0.0014 & -0.0021 \\ -0.0005 & -0.0001 & 0.0007 \end{pmatrix}$$

In summary, from these real life scenarios considered in this Chapter, we would choose the weighted methods when we are working in a unstructured covariance scenario that is not high-dimensional (Glucose Data, Section 4.1) or in an asymptotic scenario (Cardiovascular Data, Section 4.2). In large sample scenarios such as the cardiovascular data, all the three methods lead to estimators that are relatively the same in performance. On the other hand, in high-dimensional or near singular scenarios, we will always use the unweighted method, as the weighted methods are highly erratic and do not provide consistently better results than the unweighted method, which was shown in all sections of Chapter 3. Another observation worth highlighting is the difference between the **S**-weighted and MLE-weighted estimators. We saw from the gene expression data that the two methods might lead to different results. This is mainly because of $\mathbf{R}_3$ being large, which indicates that the fitted (assumed) mean trajectory over time might not be a good fit.

Chapter 5

Discussion and Future Directions

5.1 Summary and Discussion

We began this thesis by first providing context and motivation for the GMANOVA model, based on a generalisation of the MANOVA model. We then used this generalisation to introduce the motivation behind the estimators of the GMANOVA model, beginning with the Potthoff and Roy estimators, which are based on transformations of the GMANOVA model to MANOVA. In doing so, we introduced the concept of weighted or unweighted estimation. Throughout this thesis, our results indicated that the unweighted method is "fairly good" to be used under most scenarios we considered, with the only exceptions being in unstructured covariance situations. This is contrary to what was previously indicated in the literature, including in the original paper by Potthoff and Roy, where it was stated that weighted estimators that take the dependency of the variables into account should perform better in general.

The unstructured covariance matrices we considered were mainly based on real life scenarios, and included weak to strong within-individual (across time) correlations. We also considered unstructured covariance matrices generated from the Wishart distribution. In all these scenarios, the weighted methods performed better provided that we were in $n > p$ scenarios.

The unweighted method proved to be highly robust in high-dimensional scenarios (whether the covariance was structured or unstructured), and including scenarios involving high variability and scenarios with weak to strong within-individual correlations. For structured covariance matrices we considered, the unweighted method performed better across

all of the sample sizes, including near singular conditions and high-dimensional scenarios. For large samples, the performance of all of the methods were comparable.

Given that $\mathbf{S}_p$ and $\hat{\mathbf{\Sigma}}_{MLE}$ are both reasonably good estimators of the true variance-covariance matrix, and from what is indicated in the literature, we expected that the weighted methods would perform better than the unweighted, at least in most cases. Over 600 simulations were conducted, and it was indeed very evidently clear that the true variance-covariance matrix performed very well across most scenarios and was usually the best method to use. However, in real-world scenarios, this method is not feasible and thus our focus was on the other 3 methods discussed. What was surprising in our simulation results was that the unweighted method, which ignores dependency, performed almost the same as the $\mathbf{\Sigma}$-weighted method in a considerably large portion of the scenarios we considered. Although our simulation scenarios are not by any means exhaustive, it is important to see that there are scenarios in which we can simply use unweighted projections. This provides significant simplifications related to hypothesis testing and model diagnostics, especially in high-dimensional and near singular scenarios.

Our initial simulations using the variance-covariance matrices from the glucose and dental data sets made it seem as if the weighted methods performed the best. However, these initial results considered an unstructured variance-covariance matrix and laid the foundation to pose additional questions related to the performance of the methods. It provided us with the motivation for exploring high-dimensional scenarios as well as varied strengths of covariance between the individuals. Further analysis revealed that the unweighted methods continued to perform well for much of the simulations, including high-dimensional scenarios where it outperformed the weighted methods significantly. In particular, we saw that the weighted methods were highly erratic when $n \approx p$ (which we referred to as near singular scenarios). This could have been due to the nature of the Moore-Penrose generalised inverse used in the weighted estimators, since $\mathbf{S}^{-1}$ does not exist. There was a clear indicator that the unweighted method should be used in many scenarios, and would provide very good estimators.

In all the scenarios we considered, we saw that $\mathbf{S}$-weighted and MLE-weighted estimators performed the same. This may have been because we had not incorporated misfitting of the systematic component of the model into our simulations, hence we expected $\mathbf{R}_3$ (which represented the part of the residual related to model fit) to be close to zero. As such, there was no difference between the pooled variance-covariance estimate ($\mathbf{S}_p$) and the MLE of $\mathbf{\Sigma}$. The two weighted estimators led to slightly different results in high-dimensional scenarios, which was perhaps due to the shrinkage structure of the MLE, which kept it non-singular even when $\mathbf{S}$ was singular.

Finally, we would like to highlight that bias, for the majority of the scenarios considered, was randomly distributed around zero. This was the case even in high-dimensional scenarios, whether we had structured or unstructured covariance matrices, regardless of the size of variability and regardless of the strength of dependency. The only notable exception was near singularity, where as mentioned above we saw unstable behavior of the weighted estimators. The MSE for all of the methods, in all of the scenarios considered, decreased with increases to sample size, indicating consistency of the estimators, again, with the exception of MSEs from the weighted estimators in near singular scenarios.

5.2 Future Directions

In this thesis, we focused solely on equal group sizes within our simulations (group 1 has the same population as group 2). However, through analysis on real data sets, where the group sizes were different, it was evident that further exploration into the methods could be useful, in particular when there are large gaps between the groups. However, we do not expect to get different results. It will also be of interest to consider the EGCM, which will allow us to model clustered longitudinal data, where the mean trajectory across the different groups can have different shapes (different degrees of polynomials).

In practical applications, hypothesis testing involving comparisons between mean trajectories over time, and across groups is often of interest. Therefore, it will be important to investigate the level and power of the tests related to unweighted and weighted approaches. If the unweighted method is shown to perform similarly or better than the weighted approaches, at least in some scenarios, it will lead to significant mathematical simplifications in our investigation of the distributions of existing test statistics.

Another future direction is robust estimation in situations where the model is misfitted, which we saw in some of the real data applications. If the systematic component of the model (the assumed mean structure over time) is not accurately modeled, we expect $\mathbf{R}_3$ to be large, which means the $\mathbf{S}$-weighted and MLE-weighted estimators will perform differently. In fact, we hypothesize that the MLE-weighted estimators will lead to better performance, since the MLE of $\boldsymbol{\Sigma}$ is based on the full residual space (unlike the pooled-variance covariance estimator) and hence accounts for the error in model misfitting. We also expect the MLE-weighted estimator to perform better when there are significant differences between individuals in the same group or in the presence of outliers. We saw such examples in the gene expression data sets, where the two weighted methods performed differently. Further research is needed, both in large sample and high-dimensional scenarios,

to investigate the robustness of the unweighted as well as the two weighted methods in these scenarios.

Finally, our choice of distance matrices was arbitrary, and the need for more robust distance techniques for optimality matrices is needed. Although we used the Euclidean distance, we could have, for instance, considered the Mahalanobis distance, which accounts for the variance-covariance matrix between the elements of the matrix estimators. The $Matrix\,MSE$ for both vector and matrix estimators, along with the aggregation strategies we used, also require further scrutiny.

References

[1] Ahmad R., von Rosen D., and Singull M. A note on mean testing for high dimensional multivariate data under non-normality. *statistica neerlandica*, 67(1):81 – 99, 2013.

[2] Anderson T.W. *Introduction to Multivariate Statistical Analysis.* New York, John Wiley and Sons, Inc., 1958.

[3] Berry W.D. and Feldman S. *Multiple regression in practice.* SAGE Publishing, 1985.

[4] Casella G. and Berger R.L. *Statistical Inference Second Edition.* Thompson Learning, Duxbury, 2002.

[5] Chen J.T. and Gupta A.K. Matrix variate skew normal distributions. *Statistics (Berlin, DDR)*, 39(3):247 – 253, 2005.

[6] Everitt B.S. and Krzanowski W.J. Principles of multivariate analysis. *Technometrics*, 43:498 – 498, 2001.

[7] Hamid J.S. and Beyene J. A multivariate growth curve model for ranking genes in replicated time-course microarray data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):26 – 33, 2009.

[8] Hamid J.S. and von Rosen D. Residuals in the extended growth curve model. *Scandinavian Journal of Statistics*, 33(1):121 – 138, 2006.

[9] Hamid J.S., Beyene J., and von Rosen D. A novel trace test for the mean parameters in a multivariate growth curve model. *Journal of Multivariate Statistics*, 102(2):238 – 251, 2011.

[10] Hamid J.S. *Residuals in the extended growth curve model.* PhD thesis, U.U.D.M. Report 2001:29, ISSN 1101-3591, Department of Mathematics, Uppsala University, Uppsala, Sweden, 2001.

[11] Jana S., Balakrishnan N., and Hamid J.S. Bayesian growth curve model useful for high-dimensional longitudinal data. *Journal of Applied Statistics*, 46(5):814 – 834, 2019.

[12] Jana S., Balakrishnan N., and Hamid J.S. Inference in the growth curve model under multivariate skew normal distribution. *Sankhya: The Indian Journal of Statistics Series B*, 82(1):34 – 69, 2020.

[13] Jana S., Balakrishnan N., von Rosen D., and Hamid J.S. Estimation of the parameters of the extended growth curve model under multivariate skew normal distribution. *Journal of Multivariate Statistics*, 166:111 – 128, 2018.

[14] Jana S., Balakrishnan N., von Rosen D., and Hamid J.S. High dimensional extension of the growth curve model and its application in genetics. *Statistical Methods and Applications*, 26(2):273 – 292, 2017.

[15] Jana S. The growth curve model for high dimensional data and its application in genomics. Master's thesis, MacMaster University, 2013.

[16] Kaufman R. *Heteroskedasticity in Regression*. SAGE Publishing, 2013.

[17] Khatri C.G. A note on a manova model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics*, 18(1):75 – 86, 1966.

[18] Kollo T. and von Rosen D. Advanced multivariate statistics with matrices. *Springer Netherlands First Edition*, 2005.

[19] Lee D.S., Evans J.C., Robins S.J., and et al. Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk the framingham heart study. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 27(1):127 – 133, 2007.

[20] Mendoza-Blanco J.R., Tu X.M., and Gleser L.J. On the difference in inference and prediction between the joint and independent t-error models for seemingly unrelated regressions. *Technical Report, Department of Statistics, University of Pittsburg*, 28(9):2119 – 2140, 1995.

[21] Nzabanita J., von Rosen D., and Singull M. Estimation of parameters in the extended growth curve model with a linearly structured covariance matrix. *Acta et Commentationes Universitatis Tartuensis De Mathematica*, 2012.

[22] Ohlson M. and von Rosen D. Explicit estimators of parameters in the growth curve model with linearly structured covariance matrices. *Journal of Multivariate Analysis*, 101(5):1284 – 1295, 2010.

[23] Pan J.X. and Fang K.T. *Growth Curve Models and Statistical Diagnostics*. Springer New York, 2002.

[24] Potthoff R.G. and Roy S.N. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3–4):313 – 326, 1964.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[26] Rosner B. *Fundamentals of Biostatistics*. Thomson-Brooks/Cole, 2006.

[27] Srivastava M.S. and Carter E.M. *An Introduction to Applied Multivariate Statistics*. North Holland, New York, 1983.

[28] Srivastava M.S. and Khatri C.G. *An introduction to Multivariate Statistics*. Elseiver North Holland, Inc. New York. First Edition, 1979.

[29] Srivastava M.S. and von Rosen D. *Growth curve models. Statistical Textbooks and Models*. Marcel Dekker Inc., New York, 1999.

[30] Srivastava M.S. *Methods of Multivariate Statistics*. John Wiley and Sons, Inc. First Edition., 2002.

[31] Srivastava M.S. Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society*, 37(1):53 – 86, 2007.

[32] Stanek III E.J. and Koch G.G. The equivalence of parameter estimates from growth curve models and seemingly unrelated regression models. *Biometrics*, 39(2):149 – 152, 1985.

[33] Verbyla A.P. and Venables W.N. An extension of the growth curve model. *Biometrika*, 75(1):129 – 138, 1988.

[34] von Rosen D. Maximum likelihood estimators in multivariate linear normal models. *Journal of Multivariate Statistics*, 31(2):187 – 200, 1989.

[35] von Rosen D. Residuals in the growth curve model. *The Annals of the Institute of Statistical Mathematics*, 47:129–136, 1995.

[36] von Rosen D. The growth curve model: a review. *Communications in Statistics-Theory and Methods*, 20(9):2791 – 2822, 1991.

[37] Woolson R.F. and Leeper J.D. Growth curve analysis of complete and incomplete longitudinal data. *Communications in Statistics - Theory and Methods*, 9(14):1491 – 1513, 1980.

[38] Zellner A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348 – 368, 1962.

[39] Zellner A. Estimators for seemingly unrelated regression equations: Some exact finite sample results. *Journal of the American Statistical Association*, 58(304):977 – 992, 1963.

[40] Zerbe G. Randomization analysis of the completely randomized design extended to growth and response curves. *Journal of the American Statistical Association*, 74(365):215 – 221, 1979.

Appendix A

Bias and MSE

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
40	10	0.624	0.360	0.836	0.715
40	20	0.296	0.450	0.792	0.334
40	30	0.168	0.237	0.209	0.136
40	40	0.015	494.290	494.784	0.018
40	50	0.176	0.166	0.166	0.175
40	100	0.001	0.020	0.020	0.000
40	150	0.014	0.067	0.067	0.014
40	200	0.004	0.007	0.007	0.004
40	250	0.006	0.013	0.013	0.005
40	300	0.027	0.019	0.019	0.028
40	350	0.012	0.008	0.008	0.013
40	400	0.003	0.002	0.002	0.004
40	450	0.017	0.018	0.018	0.018
40	500	0.009	0.004	0.004	0.007

Table A.1: Comparison of Euclidean Distances of the Bias Matrix of the Various Weighting Methods using a Covariance Matrix with Strong Dependency

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
40	10	0.457	0.151	0.574	0.496
40	20	0.256	0.535	0.642	0.272
40	30	0.074	0.153	0.134	0.066
40	40	0.028	0.266	0.285	0.031
40	50	0.186	0.084	0.084	0.186
40	100	0.004	0.017	0.017	0.004
40	150	0.002	0.027	0.027	0.002
40	200	0.003	0.002	0.002	0.003
40	250	0.006	0.009	0.009	0.005
40	300	0.011	0.008	0.008	0.012
40	350	0.005	0.002	0.002	0.006
40	400	0.005	0.006	0.006	0.005
40	450	0.004	0.004	0.004	0.004
40	500	0.008	0.005	0.005	0.007

Table A.2: Comparison of Euclidean Distances of the Bias Matrix of the Various Weighting Methods using a Covariance Matrix with Weak Dependency

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
4	20	2.158	2.276	2.276	1.981
4	30	1.458	1.495	1.495	1.355
4	40	1.095	1.114	1.114	1.033
4	50	0.864	0.861	0.861	0.816
4	60	0.733	0.701	0.701	0.682
4	70	0.668	0.645	0.645	0.625
4	80	0.590	0.553	0.553	0.541
4	90	0.526	0.508	0.508	0.496
4	100	0.481	0.451	0.451	0.449
8	20	0.134	0.124	0.124	0.092
8	30	0.092	0.075	0.075	0.061
8	40	0.065	0.047	0.047	0.041
8	50	0.053	0.039	0.039	0.036
8	60	0.044	0.030	0.030	0.028
8	70	0.038	0.027	0.027	0.025
8	80	0.033	0.022	0.022	0.021
8	90	0.028	0.020	0.020	0.018
8	100	0.027	0.018	0.018	0.017

Table A.3: Comparison of Euclidean Distances of the MSE Matrix of the Various Weighting Methods using Covariance Matrices Generated from the Dental ($p = 4$) and Glucose data ($p = 8$).

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
40	10	0.2670	0.3106	0.3155	0.2646
40	20	0.1364	0.1746	0.1757	0.1353
40	30	0.0859	0.5844	0.5643	0.0850
40	40	0.0623	56.9031	119.0147	0.0622
40	50	0.0512	3.0855	3.0855	0.0510
40	100	0.0251	0.0402	0.0402	0.0249
40	150	0.0171	0.0224	0.0224	0.0170
40	200	0.0126	0.0155	0.0155	0.0126
40	250	0.0105	0.0122	0.0122	0.0105
40	300	0.0084	0.0100	0.0100	0.0084
40	350	0.0077	0.0084	0.0084	0.0077
40	400	0.0063	0.0071	0.0071	0.0063
40	450	0.0055	0.0063	0.0063	0.0055
40	500	0.0055	0.0055	0.0055	0.0055

Table A.4: Comparison of Euclidean Distances of the MSE Matrix of the Various Weighting Methods using a Covariance Matrix with Strong Dependency

p	n	Unweighted	$\mathbf{S}$-weighted	MLE-weighted	$\mathbf{\Sigma}$-weighted
40	10	0.1611	1.8456	1.8143	0.1602
40	20	0.0799	0.1433	0.1142	0.0797
40	30	0.0511	0.1339	0.1093	0.0508
40	40	0.0379	17.4596	67.2878	0.0379
40	50	0.0307	0.5375	0.5375	0.0307
40	100	0.0152	0.0241	0.0241	0.0152
40	150	0.0100	0.0130	0.0130	0.0100
40	200	0.0077	0.0095	0.0095	0.0077
40	250	0.0063	0.0071	0.0071	0.0063
40	300	0.0045	0.0055	0.0055	0.0045
40	350	0.0045	0.0045	0.0045	0.0045
40	400	0.0032	0.0045	0.0045	0.0032
40	450	0.0032	0.0032	0.0032	0.0032
40	500	0.0032	0.0032	0.0032	0.0032

Table A.5: Comparison of Euclidean Distances of the MSE Matrix of the Various Weighting Methods using a Covariance Matrix with Weak Dependency

Appendix B

Generating the Covariance Matrices

Below is the R code used to generate the covariance matrices, where weak, moderate and strong dependency was created by selecting appropriate elements of the Toeplitz matrix that was created.

```
library(ramify)
library(Matrix)
for (m in 1:nsim){
  S=toeplitz(320:1)/320
  set.seed(p*m)
  Sigma_gcm=matrix(rWishart(n=1,df=320,Sigma=S),nrow=320)/320
  h1=triu(Sigma_gcm[1:p,(p+1):(2*p)])
  h2=forceSymmetric(h1)
  h3=matrix(h2,byrow=TRUE,nrow=p)
  diag(h3)=diag(Sigma_gcm)[(p+1):(2*p)]
  h11=triu(Sigma_gcm[1:p,(p*3+1):(p*4)])
  h21=forceSymmetric(h11)
  h31=matrix(h21,byrow=TRUE,nrow=p)
  diag(h31)=diag(Sigma_gcm)[(p*3+1):(p*4)]
  h12=triu(Sigma_gcm[1:p,(p*5+1):(p*6)])
  h22=forceSymmetric(h12)
  h32=matrix(h22,byrow=TRUE,nrow=p)
  diag(h32)=diag(Sigma_gcm)[(p+1):(2*p)]
  h3_avg=h3_avg+h3
```

```
   h31_avg=h31_avg+h31
   h32_avg=h32_avg+h32
}
#h1: High Correlation
#h2: Moderate Correlation
#h3: Low Correlation

h3_new=h3_avg/nsim
h31_new=h31_avg/nsim
h32_new=h32_avg/nsim
```