Improving Small-Sample Inference in Group Randomized Trials and Other Sources of Correlated Binary Outcomes

by

Philip Michael Westgate

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Thomas M. Braun, Chair
Professor John D. Kalbfleisch
Professor Susan A. Murphy
Professor Peter X. Song

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor, Dr. Thomas Braun, for all of his help and advice throughout my doctoral work. I would like to thank Dr. Peter Song and Peisong Han for their very helpful advice with respect to Chapter IV of this dissertation. I would finally like to thank Dr. John Kalbfleisch, Dr. Susan Murphy, and Dr. Song for serving on my Dissertation Committee.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ABSTRACT

Improving Small-Sample Inference in Group Randomized Trials and Other Sources of Correlated Binary Outcomes

by

Philip Michael Westgate

Chair: Thomas M. Braun

Group Randomized Trials (GRTs), along with many other types of studies, commonly can be composed of a small to moderate number of independent clusters of correlated data. In this dissertation, we focus on statistical inference in these settings. Particularly, we concentrate on test size and estimation variability when a marginal model is employed.

Our first focus is in a general GRT setting in which a logistic regression only implements an indicator of treatment assignment. A Wald test, using a model-based standard error, for a marginal treatment effect can tend to have a realized test size smaller than its nominal value. We therefore propose a pseudo-Wald statistic that consistently produces test sizes at their nominal value, therefore increasing or maintaining power.

Our second focus is on the estimation performance of QIF as compared to GEE when the number of clusters is not large, with a focus on GRT settings. GEE is commonly used for the analysis of correlated data, while QIF is a newer method with the theoretical advantage of being equally or more efficient. Therefore, it would be reasonable to believe that QIF should maintain or increase power in GRTs, which typically have low power. We show, however, that QIF may not have this advantage in GRT settings, and estimates from QIF can have

greater variability than estimates from GEE due to the empirical impact of imbalance in cluster sizes and covariates, therefore concluding GEE is a more appropriate method in these settings.

We finally focus on improving the small-sample estimation performance of QIF. Specifically, we propose multiple alternative weighting matrices to use in QIF that combat its small-sample deficiencies. These weighting matrices are expected to perform better in small-sample settings, such as for GRTs, but maintain QIF's large-sample advantages. We compare the performances of the proposed QIF modifications via simulations, which show they can improve small-sample estimation. We also demonstrate that two of the proposed QIF versions work best.

CHAPTER I

Introduction

1.1 Background and Significance

This dissertation focuses on improving the analysis of correlated data, which occurs in a wide range of general scenarios such as in longitudinal studies and group randomized trials (GRTs). It is oriented toward the performance of statistical methodologies, in conjunction to fitting marginal models, commonly used in these types of studies when the sample size, or number of independent clusters, is small. Emphasis is given to GRT settings in which outcomes are binary in nature. However, much of our work can be generalized to common repeated measures scenarios with non-binary outcomes, especially methods presented in Chapters III and IV.

In these small-sample settings, some statistical methods may perform sub-optimally due to their reliance upon asymptotic theory. In this dissertation, we focus on improving test size and parameter estimation in these contexts. It is important to have these improvements, as they may have an effect on inference in costly research studies. Particularly, Chapter II focuses on improving test size in GRT settings in which outcomes are binary in nature. Chapter III discusses why a newer method, Quadratic Inference Functions (QIF) (Qu, Lindsay, and Li, 2000), which has multiple theoretical advantages over Generalized Estimating Equations (GEE) (Liang and Zeger, 1986), including estimation efficiency, can actually lead to estimates having greater variability than the corresponding estimates from GEE. Focus is given

to GRT settings in which a working exchangeable correlation structure is employed. Chapter IV proposes and studies methods that are meant to improve the small-sample estimation performance of QIF, with a general applied focus on GRTs and longitudinal study settings in which the true and working correlation structures were not restricted to be exchangeable. Chapter V summarizes the findings of this dissertation, discusses their importance, and outlines future work.

Specifically, Chapter II focuses on GRTs, which randomize groups of people to treatment or control arms instead of individually randomizing subjects. Typically, GRTs have a small number, n, of independent clusters, each of which can be quite large. When each subject has a binary outcome, over-dispersed binomial data may result, quantified as an intra-cluster correlation (ICC). Treating the ICC as a nuisance parameter, inference for a treatment effect can be done using quasi-likelihood (Wedderburn, 1974) with a logistic link. A Wald statistic, which, under standard regularity conditions, has an asymptotic standard normal distribution, can be used to test for a marginal treatment effect. However, we have found in our setting that the Wald statistic may have a variance less than one, resulting in a test size smaller than its nominal value. This problem is most apparent when marginal probabilities are close to zero or one, particularly when n is small and the ICC is not negligible. When the ICC is known, we develop a method for adjusting the estimated standard error appropriately such that the Wald statistic will approximately have a standard normal distribution. We also propose ways to handle non-nominal test sizes when the ICC is estimated. We demonstrate the utility of our methods through simulation results covering a variety of realistic settings for GRTs, and demonstrate them via the analysis of a dataset from an actual GRT.

Chapter III changes focus from test size to the variability of parameter estimates, while maintaining an applied interest in GRT scenarios. GEE, already cited, are commonly used for the analysis of correlated data. Qu *et al.* (2000) proposed the use of QIF as an alternative method to increase efficiency when the working covariance structure is misspecified. Although existing literature shows QIF has advantages over GEE, the impacts of covariates

and imbalanced cluster sizes on the estimation performance of QIF in finite samples have not been studied. This cluster size variation causes QIF's estimating equations and GEE to be in separate classes when an exchangeable correlation structure is implemented, causing QIF and GEE to be incomparable in terms of efficiency. When utilizing this structure and the number of clusters is not large, we discuss how covariates and cluster size imbalance can cause QIF, rather than GEE, to produce estimates with the larger variability. This occurrence is mainly due to the empirical nature of weighting QIF employs, rather than differences in estimating equations classes. We demonstrate QIF's lost estimation precision through simulation studies covering a variety of general GRT scenarios, and compare QIF and GEE in the analysis of data from a GRT.

Chapter IV again focuses on the estimation performance of QIF. Here, we focus on decreasing the variance in its parameter estimates, not only in GRT settings in which an exchangeable correlation structure is employed, but in general repeated measures scenarios in which the true and working structures can be different than exchangeable, such as AR-1. Particularly, we propose and compare six alternative weighting matrices for QIF, five of which asymptotically are optimally weighted combinations of the empirical covariance matrix and other matrices expected to potentially perform better when the number of independent clusters is small. These combinations are derived upon minimizing expected quadratic loss, maintain the large sample advantages QIF has over GEE, and as shown in simulations, can improve the small sample estimation performance of QIF. Additionally, two of the proposed QIF versions are shown to work best.

CHAPTER II

Improving Small-Sample Inference in Group Randomized Trials with Binary Outcomes

2.1 Introduction

Many clinical trials involve testing a new treatment or intervention versus a control, with each study participant randomly assigned to one of these study arms. However, group randomized trials (GRTs) are unique in that groups, or clusters, of people are randomized instead of each person individually, but the outcome of interest is still obtained from each subject. Due to feasibility issues, such as high costs, most likely only a relatively small number of clusters will be involved in a GRT. Additionally, cluster sizes can typically be quite large. To demonstrate, some common groups of randomization are patients with the same health care provider, communities, and schools.

An example of a GRT would be the study reported by Atri *et al.* (1997). This study aimed to discover if a two-hour training session for receptionists, who were supposed to later attempt to contact patients, would increase breast screening rates in women who failed to attend for an appointment by a certain time point. Twelve practices, and inherently the women their receptionists were to contact, were randomized to this receptionist intervention, while fourteen were randomized to be controls. The subject-level outcome of interest was an indicator of whether a given woman had received screening after failing to attend her initial appointment.

It is the purpose of this chapter to demonstrate methods for testing for a marginal

treatment effect with a Wald statistic that maintains a nominal test size when the outcomes of interest from a GRT are subject-level binary indicators of a desired outcome ("success"). These types of data often lead to over-dispersion because of unmeasured group effects that make the probability of success vary between clusters. For example, in the Atri *et al.* (1997) study, the receptionists' natural ability to get women to come in for breast screening varies, thus resulting in different success rates for each practice. Statistical methods need to take this over-dispersion into account to obtain reliable inference.

In Section 2.2 we introduce statistical notation and existing methodology. In Section 2.3, we present drawbacks with a traditional Wald statistic when using a model-based standard error (SE), which motivates the derivation of our psuedo-Wald tests. In Section 2.4, we examine the performance of our pseudo-Wald tests via simulation as well as in application to an actual GRT. Section 2.5 contains a discussion and concluding remarks.

2.2 Notation and Existing Methodology

2.2.1 Notation

Throughout this chapter, we adopt the following notation. Let X_{ij} represent the outcome for subject j in cluster i, $j = 1, 2, \ldots n_i$; $i = 1, 2, \ldots n$. $X_{ij} = 1$ denotes success and $X_{ij} = 0$ denotes the absence of the desired outcome ("failure"). We also let ρ denote the intra-cluster correlation (ICC) for any pair of outcomes from individuals within the same cluster. We assume the ICC is constant across clusters.

We let Z_i represent the indicator of treatment assignment for all individuals in cluster i, with $Z_i = 1$ indicating new treatment or intervention and $Z_i = 0$ indicating control. We let the first u clusters represent the controls, and the last $n - u$ represent groups receiving the new treatment or intervention.

We assume $\pi_i = Prob(X_{ij} = 1) = E[Prob(X_{ij} = 1 \mid p_i)]$ is constant for all clusters in the same treatment arm, where p_i is the unobserved true probability of success for any given

subject in the ith cluster. We assume a simple logistic regression model, $\text{logit}(\pi_i) = \beta_0 + \beta_1 Z_i$, where $\text{logit}(\pi_i) = log(\pi_i) - log(1 - \pi_i)$. If the ith cluster is randomized to control, then $\pi_i = \pi_C$; if randomized to the new treatment or intervention, then $\pi_i = \pi_T$. Although this model can be generalized to include cluster-level or individual-level predictors, in the current presentation we assume that such adjustments are unnecessary.

Let $Y_i = \sum_{j=1}^{n_i} X_{ij}$ represent the number of successes in cluster i, which has mean $E(Y_i) = n_i\pi_i$ and variance $Var(Y_i) = n_i\pi_i(1 - \pi_i)[1 + (n_i - 1)\rho]$. If $\rho = 0$, then Y_i has a binomial distribution. In most settings, ρ is assumed to be positive, although negative values are possible. See, for example, Prentice (1986). This positive correlation results in Y_i being overdispersed, i.e., having larger variance than what is predicted by the binomial distribution. The factor $[1 + (n_i - 1)\rho]$ is known as the variance inflation factor (VIF) for the ith cluster.

The ICC can be viewed as measuring the degree to which responses from subjects within the same cluster tend to respond "more alike" as compared with subjects from different clusters in the same arm; this effect is due to clusters potentially having different success rates, or $Var(p_i) > 0$ for $i = 1, 2, \ldots n$. Mathematically, for $1 \leq j \neq l \leq n_i$,

$$\begin{aligned} \rho &= \frac{Cov[E(X_{ij} \mid p_i), E(X_{il} \mid p_i)] + E[Cov(X_{ij}, X_{il} \mid p_i)]}{E[Var(X_{ij} \mid p_i)] + Var[E(X_{ij} \mid p_i)]} \\ &= \frac{Cov[p_i, p_i] + E(0)}{E(p_i[1 - p_i]) + Var(p_i)} \\ &= \frac{Var(p_i)}{\pi_i(1 - \pi_i)} \end{aligned}$$

Due to the constraint ICC ≤ 1, we have $Var(p_i) \leq \pi_i(1 - \pi_i)$.

Although we only work with the ICC in this chapter, it is important to note that another popular method is using pairwise odds ratios for modeling the association among subject-level outcomes within the same cluster. See Carey, Zeger, and Diggle (1993) for more detail.

2.2.2 Quasi-Likelihood

Wedderburn (1974) developed the theory of quasi-likelihood (QL), which is used when a generalized linear model (GLM) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) is desired but the true distribution for the observed data is unknown. This is important since the true distribution for over-dispersed binomial data will be unknown in practice. Using QL, only the mean and variance structures for the proportion of successes in each cluster need correct specification. More specifically, the link function with its linear predictor, $h(\mu_i) = \eta_i = \mathbf{z}_i'\boldsymbol{\beta}$, and $Var(Y_i)$ need correct specification. Here, $\mathbf{z}_i' = [1, z_{i,1},..., z_{i,p-1}]$ is a p $\times$ 1 vector of covariate values for the ith independent observation and $\boldsymbol{\beta} = [\beta_0,..., \beta_{p-1}]'$ is a p $\times$ 1 vector of corresponding regression parameters. Maximum quasi-likelihood (MQL) involves setting the following quasi-score equations equal to zero and solving for $\boldsymbol{\beta}$, where $\mathbf{D}_i = \partial\mu_i/\partial\boldsymbol{\beta} = [\partial\mu_i/\partial\beta_0,..., \partial\mu_i/\partial\beta_{p-1}]'$ and $V_i = Var(Y_i)$:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i V_i^{-1}(Y_i - \mu_i).$$

In our setting, the quasi-score equations simplify to

$$\sum_{i=1}^{n} [1, Z_i]' \frac{Y_i - n_i\pi_i}{1 + (n_i - 1)\rho}.$$

As long as a consistent estimate for the true ICC is used, the MQL estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{MQL} = [\hat{\beta}_{0MQL},..., \hat{\beta}_{(p-1)MQL}]'$, converges in probability to $\boldsymbol{\beta}$ and has an asymptotic normal distribution with covariance matrix $(\sum_{i=1}^{n} \mathbf{D}_i V_i^{-1} \mathbf{D}_i')^{-1}$, simplifying to

$$\left(\sum_{i=1}^{n} [1, Z_i]'[1, Z_i] \frac{n_i\pi_i(1 - \pi_i)}{1 + (n_i - 1)\rho} \right)^{-1} \tag{2.1}$$

in our scenario.

Our main concern was that MQL, which is popular for analyzing over-dispersed binomial data, gives statistical results relying upon asymptotic theory. Unfortunately, since GRTs

generally do not have a large number of independent clusters, asymptotic theory may not hold. We discuss three potential problems resulting from this, all of which may cause test size to not be at its nominal value when using a Wald statistic. First, when the data analyst is unsure of the correct variance or ICC structure, the sandwich, or empirical, covariance matrix estimator can be used and is given by

$$\left(\sum_{i=1}^{n} \hat{\boldsymbol{D}}_i \hat{V}_i^{-1} \hat{\boldsymbol{D}}_i'\right)^{-1} \left(\sum_{i=1}^{n} \hat{\boldsymbol{D}}_i \hat{V}_i^{-1} (Y_i - n_i \hat{\pi}_i)^2 \hat{V}_i^{-1} \hat{\boldsymbol{D}}_i'\right) \left(\sum_{i=1}^{n} \hat{\boldsymbol{D}}_i \hat{V}_i^{-1} \hat{\boldsymbol{D}}_i'\right)^{-1},$$

with consistent estimates used in place of unknown parameters (Liang and Zeger, 1986). This gives a consistent estimate for the covariance of the MQL parameter estimates. Unfortunately, the variance of this covariance estimate, along with its tendency to underestimate the true standard errors (SEs) for small n, can cause test size to be too large (Kauermann and Carroll, 2001; Mancl and DeRouen, 2001). Although there is no formal definition for what small n is, Mancl and DeRouen (2001) and Murray, Varnell, and Blitstein (2004) suggest $n < 50$ and $n < 40$, respectively. Bootstrap and jackknife methods are alternatives to the sandwich estimator, but can also be problematic, particularly when the number of successes in each cluster are zero or small (Mancl and DeRouen, 2001). Second, MQL estimates of the regression parameters tend to be biased away from zero when using the logistic link and n is small, with bias increasing as the true parameter values move further from zero (presented later). Third, since normality of the Wald statistic is an asymptotic result, combined with the need to estimate the ICC, the correct distribution from which to obtain critical values is unknown.

Methods to help reduce test size toward the nominal value when using the sandwich SE estimator have been introduced by Kauermann and Carroll (2001), Mancl and DeRouen (2001), Pan (2001), Fay and Graubard (2001), Pan and Wall (2002), Morel, Bokossa, and Neerchal (2003), and McCaffrey and Bell (2006). Drum and McCullagh (1993) argued that using the model-based SE instead of the sandwich estimate is best when n is small and there

is no reason to believe the assumed variance structure is "substantially incorrect." Liang and Hanfelt (1994) also expected the model-based SE to be more stable.

2.2.3 Bias Correction for the MQL Parameter Estimates

Similar to the formulas for the bias in maximum likelihood estimates (MLEs) from GLMs given by Cordeiro and McCullagh (1991), Cordeiro and Demetrio (2008) gave formulas for the bias to order n^{-1} for the MQL estimates. These formulas are given by

$$\mathbf{Bias}(\hat{\boldsymbol{\beta}}_{MQL}) = -0.5(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{M}_d\boldsymbol{F}\mathbf{1}.$$

Here, $\mathbf{W} = diag(n_i\pi_i(1-\pi_i)/[1+(n_i-1)\rho])$, $\mathbf{Z}' = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$, $\mathbf{M} = \mathbf{Z}(\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'$, $\mathbf{M}_d = \text{diag}[\text{M}_{ii}]$ where M_{ii} is the i_{th} diagonal element of $\mathbf{M}$,

$$\mathbf{F} = diag\left[\frac{n_i\pi_i(1-\pi_i)[2(1-\pi_i)-1]}{1+(n_i-1)\rho}\right],$$

and $\mathbf{1}$ is an $n \times 1$ vector of ones. We now write π_C and π_T as $\pi_C(\boldsymbol{\beta})$ and $\pi_T(\boldsymbol{\beta})$, respectively, since marginal probabilities are functions of $\boldsymbol{\beta}$. In our settings,

$$Bias(\hat{\beta}_{0MQL}) = \frac{2\pi_C(\boldsymbol{\beta})-1}{2\pi_C(\boldsymbol{\beta})[1-\pi_C(\boldsymbol{\beta})]\sum_{i=1}^{u} q_i} \tag{2.2}$$

and

$$Bias(\hat{\beta}_{1MQL}) = \frac{2\pi_T(\boldsymbol{\beta})-1}{2\pi_T(\boldsymbol{\beta})[1-\pi_T(\boldsymbol{\beta})]\sum_{i=u+1}^{n} q_i} - Bias(\hat{\beta}_{0MQL}), \tag{2.3}$$

where $q_i = n_i/[1+(n_i-1)\rho]$. Bias increases with decreases in n and cluster sizes and increases in ρ, $|\pi_C - 0.5|$, and $|\pi_T - 0.5|$.

The bias-corrected estimates (BCEs) are given by $\hat{\boldsymbol{\beta}}_{BC} = [\hat{\beta}_{0MQL} - \widehat{Bias}(\hat{\beta}_{0MQL}), \hat{\beta}_{1MQL} - \widehat{Bias}(\hat{\beta}_{1MQL})]' = [\hat{\beta}_{0BC}, \hat{\beta}_{1BC}]'$, where $\pi_C(\hat{\boldsymbol{\beta}}_{MQL})$ and $\pi_T(\hat{\boldsymbol{\beta}}_{MQL})$ are used in place of $\pi_C(\boldsymbol{\beta})$ and $\pi_T(\boldsymbol{\beta})$, respectively, in Equations (2.2) and (2.3). Due to using estimated marginal

probabilities, biases can be slightly overestimated as found in Bull and Greenwood (1997) and our simulations (not shown), especially when marginal probabilities are near the boundary of the parameter space and the number of clusters is small. The following iterative procedure was therefore used, which produces better bias approximations:

(1) Estimate bias as just mentioned to obtain $\widehat{Bias}^{(1)}(\hat{\beta}_{0MQL})$ and $\widehat{Bias}^{(1)}(\hat{\beta}_{1MQL})$. From these, denote the current BCEs as $\hat{\boldsymbol{\beta}}^{(1)}_{BC}$.

(2) Next, use $\pi_C(\hat{\boldsymbol{\beta}}^{(1)}_{BC})$ and $\pi_T(\hat{\boldsymbol{\beta}}^{(1)}_{BC})$ in Equations (2.2) and (2.3) to obtain updated bias estimates, $\widehat{Bias}^{(2)}(\hat{\beta}_{0MQL})$ and $\widehat{Bias}^{(2)}(\hat{\beta}_{1MQL})$. Use these to update the BCEs: $\hat{\boldsymbol{\beta}}^{(2)}_{BC} = [\hat{\beta}_{0MQL} - \widehat{Bias}^{(2)}(\hat{\beta}_{0MQL}), \hat{\beta}_{1MQL} - \widehat{Bias}^{(2)}(\hat{\beta}_{1MQL})]'$.

(3) Keep repeating until $|\hat{\beta}^{(s)}_{0BC} - \hat{\beta}^{(s-1)}_{0BC}| + |\hat{\beta}^{(s)}_{1BC} - \hat{\beta}^{(s-1)}_{1BC}| < \epsilon$, for some ϵ close to zero and $s \geq 1$. We used $\epsilon = 10^{-7}$.

Equation (2.2) shows $0 \leq |\beta_0| \leq |E(\hat{\beta}_{0MQL})|$, with β_0 and $E(\hat{\beta}_{0MQL})$ having the same sign, so $\hat{\beta}_{0MQL}$ is positively biased when $\beta_0 > 0$ and negatively biased when $\beta_0 < 0$. This implies that the BCE for β_0 will take on a value between zero and $\hat{\beta}_{0MQL}$, giving $0 \leq \hat{\beta}_{0BC}/\hat{\beta}_{0MQL} \leq 1$. Equation (2.3) shows the same relationship typically occurs for $\hat{\beta}_{1MQL}$ and $\hat{\beta}_{1BC}$. When π_C and π_T are close in value, though, there is a chance that this will not occur if $|\sum_{i=1}^{u} q_i - \sum_{i=u+1}^{n} q_i|$ is not "small". Although it makes little difference, we chose to set $\hat{\beta}_{1BC} = \hat{\beta}_{1MQL}$ if $\hat{\beta}_{1BC}/\hat{\beta}_{1MQL}$ was originally greater than one.

We note that bias corrections for maximum likelihood estimates, related to the work of Cordeiro and McCullagh (1991), have been discussed by Cox and Snell (1968) and Firth (1993). Cox and Snell (1968) gave general formulas for order n^{-1} biases of multiparameter MLEs, while Cordeiro and McCullagh (1991) extended this idea by giving the n^{-1} biases of MLEs in GLMs. Firth (1993), however, took a different approach to eliminate these n^{-1} biases by using a bias term inside the score equations.

2.3 Developing a Pseudo-Wald Test with Nominal Size

2.3.1 Quantifying the Impact of Model-Based SEs on Traditional Wald Test Size

From Equation (2.1), the model-based SE for $\hat{\beta}_{1MQL}$ is

$$SE(\hat{\beta}_{1MQL}) = SE_{\hat{\beta}_{1MQL}}[\pi_C(\boldsymbol{\beta}), \pi_T(\boldsymbol{\beta})] =$$

$$\sqrt{\left[\sum_{i=1}^{u} \frac{n_i \pi_C(\boldsymbol{\beta})[1-\pi_C(\boldsymbol{\beta})]}{1+(n_i-1)\rho}\right]^{-1} + \left[\sum_{i=u+1}^{n} \frac{n_i \pi_T(\boldsymbol{\beta})[1-\pi_T(\boldsymbol{\beta})]}{1+(n_i-1)\rho}\right]^{-1}} \tag{2.4}$$

This is estimated by $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL}) = SE_{\hat{\beta}_{1MQL}}[\pi_C(\hat{\boldsymbol{\beta}}_{MQL}), \pi_T(\hat{\boldsymbol{\beta}}_{MQL})]$. Simulations (not shown) indicated that when using an unbiased estimate, $\hat{\rho}$, for the ICC, $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$ is also fairly unbiased. For now, assume ρ is known.

The Wald statistic $W_{Reg} = \hat{\beta}_{1MQL}/\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$ is regularly used in practice to test for a marginal treatment effect. Using $\pi_C = \pi_T = \rho = 0.05$, Figure 2.1*a* shows the empirical distribution for W_{Reg} from 100,000 simulations, along with the density of the $N(0,1)$ distribution. Cluster sizes varied uniformly from 25 to 150 subjects, ten clusters were randomized to each treatment arm, and outcomes were generated by the beta-binomial distribution. Figure 2.1*a* reveals that the $N(0,1)$ distribution has heavier tails since the variance of W_{Reg}, $Var(W_{Reg})$, is less than one (0.905 in these simulations). This example implies that using W_{Reg} in conjunction with critical values from the $N(0,1)$ distribution can result in a test size that is smaller than desired, as will using any heavy-tailed distribution, such as a t-distribution.

Equation (2.1) shows that the variance of $\hat{\boldsymbol{\beta}}_{MQL}$ increases with decreases in n and cluster sizes and increases in ρ, $|\pi_C - 0.5|$, and $|\pi_T - 0.5|$. In practice, since $\hat{\boldsymbol{\beta}}_{MQL}$ is actually used to estimate its own variability, $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$ is not a fixed quantity. Its variance, $Var[\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})]$, will increase as $Var(\hat{\boldsymbol{\beta}}_{MQL})$ increases, and therefore is also a function

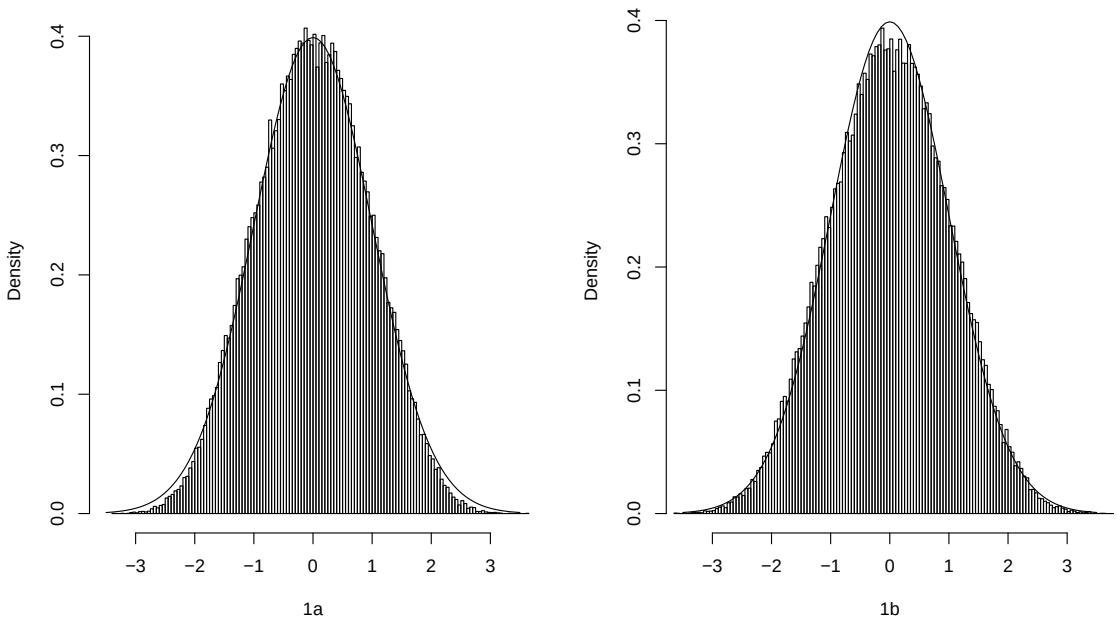

Figure 2.1: $N(0,1)$ density and empirical distributions for W_{Reg} $(1a)$ and $\widetilde{W}_{1.5}$ $(1b)$, where ten clusters were randomized to each treatment arm, and marginal probabilities and the ICC were 0.05.

of n, cluster sizes, ρ, π_C and π_T..

Due to the variability in $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$, $Var(W_{Reg})$ will depend on the variances and covariance of $\hat{\beta}_{1MQL}$ and $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$. As $Var(\hat{\beta}_{1MQL})$ increases, there are more extreme values for $\hat{\beta}_{1MQL}$, and these large values are associated with values for $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$ that are larger than the true SE. Hauck and Donner (1977) demonstrated this tendency in logistic regression. This relationship can cause W_{Reg} to be smaller than desired, therefore reducing $Var(W_{Reg})$ and making the tails in the distribution of W_{Reg} to become lighter, thus diminishing test size. Test size decreases as $Var[\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})]$ increases, implying test size decreases away from its nominal level with decreases in n and cluster sizes and increases in ρ, $|\pi_C - 0.5|$, and $|\pi_T - 0.5|$.

This phenomenon of test size being smaller than its nominal level is rather minor when marginal probabilities are not near the edge of the parameter space. Due to the curvature of Equation (2.4) with respect to the marginal probabilities, the impact from variation in the estimate of $\boldsymbol{\beta}$ used in this formula leads to increasingly larger variations in the estimated SE as $|\pi_C - 0.5|$ and $|\pi_T - 0.5|$ approach 0.5. Empirical evidence (not shown) indicates that having a large number of clusters, say thirty or more per treatment arm, will typically combat this problem quite well. For a small to moderate number of clusters, the decrease in test size may become important as $|\pi_C - 0.5|$ and $|\pi_T - 0.5|$ rise to 0.3 or higher, especially when ρ is almost as large, if not larger, than π_C and π_T. Additionally, some GRTs may not have large cluster sizes. An example of this would be a study where each cluster is actually an individual subject contributing a small number, or group, of binary outcomes. Having smaller sizes for a fixed number of clusters will increase the variation in the estimate of $\boldsymbol{\beta}$, causing a greater impact on test size. This impact may be negligible, however, unless the differences in cluster sizes are large. For instance, a study in which only a small number of observations on each subject are observed can have larger test size problems than a GRT where there are a large number of outcomes in each cluster.

2.3.2 Deriving a Pseudo-Wald Statistic with Known ICC

In practice, the ICC will need to be estimated, but the goal of this section is to show how the size of the Wald test can be adjusted closer to a nominal level, α, when ρ is known. As mentioned previously, $Var(W_{Reg})$ can be less than one, resulting in test sizes smaller than α when using $N(0,1)$ critical values. Two possible ways of fixing this would be to find critical values that will consistently produce a test size equal to α, or modify $\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$ such that the resulting SE estimate is smaller by an amount depending on n, cluster sizes, ρ, π_C and π_T, with the goal of producing a Wald statistic with a variance of one. We take the latter approach, but utilize the idea of changing the critical values when we later incorporate estimation of ρ.

Test size, $Bias(\hat{\beta}_{0MQL})$, and $Bias(\hat{\beta}_{1MQL})$ are functions of n, cluster sizes, ρ, π_C and π_T. Incorporating these relationships, we define

$$\tilde{\beta}_k^N = (\hat{\beta}_{kBC}/\hat{\beta}_{kMQL})^N \hat{\beta}_{kMQL} = (\hat{\beta}_{kBC})^N/(\hat{\beta}_{kMQL})^{N-1}, \quad k = 0, 1,$$

for any non-negative real number N. Our proposed pseudo-SE estimate is $\widetilde{SE}_N(\hat{\beta}_{1MQL}) = SE_{\hat{\beta}_{1MQL}}[\pi_C(\tilde{\boldsymbol{\beta}}^N), \pi_T(\tilde{\boldsymbol{\beta}}^N)]$, in which $\tilde{\boldsymbol{\beta}}^N = [\tilde{\beta}_0^N, \tilde{\beta}_1^N]'$. Since N is non-negative, $(\hat{\beta}_{kBC}/\hat{\beta}_{kMQL}) \in [0,1]$ implies $|\tilde{\beta}_k^N| \leq |\hat{\beta}_{kMQL}|$, which in practice will give $\widetilde{SE}_N(\hat{\beta}_{1MQL}) \leq \widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$ since $\pi_C(\tilde{\boldsymbol{\beta}}^N)$ and $\pi_T(\tilde{\boldsymbol{\beta}}^N)$ will be no further in value from 0.5 than $\pi_C(\hat{\boldsymbol{\beta}}_{MQL})$ and $\pi_T(\hat{\boldsymbol{\beta}}_{MQL})$, respectively, in realistic settings. Typically, $\widetilde{SE}_N(\hat{\beta}_{1MQL}) < \widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$. The pseudo-Wald statistic $\widetilde{W}_N = \hat{\beta}_{1MQL}/\widetilde{SE}_N(\hat{\beta}_{1MQL})$ will then be larger in absolute value than W_{Reg}. Also, decreases in n and cluster sizes and increases in ρ, $|\pi_C - 0.5|$, and $|\pi_T - 0.5|$ correspond to increases in test size when using $\widetilde{W}_N$ as the test statistic as compared to using W_{Reg}.

Using $\widetilde{W}_N$ as our test statistic, we needed to find a value for N such that the variance of $\widetilde{W}_N$ is always approximately one, with a resulting test size equal to α when using $N(0,1)$ critical values. Increasing N causes $\widetilde{SE}_N(\hat{\beta}_{1MQL})$ to decrease, and therefore test size will increase. Simulations were conducted to find an appropriate solution. Each setting was

Table 2.1: Empirical test sizes using the given Wald statistic and $N(0,1)$ critical values. *n/2* clusters were randomized to each treatment arm, and had a marginal success rate π. The ICC is known. Bold values have corresponding 95% confidence intervals covering 0.05, the nominal level. Outcomes from the first ten sets of simulations were generated using the beta-binomial distribution, while outcomes from the last ten sets of simulations were generated using a log-gamma mixture distribution, such that $-ln(p_i) \sim Gamma(\theta_i, \phi_i)$.

n/2	π	ICC	W_{Reg}	$\widetilde{W}_1$	$\widetilde{W}_{1.25}$	$\widetilde{W}_{1.5}$	$\widetilde{W}_{1.75}$	$\widetilde{W}_2$
10	0.05	0.05	0.0360	**0.0459**	**0.0482**	**0.0509**	**0.0540**	0.0557
20	0.05	0.05	0.0407	0.0440	0.0453	**0.0464**	**0.0472**	**0.0485**
10	0.10	0.05	0.0436	**0.0461**	**0.0471**	**0.0478**	**0.0482**	**0.0491**
20	0.10	0.05	**0.0470**	**0.0490**	**0.0494**	**0.0499**	**0.0505**	**0.0509**
10	0.10	0.10	0.0390	**0.0462**	**0.0483**	**0.0502**	**0.0521**	**0.0534**
20	0.10	0.10	0.0429	**0.0458**	**0.0466**	**0.0471**	**0.0487**	**0.0490**
10	0.20	0.05	**0.0481**	**0.0492**	**0.0495**	**0.0495**	**0.0496**	**0.0500**
20	0.20	0.10	**0.0496**	**0.0502**	**0.0502**	**0.0503**	**0.0504**	**0.0505**
10	0.30	0.05	**0.0518**	**0.0523**	**0.0524**	**0.0525**	**0.0527**	**0.0528**
20	0.30	0.10	**0.0514**	**0.0518**	**0.0520**	**0.0520**	**0.0520**	**0.0520**
10	0.05	0.05	0.0350	**0.0464**	**0.0491**	**0.0517**	0.0543	0.0565
20	0.05	0.05	0.0422	**0.0464**	**0.0471**	**0.0479**	**0.0492**	**0.0502**
10	0.10	0.05	0.0429	0.0452	0.0456	**0.0470**	**0.0483**	**0.0492**
20	0.10	0.05	**0.0477**	**0.0493**	**0.0496**	**0.0502**	**0.0507**	**0.0513**
10	0.10	0.10	0.0375	0.0450	**0.0476**	**0.0496**	**0.0514**	**0.0536**
20	0.10	0.10	0.0440	**0.0460**	**0.0467**	**0.0472**	**0.0483**	**0.0492**
10	0.20	0.05	**0.0479**	**0.0485**	**0.0488**	**0.0492**	**0.0496**	**0.0500**
20	0.20	0.05	**0.0483**	**0.0488**	**0.0488**	**0.0488**	**0.0488**	**0.0488**
10	0.20	0.10	**0.0459**	**0.0474**	**0.0480**	**0.0483**	**0.0490**	**0.0497**
20	0.20	0.10	**0.0492**	**0.0502**	**0.0505**	**0.0507**	**0.0511**	**0.0512**

examined in 10,000 simulations, and cluster sizes varied uniformly from 25 to 150 subjects. Empirical test size was compared for W_{Reg} and $\widetilde{W}_N$ with $N \in \{1, 1.25, 1.5, 1.75, 2\}$. Results from using a five percent significance level can be seen in Table 2.1 and show that N=1.5 performed best. However, allowing N to be any value from 1.25 to 2 would be adequate. Figure 2.1*b* shows the empirical distribution for $\widetilde{W}_{1.5}$ from the same set of simulations used to produce Figure 2.1*a*, along with the density of the $N(0,1)$ distribution. The tails from this empirical distribution match the tails from the $N(0,1)$ density, indicating that test size is at its nominal level.

Although we are dealing with scenarios involving small n, it is important to show how

$\widetilde{W}_N$ performs asymptotically. Given $N < \infty$ and $(n-u)/u$ is a constant (typically around one), both $\tilde{\boldsymbol{\beta}}^N \xrightarrow{p} \boldsymbol{\beta}$ and $\widetilde{W}_N \xRightarrow{d} N(0,1)$ as $n \to \infty$. Therefore, $\widetilde{W}_N$ and W_{Reg} should give very similar results for large n, and both will approximately behave as standard normal random variables.

2.3.3 Incorporating Estimation of ICC

The following two issues still need to be handled in practice: (1) finding a consistent estimate, $\hat{\rho}$, for the ICC, and (2) finding appropriate critical values and/or adjust Equation (2.4) to deal with the effect estimating ρ has on the distribution of the Wald statistic that is utilized. With regard to the first issue, there are many papers that have dealt with the topic of estimating the ICC, with Ridout, Demetrio, and Firth (1999) presenting an overview with simulations comparing many different estimation procedures. Some of the more well-known ways to estimate the ICC are Williams' method (Williams, 1982), Pseudolikelihood (Carroll and Ruppert, 1982; Davidian and Carroll, 1987), Extended Quasi-Likelihood (Nelder and Pregibon, 1987), and ANOVA (Donner and Donald, 1988; Reed, 2000; Jung, Kang, and Ahn, 2001). These are useful in that they can handle regression models with multiple covariates, both categorical and continuous, at the cluster level. Preliminary simulations (not shown) were done to find the most appropriate method in our settings in terms of mean squared error (MSE). Of the previously mentioned methods, ANOVA performed best. If the assumption of a common ICC is correct, a consistent estimate is given by

$$\hat{\rho}_{ANOVA} = \frac{MSB - MSE}{MSB + (K-1)MSE}$$

where

$$MSB = \frac{1}{n-2}\sum_{i=1}^{n} n_i \left(\frac{Y_i}{n_i} - \hat{\pi}_i\right)^2, \; MSE = \frac{1}{\sum_{i=1}^{n} n_i - n}\sum_{i=1}^{n} Y_i\left(1 - \frac{Y_i}{n_i}\right),$$

$$K = \frac{1}{n-2}\left[\sum_{i=1}^{n} n_i - \sum_{i=1}^{u}\frac{n_i^2}{m_C} - \sum_{i=u+1}^{n}\frac{n_i^2}{m_T}\right],\ m_C = \sum_{i=1}^{u} n_i,\ m_T = \sum_{i=u+1}^{n} n_i$$

With regard to the second issue, estimating the ICC causes $Var(W_{Reg})$ and $Var(\widetilde{W}_N)$ to increase, leading to an inflation of test size. One method to deal with increased test size would be to use critical values that are larger in absolute value; however, these values would need to depend on the bias and variance of $\hat{\rho}$, and therefore n, since increasing n will decrease these quantities, causing test size to reduce back toward α. One possible solution would be to obtain critical values from a t-distribution with $f(n)$ degrees of freedom (df), denoted as $t_{f(n)}$, assuming f$(\cdot)$ is a "correctly" chosen function. We later demonstrate the utility of $f(n) = n$.

Another way to shrink test size back toward α after estimating the ICC would be to use leverage values to inflate the estimated versions of Equation (2.4) for W_{Reg} and $\widetilde{W}_N$ while continuing to use $N(0,1)$ critical values. Leverage values, $0 \leq h_i \leq 1, i = 1, 2, \ldots n$, are the diagonal elements of the $n \times n$ matrix

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}^{1/2},$$

such that $\sum_{i=1}^{n} h_i = p$, where p is the number of regression parameters ($p = 2$ in our setting). Note that leverage values are estimated using $\pi_C(\hat{\boldsymbol{\beta}}_{MQL})$ and $\pi_T(\hat{\boldsymbol{\beta}}_{MQL})$. As $n \to \infty$, we have $h_i \to 0, i = 1, \ldots n$. We multiply the ith term in Equation (2.4) by $(1 - h_i)$, giving

$$\sqrt{\left[\sum_{i=1}^{u}\frac{n_i\pi_C(\boldsymbol{\beta})[1-\pi_C(\boldsymbol{\beta})](1-h_i)}{1+(n_i-1)\rho}\right]^{-1} + \left[\sum_{i=u+1}^{n}\frac{n_i\pi_T(\boldsymbol{\beta})[1-\pi_T(\boldsymbol{\beta})](1-h_i)}{1+(n_i-1)\rho}\right]^{-1}} \tag{2.5}$$

Using this function of the leverages inside Equation (2.4) will cause the denominator of the Wald statistic to increase as n decreases, offsetting at least part of the elevation in the variance of the Wald statistic due to the increase in variance and bias of $\hat{\rho}$.

2.4 Assessing the Utility of the Pseudo-Wald Statistic

2.4.1 Via Simulation Study

We compare the test size of W_{Reg} and $\widetilde{W}_{1.5}$ to those of two other Wald statistics. The first, $W_S = \hat{\beta}_{1MQL}/\widehat{SE}_S(\hat{\beta}_{1MQL})$ is the traditional Wald statistic using the sandwich form for the SE of $\hat{\beta}_{1MQL}$. The second, $W_{SBC} = \hat{\beta}_{1MQL}/\widehat{SE}_{SBC}(\hat{\beta}_{1MQL})$ replaces the sandwich estimate for $SE(\hat{\beta}_{1MQL})$ with the bias-corrected version proposed by Mancl and DeRouen (2001). All four statistics are compared to both $N(0,1)$ and t_n critical values. We also compute versions of W_{Reg} and $\widetilde{W}_{1.5}$ implementing Equation (2.5) which are compared to $N(0,1)$ critical values. As mentioned previously, ANOVA is used to estimate a common ICC. Therefore, our results apply to using any estimator of ρ with bias and variance similar to that of $\hat{\rho}_{ANOVA}$. Empirical test sizes under five settings using a significance level of 0.05 are displayed in Table 2.2.

Results show that using W_{Reg} leads to inconsistent test sizes. When compared to $N(0,1)$ critical values, test size is too large in scenarios with marginal probabilities of at least 0.20, unless we use the inflated SE. When using the inflated SE or comparing to t_n critical values, the null hypothesis is not rejected enough if marginal probabilities are 0.10 or less. W_S is even less desirable to use. It gives inconsistent test sizes, all tending to be too large. W_{SBC} fairs better, but also leads to inconsistent inference. W_{SBC} is more reliable when compared to t_n critical values, but test size is too large when group probabilities are 0.05 and too small when group probabilities range from 0.2 to 0.5. Comparing $\widetilde{W}_{1.5}$ and its inflated SE version to t_n and $N(0,1)$ critical values, respectively, test size is consistently at its nominal level, and so there is no need to be concerned whether inference will be liberal or conservative. This gives a valid test, which will not reject the null hypothesis too often as the previously mentioned tests do in some settings. Additionally, these two proposed pseudo-Wald tests will produce greater power over all scenarios where other tests are conservative, such as when comparing W_{SBC} to t_n critical values and the marginal probabilities, although not necessarily

Table 2.2: Empirical test sizes using the given Wald statistic and distribution for obtaining critical values when estimating a common ICC. Bold values have corresponding 95% confidence intervals covering 0.05, the nominal level. Each setting used 10,000 replications, with cluster sizes generated as in previous simulations, while data came from a beta-binomial distribution with a common ICC. Eqn. 5 denotes the use of Equation (2.5) and $N(0,1)$ critical values.

			W_{Reg}			$\widetilde{W}_{1.5}$			W_S		W_{SBC}	
$n/2$	π	ICC	$N(0,1)$	t_n	Eqn. 5	$N(0,1)$	t_n	Eqn. 5	$N(0,1)$	t_n	$N(0,1)$	t_n
10	0.05	0.05	**0.0506**	0.0376	0.0387	0.0662	**0.0478**	**0.0503**	0.1029	0.0836	0.0715	0.0548
20	0.10	0.10	**0.0509**	0.0443	0.0453	0.0557	**0.0486**	**0.0494**	0.0713	0.0635	0.0586	**0.0509**
10	0.20	0.10	0.0614	**0.0462**	**0.0484**	0.0645	**0.0491**	**0.0518**	0.0836	0.0693	0.0589	0.0452
10	0.30	0.05	0.0655	**0.0514**	**0.0537**	0.0660	**0.0520**	**0.0537**	0.0806	0.0662	0.0550	0.0418
20	0.50	0.10	0.0567	**0.0499**	**0.0511**	0.0567	**0.0499**	**0.0511**	0.0635	0.0556	**0.0501**	0.0437

equal, range from 0.20 to 0.50.

2.4.2 Via Application to Actual GRT

We illustrate an application of our pseudo-Wald tests using the study reported by Atri *et al.* (1997), the data for which are presented in Turner, Omar, and Thompson (2001). The number of women failing to attend appointments in a given practice ranged from 19 to 201. The ICC was estimated to be 0.064, indicating a small variation in estimated success rates between practices in the same treatment arm. Using the MQL regression parameters, the marginal probabilities of breast screening for intervention and control practices were estimated to be 0.101 and 0.035, respectively. Parameter and SE estimates, along with p-values and 95% confidence intervals, for the various methods presented in this chapter are given in Table 2.3. The model-based SE gave the largest SE estimate, while the sandwich SE estimate was the smallest. The use of any combination of SE estimate and distribution to obtain critical values from resulted in the rejection of the null hypothesis, implying there is strong enough evidence at the five percent significance level to conclude that the intervention was effective. The pseudo-Wald statistic appeared to give slightly stronger evidence supporting a treatment effect as compared with the use of W_{Reg}. The use of the sandwich SEs gave the strongest support for a treatment effect, although these SE estimates may be biased downward.

2.5 Concluding Remarks

Many GRTs randomize a relatively small number of clusters. When the data to be analyzed from this setting is in the form of a binary observation from each study participant, our proposed pseudo-Wald statistic, $\widetilde{W}_{1.5}$, outperforms existing Wald statistics using model-based or sandwich SEs. The Wald statistic using model-based SEs can produce a test size smaller than the nominal value, and therefore will produce less power than our pseudo-Wald statistic under the alternative hypothesis. Additionally, test size can be too large and is

Table 2.3: Estimates, p-values, and 95% confidence intervals (CIs) resulting from the analysis of the breast screening data. Critical values were obtained from the $N(0,1)$ and t_{26} distributions. P-values and CIs correspond to the Wald test using the SE estimate in the corresponding row. $\widetilde{SE}^{*}_{1.5}(\hat{\beta}_{1MQL})$ indicates the use of Equation (2.5) with our pseudo-SE method.

	Estimate	$N(0,1)$		t_{26}	
$\hat{\beta}_{1MQL}$	1.138				
		p-value	95% CI	p-value	95% CI
$\widehat{SE}_{MQL}(\hat{\beta}_{1MQL})$	0.499	0.023	(0.160, 2.116)	0.031	(0.113, 2.163)
$\widetilde{SE}_{1.5}(\hat{\beta}_{1MQL})$	0.479	0.017	(0.200, 2.076)	0.025	(0.154, 2.122)
$\widetilde{SE}^{*}_{1.5}(\hat{\beta}_{1MQL})$	0.498	0.022	(0.162, 2.114)		
$\widehat{SE}_{S}(\hat{\beta}_{1MQL})$	0.417	0.006	(0.322, 1.955)	0.011	(0.282, 1.995)
$\widehat{SE}_{SBC}(\hat{\beta}_{1MQL})$	0.448	0.011	(0.261, 2.015)	0.017	(0.218, 2.058)

inconsistent when using the sandwich-based methods. Therefore, we recommend that $\widetilde{W}_{1.5}$ should be utilized with t_n critical values, or with the proposed inflated SE given in Equation (2.5) and $N(0,1)$ critical values, for hypothesis testing and obtaining confidence intervals.

One may be interested in the validity of the use of $\widetilde{W}_{1.5}$ when the nominal level is, say, 0.01 or 0.10, rather than a traditional value of 0.05. The density corresponding to $\widetilde{W}_{1.5}$ is just a widening of the bell-shaped density of W_{Reg}, especially in the tails. $\widetilde{W}_{1.5}$ produces test sizes at the nominal level of 0.05 and also has a bell-shaped density, making it unlikely that choosing a nominal level smaller than 0.05 will yield a larger test size. With levels larger than 0.05, where the density corresponding to W_{Reg} is very similar to the $N(0,1)$ density, there is a possibility that $\widetilde{W}_{1.5}$ may lead to a slightly inflated realized test size. However, this possible increase in realized size will be of little concern for fixed nominal levels regularly used in practice.

Throughout this manuscript, we have assumed the ICC is equal for all clusters. If the correlation varies from cluster to cluster, our bias and covariance formulas for the MQL estimates take this variable correlation into account, and our method is still valid when the varying correlations are known. However, the correlations will need to be estimated in

practice. If one were to incorrectly assume a common ICC, the size of the proposed pseudo-Wald test may not be nominal if the quantities $\hat{q}_i = n_i/[1 + (n_i - 1)\hat{\rho}_{ANOVA}]$, $i = 1, 2, \ldots n$, are not close to the values that would have resulted with correctly specified cluster-specific correlation estimates. This is more likely to occur in scenarios in which cluster sizes are large and correlation varies moderately, or with smaller clusters with large variations in ICC.

Our proposed method can be useful in the analysis of rare or common events data where $|\pi_C - 0.5|$ and $|\pi_T - 0.5|$ are very close to 0.5. Due to the marginal probabilities being near the boundary of the parameter space, test size can be smaller than the nominal level even if there is no correlation. Here, our method has no additional limitations as compared to the traditional Wald statistic. Furthermore, our methods are applicable to any non-GRT setting that produces cluster-correlated binary data. These settings include teratology experiments and studies collecting repeated measures on the same subjects.

The results of this chapter focus on Wald tests for a marginal model parameter. Another popular approach for the analysis of correlated binary data is the use of a generalized linear mixed effects model (GLMM), in which the correlation is modeled as a random cluster effect, thereby making interpretation of mean parameters conditional for a given cluster. In the Atri *et al.* (1997) study, the interest was in increasing breast screening in the population, and so a marginal model and our methods would be more suitable than a conditional or random effects approach. A marginal interpretation may not be as suitable, though, if we had a scenario where a binary outcome were measured repeatedly on each patient and a subject-specific interpretation of a mean parameter were of primary interest. In this setting, our methods would not be used, although generalization of our methods to random effects model parameters is certainly worthy of research.

Our presentation did not implement $\hat{\beta}_{1BC}$ in the numerator for any of the presented Wald statistics, even though it will contain less, if any, bias than $\hat{\beta}_{1MQL}$; additionally, its variance is smaller. Neither Cordeiro and McCullagh (1991) nor Cordeiro and Demetrio (2008) proposed a variance estimator for $\hat{\boldsymbol{\beta}}_{BC}$. Cordeiro and McCullagh (1991) showed that

for logistic regression with no over-dispersion, the bias in the maximum likelihood estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{ML}$, is approximately $p\boldsymbol{\beta}/n$ for small $\boldsymbol{\beta}$. From this, King and Zeng (2001) proposed estimating the variance of their $\hat{\boldsymbol{\beta}}_{BC}$ by multiplying the model-based variance by $[n/(n+p)]^2$. We utilize a similar approach; by taking into account that the bias-corrected estimate is approximately a fraction of the MQL estimate, we suggest estimating the SE for $\hat{\beta}_{1BC}$ by multiplying the estimated SE for $\hat{\beta}_{1MQL}$ by $\hat{\beta}_{1BC}/\hat{\beta}_{1MQL}$. One can then incorporate this with the results of this chapter, i.e. use $(\hat{\beta}_{1BC}/\hat{\beta}_{1MQL})\widetilde{SE}_{1.5}(\hat{\beta}_{1MQL})$ as the pseudo-SE estimate implemented inside a Wald statistic with $\hat{\beta}_{1BC}$ in the numerator. This quantity is equivalent to $\widetilde{W}_{1.5}$; therefore, test size remains unchanged. Simulations (not shown) demonstrated that using $\hat{\beta}_{1BC}$ with this SE estimate yields approximately the same coverage probability as if $\hat{\beta}_{1MQL}$ were utilized; however, due to $\hat{\beta}_{1BC}$ being approximately unbiased and less variable, it will yield a more desirable confidence interval.

In further research on testing for a marginal treatment effect, we will study test size resulting from the typical Wald test using the model-based SE when the outcomes of interest are not binary responses. We will also extend our model to include other covariates. Further study is also needed to find more exact SE formulas for the BCEs, and to determine if these would carry more accuracy and utility than our proposed formula.

An R function that implements our proposed pseudo-Wald tests, and also outputs our suggested 95% confidence intervals using $\hat{\beta}_{1BC}$ as the point estimate, can be obtained by contacting the author at pwestgat@umich.edu.

CHAPTER III

The Effect of Cluster Size Imbalance and Covariates on the Estimation Performance of QIF

3.1 Introduction

Correlated data with imbalanced cluster sizes arise often in practice. GRTs and longitudinal studies in which the number of repeated measures is not constant across subjects are two popular examples where data are composed of independent clusters that typically are comprised of varying sizes. With these types of data, individual-level responses within any given cluster are assumed to be correlated.

We particularly focus on GRTs, which are unique from other randomized trials. They typically are comprised of a small number of independent clusters that can be quite large and variable in size. Due to these attributes, statistical power can be quite low, but the study itself can be very costly to conduct. When the desired interpretations for regression parameters are in terms of the population mean, Generalized Estimating Equations (GEE) (Liang and Zeger, 1986) are a popular tool of choice for the analysis of data arising from these trials. They require working correlation and marginal variance structures to be given, but only the mean structure needs correct specification in order to obtain consistent parameter estimates.

Due to potentially low power in these studies, the use of a more efficient method would be very beneficial, which is why we focus on the estimation performance of Quadratic Inference Functions (QIF). With the same limited requirements as GEE, Qu, Lindsay, and Li (2000)

proposed this as an alternative method, which is a combination of GEE and the Generalized Method of Moments (GMMs) (Hansen, 1982). QIF asymptotically has greater efficiency than GEE when employing the incorrect covariance structure, and is as efficient when using the correct structure (Qu *et al.*, 2000; Song, 2007; Song *et al.*, 2009). This result depends on all cluster sizes being equal if a common exchangeable correlation is implemented, such that both procedures' estimating equations are within the same class. Many papers, such as Qu *et al.* (2000), demonstrate the utility QIF has over GEE when using a working exchangeable or AR-1 correlation matrix. No paper, however, has studied the finite sample estimation precision, or reliability, of QIF as compared to GEE when cluster sizes vary and the exchangeable structure is reasonably employed, such as in common GRT settings. Additionally, the effect of covariates on QIF's estimation performance has not been considered.

Our motivating dataset comes from Yudkin and Moher (2001), who discuss issues with an ongoing GRT dealing with coronary heart disease (CHD) and promoting secondary prevention via two interventions as compared with a control that gives ordinary care to patients. They give a table of baseline results on four variables and the size of each of the twenty-one practices, or clusters, participating in the study. Using the presented data, we found the number of patients in each practice who were recently adequately assessed for three CHD risk factors. The other three variables were practice-level proportions of patients having a record of treatment with aspirin, hypotensives, or lipid-lowering drugs since their diagnosis with CHD. Practices varied in size from twenty-eight to 244.

One issue Yudkin and Moher (2001) discuss with the baseline data is how to utilize restricted randomization of practices to trial arms such that balance, in terms of adequate assessment and the three records of treatment, is achieved. Our motivation, however, is in quantifying the association between the marginal probability of a practice, which gives ordinary patient care, having recently adequately assessed any given patient and the proportion of patients in that practice having a record of treatment with any of the three drug types. Use of the logistic link would be common for this marginal model, in which the proportion of pa-

tients for any one of the three records of drug treatment could simply be used as a continuous covariate. The true, but unobserved, probabilities of adequately assessing any given patient can vary across practices about their marginal means due to unknown factors, thus inducing correlation among patients within the same practice. Therefore, a common exchangeable correlation would be a natural structure to implement in the estimation method.

We later show that GEE and QIF can produce notably different probability estimates from the analysis of this data, leading to the issue of which method gave more trustworthy estimates, and in general, which of these two methods would be best to use for the analysis of data from any GRT. As QIF theoretically is equally or more efficient than GEE, one may think that its estimates here would be more reliable. For example, the degree of correlation could depend on the proportion of patients with a record of drug treatment, and QIF should take this into account if that truly were the case, while GEE cannot. The purpose of this chapter is to give details into how QIF and GEE can give notably different estimates in this or any other GRT setting, and why GEE may actually be better to employ in GRT scenarios and many other small-sample settings.

Section 3.2 discusses GEE and QIF in more detail, including comparisons of their respective classes of estimating equations when an exchangeable correlation structure is implemented. In Section 3.3, we discuss how cluster size imbalance and covariates can cause QIF, as compared with GEE, to lose estimation precision when the number of clusters is not large. Additionally, a new version of QIF with corresponding estimating equations in the same class as GEE when clusters vary in size is presented to argue that the empirical nature of the weighting matrix used in QIF, rather than class, has the largest impact on estimation performance. In Section 3.4, we present simulation results, with emphasis on our motivating dataset and general GRTs, demonstrating the differences in the precisions of parameter estimates from GEE and QIF. Furthermore, the distinct estimation performances of these two methods are shown in application to the motivating dataset. Concluding remarks are given in Section 3.5.

3.2 Marginal Models

3.2.1 Generalized Estimating Equations

We have N independent clusters of data, and cluster $i, i = 1, 2, \ldots N$, has n_i observations, outcome vector $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^T$, and mean vector $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i)$. The marginal mean structure is specified as $h(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{x}_i\boldsymbol{\beta}$, where the jth row of $\mathbf{x}_i$, $j = 1, 2, \ldots n_i$, is $\mathbf{x}_{ij} = [x_{ij0}, x_{ij1}, \ldots, x_{ij(p-1)}]$, the vector of covariate values for the jth observation in cluster i, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^T$ is a $p\times$ 1 vector of corresponding regression parameters. The estimates for $\boldsymbol{\beta}$ are obtained by setting the GEE equal to zero,

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.1}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $\boldsymbol{V}_i$ is the working covariance structure for $\boldsymbol{Y}_i$. $\boldsymbol{V}_i$ can be written as $\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i$ is a diagonal matrix of the working marginal variances for the n_i observations, and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is their working correlation structure with parameter(s) $\boldsymbol{\alpha}$. When the covariance structure is correctly specified and a consistent estimate for $\boldsymbol{\alpha}$ is employed, GEE as given in Equation (3.1) are optimal estimating equations (Small and McLeish, 1994). If $\boldsymbol{V}_i$ is misspecified, the parameter estimates, $\hat{\boldsymbol{\beta}}$, are still consistent when the mean structure is correct.

When implementing an exchangeable correlation structure, Equation (3.1) can be rewritten as

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2}(\gamma_{1i}\boldsymbol{M}_{1i} + \gamma_{2i}\boldsymbol{M}_{2i})\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.2}$$

where $\gamma_{1i} = -[(n_i - 2)\rho_i + 1]/k_i$, $\gamma_{2i} = \rho_i/k_i$, $k_i = (n_i - 1)\rho_i^2 - (n_i - 2)\rho_i - 1$, $\boldsymbol{M}_{1i}$ is an $n_i \times n_i$ identity matrix, $\boldsymbol{M}_{2i}$ is an $n_i \times n_i$ matrix composed of zeros on the diagonal and ones elsewhere, and ρ_i is a function of $\boldsymbol{\alpha}$ and is the assumed common correlation within the ith cluster (Qu *et al.* 2000). If cluster sizes are all equal and a constant correlation is assumed

across clusters, Equation (3.2) is easily seen as being in the class of estimating equations given by

$$\sum_{r=1}^{2} \boldsymbol{B}_r \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_r \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{3.3}$$

where $\boldsymbol{B}_r$, $r = 1, 2$, are $p \times p$ arbitrary nonrandom matrices. $\boldsymbol{M}_{1i}$ and $\boldsymbol{M}_{2i}$, $i = 1, 2, \ldots N$, do not change across clusters, and therefore are denoted here as $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. They can be thought of as basis matrices since all other quantities inside the two sums over the N clusters are the same (Qu *et al.*, 2000). With respect to GEE, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are identity matrices multiplied by γ_1 and γ_2, respectively, where $\gamma_r = \gamma_{ri}$, $r = 1, 2$; $i = 1, 2, \ldots N$. When clusters vary in size, or a common correlation is no longer used across all clusters, GEE is not within this class of estimating equations and belongs to a more specific class given by

$$\sum_{r=1}^{2} \boldsymbol{O}_r \sum_{i=1}^{N} \gamma_{ri} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_{ri} \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{3.4}$$

where $\boldsymbol{O}_r$, $r = 1, 2$, are $p \times p$ arbitrary nonrandom matrices equal to the identity matrix for GEE.

3.2.2 Quadratic Inference Functions

The QIF proposed by Qu *et al.* (2000) combines the methods of GMMs and GEE. It assumes $\boldsymbol{R}_i^{-1}(\boldsymbol{\alpha}) = \sum_{r=1}^{m} \gamma_{ri} \boldsymbol{M}_{ri}$, where $\boldsymbol{M}_{ri}$, $r = 1, 2, \ldots m$, are known basis matrices and γ_{ri}, $r = 1, 2, \ldots m$, are functions of $\boldsymbol{\alpha}$ that we will refer to as correlation weights. An exchangeable structure is a specific case where the inverse of the correlation matrix can be written as the sum of weighted basis matrices, as shown in Equation (3.2). Unstructured, AR-1, and independence are the other correlation structures QIF currently supports via this assumption, each of which have inverses that can at least be approximated by using two basis matrices (Song *et al.*, 2009). We do not focus on these in this chapter since QIF and GEE lead to identical estimating equations when using independence (Qu and Song, 2004),

unstructured cannot be implemented when clusters vary in size, and exchangeable is more commonly employed than AR-1 in GRT settings. However, the discussion in Section 3.3 on the lost reliability of QIF as compared with GEE is still at least partially relevant to AR-1 and unstructured working correlations.

Equation (3.2) can be viewed as the sum of two unbiased estimating equations, each of which are used to build extended score equations defined as

$$\bar{g}_N(\boldsymbol{\beta}) = \frac{1}{N} g_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} g_i(\boldsymbol{\beta}) = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^{N} g_{1i}(\boldsymbol{\beta}) \\ \frac{1}{N} \sum_{i=1}^{N} g_{2i}(\boldsymbol{\beta}) \end{bmatrix}, \tag{3.5}$$

or

$$\bar{g}_N(\boldsymbol{\beta}) = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^{N} g_{1i}(\boldsymbol{\beta}) \\ \vdots \\ \frac{1}{N} \sum_{i=1}^{N} g_{mi}(\boldsymbol{\beta}) \end{bmatrix}$$

in a general setting. The number of extended score equations is m times the number of regression parameters, and therefore cannot be set equal to zero to obtain parameter estimates, as is done for GEE, since no identifiable solution exists. The extended score equations are used in Hansen's (1982) GMMs to create the QIF, defined as

$$Q_N(\boldsymbol{\beta}) = N \bar{g}_N^T(\boldsymbol{\beta}) C_N^{-1}(\boldsymbol{\beta}) \bar{g}_N(\boldsymbol{\beta}) = \left[\sum_{i=1}^{N} g_i^T(\boldsymbol{\beta}) \right] \left[\sum_{i=1}^{N} g_i(\boldsymbol{\beta}) g_i^T(\boldsymbol{\beta}) \right]^{-1} \left[\sum_{i=1}^{N} g_i(\boldsymbol{\beta}) \right].$$

The estimate for $\boldsymbol{\beta}$ can now be found by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta})$, which is asymptotically equivalent to solving

$$N \nabla \bar{g}_N^T(\boldsymbol{\beta}) C_N^{-1}(\boldsymbol{\beta}) \bar{g}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \nabla \bar{g}_N^T(\boldsymbol{\beta}) C_N^{-1}(\boldsymbol{\beta}) g_i(\boldsymbol{\beta}) = \mathbf{0} \tag{3.6}$$

for $\boldsymbol{\beta}$, where ∇ denotes the gradient with respect to $\boldsymbol{\beta}^T$. Here, the empirical covariance

matrix $C_N(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N} g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta})$ is used to estimate the optimal weighting matrix $\boldsymbol{\Sigma}_N = (1/N)\sum_{i=1}^{N} Cov[g_i(\boldsymbol{\beta})]$. Results using $C_N(\boldsymbol{\beta})$ are asymptotically equivalent to using $\boldsymbol{\Sigma}_N$, since $C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$ (Qu *et al.*, 2000; Pilla and Loader, 2006).

In practice, $g_{ri}(\boldsymbol{\beta}) = \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_{ri} \boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $r = 1, \ldots, m$; $i = 1, 2, \ldots N$, are regularly implemented in QIF's extended score equations. These ignore the correlation weights, implying that $\boldsymbol{\alpha}$ does not need to be estimated. When using an exchangeable structure and cluster sizes do not vary, Equation (3.6) is in the class of estimating equations given by Equation (3.3). When cluster sizes vary, Equation (3.6) is in the class of estimating equations given by Equation (3.3) with one difference: the dimensions of the basis matrices are dependent on cluster size. This is not in the class given by Equation (3.4) to which GEE belongs in this scenario. In order for Equation (3.6) and GEE to be in the same class when cluster sizes vary, the extended score equations need to incorporate the correlation weights; i.e. use $g_{ri}(\boldsymbol{\beta}) = \gamma_{ri} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_{ri} \boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $r = 1, 2$; $i = 1, 2, \ldots N$, inducing the need to estimate $\boldsymbol{\alpha}$.

From Lindsay (1982), Hansen (1982), and Small and McLeish (1994), Qu *et al.* (2000) show that, since $C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$, the estimating equations given in Equation (3.6) are fully efficient when the covariance structure is correctly specified and they are in the same class as GEE, and always optimal in the Löwner ordering among estimating equations within their given class. When cluster sizes are constant and assuming a common correlation with an exchangeable structure, Equation (3.6) and GEE are in the class of estimating equations given by Equation (3.3), and therefore QIF has the theoretical advantage of asymptotically producing parameter estimates having equal or less variance than estimates from GEE. When cluster sizes vary and an exchangeable correlation is utilized, QIF loses this theoretical advantage unless the correlation weights are used inside the extended score equations. In this case, GEE and Equation (3.6) both belong to the class given by Equation (3.4).

3.3 Empirical Weighting and QIF Estimation Precision

3.3.1 The Impacts of Imbalanced Cluster Sizes and Covariates

QIF uses an empirical weighting matrix, $C_N(\boldsymbol{\beta})$, to estimate $\boldsymbol{\Sigma}_N$, which is optimal. Asymptotically, using this matrix is the source of QIF's efficiency advantage. Even when clusters vary in size, causing QIF and GEE to not be directly comparable in terms of efficiency theory when using a common exchangeable correlation, QIF still has an advantage in the sense that it is composed of a consistent weighting matrix whether the true covariance structure is implemented or not. However, the use of $C_N(\hat{\boldsymbol{\beta}})$ can lead to lost, rather than gained, estimation precision as compared to GEE, in small to moderately sized samples, such as the GRT dataset containing only twenty-one practices. Specifically, imbalance in cluster sizes, the number of covariates, and their corresponding values, all affect the amount of empirical information from $C_N(\hat{\boldsymbol{\beta}})$ used to estimate the working weights inside the estimating equations. The weight(s) used by QIF for any given cluster's outcomes can therefore be quite variable when N is not large, which we discuss next. Conversely, GEE uses outcomes from all clusters to estimate a common correlation parameter that is used inside a fixed correlation, and thus weighting, structure. This produces weights, and thus estimating equations, that are possibly much less variable than their counterparts used by QIF, potentially leading to greater estimation reliability.

Before discussing QIF's empirical weighting nature, we emphasize that use of some $\hat{\boldsymbol{\beta}}$ to estimate $\boldsymbol{V}_i$ in GEE and empirical covariances, $g_i g_i^T$, $i = 1, 2, \ldots N$, in QIF, is required in practice. This will have little influence on parameter estimates from GEE due to the fixed weighting structure this method employs. However, for small N, not only can QIF's weights be quite variable, $Var(\hat{\boldsymbol{\beta}})$ can be large as well, thus implying that the estimated empirical covariances $(g_i(\hat{\boldsymbol{\beta}})g_i^T(\hat{\boldsymbol{\beta}}),\ i = 1, 2, \ldots N)$ can be notably different than the true empirical covariances $(g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta}),\ i = 1, 2, \ldots N)$. Due to the variable weighting nature used by QIF, these differences between estimated and true empirical covariances can lead to notable

differences between the estimated and true weights, which affects parameter estimation. Windmeijer (2000, 2005) also notes in the GMM literature that the variance of the final estimate for $\boldsymbol{\beta}$ is influenced by the use of $\hat{\boldsymbol{\beta}}$ inside the empirical weighting matrix.

We first assume $\boldsymbol{\Sigma}_N$ is known. This matrix averages the covariances of the N extended score components, $g_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots N$, such that $\bar{g}_N(\boldsymbol{\beta})$ is optimally weighted inside Equation (3.6). This typically does not allow the sole use of $Cov[g_i(\boldsymbol{\beta})]$ to weight $g_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots N$, inside the estimating equations. Rather, QIF uses $\nabla \bar{g}_N^T(\boldsymbol{\beta})\boldsymbol{\Sigma}_N^{-1}$ in its estimating equations to determine how much individual weight should be given to each $g_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots N$, by using the averaged information from sensitivities (Song, 2007) and covariances over the extended score components from all N clusters.

For a simple example, the baseline marginal probability of adequate assessment was of interest to Yudkin and Moher (2001), corresponding to an intercept-only model. We later show that the value used to weight the difference in the observed and expected number of patients adequately assessed in a given practice is a linear function of size in this simple model, due to each extended score component being linear in terms of size and the corresponding residual. Therefore, by using information from all N practices, QIF takes into account how sensitivities and covariances change on average with respect to size, and then determines the appropriate weight function. As covariates, such as drug treatment proportions, are added to the model, this type of average sensitivity and covariance trend has to be determined with respect to the numerous combinations of size and covariates, making $\nabla \bar{g}_N^T(\boldsymbol{\beta})\boldsymbol{\Sigma}_N^{-1}$ more complicated, as we will later explicitly show in Equation (3.7).

In practice, $C_N(\hat{\boldsymbol{\beta}})$ is used in place of $\boldsymbol{\Sigma}_N$. As covariances can potentially depend upon cluster size and covariates, $C_N(\hat{\boldsymbol{\beta}})$ attempts to determine the appropriate trends by averaging over all estimated empirical covariances, $g_i(\hat{\boldsymbol{\beta}})g_i^T(\hat{\boldsymbol{\beta}})$, $i = 1, 2, \ldots N$. In finite samples, estimating the effect of these factors on covariances can lead to additional weighting variability that may be detrimental even when the working covariance structure is incorrect. This is particularly true for GRTs, as N can be quite small.

The sensitive nature of weighting used by QIF via C_N is exhibited by Qu and Song (2004), as they show QIF is robust to outliers and contaminated data. Specifically, they prove that $||\nabla \bar{g}_N^T(\boldsymbol{\beta}) C_N^{-1}(\boldsymbol{\beta}) g_i(\boldsymbol{\beta})||^2 \to 0$ as $||\boldsymbol{Y}_i - \boldsymbol{\mu}_i|| \to \infty$, where $||\boldsymbol{K}|| = [tr(\boldsymbol{K}^T \boldsymbol{K})]^{1/2}$ for some arbitrary matrix $\boldsymbol{K}$. Note that large $||\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i||$ implies that the estimated empirical covariance, $g_i(\hat{\boldsymbol{\beta}}) g_i^T(\hat{\boldsymbol{\beta}})$, of the corresponding extended score component is also large, and will actually downweight the outcomes from this cluster without necessarily doing the same to other clusters. This shows that although QIF does not solely use $\widehat{Cov}[g_i(\hat{\boldsymbol{\beta}})] = g_i(\hat{\boldsymbol{\beta}}) g_i^T(\hat{\boldsymbol{\beta}})$ to individually weight g_i, $i = 1, 2, \ldots N$, in Equation (3.6), the estimated empirical covariance from any given cluster can be the most important factor into how much weight its outcomes receive in the estimating equations.

We now focus on the direct influence from the ith cluster's empirical covariance, for some $i \in [1, 2, \ldots N]$, on the estimation of weights given to outcomes from any other cluster, say cluster k. If the only difference between $g_k(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta})$ is that $\boldsymbol{Y}_i \neq \boldsymbol{Y}_k$, i.e. $\mathbf{x}_i$ and $\mathbf{x}_k$ are interchangeable, implying equivalence in terms of size and covariate values, then $\boldsymbol{Y}_k$ will be weighted in the same manner as $\boldsymbol{Y}_i$, and $E[g_i(\boldsymbol{\beta}) g_i^T(\boldsymbol{\beta})] = E[g_k(\boldsymbol{\beta}) g_k^T(\boldsymbol{\beta})]$. For instance, two practices in the motivating dataset have an equal number of patients with CHD, and therefore would be equivalently weighted in the intercept-only model. This implies, for example, if $||\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i||$ is much larger than expected, then both $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_k$ will be equally downweighted inside the estimating equations unless $||\boldsymbol{Y}_k - \hat{\boldsymbol{\mu}}_k||$, and therefore $g_k(\hat{\boldsymbol{\beta}}) g_k^T(\hat{\boldsymbol{\beta}})$, is smaller than expected by an amount such that it offsets the overestimated covariance from cluster i. However, if there are other clusters with similar covariate design matrices, their estimated empirical covariances may have a less direct impact on the estimation of weights given to $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_k$, but may partially offset some of the random variability from $g_i(\hat{\boldsymbol{\beta}}) g_i^T(\hat{\boldsymbol{\beta}})$ and $g_k(\hat{\boldsymbol{\beta}}) g_k^T(\hat{\boldsymbol{\beta}})$. For instance, if we are using the record of aspirin treatment to predict adequate assessment, the most direct influence on the weight given to outcomes from any given practice comes from other clusters that have a similar size and percentage of aspirin treatment. Therefore, unless there are numerous clusters with the same or a similar

design, thus directly improving the estimates for their outcomes' corresponding weights, QIF's estimating equations can be quite variable. Additionally, similarity is rather arbitrary in terms of covariate design matrices, and we later give specific examples that explicitly give insight into what is deemed similar with respect to obtaining weights.

As covariates are added to the regression model, such as using all three proportions of treatment records as predictors of adequate assessment in the same model, or clusters become more variable in size, there is greater dissimilarity across clusters and the weighting of their outcomes will be more complicated. In general, empirical covariances used to directly influence the weights given to outcomes from any cluster come from a group of similar clusters, although there can be notable indirect influence from dissimilar clusters due to the linear trend as shown later in Equation (3.7). Greater dissimilarity can therefore result in less direct empirical information being utilized to estimate weights for any given cluster's outcomes, due to $\nabla \bar{g}_N^T C_N^{-1}$ taking these dissimilarities in the N extended score components into account. This is potentially a good property when we have outliers to downweight, but when N is not arbitrarily large as is the case in a GRT, this potentially leads to high variability in the weights, possibly causing QIF to be a less reliable estimation method than GEE.

We now present a general scenario to clearly show the influence from dissimilarities across clusters and how outcomes are weighted. We have p cluster-level covariates, and $h(.)$ is the canonical link, allowing Equation (3.6) to simplify to

$$\sum_{i=1}^{N} \begin{bmatrix} \sum_{j=0}^{p-1} \kappa_{j0} x_{ij} + (n_i - 1) \sum_{j=p}^{2p-1} \kappa_{j0} x_{i(j-p)} \\ \vdots \\ \sum_{j=0}^{p-1} \kappa_{j(p-1)} x_{ij} + (n_i - 1) \sum_{j=p}^{2p-1} \kappa_{j(p-1)} x_{i(j-p)} \end{bmatrix} (Y_i - n_i \mu_i), \tag{3.7}$$

in which $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, $E(Y_{ij}) = \mu_i$ and x_{ij} is the value of the jth covariate, $j = 0, \ldots p-1$, for the ith cluster. The kappas are all estimated using functions of estimated parameters, cluster size, covariate values, and the N empirical covariances which have the most influence.

The number of kappa parameters increases with the number of covariates, and the amount of weight given to any cluster's outcomes relies upon linear combinations of its size and covariate values. For fixed N, as the number of kappa parameters increases, the amount of information utilized from empirical covariances to estimate any given kappa can decrease. For example, when only using one drug treatment percentage as a covariate, QIF estimates eight kappas, while this number increases to thirty-two when using all three treatments in the model.

To give specific examples clearly demonstrating the variability in QIF's estimating equations' weights due to dissimilarities across clusters, we first return to the intercept-only model, but allow clusters to be one of two possible sizes for simplicity. There are two kappas to estimate, and Equation (3.7) reduces to $\sum_{i=1}^{N}[\kappa_{00}+\kappa_{10}(n_i-1)](Y_i-n_i\mu) = \sum_{i=1}^{N} w_{QIFi}(Y_i-n_i\mu)$, where μ is the marginal mean shared by all outcomes. Here, the weight given to the ith cluster will actually be estimated using only the empirical covariances of the extended score equation components from clusters of that vary same size, rather than using the empirical variability from all N clusters as does GEE when estimating a common correlation parameter inside the corresponding fixed weighting function. This nature of weighting is advantageous for QIF when N is arbitrarily large and the true covariances do depend on cluster size in some misspecified manner, but for small N can lead to increased variability in the working weights that can more than offset this advantage in terms of estimation performance.

If we keep N fixed and extend the model to resemble a general GRT in which there is only one covariate, a cluster-level intervention indicator, then there are eight unknown kappas, each estimated with a larger variability. In this situation, QIF carries out estimation in a manner equivalent to fitting an intercept-only model for each trial arm. This implies that $\nabla\bar{g}_N^T C_N^{-1}$ accounts for dissimilarities corresponding to study arm and cluster size, and so the weight given to $(Y_i - n_i\mu_i)$, $i = 1, 2, \ldots N$, is obtained using only the empirical covariances from equivalently sized clusters within the same trial arm.

In practice, as is the case with our motivating dataset, there typically is much larger

imbalance in size across clusters. When this occurs, $\nabla \bar{g}_N^T C_N^{-1}$ has to determine if the weight given to outcomes should increase or decrease with size, which is done separately for each trial arm. For instance, suppose outcomes from clusters larger in size have combined empirical covariances smaller than their true covariances. Here, the QIF's estimating equations will then overweight larger clusters, while the weight given to smaller clusters may also be influenced due to the linear trend shown in the intercept-only model's estimating equation. Although clusters cannot be categorized into distinct groups as was done when there were only two possible sizes, the empirical covariances directly influencing the working weight for any given cluster's outcomes are from that individual cluster and other similarly sized clusters in the same trial arm. Dealing with the baseline dataset, we focus this issue toward an intercept-only model for adequate assessment. Here, the estimated empirical covariances from practices with similar numbers of CHD patients have the most direct effect on the estimated weights given to the number of adequately assessed patients from these same practices.

If we were to expand this model even further by using a continuous covariate, such as aspirin treatment percentage, rather than an indicator, or by adding even more covariates such as the other two drug treatment percentages to the model, it is easy to see by Equation (3.7) that there will be more unknown kappa parameters to estimate and potentially less information utilized per kappa estimate. Additionally, the manner of weighting shown in the general GRT example continues in that similar clusters, determined with respect to size and covariate values, have the strongest direct influence on how much estimated weight their corresponding outcomes receive in QIF's estimating equations. For small N, the variability in these estimating equations can therefore be quite large, possibly making GEE more reliable.

3.3.2 The Impact of Estimating Equations Class

In Section 3.2 we defined a new QIF version such that the rth component of its extended score equations is given by $g_{ri}(\boldsymbol{\beta}) = \gamma_{ri} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_{ri} \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $r = 1, 2$; $i = 1, 2, \ldots N$,

implying the corresponding asymptotic estimating equations are in the same class as GEE, given by Equation (3.4), when clusters vary in size. Theoretically, this QIF version will have an asymptotic efficiency advantage over GEE, particularly when the covariance structure is misspecified. When N is not large, however, this version of QIF can also be less reliable than GEE.

Assuming the same intercept-only model presented in the previous subsection, the estimating equation from this new QIF version that is in the same class as GEE is given by $\sum_{i=1}^{N}[\kappa_{00}\gamma_{1i} + \kappa_{10}(n_i - 1)\gamma_{2i}](Y_i - n_i\mu) = \sum_{i=1}^{N} w_{QIFi}(Y_i - n_i\mu)$, where κ_{00} and κ_{10} now also include γ_{1i} and γ_{2i}, $i = 1, 2, \ldots N$. When cluster sizes are similar and the correlation estimate is not large, γ_{1i} and γ_{2i} approximately cancel out inside the weights, which in turn causes the weights to closely approximate their corresponding values from the regular QIF. As the variation in cluster sizes increases, as is seen in GRTs, or the correlation estimate becomes large, there can be a distinct difference between corresponding weights from the two QIF versions. However, the empirical nature of the weighting matrix for the newly defined QIF still exists and influences weighting accuracy, possibly decreasing estimation reliability as compared with GEE.

3.4 The Impacts of Cluster Sizes and Covariates

3.4.1 Shown Via Simulation Study

3.4.1.1 Intercept-Only Simulations

Employing a common exchangeable correlation, we first demonstrate the difference between both QIF versions and GEE in the context of an intercept-only model, representing the setting in which we are only interested in estimating the marginal probability of adequate assessment. Results from ten random simulations, with outcomes generated from a beta-binomial distribution, are presented in Table 3.1, including the intercept estimates and ratios equaling the estimated weight given to a cluster of size fifty divided by the

Table 3.1: Intercept estimates and weight ratios, equaling the estimated weight given to a cluster of size fifty divided by the estimated weight given to a cluster of size 150, from GEE and both QIF versions. The first (last) five simulation results come from the analyses of randomly generated datasets in which outcomes had a marginal probability of 0.25 (0.05) and exchangeable correlation of 0.05.

	GEE		QIF		QIF in GEE Class	
Simulation	$\hat{\beta}_0$	Ratio	$\hat{\beta}_0$	Ratio	$\hat{\beta}_0$	Ratio
1	-1.18	2.49	-1.18	1.50	-1.18	2.01
2	-0.95	2.59	-0.99	0.62	-0.97	1.00
3	-1.12	2.47	-1.17	1.10	-1.15	1.52
4	-1.15	2.66	-1.13	1.87	-1.13	3.13
5	-0.99	2.64	-0.99	1.51	-0.98	4.03
6	-3.23	2.34	-3.32	2.19	-3.32	4.39
7	-3.09	2.25	-3.06	1.44	-3.09	2.20
8	-2.88	2.19	-3.02	6.74	-3.07	17.00
9	-3.26	2.50	-3.22	3.94	-3.31	5.33
10	-2.93	2.33	-2.92	1.22	-2.93	2.09

estimated weight given to a cluster of size 150. Data were generated using the model $logit(\pi) = log(\pi) - log(1-\pi) = \beta_0$, in which π is the marginal probability for any given outcome. Values for the marginal probability were 0.25 (0.05) for the first (last) five simulations, implying $\beta_0 = -1.10$ ($\beta_0 = -2.94$), while the common correlation was 0.05. Each simulated dataset consisted of twenty-one practices, with corresponding sizes generated by a normal distribution with mean 100 and standard deviation fifty, approximately representing the empirical distribution of sizes contained in the motivating dataset. Generated sizes were rounded to the nearest integer and forced to take on values between twenty-five and 250. In this setting, the optimal weight is given as $w_i = [1 + (n_i - 1)0.05]^{-1}$ inside the estimating equation $\sum_{i=1}^{N} w_i(Y_i - n_i\pi)$. The optimal ratio is therefore 2.45.

In these ten simulations, the weights used by GEE were much less variable than those used by either QIF version. For GEE, clusters of size fifty were given anywhere from 2.19 to 2.66 times more weight than clusters of size 150. QIF, however, once gave more weight to

larger clusters, and clusters of size fifty were given anywhere from thirty-eight percent less to 6.74 times more weight than clusters of size 150. The QIF with corresponding estimating equations in the same class as GEE produces notably different weight ratios, but performed similarly to QIF with respect to the variability in its working weights. Due to this variability in relative weighting, estimation precision is lost as compared to GEE here.

For each of the two examined settings, we also performed 1,000 additional simulations generated in the same manner. When the marginal probability was 0.25 (0.05), the empirical mean squared error (MSE) for GEE's intercept estimate was only eighty-two (sixty-two) and eighty-six (sixty-five) percent as large as the MSEs produced by QIF and the newly defined QIF, respectively, implying GEE was more precise. The last five simulation results in Table 3.1 show that the weights implemented by both QIF versions were more variable when the marginal probability was 0.05, leading to the smaller MSE ratios in this setting. Additionally, the newly defined QIF only slightly increased precision over the typical QIF.

3.4.1.2 Description of General Simulation Settings

We now compare the MSE of both QIF versions and GEE, all implementing a common exchangeable correlation structure, in a variety of simulations representing GRT scenarios. Table 3.2 presents empirical MSEs for each QIF version, in addition to ratios comprised of the MSEs from GEE and the respective QIF version in the numerator and denominator, respectively. The presented MSE quantity for any given method is the sum of the empirical MSEs from all non-intercept regression parameter estimates in the respective model. Five different scenarios comprised of four settings each were examined in 1,000 simulations. A beta-binomial distribution was used to generate outcomes.

In the first three scenarios, which represent general GRTs, the true model in each scenario is a logistic regression with only cluster-level covariates. Scenarios one and three only use an intervention indicator, with $N/2$ clusters in each trial arm. The second scenario uses an additional indicator and a continuous covariate, with corresponding parameters $\beta_2 =$

$\beta_3 = 1$. There were $N/4$ clusters for each of the four possible combinations for the two indicators, while the values for the continuous covariate were independently drawn from $Uniform(-1, 1)$. Cluster sizes varied uniformly and independently from 5 to 20 in the first scenario and 25 to 150 in the next two. Table 3.2 presents the number of clusters and marginal probabilities for control (π_C) and intervention (π_T) clusters when the model only includes an intervention indicator (all other covariates equal zero).

The models in the last two scenarios are representative of the analyses we later carry out that use the percentages of patients with records of drug treatment to predict adequate assessment. Scenario 4 represents the logistic regression in which the proportion of patients having a record of treatment with aspirin is used as a predictor, while the fifth scenario is representative of using proportions from all three drug treatment records as covariates. Cluster sizes were generated in the same manner as for the intercept-only model simulations. The percent with a record of aspirin treatment varied uniformly and independently from sixty-six to ninety-six in both scenarios, while the percents of patients having a record of treatment with hypotensives or lipid-lowering drugs were generated uniformly and independently on the set of integers ranging from thirty-seven to seventy-five and fourteen to fifty, respectively, in the last scenario. In the first two settings of Scenarios 4 and 5, $\beta_0 = -1$ and all other parameters were given values of zero. In the last two settings, $\boldsymbol{\beta} = [-2, 0.015]^T$ in Scenario 4 and $\boldsymbol{\beta} = [-3, 0.01, 0.02, 0.04]^T$ in Scenario 5. Table 3.2 indicates the number of practices, N.

Although its structure was not misspecified, the exchangeable correlation value was allowed to vary from cluster to cluster in some settings since this will be accounted for, at least asymptotically, by $C_N(\boldsymbol{\beta})$. In the first scenario, correlations were dependent on whether a cluster was in the control or intervention arm. Denoting correlation pairs by $(\rho_{control}, \rho_{intervention})$, we used (0.1, 0.1), (0.3, 0.1), (0.3, 0.3), and (0.15, 0.05), for the first through fourth settings, respectively. The second scenario made the ith cluster's correlation a function of its marginal probability, $log[\rho_i/(1-\rho_i)] = \lambda_1 + \lambda_2|\pi_i - 0.5|$, in which λ_2 was

-5 for the first and third settings and -2.5 for the other two, while $\lambda_1 = -2$. Correlations were given by $exp(\varphi_1 + \varphi_2 n_i)/[1 + exp(\varphi_1 + \varphi_2 n_i)]$ in the third scenario, where φ_1 and φ_2 were chosen such that ρ_i, $i = 1, 2, \ldots N$, were in the ranges (0.024, 0.079), (0.022, 0.337), (0.011, 0.119), and (0.008, 0.067) for the first through fourth settings, respectively. The correlation was fixed at 0.05 in the first two settings of Scenarios 4 and 5, but was equal to $log[\rho_i/(1-\rho_i)] = -2.94 + 0.075(x_i - 81)$ and $log[\rho_i/(1-\rho_i)] = -2.25 - 5|\pi_i - 0.5|$ in the last two settings of Scenarios 4 and 5, respectively. Here, x_i represents the practice level proportion of CHD patients having a record of aspirin treatment.

3.4.1.3 Description of Results

Empirical MSEs were dominated by empirical variances, as squared bias was negligible, and we discuss MSE results in terms of precision. The first two scenarios show that imbalance in cluster sizes, even with only one covariate in the model and a large number of clusters, can cause QIF to produce estimates with larger variance than the corresponding GEE estimates. Scenario 2 also shows that adding covariates to the model can cause QIF to lose even more precision, as expected, since this creates greater dissimilarity across clusters and more weight parameters, or kappas, to estimate. The last two scenarios also show that even when the number of clusters is large in a context representing our motivating dataset, which typically is not the case in GRTs, QIF can be considerably less precise than GEE.

Additionally, although GEE assumes a common correlation here, allowing the true correlation value to vary across clusters did not make QIF more reliable, except in three settings of the third scenario. This is an example where QIF may be advantageous over GEE. It appears, however, that this is only the case when marginal probabilities are not near zero or the number of clusters is large. The degree of dependency correlation has on cluster size also is relevant to whether QIF or GEE performs better here. Additionally, the differences in the precisions of QIF and GEE shown in this scenario are small compared to the overall results from the other scenarios, deeming GEE as a more reliable method in general GRT

Table 3.2: Empirical MSEs for both QIF versions, and ratios comparing GEE's empirical MSE to these respective quantities. Common exchangeable correlation structures were employed with these methods. The scenarios presented are general representations of GRTs and the GRT dataset of interest.

		QIF		QIF in GEE Class	
Scenario	N,π_C,π_T	MSE	MSE Ratio	MSE	MSE Ratio
1.1	40, 0.1, 0.1	0.319	0.660	0.333	0.633
1.2	40, 0.5, 0.4	0.137	0.809	0.137	0.811
1.3	200, 0.5, 0.5	0.032	0.962	0.032	0.975
1.4	200, 0.15, 0.1	0.032	0.920	0.032	0.925
2.1	20, 0.5, 0.4	0.400	0.584	0.382	0.608
2.2	20, 0.1, 0.1	0.652	0.598	0.650	0.600
2.3	100, 0.5, 0.5	0.046	0.883	0.045	0.901
2.4	100, 0.15, 0.1	0.074	0.877	0.073	0.884
3.1	20, 0.05, 0.05	0.362	0.708	0.384	0.668
3.2	20, 0.5, 0.5	0.093	1.123	0.091	1.144
3.3	100, 0.5, 0.5	0.007	1.165	0.007	1.161
3.4	100, 0.05, 0.05	0.030	1.032	0.030	1.029
4.1	21	3.1×10^{-4}	0.670	2.6×10^{-4}	0.808
4.2	100	5.7×10^{-5}	0.648	4.3×10^{-5}	0.858
4.3	21	3.6×10^{-4}	0.658	3.0×10^{-4}	0.798
4.4	100	6.4×10^{-5}	0.724	5.2×10^{-5}	0.891
5.1	21	8.0×10^{-4}	0.696	8.3×10^{-4}	0.669
5.2	100	1.2×10^{-4}	0.784	1.1×10^{-4}	0.876
5.3	21	7.9×10^{-4}	0.682	7.8×10^{-4}	0.690
5.4	100	1.2×10^{-4}	0.794	1.1×10^{-4}	0.894

scenarios.

Both QIF versions performed approximately the same in the first three scenarios. However, in all four settings of Scenario 4, along with the settings consisting of 100 clusters in the last scenario, the QIF with estimating equations in the same class as GEE performed notably better than the regular QIF. This may imply that estimating equations class may influence estimation precision in some scenarios, while the empirical weighting employed by both QIF versions still is the major influence of the differences between these two methods and GEE. In all presented simulation results, the estimated correlation used for the QIF with estimating equations in the same class as GEE was taken as the estimate for the common correlation from GEE. We also estimated correlation iteratively inside this QIF version in the same manner as GEE, which led to almost identical results.

3.4.2 Shown Via Application to the Motivating Example

We now return to our motivating dataset. By multiplying the size and adequate assessment percentage for a given practice, and then rounding this quantity to the nearest integer, our dataset had a total of 629 adequately assessed patients, while Yudkin and Moher (2001) report 627. However, this slight difference is not notable in terms of the regression results, which are presented in Table 3.3. We fit three models of the form $logit(\pi_{ij}) = logit(\pi_i) = \beta_0 + \beta_1 x_i$, in which x_i represents the proportion of patients within the ith practice who have a record of treatment with the corresponding drug type, and one model in which $logit(\pi_{ij}) = logit(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$. x_{1i}, x_{2i}, and x_{3i} correspond to the percentages having a record of treatment with aspirin, hypotensives, and lipid-lowering drugs, respectively. π_i is the marginal probability of any given CHD patient in practice i being adequately assessed.

We first demonstrate the difference in weighting between GEE and both QIF versions by estimating the overall marginal probability of adequate assessment, corresponding to an intercept-only model. The estimated weights (marginal probabilities) used (produced) by

Table 3.3: Estimated logistic regression results when analyzing the GRT dataset using the given record of drug treatment proportions as covariates inside the model. The minimum and maximum predicted marginal probabilities are also given from each model and method.

Covariate(s)	**Method**	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	**Min**	**Max**
Aspirin	GEE	-2.050	0.014			0.246	0.332
	QIF	-4.025	0.039			0.192	0.435
	QIF in GEE Class	-2.365	0.016			0.217	0.311
Hypotensives	GEE	-1.793	0.015			0.226	0.343
	QIF	-2.431	0.025			0.183	0.370
	QIF in GEE Class	-2.158	0.019			0.191	0.331
Lipid-lowering drugs	GEE	-1.736	0.031			0.214	0.453
	QIF	-2.459	0.063			0.172	0.670
	QIF in GEE Class	-2.535	0.063			0.161	0.652
All Three	GEE	-2.845	0.006	0.014	0.028	0.193	0.401
	QIF	-5.884	0.021	0.035	0.052	0.117	0.515
	QIF in GEE Class	-3.746	0.004	0.019	0.057	0.139	0.555
Results After Removing Practice 1							
Lipid-lowering drugs	GEE	-2.409	0.061			0.173	0.518
	QIF	-2.462	0.062			0.170	0.525
	QIF in GEE Class	-2.520	0.063			0.163	0.519
Results After Removing Practice 21							
Aspirin	GEE	-2.247	0.016			0.235	0.334
	QIF	-2.029	0.011			0.219	0.283
	QIF in GEE Class	-1.920	0.010			0.222	0.278

GEE, QIF, and the newly defined QIF were $[1 + 0.058(n_i - 1)]^{-1}$ (0.276), 0.273-0.001(n_i-1) (0.276), and $2.510\hat{\gamma}_{1i} + 2.697(n_i - 1)\hat{\gamma}_{2i}$ (0.257), respectively, when using the logistic link. There is no difference between the probability estimates from GEE and QIF, although the newly defined QIF does give a slightly smaller value. Obtaining parameter estimates this close in value from these three methods appears to coincide with the simulation results presented earlier when the true marginal probability was 0.25. However, when the true marginal probability is closer to zero or the number of clusters is smaller, there is a greater chance in obtaining a sample from which probability estimates can be notably different across these three methods due to a larger variability in weighting used by both QIF versions.

The proportion of patients having a record of treatment with aspirin, hypotensives, or lipid-lowering drugs ranged from sixty-six to ninety-six, thirty-seven to seventy-five, and fourteen to fifty, respectively. In each model, all three methods estimated the marginal probability of adequate assessment to be larger for practices having a greater proportion of patients with a record of drug treatments. Table 3.3 shows the range of estimated marginal probabilities over all practices used in the corresponding model. Results clearly show the difference between GEE and both QIF versions. The range in marginal probability estimates was always smaller for GEE than either QIF version, except when excluding Practice 21 from the analysis. Additionally, of the two QIF versions, parameter estimates from the version with estimating equations in the same class as GEE were notably closer to the estimates from GEE for the first two models. This type of result was especially seen in Scenario 4 of our simulation results, in which the newly defined QIF performed better than the regular QIF, but not as precisely as GEE. This gives some indication that the GEE estimates here may be most reliable. Furthermore, when the proportion of patients with a record of aspirin or lipid-lowering drug treatments were used in the model, QIF estimated the strongest association between these covariates and adequate assessment.

We now take a closer look into the strength of the estimated marginal association between lipid-lowering drugs and adequate assessment. In one practice (Practice 1), only fourteen

percent of patients were adequately assessed, which is much smaller than any of the corresponding marginal probability estimates, given in the third model presented in Table 3.3, from any of the three methods. This practice had the maximum proportion of patients with a record of being treated with lipid-lowering drugs, and therefore had the largest marginal probability estimate. The first plot in Figure 3.1 clearly shows the difference in this practice's observed and estimated probabilities. GEE does not directly take into account how far the observed proportion of adequately assessed patients is from the marginal mean, and therefore the estimated association between adequate assessment and lipid-lowering drugs is not as strong as it would be without using data from this practice. QIF, however, does directly take into account the large empirical variability and downweights this practice's outcomes, allowing the estimated association to be stronger. Explicitly, if we were to estimate these models without the first practice, QIF and GEE would produce $\hat{\boldsymbol{\beta}} = [-2.462, 0.062]$ and $\hat{\boldsymbol{\beta}} = [-2.409, 0.061]$, respectively. These estimates are only slightly different than the estimates given from QIF when including this practice, distinctly showing the robustness property of QIF. Although QIF may seem advantageous in this situation, we do not know for sure if this practice truly is an outlier. If practices such as this one that appear to be outliers truly do occur throughout the entire population of practices, then GEE may have given a better estimate of the marginal trend.

In the model in which the proportion of patients having a record of treatment with aspirin is the only covariate, the difference in estimates from GEE and QIF were not due to an outlier. Rather, the sensitivity, with respect to empirical covariances, of the estimated weights implemented inside QIF's estimating equations led to the notable differences. For instance, the largest practice (Practice 21) had a notable influence on the estimates produced by QIF, but not GEE. Table 3.3 presents the differences in parameter estimates before and after removing this practice. The second plot in Figure 3.1 shows that the overall empirical variation increases with practice size, excluding the largest practice, in our sample. This one practice actually brings down the averaged covariance trend with respect to size, estimated

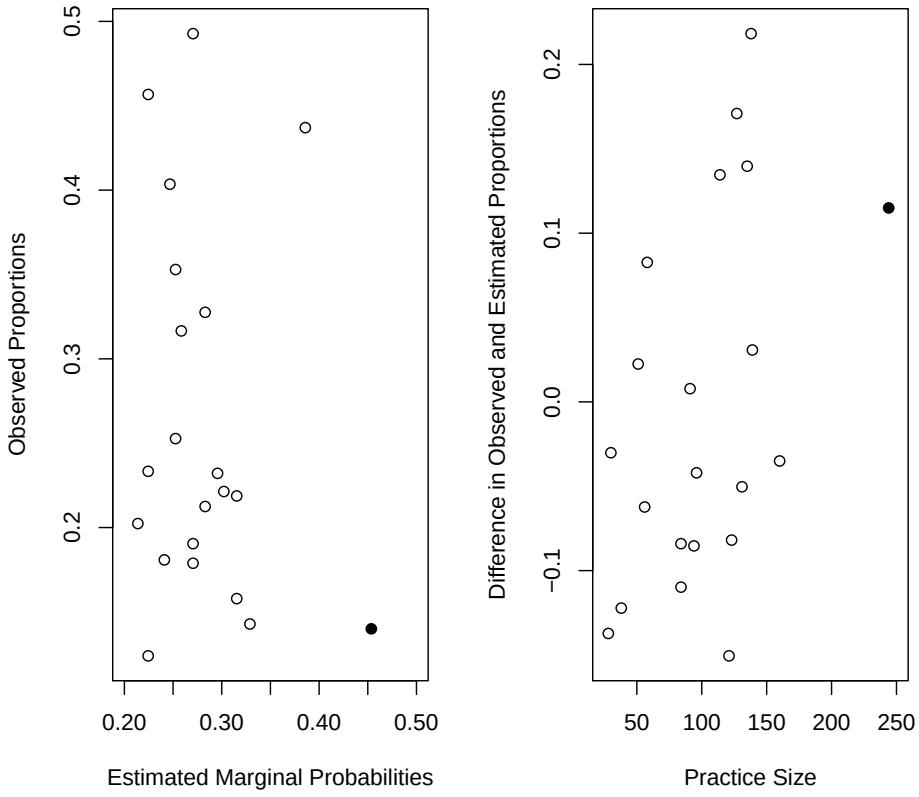

Figure 3.1: Estimated marginal probabilities in the left plot are from using GEE to estimate the model in which the proportion of patients having a record of treatment with lipid-lowering drugs is the only covariate. The bold dot corresponds to Practice 1. In the right plot, estimated marginal probabilities used to obtain differences are from using GEE and the model in which the proportion having a record of treatment with aspirin is the only covariate. The bold dot corresponds to Practice 21.

via $\nabla \bar{g}_N^T(\hat{\boldsymbol{\beta}}) C_N^{-1}(\hat{\boldsymbol{\beta}})$, as it is notably bigger than the other large practices, but its empirical variation is only moderate relative to these same practices. Therefore, when we remove Practice 21, larger practices receive less weight, while the weight given to smaller clusters increases, due to a rise in the averaged empirical covariances in larger practices.

Specifically, Equation (3.7) reduces to

$$\sum_{i=1}^{21} \begin{bmatrix} 0.004 + 0.109x_i + 0.243(n_i - 1) - 0.004(n_i - 1)x_i \\ 0.245 + 7.917x_i + 15.030(n_i - 1) - 0.233(n_i - 1)x_i \end{bmatrix} (Y_i - n_i\pi_i)$$

when using all practices, and becomes

$$\sum_{i=1}^{20} \begin{bmatrix} 0.005 + 0.257x_i + 0.183(n_i - 1) - 0.004(n_i - 1)x_i \\ 0.354 + 21.030x_i + 10.274(n_i - 1) - 0.279(n_i - 1)x_i \end{bmatrix} (Y_i - n_i\pi_i)$$

when deleting Practice 21. By plugging in values for x_i and size from the dataset, it can be seen that the proportion of weight given to smaller (larger) clusters typically increased (decreased) after removing this practice. For instance, Practice 1 (20) consisted of twenty-eight (160) CHD patients, and seventy-nine (sixty-seven) percent of patients had a record of aspirin treatment. The weight matrix given to the residual from the first practice increased from $[7.26, 534.51]^T$ to $[16.53, 1344.01]^T$ after removing Practice 21, while the weight matrix corresponding to Practice 20 decreased from $[6.53, 438.31]^T$ to $[2.65, 70.74]^T$. In comparison, GEE is given as

$$\sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \frac{Y_i - n_i\pi_i}{1 + (n_i - 1)\hat{\rho}}$$

for this example, in which $\hat{\rho}$ is estimated using the empirical variabilities from all practices, and the influence via $\hat{\rho}$ from the empirical variability of one cluster on the weights used by GEE is very minor. Before (after) removing Practice 21, ρ was estimated to be 0.055 (0.057),

giving weights that were only negligibly affected.

The GEE estimates do not differ by a large amount when including or excluding Practice 21, and so it appears that GEE produced more reliable estimates. The estimates produced by QIF were largely influenced by the empirical variability in larger clusters, especially Practice 21, showing that this method can be sensitive in settings consisting of a small number of clusters due to the variability in $C_N(\hat{\boldsymbol{\beta}})$. We note that if more practices were included in this study, the average of the empirical covariances would lessen the influence from a single cluster on weights used inside QIF's estimating equations, making them more reliable. Additionally, although we do not know the true covariances for outcomes in this dataset, implying we cannot say for sure that GEE produced more appropriate estimates than QIF, we do see here that QIF can be sensitive even to a single cluster's empirical covariance, which is the type of scenario in which QIF can be less reliable than GEE.

When using all three covariates in the same model, there were notable differences in parameter estimates across the three methods. Both QIF versions were similar in terms of their predicted ranges of marginal probabilities, which were approximately twice as wide as the range given from GEE. As with the two previously discussed models, the influences from Practices 1 and 21 were the major factors in the differences between GEE and both QIF versions. Taking into account how these practices influenced QIF in these two models, in addition to the simulation results presented in Scenario 5, it is likely that the GEE estimates are more reliable here. These estimates show a notable association between adequate assessment and the three covariates, but not nearly as strong of a relationship estimated by either QIF version due to the influence of only two practices.

3.5 Concluding Remarks

QIF has the theoretical advantage of producing regression estimates with equal or greater efficiency than GEE (Qu *et al.*, 2000). We have given details and evidence that the class of estimating equations realistically has less to do with differences in estimation precision

between QIF and GEE than the empirical nature of $C_N(\hat{\boldsymbol{\beta}})$ and how this matrix is used to weight outcomes inside QIF's estimating equations. In small to moderately sized samples, as is common with GRTs, QIF can produce estimates with lower precision than GEE, even when the incorrect covariance structure is implemented. We also showed via our motivating dataset that the weights QIF implements in its estimating equations can be sensitive to the empirical variability of even a single cluster, as the number of practices was small.

This chapter focused on an exchangeable structure, as unstructured requires a balanced design and QIF and GEE are equivalent estimation procedures under an independence assumption. Although less common when clusters vary in size, especially in GRT scenarios, an AR-1 structure would be appropriate, for example, in a setting resembling administrative censoring in which patients contribute data at the same distinct time points which are equally spaced, but with the allowance that they can drop out of the study any time before their final scheduled visit. The inverse of an AR-1 structure is the weighted sum of three basis matrices, the last of which is usually not implemented with QIF as it contains little information (Qu *et al.*, 2000). When using all three, however, the estimating equations from QIF are in the same class as GEE whether clusters vary in size or not, as $\gamma_{ri}, r = 1, 2, 3; i = 1, 2, \ldots N$, are not functions of size (Qu *et al.*, 2000). Whether using this version or only the first two basis matrices in the extended score equations should make little difference, however. Just as we showed in this chapter that the class of estimating equations has little effect on the estimation precision differences, the empirical nature of $C_N(\hat{\boldsymbol{\beta}})$ can still cause a loss in QIF's estimation reliability as compared with GEE when using a working AR-1 structure. This result will later become evident in the simulation results of Chapter IV.

As is evident from this chapter, research was required to improve the estimation performance of QIF in small to moderately sized samples, and is the focus of the next chapter. However, if estimation precision is not a concern when deciding between QIF and GEE for data analysis, QIF has distinct advantages. For example, the QIF can itself be used as a statistic in goodness-of-fit and likelihood ratio score tests (Qu *et al.*, 2000; Song *et al.*,

2009). Although we do not suggest a numerical value, as we have argued that the amount of empirical information used to estimate weights given to any cluster's outcomes depends on the setting, QIF may be a better method to employ when N is arbitrarily large. This is particularly true when the actual covariance structure is believed to possibly deviate largely from the chosen working structure. In this situation, $C_N(\hat{\beta})$ may have greater accuracy in modeling the entire true covariance structure on average, potentially leading to an improved estimation performance, in addition to the ability to make use of QIF's other advantages. In the next chapter, we propose an improvement to QIF that actually allows us to avoid having to decide if GEE or QIF will perform best in any specified scenario.

CHAPTER IV

Improved Quadratic Inference Functions for Parameter Estimation in the Analysis of Correlated Data

4.1 Introduction

We now give focus towards QIF's estimation performance in general correlated data settings in which the number of clusters is not large, the working correlation structure is not necessarily exchangeable, and clusters may or may not vary in size. Particularly, we show that QIF's estimation performance can be inferior to that of GEE's in these types of general settings, we propose multiple alternative QIF versions to improve estimation, and we suggest an improved QIF version which can be used in place of the regular QIF or GEE. An example we use in this chapter is an AIDS study in which 283 men were followed over time, each providing 1 to 14 observations. Outcomes from the same subject are assumed to be associated, although their true correlation structure is unknown, while a marginal model is fit to describe the mean time trend and the influence of baseline covariates.

This chapter develops improvements to the weighting matrix employed by QIF that are meant to eliminate potential small-sample estimation deficiencies as compared with GEE, while typically maintaining QIF's large-sample advantages. Particularly, we propose utilizing a weighted combination of the empirical covariance matrix and other matrices that are less variable in small samples, in which the corresponding weights minimize the expected quadratic losses of the resulting matrices. The proposed weighting matrices for QIF are developed in Section 4.2. Section 4.3 demonstrates QIF's potential for inferior estimation

performances as compared with GEE in general small-sample settings, and examines the utility of the multiple proposed alternative weighting matrices in these simulations. Additionally, the performances of these methods are contrasted in application to the AIDS study. Concluding remarks are then given in Section 4.4, while the Appendix presents proofs justifying the use of the proposed weighting matrices.

4.2 Improved Weighting Matrices

4.2.1 Using a Model-Based Covariance Matrix

In small to moderately sized samples, GEE can be a better estimation procedure than QIF due to employing model-based covariance structures to weight outcomes. Alternatively, QIF uses $C_N(\boldsymbol{\beta})$ to obtain the weights given to outcomes in Equation (3.6), which can cause these estimating equations to be quite variable when there is not a large number of clusters. Therefore, it makes sense that when $C_N(\boldsymbol{\beta})$ is quite variable, a possibly better estimate for $\boldsymbol{\Sigma}_N$ would be the corresponding model-based version, $\boldsymbol{M}_N = (1/N)\sum_{i=1}^{N} \widehat{Cov}[g_i(\boldsymbol{\beta})]$, which uses the working covariance structures. Particularly,

$$\begin{aligned}\widehat{Cov}[g_i(\boldsymbol{\beta})] &= \hat{E}[g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta})] = \hat{E}[\boldsymbol{B}_i\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{B}_i^T] = \boldsymbol{B}_i\boldsymbol{A}_i^{-1/2}\boldsymbol{V}_i\boldsymbol{A}_i^{-1/2}\boldsymbol{B}_i^T \\ &= \boldsymbol{B}_i\boldsymbol{A}_i^{-1/2}[\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}]\boldsymbol{A}_i^{-1/2}\boldsymbol{B}_i^T = \boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T,\end{aligned}$$

and

$$\boldsymbol{B}_i = \begin{bmatrix} \boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_{1i} \\ \vdots \\ \boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_{mi} \end{bmatrix}$$

and $\boldsymbol{e}_i = \boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$ are defined by Han and Song (2011).

Use of $\boldsymbol{M}_N$ as the weighting matrix in Equation (3.6) can also be problematic, however.

QIF has greater efficiency than GEE when the covariance structure is misspecified and both sets of estimating equations are within the same class, which is why $C_N(\boldsymbol{\beta})$ can be very helpful. If the model-based covariance for the extended score equations, $\boldsymbol{M}_N$, is misspecified and always implemented, then QIF no longer has this advantage. Additionally, even when the number of clusters is not large, if the true covariance structure is misspecified, in some settings the corresponding bias in $\boldsymbol{M}_N$ can be more detrimental than the variability in $C_N(\boldsymbol{\beta})$ with respect to parameter estimation. We therefore propose implementing a weighting matrix, C_N^*, that optimally takes into account both the variability in $C_N(\boldsymbol{\beta})$ and the bias in $\boldsymbol{M}_N$ in order to determine the best weighting matrix to utilize for parameter estimation. Particularly, C_N^* should be as close to the optimal covariance matrix, $\boldsymbol{\Sigma}_N$, as possible. In order for C_N^* to (i) take into account the bias in $\boldsymbol{M}_N$ and variability in $C_N(\boldsymbol{\beta})$, (ii) be as close in value, on average, to $\boldsymbol{\Sigma}_N$ as possible, and (iii) maintain the theoretical advantages QIF has over GEE, we propose improving QIF's estimation performance by employing

$$C_N^* = \rho_N \boldsymbol{M}_N + (1 - \rho_N) C_N(\boldsymbol{\beta}) \tag{4.1}$$

in place of $C_N(\boldsymbol{\beta})$ as the weighting matrix in Equation (3.6). Here, $\rho_N = \tau_N^2/(\alpha_N^2 + \tau_N^2) = \tau_N^2/\delta_N^2$, $\alpha_N^2 = ||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2$, $\tau_N^2 = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2]$, and $\delta_N^2 = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2]$. Here, $||\boldsymbol{K}|| = \sqrt{tr(\boldsymbol{K}\boldsymbol{K}^T)/p}$ for some arbitrary $p \times p$ matrix $\boldsymbol{K}$ (Ledoit and Wolf, 2004), and this value for ρ_N minimizes the expected quadratic loss of $||C_N^* - \boldsymbol{\Sigma}_N||$, or $E[||C_N^* - \boldsymbol{\Sigma}_N||^2]$. Here, τ_N^2 and α_N^2 take into account the variability in $C_N(\boldsymbol{\beta})$ and bias in $\boldsymbol{M}_N$, respectively. Additionally, for several conditions that are typically met in practice, $E[||C_N^* - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $C_N^* - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$ (Ledoit and Wolf, 2004; Han and Song, 2011; conditions and proof in Appendix). This result corresponds to Theorem 1 given in Han and Song (2011).

The proposed weighting matrix is related to the works of Ledoit and Wolf (2004) and Han and Song (2011). Ledoit and Wolf (2004) proposed a well-conditioned estimated covariance

matrix that is a weighted combination of the identity and sample covariance matrices. Han and Song (2011) extended this idea for use with QIF, proposing the use of $S_N = \rho_N \mu_N I + (1 - \rho_N) C_N(\boldsymbol{\beta})$, in which I is the identity matrix, μ_N is the average value for the diagonal elements of $\boldsymbol{\Sigma}_N$, and ρ_N minimizes $E[||S_N - \boldsymbol{\Sigma}_N||^2]$. They propose this alternate weighting matrix, which is referred to as the linear shrinkage estimator, as $C_N(\boldsymbol{\beta})$ may not be invertible in some study designs. However, although the use of S_N can lead to more stable results due to fixing this particular problem, it is not designed to improve QIF's estimation performance in general settings. Particularly, $\mu_N I$ is not meant to model $\boldsymbol{\Sigma}_N$, whereas this is the sole purpose of $\boldsymbol{M}_N$.

In practice, $\boldsymbol{M}_N$ and ρ_N need estimation. Similar to Ledoit and Wolf (2004) and Han and Song (2011), we propose the following:

- The estimator for $\boldsymbol{M}_N$ is $\hat{\boldsymbol{M}}_N$, in which covariance parameters need estimation
- The estimator for δ_N^2 is $d_N^2 = ||C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N||^2$
- The estimator for τ_N^2 is $t_N^2 = min[\bar{t}_N^2, d_N^2]$, $\bar{t}_N^2 = \frac{1}{N^2} \sum_{i=1}^N ||g_i(\boldsymbol{\beta}) g_i(\boldsymbol{\beta})^T - C_N(\boldsymbol{\beta})||^2$
- The estimator for α_N^2 is $a_N^2 = d_N^2 - t_N^2$
- The estimator for C_N^* is $\hat{C}_N^* = \frac{t_N^2}{d_N^2} \hat{\boldsymbol{M}}_N + \frac{a_N^2}{d_N^2} C_N(\boldsymbol{\beta}) = \hat{\rho}_N \hat{\boldsymbol{M}}_N + (1 - \hat{\rho}_N) C_N(\boldsymbol{\beta})$

The use of $\bar{t}_N^2$ is appropriate in the settings of Han and Song (2011), as they deal with balanced covariate designs. However, in many general applications, the covariances of the N extended score components will likely vary, inducing bias in $\bar{t}_N^2$. Particularly, $Bias(\bar{t}_N^2) \approx (1/N^2) \sum_{i=1}^N ||Cov[g_i(\boldsymbol{\beta})]||^2 - (1/N^3) || \sum_{i=1}^N Cov[g_i(\boldsymbol{\beta})]||^2$ (see Appendix). Bias can be estimated using the model-based covariances, $\boldsymbol{B}_i \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{B}_i^T$, $i = 1, 2, \ldots, N$, giving an alternative estimate, $\hat{t}_N^2 = max\left(0, min[\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2), d_N^2]\right)$, for τ_N^2. Results using $\hat{t}_N^2$ and t_N^2 are asymptotically equivalent.

Justifications for these estimates are given in the Appendix, and are based on the Lemma given by Han and Song (2011) with its corresponding proofs, which also use work from Ledoit

and Wolf (2004). Specifically, $a_N^2 - \alpha_N^2$, $t_N^2 - \tau_N^2$, and $d_N^2 - \delta_N^2$ all converge in quadratic mean to zero as $N \to \infty$, under the assumption that $E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4] \to 0$ as $N \to \infty$. Furthermore, corresponding to the second theorem and its proof given in Han and Song (2011), $E[||\hat{C}_N^* - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $\hat{C}_N^*$ is asymptotically optimal since $\hat{C}_N^* - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$ (Ledoit and Wolf, 2004; Han and Song, 2011; conditions and proof in Appendix).

4.2.2 An Alternative Empirical Covariance Matrix

The previous chapter explains that cluster size imbalance can be detrimental to QIF's small-sample estimation performance via $C_N(\boldsymbol{\beta})$. In this section, we therefore propose an alternate weighting matrix, $\tilde{C}_N(\boldsymbol{\beta})$, that averages out this detrimental effect due to variation in cluster sizes when implementing a working exchangeable correlation structure. Our hopes with this matrix is that it will be more stable than $C_N(\boldsymbol{\beta})$ and less biased than $\boldsymbol{M}_N$ when the working covariance structure is misspecified. $\tilde{C}_N(\boldsymbol{\beta})$ removes most of the influence cluster size imbalance has on estimating weights given to observations within Equation (3.6), increasing the amount of information used to estimate these weights, which will in turn have smaller variances. Using an exchangeable correlation structure, $m = 2$ and the empirical covariance for the ith cluster is $g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta}) =$

$$\begin{bmatrix} g_{1i}(\boldsymbol{\beta})g_{1i}^T(\boldsymbol{\beta}) & g_{1i}(\boldsymbol{\beta})g_{2i}^T(\boldsymbol{\beta}) \\ g_{2i}(\boldsymbol{\beta})g_{1i}^T(\boldsymbol{\beta}) & g_{2i}(\boldsymbol{\beta})g_{2i}^T(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} A_{11i} & A_{12i} \\ A_{21i} & A_{22i} \end{bmatrix}.$$

Denote the element in row u and column v of A_{jki} as $A_{jki}(u, v)$, $j, k = 1, 2$. The elements of this matrix depend on cluster size via sums in the following fashion:

$$A_{11i}(u, v) = \left[\sum_{l=1}^{n_i} \frac{\partial \mu_{il}}{\partial \beta_{v-1}} \cdot \frac{r_{il}}{\sigma_{il}}\right] \left[\sum_{l=1}^{n_i} \frac{\partial \mu_{il}}{\partial \beta_{u-1}} \cdot \frac{r_{il}}{\sigma_{il}}\right] \tag{4.2}$$

$$A_{12i}(u,v) = A_{21i}(v,u) = \left[\sum_{l=1}^{n_i} \frac{\partial \mu_{il}}{\partial \beta_{v-1}} \cdot \frac{r_{il}}{\sigma_{il}}\right] \left[\sum_{h=1}^{n_i} \frac{r_{ih}}{\sqrt{\sigma_{ih}}} \left(\sum_{l \neq h}^{n_i} \frac{\partial \mu_{il}}{\partial \beta_{u-1}} \sigma_{il}^{-1/2}\right)\right] \tag{4.3}$$

$$A_{22i}(u,v) = \left[\sum_{h=1}^{n_i} \frac{r_{ih}}{\sqrt{\sigma_{ih}}} \left(\sum_{l \neq h}^{n_i} \frac{\partial \mu_{il}}{\partial \beta_{v-1}} \sigma_{il}^{-1/2}\right)\right] \left[\sum_{h=1}^{n_i} \frac{r_{ih}}{\sqrt{\sigma_{ih}}} \left(\sum_{l \neq h}^{n_i} \frac{\partial \mu_{il}}{\partial \beta_{u-1}} \sigma_{il}^{-1/2}\right)\right] \tag{4.4}$$

Here, $r_{il} = (Y_{il} - \mu_{il})$ and σ_{il} is the working variance for Y_{il}.

To decrease the effect from cluster size imbalance, the values for Equations (4.2) - (4.4) could be weighted such that their magnitudes have a diminished dependency on cluster size. We propose dividing each summation by its respective number of terms. When covariances and marginal means within any given cluster do not rely upon its size, the magnitudes of these quantities in Equations (4.2) - (4.4) will depend much less on cluster size. There may still be some impact, though, since Equations (4.2) - (4.4) are comprised of more covariance components, $r_{il}r_{ih}$, $l \neq h$, than variance components, r_{il}^2, $l = 1, 2, \ldots n_i$; $i = 1, 2, \ldots N$. If covariances and variances differ in magnitude, dividing each summation by its respective number of terms cannot take this dissimilarity into account, and therefore cluster size may still have a small effect.

Define

$$\tilde{C}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=l}^{N} \tilde{C}_i(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} n_i^2 \boldsymbol{a} & n_i^2(n_i-1)\boldsymbol{b} \\ n_i^2(n_i-1)\boldsymbol{b}^T & n_i^2(n_i-1)^2\boldsymbol{c} \end{bmatrix},$$

$\boldsymbol{a} = (1/N)\sum_{i=1}^{N} A_{11i}/n_i^2$, $\boldsymbol{b} = (1/N)\sum_{i=1}^{N} A_{12i}/[n_i^2(n_i-1)]$, and $\boldsymbol{c} = (1/N)\sum_{i=1}^{N} A_{22i}/[n_i^2(n_i-1)^2]$. This can be viewed as estimating $Cov[g_i(\boldsymbol{\beta})]$ via $\tilde{C}_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots N$, by multiplying the average of the four weighted empirical covariance component matrices by the functions of cluster size that these matrices originally were divided by for obtaining the weighted average. When all clusters are equal in size, $\tilde{C}_N(\boldsymbol{\beta}) = C_N(\boldsymbol{\beta})$. If $n_i = 1$, we set the ith element in $\boldsymbol{b}$ and $\boldsymbol{c}$ equal to 0.

Two potential problems are apparent with using $\tilde{C}_N(\boldsymbol{\beta})$ as the weighting matrix. First, the effect covariates have on the estimation performance of QIF is not taken into account, and second, $\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N$ does not converge in probability to 0. Rather, $\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N \xrightarrow{p} 0$, where $\tilde{\boldsymbol{\Sigma}}_N = E[\tilde{C}_N(\boldsymbol{\beta})]$. Due to this result, the estimating equations in Equation (3.6), now using $\tilde{C}_N(\boldsymbol{\beta})$ in place of $C_N(\boldsymbol{\beta})$, are theoretically no longer optimal within their respective class. Our hope is that the corresponding equations will improve estimation performance both by approximating the optimal equations well and decreasing the variability in Equation (3.6).

In order to maintain QIF's large-sample advantages, we suggest using a weighted combination of $\tilde{C}_N(\boldsymbol{\beta})$ and $C_N(\boldsymbol{\beta})$, similar to the weighted version involving $\boldsymbol{M}_N$. We explore two different versions, both based on minimizing the expected quadratic loss of the proposed weighting matrix. In the Appendix, we prove the same asymptotic results as we did for C_N^* and $\hat{C}_N^*$, using similar arguments based on work from Han and Song (2011) and Ledoit and Wolf (2004). Specifically, we prove the Lemma and both Theorems justifying the use of the proposed weight estimates and that the proposed weighting matrices are asymptotically optimal. The first version is simplified in that it does not take into account the covariance term between $\tilde{C}_N(\boldsymbol{\beta})$ and $C_N(\boldsymbol{\beta})$ when minimizing the expected quadratic loss, while the second version does.

4.2.2.1 Not Using the Covariance Term

In order to use a combination of $\tilde{C}_N(\boldsymbol{\beta})$ and $C_N(\boldsymbol{\beta})$ as the weighting matrix, while ignoring the covariance term between these two matrices, we propose the following unestimated matrix:

$$\boldsymbol{S}_N^1 = \rho_N \tilde{\boldsymbol{\Sigma}}_N + (1 - \rho_N) C_N(\boldsymbol{\beta}) \tag{4.5}$$

Here, $\rho_N = \tau_N^2/(\alpha_N^2+\tau_N^2) = \tau_N^2/\delta_N^2$, $\alpha_N^2 = ||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2$, $\tau_N^2 = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2]$, and $\delta_N^2 = E[||C_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^2]$. This value for ρ_N minimizes the expected quadratic loss of $||\boldsymbol{S}_N^1 - \boldsymbol{\Sigma}_N||$. τ_N^2 and α_N^2 take into account the variability in $C_N(\boldsymbol{\beta})$ and bias in $\tilde{\boldsymbol{\Sigma}}_N$, respectively.

In practice, $\tilde{\boldsymbol{\Sigma}}_N$ and ρ_N need estimation. Similar to Ledoit and Wolf (2004) and Han and Song (2011), we propose the following:

- The estimator for $\tilde{\boldsymbol{\Sigma}}_N$ is $\tilde{C}_N(\boldsymbol{\beta})$
- The estimator for δ_N^2 is $d_N^2 = ||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2$
- The estimator for τ_N^2 is $t_N^2 = min[\bar{t}_N^2, d_N^2]$ or $\hat{t}_N^2 = min[\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2), d_N^2]$
- The estimator for α_N^2 is $a_N^2 = d_N^2 - t_N^2$
- The estimator for $\boldsymbol{S}_N^1$ is $\hat{\boldsymbol{S}}_N^1 = \frac{t_N^2}{d_N^2}\tilde{C}_N(\boldsymbol{\beta}) + \frac{a_N^2}{d_N^2}C_N(\boldsymbol{\beta}) = \hat{\rho}_N\tilde{C}_N(\boldsymbol{\beta}) + (1-\hat{\rho}_N)C_N(\boldsymbol{\beta})$

4.2.2.2 Using the Covariance Term

We now propose

$$\boldsymbol{S}_N^2 = \rho_N\tilde{C}_N(\boldsymbol{\beta}) + (1-\rho_N)C_N(\boldsymbol{\beta}), \tag{4.6}$$

which utilizes the covariance term between $\tilde{C}_N(\boldsymbol{\beta})$ and $C_N(\boldsymbol{\beta})$. Here, $\rho_N = \gamma_N/\delta_N^2$, $(1-\rho_N) = \lambda_N/\delta_N^2$, $\delta_N^2 = E[||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2] = \alpha_N^2 + \tau_N^2 + 2\theta_N$, $\alpha_N^2 = E[||\boldsymbol{\Sigma}_N - \tilde{C}_N(\boldsymbol{\beta})||^2]$, $\tau_N^2 = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2]$, $\theta_N = E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \tilde{C}_N(\boldsymbol{\beta}) >]$,

$$\gamma_N = \gamma_N(\tau_N^2, \theta_N, \delta_N^2) = \begin{cases} 0 & if \quad \tau_N^2 + \theta_N < 0 \\ \tau_N^2 + \theta_N & if \quad 0 \le \tau_N^2 + \theta_N \le \delta_N^2 \\ \delta_N^2 & if \quad \tau_N^2 + \theta_N > \delta_N^2 \end{cases}$$

and

$$\lambda_N = \lambda_N(\alpha_N^2, \theta_N, \delta_N^2) = \begin{cases} 0 & if \quad \alpha_N^2 + \theta_N < 0 \\ \alpha_N^2 + \theta_N & if \quad 0 \leq \alpha_N^2 + \theta_N \leq \delta_N^2 \\ \delta_N^2 & if \quad \alpha_N^2 + \theta_N > \delta_N^2 \end{cases}$$

Here, $< \boldsymbol{K}_1, \boldsymbol{K}_2 >= tr(\boldsymbol{K}_1 \boldsymbol{K}_2^T)/p$ for some arbitrary $p \times p$ matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ (Ledoit and Wolf, 2004), and this value for ρ_N minimizes the expected quadratic loss of $||\boldsymbol{S}_N^2 - \boldsymbol{\Sigma}_N||$, while maintaining the constraint $0 \leq \rho_N \leq 1$. Theoretically, this constraint is not necessarily satisfied if using $\rho_N = (\tau_N^2 + \theta_N)/\delta_N^2$, inducing the need for γ_N and $\lambda_N = \delta_N^2 - \gamma_N$.

In practice, ρ_N needs estimation. Similar to Ledoit and Wolf (2004) and Han and Song (2011), we propose the following:

- The estimator for δ_N^2 is $d_N^2 = ||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2$
- The estimator for τ_N^2 is $\bar{t}_N^2 = \frac{1}{N^2} \sum_{i=1}^N ||g_i(\boldsymbol{\beta}) g_i(\boldsymbol{\beta})^T - C_N(\boldsymbol{\beta})||^2$
- The estimator for θ_N is $\hat{\theta}_N = -0.5[\frac{1}{N^2} \sum_{i=1}^N ||\tilde{C}_i(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2 + \bar{t}_N^2]$
- The estimator for α_N^2 is $a_N^2 = \frac{1}{N^2} \sum_{i=1}^N ||\tilde{C}_i(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2 + d_N^2$
- The estimator for γ_N is $\hat{\gamma}_N = \gamma_N(\bar{t}_N^2, \hat{\theta}_N, d_N^2)$
- The estimator for λ_N is $\hat{\lambda}_N = \lambda_N(a_N^2, \hat{\theta}_N, d_N^2)$
- The estimator for $\boldsymbol{S}_N^2$ is $\hat{\boldsymbol{S}}_N^2 = \frac{\hat{\gamma}_N}{d_N^2} \tilde{C}_N(\boldsymbol{\beta}) + \frac{\hat{\lambda}_N}{d_N^2} C_N(\boldsymbol{\beta}) = \hat{\rho}_N \tilde{C}_N(\boldsymbol{\beta}) + (1 - \hat{\rho}_N) C_N(\boldsymbol{\beta})$

4.2.3 The Advantages of C_N^*

Use of C_N^*, rather than $\tilde{C}_N(\boldsymbol{\beta})$, $\boldsymbol{S}_N^1$, or $\boldsymbol{S}_N^2$, has multiple advantages. Use of $\boldsymbol{M}_N$ is applicable for any working correlation structure, while $\tilde{C}_N(\boldsymbol{\beta})$ is only designed for the exchangeable structure. Additionally, $\boldsymbol{M}_N$ not only combats the impact of cluster size imbalance, although in a different manner than $\tilde{C}_N(\boldsymbol{\beta})$, it also protects the weighting matrix from the effect of covariates on $C_N(\boldsymbol{\beta})$ as well. Related to this, $\tilde{C}_N(\boldsymbol{\beta})$, $\boldsymbol{S}_N^1$, and $\boldsymbol{S}_N^2$ are equivalent

to using $C_N(\boldsymbol{\beta})$ as the weighting matrix when all clusters are constant in size, possibly still leading to inferior estimation performance, while use of $\boldsymbol{M}_N$ in C_N^* does not revert back to using $C_N(\boldsymbol{\beta})$ in finite samples.

4.3 Assessing the Utility of the Proposed Weighting Matrices

4.3.1 Via Simulation Study

To assess the estimation performances of GEE, QIF, and QIF with the different proposed weighting matrices, we use empirical MSE quantities that are the sum of the empirical MSEs from all non-intercept parameters. Use of $\hat{C}_N^*$, $\tilde{C}_N(\boldsymbol{\beta})$, $\hat{\boldsymbol{S}}_N^1$, and $\hat{\boldsymbol{S}}_N^2$ will be referred to as QIF2, QIF3, QIF4, and QIF5, respectively. As QIF2 and QIF4 can use either t_N^2 or $\hat{t}_N^2$, we will denote these methods with a when using t_N^2 and b when implementing $\hat{t}_N^2$. Tables report MSE ratios, which take the empirical MSE quantity from GEE and divides it by the MSE value from the corresponding method, and/or the empirical mean of the estimated weights given to $\hat{\boldsymbol{M}}_N$ or $\tilde{C}_N(\boldsymbol{\beta})$. Table 4.1 presents results from general repeated measures scenarios in which clusters are constant in size, implying QIF, QIF3, QIF4, and QIF5 are equivalent. Table 4.2 presents results from general GRTs and repeated measures scenarios including settings mimicking the AIDS study. Table 4.1 (4.2 and 4.3) presents results from three (five) different scenarios, each comprised of two or four different settings. Each setting was examined via 1,000 simulations. Correlated binary data were generated using the method presented by Qaqish (2003), except for in GRT scenarios in which the beta-binomial distribution was utilized. Correlation and variance (normally distributed data) parameter estimates used in model-based covariances for QIF2 and QIF4b were obtained from GEE to reduce simulation time, although estimates could be found iteratively as is done with GEE. Additionally, $\hat{\boldsymbol{\beta}}$ was used in place of $\boldsymbol{\beta}$ inside the empirical covariance matrices, $\tilde{C}_N(\boldsymbol{\beta})$ and $C_N(\boldsymbol{\beta})$.

4.3.1.1 Description of Simulation Settings and Presentation of Results

In Scenario 1, the marginal model is given by

$$Y_{ij} = \beta_0 + \beta_1 z_{1ij} + \beta_2 z_{2ij} + \epsilon_{ij}; \epsilon_{ij} \sim N(0, j/5); j = 1, \ldots, 10.$$

The number of clusters was 25 (200) for the first (second) and third (fourth) settings. The true correlation structure was AR-1 (exchangeable) for the first (last) two settings, while the working correlation structure was always AR-1. The correlation parameter was 0.7 for each setting, while $\boldsymbol{\beta} = [0, 0, 1]^T$. Individual-level covariates were generated independently within and across clusters from $N(j/10, 1)$, similar to a design presented by Qu, Lindsay, and Li (2000).

In Scenario 2, the marginal model is given by

$$Y_{ij} = \beta_0 + \beta_1 z_{1ij} + \beta_2 z_{2i} + \epsilon_{ij}; \epsilon_{ij} \sim N(0, 1 + 10 z_{2i}); j = 1, 2, 3, 4.$$

The number of clusters was 25 (200) for the first (second) and third (fourth) settings. The true correlation structure was AR-1 (exchangeable) for the first (last) two settings, while the working correlation structure was always AR-1. The correlation parameter was 0.7 for each setting, while $\boldsymbol{\beta} = [1, 0, 1]^T$. z_{2i} was generated independently across clusters from $Uniform(0, 1)$. Of the four equally spaced time points, two were randomly and uniformly chosen to be the times at which the indicator covariate, z_{1ij}, was given a value of 1, and thus the remaining two were given a value of 0.

In Scenario 3, the marginal model is given by

$$logit(\pi_{ij}) = \beta_0 + \beta_1 z_{1i} + \beta_1 z_{2ij}; j = 1, 2, 3, 4,$$

where π_{ij} is the marginal probability for the jth response in cluster i, and $logit(\pi_{ij}) = log[\pi_{ij}/(1 - \pi_{ij})]$. The number of clusters in each of two trial arms was 25 (250) for the first

Table 4.1: Empirical means of estimated weights given to $\hat{\boldsymbol{M}}_N$ in QIF2a and QIF2b, and empirical MSE ratios comparing the three QIF versions to GEE. Scenarios are general representations of possible repeated measures studies in which clusters are constant in size, implying QIF3, QIF4, and QIF5 are equivalent to QIF.

		QIF	QIF2a		QIF2b	
Setting	N	MSE Ratio	$\hat{E}(\hat{\rho}_N)$	MSE Ratio	$\hat{E}(\hat{\rho}_N)$	MSE Ratio
(1.1)	25	0.819	0.973	0.960	0.960	0.959
(1.2)	200	0.940	0.776	0.948	0.744	0.948
(1.3)	25	0.876	0.588	0.980	0.564	0.979
(1.4)	200	0.982	0.085	0.984	0.082	0.984
(2.1)	25	0.813	0.996	0.997	0.995	0.998
(2.2)	200	0.967	0.746	0.992	0.727	0.991
(2.3)	25	0.870	0.999	1.000	0.998	1.000
(2.4)	200	1.024	0.384	1.029	0.374	1.030
(3.1)	50	0.864	0.996	0.977	0.980	0.977
(3.2)	500	0.953	0.997	0.965	0.980	0.965
(3.3)	50	0.886	0.947	0.985	0.848	0.986
(3.4)	500	1.011	0.158	1.011	0.088	1.011

(second) and third (fourth) settings. The true correlation structure was AR-1 (exchangeable) for the first (last) two settings, while the working correlation structure was always AR-1. The correlation parameter was 0.7 for each setting, while $\boldsymbol{\beta} = [0, 0, 0.1]^T$. z_{1i} was given a value of 0 or 1, depending upon the arm of the trial to which the ith cluster belonged, while $z_{2ij} \sim Uniform(0, 1)$ was generated independently from all observations within and across clusters. These settings have similarities to those used by Song *et al.* (2009).

In Scenario 4, the marginal model is given by

$$logit(\pi_{ij}) = \beta_0 + \beta_1 z_{1i}; j = 1, \ldots, n_i,$$

representing a general GRT scenario. An equal number of clusters were randomized to

the intervention and control arms of the trial, and z_{1i} was an indicator for intervention assignment. The number of clusters was 20, 200, 40, and 400 for the first through fourth settings, respectively. Clusters varied in size independently and uniformly from 10 to 50, and an exchangeable correlation structure was correctly implemented. In the first two settings, marginal probabilities and correlations were 0.1 and 0.05, respectively, across all clusters. In the last two settings, marginal probabilities (correlations) were 0.5 (0.3) and 0.3 (0.2) for control and intervention clusters, respectively.

Scenario 5 uses the same marginal model as Scenario 4, and is meant to demonstrate settings in which correlation values are impacted by cluster size, similar to simulations done in the previous chapter. Ten (fifty) clusters were randomized to each trial arm in the first (second) setting. Marginal probabilities were 0.5, an equal number of clusters were randomized to each trial arm, and cluster sizes were independently generated from $Uniform(25, 150)$. The exchangeable correlation value for the ith cluster was $exp(\omega_1 + \omega_2 * n_i)/(1 + exp(\omega_1 + \omega_2 * n_i))$, in which ω_1 and ω_2 were -0.05 (-1.5) and -0.025 (-0.02), respectively, in the first (second) setting. This allowed correlations to range from 0.02 to 0.34 in the first setting, and 0.01 to 0.12 in the second setting.

In Scenario 6, the marginal model is given by

$$logit(\pi_{ij}) = logit(\pi_i) = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i}; j = 1, \ldots, n_i,$$

representing a general GRT scenario with multiple covariates, similar to simulations done in the previous chapter. The number of clusters was 20 (40) for the first (second) setting. Cluster sizes were independently generated from $Uniform(25, 150)$, while $z_{3i} \sim Uniform(-1, 1)$. The first two covariates, z_{1i} and z_{2i}, are indicators, and there were $N/4$ clusters in each of their four corresponding combinations. The exchangeable correlation value for the ith cluster was $exp(-2 - 5|\pi_i - 0.5|)/(1 + exp(-2 - 5|\pi_i - 0.5|))$, allowing correlations to range from 0.02 to 0.12, while $\boldsymbol{\beta} = [0, 0, 1, 1]^T$.

Scenario 7 is the same as Scenario 3, except clusters now vary in size and can contribute up to eight observations, rather than four. Specifically, each cluster had two to eight observations, randomly and uniformly selected from all eight possible observation times. In the first two settings, an AR-1 correlation structure was implemented, while the true structure was exchangeable with a parameter of 0.7. The true correlation between any two observations j and k from the ith cluster was $0.7^{|j-k|}$, $j, k = 1, 2, \ldots, 8$, in the last two settings, while an exchangeable structure was utilized.

In Scenario 8, the marginal model is given by

$$Y_{it} = \beta_0 + \beta_1 time_{it} + \beta_2 time_{it}^2 + \beta_3 z_{3i} + \beta_4 z_{4i} + \beta_5 z_{5i} + \epsilon_{it}; \epsilon_{it} \sim N(0, 105);$$

$i = 1, \ldots, 283$; $t = 1, \ldots, n_i$, in which $\boldsymbol{\beta} = [37, -4.5, 0.35, 0.40, 0, 0]^T$. This is meant to mimic the scenario of the application dataset, and cluster sizes, z_{3i}, z_{4i}, and z_{5i} were independently generated from $Uniform[1, 14]$, $N(0, 64)$, $N(0, 64)$, and $Bernoulli(0.35)$, respectively. Time was generated uniformly on $[0.1, 0.2, \ldots, 5.9]$, and any given time point was not allowed to be observed more than once for any subject. The true correlation structure was $Corr(Y_{it}, Y_{ik}) = 0.6^{|time_{it} - time_{ik}|}$ (exchangeable with a common correlation of 0.6) for the first (last) two settings, while the working correlation structure was exchangeable (AR-1) for the first (second) and third (fourth) settings. Each setting presents results from fitting models with (a) only the first three covariates and (b) all five covariates. MSE ratios reported in Settings 8.2b and 8.4b are only for estimates of β_1 and β_2, as QIF's estimation performance is only superior to GEE's for estimating the marginal time trend. These ratios would be closer in value to one if the corresponding MSE quantities were for all non-intercept parameters.

4.3.1.2 Description of Results

Results show that QIF can produce estimates with greater variability than the corresponding estimates from GEE in settings consisting of a small to moderately sized sample,

Table 4.2: Empirical MSE ratios comparing the seven QIF versions to GEE. Scenarios are general representations of GRTs and repeated measures scenarios, including settings mimicking the AIDS study. MSE ratios reported in Settings 8.2b and 8.4b are only for estimates of β_1 and β_2.

		MSE Ratios						
Setting	N	QIF	QIF2a	QIF2b	QIF3	QIF4a	QIF4b	QIF5
(4.1)	20	0.747	0.969	0.922	0.978	0.914	0.879	0.840
(4.2)	200	0.924	0.993	0.991	0.965	0.951	0.948	0.934
(4.3)	40	0.812	0.966	0.919	0.981	0.959	0.936	0.901
(4.4)	400	0.969	0.972	0.971	0.978	0.978	0.977	0.974
(5.1)	20	1.153	1.148	1.190	0.988	1.146	1.192	1.202
(5.2)	100	1.175	1.103	1.145	0.939	1.176	1.175	1.176
(6.1)	20	0.622	0.946	0.843	0.858	0.808	0.762	0.736
(6.2)	40	0.746	0.973	0.927	0.932	0.909	0.884	0.847
(7.1)	50	0.808	0.953	0.931				
(7.2)	200	0.944	0.954	0.951				
(7.3)	50	0.812	0.982	0.962	0.925	0.908	0.893	0.876
(7.4)	200	0.931	0.980	0.977	0.954	0.947	0.942	0.939
(8.1a)	283	0.993	1.003	1.000	0.958	0.964	0.964	0.975
(8.1b)	283	0.936	0.979	0.979	0.878	0.889	0.889	0.917
(8.2a)	283	1.150	1.139	1.142				
(8.2b)	283	1.138	1.143	1.146				
(8.3a)	283	0.874	0.896	0.897	0.779	0.785	0.785	0.809
(8.3b)	283	0.904	0.953	0.951	0.838	0.842	0.842	0.875
(8.4a)	283	1.040	1.049	1.049				
(8.4b)	283	1.041	1.050	1.050				

Table 4.3: Empirical mean estimates for ρ_N in QIF2, QIF4, and QIF5. Scenarios are general representations of GRTs and repeated measures scenarios, including settings mimicking the AIDS study.

		$\hat{E}(\hat{\rho}_N)$				
Setting	N	QIF2a	QIF2b	QIF4a	QIF4b	QIF5
(4.1)	20	0.936	0.765	0.702	0.587	0.410
(4.2)	200	0.973	0.937	0.654	0.600	0.355
(4.3)	40	0.797	0.533	0.888	0.713	0.614
(4.4)	400	0.158	0.085	0.860	0.683	0.547
(5.1)	20	0.628	0.248	0.232	0.111	0.053
(5.2)	100	0.638	0.372	0.066	0.032	0.002
(6.1)	20	0.908	0.674	0.755	0.599	0.411
(6.2)	40	0.963	0.803	0.840	0.715	0.485
(7.1)	50	0.700	0.517			
(7.2)	200	0.161	0.107			
(7.3)	50	0.988	0.895	0.857	0.701	0.599
(7.4)	200	0.998	0.968	0.791	0.614	0.501
(8.1a)	283	0.981	0.917	0.794	0.794	0.406
(8.1b)	283	0.985	0.906	0.830	0.830	0.389
(8.2a)	283	0.957	0.920			
(8.2b)	283	0.973	0.932			
(8.3a)	283	0.983	0.918	0.877	0.877	0.509
(8.3b)	283	0.986	0.901	0.901	0.901	0.484
(8.4a)	283	0.224	0.206			
(8.4b)	283	0.287	0.264			

even when the working covariance structure is misspecified. This result was observed in general GRT and repeated measures scenarios in which working AR-1 and exchangeable correlations were utilized. Also, both QIF2a and QIF2b usually performed at least almost as well as GEE when QIF led to estimates with the largest MSE, and approximately as good as QIF when GEE worked least favorably. However, use of $\tilde{C}_N(\boldsymbol{\beta})$ in some form only worked better than QIF in GRT scenarios and Scenario 7. In Scenario 4, QIF3 was one of the best QIF versions, whereas the opposite was seen in Scenario 5. QIF4 and QIF5 performed better than QIF in Scenarios 4 and 6, and approximately just as well in Scenario 5. However, there were many settings in which QIF2 performed notably better than any of these other QIF methods.

In all but one setting in which the number of independent clusters was fifty or less, the empirical MSE from all non-intercept parameters was notably smaller for GEE than QIF. This was most obvious in the GRT settings of Scenarios 4 and 6. When N was larger, GEE and QIF worked similarly, except in Settings 5.2, 8.2, and 8.3.

QIF2a and QIF2b considerably improved the estimation performance of QIF, particularly in settings consisting of 50 clusters or less. When $N \leq 100$, the only setting in which QIF worked better than QIF2a and QIF2b was Setting 5.2, although the difference in empirical MSEs here is small. Additionally, QIF2a, QIF2b, and GEE all performed similarly in most settings, with the exception of Scenario 5 and Settings 8.2 and 8.3, as previously mentioned. These first two are examples of situations when QIF, QIF2a, and QIF2b perform better than GEE, and the opposite was seen in Setting 8.3. Also, Scenarios 4 and 6 suggest that GEE may work better than QIF2b in GRT settings in which the number of clusters is small, particularly when there are multiple covariates.

Any QIF version utilizing $\tilde{C}_N(\boldsymbol{\beta})$ also performed better than or as well as QIF in the GRT settings, with the exception of QIF3 in Scenario 5, which was also the only GRT scenario in which QIF3 did not perform better than QIF4 and QIF5. However, the overall performance of QIF2 was better than that of QIF3, QIF4, and QIF5 in these settings. Additionally, both

QIF and QIF2 performed best in the settings of Scenario 8.

QIF2a and QIF2b led to parameter estimates with almost equivalent MSEs in the majority of repeated measures scenarios, although some small differences were evident in the GRT scenarios. Estimated weights from these two methods were very similar in Table 4.1 since clusters were constant in size, whereas large variations in size increase the bias in $\bar{t}_N^2$, shown via differences in the mean weight estimates presented in Table 4.3. These differences diminished as N increased, however, since the bias in $\bar{t}_N^2$ decreased. Due to estimated bias, QIF2b outperformed QIF2a in Scenario 5 by using smaller weights on average, while QIF2a worked better than QIF2b in the other GRT settings by using larger weights on average. These same results were also evident when comparing QIF4a and QIF4b, with the exception of diminishing differences in weight estimates as N increased.

In Table 4.1, when the correlation structure was correctly specified to be AR-1 and $N \leq 50$, mean weight estimates were close to one, due to the variability in $C_N(\boldsymbol{\beta})$. However, when $N = 200$ and the marginal variances were incorrectly assumed constant, the average weight given to $C_N(\boldsymbol{\beta})$ increased due to its ability to account for these misspecifications in larger sized samples. Additionally for QIF2, Scenario 3 and Setting 8.3 show that $\hat{\rho}_N$ can be close to one even for moderate N when the entire covariance structure is correct, as $\tau_N^2 = \delta_N^2$. In Table 4.3, when correctly implementing an exchangeable structure, similar results were also particularly evident in Scenario 4 for QIF2. No notable trend was seen with QIF4 or QIF5.

In settings in which AR-1 was incorrectly implemented, more weight was given to $\hat{\boldsymbol{M}}_N$ on average for smaller sample sizes. However, the majority of weight was given to $C_N(\boldsymbol{\beta})$ when $N \geq 200$ since this empirical matrix can more accurately account for the true covariances within the data. An exception would be Setting 8.2, in which the mean estimated weights were quite large. In this setting, it is interesting that QIF and QIF2 performed better than GEE, although further study of these results indicates that $C_N(\boldsymbol{\beta})$ did not necessarily give QIF the advantage here. Rather, two possible explanations for this superior performance

could be that QIF only used two basis matrices to approximate the inverse of the AR-1 structure, and that estimation was accomplished my minimizing the value for $Q_N(\boldsymbol{\beta})$. When incorrectly implementing an exchangeable structure, Setting 8.1 and the third and fourth settings of Scenario 7 give evidence that this working correlation can still possibly lead to large values for $\hat{\rho}_N$ when N is moderate and QIF2 or QIF4 are used. Even so, this was not detrimental to parameter estimation.

To give examples of what individual values looked like, Figures 4.1 - 4.4 present histograms of $\hat{\rho}_N$ from Settings (1.3) and (5.1). Figure 4.1 shows that the majority of values for $\hat{\rho}_N$ were between 0.3 and 0.7 in Setting (1.3), with the majority being around 0.5. However, values of at least 0.9 were also seen in about fifteen percent of the simulations. Also, estimated weights of 0.2 or less were rarely given. With respect to Setting (5.1), notable peaks about $\hat{\rho}_N = 0$ and $\hat{\rho}_N = 1$ were seen. For QIF2a, $\hat{\rho}_N = 1$ in approximately half of the simulations, while this only occurred about twenty percent of the time with QIF2b. For QIF4, $\hat{\rho}_N = 1$ in less than twenty percent of the simulations, and only three percent for QIF5. In the majority of simulations, QIF was superior to GEE and QIF2b used $\hat{\rho}_N = 0$ due in part to large bias estimates. Additionally, small values for $\hat{\rho}_N$ were also seen in the majority of simulations for QIF4 and QIF5 due to using large values for d_N^2, as $\tilde{C}_N(\boldsymbol{\beta})$ is not necessarily meant to work well when the true covariances rely upon the respective cluster sizes. It is interesting that in simulations in which QIF2b used $\hat{\rho}_N = 1$, for example, QIF only performed approximately as well as GEE, while QIF2 performed slightly better. This is because QIF's estimating equations and GEE are in different classes, allowing the completely model-based QIF version to be more efficient than GEE in this Scenario. Additionally, depending on whether QIF performed better than GEE or not, use of $0 < \hat{\rho}_N < 1$ could either decrease or increase MSE as compared with QIF.

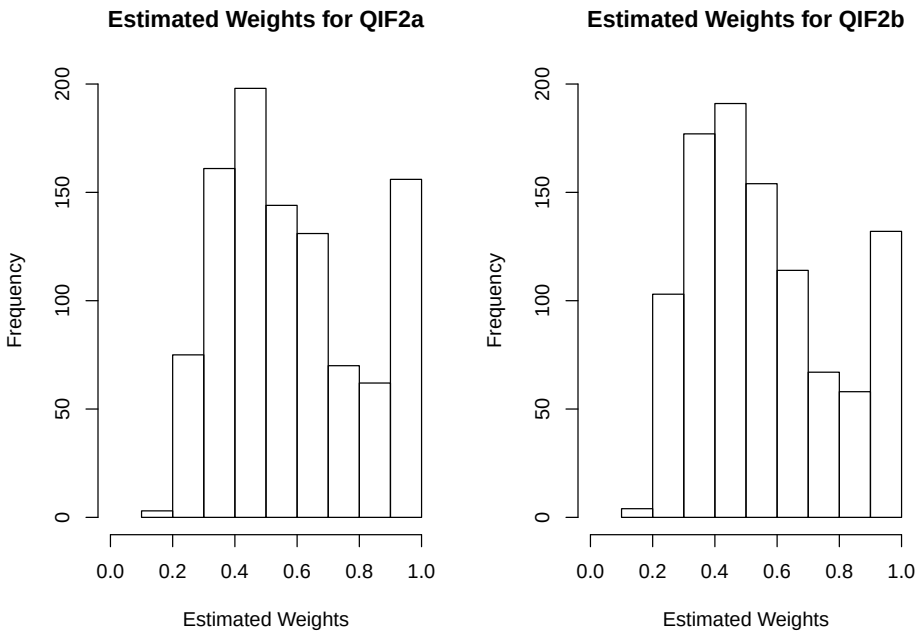

Figure 4.1: Values of $\hat{\rho}_N$ from using QIF2 to analyze the 1000 simulated datasets from Setting (1.3).

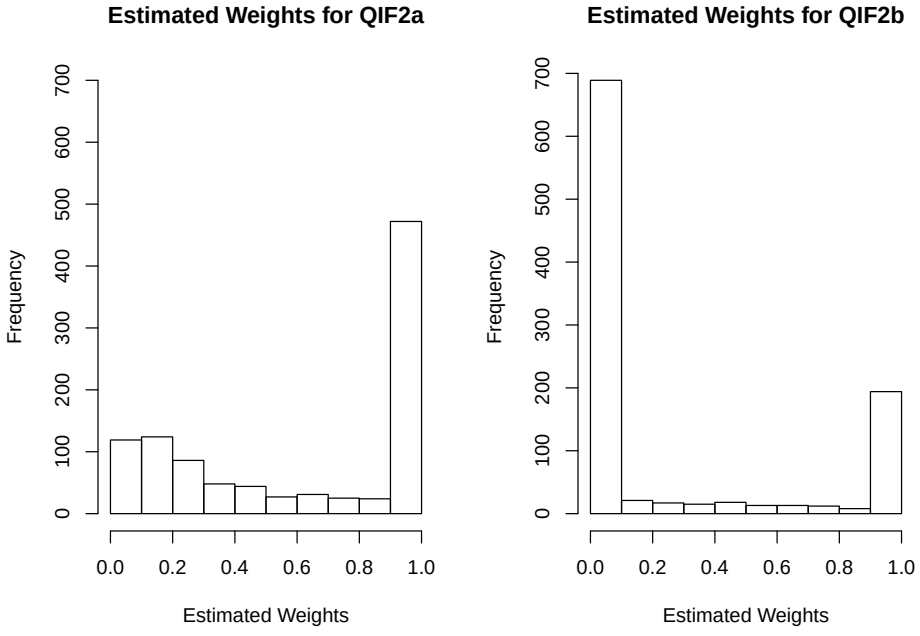

Figure 4.2: Values of $\hat{\rho}_N$ from using QIF2 to analyze the 1000 simulated datasets from Setting (5.1).

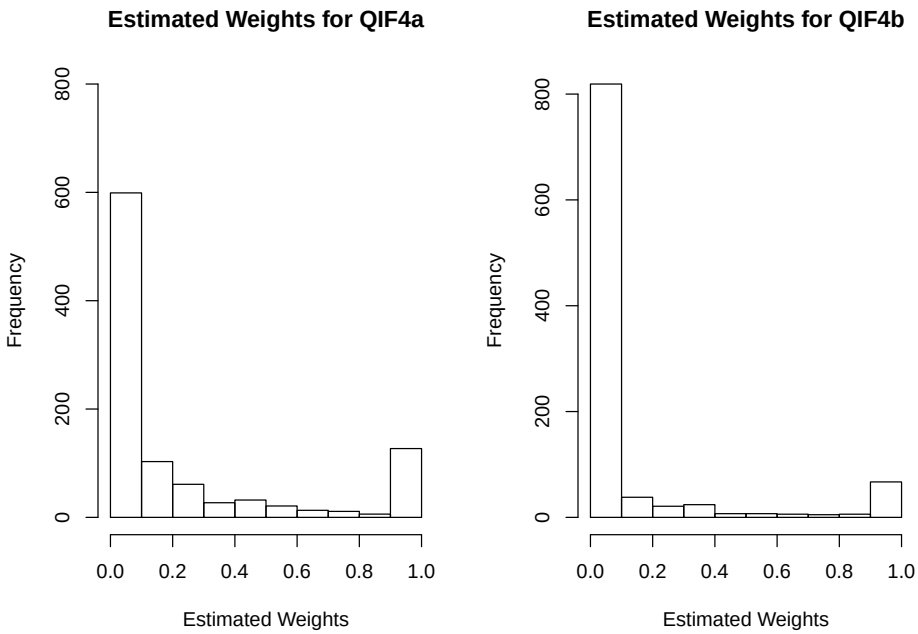

Figure 4.3: Values of $\hat{\rho}_N$ from using QIF4 to analyze the 1000 simulated datasets from Setting (5.1).

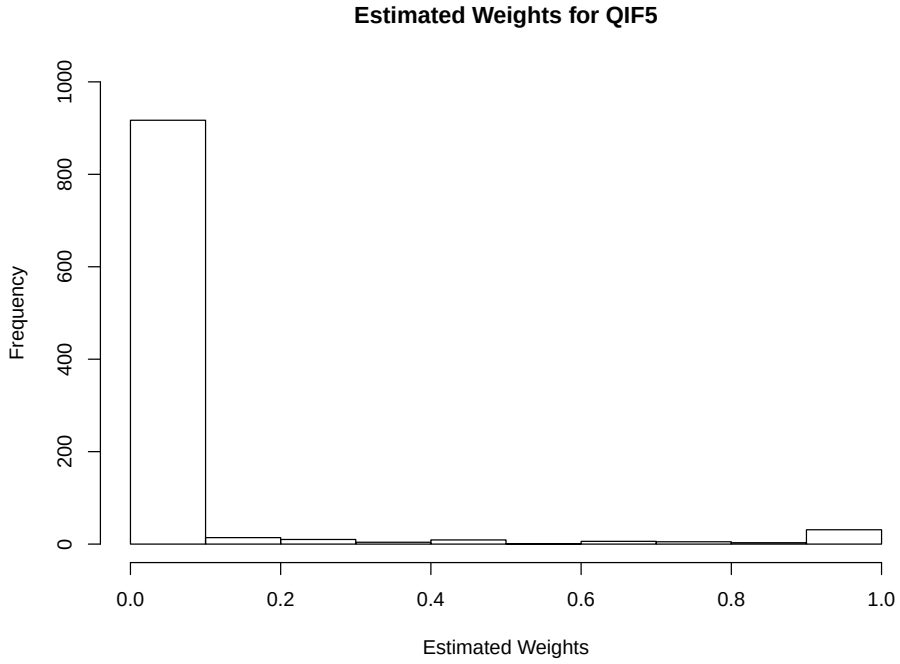

Figure 4.4: Values of $\hat{\rho}_N$ from using QIF5 to analyze the 1000 simulated datasets from Setting (5.1).

4.3.2 Via Application to an AIDS Dataset

As demonstrated in the previous chapter via a GRT dataset from Yudkin and Moher (2001), results from QIF and GEE can be notably different when the number of clusters is small. In this example, independent clusters were twenty-one medical care practices, ranging in size from 28 to 244 patients with coronary heart disease (CHD). The marginal logistic regression model used three different treatment records to predict the marginal probability of any given patient being adequately assessed for three CHD risk factors. The variability in $C_N(\boldsymbol{\beta})$ was particularly demonstrated when aspirin percentage was used as a covariate. For this model, the size of one practice was largely influential upon the weights, estimated via $C_N(\boldsymbol{\beta})$, given to outcomes in the corresponding estimating equations. Due to this large variability within $C_N(\boldsymbol{\beta})$, it turns out that $\hat{\rho}_N = 1$ for every proposed QIF version implementing a weighted combination of $C_N(\boldsymbol{\beta})$ and another matrix. This implies that full use of $\hat{\boldsymbol{M}}_N$ or $\tilde{C}_N(\boldsymbol{\beta})$ is estimated to give the smallest expected quadratic loss for the respective proposed weighting matrix. Additionally, QIF3 gives results that are more similar to GEE implementing a common exchangeable correlation structure than to QIF and QIF2, although differences were still evident.

For an illustrative example on how estimation of $\hat{\rho}_N$ works, we use an AIDS dataset (Kaslow *et al.*, 1987; Huang, Wu, and Zhou, 2002) that was utilized by Qu and Li (2006) to demonstrate their extension of QIF for varying-coefficient models. The dataset contains 283 males who became HIV-positive, each contributing one to fourteen observations at unequally spaced time points, where time is years since infection and ranges from 0.1 to 5.9. The longitudinal outcome of interest is CD4 percentage, while subjects' baseline covariates are age in years, smoking status, and CD4 percentage before infection (pre-CD4). For more information, refer to the previously cited manuscripts.

As the marginal mean CD4 percentage over time is of interest, we fit the following

parsimonious model:

$$\begin{aligned} E(CD4\%_{it}) &= \beta_0 + \beta_1 time_{it} + \beta_2 time_{it}^2 + \beta_3(preCD4\%_i - 42.69) + \\ &\quad \beta_4(age_i - 34.36) + \beta_5 smoke_i; \end{aligned}$$

$i = 1, \ldots, 283;\ t = 1, \ldots, n_i$. Here, pre-CD4 percentage and age are centered at their respective sample means, and $smoke_i$ indicates whether the ith subject smokes or not. We additionally fit a model without age and the smoking indicator, as they do not have a statistically significant impact on mean CD4 percentage.

The results from fitting these two models are shown in Table 4.4, which gives the regression parameter estimates from GEE and each QIF version, in addition to $\hat{\rho}_N$ when applicable. Each method and model combination was implemented twice, once using each of a working exchangeable and AR-1 correlation structure, with the exception of the QIF versions utilizing $\tilde{C}_N(\boldsymbol{\beta})$ that are not applicable for AR-1. Due to each subject contributing a varying number of unequally spaced observations, an exchangeable structure may be as reasonable of a guess at the true correlation structure as AR-1, and was implemented by Qu and Li (2006). The estimated variance, AR-1 correlation, and exchangeable correlation parameters from GEE were always around 106.5, 0.77, and 0.64, respectively.

Results show that this dataset is an example in which the differing weights given to outcomes in the corresponding estimating equations do not have a large impact, as parameter estimates do not vary across methods to a very notable degree. However, several parameters estimated by QIF3 did appear to be larger in magnitude than for the other methods, although not to a large degree. Values for $\hat{\rho}_N$, however, did vary across methods, models, and working correlation structures. As clusters varied in size, a notable amount of bias was estimated when using QIF2, especially seen when comparing the values for $\hat{\rho}_N$ from QIF2a and QIF2b when fitting a model with only the first three covariates and using an exchangeable structure. When using AR-1, $\hat{\rho}_N$ was relatively small in value, ranging from 0.22 to

0.46, similar to its corresponding empirical means from Setting 8.4. This likely occurred due to a misspecification in the working covariance structure, in which case $C_N(\boldsymbol{\beta})$ can be more accurate in accounting for the true covariances within the data. However, similar to Settings 8.1 and 8.3, $\hat{\rho}_N$ was large when using an exchangeable structure, except when using QIF2b with only three covariates, QIF4, or QIF5. This result may imply the exchangeable structure closely resembles reality, although Setting 8.1 indicates that this working structure can potentially lead to large values for QIF2's $\hat{\rho}_N$ even when N is this large and the true correlation in not exchangeable. Another notable trend is that $\hat{\rho}_N$ always decreased in value, except when utilizing an exchangeable structure with QIF2a or QIF5, as age and the smoking indicator were taken out of the model. This possibly occurred since the dimension of $C_N(\boldsymbol{\beta})$ reduces when decreasing the number of covariates, therefore diminishing its overall variability. Here, not using age and smoking implies that $C_N(\boldsymbol{\beta})$ does not have to estimate the effect these two covariates have on the covariance structure of the data.

4.4 Concluding Remarks

Although QIF has a theoretical efficiency advantage over GEE, its estimation performance may actually be inferior, particularly when the number of clusters is small. Simulations demonstrated that this result can be seen even in general correlated data settings and when the working correlation structure is not exchangeable, complementing the results of the previous chapter. To improve QIF in this regard, we proposed several different weighting matrices to replace $C_N(\boldsymbol{\beta})$ inside the corresponding estimating equations. One utilizes an empirical matrix that was intended to average out the effect from cluster size variation, denoted as $\tilde{C}_N(\boldsymbol{\beta})$, while the others implement a weighted combination of $C_N(\boldsymbol{\beta})$ and either $\tilde{C}_N(\boldsymbol{\beta})$ or $\hat{\boldsymbol{M}}_N$, the model-based covariance matrix. These combinations optimally take into account the bias and variability within each of these matrices, minimizing the expected quadratic loss of the proposed matrix and allowing for the implementation of an asymptotically optimal weighting matrix.

Table 4.4: AIDS Dataset Analysis Results

Exchangeable Working Correlation Structure

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\rho}_N$
GEE	36.68	-4.64	0.37	0.40	-0.04	0.69	
	36.94	-4.65	0.37	0.39			
QIF	36.77	-4.63	0.37	0.37	-0.04	0.53	
	36.93	-4.63	0.37	0.37			
QIF2a	36.84	-4.78	0.40	0.38	-0.05	0.67	1.00
	37.09	-4.78	0.40	0.37			1.00
QIF2b	36.83	-4.76	0.40	0.38	-0.05	0.65	0.97
	36.97	-4.67	0.38	0.37			0.28
QIF3	37.98	-5.11	0.46	0.39	-0.08	0.44	
	37.99	-5.01	0.44	0.37			
QIF4a	37.22	-4.79	0.40	0.38	-0.06	0.51	0.47
	37.08	-4.66	0.38	0.37			0.22
QIF4b	37.22	-4.79	0.40	0.38	-0.06	0.51	0.47
	37.08	-4.66	0.38	0.37			0.22
QIF5	36.91	-4.67	0.38	0.38	-0.05	0.55	0.19
	36.98	-4.62	0.37	0.37			0.09

AR-1 Working Correlation Structure

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\rho}_N$
GEE	36.62	-4.86	0.43	0.39	-0.03	0.48	
	36.79	-4.86	0.43	0.38			
QIF	36.54	-4.50	0.34	0.39	0.00	0.46	
	36.75	-4.57	0.34	0.39			
QIF2a	36.59	-4.69	0.38	0.39	-0.01	0.49	0.46
	36.77	-4.67	0.37	0.39			0.28
QIF2b	36.58	-4.65	0.37	0.39	-0.01	0.49	0.38
	36.77	-4.65	0.36	0.39			0.22

In practice, the optimal weight, ρ_N, for each method needs to be estimated, and the overall performance of QIF as compared with its proposed alternate versions was demonstrated via simulations in a variety of scenarios. In general, QIF2 typically improved QIF's estimation performance to be comparable to that of GEE's when necessary, and, alternatively, worked as good as QIF in settings in which GEE produced estimates with the largest variability. However, any use of $\tilde{C}_N(\boldsymbol{\beta})$ was detrimental in the repeated measures settings of Scenario 8, and although advantageous to QIF in the GRT scenarios, QIF2 had an overall superior performance as compared with QIF3, QIF4, and QIF5. Additionally, as the sample estimate for the variability in $C_N(\boldsymbol{\beta})$ contained bias, two different methods for obtaining $\hat{\rho}_N$ were proposed for use in QIF2 and QIF4. No notable differences were seen in terms of estimation performance, except in some GRT settings.

Overall, QIF2 appears to typically work approximately as well as the method, either GEE or the regular QIF, that produces estimates with the least variability in any given setting. This is advantageous in that it allows for the avoidance of having to choose between the regular QIF and GEE. For instance, one could argue that GEE should be used when N is small and the covariance structure is reasonably chosen, while QIF should be chosen when N is large or the working covariance structure has the potential to largely deviate from the truth. These are rather arbitrary aspects, and using QIF2, which optimally takes into account the model-based and empirical aspects of GEE and QIF, respectively, prevents the use of a potentially inferior method.

As mentioned in the previous chapter, the QIF can itself be used as a statistic in goodness-of-fit and likelihood ratio score-type tests (Qu *et al.*, 2000; Song *et al.*, 2009). It has also been shown to be much more robust to outliers than GEE (Qu and Song, 2004). As the proposed weighting matrices are asymptotically optimal, QIF2a and QIF2b have the same asymptotic properties as QIF, and their inference function values can also be used as test statistics. Also, they will be more robust to outliers than GEE, although further study is required to determine how influential outliers are to QIF2 as compared with QIF and

GEE. Future research is also needed to determine the validity of the proposed QIF versions as test statistics, as they will partially employ a covariance matrix that could be biased. Additionally, these matrices will also influence the validity of the empirical SE estimates the regular QIF employs. Therefore, future research on obtaining valid SEs is needed, particularly in small-sample settings.

4.5 Appendix

This Appendix has three separate sections. The first presents proofs dealing with C_N^* and $\hat{C}_N^*$, the second presents proofs dealing with $\boldsymbol{S}_N^1$ and $\hat{\boldsymbol{S}}_N^1$, and the last presents proofs dealing with $\boldsymbol{S}_N^2$ and $\hat{\boldsymbol{S}}_N^2$. This Appendix provides proofs of results given in Section 4.2. The general form for these proofs come from the work of Han and Song (2011) and Ledoit and Wolf (2004). We direct the reader to proofs in these manuscripts when our proposed method does not require any modifications to the corresponding work. Additionally, following Han and Song (2011), we let c be a finite, generic constant which can change in value.

4.5.1 Proofs for Results Using C_N^* and $\hat{C}_N^*$

We first prove that $\rho_N = \tau_N^2/(\alpha_N^2 + \tau_N^2)$ minimizes $E[||C_N^* - \boldsymbol{\Sigma}_N||^2]$, closely related to the corresponding proof given in Ledoit and Wolf (2004):

Proof.

$$\begin{aligned}
E[||C_N^* - \boldsymbol{\Sigma}_N||^2] &= E[||\rho_N \boldsymbol{M}_N + (1-\rho_N)C_N(\boldsymbol{\beta}) - \rho_N \boldsymbol{\Sigma}_N - (1-\rho_N)\boldsymbol{\Sigma}_N||^2] \\
&= E[||\rho_N[\boldsymbol{M}_N - \boldsymbol{\Sigma}_N] + (1-\rho_N)[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N]||^2] \\
&= \rho_N^2 E[||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2] + (1-\rho_N)^2 E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + \\
&\quad 2\rho_N(1-\rho_N)E[< \boldsymbol{M}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >] \\
&= \rho_N^2 \alpha_N^2 + (1-\rho_N)^2 \tau_N^2 + 0.
\end{aligned}$$

Now take the first derivative with respect to ρ_N and set equal to 0:

$$2\rho_N \alpha_N^2 - 2(1-\rho_N)\tau_N^2 = 0.$$

Solving for ρ_N,

$$\rho_N = \frac{\tau_N^2}{\alpha_N^2 + \tau_N^2} = \tau_N^2/\delta_N^2,$$

where $\delta_N^2 = \alpha_N^2 + \tau_N^2 = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2]$.

The following proofs are based on the Lemma given by Han and Song (2011) and its corresponding conditions. The first two parts of their Lemma are equivalent in our scenario, and we therefore omit the proofs here.

We now prove $\delta_N^2 = \alpha_N^2 + \tau_N^2$:

Proof.

$$\begin{aligned}
\delta_N^2 &= E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2] = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N + \boldsymbol{\Sigma}_N - \boldsymbol{M}_N||^2] \\
&= E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + E[||\boldsymbol{\Sigma}_N - \boldsymbol{M}_N||^2] + \\
&\quad 2E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \boldsymbol{M}_N >] \\
&= \tau_N^2 + \alpha_N^2 + 0 = \tau_N^2 + \alpha_N^2.
\end{aligned}$$

We now prove $||\boldsymbol{M}_N||$, δ_N^2, α_N^2, and τ_N^2 remain bounded, and $\tau_N^2 \to 0$ as $N \to \infty$:

Proof.

We first prove $||\boldsymbol{M}_N||$ remains bounded, using arguments similar to those implemented by Han and Song (2011) for proving that $||\boldsymbol{\Sigma}_N||$ remains bounded:

$$\begin{aligned}
||\boldsymbol{M}_N|| &= ||\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{B}_i \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{B}_i^T|| \leq \frac{c}{N}\sum_{i=1}^{N} ||\boldsymbol{B}_i||||\boldsymbol{R}_i(\boldsymbol{\alpha})||||\boldsymbol{B}_i^T|| \\
&\leq \frac{c}{N}\sum_{i=1}^{N} ||\boldsymbol{B}_i||||\boldsymbol{B}_i^T|| = \frac{c}{N}\sum_{i=1}^{N} ||\boldsymbol{B}_i||^2 < \infty.
\end{aligned}$$

We now show α_N^2 remains bounded:

$$\alpha_N^2 \;=\; ||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2 = ||\boldsymbol{M}_N||^2 + ||\boldsymbol{\Sigma}_N||^2 - 2 < \boldsymbol{M}_N, \boldsymbol{\Sigma}_N >$$

α_N^2 is composed of three quantities. We already have the results that $||\boldsymbol{\Sigma}_N||^2 < \infty$ (Han and Song, 2011) and $||\boldsymbol{M}_N||^2 < \infty$, so we now only need to show that $| < \boldsymbol{\Sigma}_N, \boldsymbol{M}_N > | < \infty$: $| < \boldsymbol{\Sigma}_N, \boldsymbol{M}_N > | \leq c||\boldsymbol{\Sigma}_N||||\boldsymbol{M}_N|| < \infty$. So $0 \leq \alpha_N^2 < \infty$.

Han and Song (2011) proved that τ_N^2 is bounded and that $\tau_N^2 \xrightarrow{p} 0$.

Now we prove that δ_N^2 is bounded:
$\delta_N^2 < \infty$ is implied since $\delta_N^2 = \alpha_N^2 + \tau_N^2$, and α_N and τ_N^2 remain bounded.

We now prove $a_N^2 - \alpha_N^2$, $t_N^2 - \tau_N^2$, and $d_N^2 - \delta_N^2$ all converge in quadratic mean to zero as $N \to \infty$, under the assumption that $E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4] \to 0$ as $N \to \infty$. The derivation for the bias in t_N^2 and the proof that $E[(\hat{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$ are given at the end of this Subsection.

Proof.

We first prove $E[(d_N^2 - \delta_N^2)^2] \to 0$ as $N \to \infty$:

$$\begin{aligned} d_N^2 &= ||C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N||^2 = ||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N + \boldsymbol{M}_N - \hat{\boldsymbol{M}}_N||^2 \\ &= ||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2 + ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2 - 2 < C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N, \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N > \end{aligned}$$

Therefore,

$$\begin{aligned} d_N^2 - \delta_N^2 &= ||C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N||^2 - E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2] \\ &= ||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2 + ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2 - \\ &\quad 2 < C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N, \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N > -E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2] \\ &= (||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2 - E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2]) + ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2 - \\ &\quad 2 < C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N, \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N > \end{aligned}$$

Following the procedure of the corresponding proof given in Han and Song (2011), all we need to do is show that the expected value of the square of each of these three terms goes to 0 as $N \to \infty$. In other words, show that each of these three terms converges in quadratic mean to 0.

First Term: $||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2 - E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2]$

$$\begin{aligned} &||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2 - E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2] \\ = \;& ||C_N(\boldsymbol{\beta})||^2 + ||\boldsymbol{M}_N||^2 - 2 < C_N(\boldsymbol{\beta}), \boldsymbol{M}_N > - \\ & E[||C_N(\boldsymbol{\beta})||^2 + ||\boldsymbol{M}_N||^2 - 2 < C_N(\boldsymbol{\beta}), \boldsymbol{M}_N >] \\ = \;& [||C_N(\boldsymbol{\beta})||^2 - E(||C_N(\boldsymbol{\beta})||^2)] + [||\boldsymbol{M}_N||^2 - E(||\boldsymbol{M}_N||^2)] - \\ & 2[< C_N(\boldsymbol{\beta}), \boldsymbol{M}_N > -E(< C_N(\boldsymbol{\beta}), \boldsymbol{M}_N >)] \\ = \;& [||C_N(\boldsymbol{\beta})||^2 - E(||C_N(\boldsymbol{\beta})||^2)] + 0 - 2[< C_N(\boldsymbol{\beta}), \boldsymbol{M}_N > - < \boldsymbol{\Sigma}_N, \boldsymbol{M}_N >] \\ = \;& [||C_N(\boldsymbol{\beta})||^2 - E(||C_N(\boldsymbol{\beta})||^2)] - 2 < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{M}_N > \end{aligned}$$

This shows that the first term can be rewritten as the sum of two terms. Therefore, we just need to show that both of these terms converge in quadratic mean to 0. Han and Song (2011) proved that as $N \to \infty$, $E\left[(||C_N(\boldsymbol{\beta})||^2 - E[||C_N(\boldsymbol{\beta})||^2])^2\right] \to 0$. Now we prove that $E[(2 < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{M}_N >)^2] \to 0$ as $N \to \infty$:

By the Cauchy-Schwarz Inequality,

$$0 \leq (< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{M}_N >)^2 \leq c||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2||\boldsymbol{M}_N||^2.$$

Therefore,

$$\begin{aligned} 0 &\leq E[(2 < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{M}_N >)^2] \leq cE[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2||\boldsymbol{M}_N||^2] \\ &= c||\boldsymbol{M}_N||^2 E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = c||\boldsymbol{M}_N||^2\tau_N^2 \to 0 \end{aligned}$$

as $N \to \infty$.

Second Term: $||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2$

By assumption,

$$E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4] \to 0$$

Third Term: $2< C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N, \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N >$

By the Cauchy-Schwarz Inequality, showing $E(c_N^2) \to 0$ and $E(b_N^2) \to 0$ implies $E(|c_N b_N|) \to 0$, and so

$$0 \leq E[(< C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N, \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N >)^2] \leq cE[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] \to 0,$$

which we now prove: $||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2 = ||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + ||\boldsymbol{\Sigma}_N - \boldsymbol{M}_N||^2 + 2 < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \boldsymbol{M}_N >$, and so $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2]$ can be written as the sum of three terms, each of which we now show converge to zero.

1. $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2]$: if we let $c_N = ||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2$ and $b_N = ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2$, then all we need to show is $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^4] \to 0$ and $E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4] \to 0$. Han and Song (2011) proved $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^4] \to 0$, and by assumption $E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4] \to 0$.

2. $E[||\boldsymbol{\Sigma}_N - \boldsymbol{M}_N||^2||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] = ||\boldsymbol{\Sigma}_N - \boldsymbol{M}_N||^2 E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] \leq cE[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] \to 0$ by assumption.

3. $E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \boldsymbol{M}_N > ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2]$: By the Cauchy-Schwarz Inequality, $0 \leq E(| < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \boldsymbol{M}_N > |||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2) \leq cE[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||||\boldsymbol{\Sigma}_N - \boldsymbol{M}_N||||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] \leq cE[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2]$. Letting $c_N = ||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||$ and $b_N = ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2$, then all we need is $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = \tau_N^2 \to 0$ and $E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4] \to 0$, which have already been shown and assumed, respectively.

Therefore, since all three terms converge to 0, $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{M}_N||^2||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] \to 0$ as $N \to \infty$.

We now prove $E[(t_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$:

Han and Song (2011) proved than $E[(\bar{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$, and the proof that $E[(t_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$ then follows from Ledoit and Wolf (2004).

Both $E[(d_N^2 - \delta_N^2)^2]$ and $E[(t_N^2 - \tau_N^2)^2]$ converge to 0 as $N \to \infty$, and therefore so does $E[(a_N^2 - \alpha_N^2)^2]$.

We now prove the two theorems similar to those given by Han and Song (2011), now based upon C_N^* and $\hat{C}_N^*$. According to the first theorem of Han and Song (2011), the following five conditions must be met in order for the previous proofs and both Theorems to be valid: (1) $\sup_{i \geq 1} n_i < \infty$; (2) $\sup_{i \geq 1} E||\boldsymbol{e}_i||^8 < \infty$; (3) $h(\mu)$ is differentiable; (4) $\limsup_{N \to \infty} \frac{1}{N} \sum_{i=1}^N ||X_i^{\wedge 8}||^2 < \infty$, and $\wedge$ is elementwise exponentiation; and (5) for any $\boldsymbol{\beta} \in \boldsymbol{B}$, $\limsup_{N \to \infty} \frac{1}{N} \sum_{i=1}^N ||\boldsymbol{G}_i^8 \boldsymbol{A}_i^{-4}||^2 < \infty$, where $\boldsymbol{B}$ is some subset of $\boldsymbol{R}^p$. Here, $\boldsymbol{G}_i$ is a matrix with $[\dot{h}(\mu_{ij})]^{-1}$ as diagonal elements, where the dot represents the derivative with respect to μ_{ij}, and X_i is the matrix of covariate values for the ith cluster.

Theorem 1.1. For $\boldsymbol{\beta} \in \boldsymbol{B}$, $E[||C_N^* - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $C_N^* - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$.

Proof. Following Ledoit and Wolf (2004):

$$\begin{aligned}
0 &\leq ||C_N^* - \boldsymbol{\Sigma}_N||^2 = ||\frac{\tau_N^2}{\delta_N^2}\boldsymbol{M}_N + \frac{\alpha_N^2}{\delta_N^2}C_N(\boldsymbol{\beta}) - \frac{\tau_N^2+\alpha_N^2}{\delta_N^2}\boldsymbol{\Sigma}_N||^2 \\
&= ||\frac{\tau_N^2}{\delta_N^2}[\boldsymbol{M}_N - \boldsymbol{\Sigma}_N] + \frac{\alpha_N^2}{\delta_N^2}[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N]||^2 \\
&= (\frac{\tau_N^2}{\delta_N^2})^2||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2 + (\frac{\alpha_N^2}{\delta_N^2})^2||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + \\
&\quad 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \boldsymbol{M}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N > \\
&\leq (\frac{\tau_N^2}{\delta_N^2})^2||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2 + ||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \boldsymbol{M}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >
\end{aligned}$$

The proof will be complete if we show these three terms have expectations that converge to 0.

First Term:

$$\begin{aligned}
E[(\frac{\tau_N^2}{\delta_N^2})^2||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2] &= (\frac{\tau_N^2}{\delta_N^2})^2||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2 \\
&= (\frac{\tau_N^2}{\delta_N^2})^2\alpha_N^2 = \tau_N^2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) \leq \tau_N^2 \xrightarrow{p} 0
\end{aligned}$$

Second Term:

$$E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = \tau_N^2 \xrightarrow{p} 0.$$

Third Term:

$$\begin{aligned}
& E[2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \boldsymbol{M}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >] \\
= \; & 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \boldsymbol{M}_N - \boldsymbol{\Sigma}_N, E[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N] > \\
= \; & 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \boldsymbol{M}_N - \boldsymbol{\Sigma}_N, 0 >= 0
\end{aligned}$$

Theorem 1.2. For $\boldsymbol{\beta} \in \boldsymbol{B}$, $E[||\hat{C}_N^* - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $\hat{C}_N^* - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$.

Proof. Following Ledoit and Wolf (2004):

$$
\begin{aligned}
0 \;\leq\; & ||\hat{C}_N^* - C_N^*||^2 = |||[\frac{t_N^2}{d_N^2}\hat{\boldsymbol{M}}_N + \frac{a_N^2}{d_N^2}C_N(\boldsymbol{\beta})] - [\frac{\tau_N^2}{\delta_N^2}\boldsymbol{M}_N + \frac{\alpha_N^2}{\delta_N^2}C_N(\boldsymbol{\beta})]||^2 \\
= \;& ||\frac{t_N^2}{d_N^2}\hat{\boldsymbol{M}}_N - \frac{\tau_N^2}{\delta_N^2}\boldsymbol{M}_N + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})C_N(\boldsymbol{\beta}) + \frac{\tau_N^2}{\delta_N^2}\hat{\boldsymbol{M}}_N - \frac{\tau_N^2}{\delta_N^2}\hat{\boldsymbol{M}}_N||^2 \\
= \;& ||\frac{\tau_N^2}{\delta_N^2}(\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N) + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})[C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N] + \frac{t_N^2}{d_N^2}\hat{\boldsymbol{M}}_N - \frac{\tau_N^2}{\delta_N^2}\hat{\boldsymbol{M}}_N + \\
& (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})\hat{\boldsymbol{M}}_N||^2 \\
= \;& ||\frac{\tau_N^2}{\delta_N^2}(\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N) + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})[C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N]||^2 \\
= \;& (\frac{\tau_N^2}{\delta_N^2})^2||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2 + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2||C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N||^2 + \\
& 2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2}) < \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N, C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N > \\
\leq \;& ||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2 + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2 d_N^2 + \\
& 2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2}) < \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N, C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N >
\end{aligned}
$$

Now we need to show that $E[||\hat{C}_N^* - C_N^*||^2] \to 0$, or that the expectations of each of the above three terms all converge to 0:

First Term: By assumption,

$$E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2] \to 0$$

Second Term:

$$\begin{aligned}
(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2 d_N^2 &= [\frac{a_N^4}{d_N^4} + \frac{\alpha_N^4}{\delta_N^4} - 2\frac{a_N^2\alpha_N^2}{d_N^2\delta_N^2}]d_N^2 \\
&= \frac{a_N^4}{d_N^2} + \frac{\alpha_N^4 d_N^2}{\delta_N^4} - \frac{2a_N^2\alpha_N^2}{\delta_N^2} \\
&= (a_N^4\delta_N^4 + \alpha_N^4 d_N^4 - 2a_N^2\alpha_N^2\delta_N^2 d_N^2)/(d_N^2\delta_N^4) \\
&= (a_N^2\delta_N^2 - \alpha_N^2 d_N^2)^2/(d_N^2\delta_N^4).
\end{aligned}$$

Ledoit and Wolf (2004) prove that $E[(a_N^2\delta_N^2 - \alpha_N^2 d_N^2)^2/(d_N^2\delta_N^4)] \to 0$ as $N \to \infty$.

Third Term:

$2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2}) < \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N, C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N >$

Using the Cauchy-Schwarz Inequality and denoting $c_N = 2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})$ and $b_N =< \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N, C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N >$, showing $E(b_N^2) \to 0$ and $E(c_N^2) \to 0$ will prove that the expectation of this third term goes to zero as $N \to \infty$.

$E(b_N^2)$: $0 \leq E[(< \hat{\boldsymbol{M}}_N - \boldsymbol{M}_N, C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N >)^2] \leq E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^2 ||C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N||^2] \leq$ (by the Cauchy-Schwarz Inequality) $\sqrt{E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4]}\sqrt{E[||C_N(\boldsymbol{\beta}) - \hat{\boldsymbol{M}}_N||^4]} \leq$ $c\sqrt{E[||\hat{\boldsymbol{M}}_N - \boldsymbol{M}_N||^4]} \to 0$ by assumption.

$E(c_N^2)$: $0 \leq 4(\frac{\tau_N^2}{\delta_N^2})^2 E[(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2] \to 0$ since $E[(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2]$ is bounded and $\frac{\tau_N^2}{\delta_N^2} \to 0$ under the assumption of a misspecified covariance structure.

We have now shown $E[||\hat{C}_N^* - C_N^*||^2] \to 0$ as $N \to \infty$. Using this in conjunction with Theorem 1 and the proof given by Han and Song (2011), we have $E[||\hat{C}_N^* - \boldsymbol{\Sigma}_N||^2] \to 0$.

We note that all proofs assume $||\boldsymbol{M}_N - \boldsymbol{\Sigma}_N||^2 > 0$ and $E[||\hat{\boldsymbol{M}}_N - C_N(\boldsymbol{\beta})||^2]$ does not converge to 0. Specifically, we assume that the working covariance structure is misspecified in some manner. In reality, we have $N < \infty$, and even if the covariance structure is correctly specified for all data, our method will still work. Additionally, if $E[||\hat{\boldsymbol{M}}_N - C_N(\boldsymbol{\beta})||^2] \to 0$, for large N it would not make a difference how much weight is given to $\hat{\boldsymbol{M}}_N$ and $C_N(\boldsymbol{\beta})$.

We now derive the bias in $\bar{t}_N^2$, and then prove $E[(\hat{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$.

$\bar{t}_N^2$ is a function of $mp \times mp$ matrices, and is used to estimate $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2]$. Now let $\bar{x}_N$ represent any one of the $(mp)^2$ elements comprising $C_N(\boldsymbol{\beta})$, and $\mu_N = E[\bar{x}_N]$. $||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2$ is just the sum of the square of each of the $(mp)^2$ elements comprising $C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N$, divided by mp, and so $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2]$ is just the sum of the variances of the $(mp)^2$ elements comprising $C_N(\boldsymbol{\beta})$, divided by mp.

Now $C_N(\boldsymbol{\beta})$ is the sample average over the N extended score equations, and $\bar{x}_N = (1/N)\sum_{i=1}^N x_i$ and $\mu_N = (1/N)\sum_{i=1}^N \mu_i$. $Var(\bar{x}_N) = (1/N^2)\sum_{i=1}^N Var(x_i) = (1/N^2)\sum_{i=1}^N \sigma_i^2$. $\bar{t}_N^2$ estimates $Var(\bar{x}_N)$ with $(1/N^2)\sum_{i=1}^N (x_i - \bar{x}_N)^2$. Now

$$\begin{aligned}
(1/N^2)\sum_{i=1}^N E[(x_i - \bar{x}_N)^2] &= (1/N^2)\sum_{i=1}^N E(x_i^2) - (1/N)E(\bar{x}_N^2) \\
&= (1/N^2)\sum_{i=1}^N (\sigma_i^2 + \mu_i^2) - (1/N)[Var(\bar{x}_N) + E(\bar{x}_N)^2] \\
&= Var(\bar{x}_N) + (1/N^2)\sum_{i=1}^N \mu_i^2 - (1/N)[Var(\bar{x}_N) + [(1/N)\sum_{i=1}^N \mu_i]^2] \\
&= [(N-1)/N]Var(\bar{x}_N) + (1/N^2)\sum_{i=1}^N \mu_i^2 - (1/N^3)\left(\sum_{i=1}^N \mu_i\right)^2
\end{aligned}$$

As $(N-1)/N \approx 1$, we can ignore the first term, and therefore the bias of the corresponding arbitrary element within $\bar{t}_N^2$ can approximated by $(1/N^2)\sum_{i=1}^N \mu_i^2 - (1/N^3)\left(\sum_{i=1}^N \mu_i\right)^2$. Since we need an estimate for the sum of the variances for each of the $(2p)^2$ elements comprising $C_N(\boldsymbol{\beta})$, we need to sum the biases from each of the variance estimates. This leads to $Bias(\bar{t}_N^2) \approx (1/N^2)\sum_{i=1}^N ||Cov[g_i(\boldsymbol{\beta})]||^2 - (1/N^3)||\sum_{i=1}^N Cov[g_i(\boldsymbol{\beta})]||^2$.

We now show the bias terms go to 0 as $N \to \infty$:

1. $(1/N^2)\sum_{i=1}^N ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T||^2$. $||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T||^2 \leq c||\boldsymbol{B}_i||^4 < \infty$ by assumptions, and so $(1/N)\sum_{i=1}^N ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2 < \infty$. Therefore, since $(1/N) \to 0$,
   $(1/N^2)\sum_{i=1}^N ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2 = (1/N)\left[(1/N)\sum_{i=1}^N ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2\right] \to 0.$

2. $(1/N^3)||\sum_{i=1}^{N} Cov[g_i(\boldsymbol{\beta})]||^2 = (1/N)||\boldsymbol{M}_N||^2$. We have already proven that $||\boldsymbol{M}_N||^2 < \infty$, implying $(1/N)||\boldsymbol{M}_N||^2 \to 0$.

We now prove $E[(\hat{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$:

$E[\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2)] = E[\bar{t}_N^2] - (1/N^2)\sum_{i=1}^{N} ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2 + (1/N)||\boldsymbol{M}_N||^2 \to 0$ since Han and Song (2011) show $E[\bar{t}_N^2] \to 0$, and we have already shown the other two terms converge to 0.

$$
\begin{aligned}
[\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2)]^2 \;=\; & (\bar{t}_N^2)^2 + [(1/N^2)\sum_{i=1}^{N} ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2]^2 + (1/N^2)||\boldsymbol{M}_N||^4 \\
& -2\bar{t}_N^2[(1/N^2)\sum_{i=1}^{N} ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2] + 2\bar{t}_N^2(1/N)||\boldsymbol{M}_N||^2 - \\
& 2[(1/N^2)\sum_{i=1}^{N} ||\boldsymbol{B}_i\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^T]||^2][(1/N)||\boldsymbol{M}_N||^2]
\end{aligned}
$$

Now, to prove $E([\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2)]^2) \to 0$, we need to show that the expected value of each of these six terms goes to 0 as $N \to \infty$. The second, third, and sixth terms are each comprised of two terms that do not contain random variables and have been shown to go to 0, therefore implying these three terms go to 0. Han and Song (2011) showed that $E[(\bar{t}_N^2)^2] \to 0$. The fourth and fifth terms are comprised of a respective bias term and $\bar{t}_N^2$. As we have already shown that the bias terms are bounded, and both the bias terms and $\bar{t}_N^2$ go to 0, we therefore have the result that the fourth and fifth terms also go to 0.

Following Han and Song (2011), $E[(\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2) - \tau_N^2)^2] = E([\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2)]^2) - 2\tau_N^2 E[\bar{t}_N^2 - \widehat{Bias}(\bar{t}_N^2)] + (\tau_N^2)^2 \to 0$ as $N \to \infty$, and the work by Ledoit and Wolf (2004) proves that $E[(\hat{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$.

4.5.2 Proofs for Results Using $\boldsymbol{S}_N^1$ and $\hat{\boldsymbol{S}}_N^1$

We first prove that $\rho_N = \tau_N^2/(\alpha_N^2 + \tau_N^2)$ minimizes $E[||\boldsymbol{S}_N^1 - \boldsymbol{\Sigma}_N||^2]$, closely related to the corresponding proof given in Ledoit and Wolf (2004):

Proof.

$$\begin{aligned}
E[||\boldsymbol{S}_N^1 - \boldsymbol{\Sigma}_N||^2] &= E[||\rho_N \tilde{\boldsymbol{\Sigma}}_N + (1-\rho_N) C_N(\boldsymbol{\beta}) - \rho_N \boldsymbol{\Sigma}_N - (1-\rho_N)\boldsymbol{\Sigma}_N||^2] \\
&= E[||\rho_N[\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N] + (1-\rho_N)[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N]||^2] \\
&= \rho_N^2 E[||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2] + (1-\rho_N)^2 E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + \\
&\quad 2\rho_N(1-\rho_N) E[< \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >] \\
&= \rho_N^2 \alpha_N^2 + (1-\rho_N)^2 \tau_N^2 - 0.
\end{aligned}$$

Now take the first derivative with respect to ρ_N and set equal to 0:

$$2\rho_N \alpha_N^2 - 2(1-\rho_N)\tau_N^2 = 0.$$

Solving for ρ_N, we get

$$\rho_N = \frac{\tau_N^2}{\alpha_N^2 + \tau_N^2} = \tau_N^2/\delta_N^2,$$

where $\delta_N^2 = \alpha_N^2 + \tau_N^2 = E[||C_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^2]$.

The following proofs are based on the Lemma given by Han and Song (2011) and its corresponding conditions. The first two parts of their Lemma are equivalent in our scenario, and we therefore omit the proofs here.

We now prove $\delta_N^2 = \alpha_N^2 + \tau_N^2$:

Proof.

$$\begin{aligned}\delta_N^2 &= E[||C_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^2] = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N + \boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N||^2] \\ &= E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + E[||\boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N||^2] + \\ &\quad 2E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N >] \\ &= \tau_N^2 + \alpha_N^2 + 0 = \tau_N^2 + \alpha_N^2.\end{aligned}$$

We now prove $||\tilde{\boldsymbol{\Sigma}}_N||$, δ_N^2, α_N^2, and τ_N^2 remain bounded, and $\tau_N^2 \to 0$ as $N \to \infty$:

Proof.

We first prove $||\tilde{\boldsymbol{\Sigma}}_N||$ remains bounded, using arguments similar to those implemented by Han and Song (2011) for proving that $||\boldsymbol{\Sigma}_N||$ remains bounded:

$||\tilde{\boldsymbol{\Sigma}}_N|| = ||\frac{1}{N}\sum_{i=1}^N \tilde{\boldsymbol{\Sigma}}_i||$, where $E[\tilde{C}_i(\boldsymbol{\beta})] = \tilde{\boldsymbol{\Sigma}}_i$.

$$\begin{aligned}||\tilde{\boldsymbol{\Sigma}}_N|| &\leq \frac{c}{N}\sum_{i=1}^N ||\tilde{\boldsymbol{\Sigma}}_i|| \leq \frac{cd}{N}\sum_{i=1}^N ||\boldsymbol{\Sigma}_i|| = \frac{c}{N}\sum_{i=1}^N ||\boldsymbol{B}_i\tilde{\boldsymbol{R}}_i\boldsymbol{B}_i^T|| \leq \frac{c}{N}\sum_{i=1}^N ||\boldsymbol{B}_i||||\tilde{\boldsymbol{R}}_i||||\boldsymbol{B}_i^T|| \\ &\leq \frac{c}{N}\sum_{i=1}^N ||\boldsymbol{B}_i||||\boldsymbol{B}_i^T|| = \frac{c}{N}\sum_{i=1}^N ||\boldsymbol{B}_i||^2 < \infty,\end{aligned}$$

where $\tilde{\boldsymbol{R}}_i = Cov(\boldsymbol{e}_i)$,

$$d = max\left[(\widehat{n^2}/n_i^2), \left(\widehat{n^2(n-1)}/[n_i^2(n_i-1)]\right), \left(\widehat{n^2(n-1)^2}/[n_i^2(n_i-1)^2]\right); i = 1, \ldots, N\right],$$

$$\begin{aligned}\widehat{n^2} &= (1/N)\sum_{i=1}^N n_i^2, \\ \widehat{n^2(n-1)} &= (1/N)\sum_{i=1}^N n_i^2(n_i-1), \\ \widehat{n^2(n-1)^2} &= (1/N)\sum_{i=1}^N n_i^2(n_i-1)^2\end{aligned}$$

We now show α_N^2 remains bounded:

$$\alpha_N^2 \;=\; ||\tilde{\mathbf{\Sigma}}_N - \mathbf{\Sigma}_N||^2 = ||\mathbf{\Sigma}_N||^2 + ||\tilde{\mathbf{\Sigma}}_N||^2 - 2 < \mathbf{\Sigma}_N, \tilde{\mathbf{\Sigma}}_N >$$

So α_N^2 is composed of three quantities. It has already been shown that $||\mathbf{\Sigma}_N||^2 < \infty$ and $||\tilde{\mathbf{\Sigma}}_N||^2 < \infty$, so we now only need to show that $| < \mathbf{\Sigma}_N, \tilde{\mathbf{\Sigma}}_N > | < \infty$: $| < \mathbf{\Sigma}_N, \tilde{\mathbf{\Sigma}}_N > | \leq c||\mathbf{\Sigma}_N||||\tilde{\mathbf{\Sigma}}_N|| < \infty$. So $0 \leq \alpha_N^2 < \infty$.

Han and Song (submitted) proved that τ_N^2 is bounded and that $\tau_N^2 \xrightarrow{p} 0$.

Now we prove that δ_N^2 is bounded:

$\delta_N^2 < \infty$ is implied since $\delta_N^2 = \alpha_N^2 + \tau_N^2$, and we already showed α_N and τ_N^2 remain bounded.

We now prove $a_N^2 - \alpha_N^2$, $t_N^2 - \tau_N^2$, $\hat{t}_N^2 - \tau_N^2$, and $d_N^2 - \delta_N^2$ all converge in quadratic mean to zero as $N \to \infty$.

Proof.

We first prove $E[(d_N^2 - \delta_N^2)^2] \to 0$ as $N \to \infty$:

$$\begin{aligned} d_N^2 - \delta_N^2 \;&=\; ||C_N(\boldsymbol{\beta})||^2 + ||\tilde{C}_N(\boldsymbol{\beta})||^2 - 2 < C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - E[||C_N(\boldsymbol{\beta})||^2] - ||\tilde{\mathbf{\Sigma}}_N||^2 + \\ &\quad 2 < \mathbf{\Sigma}_N, \tilde{\mathbf{\Sigma}}_N > \\ &=\; \left(||C_N(\boldsymbol{\beta})||^2 - E[||C_N(\boldsymbol{\beta})||^2]\right) + \left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - ||\tilde{\mathbf{\Sigma}}_N||^2\right) - \\ &\quad 2\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - < \mathbf{\Sigma}_N, \tilde{\mathbf{\Sigma}}_N >\right) \end{aligned}$$

Following Han and Song (2011), we need to show that the expected value of the square of each of these three terms goes to 0 as $N \to \infty$ in order to prove $E[(d_N^2 - \delta_N^2)^2] \to 0$ as

$N \to \infty$. Han and Song (2011) proved part (i) of the Lemma that for this first term,

$$E\left[\left(||C_N(\boldsymbol{\beta})||^2 - E[||C_N(\boldsymbol{\beta})||^2]\right)^2\right] \to 0.$$

To show $E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - ||\tilde{\boldsymbol{\Sigma}}_N||^2\right)^2\right] \to 0$, we first let $\bar{x}_{ij}$ represent the i,jth element of $\tilde{C}_N(\boldsymbol{\beta})$, $i,j = 1,2,\ldots,2p$. Using the fact that $E(\bar{x}_{ij}^2) = Var(\bar{x}_{ij}) + [E(\bar{x}_{ij})]^2$, we have $E(||\tilde{C}_N(\boldsymbol{\beta})||^2) = (2p)^{-1}\sum_{i=1}^{2p}\sum_{j=1}^{2p} Var(\bar{x}_{ij}) + ||\tilde{\boldsymbol{\Sigma}}_N||^2$, and so $||\tilde{\boldsymbol{\Sigma}}_N||^2 = E(||\tilde{C}_N(\boldsymbol{\beta})||^2) - (2p)^{-1}\sum_{i=1}^{2p}\sum_{j=1}^{2p} Var(\bar{x}_{ij})$. Using this, we need to show

$$\begin{aligned}
& E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - ||\tilde{\boldsymbol{\Sigma}}_N||^2\right)^2\right] \\
= \; & E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E(||\tilde{C}_N(\boldsymbol{\beta})||^2) + (2p)^{-1}\sum_{i=1}^{2p}\sum_{j=1}^{2p} Var(\bar{x}_{ij})\right)^2\right] \\
= \; & E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E[||\tilde{C}_N(\boldsymbol{\beta})||^2]\right)^2\right] + (2p)^{-2}\left(\sum_{i=1}^{2p}\sum_{j=1}^{2p} Var(\bar{x}_{ij})\right)^2 + \\
& (2p)^{-1}\left(\sum_{i=1}^{2p}\sum_{j=1}^{2p} Var(\bar{x}_{ij})\right) E\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E[||\tilde{C}_N(\boldsymbol{\beta})||^2]\right) \to 0.
\end{aligned}$$

Using the proof Han and Song (2011) used to prove part (i) of the Lemma, we have

$$E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E[||\tilde{C}_N(\boldsymbol{\beta})||^2]\right)^2\right] \to 0.$$

This is the case since each element of $\tilde{C}_i(\boldsymbol{\beta})$ is just a finite multiple of the corresponding element in $g_i(\boldsymbol{\beta})g_i(\boldsymbol{\beta})^T$. $Var(\bar{x}_{ij}) \to 0$ as $N \to \infty$, implying $[\sum_{i=1}^{2p}\sum_{j=1}^{2p} Var(\bar{x}_{ij})]^2 \to 0$. Finally, the third term is composed of a bounded term that converges to 0 and another term that has an expectation of 0, and so $E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - ||\tilde{\boldsymbol{\Sigma}}_N||^2\right)^2\right] \to 0$.

With respect to the third term,

$$\begin{aligned} & E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >\right)^2\right] \\ = \; & E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N > + E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >]\right)^2\right] \\ = \; & E\left[\left([< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - E(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >)] + [E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >]\right)^2\right], \end{aligned}$$

and so we now show

$$E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - E(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >)\right)^2\right] \to 0,$$
$$E\left[\left(E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >\right)^2\right] \to 0,$$

and

$$E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > - E(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >)\right)\left(E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >\right)\right] \to 0$$

To prove

$$E\left[\left(||C_N(\boldsymbol{\beta})||^2 - E[||C_N(\boldsymbol{\beta})||^2]\right)^2\right] \to 0,$$

Han and Song (2011) show that

$$||C_N(\boldsymbol{\beta})||^2 = (2p)^{-1} \sum_{k=1}^{2p} \sum_{h=1}^{2p} \left(\frac{1}{N^2} \sum_{i=1}^{N} g_{ik}^2 g_{ih}^2 + \frac{1}{N^2} \sum_{i,j=1, j\neq i}^{N} g_{ik} g_{ih} g_{jk} g_{jh} \right),$$

$Var[(1/N^2)\sum_{i=1}^{N} g_{ik}^2 g_{ih}^2] \to 0$, and $Var[(1/N^2)\sum_{i,j=1,j\neq i}^{N} g_{ik} g_{ih} g_{jk} g_{jh}] \to 0$.

$< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >$ can be written similar to the expression just shown for $||C_N(\boldsymbol{\beta})||^2$, with the difference being that terms coming from $\tilde{C}_N(\boldsymbol{\beta})$ need to be multiplied by their corresponding functions of the cluster sizes, all of which are bounded. Therefore, since these

terms are bounded, this proof given by Han and Song (2011) also shows

$$E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > -E(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >)\right)^2\right] \to 0.$$

$$E\left[\left(E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >\right)^2\right] = \left(E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >\right)^2.$$

We now need to show $E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N > \to 0$. Let $\bar{x}_{ij}$ and $\bar{y}_{ij}$ represent the i,jth elements of $\tilde{C}_N(\boldsymbol{\beta})$ and $C_N(\boldsymbol{\beta})$, respectively, $i,j = 1,2,\ldots,2p$. Then

$$E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] = (2p)^{-1} E\left(\sum_{i=1}^{2p}\sum_{j=1}^{2p} \bar{x}_{ij}\bar{y}_{ij}\right),$$

and $< \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >= (2p)^{-1}\sum_{i=1}^{2p}\sum_{j=1}^{2p} \tilde{\mu}_{ij}\mu_{ij}$, where $\tilde{\mu}_{ij} = E(\bar{x}_{ij})$ and $\mu_{ij} = E(\bar{y}_{ij})$.

$Cov(\bar{x}_{ij}, \bar{y}_{ij}) = E(\bar{x}_{ij}\bar{y}_{ij}) - \tilde{\mu}_{ij}\mu_{ij}$, and therefore

$$(2p)^{-1}\sum_{i=1}^{2p}\sum_{j=1}^{2p} Cov(\bar{x}_{ij}, \bar{y}_{ij}) = E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >,$$

implying we need to show $(2p)^{-1}\sum_{i=1}^{2p}\sum_{j=1}^{2p} Cov(\bar{x}_{ij}, \bar{y}_{ij}) \to 0$. For each combination of i and j, $0 \leq |Cov(\bar{x}_{ij}, \bar{y}_{ij})| \leq \sqrt{Var(\bar{x}_{ij})Var(\bar{y}_{ij})} \to 0$, since $Var(\bar{x}_{ij}), Var(\bar{y}_{ij}) \to 0$.

The last term,

$$E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > -E(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >)\right)\left(E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] - < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >\right)\right],$$

is 0 since the expectation of the first term is 0, and the second term is bounded and converges to 0.

We now prove both $E[(t_N^2 - \tau_N^2)^2] \to 0$ and $E[(\hat{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$:

Han and Song (2011) proved that $E[(\bar{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$. The proof that $E[(t_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$ then follows from Ledoit and Wolf (2004), while the proof that $E[(\hat{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$ has already been given.

$E[(d_N^2 - \delta_N^2)^2]$, $E[(t_N^2 - \tau_N^2)^2]$, and $E[(\hat{t}_N^2 - \tau_N^2)^2]$ converge to 0 as $N \to \infty$, and therefore so does $E[(a_N^2 - \alpha_N^2)^2]$.

We now prove the two theorems similar to those given by Han and Song (2011), now based upon $\boldsymbol{S}_N^1$ and $\hat{\boldsymbol{S}}_N^1$. We note that these theorems assumes there is variation in cluster sizes, such that $C_N(\boldsymbol{\beta}) \neq \tilde{C}_N(\boldsymbol{\beta})$, as we need $\delta_N^2 > 0$. Specifically, we assume $||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2 > 0$ and $E[||\tilde{C}_N(\boldsymbol{\beta}) - C_N(\boldsymbol{\beta})||^2]$ does not converge to 0.

Theorem 2.1. For $\boldsymbol{\beta} \in \boldsymbol{B}$, $E[||\boldsymbol{S}_N^1 - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $\boldsymbol{S}_N^1 - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$.

Proof. Following Ledoit and Wolf (2004):

$$
\begin{aligned}
0 \;\leq\; & ||\boldsymbol{S}_N^1 - \boldsymbol{\Sigma}_N||^2 = ||\frac{\tau_N^2}{\delta_N^2}\tilde{\boldsymbol{\Sigma}}_N + \frac{\alpha_N^2}{\delta_N^2}C_N(\boldsymbol{\beta}) - \frac{\tau_N^2 + \alpha_N^2}{\delta_N^2}\boldsymbol{\Sigma}_N||^2 \\
= \;& ||\frac{\tau_N^2}{\delta_N^2}[\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N] + \frac{\alpha_N^2}{\delta_N^2}[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N]||^2 \\
= \;& (\frac{\tau_N^2}{\delta_N^2})^2||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2 + (\frac{\alpha_N^2}{\delta_N^2})^2||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + \\
& 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N > \\
\leq \;& (\frac{\tau_N^2}{\delta_N^2})^2||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2 + ||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >
\end{aligned}
$$

The proof will be complete if we show that these three terms have expectations that converge to 0.

First Term:

$$
\begin{aligned}
E[(\frac{\tau_N^2}{\delta_N^2})^2||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2] \;&=\; (\frac{\tau_N^2}{\delta_N^2})^2||\tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N||^2 \\
&=\; (\frac{\tau_N^2}{\delta_N^2})^2\alpha_N^2 = \tau_N^2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) \leq \tau_N^2 \xrightarrow{p} 0
\end{aligned}
$$

Second Term:

$$E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = \tau_N^2 \xrightarrow{p} 0.$$

Third Term:

$$\begin{aligned} & E[2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \tilde{\mathbf{\Sigma}}_N - \mathbf{\Sigma}_N, C_N(\boldsymbol{\beta}) - \mathbf{\Sigma}_N >] \\ = \; & 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \tilde{\mathbf{\Sigma}}_N - \mathbf{\Sigma}_N, E[C_N(\boldsymbol{\beta}) - \mathbf{\Sigma}_N] > \\ = \; & 2(\frac{\tau_N^2}{\delta_N^2})(\frac{\alpha_N^2}{\delta_N^2}) < \tilde{\mathbf{\Sigma}}_N - \mathbf{\Sigma}_N, 0 >= 0 \end{aligned}$$

Theorem 2.2. For $\boldsymbol{\beta} \in \boldsymbol{B}$, $E[||\hat{\boldsymbol{S}}_N^1 - \mathbf{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $\hat{\boldsymbol{S}}_N^1 - \mathbf{\Sigma}_N \xrightarrow{p} 0$.

Proof. Following Ledoit and Wolf (2004):

$$\begin{aligned} 0 \; \leq \; & ||\hat{\boldsymbol{S}}_N^1 - \boldsymbol{S}_N^1||^2 = |||[\frac{t_N^2}{d_N^2}\tilde{C}_N(\boldsymbol{\beta}) + \frac{a_N^2}{d_N^2}C_N(\boldsymbol{\beta})] - [\frac{\tau_N^2}{\delta_N^2}\tilde{\mathbf{\Sigma}}_N + \frac{\alpha_N^2}{\delta_N^2}C_N(\boldsymbol{\beta})]||^2 \\ = \; & ||\frac{t_N^2}{d_N^2}\tilde{C}_N(\boldsymbol{\beta}) - \frac{\tau_N^2}{\delta_N^2}\tilde{\mathbf{\Sigma}}_N + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})C_N(\boldsymbol{\beta}) + \frac{\tau_N^2}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta}) - \frac{\tau_N^2}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta})||^2 \\ = \; & ||\frac{\tau_N^2}{\delta_N^2}(\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N) + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})[C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})] + \frac{t_N^2}{d_N^2}\tilde{C}_N(\boldsymbol{\beta}) - \\ & \frac{\tau_N^2}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta}) + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})\tilde{C}_N(\boldsymbol{\beta})||^2 \\ = \; & ||\frac{\tau_N^2}{\delta_N^2}(\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N) + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})[C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})]||^2 \\ = \; & (\frac{\tau_N^2}{\delta_N^2})^2||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N||^2 + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2 + \\ & 2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2}) < \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N, C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta}) > \\ \leq \; & ||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N||^2 + (\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2 d_N^2 + \\ & 2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2}) < \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N, C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta}) > \end{aligned}$$

Now we need to show that $E[||\hat{\boldsymbol{S}}_N^1 - \boldsymbol{S}_N^1||^2] \to 0$, or that the expectations of each of the above three terms all converge to 0:

First Term: We need to show $E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N||^2] \to 0$. Following the work of Han and Song (2011) showing $\tau_N^2 \xrightarrow{p} 0$, we have $E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\mathbf{\Sigma}}_N||^2] = E(||\tilde{C}_N(\boldsymbol{\beta})||^2) - ||\tilde{C}_N(\boldsymbol{\beta})||^2 +$

$(||\tilde{C}_N(\boldsymbol{\beta})|| - ||\tilde{\boldsymbol{\Sigma}}_N||)(||\tilde{C}_N(\boldsymbol{\beta})|| + ||\tilde{\boldsymbol{\Sigma}}_N||)$. These two terms converge in probability to 0 since $E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E[||\tilde{C}_N(\boldsymbol{\beta})||^2]\right)^2\right] \to 0$, $||\tilde{C}_N(\boldsymbol{\beta})||$ and $||\tilde{\boldsymbol{\Sigma}}_N||$ remain bounded, and $|\ ||\tilde{C}_N(\boldsymbol{\beta})|| - ||\tilde{\boldsymbol{\Sigma}}_N||\ | \leq ||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N|| \xrightarrow{p} 0$.

Second Term:

$$\begin{aligned}
(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2 d_N^2 &= [\frac{a_N^4}{d_N^4} + \frac{\alpha_N^4}{\delta_N^4} - 2\frac{a_N^2\alpha_N^2}{d_N^2\delta_N^2}]d_N^2 \\
&= \frac{a_N^4}{d_N^2} + \frac{\alpha_N^4 d_N^2}{\delta_N^4} - \frac{2a_N^2\alpha_N^2}{\delta_N^2} \\
&= (a_N^4\delta_N^4 + \alpha_N^4 d_N^4 - 2a_N^2\alpha_N^2\delta_N^2 d_N^2)/(d_N^2\delta_N^4) \\
&= (a_N^2\delta_N^2 - \alpha_N^2 d_N^2)^2/(d_N^2\delta_N^4).
\end{aligned}$$

Ledoit and Wolf (2004) prove that $E[(a_N^2\delta_N^2 - \alpha_N^2 d_N^2)^2/(d_N^2\delta_N^4)] \to 0$ as $N \to \infty$.

Third Term:

$2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2}) < \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N, C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta}) >$

Using the Cauchy-Schwarz Inequality and denoting $c_N = 2(\frac{\tau_N^2}{\delta_N^2})(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})$ and $b_N =< \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N, C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta}) >$, showing $E(b_N^2) \to 0$ and $E(c_N^2) \to 0$ will prove that the expectation of this third term goes to zero as $N \to \infty$.

$E(b_N^2)$: $0 \leq E[(< \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N, C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta}) >)^2] \leq$
$E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^2||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2] \leq$ (by the Cauchy-Schwarz Inequality)
$\sqrt{E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^4]}\sqrt{E[||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^4]} \leq c\sqrt{E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^4]} \to 0$.
$\sqrt{E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^4]} \to 0$ by previous work and work from Han and Song (2011).

$E(c_N^2)$: $0 \leq 4(\frac{\tau_N^2}{\delta_N^2})^2 E[(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2] \to 0$ since $E[(\frac{a_N^2}{d_N^2} - \frac{\alpha_N^2}{\delta_N^2})^2]$ is bounded and $\frac{\tau_N^2}{\delta_N^2} \to 0$ under the assumption of a misspecified covariance structure.

We have now shown $E[||\hat{\boldsymbol{S}}_N^1 - \boldsymbol{S}_N^1||^2] \to 0$ as $N \to \infty$. Using this in conjunction with Theorem 1 and the proof given by Han and Song (2011), we have $E[||\hat{\boldsymbol{S}}_N^1 - \boldsymbol{\Sigma}_N||^2] \to 0$.

4.5.3 Proofs for Results Using $\boldsymbol{S}_N^2$ and $\hat{\boldsymbol{S}}_N^2$

We first prove that $\rho_N = \gamma_N^2/\delta_N^2$ minimizes $E[||\boldsymbol{S}_N^2 - \boldsymbol{\Sigma}_N||^2]$, related to the corresponding proof given in Ledoit and Wolf (2004):

Proof.

$$\begin{aligned}
E[||\boldsymbol{S}_N^2 - \boldsymbol{\Sigma}_N||^2] &= E[||\rho_N \tilde{C}_N(\boldsymbol{\beta}) + (1-\rho_N) C_N(\boldsymbol{\beta}) - \rho_N \boldsymbol{\Sigma}_N - (1-\rho_N)\boldsymbol{\Sigma}_N||^2] \\
&= E[||\rho_N[\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N] + (1-\rho_N)[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N]||^2] \\
&= \rho_N^2 E[||\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + (1-\rho_N)^2 E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + \\
&\quad 2\rho_N(1-\rho_N) E[< \tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >] \\
&= \rho_N^2 \alpha_N^2 + (1-\rho_N)^2 \tau_N^2 - 2\rho_N(1-\rho_N)\theta_N.
\end{aligned}$$

Now take the first derivative with respect to ρ_N and set equal to 0:

$$2\rho_N \alpha_N^2 - 2(1-\rho_N)\tau_N^2 - 2(1-2\rho_N)\theta_N = 0.$$

Solving for ρ_N, we get

$$\rho_N = \frac{\tau_N^2 + \theta_N}{\alpha_N^2 + \tau_N^2 + 2\theta_N} = \frac{\tau_N^2 + \theta_N}{\delta_N^2},$$

where $\delta_N^2 = E[||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2]$. For this value of ρ_N to give the minimum expected quadratic loss, we need $\theta_N > -0.5[\alpha_N^2 + \tau_N^2]$. The next proof given shows $0 \leq \delta_N^2 = \alpha_N^2 + \tau_N^2 + 2\theta_N$, implying $\theta_N \geq -0.5[\alpha_N^2 + \tau_N^2]$.

The following proofs are based on the Lemma given by Han and Song (2011) and its corresponding conditions. The first two parts of their Lemma are equivalent in our scenario, and we therefore omit the proofs here.

We now prove $\delta_N^2 = \tau_N^2 + \alpha_N^2 + 2\theta_N$:

Proof.

$$\begin{aligned}
\delta_N^2 &= E[||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2] = E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N + \boldsymbol{\Sigma}_N - \tilde{C}_N(\boldsymbol{\beta})||^2] \\
&= E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] + E[||\boldsymbol{\Sigma}_N - \tilde{C}_N(\boldsymbol{\beta})||^2] + \\
&\quad 2E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \tilde{C}_N(\boldsymbol{\beta}) >] \\
&= \tau_N^2 + \alpha_N^2 + 2\theta_N.
\end{aligned}$$

Theoretically, $0 \leq \rho_N \leq 1$ is not necessarily satisfied, so we now use the following constraints: Let

$$\gamma_N = \gamma_N(\tau_N^2, \theta_N, \delta_N^2) = \begin{cases} 0 & if \quad \tau_N^2 + \theta_N < 0 \\ \tau_N^2 + \theta_N & if \quad 0 \leq \tau_N^2 + \theta_N \leq \delta_N^2 \\ \delta_N^2 & if \quad \tau_N^2 + \theta_N > \delta_N^2 \end{cases}$$

and

$$\lambda_N = \lambda_N(\alpha_N^2, \theta_N, \delta_N^2) = \begin{cases} 0 & if \quad \alpha_N^2 + \theta_N < 0 \\ \alpha_N^2 + \theta_N & if \quad 0 \leq \alpha_N^2 + \theta_N \leq \delta_N^2 \\ \delta_N^2 & if \quad \alpha_N^2 + \theta_N > \delta_N^2 \end{cases}$$

Since $\delta_N^2 \geq 0$, $\alpha_N^2 \geq 0$, $\tau_N^2 \geq 0$, and $\delta_N^2 = \alpha_N^2 + \tau_N^2 + 2\theta_N$, we have $\gamma_N + \lambda_N = \delta_N^2$. We then define $\rho_N = \gamma_N/\delta_N^2$ and $(1 - \rho_N) = \lambda_N/\delta_N^2$.

We now prove δ_N^2, α_N^2, τ_N^2, θ_N, γ_N, and λ_N remain bounded, with $\tau_N^2 \to 0$, $\theta_N \to 0$, and $\gamma_N \to 0$ as $N \to \infty$:

Proof.

We now show α_N^2 remains bounded:

$$\begin{aligned}
\alpha_N^2 &= E[||\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] \\
&= ||\boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N||^2 + E[||\tilde{\boldsymbol{\Sigma}}_N - \tilde{C}_N(\boldsymbol{\beta})||^2] + 2E[< \boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N, \tilde{\boldsymbol{\Sigma}}_N - \tilde{C}_N(\boldsymbol{\beta}) >] \\
&= ||\boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N||^2 + E[||\tilde{\boldsymbol{\Sigma}}_N - \tilde{C}_N(\boldsymbol{\beta})||^2],
\end{aligned}$$

so we need to show that these two terms are bounded.

First, $||\boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N||^2 = ||\boldsymbol{\Sigma}_N||^2 + ||\tilde{\boldsymbol{\Sigma}}_N||^2 - 2 < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N >$, and we have already shown that $||\boldsymbol{\Sigma}_N||^2$ and $||\tilde{\boldsymbol{\Sigma}}_N||^2$ are bounded. Similarly, we have $| < \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N > | \leq c||\boldsymbol{\Sigma}_N||||\tilde{\boldsymbol{\Sigma}}_N|| < \infty$, so $||\boldsymbol{\Sigma}_N - \tilde{\boldsymbol{\Sigma}}_N||^2 < \infty$. Second, we have already proven $E[||\tilde{\boldsymbol{\Sigma}}_N - \tilde{C}_N(\boldsymbol{\beta})||^2] \to 0$ as $N \to \infty$, implying it remains bounded.

Han and Song (2011) proved that τ_N^2 is bounded and that $\tau_N^2 \xrightarrow{p} 0$.

We now prove that θ_N is bounded and that $\theta_N \xrightarrow{p} 0$:

$\theta_N = E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \boldsymbol{\Sigma}_N - \tilde{C}_N(\boldsymbol{\beta}) >] = -E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >]$. Before taking the expectation,

$$\begin{aligned}
& < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N > \\
= \; & < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N + \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N > \\
= \; & < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N > + < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N > .
\end{aligned}$$

Now $E[< C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N >] = < \boldsymbol{\Sigma}_N - \boldsymbol{\Sigma}_N, \tilde{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}_N >= 0$. Also, $0 \leq E[| < C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, \tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N > |] \leq cE[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||] \xrightarrow{p} 0$, since $E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] \to 0$ and $E[||\tilde{C}_N(\boldsymbol{\beta}) - \tilde{\boldsymbol{\Sigma}}_N||^2] \to 0$. Therefore, $\theta_N \xrightarrow{p} 0$.

Now we prove that δ_N^2 is bounded:

$\delta_N^2 < \infty$ is implied since $\delta_N^2 = \alpha_N^2 + \tau_N^2 + 2\theta_N$, and we already showed α_N, τ_N^2, and θ_N remain bounded.

Now we prove that λ_N and γ_N remain bounded, and that $\gamma_N \xrightarrow{p} 0$:

It follows that since $\delta_N^2 < \infty$, then λ_N and γ_N also remain bounded. We have also shown that $\theta_N \xrightarrow{p} 0$ and $\tau_N^2 \xrightarrow{p} 0$, therefore $\tau_N^2 + \theta_N \xrightarrow{p} 0$, implying $\gamma_N \xrightarrow{p} 0$.

We now prove $a_N^2 - \alpha_N^2$, $\bar{t}_N^2 - \tau_N^2$, $d_N^2 - \delta_N^2$, $\hat{\theta}_N - \theta_N$, $\hat{\gamma}_N - \gamma_N$, and $\hat{\lambda}_N - \lambda_N$ all converge in quadratic mean to zero as $N \to \infty$.

Proof.

We first prove $E[(d_N^2 - \delta_N^2)^2] \to 0$ as $N \to \infty$:

$$\begin{aligned} d_N^2 - \delta_N^2 &= ||C_N(\boldsymbol{\beta})||^2 + ||\tilde{C}_N(\boldsymbol{\beta})||^2 - 2 < C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > -E[||C_N(\boldsymbol{\beta})||^2] - \\ &\quad E[||\tilde{C}_N(\boldsymbol{\beta})||^2] + 2E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >] \\ &= \left(||C_N(\boldsymbol{\beta})||^2 - E[||C_N(\boldsymbol{\beta})||^2]\right) + \left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E[||\tilde{C}_N(\boldsymbol{\beta})||^2]\right) - \\ &\quad 2\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > -E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >]\right) \end{aligned}$$

Following Han and Song (2011), all we need to show is that the expected value of the square of each of these terms goes to 0 as $N \to \infty$ in order to prove that $E[(d_N^2 - \delta_N^2)^2] \to 0$ as $N \to \infty$. Han and Song (2011) proved part (i) of their Lemma that for this first term,

$$E\left[\left(||C_N(\boldsymbol{\beta})||^2 - E[||C_N(\boldsymbol{\beta})||^2]\right)^2\right] \to 0.$$

Similarly, we have already shown

$$E\left[\left(||\tilde{C}_N(\boldsymbol{\beta})||^2 - E[||\tilde{C}_N(\boldsymbol{\beta})||^2]\right)^2\right] \to 0$$

and

$$E\left[\left(< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) > -E[< C_N(\boldsymbol{\beta}), \tilde{C}_N(\boldsymbol{\beta}) >]\right)^2\right] \to 0.$$

Han and Song (2011) proved that $E[(\bar{t}_N^2 - \tau_N^2)^2] \to 0$ as $N \to \infty$.

We now prove $E[(\hat{\theta}_N - \theta_N)^2] \to 0$ as $N \to \infty$:

$E[(\hat{\theta}_N - \theta_N)^2] = E[\hat{\theta}_N^2] - 2\theta_N E[\hat{\theta}_N] + \theta_N^2$. We already proved $\theta_N \xrightarrow{p} 0$, implying $\theta_N^2 \xrightarrow{p} 0$. Since $\hat{\theta}_N$ is bounded and $\theta_N \xrightarrow{p} 0$, we have $2\theta_N E[\hat{\theta}_N] \xrightarrow{p} 0$. Now $\hat{\theta}_N^2 = 0.25[\frac{1}{N^2}\sum_{i=1}^{N} ||\tilde{C}_i(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2 + \bar{t}_N^2]^2$, and following the argument given by Han and Song (2011) to prove $E[(d_N^2 - \delta_N^2)^2] \to 0$, to show $E[(\hat{\theta}_N - \theta_N)^2] \to 0$ as $N \to \infty$, we need to prove $E[(\frac{1}{N^2}\sum_{i=1}^{N} ||\tilde{C}_i(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2)^2] \to 0$ and $E[(\bar{t}_N^2)^2] \to 0$ as $N \to \infty$.

From Han and Song (2011), $E[(\bar{t}_N^2)^2] \to 0$. Similarly, the proof for $E[(\frac{1}{N^2}\sum_{i=1}^{N} ||\tilde{C}_i(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2)^2] \to 0$ follows their work.

$E[(a_N^2 - \alpha_N^2)^2] \to 0$ as $N \to \infty$ is now implied since $\delta_N^2 = \alpha_N^2 + \tau_N^2 + 2\theta_N$ and $d_N^2 = a_N^2 + \bar{t}_N^2 + 2\hat{\theta}_N$.

To prove $E[(\hat{\gamma}_N - \gamma_N)^2]$,$E[(\hat{\lambda}_N - \lambda_N)^2] \to 0$ as $N \to \infty$, we only need to show one of these is true, as this implies the other since $\delta_N^2 = \lambda_N + \gamma_N$. However, we have already shown $E[(d_N^2 - \delta_N^2)^2]$, $E[(\bar{t}_N^2 - \tau_N^2)^2]$, $E[(\hat{\theta}_N - \theta_N)^2]$, and $E[(a_N^2 - \alpha_N^2)^2]$ all converge to 0 as $N \to \infty$, and since $d_N^2, \bar{t}_N^2, \hat{\theta}_N$, and a_N^2 in $\hat{\gamma}_N$ and $\hat{\lambda}_N$ match up to $\delta_N^2, \tau_N^2, \theta_N$, and α_N^2, respectively, in γ_N and λ_N, we therefore have the desired results.

We now prove the two theorems given by Han and Song (2011), but now based upon $\boldsymbol{S}_N^2$ and $\hat{\boldsymbol{S}}_N^2$. We note that these theorems assume there is variation in cluster sizes, such that $C_N(\boldsymbol{\beta}) \neq \tilde{C}_N(\boldsymbol{\beta})$, as we need $\delta_N^2 > 0$. Specifically, we assume $E[||\tilde{C}_N(\boldsymbol{\beta}) - C_N(\boldsymbol{\beta})||^2]$ and $||\tilde{C}_N(\boldsymbol{\beta}) - C_N(\boldsymbol{\beta})||^2$ are always greater than 0.

Theorem 3.1. For $\boldsymbol{\beta} \in \boldsymbol{B}$, $E[||\boldsymbol{S}_N^2 - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $\boldsymbol{S}_N^2 - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$.

Proof. Following Ledoit and Wolf (2004):

$$\begin{aligned}
0 \leq ||\boldsymbol{S}_N^2 - \boldsymbol{\Sigma}_N||^2 &= ||\frac{\gamma_N}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta}) + \frac{\lambda_N}{\delta_N^2}C_N(\boldsymbol{\beta}) - \frac{\gamma_N + \lambda_N}{\delta_N^2}\boldsymbol{\Sigma}_N||^2 \\
&= ||\frac{\gamma_N}{\delta_N^2}[\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N] + \frac{\lambda_N}{\delta_N^2}[C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N]||^2 \\
&= (\frac{\gamma_N}{\delta_N^2})^2||\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + (\frac{\lambda_N}{\delta_N^2})^2||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + \\
&\quad 2(\frac{\gamma_N}{\delta_N^2})(\frac{\lambda_N}{\delta_N^2}) < \tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N > \\
&\leq \frac{\gamma_N}{\delta_N^2}||\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + ||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2 + \\
&\quad 2(\frac{\gamma_N}{\delta_N^2})(\frac{\lambda_N}{\delta_N^2}) < \tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >
\end{aligned}$$

The proof will be complete if we show that these three terms have expectations that converge to 0.

First Term: $0 \leq E[||\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = \alpha_N^2 < \infty$, which we proved earlier.

$$(\frac{\gamma_N}{\delta_N^2})^2 E[||\tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = (\frac{\gamma_N}{\delta_N^2})^2\alpha_N^2 = \gamma_N\frac{\gamma_N}{\delta_N^2}\frac{\alpha_N^2}{\delta_N^2} \leq c\gamma_N \to 0$$

Second Term:

$$E[||C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N||^2] = \tau_N^2 \xrightarrow{p} 0$$

Third Term: $2(\frac{\gamma_N}{\delta_N^2})(\frac{\lambda_N}{\delta_N^2})E[< \tilde{C}_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N, C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N >] = -(\frac{\gamma_N}{\delta_N^2})(\frac{\lambda_N}{\delta_N^2})\theta_N \xrightarrow{p} 0$

Theorem 3.2. For $\boldsymbol{\beta} \in \boldsymbol{B}$, $E[||\hat{\boldsymbol{S}}_N^2 - \boldsymbol{\Sigma}_N||^2] \to 0$ as $N \to \infty$, implying $\hat{\boldsymbol{S}}_N^2 - \boldsymbol{\Sigma}_N \xrightarrow{p} 0$.

Proof. Following Ledoit and Wolf (2004):

$$\begin{aligned}
0 \leq ||\hat{\boldsymbol{S}}_N^2 - \boldsymbol{S}_N^2||^2 &= ||(\frac{\hat{\gamma}_N}{d_N^2} - \frac{\gamma_N}{\delta_N^2})\tilde{C}_N(\boldsymbol{\beta}) + (\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})C_N(\boldsymbol{\beta}) + \frac{\gamma_N}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta}) - \frac{\gamma_N}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta})||^2 \\
&= ||(\frac{\gamma_N}{\delta_N^2} - \frac{\gamma_N}{\delta_N^2})\tilde{C}_N(\boldsymbol{\beta}) + (\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})[C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})] + \\
&\quad \frac{\hat{\gamma}_N}{d_N^2}\tilde{C}_N(\boldsymbol{\beta}) - \frac{\gamma_N}{\delta_N^2}\tilde{C}_N(\boldsymbol{\beta}) + (\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})\tilde{C}_N(\boldsymbol{\beta})||^2 \\
&= ||0 + (\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})[C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})] + 0||^2 \\
&= (\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})^2||C_N(\boldsymbol{\beta}) - \tilde{C}_N(\boldsymbol{\beta})||^2 = (\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})^2 d_N^2
\end{aligned}$$

We now need to show that $E[(\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})^2 d_N^2] \to 0$ as $N \to \infty$. The following will be used for Lemma A.1 in Ledoit and Wolf (2004):

$$\begin{aligned}
(\frac{\hat{\lambda}_N}{d_N^2} - \frac{\lambda_N}{\delta_N^2})^2 d_N^2 &= \left[\frac{\hat{\lambda}_N^2}{d_N^4} + \frac{\lambda_N^2}{\delta_N^4} - 2\frac{\hat{\lambda}_N\lambda_N}{d_N^2\delta_N^2}\right] d_N^2 \\
&= (\hat{\lambda}_N^2\delta_N^4 + \lambda_N^2 d_N^4 - 2\hat{\lambda}_N\lambda_N d_N^2\delta_N^2)/(d_N^2\delta_N^4) \\
&= (\hat{\lambda}_N\delta_N^2 - \lambda_N d_N^2)^2/(d_N^2\delta_N^4)
\end{aligned}$$

Ledoit and Wolf (2004) prove that $E[(a_N^2\delta_N^2 - \alpha_N^2 d_N^2)^2/(d_N^2\delta_N^4)] \to 0$ as $N \to \infty$. By simply replacing a_N^2 and α_N^2 with $\hat{\lambda}_N$ and λ_N, respectively, in their proof, we have $E[(\hat{\lambda}_N\delta_N^2 - \lambda_N d_N^2)^2/(d_N^2\delta_N^4)] \to 0$ as $N \to \infty$.

We have now shown $E[||\hat{\boldsymbol{S}}_N^2 - \boldsymbol{S}_N^2||^2] \to 0$ as $N \to \infty$. Using this in conjunction with Theorem 1 and the proof given by Han and Song (2011), we have $E[||\hat{\boldsymbol{S}}_N^2 - \boldsymbol{\Sigma}_N||^2] \to 0$.

CHAPTER V

Summary

This dissertation studied the small-sample deficiencies of two different popular techniques for statistical inference when using a marginal model, and proposed modifications to improve their performances. Particularly, our topics of interest were test size and estimation variability. Additionally, an applied focus was directed toward GRT settings with binary data, as demonstrated by the breast screening and CHD studies, but attention was also given to marginal models in general repeated measures settings, such as in the AIDS study example.

With respect to sub-optimal inference performance, a Wald statistic in the settings of Chapter 2 does not necessarily have a standard normal distribution when the number of independent clusters is small, leading to a decreased test size, and therefore diminished power. Additionally, the estimating equations for QIF were shown to be in a different class than GEE when using an exchangeable correlation structure and clusters vary in size, implying QIF does not necessarily have an efficiency advantage over GEE in this situation. However, even after modifying QIF's estimating equations to be within the same class as GEE, corresponding parameter estimates could still contain greater variability than the corresponding estimates resulting from the use of GEE. This inferior performance was especially evident in GRT scenarios, but was also seen in repeated measures designs and was not restricted to a working exchangeable correlation structure.

To improve inference with the Wald statistic, we proposed a modified standard error,

creating a pseudo-Wald statistic, $\widetilde{W}_{1.5}$, which led to test sizes at the nominal value when the true exchangeable correlation, or ICC, was known. We also suggested two techniques to yield nominal sizes when estimating the correlation parameter, which needs to be done in practice. The first uses $\widetilde{W}_{1.5}$ and critical values from a t-distribution with degrees of freedom equalling the number of clusters in the study, and the second uses an inflated pseudo-standard error, via leverage values, inside $\widetilde{W}_{1.5}$ and standard normal critical values.

To improve inference with QIF, we proposed multiple different weighting matrices to use in place of the extended score equations' empirical covariance matrix, $C_N(\boldsymbol{\beta})$. The majority of these matrices were based on the weighted combination of $C_N(\boldsymbol{\beta})$ and either the model-based covariance matrix, $\boldsymbol{M}_N$, or another empirical matrix we proposed in order to address cluster size variation, denoted as $\tilde{C}_N(\boldsymbol{\beta})$. These weighted combinations were based on minimizing the expected quadratic loss of the resulting weighting matrix, which is asymptotically optimal. Simulations showed that the weighted combination of $C_N(\boldsymbol{\beta})$ and $\boldsymbol{M}_N$, defined as QIF2 in this dissertation, typically improved QIF's estimation performance to approximately the level of GEE. However, QIF2 and QIF worked similarly when GEE produced estimates having the greatest variability.

Although the proposed pseudo-Wald test yields sizes at the nominal value when using a correctly specified correlation, future work is needed to compare the sensitivities of $\widetilde{W}_{1.5}$ and the Wald statistic using the bias-corrected standard error to misspecified correlations, such as incorrectly assuming the ICC is constant across clusters. The power of each test statistic should be studied as well. Future work should also deal with extending our proposed standard error modification for implementation when additional covariates are used in the regression model, and study is needed to determine if the issue of non-nominal test size occurs only when outcomes are binary in nature.

With respect to QIF2, future work is needed to determine the validity of the corresponding quadratic inference function as a test statistic in small-sample settings, as $\boldsymbol{M}_N$ will be biased when implementing an incorrect covariance structure. Additionally, the sensitivity of

QIF2 to outliers, as compared with GEE and QIF, needs to be studied. Use of $\boldsymbol{M}_N$ will also effect the validity of standard error estimates produced by QIF2, and therefore future work is needed to determine an appropriate method for obtaining standard errors with good properties that lead to valid Wald tests for any sample size.

Finally, the applied importance of our work is that it provides insight into the potential problems with the methods of focus, and proposes corresponding remedies to improve statistical inference. Typically, studies are very costly to carry out, especially GRTs. Increasing size to its nominal value for the Wald test inherently increases the power of the study, which is important since the true statistical power for GRTs can be quite small to begin with. Additionally, estimates with increased variability lead to a study with less reliable results. This not only influences hypothesis testing, but also population-average estimates that will be reported, justifying the importance of improving QIF's performance.

BIBLIOGRAPHY

BIBLIOGRAPHY

Atri, J., Falshaw, M., Gregg, R., Robson, J., Omar, R. Z., Dixon, S. (1997). Improving uptake of breast screening in multiethnic populations: a randomised controlled trial using practice reception staff to contact non-attenders. *British Medical Journal* **315**:1356–1359.

Bull, S. B. and Greenwood, C. M. T. (1997). Jackknife bias reduction for polychotomous logistic regression. *Statistics in Medicine* **16,** 545–560.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80,** 517–526.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The Annals of Statistics* **10,** 429–441.

Cordeiro, G. M. and Demetrio, C. G. B. (2008). Corrected estimators in extended quasi-likelihood models. *Communications in Statistics - Theory and Methods* **37,** 873–880.

Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B* **53,** 629–643.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B* **30,** 248–275.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82,** 1079–1091.

Donner, A. and Donald, A. (1988). The statistical analysis of multiple binary measurements. *Journal of Clinical Epidemiology* **41,** 899–905.

Drum, M. and McCullagh, P. (1993). Comment on 'regression models for discrete longitudinal responses'. *Statistical Science* **8,** 300–301.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics* **57,** 1198–1206.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80,** 27–38.

Han, P. and Song, P. X.-K. (2011). A note on improving quadratic inference functions using a linear shrinkage approach. *Statistics and Probability Letters* **81,** 438–445.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50,** 1029–1054.

Hauck, W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* **72,** 851–853.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89,** 111–128.

Jung, S. H., Kang, S. H., and Ahn, C. (2001). Sample size calculations for clustered binary data. *Statistics in Medicine* **20,** 1971–1982.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. (1987). The multicenter aids cohort study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology* **126,** 310–318.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96,**:1387–1396.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis* **9,** 137–163.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88,** 365–411.

Liang, K.-Y. and Hanfelt, J. (1994). On the use of the quasi-likelihood method in teratological experiments. *Biometrics* **50,** 872–880.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lindsay, B. G. (1982). Conditional score functions: Some optimality results. *Biometrika* **69,** 503–512.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57,** 126–134.

McCaffrey, D. F. and Bell, R. M. (2006). Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters. *Statistics in Medicine* **25,** 4081–4098.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edn. Chapman and Hall: London, 1989.

Morel, J. G., Bokossa, M. C., and Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal* **45,** 395–409.

Murray, D. M., Varnell, S. P., and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* **94,** 423–432.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74,** 221–232.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.

Pan, W. (2001). On the robust variance estimator in generalised estimating equations. *Biometrika* **88,** 901–906.

Pan, W. and Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* **21,** 429–1441.

Pilla, R. S. and Loader, C. (2006). On large-sample estimation and testing via quadratic inference functions for correlated data. *Available online at http://arxiv.org/abs/math/0505360*, 1–32.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* **81,** 321–327.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* **90,** 455–463.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics* **62,** 379–391.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87,** 823–836.

Qu, A. and Song, P. X.-K. (2004). Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika* **91,** 447–459.

Reed, J. F. (2000). Eliminating bias in randomized cluster trials with correlated binomial outcomes. *Computer Methods and Programs in Biomedicine* **61,** 119–123.

Ridout, M. S., Demetrio, C. G. B., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55,** 137–148.

Small, C. G. and McLeish, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. New York: Wiley.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. New York: Springer.

Song, P. X.-K., Jiang, Z., Park, E., and Qu, A. (2009). Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine* **28,** 3683–3696.

Turner, R. M., Omar, R. Z., and Thompson, S. G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* **20,** 453–472.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61,** 439–447.

Williams, D. A. (1982). Extra-binomial variation in logistic-linear models. *Journal of the Royal Statistical Society, Series C* **31,** 144–148.

Windmeijer, F. (2000). A finite sample correction for the variance of linear two-step GMM estimators. Institute for Fiscal Studies Working Paper Series No. W00/19, London.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* **126,** 25–51.

Yudkin, P. L. and Moher, M. (2001). Putting theory into practice: a cluster randomized trial with a small number of clusters. *Statistics in Medicine* **20,** 341–349.