Robust Methods for Estimating the Intraclass Correlation Coefficient and for Analyzing Recurrent Event Data

Robust Methods for Estimating the Intraclass Correlation Coefficient and for Analyzing Recurrent Event Data

A dissertation presented by

Tom Chen

to

The Department of Biostatistics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biostatistics

Harvard University
Cambridge, Massachusetts
September 2018

Dissertation Advisors: Tom Chen
Rui Wang
Eric Tchetgen Tchetgen

Robust Methods for Estimating the Intraclass Correlation Coefficient and for Analyzing Recurrent Event Data

Abstract

Robust statistics have emerged as a family of theories and techniques for estimating parameters of a model while dealing with deviations from idealized assumptions. Examples of deviations include misspecification of parametric assumptions or missing data; while there are others, these two deviations are the focus of this work. When unaccounted for, naive analysis with existing techniques may lead to biased estimators and/or undercovered confidence intervals.

At the same time, research poured into clustered/correlated data is extensive and a large body of methods have been developed. Many works have already connected topics within robust statistics and correlated data, but a plethora of open problems remain. This dissertation investigates a few of these open problems.

Chapter 1 combines second-order generalized estimating equations (GEE2), inverse probability weighting (IPW), and semiparametric theory in order to estimate the intraclass correlation coefficient (ICC) in the presence of informative missing data.

Chapter 2 approaches linear models with correlated outcomes from the mixed models (MM) perspective instead of GEE. In addition to the estimation of 2nd moments, this framework also allows estimation of the skewness and kurtosis of the distributions of the random effects/subject-specific error terms and tests for normality in both the random effects and the error terms.

Chapter 3 addresses analytical challenges in the unique structure of "evolving clustered randomized trials" in HIV prevention trials. In evolving CRTs, subjects are socially/sexually linked to an index partner, provided intervention based on the randomized arm this index partner is assigned to, and followed until HIV infection occurs or the end of

study. We view phylogenetically-linked partners over time as recurrent events to the index and assess the intervention effect through the use of recurrent event analysis. However, subjects may refuse to participate or drop-out, leading to a statistical problem of potentially informative missing and/or censored events in a recurrent event process. We address this issue with embedding IPW within the recurrent event estimating equations.

Contents

1 A stochastic second-order generalized estimating equations approach for estimating association parameters in the presence of informative missingness **1**

1.1 Notation and standard GEE2 4

1.2 Proposal I: Stochastic GEE2 6

1.2.1 Background: Robbins-Monro Algorithm 6

1.2.2 Subsampling 7

1.2.3 Exploiting sparsity 8

1.2.4 Par-S-GEE2 10

1.3 Proposal II: IPW & DR GEE2 11

1.3.1 IPW-GEE2 11

1.3.2 DR-GEE2 12

1.3.3 Inference 13

1.3.4 Embedding IPW-GEE2 and DR-GEE2 with S-GEE2 13

1.4 Simulation 14

1.4.1 Consistency and efficiency of IPW-GEE2 & DR-GEE2 schemes . . 16

1.4.2 Algorithmic Characteristic of DR-GEE2 vs S-DR-GEE2 21

1.5 Application to Sanitation Data 23

1.6 Discussion 26

2 Linear mixed effects models with Fleishman-distributed variance compo-

nents **28**

2.1 Introduction 29

2.2 Proposal I: REMM-F for non-normal mixed models 32

2.2.1 Background: Fleishman distribution 32

2.2.2 Notation and Methods 33

2.2.3 Testing normality of U_i and ϵ_{ij} 39

2.3 Simulation 41

2.4 Proposal II: QN-ICC for agreement studies 45

2.5 Application to CHAT Signal Data 48

2.6 Discussion 53

3 Robust Estimation of Recurrent Event Mean Functions in the Presence of Informative Event Censoring **55**

3.1 Introduction 55

3.2 Methods 59

3.2.1 Data Structure & Model Set-Up 59

3.2.2 Recurrent Events with No Missing Events 60

3.2.3 Recurrent Events with Inverse-Probability Weighted Events 60

3.3 Simulations 63

3.4 Discussion 66

References **69**

Appendices **81**

A Appendix for Chapter 1 **81**

A.1 Pseudocode for Stochastic Algorithms 81

A.2 Time Complexity Proofs 83

A.3 Proof of CAN for DR estimator 90

List of Figures

2.1 Skewness-elongation bounds and locations of select distributions within the skewness-elongation plane. Severe platykurtic distributions, such as the uniform distribution, fall under the Fleishman bound. 36

2.2 Visualization of KDE-boot in action. The shaded dark gray area represents regions of more extreme values than our test statistic under H_0, which be the integration region in calculating the p-value. 40

2.3 + REMM-F × Bootstrap REMM-N ⋄ Smith ∇GEE2 Empirical coverage levels and lengths for each of the ICC confidence interval methods under several scenarios, averaged over 1500 replicate simulations. 42

2.4 + KDE-boot × Shapiro-Wilk ⋄ Anderson-Darling Type I error rates (when distribution is normal) or power (when distribution is non-normal) for each normality test under several scenarios, averaged over 1500 replicate simulations. Dashed lines indicate nominal Type I error rate $\alpha = 0.05$, while dashed-dot lines indicate maximum power $\beta = 1$. 44

2.5 $+U_i$ Skewness × ϵ_{ij} Skewness ⋄ U_i Kurtosis ∇ϵ_{ij} Kurtosis Empirical coverage levels and lengths for U_i, ϵ_{ij} skewness/kurtosis combinations under several scenarios, averaged over 1500 replicate simulations. 46

2.6 Scatterplot matrix displaying pairwise scatterplots of log spectral wave density among the six available channels on the off-diagonals and kernel density plots of log spectral density for each channel on the diagonal. Note the unnormalized density plots exhibit right-skewness for each channel, even so after a log transform. The normalized density plots are uniformly distributed and should resemble a rectangle, but the nature of Gaussian kernels always give density to points outside a finite support. 49

3.1 Four types of data scenarios in the RIING study 57

3.2 Growth curves for treatment and control groups. The complete-case and IPW fittings shown above are a single draw from the 1000 replicate simulations, shown for demonstration purposes. 65

List of Tables

1.1 Time complexities for S-GEE2 algorithms under various working covariance structures. 9

1.2 Information regarding the generation process 16

1.3 Biases & Standard Errors from 1000 replicate simulations with both Y_{ij}, R_{ij} simulated with Parzen's method. 18

1.4 Biases & Standard Errors from 1000 replicate simulations with R_{ij} simulated using Parzen's method and Y_{ij} simulated using random-intercept method. . 20

1.5 Comparison of statistical characteristics of full DR-GEE2 vs S-DR-GEE2. $\mathcal{R} = 2000$ replicate simulations. 22

1.6 Algorithmic analysis of standard and stochastic DR-GEE2. $\mathcal{R} = 2000$ replicate simulations. Run-time values are computed on runs which converged. The conditional TM error is the error rate among simulations whence PSM and OM converged. † Each replicate simulation was executed in R on a dual-core node on the Orchestra cluster supported by the Harvard Medical School Research Information Technology Group. 22

1.7 Effects of the supply side-market vs. control on the probability of hygienic latrine ownership in the sanitation data analysis [Guiteras et al., 2015] using the complete-case GEE2, IPW-GEE2 adjustment (non-adjusting and adjusting for missingness ICC), and DR-GEE2, assuming outcomes are rMAR. * Fitted with 50 parallel stochastic GEE2, and averaging convergent estimates. Reported are median times among convergent estimates. † Executed in R on a desktop with Intel(R) Core(TM) i5-4460 CPU 3.20GHz 25

2.1 Analysis results of CHAT EEG log spectral density with Studentized values within columns. Top panel presents point and 95% CI's for the ICC; middle panel presents p-values from normality tests on U_i and ϵ_{ij}; bottom panel presents point and 95% CI's for the skewness and kurtosis of U_i and ϵ_{ij}. Each column represents a specific EEG wave. 50

2.2 Analysis results of CHAT EEG log spectral density with uniform QN. Top panel presents point and 95% CI's for the ICC; middle panel presents p-values from normality tests on U_i and ϵ_{ij}; bottom panel presents point and 95% CI's for the skewness and kurtosis of U_i and ϵ_{ij}. Each column represents a specific EEG wave. 51

2.3 Analysis results of CHAT EEG log spectral density with normal QN. Top panel presents point and 95% CI's for the ICC; middle panel presents p-values from normality tests on U_i and ϵ_{ij}; bottom panel presents point and 95% CI's for the skewness and kurtosis of U_i and ϵ_{ij}. Each column represents a specific EEG wave. 52

3.1 Bias & coverage for the crude and IPW estimators for growth curves in each intervention arm, and their differences. Number of replicate simulations = 1000. 66

Acknowledgments

I cannot express enough gratitude to all those involved in my journey. Above all, I want to thank my advisor Dr. Rui Wang, who has always offered her guidance and support throughout my doctoral studies. She has molded me into the researcher that I am today. Without her advice and insights, it would be impossible for me to complete this dissertation. Her advice extends beyond research into life as well.

I am grateful to have Dr. Eric Tchetgen Tchetgen and Dr. Victor DeGruttola serving as my committee members and providing thoughtful and invaluable comments. My PhD studies would have also been impossible without the help of department faculty and staff, who are the silent force in the background propelling my educational development.

Many thanks go to my friends for providing support, encouragement, and amusement, especially my wife Xuan Zhang, who has always been there to inspire me, listen to me, and write code for me whenever I can't find a solution on StackExchange. Last but certainly not least, I would like to thank my family, who always will be the foundation of my achievements.

Chapter 1

A stochastic second-order generalized estimating equations approach for estimating association parameters in the presence of informative missingness

Cluster randomized trials (CRTs), in which individuals are randomly assigned to intervention in groups, have been increasingly implemented to evaluate efficacy and effectiveness of various intervention programs. The intraclass correlation coefficient (ICC) characterizes the degree of similarity of individuals within a community and is crucial in accurately computing sample sizes needed to achieve a certain power level in a CRT. The statistical power and required sample size for a CRT can change substantially depending on the ICC. For example, in a matched-pair CRT with 15 pairs and a sample size of 300 within each cluster as in the Botswana Combination Prevention Project (BCPP) [Gaolathe et al., 2016; Wang et al., 2014], the power to detect a 40% reduction in 3-year cumulative incidence from 2.5% to 1.5% decreases from 80% to 52% as the ICC increases from 0.001 to 0.005. To achieve 80% power with an ICC of 0.005, assuming all else being fixed, the number of clusters required almost doubles (15 pairs to 27 pairs). When analyzing data from CRTs, a commonly used and robust approach is based on comparisons of a community-level measure of the end of interest. Tests constructed by giving equal weight to each cluster may

not be fully efficient, especially when the sizes of clusters vary substantially. The optimal weights depend crucially on the ICC for both parametric test (e.g., t-test) [Hayes and Moulton, 2009] and nonparametric permutation tests [Braun and Feng, 2001; Wang and DeGruttola, 2017]. Despite its importance, obtaining reliable estimates of ICC remains a major problem in designing CRTs [Donner and Klar, 2000; Hayes and Bennett, 1999; Klar and Donner, 2001]. Furthermore, ICC can vary considerably by intervention group and community characteristics (e.g. community size) [Crespi et al., 2009; Wu et al., 2012].

In CRTs, interest often lies in estimating the causal effect of intervention on the cluster – the difference between outcomes for the treated cluster vs the untreated cluster [Carnegie et al., 2016]. Mixed models and their extensions [Anderson and Aitkin, 1988; Laird and Ware, 1982b; Stiratelli et al., 1984] are a popular choice in estimating the causal effect accounting for the clustering effect. Variance modeling for continuous outcomes have also been proposed [Leckie et al., 2014; Paik, 1992; West et al., 2018]. These methods require parametric assumptions and provide a conditional interpretation of causal effects. Here we focus on the generalized estimating equations (GEE) [Liang and Zeger, 1986] approach. This estimation procedure is semiparametric in that it assumes correct specification of a marginal mean model, instead of a full likelihood, in order to yield a consistent and asymptotically normal (CAN) estimator of the treatment effect; while the correct specification of the correlation structure may improve efficiency, its misspecification does not affect valid inference [Zeger et al., 1988]. As a result of this flexible feature, one typically uses a rough approximation of the ICC through a method of moments approach. When ICC itself is of primary interest, the method of moments approach is inefficient and inaccurate. This gives rise to second-order generalized estimating equations (GEE2) [Liang and Zeger, 1992; Zhao and Prentice, 1990], which includes an extra stack of estimating equations specifically directed at estimating the ICC.

Several authors [Carey et al., 1993; Ziegler et al., 1998] have noted of convergence problems regarding GEE2s, and we later demonstrate similar problems as well. GEE2 are notoriously hard to solve due to the far larger stack of estimating equations for the association parameters, leading to excessive computing time for obtaining solutions to these

equations. Carey et al. [1993] championed the use of odds ratios in measuring association, for which they propose fitting with alternating logistic regression, which is computationally efficient but very specific to just the odds ratio setting. In our context, the ICC is a means to an end (i.e. sample size calculations), which the odds ratio is unsuited for. We develop stochastic methods to alleviate the computational challenges associated with solving GEE2. These stochastic algorithms involve running Fisher scoring / Newton-Raphson on a different subset (minibatch) of the data at each iteration, in the spirit of minibatch stochastic gradient descent (mbSGD) and the more general class of Robbins-Monro (RM) algorithms. Under mild regularity conditions [Blum, 1954], the algorithm almost surely converges to the same solution as if we performed standard Fisher scoring on GEE2.

It is common to encounter missing outcomes in practice. A second purpose of this paper is to investigate methods in accounting for missing outcome data. When outcomes are assumed missing completely at random [Rubin, 1976] (MCAR; the outcomes are missing independently of both observed and unobserved data), GEE2 analysis performed on complete-case CRT data provides CAN estimators for the treatment and ICC parameters. In the case of missing at random (MAR; outcome missingness is independent of the unobserved variables conditional on the observed variables), GEE2 produces inconsistent estimates unless all factors contributing to the propensity of being missing are included in a correctly-specified outcome model.

Currently, methods are available to account for a restricted missing at random mechanism (rMAR; outcome missingness depends only on observed covariates but not on observed outcomes) in the GEE1 case for the estimation of marginal treatment effects [Prague et al., 2016]. The strategy is based on the standard inverse-probability weighting (IPW) and outcome model (OM) augmentation which has been made popular by the semiparametric methods field [Robins et al., 1994; Tsiatis, 2007b; Van der Laan and Robins, 2003]. The resulting estimators are doubly-robust (DR) in the sense that they only require either the IPW model or OM to be correctly specified in order to produce consistent treatment effect estimates. However, properly incorporating this framework for association parameters requires modeling the correlation among missingness indicators for correlated units within

a cluster, a potential complication which to the best of our knowledge has previously not been addressed in the literature on semiparametric methods for missing clustered data. In the context of CRTs, there is no natural ordering of the outcomes within a community and the missingness pattern is non-monotone, making the problem much more intractable [Tsiatis, 2007b].

Section 1.1 gives background on the standard GEE2 (non-stochastic and no missing data). In Section 1.2, we introduce the RM algorithm and expand on the stochastic paradigm to model fitting, and adapt this approach to fitting GEE2, which we coin as stochastic GEE2 (S-GEE2). Issues such as computational complexity, efficient implementation, and parallelization as a further mechanism in reducing computing time and numerical errors are explored here. In Section 1.3, we draw from the semiparametric methods discipline to construct IPW and DR variants of GEE2, which we call IPW-GEE2 and DR-GEE2, and explain how to further adapt these procedures with stochastic GEE2. We evaluate the performance of the proposed estimators and the proposed computational algorithms with simulations in Section 1.4 and apply the new estimators and algorithms to analyze Bangladeshi sanitation data [Guiteras et al., 2015] in Section 1.5. We end with a discussion in Section 1.6. Proofs are relegated to Appendix.

1.1 Notation and standard GEE2

Henceforth, we work with binary outcomes $Y_{ij} \in \{0,1\}$ for subject $j = 1, \cdots, n_i$ in cluster $i = 1, \cdots, I$; the framework is readily generalizable to continuous outcomes. Let $A_i \in \{0,1\}$ denote the treatment randomized at the cluster level with $\mathbb{P}(A_i = 1) = p_A$, $\mathbf{Z}_i \in \mathbb{R}^q$ as baseline cluster-level covariates, $\mathbf{X}_{ij} \in \mathbb{R}^m$ as baseline subject-level covariates, and $\mathbf{X}_i = \{\mathbf{X}_{ij}\}_{j=1}^{n_i}$. We denote $P(\cdot)$ as the probability measure associated with the argument i.e. $P(a), P(\mathbf{z}, \mathbf{x})$. Let $\pi_{ij} = \mathbb{E}[Y_{ij}|A_i, \mathbf{Z}_i, \mathbf{X}_i]$ denote the conditional mean outcome and

$$\rho_{ijj'} \stackrel{\text{def}}{=} \text{Cov}(Y_{ij}, Y_{ij'}|A_i, \mathbf{Z}_i, \mathbf{X}_i) \Big/ \sqrt{\text{Var}(Y_{ij}|A_i, \mathbf{Z}_i, \mathbf{X}_i)\text{Var}(Y_{ij'}|A_i, \mathbf{Z}_i, \mathbf{X}_i)}$$

denote the conditional ICC. The quantities of interest are $\pi_i^* = \mathbb{E}[Y_{ij}|A_i]$ and $\rho_i^* = \text{Corr}(Y_{ij}, Y_{ij'}|A_i)$, which are the treatment-specific mean outcome and ICC. It is clear that π_i^* is a marginalization of π_{ij} in the sense that $\pi_i^* = \mathbb{E}[\pi_{ij}|A_i] = \int \pi_{ij} dP(\mathbf{z}_i, \mathbf{x}_i)$. But, $\rho_i^* \neq \mathbb{E}[\rho_{ijj'}|A_i]$ in general. Indeed, it is easy to confirm that $\rho_i^* = \mathbb{E}[\rho_{ijj'}^\dagger|A_i]$, where

$$\rho_{ijj'}^\dagger \stackrel{\text{def}}{=} \mathbb{E}\left[\frac{(Y_{ij} - \pi_i^*)(Y_{ij'} - \pi_i^*)}{\pi_i^*(1-\pi_i^*)}\middle| A_i, \mathbf{Z}_i, \mathbf{X}_i\right] = \frac{(\pi_{ij} - \pi_i^*)(\pi_{ij'} - \pi_i^*) + \rho_{ijj'}\sqrt{\mathcal{V}_{ijj'}}}{\pi_i^*(1-\pi_i^*)} \quad (1.1)$$

where $\mathcal{V}_{ijj'} = \pi_{ij}(1-\pi_{ij})\pi_{ij'}(1-\pi_{ij'})$.

Let $\widehat{\pi}_{ij}$ be an estimator of π_{ij}, converging to the limit $\overline{\pi}_{ij}$, which may or may not equal the true π_{ij}. Likewise, define $\widehat{\rho}_{ijj'}$ and $\overline{\rho}_{ijj'}$. Standard models for $\widehat{\pi}_{ij}$ include logistic or probit regression, while a model for $\widehat{\rho}_{ijj'}$ could be a generalized linear model with link function $g(x) = \text{atanh}(x)$, the Fisher z-transform. The Fisher z-transform is commonly used as a variance-stabilizing transformation for the sample correlation coefficient, but we apply it here to map the $[-1, 1]$ support of ρ_i^* onto $\mathbb{R}$.

Similarly, let $\widehat{\pi}_i^*$ and $\widehat{\rho}_i^*$ be estimators for π_i^* and ρ_i^* with limits $\overline{\pi}_i^*$ and $\overline{\rho}_i^*$, respectively. For example, inference for the causal effect of A_i can be estimated under the model

$$\text{logit}(\pi_i^*(\boldsymbol{\beta}_Y^*; A_i)) = \beta_{0Y}^* + \beta_{AY}^* A_i, \qquad \text{atanh}(\rho_i^*(\boldsymbol{\alpha}_Y^*; A_i)) = \alpha_{0Y}^* + \alpha_{AY}^* A_i \quad (1.2)$$

to produce estimators $(\widehat{\boldsymbol{\beta}}_Y^*, \widehat{\boldsymbol{\alpha}}_Y^*)$. Eq 1.2 will be referred to as the canonical treatment model (TM). In the absence of missing data, and since A_i is binary, the canonical TM is guaranteed to yield consistent $\overline{\pi}_i^* = \pi_i^*$ and $\overline{\rho}_i^* = \rho_i^*$. In the standard GEE2 framework, we would estimate $(\widehat{\boldsymbol{\beta}}_Y^*, \widehat{\boldsymbol{\alpha}}_Y^*)$ as the solution to the equations

$$\mathbf{0} = \sum_{i=1}^{I} D_i^\intercal V_i^{-1} E_i \stackrel{\text{def}}{=} \sum_{i=1}^{I} \mathbf{S}_i^Y(A_i, \boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*) \quad (1.3)$$

where

$$D_i = \frac{\partial(\boldsymbol{\pi}_i^*(\boldsymbol{\beta}_Y^*; A_i), \mathbf{r}_i^*(\boldsymbol{\alpha}_Y^*; A_i))}{\partial(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)^\intercal}, \quad V_i = \text{Cov}\begin{pmatrix}\mathbf{Y}_i \\ \mathcal{E}(\mathbf{Y}_i)\end{pmatrix}, \quad E_i = \begin{pmatrix}\mathbf{Y}_i - \boldsymbol{\pi}_i^*(\boldsymbol{\beta}_Y^*) \\ \mathcal{E}(\mathbf{Y}_i) - \mathbf{r}_i^*(\boldsymbol{\alpha}_Y^*)\end{pmatrix}$$

$$\mathcal{E}(\mathbf{Y}_i) = \left[\frac{(Y_{ij} - \pi_i^*)(Y_{ij'} - \pi_i^*)}{\pi_i^*(1-\pi_i^*)}\right]_{j<j'}$$

Note that the working covariance matrix V_i need not be correctly specified to produce consistent estimates, but doing so may lead to improved efficiency. The expression $\mathcal{E}(\mathbf{Y}_i)$ involves standardized residuals and is one particular parametrization of GEE2 [Ziegler et al., 2000], but we note there are others [Liang and Zeger, 1992; Zhao and Prentice, 1990]. We pick the above parametrization because it specifically targets estimating the treatment-specific ICC ρ_i^* instead of, say, the cross moments or covariances, but our proposed framework is just as applicable to these other parametrizations.

1.2 Proposal I: Stochastic GEE2

1.2.1 Background: Robbins-Monro Algorithm

GEE2 is ordinarily solved using Newton-Raphson, which has iterations $\theta_0, \theta_1, \cdots$ of the form

$$\theta_{\omega+1} = \theta_\omega + H_{(\omega)}^{-1} G_{(\omega)} \tag{1.4}$$

where

$$H_{(\omega)} = \sum_{i=1}^{I} D_{i(\omega)}^\intercal V_{i(\omega)}^{-1} D_{i(\omega)}, \qquad G_{(\omega)} = \sum_{i=1}^{I} D_{i(\omega)}^\intercal V_{i(\omega)}^{-1} E_{i(\omega)} \tag{1.5}$$

Here, a subscript (ω) indicates evaluation at parameter values $\theta_\omega = (\boldsymbol{\beta}_\omega, \boldsymbol{\alpha}_\omega)$, and we are using letters H and G to invoke the "Hessian" and "gradient" terminologies prevalent in the stochastic approximation literature. In GEE1, computation is dominated by the inversion of $V_{i(\omega)}^{-1}$, which has computational order of $\mathcal{O}(\max n_i^3)$. With GEE2, this increases to $\mathcal{O}(\max n_i^6)$. To reduce this computational burden, we propose a Robbins-Monro (RM) [Robbins and Monro, 1951] variant to fitting GEE2.

In general, the RM algorithm states that, in solving for θ_0 in the equation $G(\theta) = 0$, if we instead possess a random variable $\tilde{G}(\theta)$ such that $\mathbb{E}[\tilde{G}(\theta)] = G(\theta)$ and iterate $\theta_{\omega+1} = \theta_\omega - \gamma_\omega \tilde{G}(\theta_\omega)$, where learning rates $\gamma_\omega > 0$ satisfy $\sum_\omega \gamma_\omega = \infty$ and $\sum_\omega \gamma_\omega^2 < \infty$, then $\theta_\omega \to \theta_0$ in L^2-mean [Robbins and Monro, 1951] and almost surely [Blum, 1954], subject to a few additional regularity conditions. The RM algorithm is useful whenever we can find

a $\tilde{G}$ which is significantly faster to compute than G. For example, consider the general M-estimation problem (for which GEE is a special case) and suppose our estimating equation takes the form $G(\theta) = \sum_{i=1}^{I} G_i(\theta)$. It is easy to confirm that $\tilde{G}(\theta) = \sum_{i \in s} G_i(\theta)/p_i$ satisfies $\mathbb{E}[\tilde{G}(\theta)] = G(\theta)$, where s is a randomly chosen subset of $U = \{1, \cdots, I\}$ according to some sampling design $\mathbb{D}$ with $p_i = \mathbb{P}(i \in s)$. Here, instead of performing I function evaluations, we only need to perform $|s|$ evaluations at each iteration. If we take $\mathbb{D}$ to be a simple random sample without replacement (SRSWOR) of size υ, this reduces to minibatch stochastic gradient descent (mbSGD) [Clémençon et al., 2015]. Our focus is slightly different, because what plagues computation is not many summands corresponding to *many* clusters, but rather the computation of each summand corresponding to *large* clusters, which is commonplace in large-scale CRTs [Gaolathe et al., 2016]. The design of the proposed class of stochastic GEE2 (S-GEE2) algorithm differs from the standard mbSGD in that we are improving iteration speed not through evaluating fewer of the functional summands $\{G_i\}_{i=1}^{I}$ (i.e. evaluating fewer clusters), but rather evaluating an unbiased and computationally-easier estimate of each summand G_i (done through sampling a subset of individuals per cluster). Also, we shall incorporate stochastic Hessians $\theta_{\omega+1} = \theta_\omega + \gamma_\omega \tilde{H}^{-1}(\theta_\omega)\tilde{G}(\theta_\omega)$, which Byrd et al. [2016] proves to convergence almost surely as well. Unlike the stochastic gradient $\tilde{G}$, the accuracy for a stochastic Hessian $\tilde{H}$ is more forgiving, hence cruder approximations are often used to improve speed and memory allocation. In the context of GEE2, the Hessians have a palatable closed-form, so we need not resort to this tactic.

1.2.2 Subsampling

For what we define as the standard S-GEE2, let $U = (U_1, \cdots, U_I)$, where each U_i correspond to the indices of the outcomes in cluster i, with $|U_i| = n_i$. At each iteration ω, sample $s_i \sim \text{SRSWOR}(U_i, \upsilon_i)$, and concatenate $s = (s_1, \cdots, s_I)$. That is, each cluster sample s_i is a simple random sample without replacement of υ_i indices of the ith cluster. Then, perform the Newton-Raphson iteration with just the subsampled units, also known as the

mini-batch in the SGD literature. Notationally, this just replaces $H_{(\omega)}, G_{(\omega)}$ in Eq 1.4 with

$$\tilde{H}_{(\omega)} = \sum_{i=1}^{I} D_{i(\omega)}^{\intercal} V_{i(\omega)}^{-1} W_{i(\omega)}^{S} D_{i(\omega)}, \qquad \tilde{G}_{(\omega)} = \sum_{i=1}^{I} D_{i(\omega)}^{\intercal} V_{i(\omega)}^{-1} W_{i(\omega)}^{S} E_{i(\omega)} \tag{1.6}$$

where $W_{i(\omega)}^{S}$ is a 0-1 weighted diagonal matrix indicating whether an observation is included in s_i, including pairwise indicators for the GEE2 portion. It is easy to verify that $\tilde{H}_{(\omega)}, \tilde{G}_{(\omega)}$ are unbiased estimators for $H_{(\omega)}, G_{(\omega)}$, respectively. Hence, by the RM conditions, we have that S-GEE2 produces estimates $(\tilde{\boldsymbol{\beta}}_{\omega}, \tilde{\boldsymbol{\alpha}}_{\omega}) \to (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$ almost surely. Because Hessians are embedded within our procedure, we should selected a sizeable subsample / mini-batch to ensure a reliable estimate of the Hessian; we recommend $\upsilon_i \geq 5$. We present the full details in pseudocode of S-GEE2 in Algorithm 1 in Appendix A.1.

1.2.3 Exploiting sparsity

Currently, the general structure of S-GEE2 solves two issues: instability and memory demands. Even for simple functions, Newton-Raphson is known for divergence issues due to evaluations near stationary points, where the Hessian is nearly non-invertible. S-GEE2 naturally solves this issue because stochasticity makes it very likely to "jump" off the path of divergence. Secondly, programming the expressions in Eq 1.6 need only store a subset of the rows in $D_{i(\omega)}$ or $E_{i(\omega)}$ and columns of $V_{i(\omega)}^{-1}$, hence greatly freeing up RAM on a computer.

One issue that can be improved, depending on the structure of the working covariance matrix V_i, is computational speed. For example, if we were to assume the off-diagonals $\text{Cov}(\mathbf{Y}_i, \mathcal{E}(\mathbf{Y}_i)) = \mathbf{0}$, then iterations for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be separated, with the GEE2 portion Newton-Raphson iterations taking the form $\boldsymbol{\alpha}_{\omega+1} = \boldsymbol{\alpha}_{\omega} + H_{\boldsymbol{\alpha}(\omega)}^{-1} G_{\boldsymbol{\alpha}(\omega)}$. Then, if we take the GEE2 portion working covariance $V_{\alpha i} = \text{Var}(\mathcal{E}(\mathbf{Y}_i))$ to be diagonal, we can show (see Appendix A.2) that each iteration has complexity $\mathcal{O}(1)$ (constant time!), given subset sizes υ_i are not growing with respect to n_i.

We summarize the scenarios and resulting computational complexities in the theorem below:

Theorem: Let $\text{Cov}(\mathbf{Y}_i, \mathcal{E}(\mathbf{Y}_i)) = \mathbf{0}$ and $\upsilon_i = \mathcal{O}(1)$ (with respect to cluster sizes n_i). In the presence of standard Newton-Raphson, an iteration of the GEE1 portion with (i) arbitrary covariance, (ii) compound symmetry covariance (equicorrelation), and (iii) independence matrices are of complexities (i) $\mathcal{O}(\max_i n_i^3)$, (ii) $\mathcal{O}(\max_i n_i)$, and (iii) $\mathcal{O}(\max_i n_i)$. Similarly, standard Newton-Raphson on the GEE2 portion yields (i) $\mathcal{O}(\max_i n_i^6)$, (ii) $\mathcal{O}(\max_i n_i^2)$, and (iii) $\mathcal{O}(\max_i n_i^2)$; stochastic Newton-Raphson on the GEE1 portion yields (i) $\mathcal{O}(\max_i n_i^3)$, (ii) $\mathcal{O}(\max_i n_i)$, and (iii) $\mathcal{O}(1)$; stochastic Newton-Raphson on the GEE2 portion yields (i) $\mathcal{O}(\max_i n_i^6)$, (ii) $\mathcal{O}(\max_i n_i^2)$, and (iii) $\mathcal{O}(1)$.

See proofs in Appendix A.2. Table 1.1 expresses a clearer schematic of the theorem.

	Full		Stochastic	
	GEE1 portion	GEE2 portion	GEE1 portion	GEE2 portion
Arbitrary covariance	$\mathcal{O}(\max_i n_i^3)$	$\mathcal{O}(\max_i n_i^6)$	$\mathcal{O}(\max_i n_i^3)$	$\mathcal{O}(\max_i n_i^6)$
Equicorrelation	$\mathcal{O}(\max_i n_i)$	$\mathcal{O}(\max_i n_i^2)$	$\mathcal{O}(\max_i n_i)$	$\mathcal{O}(\max_i n_i^2)$
Independence	$\mathcal{O}(\max_i n_i)$	$\mathcal{O}(\max_i n_i^2)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$

Table 1.1: Time complexities for S-GEE2 algorithms under various working covariance structures.

If we choose to model with equicorrelated $\rho_{ijj'} = \rho_i$, as commonly done in CRT's [Crespi et al., 2009; Hayes and Moulton, 2009] and assume independence working covariance for the GEE2 portion, then standard Newton-Raphson on GEE2 would have $\mathcal{O}(\max_i n_i)$ for the GEE1 portion and $\mathcal{O}(\max_i n_i^2)$ for the GEE2 portion, hence the overall complexity is $\mathcal{O}(\max_i n_i^2)$. With S-GEE2, while the GEE1 portion remains at $\mathcal{O}(\max_i n_i)$, the GEE2 portion now becomes $\mathcal{O}(1)$, and hence S-GEE2 has overall complexity of $\mathcal{O}(\max_i n_i)$. Therefore, S-GEE2 cuts down the computation per iteration from roughly a quadratic rate to roughly a linear rate. If we allow the GEE1 portion to also have an independence covariance structure, then the effect of S-GEE2 is even more dramatic, cutting complexity from

$\mathcal{O}(\max_i n_i^2)$ to $\mathcal{O}(1)$.

1.2.4 Par-S-GEE2

While S-GEE2 algorithms allow faster computations in its iterative fitting procedure, each iteration is not as informative due to variation from the induced missingness. Hence, more iterations of S-GEE2 are needed in order to solve the estimating equations, although in practice the additional time in running more iterations is far less significant than the computational savings per iteration. Nevertheless, in pursuit of a S-GEE2 variant requiring fewer iterations, we propose the Parallel S-GEE2 (Par-S-GEE2) class of algorithms. The general technique of parallelized SGD is expanded upon in Zinkevich et al. [2010]. The basic idea is, after sufficiently enough iterations of S-GEE2, the stochastic estimates will become unbiased and further iterations are meant to reduce variation from the stochastic nature of the algorithm. Rather, one can run K independent chains of S-GEE2 and average the resulting convergent estimates. Both running more iterations on a single chain or averaging over multiple chains has the same effect in reducing the variation in estimates, but with the former, the iterations must be done sequentially and hence the user must wait, while with the latter, the chains can be run in parallel. Pseudocode is provided in Algorithm 2 in Appendix A.1.

S-GEE2 reduces the frequency of divergence, but generally not all of it; there remains a non-negligible probability that the algorithm may diverge. Par-S-GEE2 inherently solves the convergence issue because at least some of the chains would have converged. The average of these convergent solutions is one estimator, or better yet, one can then feed this estimator as an initial value on another run of Par-S-GEE2, since the provided estimate would act as a better initialization and reduce the number of divergences. In a sense, Par-S-GEE2 is very similar to multistart search [Ugray et al., 2007] because each chain initially fluctuates around the search space, effectively acting as a scattering of starting values. At the same time, this scattering is informative because each chain is still trying to fit on a subset of data. Hence, Par-S-GEE2 offers an advantage in intrinsically incorporating

information in its multistart search rather than truly random scattering.

1.3 Proposal II: IPW & DR GEE2

1.3.1 IPW-GEE2

Accounting for missing outcome data in CRTs is challenging under the missing at random (MAR) assumption because there is no natural ordering of the outcomes within a cluster and the missingness can not be considered as monotone. We consider a submodel of MAR, restricted MAR (rMAR). Let $R_{ij} \in \{0,1\}$ with $R_{ij} = 0$ indicating Y_{ij} is missing. Then, rMAR is defined as $\mathbb{P}(R_{ij} = 1|\mathbf{Y}_i, A_i, \mathbf{Z}_i, \mathbf{X}_i) = \mathbb{P}(R_{ij} = 1|A_i, \mathbf{Z}_i, \mathbf{X}_i)$; that is, $R_{ij} \perp\!\!\!\perp Y_{ij}|A_i, \mathbf{Z}_i, \mathbf{X}_i$. To continue with valid inference, we assume that $\mathbb{P}(R_{ij} = 1|A_i, \mathbf{Z}_i, \mathbf{X}_i) > 0$, commonly known as the positivity assumption (PO). We propose the inverse-probability weighting second-order generalized estimating equations (IPW-GEE2) as

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{I} D_i^\intercal V_i^{-1} W_i^R E_i \stackrel{\text{def}}{=} \sum_{i=1}^{I} \Phi_i^Y(A_i, \boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R) \\ \mathbf{0} &= \sum_{i=1}^{I} \mathbf{S}_i^R(A_i, \mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R) \end{aligned} \tag{1.7}$$

where we have incorporated the following inverse-probability weighting matrix:

$$W_i^R = \text{diag}\left(\underbrace{\frac{R_{i1}}{\overline{\pi}_{i1}^R(\boldsymbol{\beta}_R)}, \cdots, \frac{R_{in_i}}{\overline{\pi}_{in_i}^R(\boldsymbol{\beta}_R)}}_{\text{IPW1}}, \underbrace{\frac{R_{i1}R_{i2}}{\overline{\eta}_{i12}^R(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)}, \cdots, \frac{R_{i(n_i-1)}R_{in_i}}{\overline{\eta}_{i(n_i-1)n_i}^R(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)}}_{\text{IPW2}} \right)$$

The summands involving $\mathbf{S}_i^R$ are needed to estimate nuisance parameters $(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)$ guiding the missingness process $\mathbf{R}_i$, but the parameters themselves are of no interest for inference. Within the IPW matrix, $\overline{\pi}_{ij}^R(\boldsymbol{\beta}_R)$ is a model (parametrized by $\boldsymbol{\beta}_R$) for $\pi_{ij}^R = \mathbb{P}(R_{ij} = 1|A_i, \mathbf{Z}_i, \mathbf{X}_i)$ and $\overline{\eta}_{ijj'}(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)$ is a model (parametrized by $\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R$) for $\eta_{ijj'}^R = \mathbb{P}(R_{ij} = R_{ij'} = 1|A_i, \mathbf{Z}_i, \mathbf{X}_i)$; we shall refer to them as the first-order and second-order propensity scores (PS1 & PS2), respectively. Since $\eta_{ijj'}^R$ is a function of $\pi_{ij}^R, \pi_{ij'}^R, \rho_{ijj'}^R$, it suffices to fit a

model for $\rho^R_{ijj'}$. The matrix W_i^R is the inverse-probability weighting (IPW) matrix, which can be decomposed into IPW1 and IPW2 portions. We refer to the first equation of Eqs 1.7 as the treatment model estimating equation (TMEE) portion, while the second equation of Eqs 1.7, which produce estimators $\widehat{\pi}^R_{ij}$ (converging to $\overline{\pi}^R_{ij}$) and $\widehat{\rho}^R_{ijj'}$ (converging to $\overline{\rho}^R_{ijj'}$), as the propensity score estimating equation (PSEE) portion.

The IPW2 portion is derived by considering that the (j, j')th element of $\mathcal{E}(\mathbf{Y}_i)$ is missing when either Y_{ij} or $Y_{ij'}$ is missing; this is exactly represented by the product of their missingness indicators, $R_{ij}R_{ij'}$, for which we would then need to model $\eta^R_{ijj'}(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)$. To the best of our knowledge, this is the first instance in which a model is required for the joint missingness indicator $R_{ij}R_{ij'}$ in the context of clustered data. Not properly accounting for the correlation among missingness indicators will in general lead to biased estimates for the association parameters. Unlike the treatment model, the PS can possibly be misspecified; if so, then estimators $(\widehat{\boldsymbol{\beta}}^*_Y, \widehat{\boldsymbol{\alpha}}^*_Y)$ may not be consistent.

1.3.2 DR-GEE2

The augmented GEE (AUG) methods, which adds a term to the standard GEE that relates the outcome to covariates and treatment, have been proposed to improve estimation efficiency by leveraging baseline covariates in the setting of CRTs [Stephens et al., 2012]. We apply this to IPW-GEE2, forming what we call DR-GEE2:

$$\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{I} [D_i^\intercal V_i^{-1} W_i^R E_i' + \zeta_i] \stackrel{\text{def}}{=} \sum_{i=1}^{I} \tilde{\Phi}_i^Y(\mathbf{Z}_i^*, \mathbf{X}_i, \mathbf{R}_i, \boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y) \\
\mathbf{0} &= \sum_{i=1}^{I} \mathbf{S}_i^R(\mathbf{Z}_i^*, \mathbf{X}_i, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R), \qquad \mathbf{0} = \sum_{i=1}^{I} \mathbf{S}_i^Y(\mathbf{Z}_i^*, \mathbf{X}_i, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y)
\end{aligned} \tag{1.8}$$

where

$$E_i' = \begin{pmatrix} \mathbf{Y}_i - \overline{\boldsymbol{\pi}}_i(\boldsymbol{\beta}_Y) \\ \mathcal{E}(\mathbf{Y}_i) - \overline{\mathbf{r}}_i^\dagger(\boldsymbol{\alpha}_Y) \end{pmatrix}, \; E_i'' = \begin{pmatrix} \overline{\boldsymbol{\pi}}_i(\boldsymbol{\beta}_Y) - \boldsymbol{\pi}_i^*(\boldsymbol{\beta}_Y^*) \\ \overline{\mathbf{r}}_i^\dagger(\boldsymbol{\alpha}_Y) - \mathbf{r}_i^*(\boldsymbol{\alpha}_Y^*) \end{pmatrix},$$

$$\zeta_i = \sum_{a=0}^{1} p_A^a (1 - p_A)^{1-a} D_i^\intercal (A = a) V_i^{-1} E_i''(A = a)$$

where $\overline{\pi}_{ij}$ is a model for π_{ij} and

$$\overline{\rho}^{\dagger}_{ijj'} = \frac{(\overline{\pi}_{ij} - \overline{\pi}^*_i)(\overline{\pi}_{ij'} - \overline{\pi}^*_i) + \overline{\rho}_{ijj'}\sqrt{\overline{\mathcal{V}}_{ijj'}}}{\overline{\pi}^*_i(1 - \overline{\pi}^*_i)}$$

akin to Eq 1.1, with models replacing each population quantity. The third set of equations in Eq 1.8, which we refer to as the outcome model estimating equations (OMEE), fits $\widehat{\pi}_{ij}$ (converging to $\overline{\pi}_{ij}$) and $\widehat{\rho}_{ijj'}$ (converging to $\overline{\rho}_{ijj'}$), collectively known as the outcome models. If the OMs are correctly specified, then under the rMAR assumption, $(\boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y)$ can be consistently estimated based on the complete-case data. The DR estimator is doubly robust in the sense that it is CAN under correct specification of either the OM [i.e. $\overline{\pi}_{ij} = \pi_{ij}$ and $\overline{\rho}_{ijj'} = \rho_{ijj'}$] or PS [i.e. $\overline{\pi}^R_{ij} = \pi^R_{ij}$ and $\overline{\rho}^R_{ijj'} = \rho^R_{ijj'}$] (see proof in Appendix A.3).

1.3.3 Inference

Concatenate $\boldsymbol{\kappa} = (\boldsymbol{\beta}^*_Y, \boldsymbol{\alpha}^*_Y, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y)$ as the collection of parameters from the TM, PSM, and OM. A direct application of the theory of M-estimators [Van der Vaart, 2000] with score equation summands $\Psi(\boldsymbol{\kappa}) = (\tilde{\Phi}^Y_i, \mathbf{S}^R_i, \mathbf{S}^Y_i)^\intercal$ yield that $\sqrt{I}(\widehat{\kappa} - \kappa) \xrightarrow{\mathcal{D}} N(0, \Sigma)$, where $\Sigma = \Gamma^{-1}\Delta(\Gamma^{-1})^\intercal$ is the sandwich estimator with meat $\Delta(\boldsymbol{\kappa}) = \mathbb{E}\left[\Psi(\boldsymbol{\kappa})\Psi(\boldsymbol{\kappa})^\intercal\right]$ and breads $\Gamma(\boldsymbol{\kappa}) = \mathbb{E}\left[\partial\Psi(\boldsymbol{\kappa})/\partial\boldsymbol{\kappa}^\intercal\right]$, from which we can extract components corresponding to just $(\widehat{\boldsymbol{\beta}}^*_Y, \widehat{\boldsymbol{\alpha}}^*_Y)$, the parameters of interest. An estimator $\widehat{\Sigma}$ can be obtained by replacing Δ with $\widehat{\Delta} = \frac{1}{I}\sum_{i=1}^I \widehat{\Psi}(\widehat{\boldsymbol{\kappa}})\widehat{\Psi}(\widehat{\boldsymbol{\kappa}})^\intercal$ and Γ with $\widehat{\Gamma} = \frac{1}{I}\sum_{i=1}^I \partial\widehat{\Psi}(\widehat{\boldsymbol{\kappa}})/\partial\boldsymbol{\kappa}$.

1.3.4 Embedding IPW-GEE2 and DR-GEE2 with S-GEE2

Fitting the PSM and OM requires just the standard GEE2, so therefore no adjustments are needed from S-GEE2 defined in Section 1.2.2. We do, however, need to adjust the TM for IPW-GEE2 and DR-GEE2.

Structurally speaking, the inclusion of subsampling matrix W^S in S-GEE2 is similar to the IPW matrix W^R in IPW-GEE2. Indeed, in fitting IPW-GEE2 with a stochastic variant (which we call S-IPW-GEE2), we simply adjust gradient and Hessian iterations in Eq 1.6 with $W^S \mapsto W^R W^S$. There are two possible candidates for the new W^S. We could use

the exact W^S scheme in the original S-GEE2 defined in Section 1.2.2. But, this is likely to sample missing outcomes, and if there is significant missingness, the iterations would fail due to singular Hessians. A more stable procedure would be to sample $s_i \sim \text{SRSWOR}(U_i^{\text{obs}}, \upsilon_i)$, where $U^{\text{obs}} = (U_1^{\text{obs}}, \cdots, U_I^{\text{obs}})$ and each U_i^{obs} correspond to the indices of the *non-missing* outcomes in cluster i. Let $m_i = |U_i^{\text{obs}}|$ be the number of non-missing observations per cluster. Then, define $W^{S\text{obs}}_{\beta i(\omega)} = \frac{m_i}{\upsilon_i}[s_i]$ and $W^{S\text{obs}}_{\alpha i(\omega)} = \frac{m_i(m_i-1)}{\upsilon_i(\upsilon_i-1)}[(s_i)_2]$, a weighted indicator matrix for a subsample of the non-missing outcomes.

Forming S-DR-GEE2 (DR-GEE2 + S-GEE2) is not so clear-cut. In Eq 1.8, the TM score equations consist of $D_i^\intercal V_i W_i^R E_i'$ and ζ_i. For the $D_i^\intercal V_i W_i^R E_i'$ component, we have missing outcomes, and therefore might seem prudent to use $W^{S\text{obs}}$ for the reasons above. But, an analogous version for ζ_i will *not* yield unbiased estimators $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$, since $W^{S\text{obs}}$ contain conditional weights given $\mathbf{R}$ (i.e. being observed), yet there are no missing elements in ζ_i. Therefore, we propose simultaneous, independent subsampling schemes for $D_i^\intercal V_i W_i^R E_i'$ and ζ_i, the latter akin to $W^{S\text{obs}}$ and the former akin to W^S. Details are presented in Algorithm 3 in Appendix A.1.

1.4 Simulation

We perform two sets of experiments. The first set explores the statistical properties of IPW-GEE2 and DR-GEE2 under combinations of correctly specified / misspecified PS model and correctly specified / misspecified OM, all of which include the ICC estimates embedded in the working covariance structure in the GEE1 portion, and assuming $\text{Cov}(\mathbf{Y}_i, \mathcal{E}(\mathbf{Y}_i)) = \mathbf{0}$ and $\text{Var}(\mathcal{E}(\mathbf{Y}_i)) = \mathbf{I}$. We include analogous estimates from a parametric mixed effects model and GEE1 with independence working covariance structure for comparison. In the second set of simulations, we compare the algorithmic properties (convergence & run-time) of S-DR-GEE2 and standard DR-GEE2 under various cluster size / number of cluster combinations.

We consider the following two data generation processes for binary data Y_{ij} (or R_{ij}):

$$\text{Parzen's method}\begin{cases}\text{logit}(\pi_{ij}) & = (\beta_{0Y}+\beta_{0AY}A_i)+(\boldsymbol{\beta}_{ZY}+\boldsymbol{\beta}_{ZAY}A_i)^\intercal\mathbf{Z}_i \\ & \quad +(\boldsymbol{\beta}_{XY}+\boldsymbol{\beta}_{XAY}A_i)^\intercal\mathbf{X}_{ij} \\ \text{atanh}(\rho_i) & = (\alpha_{0Y}+\alpha_{0AY}A_i)+(\boldsymbol{\alpha}_{ZY}+\boldsymbol{\alpha}_{ZAY}A_i)^\intercal\mathbf{Z}_i \\ (\mathfrak{L}_i,\mathfrak{U}_i) & = \left(-\sqrt{\frac{\min(\boldsymbol{\pi}_i)}{1-\min(\boldsymbol{\pi}_i)}},\sqrt{\frac{1-\max(\boldsymbol{\pi}_i)}{\max(\boldsymbol{\pi}_i)}}\right) \\ (\delta_i,\epsilon_i) & = \left(\frac{\mathfrak{U}_i(-\mathfrak{U}_i\mathfrak{L}_i-\rho_i)}{(\mathfrak{U}_i-\mathfrak{L}_i)\rho_i},\frac{-\mathfrak{L}_i(-\mathfrak{U}_i\mathfrak{L}_i-\rho_i)}{(\mathfrak{U}_i-\mathfrak{L}_i)\rho_i}\right) \\ \xi_i|A_i,\mathbf{Z}_i & \sim (\mathfrak{U}_i-\mathfrak{L}_i)\text{Beta}(\delta_i,\epsilon_i)+\mathfrak{L}_i \\ Y_{ij}|A_i,\mathbf{Z}_i,\mathbf{X}_i,\xi_i & \sim \text{Bernoulli}\left(\pi_{ij}+\xi_i\sqrt{\pi_{ij}(1-\pi_{ij})}\right)\end{cases} \tag{1.9}$$

$$\text{Random intercept}\begin{cases}\text{logit}(\pi_{ij}) & = (\beta_{0Y}+\beta_{0AY}A_i)+(\boldsymbol{\beta}_{ZY}+\boldsymbol{\beta}_{ZAY}A_i)^\intercal\mathbf{Z}_i \\ & \quad +(\boldsymbol{\beta}_{XY}+\boldsymbol{\beta}_{XAY}A_i)^\intercal\mathbf{X}_{ij} \\ \xi_i|A_i & \sim N(0,(\tfrac{1}{3}+\tfrac{1}{2}A_i)^2) \\ \text{logit}(p_{ij}) & = \xi_i+\text{logit}(\pi_{ij}) \\ Y_{ij}|A_i,\mathbf{Z}_i,\mathbf{X}_i,\xi_i & \sim \text{Bernoulli}\left(p_{ij}\right)\end{cases}$$

Parzen's method [Parzen, 2009] offers a random-effects form that attains nominal levels of π_{ij} and ρ_i (i.e. $\mathbb{P}(Y_{ij}|A_i,\mathbf{Z}_i,\mathbf{X}_i)=\pi_{ij}$ and $\text{Corr}(Y_{ij},Y_{ij'}|A_i,\mathbf{Z}_i)=\rho_i$) and specifically generates equicorrelated data. To ensure $0\leq\pi_{ij}+\xi_i\sqrt{\pi_{ij}(1-\pi_{ij})}\leq 1$, one must ensure that $-\mathfrak{U}_i\mathfrak{L}_i-\rho_i\geq 0$ for all i. The random intercept is the traditional approach in inducing correlation among observations in a cluster. With a normal random intercept, the marginal probability of success

$$\mathbb{P}(Y_{ij}=1|A_i,\mathbf{Z}_i,\mathbf{X}_i)=\int\frac{e^{\xi_i+L(\boldsymbol{\beta};A_i,\mathbf{Z}_i,\mathbf{X}_i)}}{1+e^{\xi_i+L(\boldsymbol{\beta};A_i,\mathbf{Z}_i,\mathbf{X}_i)}}dP(\xi_i) \tag{1.10}$$

where $L(\boldsymbol{\beta};A_i,\mathbf{Z}_i,\mathbf{X}_i)$ is the linear function, is not of the logistic form and will not have a closed-form. Furthermore, the ICC is induced linearly on the logit scale, yet the manifested ICC after performing the expit function and appropriate marginalization will vary within-cluster and hence is unsuitable for simulation of equicorrelated data. We use Parzen's method to generate the ideal case of equicorrelated outcomes, while we use random intercept

to induce non-equicorrelated outcomes. Furthermore, since the normal random intercept is not of the logistic form, any OM we fit with logistic regression is necessarily a misspecified model, yet we show that the marginalization interpretation $\rho_i^* = \mathbb{E}[\rho_{ijj'}^\dagger | A_i]$ holds.

1.4.1 Consistency and efficiency of IPW-GEE2 & DR-GEE2 schemes

Let $\mathcal{U}(a, b)$ denote the continuous uniform distribution on (a, b), and let $\mathcal{U}\{a, b\}$ denote the discrete uniform distribution on $\{a, a+1, \cdots, b-1, b\}$. To evaluate the asymptotic properties of GEE2, we set the number of clusters to an unrealistic $I = 2000$ with cluster sizes $n_i \sim \mathcal{U}\{80, 140\}$ so that we have average cluster size $\mathbb{E}[n_i] = 110$. The setting with large number of clusters allows us to observe asymptotic properties more quickly and to avoid computational issues that will be explored in Section 1.4.2. We generate $A_i \sim \text{Ber}(1/2)$ and choose $\mathbf{X}_{ij} \in \mathbb{R}^3$ and $\mathbf{Z}_i \in \mathbb{R}$. Details regarding generation of $\mathbf{X}_{ij}$, $\mathbf{Z}_i$ and choice of coefficients for Y_{ij} is presented in Table 1.2. We also generate R_{ij} with these same covariates and coefficients for simplicity.

Covariate	Intercept	$\mathbf{X}_{ij}$			$\mathbf{Z}_i$
Generation	–	$\mathcal{U}(20, 60)$	$\mathcal{U}(1, 10)$	$\mathcal{U}(4, 25)$	$\mathcal{U}\{80, 140\}$
Main-effects $\boldsymbol{\beta}_{\cdot Y}$	0.11	-0.007	-0.020	-0.040	0.009
Interaction $\boldsymbol{\beta}_{\cdot AY}$	0.67	0.012	0.030	0.060	-0.018
Main-effects $\boldsymbol{\alpha}_{\cdot Y}$	-0.32	–	–	–	0.004
Interaction $\boldsymbol{\alpha}_{\cdot Y}$	0.96	–	–	–	-0.008

Table 1.2: Information regarding the generation process

The values in Table 1.2 are carefully chosen to guarantee $-\mathfrak{U}_i\mathfrak{L}_i - \rho_i \geq 0$ in Parzen's method. The resulting values for $\mathbb{P}(Y_{ij} = 1 | A_i, \mathbf{Z}_i, \mathbf{X}_i)$ and $\text{Corr}(Y_{ij}, Y_{ij'} | A_i, \mathbf{Z}_i, \mathbf{X}_i)$, after marginalizing out ξ_i, are in the range [0.324, 0.733] and [0.004, 0.306], respectively. For the random-intercept method, the values of $\mathbb{P}(Y_{ij} = 1 | A_i, \mathbf{Z}_i, \mathbf{X}_i)$ and $\text{Corr}(Y_{ij}, Y_{ij'} | A_i, \mathbf{Z}_i, \mathbf{X}_i)$ are in the range [0.333, 0.738] and [0.022, 0.134], respectively. The true treatment coefficients $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$ in the canonical TM can be calculated by numerically integrating out all

other covariates except for A_i in π_{ij} and $\rho^{\dagger}_{ijj'}$:

$$\begin{aligned}\text{expit}(\beta^*_{0Y} + \beta^*_{AY} A_i) &= \int_{\mathbb{R}^4} \pi_{ij} dP(\mathbf{x}_{ij}) dP(\mathbf{z}_i) \\ \tanh(\alpha^*_{0Y} + \alpha^*_{AY} A_i) &= \int_{\mathbb{R}^7} \rho^{\dagger}_{ijj'} dP(\mathbf{x}_{ij}) dP(\mathbf{x}_{ij'}) dP(\mathbf{z}_i)\end{aligned} \tag{1.11}$$

Under Parzen's method, we obtain the values $(\boldsymbol{\beta}^*_Y, \boldsymbol{\alpha}^*_Y) = (0.1413, 0.1808, 0.1238, 0.0755)$, and under random intercept, we obtain $(\boldsymbol{\beta}^*_Y, \boldsymbol{\alpha}^*_Y) = (0.1378, 0.1429, 0.0307, 0.1032)$.

The results in Table 1.3 display biases, replicate standard errors, and average sandwich standard errors of estimated parameters from several models with $\mathcal{R} = 1000$ replicate generations of missingness and outcome, both using Parzen's method. For the mixed effects model, we fit the following on the complete-case data:

$$\text{logit}\{\mathbb{P}(Y_{ij} = 1 | A_i, \xi_i)\} = \tilde{\beta}_0 + \tilde{\beta}_A A_i + \xi_i \quad \text{with} \quad \xi_i | A_i \sim N(0, \tilde{\sigma}^2_{A_i}) \tag{1.12}$$

which takes nearly the functional form of the random intercept generation process in Eq 1.9, less the baseline covariates. Using the marginalizations in Eqs 1.10 and 1.11, we can obtain $(\beta^*_{0Y}, \beta^*_{AY}, \alpha^*_{0Y}, \alpha^*_{AY})$ from $(\tilde{\beta}_0, \tilde{\beta}_A, \tilde{\sigma}^2_0, \tilde{\sigma}^2_1)$ and standard errors for $\beta^*_{0Y}, \beta^*_{AY}$ from the standard errors of $\tilde{\beta}_0, \tilde{\beta}_A$ through the delta method. Unfortunately, analytical standard errors for $\alpha^*_{0Y}, \alpha^*_{AY}$ require standard errors of $\tilde{\sigma}^2_0, \tilde{\sigma}^2_1$, for which methods are less well-developed [Bates, 2010; McCulloch and Searle, 2001; Wu et al., 2012]. Hence, while we report replicate standard errors for $\tilde{\sigma}^2_0, \tilde{\sigma}^2_1$, we omit sandwich error standard errors. Mixed effects models naturally handle MAR if the true generation process follows the form in Eq 1.12. Certainly, both generation processes in Eq 1.9 do not. Parzen's method takes a functional form that is different from the random intercept method, and the random intercept method includes additional covariates for which Eq 1.12 omits.

For the IPW-GEE2 fits, we distinguish $\mathcal{G}_1(\mathbf{R})$ IPW and $\mathcal{G}_2(\mathbf{R})$ IPW as the IPW models with and without accounting for the correlation among the missingness indicators, respectively, as discussed in Section 1.3.1. For GEE1, there is no model for correlated missingness, and that block is omitted. The fitted OM and correctly-specified PSM are

$$\begin{aligned}\text{logit}(\pi_{ij}) &= (\beta_{0Y} + \beta_{0AY} A_i) + (\boldsymbol{\beta}_{ZY} + \boldsymbol{\beta}_{ZAY} A_i)^{\intercal} \mathbf{Z}_i + (\boldsymbol{\beta}_{XY} + \boldsymbol{\beta}_{XAY} A_i)^{\intercal} \mathbf{X}_{ij} \\ \text{atanh}(\rho_{ijj'}) &= (\alpha_{0Y} + \alpha_{0AY} A_i) + (\boldsymbol{\alpha}_{ZY} + \boldsymbol{\alpha}_{ZAY} A_i)^{\intercal} \mathbf{Z}_i\end{aligned} \tag{1.13}$$

	Averaged bias (Replicate SE) (Averaged sandwich SE)				Averaged bias (Replicate SE) (Averaged sandwich SE)	
	β^*_{0Y}	β^*_{AY}	α^*_{0Y}	α^*_{AY}	β^*_{0Y}	β^*_{AY}
Complete Case Mixed Effects						
	0.0421	−0.0238	0.0016	−0.0009		
	(0.0227)	(0.0364)	(0.0053)	(0.0088)	———	
	(0.0238)	(0.0373)	—	—		
GEE		*GEE2*			*GEE1*	
Complete Case						
	0.0349	−0.0239	0.0113	−0.0016	0.0413	−0.0228
	(0.0245)	(0.0379)	(0.0070)	(0.0121)	(0.0262)	(0.0404)
	(0.0238)	(0.0380)	(0.0069)	(0.0117)	(0.0260)	(0.0416)
PSM Correctly Specified						
	−0.0006	0.0020	0.0024	−0.0008	−0.0003	0.0010
$\mathcal{G}_1(\mathbf{R})$ IPW	(0.0257)	(0.0398)	(0.0064)	(0.0112)	(0.0252)	(0.0391)
	(0.0249)	(0.0400)	(0.0064)	(0.0111)	(0.0252)	(0.0405)
	−0.0005	0.0019	−0.0001	0.0002		
$\mathcal{G}_2(\mathbf{R})$ IPW	(0.0258)	(0.0399)	(0.0066)	(0.0112)	———	
	(0.0249)	(0.0401)	(0.0063)	(0.0109)		
	−0.0006	0.0018	−0.0003	0.0003	−0.0004	0.0010
Doubly-Robust	(0.0262)	(0.0399)	(0.0061)	(0.0111)	(0.0251)	(0.0391)
	(0.0297)	(0.0389)	(0.0060)	(0.0108)	(0.0246)	(0.0404)
PSM Misspecified						
	0.0341	−.0124	0.0112	−0.0018	0.0341	−0.0121
$\mathcal{G}_1(\mathbf{R})$ IPW	(0.0255)	(0.0414)	(0.0068)	(0.0116)	(0.0264)	(0.0401)
	(0.0255)	(0.0411)	(0.0068)	(0.0117)	(0.0260)	(0.0416)
	0.0326	−0.0092	0.0089	0.0022		
$\mathcal{G}_2(\mathbf{R})$ IPW	(0.0252)	(0.0411)	(0.0067)	(0.0117)	———	
	(0.0255)	(0.0411)	(0.0067)	(0.0117)		
	0.0000	0.0005	−0.0002	−0.0001	−0.0002	0.0007
Doubly-Robust	(0.0251)	(0.0401)	(0.0061)	(0.0107)	(0.0252)	(0.0392)
	(0.0303)	(0.0397)	(0.0064)	(0.0114)	(0.0253)	(0.0415)

Table 1.3: Biases & Standard Errors from 1000 replicate simulations with both Y_{ij}, R_{ij} simulated with Parzen's method.

i.e. the exact model used to generate R_{ij}, Y_{ij} from Parzen's method. The fitted misspecified PSM is

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \beta_{0Y} + \beta_{AY} A_i + \boldsymbol{\beta}_{ZY}^\intercal \mathbf{Z}_i + \boldsymbol{\beta}_{XY}^\intercal \mathbf{X}_{ij} \\ \text{atanh}(\rho_{ijj'}) &= \alpha_{0Y} + \alpha_{AY} A_i + \boldsymbol{\alpha}_{ZY}^\intercal \mathbf{Z}_i \end{aligned} \tag{1.14}$$

i.e. the model with interaction terms of A_i with $\mathbf{Z}_i, \mathbf{X}_i$ are omitted.

We compare the performance of each estimation procedure based on the replicate Wald statistic $W = \sqrt{\mathcal{R}} \cdot \frac{\text{Bias}}{\text{Std Error}}$ and checking whether $|W| > 2$, where $\mathcal{R}$ is the number of replicate simulations. Using this metric and the information from Table 1.3, when PSM is correctly specified, complete-case analysis (for both mixed effects, GEE1, and GEE2) leads to severe bias in estimating all parameters. While mixed models are expected to account for missingness when confounding covariates are included, the focus is on marginal effects and hence inclusion would provide entirely different interpretations. $\mathcal{G}_1(\mathbf{R})$ IPW-GEE2 and IPW-GEE1 provide consistent estimates for the mean parameters β_{0Y}^* and β_{AY}^*, although the former still fails to correctly estimate the association parameters α_{0Y}^* and α_{AY}^*. $\mathcal{G}_2(\mathbf{R})$ IPW-GEE2 and doubly-robust GEE2 and GEE1 produce consistent estimates for all parameters estimable under their respective models. When PSM is misspecified, we note that only DR-GEE2 and DR-GEE1 produce consistent estimates. Note that the sandwich variance estimators in general are close to the true sampling variance with the exception of β_{0Y} under the DR-GEE2 model, for which it is somewhat conservative. We also observe that DR-GEE1 (with independence covariance structure) standard errors of the mean parameters $\beta_{0Y}^*, \beta_{AY}^*$ are smaller than the DR-GEE2 standard errors of $\beta_{0Y}^*, \beta_{AY}^*$.

The results in Table 1.4 display biases, replicate standard errors, and sandwich standard errors of estimated parameters from several models with $\mathcal{R} = 1000$ replicate generations of missingness using Parzen's method and outcome using random intercepts. We still fit the correct OM and PSM using Eq 1.13 and incorrect PSM using Eq 1.14. Note that the true OM is no longer of the logistic form, and hence the fitted OM will be misspecified. Nevertheless, we reach nearly identical conclusions regarding the validity of models as done with Table 1.3. Especially noteworthy is that, even when the PSM is misspecified, the DR-GEE2 produces consistent estimates of all its parameters. Consistent estimation of the

	Averaged bias (Replicate SE) (Averaged sandwich SE)				Averaged bias (Replicate SE) (Averaged sandwich SE)	
	β^*_{0Y}	β^*_{AY}	α^*_{0Y}	α^*_{AY}	β^*_{0Y}	β^*_{AY}
Complete Case Mixed Effects						
	0.0343	−0.0244	−0.0005	−0.0001		
	(0.0144)	(0.0290)	(0.0020)	(0.0058)	———	
	(0.0139)	(0.0279)	—	—		
GEE	*GEE2*				*GEE1*	
Complete Case						
	0.0340	−0.0266	−0.0005	−0.0004	0.0400	−0.0239
	(0.0143)	(0.0291)	(0.0022)	(0.0071)	(0.0145)	(0.0303)
	(0.0140)	(0.0284)	(0.0022)	(0.0070)	(0.0143)	(0.0299)
PSM Correctly Specified						
	−0.0001	−0.0020	−0.0002	0.0003	−0.0002	0.0003
$\mathcal{G}_1(\mathbf{R})$ IPW	(0.0148)	(0.0295)	(0.0023)	(0.0070)	(0.0143)	(0.0297)
	(0.0143)	(0.0297)	(0.0022)	(0.0071)	(0.0143)	(0.0299)
	−0.0001	−0.0021	−0.0001	0.0002		
$\mathcal{G}_2(\mathbf{R})$ IPW	(0.0150)	(0.0296)	(0.0023)	(0.0070)	———	
	(0.0143)	(0.0297)	(0.0022)	(0.0071)		
	−0.0001	−0.0020	−0.0001	0.0003	0.0000	0.0003
Doubly-Robust	(0.0149)	(0.0294)	(0.0023)	(0.0070)	(0.0139)	(0.0297)
	(0.0212)	(0.0248)	(0.0022)	(0.0071)	(0.0137)	(0.0299)
PSM Misspecified						
	0.0328	−0.0157	−0.0005	−0.0003	0.0327	−0.0134
$\mathcal{G}_1(\mathbf{R})$ IPW	(0.0145)	0.0303	(0.0022)	(0.0071)	(0.0145)	(0.0302)
	(0.0143)	(0.0297)	(0.0022)	0.0070	(0.0143)	(0.0299)
	0.0313	−0.0128	−0.0005	−0.0005		
$\mathcal{G}_2(\mathbf{R})$ IPW	(0.0145)	(0.0304)	(0.0022)	(0.0071)	———	
	(0.0142)	(0.0297)	(0.0022)	(0.0071)		
	−0.0006	−0.0006	−0.0001	−0.0001	−0.0008	0.0013
Doubly-Robust	(0.0145)	(0.0296)	(0.0022)	(0.0070)	(0.0141)	(0.0302)
	(0.0211)	(0.0247)	(0.0022)	(0.0069)	(0.0137)	(0.0299)

Table 1.4: Biases & Standard Errors from 1000 replicate simulations with R_{ij} simulated using Parzen's method and Y_{ij} simulated using random-intercept method.

mean parameters may be due to the fact that random intercept generation is still "linear enough" with respect to the covariates. Consistent estimation of the association parameters suggested that, even when the outcome is non-equicorrelated, we may still model it with an equicorrelated OM and still produce roughly consistent estimates of the ICC.

1.4.2 Algorithmic Characteristic of DR-GEE2 vs S-DR-GEE2

Having established the consistency of DR-GEE2, in our second set of experiments we now compare against S-DR-GEE2. We generate both missingness and outcome using Parzen's method and the information from Table 1.2, and we fit with both PSM and OM correctly specified. We now vary the number of cluster I and cluster sizes n_i, and consider the following three scenarios: $(I, \mathbb{E}[n_i]) = (30, 30), (300, 30), (30, 300)$. Tables 1.5 and 1.6 present the statistical and algorithmic results, respectively, of DR-GEE2 vs S-DR-GEE2. For $\mathbb{E}[n_i] = 30$, we set the subsample sizes $\mathbb{E}[\upsilon_i] = 9$ and for $\mathbb{E}[n_i] = 300$, we set $\mathbb{E}[\upsilon_i] = 45$; all S-DR-GEE2 scenarios used learning rates $\gamma_\omega = (\omega + 1)^{-1}$.

From Table 1.5, and using the Wald statistic metric to evaluate model validity, the association parameters from the $I = 30$ sub-experiments all are biased. This is readily explained by the fact that the asymptotics for the association parameters depend on I rather than $\sum_{i=1}^{I} n_i$, and hence at these small number of clusters, asymptotics haven't fully kicked in. Other than this, overall, the parameter estimates and standard errors are very similar between DR-GEE2 and S-DR-GEE2, albeit the standard errors under S-DR-GEE2 are slightly higher. This slightly higher variability can be eliminated by simply asking for a few more iterations. Even so, at a small cost of higher variability, the computational savings of S-DR-GEE2 are apparent. From Table 1.6, even at small cluster sizes, which S-DR-GEE2 was not designed to be optimal, we still see moderately higher convergent solutions and somewhat less time to fit each model. We see these results further accentuated when expected cluster size is 300. For the OM, PSM, and TM, we see that S-DR-GEE2 provides up to 80% reduction in returned errors (i.e. divergence, large condition numbers of Hessians) and approximately 90% reduction in run-time.

Scenarios	Full DR-GEE2 Averaged bias (Replicate SE) (Averaged sandwich SE)				S-DR-GEE2 Averaged bias (Replicate SE) (Averaged sandwich SE)			
	β^*_{0Y}	β^*_{AY}	α^*_{0Y}	α^*_{AY}	β^*_{0Y}	β^*_{AY}	α^*_{0Y}	α^*_{AY}
$(I, \mathbb{E}[n_i]) = (30, 30)$	0.0067	−0.0082	−0.0153	0.0010	0.0025	0.0071	−0.0041	−0.0095
	(0.2563)	(0.3973)	(0.0629)	(0.1140)	(0.2724)	(0.4084)	(0.0715)	(0.1203)
	(0.2541)	(0.3516)	(0.0535)	(0.0983)	(0.2533)	(0.3513)	(0.0580)	(0.1012)
$(I, \mathbb{E}[n_i]) = (300, 30)$	−0.0004	−0.0004	−0.0021	0.0004	0.0015	0.0046	−0.0009	−0.0002
	(0.0707)	(0.1144)	(0.0199)	(0.0338)	(0.0759)	(0.1188)	(0.0218)	(0.0362)
	(0.0840)	(0.1106)	(0.0199)	(0.0339)	(0.0842)	(0.1109)	(0.0201)	(0.0339)
$(I, \mathbb{E}[n_i]) = (30, 300)$	−0.0005	0.0034	−0.0124	−0.0010	−0.0051	0.0067	−0.0083	−0.0029
	(0.2103)	(0.3364)	(0.0552)	(0.1033)	(0.2141)	(0.3486)	(0.0468)	(0.0872)
	(0.2155)	(0.2970)	(0.0388)	(0.0782)	(0.2170)	(0.2952)	(0.0388)	(0.0737)

Table 1.5: Comparison of statistical characteristics of full DR-GEE2 vs S-DR-GEE2. $\mathcal{R} = 2000$ replicate simulations.

	`geese`			Full DR-GEE2			S-DR-GEE2		
$(I, \mathbb{E}[n_i])$	$(30, 30)$	$(300, 30)$	$(30, 300)$	$(30, 30)$	$(300, 30)$	$(30, 300)$	$(30, 30)$	$(300, 30)$	$(30, 300)$
Convergence									
% PSM error only	—	—	—	4.22%	0.41%	7.97%	0.58%	0.10%	1.68%
% OM error only	—	—	—	9.03%	0.86%	11.80%	9.38%	0.77%	6.30%
% PSM or OM error	—	—	—	0.36%	0.00%	0.49%	0.12%	0.00%	0.11%
% Conditional TM error	0%	0%	26%	2.13%	0.00%	3.97%	1.23%	0.00%	0.41%
Run-time (sec)†									
PSM fitting	—	—	—	0.38	3.88	25.69	0.29	2.84	1.76
OM fitting	—	—	—	0.20	2.05	8.01	0.25	2.33	0.81
TM fitting	0.10	0.86	1174	0.40	4.24	27.59	0.31	3.14	1.53

Table 1.6: Algorithmic analysis of standard and stochastic DR-GEE2. $\mathcal{R} = 2000$ replicate simulations. Run-time values are computed on runs which converged. The conditional TM error is the error rate among simulations whence PSM and OM converged.

† Each replicate simulation was executed in R on a dual-core node on the Orchestra cluster supported by the Harvard Medical School Research Information Technology Group.

We also fit a complete-case TM in each replicate simulation using the `geese` command from the `geepack` package. We see that `geese` fits faster than our algorithms in the (30, 30) and (300, 30) cases, while our code runs far faster and leads to fewer errors in the (30, 300) case. Granted, the comparisons are not the most commensurate: `geese` performs all calculations in the C programming and wraps the results into R, while our implementation is fully in R, not to mention the additional time in incorporating the IPW or DR portions. On the other hand, our use of `geese` specifies a custom covariance structure for each cluster to handle the different treatment arms, while our implementation fully exploits analytical inverses of the equicorrelation structure.

1.5 Application to Sanitation Data

Guiteras et al. [2015] investigated the efficacy of alternative policies in encouraging use of hygienic latrines in developing countries. A total of 380 communities in rural Bangladesh were assigned to different marketing interventions – community motivation, subsidies, supply-side market, a combination of the three and a control group. The dataset contains 4768 individuals across 100 clusters with ten individual-level covariates (report diarrhea indicator X_1, male indicator X_2, age X_3, education indicator X_4, Muslim indicator X_5, Bengali indicator X_6, agricultor indicator X_7, stove indicator X_8, water pipes indicator X_9, phone indicator X_{10}) and five (excluding marketing intervention) cluster-level covariates (village population Z_1, # of doctors Z_2, % landless Z_3, % almost landless Z_4, % access electricity Z_5). The overall outcome missingness is 3.4%. Results based on a mixed-effect model suggested supply-side market alone did not increase hygienic latrine ownership (+0.3% points, p-value = 0.90). We reanalyzed this dataset with GEE2 approaches assuming that the outcomes are rMAR, letting $A_i = 1$ for supply-side market alone and $A_i = 0$ for control group. Due to the low outcome missingness, and in order to fully illustrate the strengths of our proposed methods, we induced additional missingness with Parzen's

method under the model

$$\begin{aligned}\text{logit}(\pi_{ij}^R) &= \beta_{0R} + \boldsymbol{\beta}_{XR}^\intercal \mathbf{X}_{ij} + \boldsymbol{\beta}_{ZR}^\intercal \mathbf{Z}_i + A_i(\beta_{AR} + \boldsymbol{\beta}_{AXR}^\intercal \mathbf{X}_{ij} + \boldsymbol{\beta}_{AZR}^\intercal \mathbf{Z}_i) \\ \tanh^{-1}(\rho_i^R) &= \alpha_{0R} + \boldsymbol{\alpha}_{ZR}^\intercal \mathbf{Z}_i + A_i(\boldsymbol{\alpha}_{AR} + \boldsymbol{\alpha}_{AZR}^\intercal \mathbf{Z}_i)\end{aligned}$$

with

$$\begin{aligned}\beta_{0R} &= 0.7, & \boldsymbol{\beta}_{XR}^\intercal &= (0, 0.3, 0.003, 0, 0.3, 0.3, 0.3, 0.3, 0, 0.3), \\ \boldsymbol{\beta}_{ZR}^\intercal &= (-0.00048, -0.0014, -0.407, -0.555), & \beta_{AR} &= -0.7, \\ \boldsymbol{\beta}_{AXR}^\intercal &= (0, 0, 0, 0, -0.6, -0.6, 0, -0.6, 0, 0), & \boldsymbol{\beta}_{AZR}^\intercal &= (0, 0.022, 0.904, 1.11), \\ \alpha_{0R} &= 0.35, & \boldsymbol{\alpha}_{ZR}^\intercal &= (0.0000645, 0.000190, 0.0543, 0.074), \\ \alpha_{AR} &= -0.10, & \boldsymbol{\alpha}_{AZR}^\intercal &= (0, -0.00291, -0.120, -0.148)\end{aligned}$$

The overall missingness is now 26%. Table 1.7 presents results upon fitting complete-case, $\mathcal{G}_1$ IPW-, $\mathcal{G}_2$ IPW-, AUG-, DR- GEE2. AUG-GEE2 is the augmentation of complete-case GEE2 instead of IPW-GEE2, and is included to provide insight to changes in parameter estimates.

Variables selected for the PSM and OM of the main effects were determined by backward stepwise logistic regression based on AIC, where the full model is a linear function of all covariates and the interactions terms between market intervention and all other covariates. We include all selected cluster-level covariates in the PSM and OM for the ICC (see Table 1.7). We experienced convergence issues in fitting the PSM and OM to the data when using full GEE2. To overcome this, we fitted 50 parallel stochastic GEE2 (described in Section 1.2.4), and averaged the convergent estimates. Complete-case analysis suggests non-significant supply-side causal effect (p-value ≈ 0.34), yet significantly different ICC between the two interventions (p-value ≈ 0.046). The two IPW-GEE2 methods see an increased magnitude in the causal effect, although still not significant (p-value ≈ 0.13 for both), and a decreased difference in ICC between the two interventions, which now becomes non-significant (p-value ≈ 0.18 for both) compared to complete-case analysis. AUG-GEE2 and DR-GEE2 both see another increase in the causal effect, which now results in significance (p-value < 0.01 in both cases), and remains non-significant for difference

	Estimates			Sandwich SE			p-value			Run-time (sec)†		
	β^*_{AY}	α^*_{0Y}	α^*_{AY}	β^*_{AY}	α^*_{0Y}	α^*_{AY}	β^*_{AY}	α^*_{0Y}	α^*_{AY}	PS	OM	TM
CC GEE2	0.154	0.082	0.093	0.163	0.017	0.046	0.344	< 0.01	0.046	—	—	0.82
$\mathcal{G}_1(\mathbf{R})$ IPW-GEE2	0.264	0.091	0.062	0.178	0.019	0.047	0.138	< 0.01	0.185	0.15	—	3.19
$\mathcal{G}_2(\mathbf{R})$ IPW-GEE2	0.275	0.095	0.065	0.182	0.019	0.048	0.130	< 0.01	0.170	2.75*	—	3.57
AUG-GEE2	0.466	0.091	0.022	0.073	0.014	0.017	< 0.01	< 0.01	0.208	—	2.52*	4.62
DR-GEE2	0.481	0.095	0.042	0.095	0.017	0.032	< 0.01	< 0.01	0.187	2.75*	2.52*	4.86

TM: $\text{logit}(\pi^*_i) = \beta^*_{0Y} + \beta^*_{AY} A_i$

$\text{atanh}(\rho^*_i) = \alpha^*_{0Y} + \alpha^*_{AY} A_i$

PSM: $\text{logit}(\pi^R_{ij}) = \beta_{0R} + \beta_{AR} A_i + \sum_{k\in\{2,3,5,6,7,8,10\}} \beta^{(k)}_{XR} X_{ijk} + \sum_{k\in\{1,2,3,4\}} \beta^{(k)}_{ZR} Z_{ik}$
$+ A_i \sum_{k\in\{5,6,8\}} \beta^{(k)}_{AXR} X_k + A_i \sum_{k\in\{2,3,4\}} \beta^{(k)}_{AZR} Z_{ik}$

$\text{atanh}(\rho^R_i) = \alpha_{0R} + \alpha_{AR} A_i + \sum_{k\in\{1,2,3,4\}} \alpha^{(k)}_{ZR} Z_{ik} + A_i \sum_{k\in\{2,3,4\}} \alpha^{(k)}_{AZR} Z_{ik}$

OM: $\text{logit}(\pi_{ij}) = \beta_{0Y} + \beta_{AY} A_i + \sum_{k\in\{1,2,3,4,5,8,9,10\}} \beta^{(k)}_{XY} X_{ijk} + \sum_{k\in\{1,2,3,4,5\}} \beta^{(k)}_{ZY} Z_{ik}$
$+ A_i \sum_{k\in\{1,3,8\}} \beta^{(k)}_{AXY} X_k + A_i \beta^{(5)}_{AZY} Z_{i5}$

$\text{atanh}(\rho_i) = \alpha_{0Y} + \alpha_{AY} A_i + \sum_{k\in\{1,2,3,4,5\}} \alpha^{(k)}_{ZY} Z_{ik} + A_i \alpha^{(5)}_{AZY} Z_{i5}$

Table 1.7: Effects of the supply side-market vs. control on the probability of hygienic latrine ownership in the sanitation data analysis [Guiteras et al., 2015] using the complete-case GEE2, IPW-GEE2 adjustment (non-adjusting and adjusting for missingness ICC), and DR-GEE2, assuming outcomes are rMAR.

* Fitted with 50 parallel stochastic GEE2, and averaging convergent estimates. Reported are median times among convergent estimates.

† Executed in R on a desktop with Intel(R) Core(TM) i5-4460 CPU 3.20GHz

ICCs between the two interventions. The results from AUG-GEE2 suggests that, even after randomization, a significant imbalance of baseline covariates remains, which is addressed by augmenting the outcome model. The resulting DR-GEE2 estimates are most affected by augmentation of the OM rather than adjustments for missingness. The inferences reached from DR-GEE2 on this induced dataset are the same as that of DR-GEE2 performed on the original dataset.

1.6 Discussion

In this paper, we proposed a stochastic algorithm to obtain the solutions to GEE2. This new algorithm substantially increases convergence rate and reduces run-times. It is in particular useful in settings whence either the number of clusters or the size of clusters is large. Accurate estimation of ICCs in general requires adequate number of clusters relative to the cluster size. When the cluster size is large relative to the number of clusters, the standard algorithm suffers from convergence issues. The stochastic algorithm alleviates this problem by performing the estimation on a subsample from each cluster for each iteration.

Another feature of S-GEE2 is the inclusion of the Hessian. Much of the literature derived from the Robbins-Monro framework does not incorporate the Hessian matrix into the iterations, instead relying on adaptive gradients and adaptive learning rates [Duchi et al., 2011; Nesterov, 1983; Zeiler, 2012]. Traditionally, Hessians are omitted because they are computationally intractable [Bottou, 2012]. But in the GEE2 framework, the Hessians are readily computable, and so are its stochastic variants. Each of these frameworks can be built upon each other to form hybrid methods, and indeed, comparisons of these different combinations would be interesting for future works.

We also proposed DR-GEE2 for estimating the marginal treatment effect and treatment-specific ICCs in cluster randomized trials. Our estimators are most useful in the settings where estimation of ICCs is the focus. If the interest is solely on the treatment effect on the outcomes, using working independence covariance matrix is an attractive approach due to its high efficiency in many settings and its simplicity in avoiding the need to estimate high-order association parameters. In the absence of missing data, standard GEE2 is highly efficient with a correctly specified working covariance structure. More concretely, the class of estimating functions which satisfy the canonical TM in Eq 1.2 and are regular asymptotically linear (RAL) must be of the form $\mathbf{0} = \sum_{i=1}^{I} h(A_i)E_i$. The choice of index function $h(A_i) = D_i^{\intercal}V_i^{-1}$, which reduces back to GEE2, results in the efficient score for the canonical TM, hence attaining the minimum asymptotic variance RAL estimator for $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$ [Chamberlain, 1986]. However, in the case of IPW-GEE2 or DR-GEE2, this choice is no

longer optimal and the actual $h_{\text{opt}}(A_i)$ to achieve the efficient score is far more complicated [Stephens et al., 2014]. Stephens et al. [2014] showed in simulation studies the efficiency gains from using $h_{\text{opt}}(A_i)$ are modest and very sensitive to the correct specification of all components that comprise $h_{\text{opt}}(A_i)$, which in practice is nearly impossible to achieve. With little computational support for $h_{\text{opt}}(A_i)$ and no theoretical support for $h(A_i) = D_i^{\intercal} V_i^{-1}$, one might just simplify the entire process by letting V_i have an independence covariance structure altogether. Our simulation studies in Section 1.4 also provide corroborative evidence supporting the use of an independence covariance structure when estimating the first-order effects.

Although the discussion centered around cluster randomized trials, the DR-GEE2 estimator can be used in other settings when estimation of ICCs is of interest such as in reliability and agreement studies. We focused our discussion on binary outcomes, but the approach can be adapted to other types of exponential family outcomes in a straightforward manner by modifying the link function and variance function for the likelihood in question. When outcomes within clusters are not equicorrelated, our ICC estimators marginalize out factors which contribute to the non-exchangeable structure and returns an estimate which can be construed as an "average" correlation.

In the presence of informative missing data, the correlation among missingness indicators needs to properly accounted for to arrive at the consistent estimators for the association parameters. We assumed rMAR in the current work. Future research on further relaxing this assumption would be useful.

Chapter 2

Linear mixed effects models with Fleishman-distributed variance components

When concerned with estimation and inference of variance components and functions thereof (e.g. the intraclass correlation, higher moments, etc), the standard random effects model relies heavily on the normality of the random effects and error distributions. We relax these assumptions by endowing each variance component with a Fleishman distribution, a flexible distribution which accounts for the third and fourth cumulants of a random variable. The simplicity and speed in simulating from the Fleishman distribution allow us to construct confidence intervals based on a Fleishman parametric bootstrap on the variance components. We also develop a test of normality for each of the variance components akin to the Jarque-Bera test by comparing the third and fourth cumulants to that of a normal distribution, which flows organically from the presented framework due to the need in estimating the higher moments of the variance components. We compare the performance of our methodology with existing techniques in simulation studies and illustrate our methods to the Childhood Adenotonsillectomy Trial (CHAT) sleep electroencephalogram (EEG) data in quantifying the agreement among different signal densities.

2.1 Introduction

In many biomedical investigations, parameters of interest are functions of higher order moments reflecting finer distributional characteristics. For example, the intraclass correlation coefficient (ICC), a function of between- and with- cluster variances, is an essential parameter in the design and analysis of cluster randomized trials and for assessing reliability of ratings or agreement of multiple measurements. Other examples are skewness and kurtosis, which can be used to assess departures from normality [Kim and White, 2004]. One of the popular methods in handling correlated outcome data is linear mixed models (LMM) [Laird and Ware, 1982a]. The one-way LMM, also known as the variance components model, takes the form

$$Y_{ij} = \beta^\intercal X_{ij} + U_i + \epsilon_{ij}, \qquad i = 1, \cdots, I; \quad j = 1, \cdots, J_i \tag{2.1}$$

It is commonly assumed that the random effects $U_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$ and subject-specific errors $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$. Inference on the fixed effect coefficients β have been shown to be robust to nonnormality of U_i or ϵ_{ij} [Butler and Louis, 1992; McCulloch and Neuhas, 2011]. This, however, is not true for higher-order quantities, such as the ICC $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$ or skewness and kurtosis (third and fourth cumulants) of U_i and ϵ_{ij}. One robust approach in estimating higher-order quantities is to allow a flexible distribution on U_i or ϵ_{ij} in a LMM. There is a rich literature on fitting flexible distributions for U_i while letting ϵ_{ij} remain normally distributed, all of which resort to fitting the REML or MLE using an EM-type algorithm; Verbeke and Lesaffre [1996] and Ghidey et al. [2004] considered a mixture of normals, while Zhang and Davidian [2001] and Lin and Lee [2008] considered a skew-normal distribution. Nevertheless, these methods do not allow a flexible distribution on ϵ_{ij}. Arellano-Valle et al. [2005] considered a skew-normal distribution on both random effects and errors, but found no tractable MLE nor REML algorithm to fit both distributions, only on one or the other.

Another robust approach in estimating ICCs (or other higher-order moments) is based on second-order (or higher-order) generalized estimating equations (GEE2) [Liang and Zeger, 1992; Zhao and Prentice, 1990], and in general M-estimation. GEE-type estima-

tors are attractive because it does not require specification of the full likelihood, and is semiparametric efficient when the working covariance matrix is correctly specified [Newey, 1990]. However, GEE2's may perform poorly when the number of clusters is small [Huang et al., 2016]. Furthermore, its use has been hindered by considerable computational burden and poor convergence rates [Evans et al., 2001; Sutradhar, 2003; Ziegler et al., 1998].

Asymptotic results for the aforementioned flexible LMM and GEE2 are available, and thus we may construct CI's for the ICC, or other estimators based off the second moments, through analytical means. In practice, distributions of these flexible LMM and GEE2 estimators tend to be skewed and require either large samples or *a priori* stabilizing transformations to better approximate normality. One solution is bootstrap, since this captures the inherent skewness through the empirical distribution of the data, but doing so with GEE2 or flexible LMM MLE is far too time consuming, the former due to solving a large stack of estimating equations, and the latter due to performing several levels of optimization in order to fit model parameters. Our proposed method is an amalgamation of flexible distributions and bootstrap, but instead of MLE, we use method of moments. This makes the bootstrapping far faster, while potentially sacrificing some efficiency loss compared to MLE. But, as we will demonstrate in extensive simulation studies, our method can lead to significantly shorter CI's than that produced from GEE2.

In this paper, we propose a parametric bootstrap approach based on the Fleishman distribution to make inferences about estimators derived from clustered data using methods of moments. We choose the Fleishman distribution because it is a distribution that is flexible to accomondate many parametric distributions and with moments readily computable and easy to simulate from. This approach is advantageous over non-parametric bootstrap for clustered data in the following ways: 1) Nonparametric bootstrap in general fails to capture more extreme observations that do not appear in the observed data; this point is not as vital for first-order estimates such as the median, but would be for the higher moments or extreme quantiles due to their sensitivity to outliers. Because our method is inherently parametric, these extreme values are given non-zero probability to be occur in the bootstrap sample. 2) One commonly-used bootstrap approach for clustered data is to sample

the clusters (and all observations within the cluster) with replacement. In the imbalanced case where cluster sizes vary, this does not preserve the structure of the original data and leads to induced variability from varying cluster sizes.

A second purpose of our paper proposes improved methods for assessing agreement. The ICC, or other moment-based agreement measure such as Pearson's correlation coefficient r, operates on a linear scale and may provide spurious values of high agreement not necessarily due to genuine agreement, but possibly as a result of outliers or different scales of the metrics. For example, given $\mathbf{y}_1 = (-10, -9, \cdots, 9, 10)$ and $\mathbf{y}_2 = (1, 0, 1, 0, \cdots, 0, 1)$, we may compute the sample Pearson's $\hat{r}_{\mathbf{y}_1,\mathbf{y}_2} = 0$. However, if we append $\mathbf{y}_1' = (\mathbf{y}_1, 100)$ and $\mathbf{y}_2' = (\mathbf{y}_2, 100)$, then $\hat{r}_{\mathbf{y}_1',\mathbf{y}_2'}$ jumps to 0.962. Various agreement indices have been proposed to guard against such undesirable cases. For example, Cohen's κ and Scott's π define agreement in terms of probabilities instead of moments, and thereby are more robust to outliers. However, Gwet [2002] and Strijbos et al. [2006] detail how these probabilistic indices can be misleading due to an inappropriate way these indices compute the probability of agreement. Krippendorff's α [Krippendorff, 2004] is a very general agreement index which allows for missing data and includes a weighting function. Many other agreement indices, such as Cohen's κ, Scott's π, and the ICC, are recovered through an appropriately defined weighting function. However, the form of Krippendorff's α includes several layers of nested sums, hence is computationally expensive to compute for large samples, let alone CI's. Our proposed index combines elements of Krippendorff's α to the ICC in order to be robust to outliers and flexible, but computationally simple.

Section 2.2 presents our proposed inference procedure under a linear mixed effects model. We first provide a brief overview of our chosen flexible distribution (Fleishman distribution) in Section 2.2.1, describe the inferential procedure in Section 2.2.2, and finally describe a normality test for U_i and ϵ_{ij} in Section 2.2.3. In Section 2.4, we describe our modified ICC for agreement studies. We evaluate the performance of our methods with simulations in Section 2.3 and apply our methods to analyze Electroencephalography (EEG) data from the Childhood Adenotonsillectomy Trial (CHAT) in Section 2.5. An EEG records the electrical signals and is commonly to used to diagnose sleep disorders. EEG

Signals were obtained through several electrodes on different locations (called "channels") on the head, which measure power density of different wave frequencies $(\delta, \theta, \alpha, \sigma, \beta)$; the frequencies listed are ordered from more sleepy to more alert. Our goal is to quantify agreement of signals from electrodes at different locations. Power density is heavily right-skewed and leptokurtic (i.e heavy-tailed), even on the log scale, and thus normality is unlikely to hold and transformations are needed to properly quantify agreement. We end with a discussion in Section 2.6.

2.2 Proposal I: REMM-F for non-normal mixed models

2.2.1 Background: Fleishman distribution

Because we use method of moments, we select a flexible distribution with moments readily computable and easy to simulate from, rather than carefully devising a distribution with a convenient density, as one would with MLE in mind. One candidate is the so-called Fleishman distribution [Fleishman, 1978]. A random variable Y is said to follow the Fleishman distribution, Fleish(a, b, c, d), if $Y \stackrel{\mathcal{D}}{=} a + bZ + cZ^2 + dZ^3$, where Z is standard normal and $b > 0$ for identifiability. This distribution reduces back to a normal distribution upon setting $c = d = 0$. Fleishman's distribution has historically been used in simulation studies due to its flexibility in choosing random variables with desired first four moments [Headrick, 2009], and is perhaps the most popular due to its ease and speed to implement and simulate, requiring just the generation of normal random variables. To the best of our knowledge, Demirtas and Hedeker [2008] were the first to apply this distribution as an inferential tool rather than a simulation tool, where it was used to generate the errors in a multiple imputation procedure. Our work will utilize the Fleishman distribution in the context of bootstrapping and extend to both random effects and errors. The flexibility of the Fleishman distribution is illustrated in Section 2.2.2, where we observe that the Fleishman distribution covers a large portion of the skewness-elongation plane and is able

to approximate many common distributions.

2.2.2 Notation and Methods

We proceed in describing our proposed method, which we call restricted method of moments with Fleishman-distributed random effects and errors (REMM-F). Let us generalize the LMM in Eq 2.1 with the following one-way semiparametric mixed model:

$$Y_{ij} = g(X_{ij}) + U_i + \epsilon_{ij}, \quad i = 1, \cdots, I; j = 1, \cdots, J_i \tag{2.2}$$

where $g : \mathbb{R}^p \to \mathbb{R}$ is an unknown, smooth mean function. The first step, like in REML, is to subtract away the fixed effects $g(X)$ and work with the residuals $\nu_{ij} = Y_{ij} - g(X_{ij})$. Unlike REML, which performs MLE on the calculated residuals, we carry out method of moments instead. We call inference based on the normality assumption as REMM-N. If U_i and ϵ_{ij} were both normally distributed, then the REMM estimators (based on the sample variances) are REML. If not, then we propose the use of REMM-F, delineated as follows: (1) estimation of the fixed-effects, (2) estimation of the first four moments of U_i and ϵ_{ij}, (3) estimation of Fleishman parameters (a, b, c, d) for U_i and ϵ_{ij}, and (4) inference.

1. Estimation of the fixed effects

Setting $g(X) = X\beta$ reduces to the standard one-way LMM, from which we can use (weighted) least-squares. We can also nonparametrically estimate g, for example, with kernel linear regression. See Henderson and Ullah [2005]; Ke and Wang [2001]; Lin and Carroll [2006]; Wang [2003]; Wu and Zhang [2002] for many other nonparametric estimators for g. If $\sigma_u^2, \sigma_\epsilon^2$ are finite, then the estimators from all these works have finite standard errors. Provided a consistent estimator $\hat{g}$, we compute the residuals $\hat{\nu}_{ij} = Y_{ij} - \hat{g}(X_{ij})$ and proceed.

2. Estimation of the first four moments of U_i and ϵ_{ij}

The standard ANOVA estimators for the variances σ_ϵ^2 and σ_u^2 are

$$s_\epsilon^2 = \frac{\mathrm{SE}_2}{J-I}, \qquad s_u^2 = \max\left\{\frac{I-1}{Q_0}\left(\frac{\mathrm{ST}_2}{I-1} - s_\epsilon^2\right), 0\right\} \tag{2.3}$$

respectively, where $J = \sum_{i=1}^I J_i$, $Q_0 = J - J^{-1}\sum_{i=1}^I J_i^2$, and

$$\begin{aligned} \mathrm{SE}_p &= \sum_{i=1}^I \sum_{j=1}^{J_i} (\hat{\nu}_{ij} - \bar{\nu}_{i\cdot})^p, & \mathrm{ST}_p &= \sum_{i=1}^I J_i (\bar{\nu}_{i\cdot} - \bar{\nu}_{\cdot\cdot})^p \\ \bar{\nu}_{i\cdot} &= J_i^{-1} \sum_{j=1}^{J_i} \hat{\nu}_{ij}, & \bar{\nu}_{\cdot\cdot} &= J^{-1} \sum_{i=1}^I J_i \bar{\nu}_{i\cdot} \end{aligned}$$

for $p = 2, 3, 4$. Furthermore, Teuscher et al. [1994] derived the following biased-adjusted estimators for the skewness and kurtosis:

$$\begin{aligned} \hat{\gamma}_\epsilon &= \frac{\mathrm{SE}_3}{s_\epsilon^3 Q_1} \\ \hat{\gamma}_u &= \frac{1}{s_u^3}\left(\frac{\mathrm{ST}_3}{Q_3} - \frac{Q_2}{Q_1 Q_3}\mathrm{SE}_3\right) \\ \hat{\kappa}_\epsilon &= \max\left\{\frac{\mathrm{SE}_4}{s_\epsilon^4 Q_4} - \frac{Q_5}{Q_4}, \hat{\gamma}_\epsilon^2 - 2\right\} \\ \hat{\kappa}_u &= \max\left\{\frac{1}{s_u^4 Q_6}\left(\mathrm{ST}_4 - \frac{Q_8}{Q_4}\mathrm{SE}_4\right) + \frac{s_\epsilon^4 Q_5 Q_8}{s_u^4 Q_4 Q_6} - \frac{s_\epsilon^4 Q_9 + s_u^2 s_\epsilon^2 Q_{10} + s_u^4 Q_7}{s_u^4 Q_6}, \hat{\gamma}_u^2 - 2\right\} \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} Q_1 &= J - 3I + 2\sum_i \frac{1}{J_i}, & Q_2 &= \frac{2-3I}{J} + \sum_i \frac{1}{J_i} \\ Q_3 &= J - \frac{3}{J}\sum_i J_i^2 + \frac{2}{J^2}\sum_i J_i^3, & Q_4 &= J - 4I + 6\sum_i \frac{1}{J_i} - 3\sum_i \frac{1}{J_i^2} \\ Q_5 &= 3\left(J - 2I + \sum_i \frac{1}{J_i}\right), & Q_6 &= J - \frac{3}{J^3}\sum_i J_i^4 + \frac{6}{J^2}\sum_i J_i^3 - \frac{4}{J}\sum_i J_i^2 \\ Q_7 &= 3\left(J - \frac{3}{J^3}\left(\sum_i J_i^2\right)^2 + \frac{4}{J^2}\sum_i J_i^3 - \frac{2}{J}\sum_i J_i^2\right), & Q_8 &= 3\frac{2I-1}{J^2} - \frac{4}{J}\sum_i \frac{1}{J_i} + \sum_i \frac{1}{J_i^2} \\ Q_9 &= 3\left(\sum_i \frac{1}{J_i} + \frac{1-2I}{J}\right), & Q_{10} &= 6\left(\frac{I+1}{J^2}\sum_i J_i^2 + I - 3\right) \end{aligned}$$

The skewness and kurtosis estimators in Eq 2.4 are consistent as $I \to \infty$, but they are not unbiased, since they are a ratio of random variables. Had we replaced s_u, s_ϵ in the denominators of Eq 2.4 with population values $\sigma_u, \sigma_\epsilon$, then the equations would be unbiased. We require at least a few of the $J_i \geq 3$ to obtain estimates of the skewness, otherwise $\mathrm{SE}_3 = \mathrm{ST}_3 = Q_1 = 0$. A counterintuitive, yet welcomed, property of Teuscher's estimators is that we do not require $J_i \geq 4$, as one might expect to obtain estimates of the kurtosis.

3. Estimation of Fleishman parameters (a, b, c, d) for U_i and ϵ_{ij}

Let $W \sim \mathrm{Fleish}(a, b, c, d)$. Without loss of generality, assume $\mathbb{E}[W] = 0, \mathrm{Var}(W) = 1$. Then the moment conditions [Fleishman, 1978] are

$$\begin{aligned} f_1(a,b,c,d) &= a + c &&= 0 \\ f_2(a,b,c,d) &= b^2 + 6bd + 2c^2 + 15d^2 - 1 &&= 0 \\ f_3(a,b,c,d) &= 2c(b^2 + 24bd + 105d^2 + 2) - \gamma &&= 0 \\ f_4(a,b,c,d) &= 24[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)] - \kappa &&= 0 \end{aligned} \tag{2.5}$$

where $\gamma = \mathbb{E}[(W - \mathbb{E}[W])^3]/\sigma^3$ and $\kappa = \mathbb{E}[(W - \mathbb{E}[W])^4]/\sigma^4 - 3$ are the skewness and kurtosis of W, respectively. These equations do not have a solution for all values of (γ, κ); indeed, the region for which a solution exists is called the *skewness-elongation region*, which in the case of the Fleishman distribution, is approximately

$$\kappa \geq 0.042717|\gamma|^4 - 0.129624|\gamma|^3 + 1.661833|\gamma|^2 - 1.147301, \qquad -3 \leq \gamma \leq 3 \tag{2.6}$$

This approximation is the degree four polynomial least-squares fit on the values enumerated in Headrick and Sawilowsky [2000]. In judging the flexibility of a distribution, one visual metric is to compare its skewness-elongation bound against that of the theoretical lower bound for all probability distributions: $\kappa \geq \gamma^2 - 2$ (in fact, equality is impossible for continuous distributions). Refer to Figure 2.1 for a plot of the theoretical bound, the Fleishman bound, and locations of selected distributions in the skewness-elongation plane; the

Fleishman distribution covers a large portion of the plane and can be used to approximate many common distributions.

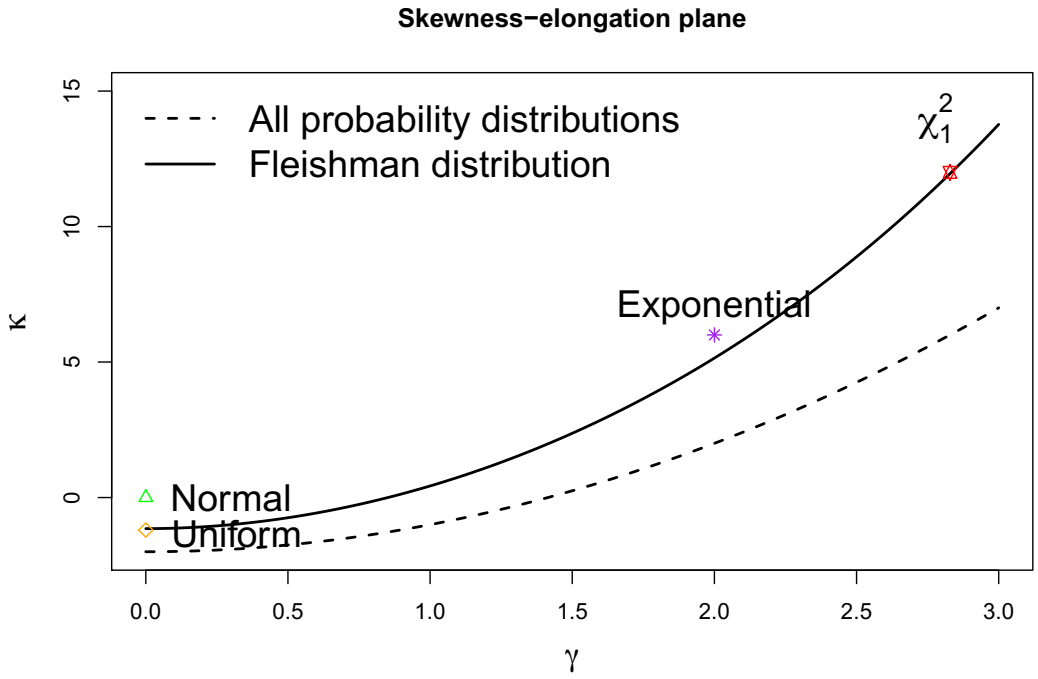


Figure 2.1: Skewness-elongation bounds and locations of select distributions within the skewness-elongation plane. Severe platykurtic distributions, such as the uniform distribution, fall under the Fleishman bound.

Let $U_i \sim \text{Fleish}(\theta_u)$, where $\theta_u = (a_u, b_u, c_u, d_u)$, and similarly $\epsilon_{ij} \sim \text{Fleish}(\theta_\epsilon)$. If a solution to Eq 2.5 exists, parameter estimation from the moments is straightforward. If not, we could simply increase the kurtosis $\kappa \mapsto \kappa'$ until (γ, κ') lies on the Fleishman bound, as in Headrick [2009]. Luo [2011] generalizes the calculation to an optimization problem

$$\underset{b,c,d}{\text{argmin}} \left\{ f_2^2(a,b,c,d) + f_3^2(a,b,c,d) + f_4^2(a,b,c,d) \right\} \tag{2.7}$$

where f_2, f_3, f_4 are defined in Eq 2.5, and $a = c$ per the restriction given by f_1. Leapfrogging from Luo's idea, a more equitable method to "share the bias" is akin to a generalized method

of moments procedure:

$$\underset{b,c,d}{\operatorname{argmin}} \left\{ w_2 f_2^2(a,b,c,d) + w_3 f_3^2(a,b,c,d) + w_4 f_4^2(a,b,c,d) \right\} \tag{2.8}$$

Here, the weights w_2, w_3, w_4 determine the importance one puts on each equation within the system; a higher weight for f_i reduces the bias for the ith cumulant. Luo's method is recovered by setting $(w_2, w_3, w_4) = (1, 1, 1)$, while Headrick's method is recovered with $(w_2, w_3, w_4) = (\infty, \infty, 1)$, where we define $\infty \times 0 = 0$. Since estimators become increasingly less reliable for higher moments, we recommend selecting $w_2 \geq w_3 \geq w_4$. For the examples to follow, we set $w_1 = 3\max(1, 0.5|\hat{\gamma}_u|, 0.25|\hat{\kappa}_u|), w_2 = 0.5$ and $w_3 = 0.25$ so that at least $3/(3 + 0.5 + 0.25) = 80\%$ of the weighting is towards the variance, and twice the weighting on the skewness over the kurtosis.

4. Inference

After performing moment matching, we are equipped with estimators $\hat{g}$, $\hat{\theta}_u = (\hat{a}_u, \hat{b}_u, \hat{c}_u, \hat{d}_u)$, and $\hat{\theta}_\epsilon = (\hat{a}_\epsilon, \hat{b}_\epsilon, \hat{c}_\epsilon, \hat{d}_\epsilon)$. We proceed with a parametric bootstrap. For the bth replicate, $b = 1, \cdots, B$:

1. Sample $g^{(b)} \sim F_{\hat{g}}$, $U_i^{(b)} \sim \text{Fleish}(\hat{\theta}_u)$, $\epsilon_{ij}^{(b)} \sim \text{Fleish}(\hat{\theta}_\epsilon)$, where the indices i, j range according to the sizes of the original data.

2. Construct pseudo-data $Y_{ij}^{(b)} = g^{(b)}(X_{ij}) + U_i^{(b)} + \epsilon_{ij}^{(b)}$.

3. Compute statistic of interest $T^{(b)} = T(\mathbf{Y}^{(b)}, \mathbf{X})$.

From $\{T^{(1)}, \cdots, T^{(B)}\}$, we can calculate bootstrapped variance estimates and construct CI's. Standard choices for interested statistics T are the main-effect estimator $\hat{g}$, the variance / skewness / kurtosis estimators in Eqs 2.3 and 2.4, or functions of higher moments, such as the ICC.

We provide heuristics on situations where REMM-F would be preferred over REMM-N. For simplicity, assume the balanced case $J_i = J$. As discussed in the Introduction, the main effects are robust to misspecification of the distributions for U_i and ϵ_{ij}; specifically,

$\hat{\beta}$ fitted using REML in the one-way LMM in Eq 2.1 has the asymptotic distribution $\sqrt{I}(\hat{\beta}-\beta) \xrightarrow{\mathcal{D}} N(0, (\mathbb{E}[X_{ij}]^\intercal \Sigma^{-1} \mathbb{E}[X_{ij}])^{-1})$, where $\Sigma = \sigma_u^2 \mathbf{I} + \sigma_\epsilon^2 \mathbf{1}\mathbf{1}^\intercal$, $\mathbf{I}$ is the identity matrix, and $\mathbf{1}$ is a vector of ones. Hence, the asymptotic distribution of $\hat{\beta}$ depends on U_i, ϵ_{ij} only through $\sigma_u^2, \sigma_\epsilon^2$ and not any other finer distributional characteristics of U_i, ϵ_{ij}. In the general case, Li and Xue [2015] show that kernel regression for g also depends on U_i, ϵ_{ij} only through $\sigma_u^2, \sigma_\epsilon^2$. We conclude that the choice of distribution for U_i, ϵ_{ij} needs to account only up to the second moments in order to provide asymptotically correct inference for β, which normal LMM does account for.

So, what quantities would REMM-F be adequate in estimating that REMM-N would not? For σ_u^2 and σ_ϵ^2, it has been shown [Jiang, 2005; Li and Xue, 2015] that

$$\sqrt{I}(s_u^2 - \sigma_u^2) \xrightarrow{\mathcal{D}} N\left(0, \mathrm{Var}(U_i^2) + \frac{4\sigma_u^2\sigma_\epsilon^2(J-1) + 2\sigma_\epsilon^4}{J(J-1)}\right)$$
$$\sqrt{IJ}(s_\epsilon^2 - \sigma_\epsilon^2) \xrightarrow{\mathcal{D}} N\left(0, \mathrm{Var}(\epsilon_{ij}^2) + \frac{2\sigma_\epsilon^4}{J-1}\right)$$

as $I \to \infty$ with J fixed. Since $\mathrm{Var}(U_i^2)$ and $\mathrm{Var}(\epsilon_{ij}^2)$ depend on quantities up to the fourth moments, valid inferences for $\sigma_u^2, \sigma_\epsilon^2$ (and therefore, the ICC $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$) require accurate information up to the fourth moments, which REMM-F does. Our approach strikes a natural balance between validity and efficiency, accounting for just enough additional parameters for valid inference on second-order parameters, which would presumably provide shorter CI lengths while remaining at nominal coverage levels, even in non-normal situations. Since the Fleishman distribution incorporates the normal distribution as a special case, we anticipate that CI's for ρ would not be substantially longer when U_i and ϵ_{ij} are truly normal.

Based on the moments paradigm, we should not expect valid CI's for higher moment quantities, such as $\gamma_u, \gamma_\epsilon, \kappa_u$, and κ_ϵ, since the variances of their respective estimators $\hat{\gamma}_u, \hat{\gamma}_\epsilon, \hat{\kappa}_u$, and $\hat{\kappa}_\epsilon$ are a function of moments up to the 6th or 8th order, for which the Fleishman distribution does not account for, so coverage levels depend heavily on how well the Fleishman distribution approximates the underlying generating distribution. Nevertheless, the relative performance is superior to that of REMM-N, since estimators for the skewness and kurtosis are not even consistent in non-normal situations, let alone possessing

asymptotically correct coverages. Little literature exists on the topic in constructing CI's for the skewness and kurtosis for U_i and ϵ_{ij}, thus REMM-F provides a good starting point.

2.2.3 Testing normality of U_i and ϵ_{ij}

While the literature on normality tests is plentiful for cross-sectional or time series models, results for mixed models are quite scarce. A natural complication is that, unlike their cross-section or time-series counterparts with just errors ϵ_{ij}, lack of normality in mixed models may arise from ϵ_{ij} or the unobserved U_i, or both. Previous works include that of Meintanis [2011] and Galvao et al. [2013], but each of these methods only handle balanced data. One workaround for imbalanced designs is to obtain within-subject residuals $\hat{\epsilon}_{ij}$ and conditional modes $\hat{U}_i$ from maximum-likelihood estimation, and then conduct a Shapiro-Wilk (SW) or Anderson-Darling (AD) test. While this method works well when both U_i and ϵ_{ij} are normal, inflated Type I error rates arise when at least one is non-normal. This is due to, when one of U_i or ϵ_{ij} is non-normal, fitting with a normal induces a "spill-over" of the non-normality onto the other.

Our solution to both imbalance and "spill-over" makes use of estimators $\hat{\gamma}_u, \hat{\kappa}_u, \hat{\gamma}_\epsilon, \hat{\kappa}_\epsilon$. That is, we test if random effects or subject-specific errors share the same skewness and kurtosis as that of a normal random variable: $H_0 : \gamma = 0$ & $\kappa = 0$ vs $H_1 : \gamma \neq 0$ or $\kappa \neq 0$. This hypothesis is the same one considered by Galvao et al. [2013], where they derived the asymptotic distributions for estimators similar $\hat{\gamma}_u, \hat{\kappa}_u, \hat{\gamma}_\epsilon, \hat{\kappa}_\epsilon$, but only for balanced data. Our strategy will instead bootstrap p-values, which have the benefit in also accounting for unbalanced data. To generate the distribution of $(\hat{\gamma}, \hat{\kappa})$ under H_0, we first compute s_u^2 and s_ϵ^2 and perform a normal distribution parametric bootstrap for B replicates to obtain boostrapped skewness and kurtosis values for U_i and ϵ_{ij} under H_0, say $\{(\hat{\gamma}^0_{ub}, \hat{\kappa}^0_{ub})\}_{b=1}^B$ and $\{(\hat{\gamma}^0_{\epsilon b}, \hat{\kappa}^0_{\epsilon b})\}_{b=1}^B$. We estimate the joint density under H_0, say $f^0_{\gamma_u,\kappa_u}$ and $f^0_{\gamma_\epsilon,\kappa_\epsilon}$, with a kernel-density estimator (KDE) based off the bootstrapped samples, say $\hat{f}^0_{\gamma_u,\kappa_u}$ and $\hat{f}^0_{\gamma_\epsilon,\kappa_\epsilon}$; see Scott [1995] and Wand and Jones [1992] for information on KDE's. Henceforth, we will fit our KDE's with a Gaussian kernel with bandwidth matrix computed with plug-in selectors.

Finally, we compute p-values

$$p_\epsilon = \int_{(\gamma,\kappa):\hat{f}^0_{\gamma_\epsilon,\kappa_\epsilon}(\gamma,\kappa)<\hat{f}^0_{\gamma_\epsilon,\kappa_\epsilon}(\hat{\gamma}_\epsilon,\hat{\kappa}_\epsilon)} \hat{f}^0_{\gamma_\epsilon,\kappa_\epsilon}(\gamma,\kappa)d\gamma d\kappa$$

and similarly for p_u. Note that this method is generalizable to any parameter θ with statistic $\hat{\theta}$ and boostrapped statistics under the null $\{\hat{\theta}^0_b\}_{b=1}^B$. We term our specific procedure as KDE-boot.

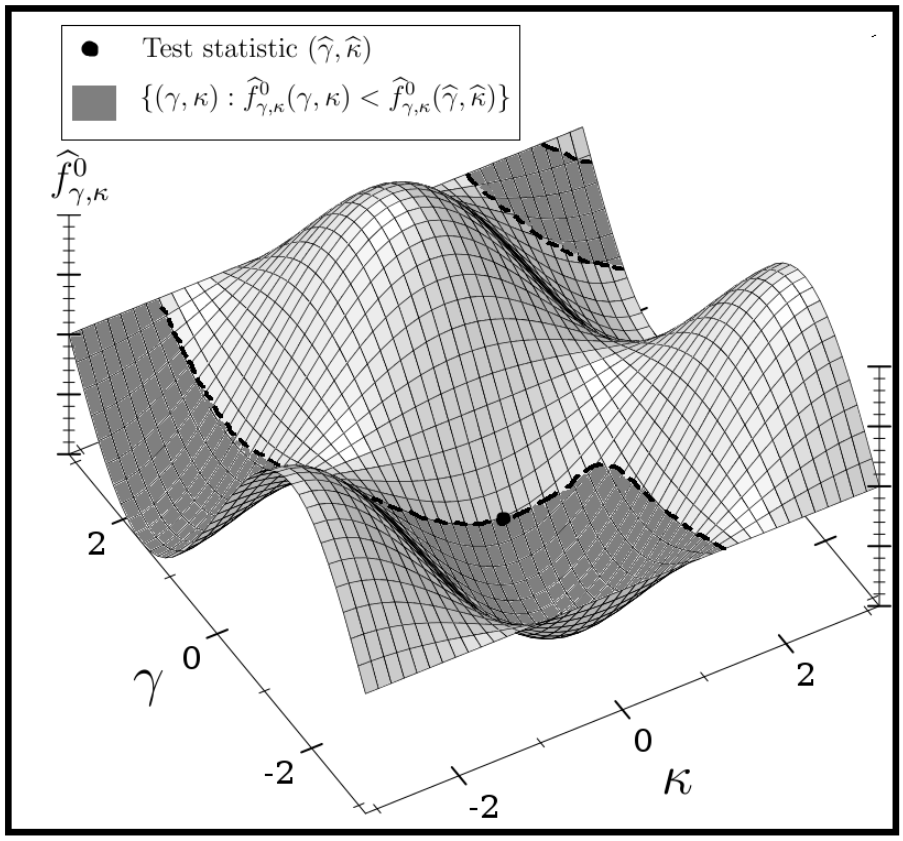


Figure 2.2: Visualization of KDE-boot in action. The shaded dark gray area represents regions of more extreme values than our test statistic under H_0, which be the integration region in calculating the p-value.

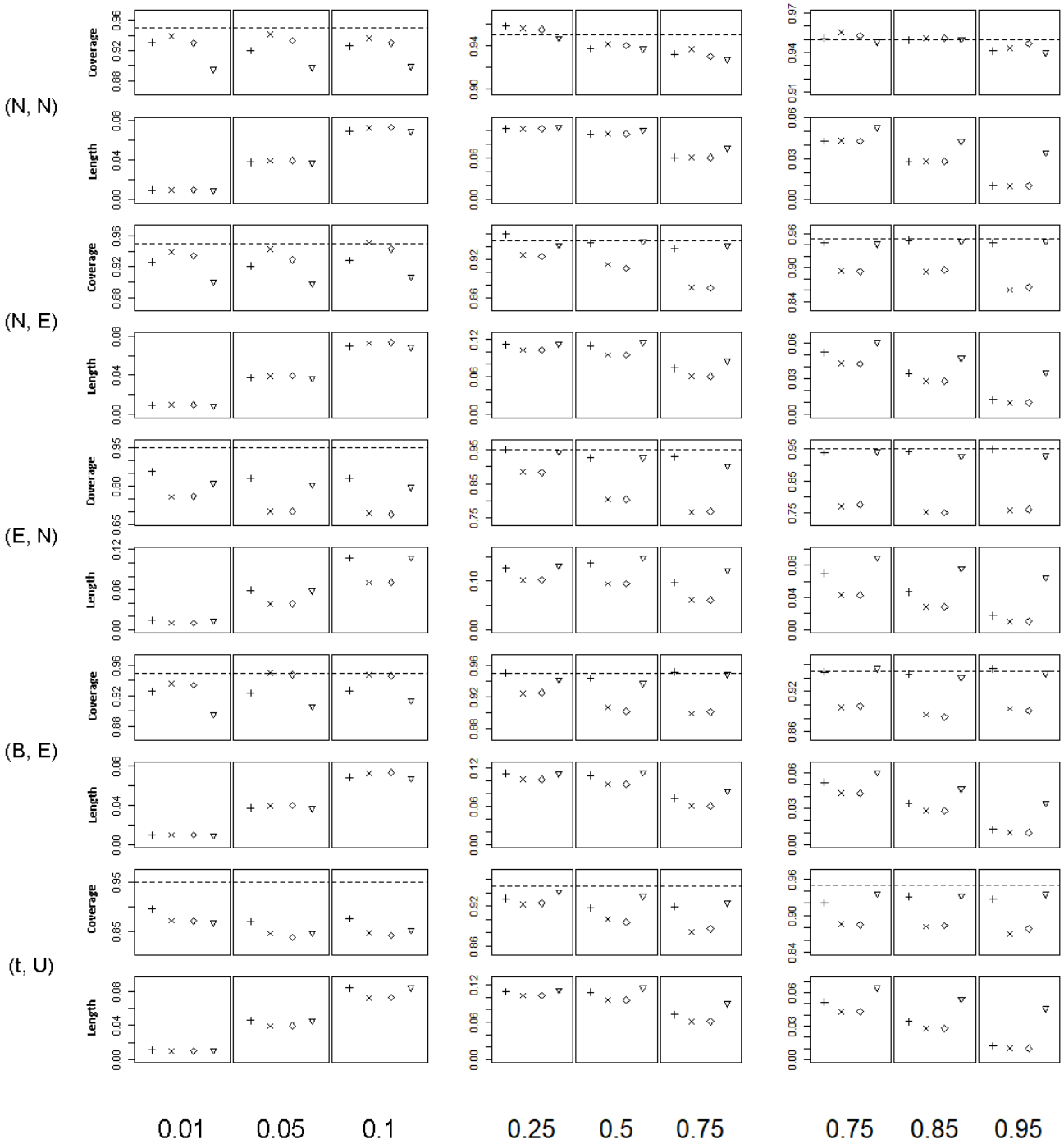
2.3 Simulation

We conduct simulation studies to investigate the performance of our proposed methods. We report here results for the model $Y_{ij} = \beta_0 + \beta_1 X_{ij} + U_i + \epsilon_{ij}$, with $(\beta_0, \beta_1) = (1, 2)$ and cluster scenarios (I, J_i) = (50, Unif{350,450}), (500, Unif{3,4}), (1000, Unif{3,4}), where Unif$\{a, b\}$ denotes the discrete uniform distribution over $\{a, a+1, \cdots, b\}$. Within each cluster scenario, we generate U_i and ϵ_{ij} to produce ICC levels $\rho = (0.01, 0.05, 0.10)$, $(0.25, 0.50, 0.75)$, $(0.75, 0.85, 0.95)$, respectively, with mean-zero centered combinations (U_i, ϵ_{ij}) of

1. (Normal, Normal), corresponding to $(\gamma_u, \kappa_u) = (0, 0)$ and $(\gamma_\epsilon, \kappa_\epsilon) = (0, 0)$
2. (Normal, Exponential), corresponding to $(\gamma_u, \kappa_u) = (0, 0)$ and $(\gamma_\epsilon, \kappa_\epsilon) = (2, 6)$
3. (Exponential, Normal), corresponding to $(\gamma_u, \kappa_u) = (2, 6)$ and $(\gamma_\epsilon, \kappa_\epsilon) = (0, 0)$
4. (Beta(5,2), Exponential), corresponding to $(\gamma_u, \kappa_u) = (-0.60, -0.12)$ and $(\gamma_\epsilon, \kappa_\epsilon) = (2, 6)$
5. (t_7, Uniform), corresponding to $(\gamma_u, \kappa_u) = (0, 2)$ and $(\gamma_\epsilon, \kappa_\epsilon) = (0, -1.2)$

There are a total of $3 \times 3 \times 5 = 45$ scenarios. The three cluster scenarios represent roughly a large-scale cluster randomized trial (CRT), a medium-scale reliability/agreement study (MRS), and a large-scale reliability/agreement study (LRS).

Figure 2.3 displays the average lengths and empirical coverages of ICC CI's of our proposed REMM-F, bootstrap REMM-N, Smith's method [Smith, 1956], and GEE2. Smith's method assumes normality of observations and constructs CI's based on the asymptotic distribution of $\hat{\rho} = s_u^2/(s_u^2 + s_\epsilon^2)$; it was demonstrated to provide the best overall coverage and interval lengths under normality in simulation studies performed in Donner [1986]. Across a wide range of settings considered in Figure 2.3, REMM-F exhibits good performance in terms of actual coverage and lengths of the CI's. For example, in the MRS and LRS settings (Figures 2.3(b) and (c)), REMM-F coverages and lengths are nearly equal to that of the bootstrap REMM-N and Smith's in the normal-normal settings, despite fitting



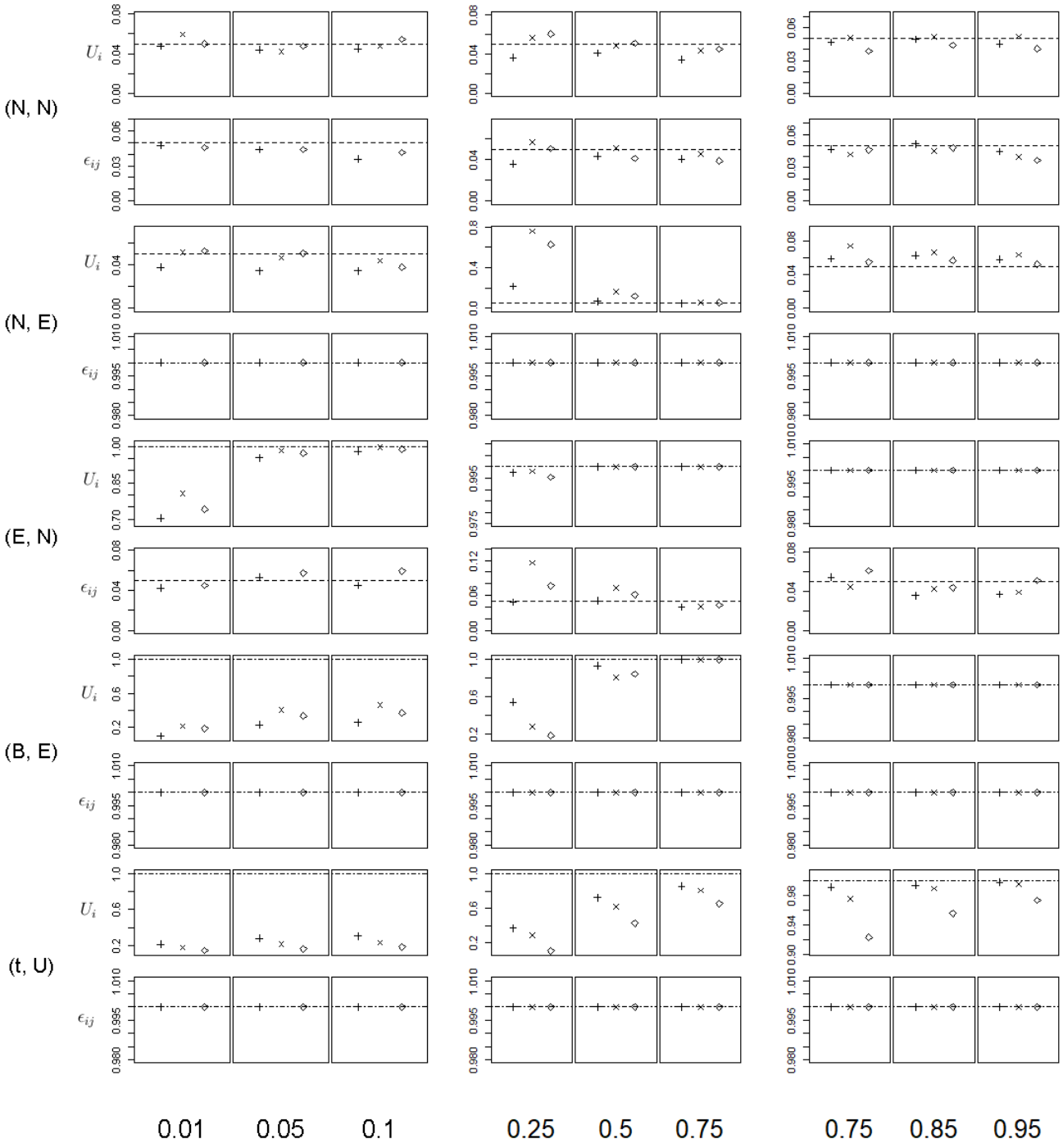
(a) $I = 50, J_i \sim \text{Unif}\{350, 450\}$ (b) $I = 500, J_i \sim \text{Unif}\{3, 4\}$ (c) $I = 1000, J_i \sim \text{Unif}\{3, 4\}$

Figure 2.3: + REMM-F × Bootstrap REMM-N ◇ Smith ∇GEE2 Empirical coverage levels and lengths for each of the ICC confidence interval methods under several scenarios, averaged over 1500 replicate simulations.

additional parameters, yet REMM-F achieves nominal coverage in the non-normal settings whence bootstrap REMM-N and Smith's fall short. GEE2, while achieving nominal coverage, consistently produces CI's longer than that of REMM-F (up to 70% longer in some settings). In the CRT settings (Figure 2.3(a)) where the number of clusters is not too large, REMM-F displays more stable behavior than the other three methods in non-normal settings, and still has competitive CI coverage and lengths in the normal-normal settings against bootstrap REMM-N and Smith's; the coverage of the GEE2 approach is lower than the nominal level in all settings.

Figure 2.4 displays the test of normality p-values for the random effects and errors with our proposed KDE-boot method against those of SW and AD on the conditional modes and fitted residuals. Note that the SW test can handle up to sample sizes of 5000 [Rahman and Govindarajulu, 1997], as implemented in the `shapiro.test` function in R, hence the SW test was omitted for ϵ_{ij} within the CRT scenarios, where we had more than 5000 residuals. Throughout all scenarios, the proposed KDE-boot test controls for the nominal type I error rates better than SW or AD. The MRS scenario in Figure 2.4 (b) best exhibits the "spill-over" phenomenon mentioned in Section 2.2.3. In the normal-exponential case, the Type I error rates of the SW and AD tests on U_i are severally inflated (nearly 80% for ICC $= 0.25$), only settling down to nominal levels when ICC $= 0.75$. The KDE-boot test is much closer to nominal rejection rate in all MRS cases. A similar phenomenon is observed for the Type I error rates on ϵ_{ij} in the exponential-normal case. For the LRS scenario in Figure 2.4 (c), we continue to observe inflated Type I error rates for the SW and AD tests for U_i when the distribution of ϵ_{ij} is non-normal and similarly for ϵ_{ij} when the distribution of U_i is non-normal. KDE-boot demonstrates more power than both SW and AD in the MRS (Beta$(5, 2)$, Exponential) scenarios and MRS/LRS (t_7, Uniform) scenario.

Figure 2.5 displays the actual coverage and length of CI's for the skewness and kurtosis of U_i and ϵ_{ij} using the proposed REMM-F method. The CI coverage for the skewness and kurtosis of ϵ_i is near nominal level in all scenarios. Among the non-normal scenarios chosen for ϵ_{ij}, it appears the Fleishman distribution provides a good approximation to the exponential distribution, and an over-estimation of higher moments for the uniform distribution.



(a) $I = 50, J_i \sim \text{Unif}\{350, 450\}$ (b) $I = 500, J_i \sim \text{Unif}\{3, 4\}$ (c) $I = 1000, J_i \sim \text{Unif}\{3, 4\}$

Figure 2.4: + KDE-boot × Shapiro-Wilk ◇ Anderson-Darling Type I error rates (when distribution is normal) or power (when distribution is non-normal) for each normality test under several scenarios, averaged over 1500 replicate simulations. Dashed lines indicate nominal Type I error rate $\alpha = 0.05$, while dashed-dot lines indicate maximum power $\beta = 1$.

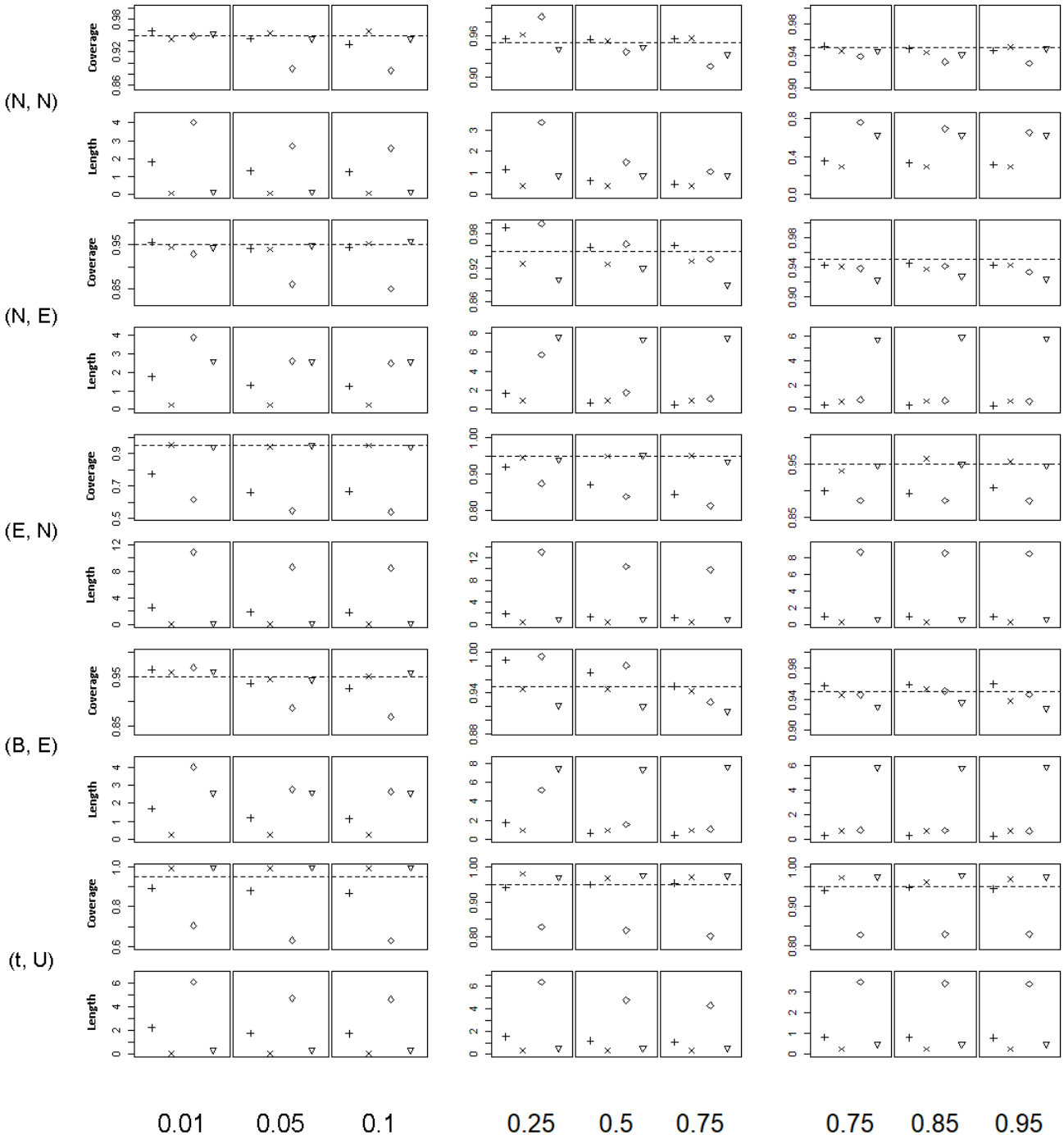
Indeed, denoting $\tilde{\mu}_n = \mathbb{E}[(X-\mu)^n]/\sigma^n$ as the nth standardized central moments, the 5th to 8th $\tilde{\mu}_n$ of an exponential distribution are (44, 265, 1854, 14833), respectively, while the 5th to 8th $\tilde{\mu}_n$ of a Fleishman distribution with first four moments equal to an exponential are approximately (44.5, 272, 1957, 16231), respectively. By previous heuristics, we require our 6th and 8th moments to be approximately matched in order to establish reliable CI's for the skewness and kurtosis, which is satisfied in this case. For the uniform distribution, the 5th to 8th $\tilde{\mu}_n$ are $(0, 3.86, 0, 9)$, while the 5th to 8th $\tilde{\mu}_n$ of a Fleishman distribution with first four moments equal to a uniform are $(0, 8, 0, 774)$. Here, estimates for the 6th and 8th moments based on the Fleishman distribution substantially overestimates those from a uniform, leading to conservative confidence intervals for the skewness and kurtosis.

Inference for the skewness (γ_ϵ) and kurtosis (κ_ϵ) of the error distribution is effectively based on $\sum_i J_i$ residuals, while inference for the skewness (γ_u) and kurtosis (κ_u) of the random effects distribution is effectively based on I cluster-level summaries. Therefore, inference based on $\hat{\gamma}_u$ and $\hat{\kappa}_u$ depends more heavily on their small-sample performance. Simulation studies by Lehmann et al. [2013] showed that at least 1000 data points are needed to reliably diminish bias in the sample skewness for mildly skewed distributions. We observe here REMM-F yields adequate coverage for the skewness of the random effects distribution and mixed results for the kurtosis.

Overall, these simulations suggest that REMM-F CI's for the skewness and kurtosis of ϵ_{ij} are quite good, but we should remain vigilant when constructing CI's for the skewness and kurtosis of U_i, only seriously considering them when γ_u, κ_u are close to 0 or I is large, say $I \geq 1000$.

2.4 Proposal II: QN-ICC for agreement studies

The ICC is commonly used to quantify agreements among different measurements. As mentioned in the Introduction, it operates on a linear scale and is sensitive to outlines and underlying distributions. In this section, we propose a modified ICC, the quantile-normalized ICC (QN-ICC) with respect to a reference distribution, to overcome these limitations. Let



(a) $I = 50, J_i \sim \text{Unif}\{350, 450\}$ (b) $I = 500, J_i \sim \text{Unif}\{3, 4\}$ (c) $I = 1000, J_i \sim \text{Unif}\{3, 4\}$

Figure 2.5: $+U_i$ Skewness $\times\, \epsilon_{ij}$ Skewness $\diamond\, U_i$ Kurtosis $\nabla \epsilon_{ij}$ Kurtosis Empirical coverage levels and lengths for U_i, ϵ_{ij} skewness/kurtosis combinations under several scenarios, averaged over 1500 replicate simulations.

Y_{ij} be a measurement on the ith subject with metric j. If no measurement is recorded for the (i, j)th element, either due to missing data or structural reasons of the design, encode with NA. We assume the missingness process is independent of the measurements and other baseline covariates, i.e. the missing completely at random (MCAR) assumption [Rubin, 1976]. Let J be the number of metrics, J_i the number of valid measurements for subject i, I as the total number of subjects, and I_j as the number of subjects with a valid measurement for metric j. The procedure for QN-ICC is as follows:

1. Compute the *standardized ranks* R_{ij}, where $R_{ij} = \text{NA}$ if $Y_{ij} = \text{NA}$, otherwise $R_{ij} = \frac{1}{I_j+1} \sum_{\ell=1}^{I_j} \mathbb{I}(Y_{ij} \geq Y_{i\ell}, Y_{ij} \neq \text{NA}, Y_{i\ell} \neq \text{NA})$, where $\mathbb{I}$ is the indicator function. In other words, if we lay Y_{ij} in an $I \times J$ table, we rank the observations in each column, leaving the NA's alone, and then divide each column by $I_j + 1$ to ensure $R_{ij} \in (0, 1)$.

2. Compute the *quantile-normalized observations* $\tilde{Y}_{ij}$, where $\tilde{Y}_{ij} = \text{NA}$ if $R_{ij} = \text{NA}$, otherwise $\tilde{Y}_{ij} = G^{-1}(R_{ij})$, where G is the distribution function of the *reference distribution*.

3. Ignore the NA's, since we are assuming MCAR, and compute point estimates and CI's of the ICC for the quantile-normalized observations with REMM-F.

QN-ICC possesses several desirable properties, including transformation invariance, accounting for missing data, and choice of reference distribution to target agreement in the most relevant range. Steps 1 and 2 comprise the quantile normalization step, which is one of the most frequently used techniques of data preprocessing in microarray analysis [Bolstad et al., 2003]. These steps standardize each of the metrics to a common reference distribution G, at which point it makes sense to compute agreement in the form of, say, ICC. The canonical reference is the uniform distribution $G(r) = r$, resulting in equal weights for all observations. Other scenarios might warrant a different reference distribution. For example, it may be of interest to quantify the agreement among various air quality metrics such as $PM_{2.5}$ concentration, PM_{10} concentration, SO_x concentration and NO_x concentration. These metrics are often modeled by a lognormal distribution; the lognormal nature

would naturally lead to high values of ICC due to its heavy-tailness and right-skewness. In assessing agreement among these metrics, it may be desired to weight higher concentrations more, since (1) air quality monitors are less accurate at lower concentrations and (2) higher concentrations are more likely to elicit immune response and hence they are the quantities of interest. At the same time, weighing according to a lognormal distribution may be too extreme, and one may prefer a reference distribution with mitigated right-skew and leptokurtic properties. Therefore, a reference distribution such as Gamma$(4, 6)$ may be advantageous.

2.5 Application to CHAT Signal Data

The Childhood Adenotonsillectomy Trial (CHAT) [Marcus et al., 2013] is a multi-center, single-blind, randomized, controlled trial designed to test whether after a 7-month observation period, children, ages 5 to 9.9 years, with mild to moderate obstructive sleep apnea randomized to early adenotonsillectomy (eAT) will show greater levels of neurocognitive functioning, specifically in the attention-executive functioning domain, than children randomized to watchful waiting plus supportive care (WWSC). Physiological measures of sleep were assessed at baseline and at 7-months with standardized full polysomnography with central scoring at the Brigham and Women's Sleep Reading Center. In total, 1,447 children had screening polysomnographs and 464 were randomized to treatment. Data from EEG spectral analysis were available on a subset of subjects ($I = 409$) at baseline. The data include log power spectral density Y_{ij} for the ith subject at the jth channel (C3, C4, F3, F4, O1, O2) for each wave $(\delta, \theta, \alpha, \sigma, \beta)$.

We would like to characterize the agreement within each wave across various channels. Figure 2.6 displays the two-way scatter plots for the 15 unique pairs of the six channels within the δ wave. We see that the log power spectral density remains right-skewed, hence the agreement will be large due to a few influential points. On the ranked scale, agreement is more ambiguous (i.e. f3 vs o2). We fit the models $\tilde{Y}_{ij} = \mu + U_i + \epsilon_{ij}$, where $\tilde{Y}_{ij}$ are the Studentized (i.e. linearly transformed log power densities within each channel to have

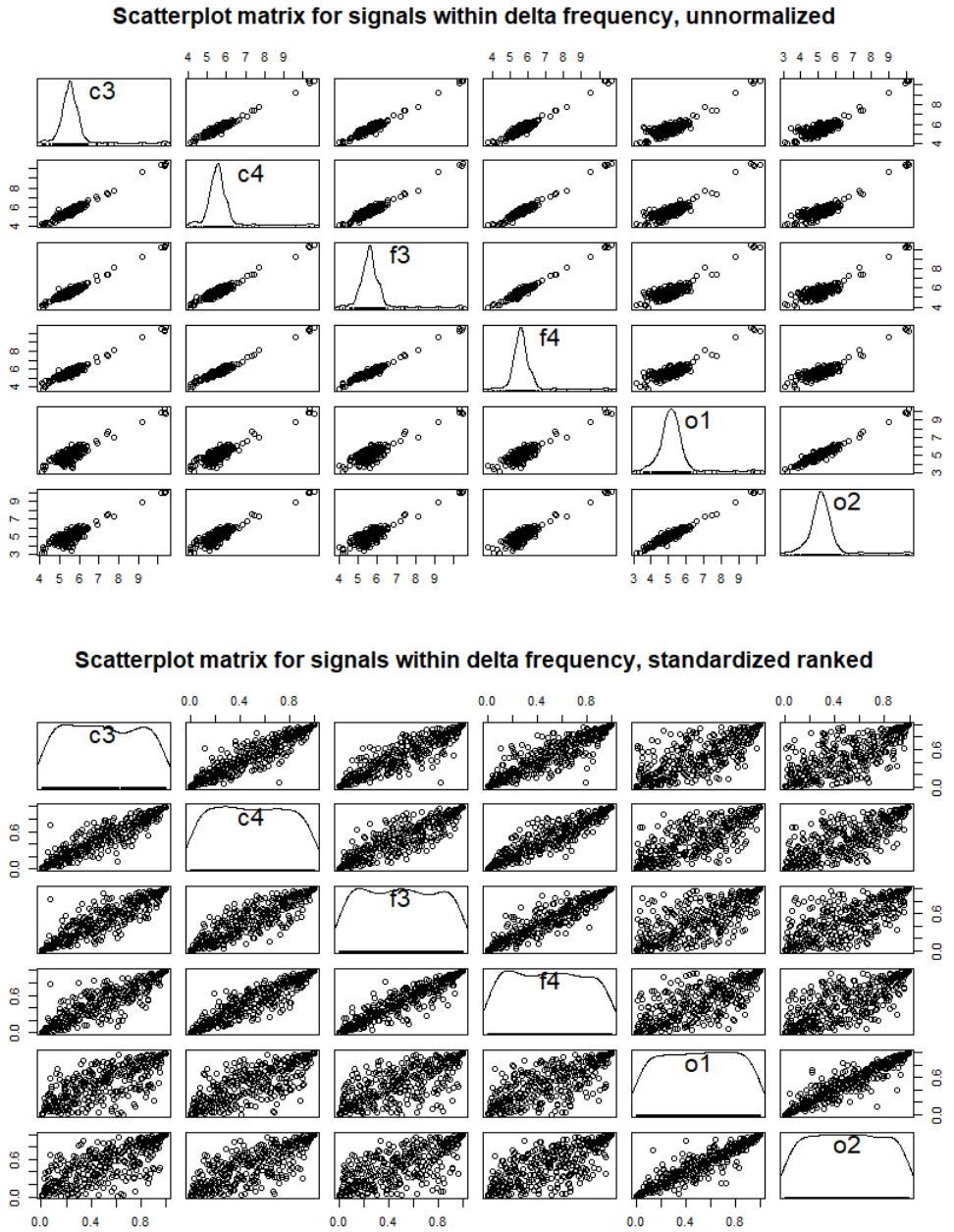


Figure 2.6: Scatterplot matrix displaying pairwise scatterplots of log spectral wave density among the six available channels on the off-diagonals and kernel density plots of log spectral density for each channel on the diagonal. Note the unnormalized density plots exhibit right-skewness for each channel, even so after a log transform. The normalized density plots are uniformly distributed and should resemble a rectangle, but the nature of Gaussian kernels always give density to points outside a finite support.

mean 0 & standard deviation 1), Uniform QN, and Normal QN versions of Y_{ij}. Tables 2.1, 2.2, and 2.3 display the p-values of the normality tests and estimates (CI's) for the ICC / skewness / kurtosis from these three models, respectively. ICC was estimated using REMM-F, bootstrap REMM-N, Smith's method, and GEE2. Normality tests were conducted with KDE-boot, SW, and AD. Skewness and kurtosis was estimated based on REMM-F. We use $B = 10000$ bootstrap samples for REMM-F and bootstrap REMM-N.

	δ	θ	α	σ	β
ICC					
REMM-F	0.921 [0.875, 0.949]	0.923 [0.884, 0.949]	0.913 [0.869, 0.943]	0.907 [0.857, 0.939]	0.937 [0.898, 0.961]
REMM-N (Bootstrap)	0.921 [0.909, 0.931]	0.923 [0.911, 0.933]	0.913 [0.899, 0.924]	0.907 [0.892, 0.919]	0.937 [0.927, 0.945]
REMM-N (Smith)	0.921 [0.91, 0.932]	0.923 [0.912, 0.934]	0.913 [0.901, 0.925]	0.907 [0.894, 0.92]	0.937 [0.928, 0.946]
GEE2	0.936 [0.89, 0.982]	0.934 [0.889, 0.979]	0.924 [0.876, 0.972]	0.918 [0.866, 0.971]	0.947 [0.901, 0.993]
$\frac{\text{CI}^{\text{GEE2}}_{\text{length}} - \text{CI}^{\text{REMM-F}}_{\text{length}}}{\text{CI}^{\text{REMM-F}}_{\text{length}}}$	0.252	0.383	0.306	0.279	0.473
Normality					
U_i KDE-boot	0	0	0	0	0
U_i SW	0	0	0	0	0
U_i AD	0	0	0	0	0
ϵ_{ij} KDE-boot	0	0	0	0	0
ϵ_{ij} SW	0	0	0	0	0
ϵ_{ij} AD	0	0	0	0	0
U_i Skewness	3.886 [2.457, 5.863]	3.318 [1.988, 5.316]	3.468 [2.184, 5.281]	3.811 [2.446, 5.708]	4.269 [2.617, 6.058]
U_i Kurtosis	24.22 [6.555, 52.898]	19.32 [5.049, 46.963]	19.971 [5.343, 44.461]	22.514 [6.569, 50.406]	26.976 [7.523, 55.655]
ϵ_{ij} Skewness	0.079 [0.072, 0.087]	0.077 [0.071, 0.084]	0.087 [0.08, 0.094]	0.093 [0.086, 0.101]	0.063 [0.057, 0.07]
ϵ_{ij} Kurtosis	-0.618 [-0.999, -0.281]	-0.414 [-0.768, -0.094]	-0.071 [-0.297, 0.153]	-0.069 [-0.321, 0.174]	0.163 [-0.288, 0.617]

Table 2.1: Analysis results of CHAT EEG log spectral density with Studentized values within columns. Top panel presents point and 95% CI's for the ICC; middle panel presents p-values from normality tests on U_i and ϵ_{ij}; bottom panel presents point and 95% CI's for the skewness and kurtosis of U_i and ϵ_{ij}. Each column represents a specific EEG wave.

	δ	θ	α	σ	β
ICC					
REMM-F	0.795 [0.772, 0.815]	0.826 [0.807, 0.844]	0.793 [0.77, 0.813]	0.756 [0.73, 0.78]	0.778 [0.753, 0.8]
REMM-N (Bootstrap)	0.795 [0.767, 0.819]	0.826 [0.802, 0.847]	0.793 [0.765, 0.817]	0.756 [0.724, 0.784]	0.778 [0.748, 0.803]
REMM-N (Smith)	0.795 [0.769, 0.821]	0.826 [0.803, 0.849]	0.793 [0.767, 0.819]	0.756 [0.727, 0.786]	0.778 [0.750, 0.805]
GEE2	0.799 [0.767, 0.831]	0.83 [0.802, 0.858]	0.793 [0.76, 0.825]	0.755 [0.718, 0.791]	0.78 [0.746, 0.814]
$\frac{\text{CI}^{\text{GEE2}}_{\text{length}} - \text{CI}^{\text{REMM-F}}_{\text{length}}}{\text{CI}^{\text{REMM-F}}_{\text{length}}}$	0.464	0.521	0.506	0.488	0.415
Normality					
U_i KDE-boot	0	0	0	0	0
U_i SW	7.0×10^{-7}	0	4.3×10^{-7}	1.7×10^{-6}	1.8×10^{-6}
U_i AD	9.6×10^{-6}	0	4.0×10^{-6}	2.3×10^{-5}	1.9×10^{-5}
ϵ_{ij} KDE-boot	0	0	0	0	0
ϵ_{ij} SW	0	0	0	0	0
ϵ_{ij} AD	0	0	0	0	0
U_i Skewness	0.041 [-0.111, 0.196]	0.044 [-0.114, 0.198]	0.053 [-0.1, 0.206]	0.043 [-0.111, 0.201]	-0.009 [-0.161, 0.148]
U_i Kurtosis	-1.095 [-1.243,-0.932]	-1.158 [-1.319, -0.993]	-1.104 [-1.253, -0.939]	-1.08 [-1.237, -0.905]	-1.077 [-1.23, -0.905]
ϵ_{ij} Skewness	0.017 [0.016, 0.019]	0.014 [0.013, 0.016]	0.017 [0.016, 0.019]	0.02 [0.019, 0.022]	0.018 [0.017, 0.02]
ϵ_{ij} Kurtosis	-0.294 [-0.583, -0.023]	-0.069 [-0.345, 0.203]	-0.021 [-0.264, 0.222]	-0.086 [-0.386, 0.195]	0.047 [-0.265, 0.363]

Table 2.2: Analysis results of CHAT EEG log spectral density with uniform QN. Top panel presents point and 95% CI's for the ICC; middle panel presents p-values from normality tests on U_i and ϵ_{ij}; bottom panel presents point and 95% CI's for the skewness and kurtosis of U_i and ϵ_{ij}. Each column represents a specific EEG wave.

From Table 2.1, we see that, even after the log transform, the skewness and kurtosis of U_i are quite high for each wave ($\hat{\gamma}_u \in [3.3, 4.3]$ and $\hat{\kappa}_u \in [19, 27]$). Therefore, the ICC estimates are likely to be greatly affected by a few, large influential points. Based on the naive estimates, the wave-specific log power density ICC ≥ 0.90 for each method. In any case, GEE2 provides CI lengths which are 25% to 47% longer REMM-F CI lengths for each wave. We see that the wave-specific skewness and kurtosis estimates of U_i and ϵ_{ij}

	δ	θ	α	σ	β
ICC					
REMM-F	0.823 [0.796, 0.846]	0.846 [0.822, 0.866]	0.821 [0.794, 0.844]	0.791 [0.76, 0.817]	0.816 [0.788, 0.841]
REMM-N (Bootstrap)	0.823 [0.798, 0.845]	0.846 [0.823, 0.864]	0.821 [0.796, 0.843]	0.791 [0.763, 0.815]	0.816 [0.792, 0.838]
REMM-N (Smith)	0.823 [0.8, 0.846]	0.846 [0.825, 0.866]	0.821 [0.798, 0.844]	0.791 [0.764, 0.817]	0.816 [0.793, 0.84]
GEE2	0.829 [0.797, 0.861]	0.851 [0.821, 0.88]	0.822 [0.789, 0.854]	0.791 [0.754, 0.829]	0.819 [0.785, 0.854]
$\frac{\text{CI}^{\text{GEE2}}_{\text{length}} - \text{CI}^{\text{REMM-F}}_{\text{length}}}{\text{CI}^{\text{REMM-F}}_{\text{length}}}$	0.291	0.350	0.301	0.319	0.301
Normality					
U_i KDE-boot	0.509	0.662	0.510	0.357	0.307
U_i SW	0.568	0.676	0.757	0.713	0.530
U_i AD	0.766	0.426	0.619	0.660	0.376
ϵ_{ij} KDE-boot	0	0	0	0	0
ϵ_{ij} SW	0	0	0	0	0
ϵ_{ij} AD	0	0	0	0	0
U_i Skewness	0.096 [-0.178, 0.372]	0.097 [-0.158, 0.36]	0.094 [-0.184, 0.373]	0.097 [-0.188, 0.402]	-0.003 [-0.315, 0.296]
U_i Kurtosis	0.135 [-0.386, 0.841]	0.051 [-0.415, 0.686]	0.136 [-0.379, 0.84]	0.245 [-0.342, 1.062]	0.298 [-0.296, 1.158]
ϵ_{ij} Skewness	0.173 [0.158, 0.188]	0.15 [0.138, 0.164]	0.175 [0.162, 0.188]	0.204 [0.188, 0.221]	0.179 [0.164, 0.196]
ϵ_{ij} Kurtosis	-0.382 [-0.679, -0.114]	-0.295 [-0.612, 0.013]	-0.116 [-0.342, 0.105]	-0.151 [-0.42, 0.103]	-0.026 [-0.36, 0.305]

Table 2.3: Analysis results of CHAT EEG log spectral density with normal QN. Top panel presents point and 95% CI's for the ICC; middle panel presents p-values from normality tests on U_i and ϵ_{ij}; bottom panel presents point and 95% CI's for the skewness and kurtosis of U_i and ϵ_{ij}. Each column represents a specific EEG wave.

are quite similar among each other, so quantile normalization is not required if one were concerned with matching the distributions as closely to each other as possible. But if one were concerned with the inequitable weighting from this Studentized distribution, since it gives far more weights to agreement of larger log power densities than smaller, then QN-ICC may be preferred. If we favor equal weighting, than we shall refer to the Uniform QN-ICC results in Table 2.2. Here, we see that the agreement is less ($\hat{\rho} \in [0.75, 0.83]$ per wave class). Also, REMM-F CI's for the ICC are shorter than intervals produced from bootstrap REMM-N or Smith's method; this is due to REMM-F leveraging the fact that $\hat{\kappa}_u < 0$. The relative lengths of GEE2 CI's of ICC over that of REMM-F is even more pronounced with this normalization, ranging around 40% - 50% longer.

Finally, we use a normal reference distribution in QN-ICC in Table 2.3, which gives more weight to agreement among larger and smaller values than a uniform reference. We see now see these calculated ICC's for the various waves are between that of the Studentized and uniform QN-ICC, indicating that there isn't as much agreement among lower values of log power density, but not enough to overpower the agreement among higher values. As before, GEE2 provides significantly longer CI's than REMM-F for the ICC, ranging from 25% to 30% longer.

2.6 Discussion

In this paper, we propose REMM-F for estimating the distributions of the random effects U_i and errors ϵ_{ij} in a one-way LMM. Our methods are especially useful for inference on second-order quantities, such as the ICC, allowing flexible distributional assumptions on U_i and ϵ_{ij}. If interest is solely on the main effect terms, then the estimators obtained through fitting with a normal LMM are robust to misspecification of the distributions of U_i and ϵ_{ij}. Methods in characterizing finer distributional characteristics, such as skewness and kurtosis, are lacking, and REMM-F provides a starting point in producing reasonable CI's for these quantities in certain situations.

Unlike the analytical methods of Smith and GEE2 in constructing ICC CI's, our pro-

posed REMM-F method is generative in nature. An early motivation of our work is the construction of ICC CI's under a missing data framework using multiple imputation, for which a generative method is required. We aspire to demonstrate in future works the validity and effectiveness of our proposed methods under the various missingness mechanisms.

We propose a test of normality, called the KDE-boot, for the distribution of the random effects and that of the subject-specific errors. The idea behind this test is to construct a kernel-density estimator for the joint distribution of a multivariate test statistic, i.e. the skewness and kurtosis. This test can help guide the choice between the use of normal LMM and REMM-F. We also propose QN-ICC, a modified ICC estimator to measure agreement among different metrics. QN-ICC enjoys the advantages of transformation invariance and flexibility in reference distribution to target specific regions that of particular interests.

Our proposed framework is customizable. For example, the Fleishman distribution can be replaced with any other flexible distribution, preferably one which is easy to simulate from. We could replace the moment equations in Eq 2.4 with ones that are less biased or more efficient when such equations become available. Although we focused on ICC on the linear scale, estimation and inference can be extended to some cases within generalized linear mixed models (GLMM). For example, in time-to-event data, we model $Y_{ij} = \exp(g(X_{ij}) + \mu + U_i + \epsilon_{ij})$ and perform our procedure on the observations $\log Y_{ij}$, which is then backtracked to the manifested scale as follows:

$$\rho = \text{Corr}(Y_{ij}, Y_{ij'}|X_{ij}, X_{ij'}) = \frac{\mathbb{E}e^{2U_i}(\mathbb{E}e^{\epsilon_{ij}})^2 - (\mathbb{E}e^{U_i})^2(\mathbb{E}e^{\epsilon_{ij}})^2}{\mathbb{E}e^{2U_i}\mathbb{E}e^{2\epsilon_{ij}} - (\mathbb{E}e^{U_i})^2(\mathbb{E}e^{\epsilon_{ij}})^2}$$

Adapting this to the binary case is more challenging, and more investigation is needed and would be useful, as inference even on the main effects can be severely affected by the misspecification in the distributions of the random effect and error distributions in such cases [Litière et al., 2007].

Chapter 3

Robust Estimation of Recurrent Event Mean Functions in the Presence of Informative Event Censoring

Motivated by a novel "evolving cluster randomized trial" for HIV prevention, where transmission clusters centered on newly HIV diagnosed individuals are established over time through phylogenetic analyses, we develop an estimating procedure for the intervention effect on patterns of HIV transmission in terms of the cluster sizes over time of the "evolving rings". We view each contact linked to the index case as a recurrent event to the index case and estimate treatment effects based on marginal rate and mean functions. A difficulty that arises is informative censoring of these contacts, which equates to missing events within the recurrent event process. We account for this dependent censoring through the use of inverse probability censoring weights.

3.1 Introduction

HIV transmission network analysis has been used to describe viral transmission dynamics, emerging epidemics, cross-national transmission and cluster growth dynamics. Identifying transmission clusters with higher growth rates is crucial from a public health perspective,

where public health interventions could improve care outcomes and prevent new infections. Several molecular HIV surveillance techniques have been introduced in order to identify HIV transmission networks, including phylogenetic [Dennis et al., 2012; Grabowski and Redd, 2014], genetic distance [Campbell et al., 2017; Wertheim et al., 2017], and combination methods. The RIING study is one such study which employs phylogenetic linkage and a novel "evolving cluster randomized trial" design to evaluate interventions aimed at reducing HIV incidence in the study population. This design administers immediate antiretroviral therapy (I-ART) to all infected index cases (i.e. participants identified and not phylogenetically linked to any previous cases). Uninfected participants who are sexually or socially in contact, either directly or through intermediary contacts, to existing indexes are referred to as contacts of index (COI). They are discovered through, for example, disease intervention specialists, health care providers, health department staff, community members, or from the indexes themselves. COIs receive intervention based off their shared indexes; that is, the social or sexual contact clusters induced from each index are the unit of randomization. In the Intervention arm, these uninfected contacts will receive immediate pre-exposure prophylaxis (I-PrEP). In the Standard of Care (SOC) arm, the uninfected contacts of the index will receive SOC linkage to available testing and prevention services. This study aims to characterize transmission networks of the trial participants and assess the effect of the treatment on patterns of HIV transmission to inform the design of the future efficacy studies.

To facilitate the characterization of the transmission network patterns, we consider a linked contact to the index as a recurrent event and the transmission cluster as comprised of these linked individuals to the index, excluding the index himself. Under this framework, the size of transmission cluster corresponds to the number of recurrent events and the growth rate of the cluster corresponds to the rate function of the recurrent events. More specifically, consider a single homogeneous group of subjects, and let $N_i(t)$ denote the number of COIs linked to index i over the time period $(0, t]$. Then the event mean and rate functions are, respectively, $\mu(t) = \mathbb{E}[N_i(t)]$ and $\rho(t) = \mu'(t)$, where we assume that events occur in continuous time. To compare two treatment groups, say SOC group 0

and Intervention group 1, we can conveniently consider ratios $\mu_1(t)/\mu_0(t)$ or $\rho_1(t)/\rho_0(t)$, or differences $\mu_1(t) - \mu_0(t)$ or $\rho_1(t) - \rho_0(t)$. We will focus on treatment difference $TD(t) \stackrel{\text{def}}{=} \mu_1(t) - \mu_0(t)$ throughout this paper, for it directly characterizes the average number of HIV infections prevented per index in using Intervention over SOC.

There have been considerable advances in the past few decades on statistical methods for the analysis of recurrent events. Perhaps the most popular approach for analysis of survival data is the Cox proportional hazards model [Cox, 1992]. Due to the independence assumption, the original Cox model is only appropriate for modeling the time to the first event, which is an inefficient use of data because data from the later events are discarded. Extensions of the original Cox model have been proposed for analyses of recurrent event data such as Andersen-Gill (AG) [Andersen and Gill, 1982], Prentice, Williams and Peterson (PWP) [Prentice et al., 1981], Wei, Lin and Weissfeld (WLW) [Wei et al., 1989] and frailty models [Therneau, 1997]. The analysis strategy taken in this work is through modeling the mean number of events [Cook and Lawless, 2007; Lin et al., 2000; Pepe and Cai, 1993].

The issue of missing data complicates effect estimates. The diagram below displays four data patterns we may observe for each COI, henceforth referred to as Scenarios 1 – 4, respectively.

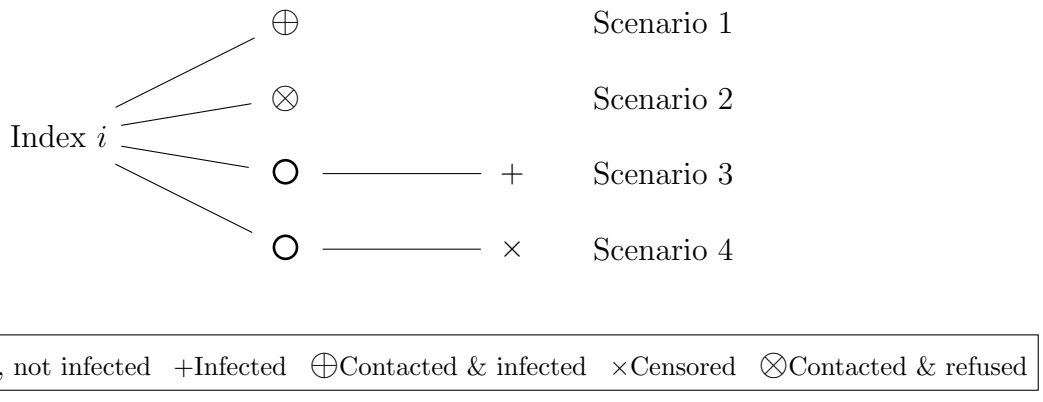


Figure 3.1: Four types of data scenarios in the RIING study

Scenario 1 represents a setting where a COI had already been infected when identified. Here, the timing of the infection is ambiguous: the COI may have been infected before the index and the index was simply observed first in the study, or the COI was infected afterwards. In the former, time to infection is not well-defined. To avoid this, our characterization will work with time to phylogenetic linkage rather than time to infection. While this metric is not a perfect proxy for infection time, it captures the general trends for infection times between two treatment groups. In Scenario 2, a COI was unable to get in touch or refused to participate. In Scenarios 3 and 4, COIs participated in the study and were followed over time for their infection status, and they either become infected (Scenario 3) or remain uninfected until the end of study or loss to follow-up (Scenario 4). Contacts of contacts are treated similarly as in Figure 3.1. Transmission clusters grow as new diagnosed individuals become linked to the index cases through phylogenetic analysis. The goal of investigation is to characterize the growth of the transmission clusters and to estimate treatment effect on features of transmission clusters, accounting for refusal into the study and informative censoring. Numerous works [Cole and Hernán, 2004; Rotnitzky and Robins, 2005; Sugihara, 2010] have discussed inverse probability weighting techniques to accounting for informative censoring in the survival analysis context. Cook et al. [2009] addressed a similar problem in accounting for informative censoring of an entire counting process $N_i(t)$. But, this is akin to accounting for all contacts in a cluster refusing or dropping out, while we need to address each event (individual) who refuses entry or drops out.

In Section 3.2, we begin with a background on recurrent event analysis and the model set-up. From there, we state the naive estimator for mean number of recurrent events and our proposed estimator which adjusts for informative censoring of events. We prove the consistency of our estimator here and advocate the use of bootstrap for inference. We evaluate the performance of our methods with simulations in Section 3.3 and end with a discussion in Section 3.4.

3.2 Methods

3.2.1 Data Structure & Model Set-Up

For each index case i ($i = 1, \cdots, I$), denote COI_{in} ($n = 1, \cdots, N_i$) as the nth contact (direct or indirect) of index. Let T_{in} denote time to phylogenetic linkage, $A_i \in \{0, 1\}$ denote intervention, and $\mathbf{X}_{in}$ denote other baseline covariates, which may include covariates of the index or covariates of direct contacts to COI_{in} if relevant. Let C_{in} be the time to contact (i.e. time to be socially/sexually linked to index) in the study, $D_{in} \in \{0, 1\}$ indicate if a contact participates in the study, R_{in} as time to loss-to-follow up, and τ_i as time to end of observation. For each contact, we observe $(U_{in}, C_{in}, \tau_i, D_{in}, \delta_{in}^T, \delta_{in}^R, A_i, \mathbf{X}_{in})$, where $U_{in} = (T_{in} \wedge R_{in} \wedge \tau_i)D_{in} + C_{in}(1 - D_{in})$, $\delta_{in}^T = \mathbb{I}(U_{in} = T_{in})D_{in}$, and $\delta_{in}^R = \mathbb{I}(U_{in} = R_{in})$. Because phylogenetic linkages can only be made after contact, we assume $T_{in} \geq C_{in}$ throughout. If context is clear, we will omit the index i to refer to a general index.

For a sequence of events $T_1, T_2, \cdots$, let $N(t) = \sum_{n=1}^{\infty} \mathbb{I}(T_n \leq t)$. Then, $N(t)$ is a right continuous process for the number of events over time interval $(0, t]$ for a general individual, with $dN(t) = 1$ if an event occurs at time t and 0 otherwise. Let $\mathbf{Z}(t)$ denote a covariate process. The history process attached to each individual is denoted by $H(t) = \{N(u), \mathbf{Z}(u) : 0 < u < t\}$. In practice, processes are observed over a finite period of time $[0, \tau_i)$, so we also let $\mathcal{T}_i(t) = \mathbb{I}(t \leq \tau_i)$ indicate whether the process is under observation at time t. The intensity function and rate function of a recurrent event process are defined, respectively, as

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(\Delta N(t) = 1|H(t))}{\Delta t} \qquad \text{and} \qquad \rho(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(\Delta N(t) = 1)}{\Delta t}$$

The rate function is conceptually and quantitatively different from the intensity function; the rate function is defined as the occurrence rate of recurrent events unconditional on the event history and covariates, whereas the intensity function is the occurrence rate conditional on the event history. In general, the rate function gives more direct interpretations for identifying risk factors, and the use of it is preferred over the intensity function in many applications [Cook et al., 2009; Wang et al., 2001]. Specifically, the quantity $\mu(t) = \int_0^t \rho(u)du$

is the expected number of events by time t, which is computed in each treatment group to provide us the treatment difference.

3.2.2 Recurrent Events with No Missing Events

We consider nonparametric estimation of $\mu(t)$, although a similar development applies to parametric models. As do many other authors [Lin et al., 2000; Miloslavsky et al., 2004], we write, informally, $\mathbb{E}[dN_i(t)] = d\mu(t)$ and, correspondingly, $\mathbb{E}[dN_i(t) - d\mu(t)] = 0$. We can then consider estimating equations of the form

$$\sum_{i=1}^{I} \mathcal{T}_i(t)(dN_i(t) - d\mu(t)) = 0 \tag{3.1}$$

Eq 3.1 is the maximum likelihood score equations derived under a Poisson model (Lawless and Nadeau 1995), but are unbiased more generally for the mean function whenever $\mathbb{E}[dN_i(t)|\mathcal{T}_i(t) = 1, \sum_{i=1}^{I} \mathcal{T}_i(t)] = d\mu(t)$. They produce the estimator

$$d\mu(t) = \frac{\sum_{i=1}^{I} \mathcal{T}_i(t) dN_i(t)}{\sum_{i=1}^{I} \mathcal{T}_i(t)}$$

from which we can obtain the mean function $\mu(t) = \int_0^t d\mu(u)$. Analogously, with no missing data, the estimator to characterize the rate function in each intervention arm $a \in \{0, 1\}$ is given by the estimating equation

$$\sum_{i=1}^{I} \mathbb{I}(A_i = a)\mathcal{T}_i(t)(dN_i(t) - d\mu_a(t)) = 0 \tag{3.2}$$

and treatment difference $TD(t) = \mu_1(t) - \mu_0(t)$.

3.2.3 Recurrent Events with Inverse-Probability Weighted Events

With missing data, the crude, complete-case estimating equation is

$$\sum_{i=1}^{I} \mathbb{I}(A_i = a)\mathcal{T}_i(t)(\delta_i^T(t) dN_i(t) - d\mu_a^{\mathrm{CC}}(t)) = 0$$

where $\delta_i^T(t) = \mathbb{I}(t \in \{U_{in} : \delta_{in} = 1\})$. Under dependent censoring, this estimating equation is no longer unbiased. We develop an inverse probability censoring weighted estimators to account for informative missingness and censoring. We make the following assumptions:

(A1) (Conditionally uninformative refusal) $\mathbf{T}_i \perp\!\!\!\perp \mathbf{D}_i | \mathbf{C}_i, A_i, \mathbf{X}_i$

(A2) (Conditionally uninformative drop-out) $\mathbf{T}_i \perp\!\!\!\perp \mathbf{R}_i | \mathbf{C}_i, A_i, \mathbf{X}_i$

(A3) (Conditional independence between refusal and drop-out) $\mathbf{D}_i \perp\!\!\!\perp \mathbf{R}_i | \mathbf{C}_i, A_i, \mathbf{X}_i$

(A4) (Independent end-of-observation) $\mathcal{T}_i(t) \perp\!\!\!\perp (\mathbf{T}_i, \mathbf{D}_i, \mathbf{R}_i) | \mathbf{C}_i, A_i, \mathbf{X}_i$

A1 and A2 are more generally known as the restricted missing at random (rMAR) [Prague et al., 2016] and no unmeasured confounders for censoring [Robins and Finkelstein, 2000] assumptions. They state that whatever process that may guide each subject's time to phylogenetic linkage and propensity to refuse or drop-out can be explained by observable, baseline covariates. Hence, they provide a framework that leverages baseline covariates to correct for bias. Since $N_i(t) = \sum_{n=1}^{\infty} \mathbb{I}(T_{in} \leq t)$, A1 and A2 imply

(A1*) $N_i(t) \perp\!\!\!\perp \mathbf{D}_i | \mathbf{C}_i, A_i, \mathbf{X}_i$

(A2*) $N_i(t) \perp\!\!\!\perp \mathbf{R}_i | \mathbf{C}_i, A_i, \mathbf{X}_i$

The mechanism in which each subject's refusal D_{in} and drop-out R_{in} interacts could be very complicated. A3 provides a convenient decoupling

(A3*) $\mathbb{P}(\delta_i^T(t) = 1 | \mathbf{Z}_i(t)) = \pi_D(\mathbf{Z}_i(t)) S_R(t | \mathbf{Z}_i(t))$

where $\mathbf{Z}_i(t) \stackrel{\text{def}}{=} (\mathbf{C}_i(t), A_i(t), \mathbf{X}_i(t)) \stackrel{\text{def}}{=} (C_{in}, A_i, \mathbf{X}_{in} : t \in \{U_{in}\})$ are any relevant covariates of the COI who experiences censoring or an event at time t. The propensity scores $\pi_D(\mathbf{Z}(t)) = \mathbb{P}(D = 1 | \mathbf{Z}(t))$ is the probability of accepting entry into the study, and censoring distribution $S_R(t | \mathbf{Z}(t)) = \mathbb{P}(R > t | \mathbf{Z}(t))$ is the drop-out survival function. A model for $\pi_D(\mathbf{Z}(t))$ can be fit from the data $(D_{in}, C_{in}, A_i, \mathbf{X}_{in})$, and a model for $S_R(t | \mathbf{Z}(t))$ can be fit from the data $(U_{in}, \delta_{in}^R, C_{in}, A_i, \mathbf{X}_{in})$. A sensible model for the drop-out process should

take the form $S_R(t|\mathbf{Z}(t)) = S_R(t - C_{in}|A_i, \mathbf{X}_{in})$, since drop-out can only accumulate risk if a subject accepts participation into the study, and by convention of setting $U_{in} = C_{in}$ when subject refuses, then $S_R(t - C_{in}|A_i, \mathbf{X}_{in}) = 1$ and thus contributes no weighting from drop-out, as expected. A4 just requires the observation period to be conditionally independent of the time to phylogenetic linkage, refusal process, and drop-out process. In practice, it is often the case that end-of-observation is determined administratively, and completely independent.

The proposed IPW adjusted estimating equation takes the form

$$\sum_{i=1}^{I} \mathbb{S}(t|\mathbf{Z}_i(t)) \stackrel{\text{def}}{=} \sum_{i=1}^{I} \mathbb{I}(A_i = a)\mathcal{T}_i(t) \left(\frac{\delta_i^T(t)}{\pi_D(\mathbf{Z}_i(t))S_R(t|\mathbf{Z}_i(t))} dN_i(t) - d\mu_a^{\text{IPW}}(t) \right) = 0 \quad (3.3)$$

or equivalently,

$$d\mu_a^{\text{IPW}}(t) = \frac{\sum_{i=1}^{I} \mathbb{I}(A_i = a)\mathcal{T}_i(t) \frac{\delta_i^T(t)}{\pi_D(\mathbf{Z}_i(t))S_R(t|\mathbf{Z}_i(t))} dN_i(t)}{\sum_{i=1}^{I} \mathbb{I}(A_i = a)\mathcal{T}_i(t) \frac{\delta_i^T(t)}{\pi_D(\mathbf{Z}_i(t))S_R(t|\mathbf{Z}_i(t))}}$$

If models for $\pi_D(\mathbf{Z}(t))$ and $S_R(t|\mathbf{Z}(t))$ are correctly specified, then the estimator $d\mu_a^{\text{IPW}}(t)$ in Eq 3.3 is consistent and asymptotically normal. Proof of consistency follows by showing the summands in Eq 3.3 are unbiased and by applying standard results regarding M-estimators [Van der Vaart, 2000]. Indeed,

$$\begin{aligned}
\mathbb{E}[\mathbb{S}(t|\mathbf{Z}_i(t))] &= \mathbb{E}\left[\mathbb{I}(A_i = a)\mathcal{T}_i(t) \left(\frac{\delta_i^T(t)}{\pi_D(\mathbf{Z}_i(t))S_R(t|\mathbf{Z}_i(t))} dN_i(t) - d\mu_a^{\text{IPW}}(t) \right) \right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\mathbb{I}(A_i = a)\mathcal{T}_i(t) \left(\frac{\delta_i^T(t)}{\pi_D(\mathbf{Z}_i(t))S_R(t|\mathbf{Z}_i(t))} dN_i(t) - d\mu_a^{\text{IPW}}(t) \right) \middle| \mathbf{Z}_i(t) \right) \right] \\
&= \mathbb{E}\left[\mathbb{I}(A_i = a)\mathbb{E}[\mathcal{T}_i(t)|\mathbf{Z}_i(t)] \left(\frac{\mathbb{E}[\delta_i^T(t)|\mathbf{Z}_i(t)]}{\pi_D(\mathbf{Z}_i(t))S_R(t|\mathbf{Z}_i(t))} \mathbb{E}[dN_i(t)|\mathbf{Z}_i(t)] - d\mu_a^{\text{IPW}}(t)) \right) \right] \quad \text{(A1}^*\text{, A2}^*\text{, A4)} \\
&= \mathbb{E}\left[\mathbb{I}(A_i = a)\mathbb{E}[\mathcal{T}_i(t)|\mathbf{Z}_i(t)] \left(\mathbb{E}[dN_i(t)|\mathbf{Z}_i(t)] - d\mu_a^{\text{IPW}}(t)) \right) \right] \quad \text{(A3}^*\text{)} \\
&= \mathbb{E}\left[\mathbb{E}[\mathbb{I}(A_i = a)\mathcal{T}_i(t) \left(dN_i(t) - d\mu_a^{\text{IPW}}(t) \right) |\mathbf{Z}_i(t)] \right] \quad \text{(A1}^*\text{, A2}^*\text{, A4)} \\
&= \mathbb{E}\left[\mathbb{I}(A_i = a)\mathcal{T}_i(t) \left(dN_i(t) - d\mu_a^{\text{IPW}}(t) \right) \right]
\end{aligned}$$

So, the expectation of the summands $\mathbb{S}(t|\mathbf{Z}_i(t))$ reduce down to the unbiased estimating equation in the no missing data scenario in Eq 3.2, as desired.

It may be possible to derive variance estimates for $\widehat{\mu}_a^{\text{IPW}}(t)$, but the resulting form would be very complicated and it is simpler to employ bootstrap methods. The complication lies

in having to estimate weights $\pi_D(\mathbf{Z}_i(t))$ and $S_R(t|\mathbf{Z}_i(t))$. Had we known the exact weights, then we could use the asymptotics derived in Andersen and Gill [1982] to construct confidence intervals for $\widehat{\mu}_a^{\text{IPW}}(t)$. Since the weights need to be estimated, a direct method in combining the variation in the estimated weights $\widehat{\pi}_D(\mathbf{Z}_i(t))$ and $\widehat{S}_R(t|\mathbf{Z}_i(t))$ and variation in the other expressions in $\widehat{\mu}_a^{\text{IPW}}(t)$ is unclear. The standard method would be to stack estimating equations for $\pi_D(\mathbf{Z}_i(t))$ and $S_R(t|\mathbf{Z}_i(t))$ with that of $d\mu_a^{\text{IPW}}(t)$ in Eq 3.3, but the asymptotics derived in Andersen and Gill [1982] is based off martingale theory instead of estimating equations, and it is unclear how to combine these two approaches. Numerous empirical studies [Austin, 2016; Cook et al., 2009; Miloslavsky et al., 2004] have also advocated for use of bootstrap variance estimators, and we will proceed with its use as well. For the bth ($b = 1, \cdots, B$) bootstrap replicate, we perform the following steps:

1. Sample I index cases and their respective clusters with replacement. Alternatively, one could sample with replacement within stratas $\{i : A_i = 0\}$ and $\{i : A_i = 1\}$ and combine to ensure balance between Intervention and SOC. Call the resulting bootstrap dataset $\mathcal{D}^{(b)}$.

2. Fit models $\widehat{\pi}_D^{(b)}(\cdot)$ and $\widehat{S}_R^{(b)}(t|\cdot)$ from $\mathcal{D}^{(b)}$.

3. Fit $\mu_0^{\text{IPW}(b)}(t)$ and $\mu_1^{\text{IPW}(b)}(t)$ from Eq 3.3 with $\widehat{\pi}_D^{(b)}(\cdot)$ and $\widehat{S}_R^{(b)}(t|\cdot)$ as weights, which then can be used to calculate $TD^{(b)}(t)$.

From here, we can calculate bootstrap standard errors and confidence intervals from the bootstrap estimators.

3.3 Simulations

To evaluate the performance of our framework, we simulate with $I = 100$ indexes and sample the number of COIs, N_i, from a zero-truncated negative binomial with $r = 5$ and $p = 0.3$, resulting in a mean of approximately 12 total COIs per index. Treatment is randomized as $A_i \sim \text{Ber}(\frac{1}{2})$ and two baseline covariates are distributed as $\mathbf{X}_{in} \sim N(\mathbf{0}, 3^{-1/2}\mathbf{I}_{2\times 2})$. End

of observation τ_i are sampled from Unif$(18, 25)$ and contact times C_{in} from a truncated exponential with rate $\lambda = 2$ on the interval $(0, \overline{\tau})$, where $\overline{\tau} = I^{-1} \sum \tau_i$; this is the ensure that all contacts happen before end of observation. Event times T_{in} are simulated from the hazard

$$\lambda_T(t|A_i, \mathbf{X}_{in}, \eta_i) = 0.72t^{0.2}e^{-A_i+X_{1in}-X_{2in}+\eta_i}$$

where frailty term e^{η_i} follows from a positive stable distribution with parameter $\alpha = 0.9$. The term η_i induces correlation among the individuals in the same cluster, which mimics the heterogeneities of event times one might expect to see with genuine HIV transmission clusters such as that studied in the RIING study. A subject's willingness to participate in the study is simulated from

$$D|A_i, \mathbf{X}_{in} \sim \text{Bernoulli}(\text{expit}(0.5 + A_i + 0.7X_{1in} - 0.5X_{2in}))$$

where $\text{expit}(x) = (1+e^{-x})^{-1}$. This results in about 81% probability of participation within the treatment arm, and 62% of participation within the control arm. Finally, the drop-out process is simulated as

$$\lambda_R(r|A_i, \mathbf{X}_{in}, C_{in}) = 0.08(r - C_{in})^{-0.2}e^{0.5A-0.5X_1+0.5X_2}$$

Within those who accept participating into the study, the drop-out process retains a further 71% within the treatment arm, and 89% within the control arm. We define the true growth curves as the mean function of the actual time to phylogenetic linkages. We perform 1000 replicate simulations and construct confidence intervals from 1000 bootstrap resamples.

Figure 3.2 displays the true growth curves among treated (solid orange) and control (solid blue). For demonstration purposes, we plot non-solid lines to display estimates based off a single replicate dataset from the 1000 replicate simulations. We see that the crude estimators, for both treated and control, displays severe underestimation of the true growth curves, with underestimation intensifying as time grows. This is behavior is to be expected, for the crude estimators completely ignore the refused and censored individuals in the study. The IPW estimators much better align with the true growth curves, since

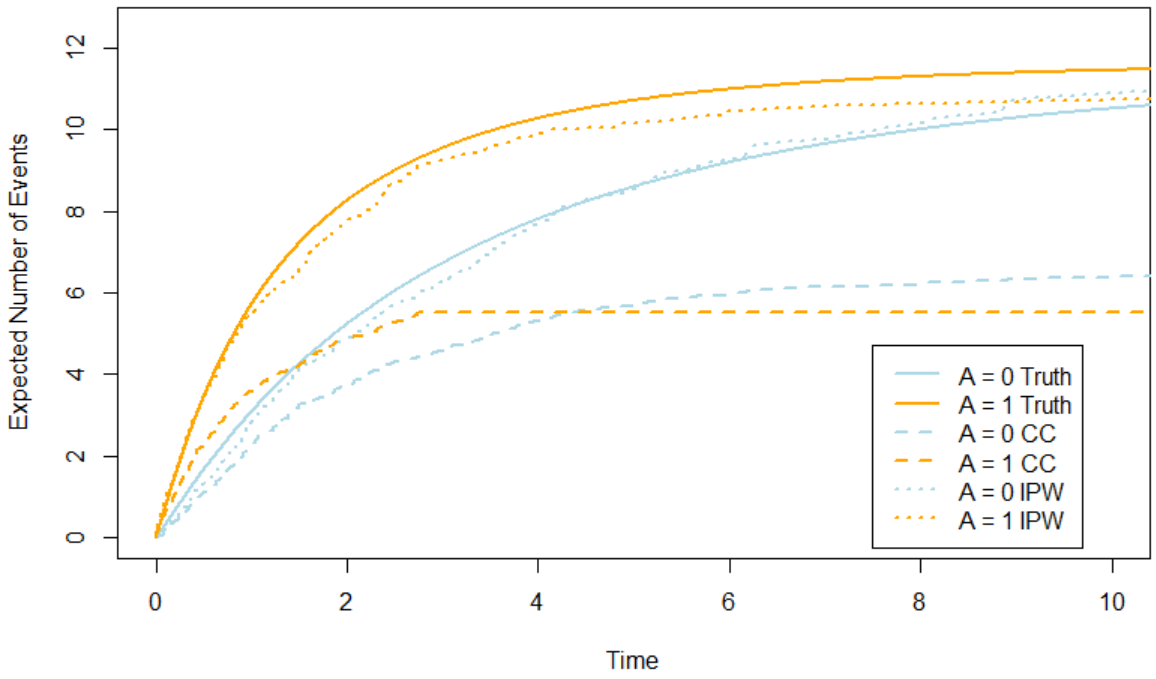


Figure 3.2: Growth curves for treatment and control groups. The complete-case and IPW fittings shown above are a single draw from the 1000 replicate simulations, shown for demonstration purposes.

these estimators weight existing events to match the expected number of total events within each intervention.

Table 3.1 displays summary statistics from the 1000 replicate simulations. The crude estimators are severally biased and undercovered for each intervention group and on the difference level. For some contexts in causal inference (i.e. matched designs), it may be possible for biases to cancel out when considered on the difference level [DiPrete and Engelhardt, 2004], but this is not the case here and still observe substantial bias. The IPW estimators have negligible bias and attain nominal coverage levels for the means in both groups and treatment difference.

Time	1	2	3	4	5	6	7	8	9	10
Bias										
Crude mean # of linkages, $A = 1$	-0.55	-1.15	-1.73	-2.21	-2.62	-2.98	-3.28	-3.53	-3.74	-3.91
Crude mean # of linkages, $A = 0$	-2.01	-3.15	-3.79	-4.21	-4.48	-4.68	-4.81	-4.91	-4.98	-5.03
Crude treatment difference	1.46	1.99	2.00	2.00	1.86	1.70	1.53	1.38	1.24	1.12
IPW mean # of linkages, $A = 1$	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02
IPW mean # of linkages, $A = 0$	0.01	0.01	-0.02	-0.01	-0.01	0.00	0.01	0.01	0.00	0.00
IPW difference	-0.02	-0.02	0.00	0.00	0.00	-0.01	-0.02	-0.02	-0.01	-0.02
Coverage (%)										
Crude mean # of linkages, $A = 1$	0	0	0	0	0	0	0	0	0	0
Crude mean # of linkages, $A = 0$	0	0	0	0	0	0	0	0	0	0
Crude treatment difference	0	0	0.1	0.3	1.9	6.8	16.6	27.1	40.0	50.2
IPW mean # of linkages, $A = 1$	93.5	93.7	93.9	92.8	93.1	93.0	93.5	93.0	93.5	93.6
IPW mean # of linkages, $A = 0$	93.4	94.2	93.3	91.3	92.6	91.2	91.6	91.7	92.2	92.4
IPW difference	93.0	93.4	94.0	93.9	93.4	94.3	94.5	93.9	93.7	93.8

Table 3.1: Bias & coverage for the crude and IPW estimators for growth curves in each intervention arm, and their differences. Number of replicate simulations = 1000.

3.4 Discussion

In this work, we describe the RIING study and the unique problems it poses. From a public health perspective, the ultimate purpose is controlling the spread of an epidemic and therefore the direct metric to characterize effectiveness is number of infected, which we approximate with the number of phylogenetically linked. This motivates our use of a recurrent event framework to model these counts, but we encounter missing events and therefore devise a solution based off inverse probability weighting.

It would be useful to more accurately characterize the contact network by differentiating between direct contacts and contacts of contacts. Staples et al. [2016] provides a model-based description of evolving clusters, and while such models are computationally demanding, they could provide efficiency gains. Comparing between methods which ignore and methods which take into account the evolving structure would be useful.

Time to phylogenetic linkage, by definition, would always occurs after time to infection.

But, the ultimate number of phylogenetically linked can remain a very good approximation for number of infected. Biologically, the number socially linked to index has an upper bound and therefore must plateau over time. If, for example, the discovery process in identifying COIs were very effective, or the study period were longer, then we would expect the number of phylogenetically linked by the end of study would nearly equal the number of infected. That is, the phylogenetic linkage process $N(t)$ may underestimate the true infection process for small t, but ultimately would catch up for larger values of t. Nevertheless, there are scenarios which no amount of effort into discovery nor waiting can reveal contacts, most notably competing risks which incapacitate a contact before identification. When improved HIV surveillance techniques become available and time to infection can be more accurately estimated, our method remains valid by simply replacing time to phylogenetic linkage with time to infection.

So far, the methods discussed in this work only accounting for "known missingness"; that is, for subjects which are known to be in contact with index, but then either refuse to participate to drop-out. However, these methods do not account for "unknown missingness", which are subjects who are never identified before the end of the observation period. We aspire to extend the methods in this current work to account for this second type of missingness.

For our methods to work, we require correction specification of the refusal and drop-out models. Another improvement is to include a doubly-robust [Robins et al., 1994; Tsiatis, 2007a] scheme within the recurrent event framework, which includes an additional outcome model (modeling recurrent events conditional on covariates) to guard against the misspecification of the propensity score (refusal and drop-out) models. Normally, a correctly-specified outcome model can be fit just on the complete case data because the missingness portion in the likelihood factors out under MAR assumptions. In our setting, the manifestation of the events on the count scale collapses over the individual events for which the MAR assumptions are assumed, therefore making it difficult to include such an outcome model. In fact, it seems more tenable to treat the recurrent events as a measurement error problem and form an outcome model this way. We hope to devise the

assumptions of such a measurement error outcome model and derive the resulting doubly-robust estimators.

While our recurrent event framework was applied to a cluster setting, it can also be applied to longitudinal follow-up studies. In such a study, the observation of recurrent events could be terminated at or before the end of the study. For example, the recurrent events could be multiple occurrences of hospitalizations from a group of patients, and the observation of the repeated hospitalization process could be terminated by the end of the study, patient dropout, loss to follow-up, or patient death. Methods such as those developed by Wang et al. [2001] account for informative missingness on the patient scale, but not the event scale. That is, patients could skip intermittent hospitalizations, but reappear later. Our methods can be applied in these cases.

References

Andersen, P. K. and Gill, R. D. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Anderson, D. and Aitkin, M. Variance component models with binary response: Interviewer variability. *JRSS B*, 47:203–210, 1988.

Arellano-Valle, R. B., Heleno, B., and Lachos, V. H. Skew-normal linear mixed models. *Journal of Data Science*, 3.4:415–438, 2005.

Austin, P. C. Variance estimation when using inverse probability of treatment weighting (iptw) with survival analysis. *Statistics in medicine*, 35(30):5642–5655, 2016.

Bates, D. *Mixed-effects modeling with R.* Springer, New York, 2010.

Blum, J. R. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

Bottou, L. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, New York, 2012.

Braun, T. M. and Feng, Z. Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association*, 96(456):1424–1432, 2001.

Butler, S. M. and Louis, T. A. Random effects models with nonparametric priors. *Statistics in medicine*, 11.1415:1981 – 2000, 1992.

Byrd, R. H., Hansen, S. L., Nocedal, J., and Singer, Y. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

Campbell, E. M., Jia, H., Shankar, A., Hanson, D., Luo, W., Masciotra, S., Owen, S. M., Oster, A. M., Galang, R. R., Spiller, M. W., et al. Detailed transmission network analysis of a large opiate-driven outbreak of hiv infection in the united states. *The Journal of infectious diseases*, 216(9):1053–1062, 2017.

Carey, V., Zeger, S. L., and Diggle, P. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526, 1993.

Carnegie, N. B., Wang, R., and De Gruttola, V. Estimation of the overall treatment effect in the presence of interference in cluster-randomized trials of infectious disease prevention. *Epidemiologic Methods*, 5:57 – 68, 2016.

Chamberlain, G. Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics*, 32:189– 218, 1986.

Clémençon, S., Bertail, P., Chautru, E., and Papa, G. Survey schemes for stochastic gradient descent with applications to m-estimation. *arXiv preprint arXiv:1501.02218*, 2015.

Cole, S. R. and Hernán, M. A. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

Cook, R. J. and Lawless, J. *The statistical analysis of recurrent events*. Springer Science & Business Media, 2007.

Cook, R. J., Lawless, J. F., Lakhal-Chaieb, L., and Lee, K.-A. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: Application to skeletal complications in cancer metastatic to bone. *JASA*, 104.485:60–75, 2009.

Cox, D. R. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

Crespi, C. M., Wong, W. K., and Mishra, S. I. Using second–order generalized estimating equations to model heterogeneous intraclass correlation in cluster–randomized trials. *Statistics in Medicine*, 28.5:814–827, 2009.

Demirtas, H. and Hedeker, D. Multiple imputation under power polynomials. *Communications in Statistics–Simulation and Computation*, 37.8:1682–1695, 2008.

Dennis, A. M., Hué, S., Hurt, C. B., Napravnik, S., Sebastian, J., Pillay, D., and Eron, J. J. Phylogenetic insights into regional hiv transmission. *AIDS (London, England)*, 26 (14):1813, 2012.

DiPrete, T. A. and Engelhardt, H. Estimating causal effects with matching methods in the presence and absence of bias cancellation. *Sociological Methods & Research*, 32(4): 501–528, 2004.

Donner, A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, pages 67–82, 1986.

Donner, A. and Klar, N. *Design and analysis of cluster randomization trials in health research*, volume 1. Wiley, New York, 2000.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, pages 2121–2159, 2011.

Evans, B. A., Feng, Z., and Peterson, A. V. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in medicine*, 20(22):3353–3373, 2001.

Fleishman, A. A method for simulating non-normal distributions. *Psychometrika*, 43.4: 521–532, 1978.

Galvao, A. F., Montes-Rojas, G., Sosa-Escudero, W., and Wang, L. Tests for skewness and kurtosis in the one-way error component model. *Journal of Multivariate Analysis*, 122: 35–52, 2013.

Gaolathe, T., Wirth, K. E., Holme, M. P., Makhema, J., Moyo, S., Chakalisa, U., Yankinda, E. K., Lei, Q., Mmalane, M., Novitsky, V., et al. Botswana's progress toward achieving the 2020 unaids 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. *The Lancet HIV*, 3(5):e221–e230, 2016.

Ghidey, W., Lesaffre, E., and Eilers, P. Smooth random effects distribution in a linear mixed model. *Biometrics*, 60.4:945–953, 2004.

Grabowski, M. K. and Redd, A. D. Molecular tools for studying hiv transmission in sexual networks. *Current Opinion in HIV and AIDS*, 9(2):126, 2014.

Guiteras, R., Levinsohn, J., and Mobarak, A. M. Encouraging sanitation investment in the developing world: A cluster-randomized trial. *Science*, 348(6237):903–906, 2015.

Gwet, K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6):1–6, 2002.

Hayes, R. J. and Bennett, S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, 28.2:319–326, 1999.

Hayes, R. and Moulton, L. *Cluster randomised trials*. Chapman & Hall/CRC, Boca Raton, 2009.

Headrick, T. C. *Statistical simulation: power method polynomials and other transformations*. CRC Press, 2009.

Headrick, T. C. and Sawilowsky, S. S. Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, 25.4:417–436, 2000.

Henderson, D. J. and Ullah, A. A nonparametric random effects estimator. *Economics Letters*, 88.3:403–407, 2005.

Huang, S., Fiero, M. H., and Bell, M. L. Generalized estimating equations in cluster randomized trials with a small number of clusters: Review of practice and simulation study. *Clinical Trials*, 13(4):445–449, 2016.

Jiang, J. Partially observed information and inference about non-guassian mixed linear models. *Annals of Statistics*, 33:2695–2731, 2005.

Ke, C. and Wang, Y. Semiparametric nonlinear mixed-effects models and their applications. *JASA*, 96:1272 – 1281, 2001.

Kim, T.-H. and White, H. On more robust estimation of skewness and kurtosis. *Finance Research Letters*, 1(1):56–73, 2004.

Klar, N. and Donner, A. Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine*, 20.24:3729–3740, 2001.

Krippendorff, K. *Content analysis: An introduction to its methodology*. Sage, 2004.

Laird, N. M. and Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, pages 963 – 974, 1982a.

Laird, N. M. and Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982b.

Leckie, G., French, R., Charlton, C., and Browne, W. Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, 39(5):307–332, 2014.

Lehmann, N., Finger, R., Klein, T., and Calanca, P. Sample size requirements for assessing statistical moments of simulated crop yield distributions. *Agriculture*, 3(2):210–220, 2013.

Li, W. and Xue, L. Efficient inference in a generalized partially linear model with random effect for longitudinal data. *Communications in Statistics-theory and Methods*, 44(2): 241–260, 2015.

Liang, K. Y. and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73.1:13–22, 1986.

Liang, K. Y. and Zeger, S. L. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–40, 1992.

Lin, D., Wei, L., Yang, I., and Ying, Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730, 2000.

Lin, T. I. and Lee, J. C. Estimation and prediction in linear mixed models with skewnormal random effects for longitudinal data. *Statistics in medicine*, 27.9:1490 – 1507, 2008.

Lin, X. and Carroll, R. Semiparametric estimation in general repeated measures problems. *JASA B*, 68:68 – 88, 2006.

Litière, S., Alonso, A., and Molenberghs, G. Type i and type ii error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4):1038–1044, 2007.

Luo, H. Generation of non-normaldata: A study of fleishmans powermethod. *Working Paper, Department of Statistics, Uppsala University*, 2011.

Marcus, C. L., Moore, R. H., Rosen, C. L., Giordani, B., Garetz, S. L., Taylor, H. G., Mitchell, R. B., Amin, R., Katz, E. S., Arens, R., et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 368(25): 2366–2376, 2013.

McCulloch, C. and Searle, S. *Generalized, linear, and mixed models*. John Wiley & Sons, New York, 2001.

McCulloch, C. E. and Neuhas, J. M. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical science*, pages 388 – 402, 2011.

Meintanis, S. G. Testing for normality with panel data. *Journal of Statistical Computation and Simulation*, 81(11):1745–1752, 2011.

Miloslavsky, M., Keleş, S., Laan, M. J., and Butler, S. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):239–257, 2004.

Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). *Doklady ANSSSR (translated as Soviet.Math.Docl.)*, 269:543 – 547, 1983.

Newey, W. K. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2): 99–135, 1990.

Paik, M. C. Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics*, pages 19–30, 1992.

Parzen, M. Random effects model for simulating clustered binary data. *unpublished*, 2009.

Pepe, M. S. and Cai, J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American statistical Association*, 88(423):811–820, 1993.

Prague, M., Wang, R., Stephens, A., Tchetgen Tchetgen, E., and DeGruttola, V. Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. *Biometrics*, 72(4):1066–1077, 2016.

Prentice, R. L., Williams, B. J., and Peterson, A. V. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

Rahman, M. M. and Govindarajulu, Z. A modification of the test of shapiro and wilk for normality. *Journal of Applied Statistics*, 24.2:219 – 236, 1997.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Robins, J. M. and Finkelstein, D. M. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89.427:846–866, 1994.

Rotnitzky, A. and Robins, J. Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*, 4:2619–2625, 2005.

Rubin, D. B. Inference and missing data. *Journal of the American Statistical Association*, 63.3:581–592, 1976.

Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Chapman and Hall, 1995.

Smith, C. On the estimation of intraclass correlation. *Annals of Human Genetics*, 21:363 – 373, 1956.

Staples, P., Prague, M., Victor, D. G., and Onnela, J.-P. Leveraging contact network information in clustered randomized trials of infectious processes. *arXiv preprint arXiv:1610.00039*, 2016.

Stephens, A. J., Tchetgen, E. J. T., and De Gruttola, V. Augmented gee for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-and individual-level covariates. *Statistics in Medicine*, 31(10):915, 2012.

Stephens, A. J., Tchetgen Tchetgen, E. J., and DeGruttola, V. D. Locally efficient estimation of marginal treatment effects when outcomes are correlated: is the prize worth the chase? *The International Journal of Biostatistics*, 10.1:59–75, 2014.

Stiratelli, R., Laird, N., and Ware, J. Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971, 1984.

Strijbos, J.-W., Martens, R. L., Prins, F. J., and Jochems, W. M. Content analysis: What are they talking about? *Computers & Education*, 46(1):29–48, 2006.

Sugihara, M. Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 9(1):21–34, 2010.

Sutradhar, B. C. An overview on regression models for discrete longitudinal responses. *Statistical Science*, 18.3:377–393, 2003.

Teuscher, F., Herrendörfer, G., and Guiard, V. The estimation of skewness and kurtosis of random effects in the linear model. *Biometrical journal*, 36(6):661–672, 1994.

Therneau, T. M. Extending the cox model. In *Proceedings of the First Seattle symposium in biostatistics*, pages 51–84. Springer, 1997.

Tsiatis, A. *Semiparametric theory and missing data.* Springer Science & Business Media, 2007a.

Tsiatis, A. *Semiparametric theory and missing data.* Springer Science & Business Media, New York, 2007b.

Ugray, Z., Lasdon, L., Plummer, J. C., Glover, F., Kelly, J., and Marti, R. Scatter search and local nlp solvers: A multistart framework for global optimization. *INFORMS Journal on Computing*, 19.3:328 – 340, 2007.

Van der Laan, M. J. and Robins, J. M. *Unified methods for censored longitudinal data and causality.* Springer Science & Business Media, New York, 2003.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge University Press, Cambridge, 2000.

Verbeke, G. and Lesaffre, E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91.433:217–221, 1996.

Wand, M. and Jones, M. *Comparison of smoothing parameterizations in bivariate kernel density estimation*. New York: Wiley, 1992.

Wang, M.-C., Qin, J., and Chiang, C.-T. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):1057–1065, 2001.

Wang, N. Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90:43 – 52, 2003.

Wang, R. and DeGruttola, V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster randomized trials. 2017.

Wang, R., Goyal, R., Lei, Q., Essex, M., and De Gruttola, V. Sample size considerations in the design of cluster randomized trials of combination hiv prevention. *Clinical trials*, 11(3):309–318, 2014.

Wei, L.-J., Lin, D. Y., and Weissfeld, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408):1065–1073, 1989.

Wertheim, J. O., Pond, S. L. K., Forgione, L. A., Mehta, S. R., Murrell, B., Shah, S., Smith, D. M., Scheffler, K., and Torian, L. V. Social and genetic networks of hiv-1 transmission in new york city. *PLoS pathogens*, 13(1):e1006000, 2017.

West, B. T., Conrad, F. G., Kreuter, F., and Mittereder, F. Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1):181–203, 2018.

Wu, H. and Zhang, J. Local polynomial mixed-effects models for longitudinal data. *JASA*, 97:883 – 897, 2002.

Wu, S., Crespi, C. M., and Wong, W. K. Comparison of methods for estimating the intra-class correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, 33.5:869–880, 2012.

Zeger, S. L., Liang, K. Y., and Albert, P. S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint*, arXiv:1212.5701, 2012.

Zhang, D. and Davidian, M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57.3:795 – 802, 2001.

Zhao, L. P. and Prentice, R. L. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77.3:642–648, 1990.

Ziegler, A., Kastner, C., and Blettner, M. The generalised estimating equations: An annotated bibliography. *Biometrical Journal*, 40.2:115–139, 1998.

Ziegler, A., Kastner, C., and Blettner, M. Familial associations of lipid profiles: A generalised estimating equations approach. *Statistics in Medicine*, 19.24:3345–3357, 2000.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. J. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.

Appendices

Appendix A

Appendix for Chapter 1

A.1 Pseudocode for Stochastic Algorithms

Algorithm 1 S-GEE2 algorithm

Require: $\mathbf{Y}, A_i, \mathbf{Z}_i, \mathbf{X}, \mathbf{W}^R, \upsilon_i, \boldsymbol{g}, \Omega$

$\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0 \leftarrow \mathbf{0}$

for $\omega = 0 : (\Omega - 1)$ **do**

  $U_i \leftarrow$ indices of $\mathbf{Y}_i$ for $i = 1 : I$

  $s_i \sim \text{SRSWOR}(U_i^{\text{obs}}, \upsilon_i)$ for $i = 1 : I$

  $W^S_{\beta i(\omega)} \leftarrow \frac{m_i}{\upsilon_i}[s_i]$ for $i = 1 : I$

  $W^S_{\alpha i(\omega)} \leftarrow \frac{m_i(m_i-1)}{\upsilon_i(\upsilon_i-1)}[(s_i)_2]$ for $i = 1 : I$

  $\tilde{H}_{\beta i(\omega)} \leftarrow \sum_{i=1}^{I} D^{\intercal}_{\beta i(\omega)} V^{-1}_{\beta i(\omega)} W^S_{\beta i(\omega)} D_{\beta i(\omega)}$

  $\tilde{G}_{\beta i(\omega)} \leftarrow \sum_{i=1}^{I} D^{\intercal}_{\beta i(\omega)} V^{-1}_{\beta i(\omega)} W^S_{\beta i(\omega)} E_{\beta i(\omega)}$

  $\tilde{H}_{\alpha i(\omega)} \leftarrow \sum_{i=1}^{I} D^{\intercal}_{\alpha i(\omega)} W^S_{\alpha i(\omega)} D_{\alpha i(\omega)}$

  $\tilde{G}_{\alpha i(\omega)} \leftarrow \sum_{i=1}^{I} D^{\intercal}_{\alpha i(\omega)} W^S_{\alpha i(\omega)} E_{\alpha i(\omega)}$

  $\boldsymbol{\beta}_{(\omega+1)} \leftarrow \boldsymbol{\beta}_{(\omega)} + \gamma_\omega \tilde{H}^{-1}_{\beta i(\omega)} \tilde{G}_{\beta i(\omega)}$

  $\boldsymbol{\alpha}_{(\omega+1)} \leftarrow \boldsymbol{\alpha}_{(\omega)} + \gamma_\omega \tilde{H}^{-1}_{\alpha i(\omega)} \tilde{G}_{\alpha i(\omega)}$

end for

return $\boldsymbol{\beta}_{(\Omega)}, \boldsymbol{\alpha}_{(\Omega)}$

Algorithm 2 Par-S-GEE2 algorithm

Require: $\mathbf{Y}, A_i, \mathbf{Z}_i, \mathbf{X}, \mathbf{W}^R, \upsilon_i, \boldsymbol{g}, \Omega, K$

1: **for** $k = 1 : K$ **do**

2: $\quad (\boldsymbol{\beta}^{(k)}, \boldsymbol{\alpha}^{(k)}) \leftarrow$ S-GEE2$(\mathbf{Y}, \mathbf{Z}^*, \mathbf{X}, \mathbf{W}^R, \boldsymbol{\pi}, \boldsymbol{\rho}^\dagger, \upsilon_i, \boldsymbol{g}, \Omega)$

3: **end for**

4: **return** $\boldsymbol{\beta} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\beta}^{(k)}, \boldsymbol{\alpha} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\alpha}^{(k)}$

Algorithm 3 S-DR-GEE2 algorithm

Require: $\mathbf{Y}, A_i, \mathbf{Z}_i, \mathbf{X}, \mathbf{W}^R, \boldsymbol{\pi}, \boldsymbol{\rho}^\dagger, \upsilon_i, \upsilon_i^{\text{obs}}, \boldsymbol{g}, \Omega$

$\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0 \leftarrow \mathbf{0}$

for $\omega = 0 : (\Omega - 1)$ **do**

  $U_i^{\text{obs}} \leftarrow$ indices of observed $\mathbf{Y}_i$ for $i = 1 : I$

  $U_i \leftarrow$ indices of all $\mathbf{Y}_i$ for $i = 1 : I$

  $s_i^{\text{obs}} \sim \text{SRSWOR}(U_i^{\text{obs}}, \upsilon_i^{\text{obs}})$ for $i = 1 : I$

  $s_i \sim \text{SRSWOR}(U_i, \upsilon_i)$ for $i = 1 : I$

  $W^{RS}_{\beta i(\omega)} \leftarrow \frac{m_i}{\upsilon_i} W^R_{\beta i(\omega)}[s_i^{\text{obs}}]$ for $i = 1 : I$

  $W^{RS}_{\alpha i(\omega)} \leftarrow \frac{m_i(m_i-1)}{\upsilon_i(\upsilon_i-1)} W^R_{\alpha i(\omega)}[(s_i^{\text{obs}})_2]$ for $i = 1 : I$

  $W^{S}_{\beta i(\omega)} \leftarrow \frac{n_i}{\upsilon_i'}[s_i]$ for $i = 1 : I$

  $W^{S}_{\alpha i(\omega)} \leftarrow \frac{n_i(n_i-1)}{\upsilon_i'(\upsilon_i'-1)}[(s_i)_2]$ for $i = 1 : I$

  $\tilde{\zeta}_{\beta i(\omega)} \leftarrow \sum_{a=0}^{1} p^a(1-p)^{1-a} D^\intercal_{\beta i(\omega)}(A = a) V^{-1}_{\beta i(\omega)} W^S_{\beta i(\omega)} E''_{\beta i(\omega)}(A = a)$ for $i = 1 : I$

  $\tilde{\zeta}_{\alpha i(\omega)} \leftarrow \sum_{a=0}^{1} p^a(1-p)^{1-a} D^\intercal_{\alpha i(\omega)}(A = a) W^S_{\alpha i(\omega)} E''_{\alpha i(\omega)}(A = a)$ for $i = 1 : I$

  $\tilde{H}_{\beta i(\omega)} \leftarrow \sum_{i=1}^{I} \sum_{a=0}^{1} p^a(1-p)^{1-a} D^\intercal_{\beta i(\omega)}(A = a) V^{-1}_{\beta i(\omega)} W^S_{\beta i(\omega)} D_{\beta i(\omega)}(A = a)$

  $\tilde{G}_{\beta i(\omega)} \leftarrow \sum_{i=1}^{I} [D^\intercal_{\beta i(\omega)} V^{-1}_{\beta i(\omega)} W^{RS}_{\beta i(\omega)} E'_{\beta i(\omega)} + \tilde{\zeta}_{\beta i(\omega)}]$

  $\tilde{H}_{\alpha i(\omega)} \leftarrow \sum_{i=1}^{I} \sum_{a=0}^{1} p^a(1-p)^{1-a} D^\intercal_{\alpha i(\omega)}(A = a) W^S_{\alpha i(\omega)} D_{\alpha i(\omega)}(A = a)$

  $\tilde{G}_{\alpha i(\omega)} \leftarrow \sum_{i=1}^{I} [D^\intercal_{\alpha i(\omega)} W^{RS}_{\alpha i(\omega)} E'_{\alpha i(\omega)} + \tilde{\zeta}_{\alpha i(\omega)}]$

  $\boldsymbol{\beta}_{(\omega+1)} \leftarrow \boldsymbol{\beta}_{(\omega)} + \gamma_\omega \tilde{H}^{-1}_{\beta i(\omega)} \tilde{G}_{\beta i(\omega)}$

  $\boldsymbol{\alpha}_{(\omega+1)} \leftarrow \boldsymbol{\alpha}_{(\omega)} + \gamma_\omega \tilde{H}^{-1}_{\alpha i(\omega)} \tilde{G}_{\alpha i(\omega)}$

end for

return $\boldsymbol{\beta}_{(\Omega)}, \boldsymbol{\alpha}_{(\Omega)}$

A.2 Time Complexity Proofs

In proving the time-complexities associated with iterations of standard Newton-Raphson or stochastic Newton-Raphson, we make many uses of the following facts:

Fact 1: The time complexity of multiplying matrix $A_{n\times m}$ and $B_{m\times p}$ is $\mathcal{O}(nmp)$.

Fact 2: The complexity of inverting an $n \times n$ matrix is $\mathcal{O}(n^3)$.
Fact 3: $\mathcal{O}(f(n)) + \mathcal{O}(g(n)) = \mathcal{O}(\max(f, g)(n))$.

Omit the R and Y indices, for the computational complexity results are the same in both cases. Let $d_\beta = \dim(\beta), d_\alpha = \dim(\alpha)$. We make the assumptions that d_β, d_α, I are fixed; hence $\mathcal{O}(d_\beta) = \mathcal{O}(d_\alpha) = \mathcal{O}(I) = \mathcal{O}(1)$. Furthermore, we conduct the proofs as if we have no natural missingness in data, for proofs with the latter return the same complexities. We can decompose a covariance matrix $V = U^{1/2}CU^{1/2}$, where C is a correlation matrix, and U is a diagonal matrix with variance entries.

Table 1.1 contains a total of 12 complexities. We break them down into four sub-theorems. Additionally, we require the assumption that $\upsilon_i = \mathcal{O}(1)$; that is, our subsample size does not grow with respect to n_i.

Sub-theorem 1

In the presence of standard Newton-Raphson, an iteration of the GEE1 portion with

(i) Arbitrary covariance matrix

(ii) Equicorrelation matrix

(iii) Independence covariance matrix

are of complexities $\mathcal{O}(\max_i n_i^3), \mathcal{O}(\max_i n_i), \mathcal{O}(\max_i n_i)$ respectively.

Proof. (i) Let us list the steps required in the computation:

1. Computing $V_{\beta i\omega}^{-1}$:

   (a) Compute $C_{\beta i\omega}^{-1}$ and $U_{\beta i\omega}^{-1/2}$, which are of complexities $\mathcal{O}(n_i^3)$ and $\mathcal{O}(n_i)$, since $U_{\beta i\omega}$ is diagonal. The time complexity in computing $C_{\beta i\omega}^{-1}$, through either Gauss-Jordan elimination or Cholesky decomposition, is $\mathcal{O}(n_i^3)$ and cannot be sped up

except through highly specialized numerically-optimized matrix algorithms (i.e. Coppersmith–Winograd algorithm).

(b) Compute $C^{-1}_{\beta i\omega} U^{-1/2}_{\beta i\omega}$. Because $U^{1/2}_{\beta i\omega}$ is diagonal, this becomes just multiplying the diagonal of $U^{-1/2}_{\beta i\omega}$ against each row of $C^{-1}_{\beta i\omega}$, and has complexity $\mathcal{O}(n_i^2)$.

(c) Left-multiply $C^{-1}_{\beta i\omega} U^{-1/2}_{\beta i\omega}$ with $U^{-1/2}_{\beta i\omega}$. This is also $\mathcal{O}(n_i^2)$.

Hence, computing $V^{-1}_{\beta i\omega}$ has complexity $\mathcal{O}(n_i^3)$.

2. Computing $H^{-1}_{\beta i\omega}$, having already computed $V^{-1}_{\beta i\omega}$:

   (a) Compute $V^{-1}_{\beta i\omega} D_{\beta i\omega}$. This has complexity $\mathcal{O}(d_\beta n_i^2) = \mathcal{O}(n_i^2)$.

   (b) Left-multiply $V^{-1}_{\beta i\omega} D_{\beta i\omega}$ by $D^{\intercal}_{\beta i\omega}$; this has complexity $\mathcal{O}(d_\beta^2 n_i) = \mathcal{O}(n_i)$.

   (c) Invert the resulting $D^{\intercal}_{\beta i\omega} V^{-1}_{\beta i\omega} D_{\beta i\omega}$. This is time complexity $\mathcal{O}(d_\beta^3) = \mathcal{O}(1)$.

   Hence, complexity in computing $H_{\beta i\omega}$ is $\mathcal{O}(n_i^2)$.

3. Computing $G_{\beta i\omega}$, having already computed $V^{-1}_{\beta i\omega}$:

   (a) All steps are almost the same as computing $H_{\beta i\omega}$, except for 2(a), where we have $V^{-1}_{\beta i\omega} E_{\beta i\omega}$, which is still $\mathcal{O}(n_i^2)$

   Overall, computing $G_{\beta i\omega}$ is $\mathcal{O}(n_i^2)$

4. Computing $H^{-1}_{\beta i\omega} G_{\beta i\omega}$, having already computed $H^{-1}_{\beta i\omega}$ and $G_{\beta i\omega}$, is just $\mathcal{O}(d_\beta) = \mathcal{O}(1)$.

Overall, steps 1 – 4 is of $\mathcal{O}(n_i^3)$, due to computing $V^{-1}_{\beta i\omega}$.

5. Perform steps 1 – 4 for each i. The time complexity is $\sum_{i=1}^{I} \mathcal{O}(n_i^3) = \mathcal{O}(\max_i n_i^3)$.

6. Summing up $H^{-1}_{\beta i\omega} G_{\beta i\omega}$ is $\mathcal{O}(I) = \mathcal{O}(1)$, and then adding this resulting quantity is $\mathcal{O}(1)$.

Overall, we have $\mathcal{O}(\max_i n_i^3)$.

(ii) Since $C_{\beta i\omega}$ is equicorrelated, we have that

$$C_{\beta i\omega}^{-1} = (1-\rho_i)^{-1}\left(\mathbf{I}_{n_i} - \frac{\rho_i}{1+(n-1)\rho_i} J_{n_i}\right)$$

by Woodbury's formula, where J_{n_i} is an $n_i \times n_i$ matrix of 1's. Hence, in computing $H_{\beta i\omega} = D_{\beta i\omega}^\intercal V_{\beta i\omega}^{-1} D_{\beta i\omega}$, we would compute

$$\underbrace{(1-\rho_i)^{-1} D_{\beta i\omega}^\intercal U_{\beta i\omega}^{-1} D_{\beta i\omega}}_{Q_1} - \underbrace{\frac{\rho_i}{(1+(n_i-1)\rho_i)(1-\rho_i)} D_{\beta i\omega}^\intercal U_{\beta i\omega}^{-1/2} J_{n_i} U_{\beta i\omega}^{-1/2} D_{\beta i\omega}}_{Q_2}$$

Since $U_{\beta i\omega}^{-1}$ is diagonal, we can perform an element-wise product with the diagonal, and hence computation of Q_1 is $\mathcal{O}(n_i)$. In computing Q_2, notice that to compute $J_{n_i} U_{\beta i\omega}^{-1/2} D_{\beta i\omega}$ is to

1. Perform $U_{\beta i\omega}^{-1/2} D_{\beta i\omega}$, which can be done through element-wise product.
2. Sum each column of the resulting $U_{\beta i\omega}^{-1/2} D_{\beta i\omega}$ into a row vector.
3. Repeat each row n_i times into a matrix.

This has time complexity $\mathcal{O}(n_i)$. Then, left-multiplying this quantity by $U_{\beta i\omega}^{-1/2}$ and then again by $D_{\beta i\omega}^\intercal$ is $\mathcal{O}(n_i)$ and $\mathcal{O}(d_\beta^2 n_i) = \mathcal{O}(n_i)$. Overall, computing $H_{\beta i\omega}^{-1}$ is now $\mathcal{O}(n_i)$. Analogous steps can be done to calculate $G_{\beta i\omega}$, which is now $\mathcal{O}(n_i)$. The rest of the proof follows steps 4 – 6 of (i), which results in $\mathcal{O}(\max_i n_i)$.

(iii) For no correlation, inverting $V_{\beta i\omega}$ requires inverting the diagonal entries; this is still of complexity $\mathcal{O}(n_i)$. Rest of the proof follows as (i). □

Sub-theorem 2

In the presence of standard Newton-Raphson, an iteration of the GEE2 portion with

(i) Arbitrary covariance matrix

(ii) Equicorrelation matrix

(iii) Independence covariance matrix

are of complexities $\mathcal{O}(\max_i n_i^6)$, $\mathcal{O}(\max_i n_i^2)$, $\mathcal{O}(\max_i n_i^2)$ respectively.

Proof. All rows and columns in the proofs for GEE1 now have lengths $\binom{n_i}{2} \sim n_i^2$ in place of n_i. Hence, all exponents in computational complexities in Theorem A.2 are doubled. □

Now, let's continue with stochastic Newton-Raphson. Define $D^{\text{sub}}_{\beta i\omega}, E^{\text{sub}}_{\beta i\omega}$ as the resulting $D_{\beta i\omega}, E_{\beta i\omega}$ with only rows corresponding to subsample s_i; we see that, the dimensions of these matrices are now $\upsilon_i \times d_\beta$ and $\upsilon_i \times 1$, respectively. Let $\tilde{W}^{R\text{sub}}_{\beta i(\omega)}$ equal $\tilde{W}^{R}_{\beta i(\omega)}$ except with both rows and columns associated with zero diagonal elements removed; this has dimension $\upsilon_i \times \upsilon_i$. We can analogously define this for $D^{\text{sub}}_{\alpha i\omega}, E^{\text{sub}}_{\alpha i\omega}, \tilde{W}^{R\text{sub}}_{\alpha i\omega}$, where any dimension with a $\binom{n_i}{2}$ is replaced with $\binom{\upsilon_i}{2}$.

Sub-theorem 3

In the presence of stochastic Newton-Raphson, an iteration of the GEE1 portion with

(i) Arbitrary covariance matrix

(ii) Equicorrelation matrix

(iii) Independence covariance matrix

will be of complexities $\mathcal{O}(\max_i n_i^3), \mathcal{O}(\max_i n_i), \mathcal{O}(1)$ respectively.

Proof. (i) We cannot exploit sparsity here, for the largest complexity object, $V^{-1}_{\beta i\omega}$, would still need to be computed, which is $\mathcal{O}(n_i^3)$.

(ii) Let's list again the steps in computing the quantities.

1. Computing $\tilde{H}_{\beta i\omega}^{-1}$: Using Woodbury's formula, the computation of $\tilde{H}_{\beta i\omega}$ would be

$$\underbrace{(1-\rho_i)^{-1}D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1}\tilde{W}_{\beta i\omega}^{R}D_{\beta i\omega}}_{\tilde{Q}_1} - \underbrace{\frac{\rho_i}{(1+(n_i-1)\rho_i)(1-\rho_i)}D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1/2}J_{n_i}U_{\beta i\omega}^{-1/2}\tilde{W}_{\beta i\omega}^{R}D_{\beta i\omega}}_{\tilde{Q}_2}$$

Exploiting sparsity, each term is equivalent to

$$\tilde{Q}_1 = (1-\rho_i)^{-1}D_{\beta i\omega}^{\intercal}(U_{\beta i\omega}^{\text{sub}})^{-1}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}$$
$$\tilde{Q}_2 = \frac{\rho_i}{(1+(n_i-1)\rho_i)(1-\rho_i)}D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1/2}J_{n_i\times\upsilon_i}(U_{\beta i\omega}^{\text{sub}})^{-1/2}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}$$

   (a) Computing $\tilde{Q}_1$ first performs the following steps:

   $$\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}} \mapsto U_{\beta i\omega}^{-1}\tilde{W}_{\beta i\omega}^{S}D_{\beta i\omega} \mapsto D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1}\tilde{W}_{\beta i\omega}^{S}D_{\beta i\omega} \mapsto (1-\rho_i)^{-1}D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1}\tilde{W}_{\beta i\omega}^{S}D_{\beta i\omega}$$

   which sequentially, conditioned on performing the previous computation, is $\mathcal{O}(d_\beta\upsilon_i)$, $\mathcal{O}(d_\beta\upsilon_i)$, $\mathcal{O}(d_\beta^2\upsilon_i)$, and $\mathcal{O}(d_\beta^2)$. The sum of these three complexities is $\mathcal{O}(\upsilon_i)$.

   (b) Computing Q_2 first performs the following steps:

   $$\begin{aligned}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}} &\mapsto (U_{\beta i\omega}^{\text{sub}})^{-1/2}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}\\ &\mapsto J_{n_i\times\upsilon_i}(U_{\beta i\omega}^{\text{sub}})^{-1/2}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}\\ &\mapsto U_{\beta i\omega}^{-1/2}J_{n_i\times\upsilon_i}(U_{\beta i\omega}^{\text{sub}})^{-1/2}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}\\ &\mapsto D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1/2}J_{n_i\times\upsilon_i}(U_{\beta i\omega}^{\text{sub}})^{-1/2}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}\\ &\mapsto \frac{\rho_i}{(1+(n_i-1)\rho_i)(1-\rho_i)}D_{\beta i\omega}^{\intercal}U_{\beta i\omega}^{-1/2}J_{n_i\times\upsilon_i}(U_{\beta i\omega}^{\text{sub}})^{-1/2}\tilde{W}_{\beta i\omega}^{R\text{sub}}D_{\beta i\omega}^{\text{sub}}\end{aligned}$$

   The time complexities of each step is $\mathcal{O}(d_\beta\upsilon_i)$, $\mathcal{O}(d_\beta\upsilon_i)$, $\mathcal{O}(d_\beta\upsilon_i)$, $\mathcal{O}(d_\beta n_i)$, $\mathcal{O}(d_\beta^2 n_i)$, and $\mathcal{O}(d_\beta^2)$. Notice that the third step cannot be simplified due to the $J_{n_i\times\upsilon_i}$ matrix separating $D_{\beta i\omega}^{\intercal}$ and $\tilde{W}_{\beta i\omega}^{R\text{sub}}$.

   (c) Inverting $H_{\beta i\omega}$ is again $\mathcal{O}(d_\beta^3)$, which is dominated by the other steps.

   Hence, calculating $H_{\beta i\omega}^{-1}$ is $\mathcal{O}(n_i)$.

2. Steps in computing $G_{\beta i\omega}^{-1}$ are analogous to step 1, and also $\mathcal{O}(n_i)$

Repeat steps 4 – 6 of Theorem A.2 (i), we again have $\mathcal{O}(\max_i n_i)$.

Remark: For the cases of a general or equicorrelated $C_{\beta i\omega}$, the time complexities of standard and stochastic Newton-Raphson iterations are the same. Intuitively, although we want to feed a subset of the data into the scoring equations, we cannot make full use of sparsity because the inverse-covariance matrix $V_{\beta i\omega}^{-1}$ forces a "mixing" of all the observations, including into missing vector slots. The next two settings no longer have any correlations, and hence we can make full use of sparsity.

(iii) We present just the proof of computing $\tilde{H}_{\beta i\omega}$, since this and $\tilde{G}_{\beta i\omega}$ are bottlenecks in the computation, and both have the same complexities. We now just need to compute

$$D_{\beta i\omega}^{\intercal} U_{\beta i\omega} \tilde{W}_{\beta i\omega}^{R} D_{\beta i\omega} = (D_{\beta i\omega}^{\text{sub}})^{\intercal} U_{\beta i\omega}^{\text{sub}} \tilde{W}_{\beta i\omega}^{R\text{sub}} D_{\beta i\omega}^{\text{sub}}$$

Sequentially, the steps in computing

$$\tilde{W}_{\beta i\omega}^{R\text{sub}} D_{\beta i\omega}^{\text{sub}} \mapsto U_{\beta i\omega}^{\text{sub}} \tilde{W}_{\beta i\omega}^{R\text{sub}} D_{\beta i\omega}^{\text{sub}} \mapsto (D_{\beta i\omega}^{\text{sub}})^{\intercal} U_{\beta i\omega}^{\text{sub}} \tilde{W}_{\beta i\omega}^{R\text{sub}} D_{\beta i\omega}^{\text{sub}}$$

are of $\mathcal{O}(d_\beta \upsilon_i), \mathcal{O}(d_\beta \upsilon_i), \mathcal{O}(d_\beta^2 \upsilon_i)$; overall, this is of time complexity $\mathcal{O}(\upsilon_i) = \mathcal{O}(1)$. □

Sub-theorem 4

In the presence of stochastic Newton-Raphson, an iteration of the GEE2 portion with

(i) Arbitrary covariance matrix

(ii) Equicorrelation matrix

(iii) Independence covariance matrix

will be of complexities $\mathcal{O}(\max_i n_i^6), \mathcal{O}(\max_i n_i^2), \mathcal{O}(1)$ respectively.

Proof. Apply Sub-theorem 3 with υ_i replaced with $\binom{\upsilon_i}{2} \sim \upsilon_i^2$, and we are done. □

A.3 Proof of CAN for DR estimator

It suffices to show $\mathbb{E}[\tilde{\Phi}_i^Y(\mathbf{Z}_i^*, \mathbf{X}_i, \mathbf{R}_i, \boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y)] = 0$ from Eq 1.8 whenever the OM or PS is correctly specified.

Case 1: OM is correctly specified

Under this case, we have $\overline{\pi}_{ij} = \pi_{ij}$ and $\overline{\rho}_{ijj'} = \rho_{ijj'}$, so we have that $\mathbb{E}[\overline{\pi}_{ij}|A_i] = \pi_i^*$ and $\mathbb{E}[\overline{\rho}_{ijj'}^{\dagger}|A_i] = \rho_i^*$. From this, it is easy to verify $\mathbb{E}[E_i'|\mathbf{R}_i, \mathbf{X}_i, \mathbf{Z}_i, A_i] = \mathbf{0}$ and $\mathbb{E}[\zeta_i] = \mathbf{0}$. Hence,

$$\begin{aligned}\mathbb{E}[\tilde{\Phi}_i^Y] &= \mathbb{E}[D_i^\intercal V_i^{-1} W_i^R E_i' + \zeta_i] = \mathbb{E}[\mathbb{E}[D_i^\intercal V_i^{-1} W_i^R E_i'|\mathbf{R}_i, \mathbf{X}_i, \mathbf{Z}_i, A_i]] + \mathbb{E}[\zeta_i] \\ &= \mathbb{E}[D_i^\intercal V_i^{-1} W_i^R \mathbb{E}[E_i'|\mathbf{R}_i, \mathbf{X}_i, \mathbf{Z}_i, A_i]] + \mathbf{0} = E[D_i^\intercal V_i^{-1} W_i^R \cdot \mathbf{0}] = \mathbf{0}\end{aligned}$$

Case 2: PS is correctly specified

Under this case, we have $\overline{\pi}_{ij}^R = \pi_{ij}^R$ and $\overline{\rho}_{ijj'}^R = \rho_{ijj'}^R$; together, this implies that $\mathbb{E}[W_i^R] = \mathbf{I}$. First, using the fact that $E_i' + E_i'' = E_i$, we may express

$$\begin{aligned}\tilde{\Phi}_i^Y &= D_i^\intercal V_i^{-1} W_i^R E_i - D_i^\intercal V_i^{-1} W_i^R E_i'' - D_i^\intercal V_i^{-1} E_i'' + D_i^\intercal V_i^{-1} W_i^R E_i'' + \zeta_i \\ &= \underbrace{D_i^\intercal V_i^{-1} W_i^R E_i}_{\mathbb{Q}_1} + \underbrace{D_i^\intercal (V_i^{-1} - V_i^{-1} W_i^R) E_i''}_{\mathbb{Q}_2} + \underbrace{\zeta_i - D_i^\intercal V_i^{-1} E_i''}_{\mathbb{Q}_3}\end{aligned}$$

It now suffices to show $\mathbb{E}[\mathbb{Q}_1], \mathbb{E}[\mathbb{Q}_2], \mathbb{E}[\mathbb{Q}_3] = \mathbf{0}$. We have $\mathbb{E}[\mathbb{Q}_1] = \mathbf{0}$ by standard IPW-GEE2. Next,

$$\mathbb{E}[\mathbb{Q}_2] = \mathbb{E}[D_i^\intercal V_i^{-1} \mathbb{E}[\mathbf{I} - W_i^R|\mathbf{X}_i, \mathbf{Z}_i^*] E_i''] = \mathbb{E}[D_i^\intercal V_i^{-1} (\mathbf{I} - \mathbf{I}) E_i''] = \mathbf{0}$$

Finally,

$$\begin{aligned}\mathbb{E}[\mathbb{Q}_3] &= \mathbb{E}[\zeta_i] - \mathbb{E}[D_i^\intercal V_i^{-1} E_i''] \\ &= \mathbb{E}[\mathbb{E}[D_i^\intercal V_i^{-1} E_i''|\mathcal{D}_i \setminus A_i]] - \mathbb{E}[D_i^\intercal V_i^{-1} E_i''] \\ &= \mathbb{E}[D_i^\intercal V_i^{-1} E_i''] - \mathbb{E}[D_i^\intercal V_i^{-1} E_i''] \\ &= \mathbf{0}\end{aligned}$$

Under certain regularity assumption defined in Van der Vaart [2000], we can demonstrate with the Slutsky's theorem and the central limit theorem that any estimator solving this Doubly Robust estimating equation is CAN.