Published in final edited form as:

J Comput Graph Stat. 2020 ; 29(3): 547–561. doi:10.1080/10618600.2019.1710156.

A stochastic second-order generalized estimating equations approach for estimating association parameters

Tom Chen¹, Eric J. Tchetgen Tchetgen², Rui Wang^{1,3,*}

¹Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

²Department of Statistics, The Wharton School, The University of Pennsylvania

³Department of Biostatistics, Harvard T.H.Chan School of Public Health

Abstract

Design and analysis of cluster randomized trials must take into account the intraclass correlation coefficient (ICC), which quantifies the correlation among outcomes from the same cluster. Second-order generalized estimating equations (GEE2) provides a statistically robust way in estimating this quantity and other association parameters. However, GEE2 becomes computationally infeasible as cluster sizes grow. This paper proposes a stochastic variant to fitting GEE2 which alleviates reliance on parameter starting values and provides substantially faster speeds and higher convergence rates than the widely used deterministic Newton-Raphson method. We also propose new estimators for the ICC which account for informative missing outcome data through the use of GEE2, for which we incorporate a “second-order” inverse probability weighting scheme and “second-order” doubly robust (DR) estimating equations that guard against partial model misspecification. Our proposed methods are evaluated through simulations and applied to data from a cluster randomized trial in Bangladesh evaluating the effect of different marketing interventions on the use of hygienic latrines.

Keywords

Clustered data; doubly robust; GEE2; Robbins-Monro

1 Introduction

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

*Corresponding author: rwang@hsph.harvard.edu.

SUPPLEMENTAL MATERIALS

supplemental.pdf: This pdf file contains a glossary of acronyms, pseudo code for implementing stochastic algorithms listed throughout this paper, proofs of the time complexities listed in Table 1, and proof of consistency and asymptotic normality of DR-GEE2 in Section 4.3.

ICC_Code.zip: Contains R code used in simulations and the Bangladeshi sanitation application.

accurately computing sample sizes needed to achieve a certain power level in a CRT. For example, in a matched-pair CRT with 15 pairs and a sample size of 300 within each cluster as seen in the Botswana Combination Prevention Project (Gaolathe et al., 2016; Wang et al., 2014), the power to detect a 40% reduction in the 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 outcome measure 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 tests (e.g., t-test) (Hayes and Moulton, 2009) and nonparametric permutation tests (Braun and Feng, 2001; Wang and De Gruttola, 2017). Despite its importance, obtaining reliable estimates of the 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 intervention effect on the cluster – the difference between outcomes in the treated cluster versus the untreated cluster (Carnegie et al., 2016). Mixed models and their extensions (Anderson and Aitkin, 1988; Laird and Ware, 1982; Stiratelli et al., 1984) are a popular choice in estimating the intervention effect while accounting for the clustering effect. Variance modeling for continuous outcomes (Leckie et al., 2014; Paik, 1992; West et al., 2018) have also been proposed. These methods require parametric assumptions and provide a conditional interpretation of intervention effects. Here, we focus on the generalized estimating equations (GEEs) (Liang and Zeger, 1986) approach. This estimation procedure is semiparametric in that it assumes the correct specification of a marginal mean model, instead of a full likelihood, in order to yield a consistent and asymptotically normal estimator of the treatment effect. The GEE approach with correct specification of the working correlation structure is theoretically optimal and has been demonstrated in simulations to have efficiency gains (Fitzmaurice, 1995), while others noted that using independence correlation structure is often adequate (McDonald, 1993; Zeger, 1988). In either case, 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 can be inaccurate. This has motivated the development of 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. Lipsitz and Fitzmaurice (1996) proposed a conditional estimating equations approach which was demonstrated to have additional efficiency gains.

Several authors (Carey et al., 1993; Ziegler et al., 1998) have noted of convergence problems regarding GEE2. Solving GEE2 is difficult due to the far larger stack of estimating equations for the association parameters, leading to excessive computing time in obtaining solutions to these equations. Carey et al. (1993) championed the use of odds ratios in measuring association, for which they proposed fitting with alternating logistic regression. Their algorithm is computationally efficient because it exploits a clever relationship in the

conditional residuals, but is specific to the odds ratio setting. In our context, as the ICC is our parameter of interest, we focus on approaches that model ICCs directly. 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 and the more general class of Robbins-Monro algorithms (Robbins and Monro, 1951). 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 that account for missing outcome data. When outcomes are assumed missing completely at random (Rubin, 1976) (outcomes are missing independently of both observed and unobserved data), the GEE2 analysis performed on complete-case CRT data provides consistent and asymptotically normal estimators for the treatment effect 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 estimators unless all factors contributing to the propensity of being missing are included in a correctly-specified outcome model. Extensions (Hall and Severini, 1998) have been proposed where estimating equations are derived assuming multivariate normality and have been shown to retain high efficiency and have minimal bias even under MAR (Lipsitz et al., 2000). Furthermore, Le Cessie and Van Houwelingen (1994) and Parzen et al. (2007) have proposed pseudo-likelihood methods which model both missingness and outcome processes. Our work is based on inverse probability weighting (IPW) and outcome model (OM) augmentation made popular by the semiparametric methods field (Robins et al., 1994; Tsiatis, 2007; Van der Laan and Robins, 2003). In the context of CRTs, there is no natural ordering of the outcomes within a community and the missingness pattern is non-monotone. 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 resulting estimators are doubly-robust (DR) in the sense that they require either the IPW or outcome model 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.

Section 2 gives background on the standard GEE2 (non-stochastic and no missing data). In Section 3, we introduce the Robbins-Monro algorithm and expand on the stochastic paradigm to fitting GEE2, which we coin as stochastic 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 4, we draw from the semiparametric methods to construct IPW and DR variants of GEE2, which we call IPW-GEE2 and DR-GEE2, and explain how to further adapt stochastic GEE2 to these procedures. We evaluate the performance of the proposed estimators and the proposed computational algorithms with simulations in Section 5 and apply the new estimators and algorithms to

analyze Bangladeshi sanitation data (Guiteras et al., 2015) in Section 6. We end with a discussion in Section 7. Proofs are relegated to Supplemental Materials.

2 Notation and standard GEE2

Henceforth, we work with binary outcomes $Y_{ij} \in \{0, 1\}$ for subject $j = 1, \dots, n_i$ in cluster $i = 1, \dots, I$; 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 its 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

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

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 verify that $\rho_i^* = \mathbb{E}[\rho_{ijj'}^\dagger | A_i]$, where

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

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

Let $\hat{\pi}_{ij}$ be an estimator of π_{ij} converging to the limit $\bar{\pi}_{ij}$, which may or may not equal the true π_{ij} . Likewise, let $\hat{\rho}_{ijj'}$ be an estimator of $\rho_{ijj'}$ with limit $\bar{\rho}_{ijj'}$. Standard models for $\hat{\pi}_{ij}$ include logistic or probit regression, while a model for $\hat{\rho}_{ijj'}$ 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, we apply it here to map the $[-1, 1]$ support of ρ_i^* onto $\mathbb{R}$.

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

$$\text{logit}(\pi_i^*(\beta_Y^*; A_i)) = \beta_{0Y}^* + \beta_{AY}^* A_i, \quad \text{atanh}(\rho_i^*(\alpha_Y^*; A_i)) = \alpha_{0Y}^* + \alpha_{AY}^* A_i \quad (2)$$

to produce estimators (β_Y^*, α_Y^*) . Eq 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 $\bar{\pi}_i^* = \pi_i^*$ and $\bar{\rho}_i^* = \rho_i^*$. In the standard GEE2 framework, we would estimate (β_Y^*, α_Y^*) as the solution to the equations

$$\mathbf{0} = \sum_{i=1}^I \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{E}_i \stackrel{\text{def}}{=} \sum_{i=1}^I \mathbf{S}_i^Y(A_i, \boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*) \quad (3)$$

where

$$\mathbf{D}_i = \frac{\partial(\pi_i^*(\boldsymbol{\beta}_Y^*; A_i), \rho_i^*(\boldsymbol{\alpha}_Y^*; A_i))}{\partial(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)^\top}, \quad \mathbf{V}_i = \text{Cov}\left(\begin{matrix} \mathbf{Y}_i \\ \mathcal{E}(\mathbf{Y}_i) \end{matrix}\right), \quad \mathbf{E}_i = \begin{pmatrix} \mathbf{Y}_i - \pi_i^*(\boldsymbol{\beta}_Y^*) \\ \mathcal{E}(\mathbf{Y}_i) - \rho_i^*(\boldsymbol{\alpha}_Y^*) \end{pmatrix}$$

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

Note that the working covariance matrix $\mathbf{V}_i$ 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 there are others (Liang and Zeger, 1992; Zhao and Prentice, 1990). We choose the above parametrization because it specifically targets the treatment-specific ICC ρ_i^* instead of, say, the cross moments or covariance, but the proposed framework is applicable to these other parametrizations.

In what follows, Section 3 and 4 will delineate two adaptations of GEE2. The first proposal provides a scheme in reducing the computation time and numerical errors which arise when fitting GEE2. The second proposal devises adjustments to GEE2 which account for missing data through the use of inverse-probability weighting and outcome model augmentation.

3 Proposal I: Stochastic GEE2

3.1 Background: Robbins-Monro Algorithm

GEE2 is ordinarily solved using Newton-Raphson, which has iterations $\boldsymbol{\theta}_{(0)}, \boldsymbol{\theta}_{(1)}, \dots$ of the form

$$\boldsymbol{\theta}_{(\omega+1)} = \boldsymbol{\theta}_{(\omega)} + \mathbf{H}_{(\omega)}^{-1} \mathbf{G}_{(\omega)} \quad (4)$$

where

$$\mathbf{H}_{(\omega)} = \sum_{i=1}^I \mathbf{D}_{i(\omega)}^\top \mathbf{V}_{i(\omega)}^{-1} \mathbf{D}_{i(\omega)}, \quad \mathbf{G}_{(\omega)} = \sum_{i=1}^I \mathbf{D}_{i(\omega)}^\top \mathbf{V}_{i(\omega)}^{-1} \mathbf{E}_{i(\omega)} \quad (5)$$

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

In general, the Robbins-Monro algorithm states that, when solving for θ in the equation $\mathbf{G}(\theta) = 0$, if we instead possess a random variable $\mathbf{G}(\theta)$ such that $\mathbb{E}[\mathbf{G}(\theta)] = \mathbf{G}(\theta)$ and iterate $\theta_{(\omega+1)} = \theta_{(\omega)} - \gamma_{(\omega)} \mathbf{G}_{(\omega)}$, where learning rates $\gamma_{(\omega)} > 0$ satisfy $\sum_{(\omega)} \gamma_{(\omega)} = \infty$ and $\sum_{(\omega)} \gamma_{(\omega)}^2 < \infty$, then $\theta_{(\omega)} \rightarrow \theta_0$ in L^2 -mean (Robbins and Monro, 1951) and almost surely (Blum, 1954), subject to a few additional regularity conditions. The standard choice for learning rates is $\gamma_{(\omega)} = (1 + \omega)^{-1}$. The Robbins-Monro algorithm is useful whenever we can find a $\mathbf{G}$ which is significantly faster to compute than $\mathbf{G}$. For example, consider the general M -estimation problem (for which GEE is a special case) and suppose our estimating equation takes the form $\mathbf{G}(\theta) = \sum_{i=1}^I \mathbf{G}_i(\theta)$. It is easy to confirm that $\mathbf{G}(\theta) = \sum_{i \in s} \mathbf{G}_i(\theta)/p_i$ satisfies $\mathbb{E}[\mathbf{G}(\theta)] = \mathbf{G}(\theta)$, where s is a randomly chosen subset of $U = \{1, \dots, I\}$ according to any sampling design with inclusion probabilities $p_i = \mathbb{P}(i \in s)$. Here, instead of performing I function evaluations, we only need to perform $|s|$ evaluations at each iteration. If we take our sampling scheme to be a simple random sample without replacement (SRSWOR) of size v , this reduces to minibatch stochastic gradient descent (Clmenon et al., 2015). Our focus is different here, 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 (see for example, Gaolathe et al. (2016)). The design of the proposed class of stochastic GEE2 algorithms differs from minibatch stochastic gradient descent in that we are improving iteration speed not through evaluating fewer of the functional summands $\{\mathbf{G}_i\}_{i=1}^I$ (i.e. evaluating fewer clusters), but rather evaluating an unbiased and computationally-easier estimate of each summand $\mathbf{G}_i$ (done through sampling a subset of individuals per cluster). Also, we shall incorporate stochastic Hessians to form stochastic Fisher scoring / Newton-Raphson iterations $\theta_{(\omega+1)} = \theta_{(\omega)} + \gamma_{(\omega)} \mathbf{H}_{(\omega)}^{-1} \mathbf{G}_{(\omega)}$, which Byrd et al. (2016) proves to convergence almost surely as well. Unlike the stochastic gradient $\mathbf{G}$, the accuracy for a stochastic Hessian $\mathbf{H}$ 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.

3.2 Subsampling

For what we define as the standard stochastic GEE2 algorithm, let $U = (U_1, \dots, U_I)$, where each U_i corresponds to the indices of the outcomes in cluster i , with $|U_i| = n_i$. At each iteration ω , sample $s_i \sim \text{SRSWOR}(U_i, v_i)$, and concatenate $s = (s_1, \dots, s_I)$. That is, each cluster sample s_i is a simple random sample without replacement of v_i indices of the i th cluster. Then, perform the Newton-Raphson iteration with just the subsampled units. Notationally, this just replaces $\mathbf{H}_{(\omega)}$, $\mathbf{G}_{(\omega)}$ in Eq 4 with

$$\mathbf{H}_{(\omega)} = \sum_{i=1}^I \mathbf{D}_{i(\omega)}^T \mathbf{V}_{i(\omega)}^{-1} \mathbf{D}_{i(\omega)}, \quad \mathbf{G}_{(\omega)} = \sum_{i=1}^I \mathbf{D}_{i(\omega)}^T \mathbf{V}_{i(\omega)}^{-1} \mathbf{W}_{i(\omega)}^S \mathbf{E}_{i(\omega)} \quad (6)$$

where $\mathbf{W}_{i(\omega)}^S$ 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 $\mathbf{H}_{(\omega)}$, $\mathbf{G}_{(\omega)}$ are unbiased estimators for $\mathbf{H}_{(\omega)}$, $\mathbf{G}_{(\omega)}$, respectively. Hence, by the Robbins-Monro

conditions, the stochastic iterations $(\boldsymbol{\beta}_{(\omega)}, \boldsymbol{\alpha}_{(\omega)}) \rightarrow (\boldsymbol{\beta}, \boldsymbol{\alpha})$ almost surely as the number of iterations $\omega \rightarrow \infty$, where $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are the solutions to estimating equations specified in GEE2. Because Hessians are embedded within our procedure, we should select a sizeable subsample to ensure a reliable estimate of the Hessian; we recommend $v_i = 5$. We present full details in pseudo code for stochastic GEE2 in Algorithm 1 of Supplemental Materials 2.

3.3 Exploiting sparsity

Stochastic GEE2 addresses 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. Our proposal naturally alleviates this issue because stochasticity makes it very likely to “jump” off the path of divergence. Secondly, programming the expressions in Eq 6 only require storing a subset of the rows in $\mathbf{D}_{i(\omega)}$ or $\mathbf{E}_{i(\omega)}$ and columns of $\mathbf{V}_{i(\omega)}^{-1}$, hence greatly freeing up RAM on a computer.

One aspect that can be improved, depending on the structure of the working covariance matrix $\mathbf{V}_i$, is computational speed. For example, if we were to assume the off-diagonals $\text{Cov}(\mathbf{Y}_i, \mathcal{Z}(\mathbf{Y}_i)) = \mathbf{0}$, then iterations for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be separated, with the GEE2 portion Newton-Raphson iterations taking the form $\boldsymbol{\alpha}_{(\omega+1)} = \boldsymbol{\alpha}_{(\omega)} + \mathbf{H}_{\boldsymbol{\alpha}(\omega)}^{-1} \mathbf{G}_{\boldsymbol{\alpha}(\omega)}$. Then, if we take the GEE2 portion working covariance $V_{\alpha i} = \text{Var}(\mathcal{Z}(\mathbf{Y}_i))$ to be diagonal, we can show (see Supplemental Materials 3) that each iteration has complexity $\mathcal{O}(1)$ (constant time!), given that subset sizes v_i are not growing with respect to n_i .

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

Theorem: Let $\text{Cov}(\mathbf{Y}_i, \mathcal{Z}(\mathbf{Y}_i)) = \mathbf{0}$ and $v_i = \mathcal{O}(1)$ (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 Supplemental Materials 3. Table 1 expresses a clearer schematic of the theorem.

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 stochastic GEE2, while the GEE1 portion remains at $\mathcal{O}(\max_i n_i)$, the GEE2 portion now becomes $\mathcal{O}(1)$, and hence the overall complexity is $\mathcal{O}(\max_i n_i)$. Therefore, stochastic 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 effects are even more pronounced, cutting the complexity from $\mathcal{O}(\max_i n_i^2)$ to $\mathcal{O}(1)$.

3.4 Parallel Stochastic GEE2

While stochastic GEE2 allows faster computations in its iterative fitting procedure, each iteration is not as informative due to additional variation from subsampling. Hence, more iterations are needed to solve the estimating equations, although the additional time in running more iterations is far less significant than the computational savings per iteration. Nevertheless, in pursuit of a variant requiring fewer iterations, we propose the parallel stochastic GEE2 class of algorithms. The general technique of parallelization is expanded upon in Zinkevich et al. (2010). The basic idea is, after sufficiently enough iterations, the stochastic estimates will become unbiased and further iterations are meant to reduce variation from the stochastic nature of the algorithm. Instead, one can run K independent chains of stochastic 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 Supplemental Materials 2.

Stochastic GEE2 reduces the frequency of divergence, but generally not all of it; there remains a non-negligible probability that the algorithm may diverge. Parallelization 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 parallel stochastic GEE2, since the provided estimate would act as a better initialization. In a sense, parallelization is 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, parallel stochastic GEE2 offers an advantage in intrinsically incorporating information in its multistart search rather than truly random scattering.

4 Proposal II: IPW- & DR-GEE2

Our main objective is a robust and efficient method for estimating association parameters, from which we proceed with the GEE2 framework. The previous section addresses the computational difficulties with GEE2; in this section we propose estimators for settings with missing outcomes. Our proposed solutions to both computational challenges and informative missingness share a similar structure: the stochastic algorithms and missingness-adjusted schemes for GEE2 both utilize weighting, the former to form unbiased stochastic Hessians and gradients, and the latter to form unbiased estimating equations.

4.1 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 cannot 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 $\mathbb{P}(R_{ij} = 1 | Y_i, A_i, \mathbf{Z}_i, \mathbf{X}_i) = \mathbb{P}(R_{ij} = 1 | A_i, \mathbf{Z}_i, \mathbf{X}_i)$; that is, $R_{ij} \perp\!\!\!\perp Y_{ij} | A_i, \mathbf{Z}_i, \mathbf{X}_i$. To continue with valid inference, we assume that

$$\pi_{ij}^R \stackrel{\text{def}}{=} \mathbb{P}(R_{ij} = 1 | A_i, \mathbf{Z}_i, \mathbf{X}_i) > 0 \quad \text{and} \quad \eta_{ijj'}^R \stackrel{\text{def}}{=} \mathbb{P}(R_{ij}R_{ij'} = 1 | A_i, \mathbf{Z}_i, \mathbf{X}_i) > 0,$$

where π_{ij}^R and $\eta_{ijj'}^R$ are the first- and second-order propensity scores (PS) that characterize the individual and joint missingness processes; inverting these PS gives the inverse-probability weights (IPW). Define the IPW matrix as

$$\mathbf{W}_i^R = \text{diag} \left(\underbrace{\frac{R_{i1}}{\pi_{i1}^R(\boldsymbol{\beta}_R)}, \dots, \frac{R_{in_i}}{\pi_{in_i}^R(\boldsymbol{\beta}_R)}}_{\text{IPW1}}, \underbrace{\frac{R_{i1}R_{i2}}{\eta_{i12}^R(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)}, \dots, \frac{R_{i(n_i-1)}R_{in_i}}{\eta_{i(i-1)n_i}^R(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)}}_{\text{IPW2}} \right)$$

Within the IPW matrix, $\pi_{ij}^R(\boldsymbol{\beta}_R)$ is a model (parametrized by $\boldsymbol{\beta}_R$) for π_{ij}^R and $\eta_{ijj'}^R(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)$ is a model (parametrized by $\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R$) for $\eta_{ijj'}^R$; we will henceforth refer to these as the propensity score models (PSM). Since $\eta_{ijj'}^R$ is a function of $\pi_{ij}^R, \pi_{ij'}^R, \rho_{ijj'}^R$, it suffices to fit a model for $\rho_{ijj'}^R$. Note that $(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)$ are nuisance parameters: they are not parameters of interest, but are needed in estimating our target parameters $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$. To this end, we propose the inverse probability weighting second-order generalized estimating equations (IPW-GEE2) as

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^I \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{W}_i^R \mathbf{E}_i \\ \mathbf{0} &= \sum_{i=1}^I \mathbf{S}_i^R(A_i, \mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R) \end{aligned} \quad (7)$$

The second line in Eq 7 is used to estimate $(\boldsymbol{\beta}_R, \boldsymbol{\alpha}_R)$, while the first line is used to estimate $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$.

The IPW2 portion in $\mathbf{W}_i^R$ is derived by considering that the (j, j') th element of $\mathcal{Z}(\mathbf{Y}_i)$ is missing when either Y_{ij} or $Y_{ij'}$ is missing, which is exactly represented by the product of their missingness indicators $R_{ij}R_{ij'}$. We note that the propensity score model for $R_{ij}R_{ij'}$ must properly account for the correlation among missingness indicators. To the best of our knowledge, these estimating equations are 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.

4.2 AUG-GEE2

In the absence of missing data, Stephens et al. (2012) proposed the augmented (AUG) estimator to improve efficiency relative to the standard GEE1 by incorporating baseline covariates. The AUG-GEE is constructed by subtracting from GEE its orthogonal projection onto the linear space spanned by score functions corresponding to all smooth parametric models for the treatment assignment mechanism given covariates. Applying this projection to GEE2 in Eq 3, the AUG-GEE2 is given by

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^I [\mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{E}_{i'} + \boldsymbol{\zeta}_i] \\ \mathbf{0} &= \sum_{i=1}^I \mathbf{S}_i^Y(\mathbf{Z}_i^*, \mathbf{X}_i, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y) \end{aligned} \quad (8)$$

where

$$\begin{aligned} \mathbf{E}_i' &= \begin{pmatrix} \mathbf{Y}_i - \bar{\pi}_i(\boldsymbol{\beta}_Y) \\ \mathcal{G}(\mathbf{Y}_i) - \bar{\rho}_i^\dagger(\boldsymbol{\alpha}_Y) \end{pmatrix}, \mathbf{E}_i'' = \begin{pmatrix} \bar{\pi}_i(\boldsymbol{\beta}_Y) - \pi_i^*(\boldsymbol{\beta}_Y^*) \\ \bar{\rho}_i^\dagger(\boldsymbol{\alpha}_Y) - \rho_i^*(\boldsymbol{\alpha}_Y^*) \end{pmatrix}, \\ \boldsymbol{\zeta}_i &= \sum_{a=0}^1 p_A^a (1 - p_A)^{1-a} \mathbf{D}_i^\top (A = a) \mathbf{V}_i^{-1} (A = a) \mathbf{E}_i'' (A = a) \end{aligned}$$

where $\mathbf{D}_i(A = a)$, $\mathbf{V}_i^{-1}(A = a)$, $\mathbf{E}_i''(A = a)$ are values of $\mathbf{D}_i$, $\mathbf{V}_i^{-1}$, $\mathbf{E}_i''$ respectively, evaluated at $A = a$; $\bar{\pi}_{ij}$ is a model for π_{ij} and the components of $\bar{\rho}_i^\dagger$ are

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

akin to Eq 1, with models replacing each population quantity. The last set of equations in Eq 8 containing $\mathbf{S}_i^Y$ yields $\hat{\pi}_{ij}$ (converging to $\bar{\pi}_{ij}$) and $\hat{\rho}_{ijj'}$ (converging to $\bar{\rho}_{ijj'}$), collectively known as the outcome models (OM). In the absence of missing data, AUG-GEE2 estimators achieve semiparametric efficiency bound when the OM is correctly specified and remain consistent even when the OM is not correctly specified (Tsiatis, 2007). However, in the presence of rMAR, AUG-GEE2 based on complete-case only does not ensure consistent estimation, even when the OMs are correctly specified (more details are included in Supplemental Materials 4). We note that this AUG-GEE2 based on complete-case only is different from a special case of the DR-GEE2 (as described below in Section 4.3) by setting the weights to 1 for observed outcomes and to 0 for missing outcomes. The latter would enjoy the doubly-robust property and produce consistent estimates when outcome models are correctly specified.

4.3 DR-GEE2

Instead of subtracting the projection using the estimating equations given in Eq 3, we apply this projection on IPW-GEE2 in Eq 7 to also account for the missing data process. The

resulting form is thereby called the doubly-robust second-order generalized estimating equations (DR-GEE2):

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^I [\mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{W}_i^R \mathbf{E}_{i'} + \boldsymbol{\zeta}_i] \\ &\stackrel{\text{def}}{=} \sum_{i=1}^I \boldsymbol{\Phi}_i^Y(\mathbf{Z}_i^*, \mathbf{X}_i, \mathbf{R}_i, \boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y) \\ \mathbf{0} &= \sum_{i=1}^I \mathbf{S}_i^R(\mathbf{Z}_i^*, \mathbf{X}_i, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R), \quad \mathbf{0} = \sum_{i=1}^I \mathbf{S}_i^Y(\mathbf{Z}_i^*, \mathbf{X}_i, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y) \end{aligned} \quad (9)$$

If the OM is correctly specified, then under the rMAR assumption, $(\boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y)$ can be consistently estimated based on the complete-case data. DR-GEE2 is doubly robust in the sense that it is consistent and asymptotically normal for $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$ under correct specification of either the OM [i.e. $\bar{\pi}_{ij} = \pi_{ij}$ and $\bar{\rho}_{ijj'} = \rho_{ijj'}$] or PS [i.e. $\bar{\pi}_{ij}^R = \pi_{ij}^R$ and $\bar{\rho}_{ijj'}^R = \rho_{ijj'}^R$] (see proof in Supplemental Materials 4).

As a special case of DR-GEE2, consider the estimator obtained by setting the weights to 1 for observed outcomes and to 0 for missing outcomes. This estimator is consistent when the OM is correctly specified due to its doubly-robust property, but may lead to biased estimation when the OM is not correctly specified. As mentioned before, this estimator is different from the AUG-GEE2 described in Section 4.2.

4.4 Inference for IPW-, AUG-, and DR-GEE2

Let $\boldsymbol{\kappa}$ be a concatenation of all required parameters for any of the previously described estimation schemes in Sections 4.1 – 4.3. For example, $\boldsymbol{\kappa} = (\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*, \boldsymbol{\beta}_R, \boldsymbol{\alpha}_R, \boldsymbol{\beta}_Y, \boldsymbol{\alpha}_Y)$ for DR-GEE2 in Eq 9. Also, let $\Psi(\cdot)$ be a single, stacked summand from one of these estimating equations; for example, $\Psi = (\Phi^Y, \mathbf{S}^R, \mathbf{S}^Y)^\top$ for DR-GEE2. A direct application of the theory of M -estimators (Van der Vaart, 2000) yields that $\sqrt{I}(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma)$, where $\Sigma = \Gamma^{-1} (\Gamma^{-1})^\top$ is the sandwich estimator with “meat” $\Delta = \mathbb{E}[\Psi(\boldsymbol{\kappa}_0)\Psi(\boldsymbol{\kappa}_0)^\top]$ and “breads” $\Gamma = \mathbb{E}[\partial\Psi(\boldsymbol{\kappa}_0)/\partial\boldsymbol{\kappa}^\top]$ evaluated at the true parameter values $\boldsymbol{\kappa}_0$, from which we can extract components corresponding to $(\boldsymbol{\beta}_Y^*, \boldsymbol{\alpha}_Y^*)$, the parameters of interest. An estimator $\hat{\Sigma}$ can be obtained by replacing Δ with $\hat{\Delta} = \frac{1}{I} \sum_{i=1}^I \Psi_i(\hat{\boldsymbol{\kappa}})\Psi_i(\hat{\boldsymbol{\kappa}})^\top$ and Γ with $\hat{\Gamma} = \frac{1}{I} \sum_{i=1}^I \partial\Psi_i(\hat{\boldsymbol{\kappa}})/\partial\boldsymbol{\kappa}$.

4.5 Embedding IPW-, AUG-, and DR-GEE2 with Stochastic GEE2

Fitting the PSM and OM requires just the standard GEE2, so the stochastic analog is the same as defined in Section 3.2. We do, however, need to adjust the TM for IPW-, AUG-, and DR-GEE2.

Structurally speaking, the inclusion of subsampling matrix $\mathbf{W}^S$ in stochastic GEE2 is similar to the IPW matrix $\mathbf{W}^R$ in IPW-GEE2. Indeed, in fitting IPW-GEE2 with a stochastic variant, we simply adjust gradient and Hessian iterations in Eq 6 with $\mathbf{W}^S \mapsto \mathbf{W}^R \mathbf{W}^S$. There are two

possible candidates for the new $\mathbf{W}^S$. We could use the exact $\mathbf{W}^S$ scheme in the original stochastic GEE2 defined in Section 3.2. 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 $s_i \sim \text{SRSWOR}(U_i^{\text{obs}}, v_i)$, where $U^{\text{obs}} = \{U_1^{\text{obs}}, \dots, U_I^{\text{obs}}\}$ 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, iterations are performed with $\mathbf{W}_{\beta i(\omega)}^{S\text{obs}} = \frac{m_i}{v_i}[s_i]$ and $\mathbf{W}_{\alpha i(\omega)}^{S\text{obs}} = \frac{m_i(m_i - 1)}{v_i(v_i - 1)}[(s_i)_2]$, a weighted indicator matrix for a subsample of the non-missing outcomes. Details are presented in Algorithm 3 of Supplemental Materials 2.

For TM estimating equation portion of AUG-GEE2 in Eq 8, we have components $\mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{E}_i$, and ζ_i . Since AUG-GEE2 deals with the situation without missing data, the subsampling here is straightforward; we embed the same full data subsampling matrix $\mathbf{W}^S$ within both $\mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{E}_i$, and ζ_i and perform the stochastic iterations. However, forming stochastic DR-GEE2 is not so clear-cut. In Eq 9, the TM estimating equations consist of $\mathbf{D}_i^T \mathbf{V}_i \mathbf{W}_i^R \mathbf{E}_i$, and ζ_i . For the $\mathbf{D}_i^T \mathbf{V}_i \mathbf{W}_i^R \mathbf{E}_i$, component, we have missing outcomes, and therefore might seem prudent to use $\mathbf{W}^{S\text{obs}}$ for the reasons above. But, an analogous version for ζ_i will *not* yield unbiased estimators (β_Y^*, α_Y^*) , since $\mathbf{W}^{S\text{obs}}$ 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 $\mathbf{D}_i^T \mathbf{V}_i \mathbf{W}_i^R \mathbf{E}_i$, and ζ_i , with the former using observed data subsampling weights $\mathbf{W}^{S\text{obs}}$ and the latter using full data subsampling weights $\mathbf{W}^S$. Details are presented in Algorithm 4 in Supplemental Materials 2.

The stochastic subsampling procedure is flexible and can be applied to different models involved in IPW-, AUG-, or DR-GEE2. For example, if one encounters computational problems when fitting OM or PSM, but not TM, the stochastic variants need only be applied to the estimating equations in line 2 of Eq 9, but not line 1.

5 Simulation

We performed two sets of experiments. The first set explored the statistical properties of IPW-, AUG-, and DR-GEE2 under combinations of correctly specified / misspecified PSMs and correctly specified / misspecified OMs, all of which include the ICC estimates embedded in the working covariance structure in the GEE1 portion, and assuming $\text{Cov}(\mathbf{Y}_i, \mathcal{E}(\mathbf{Y}_i)) = 0$ and $\text{Var}(\mathcal{E}(\mathbf{Y}_i)) = \mathbf{I}$. We included analogous estimates from a parametric mixed effects model, GEE1 with independence working covariance structure, the conditional-GEE method of Lipsitz and Fitzmaurice (1996), the Gaussian-GEE method of Lipsitz et al. (2000), and the pseudo-likelihood method of Parzen (2009) for comparison. In the second set of simulations, we compared the algorithmic properties (convergence & run-time) of stochastic DR-GEE2 and standard DR-GEE2 under various cluster size / number of cluster combinations. For stochastic DR-GEE2, the estimating equations for the PSM, OM, and TM were all solved stochastically.

The following two data generation processes for binary data Y_{ij} (or R_{ij}) were used:

$$\begin{aligned}
 \text{Parzen's method} \quad & \begin{cases} \text{logit}(\pi_{ij}) &= (\beta_{0Y} + \beta_{0AY}A_i) + (\beta_{ZY} + \beta_{ZAY}A_i)^T \mathbf{Z}_i + (\beta_{XY} + \beta_{XAY}A_i)^T \mathbf{X}_{ij} \\ \text{atanh}(\rho_i) &= (\alpha_{0Y} + \alpha_{0AY}A_i) + (\alpha_{ZY} + \alpha_{ZAY}A_i)^T \mathbf{Z}_i \\ (\mathfrak{Q}_i, \mathfrak{U}_i) &= \left(-\sqrt{\frac{\min(\pi_i)}{1 - \min(\pi_i)}}, \sqrt{\frac{1 - \max(\pi_i)}{\max(\pi_i)}} \right) \\ (\delta_i, \epsilon_i) &= \left(\frac{\mathfrak{U}_i(-\mathfrak{U}_i\mathfrak{Q}_i - \rho_i)}{(\mathfrak{U}_i - \mathfrak{Q}_i)\rho_i}, \frac{-\mathfrak{Q}_i(-\mathfrak{U}_i\mathfrak{Q}_i - \rho_i)}{(\mathfrak{U}_i - \mathfrak{Q}_i)\rho_i} \right) \\ \xi_i | A_i, \mathbf{Z}_i &\sim (\mathfrak{U}_i - \mathfrak{Q}_i) \text{Beta}(\delta_i, \epsilon_i) + \mathfrak{Q}_i \\ Y_{ij} | A_i, \mathbf{Z}_i, \mathbf{X}_i, \xi_i &\sim \text{Bernoulli}(\pi_{ij} + \xi_i \sqrt{\pi_{ij}(1 - \pi_{ij})}) \end{cases} \\
 \\
 \text{Random intercept} \quad & \begin{cases} \text{logit}(\pi_{ij}) &= (\beta_{0Y} + \beta_{0AY}A_i) + (\beta_{ZY} + \beta_{ZAY}A_i)^T \mathbf{Z}_i + (\beta_{XY} + \beta_{XAY}A_i)^T \mathbf{X}_{ij} \\ \xi_i | A_i &\sim N\left(0, \left(\frac{1}{3} + \frac{1}{2}A_i\right)^2\right) \\ \text{logit}(p_{ij}) &= \xi_i + \text{logit}(\pi_{ij}) \\ Y_{ij} | A_i, \mathbf{Z}_i, \mathbf{X}_i, \xi_i &\sim \text{Bernoulli}(p_{ij}) \end{cases}
 \end{aligned}$$

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 $-\mathfrak{U}_i\mathfrak{Q}_i - \rho_i \geq 0$ for all i . The random-intercept method is the traditional approach to induce correlation among observations in a cluster. With a normal random intercept, the marginal probability of success

$$\begin{aligned}
 \mathbb{P}(Y_{ij} = 1 | A_i, \mathbf{Z}_i, \mathbf{X}_i) &= \int \mathbb{P}(Y_{ij} = 1 | \xi_i, A_i, \mathbf{Z}_i, \mathbf{X}_i) dP(\xi_i) = \\
 &= \int \frac{e^{\xi_i + L(\beta; A_i, \mathbf{Z}_i, \mathbf{X}_i)}}{1 + e^{\xi_i + L(\beta; A_i, \mathbf{Z}_i, \mathbf{X}_i)}} dP(\xi_i)
 \end{aligned} \tag{10}$$

where $L(\beta; A_i, \mathbf{Z}_i, \mathbf{X}_i)$ is the linear function, is not of the logistic form and does 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 varies within-cluster. Hence, it is ill-suited for simulation of equicorrelated binary data. We used Parzen's method to generate the ideal case of equicorrelated outcomes, while we used the random-intercept method to generate correlated outcomes with a non-exchangeable correlation structure.

5.1 Statistical Properties of IPW-, AUG-, & 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 $\{a, a + 1, \dots, b - 1, b\}$. To evaluate the asymptotic properties of GEE2, we set the number of clusters to $I = 2000$ with cluster sizes $n_i \sim \mathcal{U}\{80, 140\}$ so that we have average cluster size $\mathbb{E}[n_i] = 110$. We choose settings with a large number of clusters

to evaluate the asymptotic properties and to avoid computational issues that will be explored in Section 5.2. We generated $A_i \sim \text{Bernoulli}(1/2)$ and chose $\mathbf{X}_{ij} \in \mathbb{R}^3$ and $\mathbf{Z}_i \in \mathbb{R}$. Details regarding generation of $\mathbf{X}_{ij}$, $\mathbf{Z}_i$ and choice of coefficients for Y_{ij} are presented in Table 2. We also generated R_{ij} with these same covariates and coefficients for simplicity.

The values in Table 2 were carefully chosen to guarantee $-\mathcal{U}_i \mathcal{Q}_i - \rho_i \geq 0$ 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 , were 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 $\text{Corr}(Y_{ij}, Y_{ij'} | A_i, \mathbf{Z}_i, \mathbf{X}_i)$ were in the range $[0.333, 0.738]$ and $[0.022, 0.134]$, respectively. Since R_{ij} 's were generated from models with the same coefficients, the probabilities of being observed from these two methods are also in the range $[0.324, 0.733]$ and $[0.333, 0.738]$, respectively. The true treatment coefficients (β_Y^*, α_Y^*) in the canonical TM were calculated by numerically integrating out all other covariates except for A_i in π_{ij} and $\rho_{ijj'}^\dagger$:

$$\begin{aligned} \text{expit}(\beta_{0Y}^* + \beta_{AY}^* A_i) &= \int_{\mathbb{R}^4} \pi_{ij} dP(\mathbf{x}_{ij}) dP(\mathbf{z}_i) \\ \tanh(\alpha_{0Y}^* + \alpha_{AY}^* A_i) &= \int_{\mathbb{R}^7} \rho_{ijj'}^\dagger dP(\mathbf{x}_{ij}) dP(\mathbf{x}_{ij'}) dP(\mathbf{z}_i) \end{aligned} \quad (11)$$

Using Parzen's method, we obtained the values $(\beta_Y^*, \alpha_Y^*) = (0.1413, 0.1808, 0.1238, 0.0755)$, and using the random-intercept method, we obtained $(\beta_Y^*, \alpha_Y^*) = (0.1378, 0.1429, 0.0307, 0.1032)$.

Figure 1 displays the average of parameter estimates and their respective 95% Monte Carlo confidence intervals $\bar{\theta} \pm 1.96 \cdot \widehat{\text{SE}}(\bar{\theta})$ over $\mathcal{R} = 1000$ simulation experiments. To assess the accuracy of our derived sandwich standard errors, we present two types of Monte Carlo confidence intervals based on $\widehat{\text{SE}}(\bar{\theta})$ calculated using (i) the average of the calculated sandwich standard errors divided by $\sqrt{\mathcal{R}}$ and (ii) the empirical standard error over the $\mathcal{R}$ simulations (i.e. empirical standard deviation divided by $\sqrt{\mathcal{R}}$). Conditional-GEE (Lipsitz and Fitzmaurice, 1996), Gaussian-GEE (Lipsitz et al., 2000), and the pseudo-likelihood methods (Parzen et al., 2007) were fitted directly to the complete-case data with the canonical TM (Eq 2), while the mixed effects model was fitted to the complete-case data using

$$\text{logit}\{\mathbb{P}(Y_{ij} = 1 | A_i, \xi_i)\} = \tilde{\beta}_0 + \tilde{\beta}_A A_i + \xi_i \quad \text{with} \quad \xi_i | A_i \sim N(0, \tilde{\sigma}_{A_i}^2). \quad (12)$$

Estimates of $(\beta_{0Y}^*, \beta_{AY}^*, \alpha_{0Y}^*, \alpha_{AY}^*)$ were obtained from $(\tilde{\beta}_0, \tilde{\beta}_A, \tilde{\sigma}_0^2, \tilde{\sigma}_1^2)$ through marginalizing out the random effects (see Eqs 10). We display Monte Carlo intervals based on the average of estimated analytical standard errors for all except the Gaussian-GEE and mixed effects scenarios. Analytical standard error estimators for $\hat{\alpha}_{0Y}^*$ and $\hat{\alpha}_{AY}^*$ from Gaussian-GEE were

not provided in Lipsitz et al. (2000) nor are they immediately derivable. In mixed effects model settings, estimation of standard errors of $\hat{\alpha}_{0Y}^*$, $\hat{\alpha}_{AY}^*$ requires standard error estimates of $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, for which methods are less developed (Bates, 2010; McCulloch and Searle, 2001; Wu et al., 2012). Mixed effects models naturally handle MAR if the true generation process follows the form in Eq 12. However, both Parzen's and the random-intercept data generation processes do not: data generated using Parzen's method do not follow a traditional mixed effects model, and data generated using the random-intercept method include additional covariates.

For the IPW-GEE2 estimates, we used $\mathcal{G}_1(\mathbf{R})$ IPW and $\mathcal{G}_2(\mathbf{R})$ IPW to distinguish IPW-GEE2 under the settings where the PSMs do not and do account for the correlation among the missingness indicators, respectively, as discussed in Section 4.1. For GEE1 methods, there is no model for correlated missingness, and therefore intervals were omitted. The correctly-specified OM and PSM were

$$\begin{aligned}\text{logit}(\pi_{ij}) &= (\beta_{0Y} + \beta_{0AY}A_i) + (\beta_{ZY} + \beta_{ZAY}A_i)^\top \mathbf{Z}_i + (\beta_{XY} + \beta_{XAY}A_i)^\top \mathbf{X}_{ij} \\ \text{atanh}(\rho_{ijj'}) &= (\alpha_{0Y} + \alpha_{0AY}A_i) + (\alpha_{ZY} + \alpha_{ZAY}A_i)^\top \mathbf{Z}_i\end{aligned}\quad (13)$$

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

$$\begin{aligned}\text{logit}(\pi_{ij}) &= \beta_{0Y} + \beta_{AY}A_i + \beta_{ZY}^\top \mathbf{Z}_i + \beta_{XY}^\top \mathbf{X}_{ij} \\ \text{atanh}(\rho_{ijj'}) &= \alpha_{0Y} + \alpha_{AY}A_i + \alpha_{ZY}^\top \mathbf{Z}_i\end{aligned}\quad (14)$$

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

As seen from Figure 1, complete-case analysis (from GEE1, GEE2, conditional-GEE, pseudolikelihood, and mixed effects) led to severe bias in estimating all parameters, with the exception of Gaussian-GEE, where the intervals for association parameters captured the truth. We note that these methods are expected to account for missingness when confounding covariates are included in the model, but the focus here is on marginal intervention effects and hence inclusion of additional covariates would in general target different treatment effect parameters for models with non-identity link functions (marginal vs. conditional). When PSM & OM were correctly specified, $\mathcal{G}_1(\mathbf{R})$ IPW-, $\mathcal{G}_2(\mathbf{R})$ IPW-, and DR-GEE2 all provided consistent estimates for the mean parameters β_{0Y}^* and β_{AY}^* , but $\mathcal{G}_1(\mathbf{R})$ IPW-GEE2 failed to estimate association parameters α_{0Y}^* and α_{AY}^* consistently. When the PSM was misspecified but the OM was correctly specified, only DR-GEE2 and DR-GEE1 produced consistent estimates of all parameters. AUG-GEE1 and AUG-GEE2 based on complete data only led to biased estimates. On the other hand, when the OM was misspecified, $\mathcal{G}_2(\mathbf{R})$ IPW- and DR-GEE2 produced consistent estimates of all parameters, and the bias from AUG-GEE1 and AUG-GEE2 for the mean parameters appeared to be more pronounced. In all scenarios, sandwich variance estimates were close to the true sampling variance. We also observed that the standard error estimates for $\hat{\beta}_{0Y}^*$ and $\hat{\beta}_{AY}^*$

obtained from DR-GEE1 with independence covariance structure were approximately equal to those obtained from DR-GEE2.

Figure 2 presents analogous results for settings where outcomes were generated using the random-intercept method and missingness indicators were generated using Parzen's method. As before, the correctly specified and misspecified PSMs were Eqs 13 and 14, respectively. Since data generated under a random intercept logistic model do not follow a marginal logistic model, the OM was misspecified. Nevertheless, we reached similar conclusions regarding the performance of estimators as before. It is interesting to note that DR-GEE2 performed well even when both PSM and OM were misspecified. This may be due to the fact that the misspecification of the OM reflected a deviation from the true model only through an omission of the random intercept; this deviation could be modest, especially if the variance of random effects was small.

5.2 Algorithmic Characteristic of Standard vs Stochastic DR-GEE2

Having established the consistency of the standard DR-GEE2, in our second set of experiments we compared stochastic and parallel stochastic DR-GEE2. Data were generated in the same way as before, where both outcomes and missingness indicators were generated using Parzen's method. When fitting DR-GEE2, both PSM and OM were correctly specified. We varied the number of clusters I and cluster sizes n_i , and considered the following three scenarios: $(I, \mathbb{E}[n_i]) = (30, 30), (300, 30), (30, 300)$. Figures 3 and 4 present the statistical and algorithmic results respectively. For $\mathbb{E}[n_i] = 30$, we set the subsample sizes $\mathbb{E}[v_i] = 9$ and for $\mathbb{E}[n_i] = 300$, we set $\mathbb{E}[v_i] = 45$; all stochastic scenarios used learning rates $\gamma(\omega) = (\omega + 1)^{-1}$, where ω indexes the iteration step. For the comparison of running longer iteration chains versus averaging shorter chains, parallel stochastic DR-GEE2 averaged 20 chains, each with one-third of the number of iterations per chain used in stochastic DR-GEE2.

As seen in Figure 3, all algorithms yield consistent estimates for the mean parameters (β_0^* and β_1^*). Overall, the parameter and standard error estimates were similar across standard, stochastic, and parallel stochastic DR-GEE2. When the number of clusters $I = 30$, estimates for the association parameters (α_0^* and α_1^*) can be biased. Poor performance of GEEs when the number of clusters is small has been noted previously (Mancl and DeRouen, 2001; Morel et al., 2003). For estimating association parameters, the number of clusters needs to be sufficiently large.

Figure 4 presents the run times and error rates associated with standard, stochastic, and parallel stochastic algorithms when fitting DR-GEE2. There were substantial reductions in error rates and run times when stochastic algorithms were used. For the OM, PSM, and TM, stochastic DR-GEE2 provided up to 80% reduction in returned errors (i.e. divergence, large condition numbers of Hessians) and approximately 90% reduction in run times. Parallel stochastic DR-GEE2, as expected, cut computation times by nearly one-third compared to stochastic DR-GEE2 and eliminated all convergence errors. For the purpose of a direct comparison, we did not include the time it took the server to set up the parallel cores for

computation. With large datasets, this set-up time would be negligible compared to the overall computing time of days or weeks. In the current settings, this set up time would offset the advantage of parallel stochastic algorithms in terms of run times; nonetheless, the advantage of reducing convergence error rates remains.

We also fitted a complete-case TM using the `geese` command from R's `geepack` package. While `geese` performed well in the (30, 30) and (300, 30) cases, our stochastic algorithms were associated with faster run times and few errors in the (30,300) cases. Granted, the comparisons were not the most commensurate: `geese` performed all calculations in the C programming language and wrapped the results into R, while our implementation was entirely in R. Our implementation also included additional time for incorporating the IPW or AUG portions.

6 Application to Sanitation Data

Guiteras et al. (2015) investigated the efficacy of alternative policies to encourage the use of hygienic latrines in developing countries. Communities in rural Bangladesh were assigned to different marketing interventions. Here we focused on comparisons between a supply-side market intervention and a control group. The dataset (accessible at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GJDUTV>) contains 4837 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 cluster-level covariates (village population Z_1 , # of doctors Z_2 , % landless Z_3 , % almost landless Z_4 , % access electricity Z_5). Analyses were performed using data from 4768 subjects with complete covariate information.

The overall missingness of the outcome, whether or not hygienic latrines was used, was 3.4%. Results based on a mixed effects model suggested that the supply-side market intervention alone did not increase hygienic latrine ownership (+0.3 percentage points, p -value = 0.90) (Guiteras et al., 2015). We reanalyzed this dataset with GEE2 approaches assuming outcomes were rMAR; covariates for the PS and outcome models were chosen using a backwards stepwise procedure based on the AIC. The resulting models are

$$\begin{aligned} \text{logit}(\pi_{ij}^R) &= \beta_{0R} + \beta_{AR}A_i + \sum_{k \in \{2, 3, 5, 6, 7, 8, 10\}} \beta_{XR}^{(k)}X_{ijk} + \sum_{k \in \{1, 2, 3, 4\}} \beta_{ZR}^{(k)}Z_{ik} + \\ &A_i \sum_{k \in \{5, 6, 8\}} \beta_{AXR}^{(k)}X_k + A_i \sum_{k \in \{2, 3, 4\}} \beta_{AZR}^{(k)}Z_{ik} \\ \text{atanh}(\rho_i^R) &= \alpha_{0R} + \alpha_{AR}A_i + \sum_{k \in \{1, 2, 3, 4\}} \alpha_{ZR}^{(k)}Z_{ik} + A_i \sum_{k \in \{2, 3, 4\}} \alpha_{AZR}^{(k)}Z_{ik} \end{aligned}$$

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \beta_{0Y} + \beta_{AY}A_i + \sum_{k \in \{1, 2, 3, 4, 5, 8, 9, 10\}} \beta_{XY}^{(k)}X_{ijk} + \sum_{k \in \{1, 2, 3, 4, 5\}} \beta_{ZY}^{(k)}Z_{ik} + \\ &A_i \sum_{k \in \{1, 3, 8\}} \beta_{AXY}^{(k)}X_k + A_i \beta_{AZY}^{(5)}Z_{i5} \\ \text{atanh}(\rho_i) &= \alpha_{0Y} + \alpha_{AY}A_i + \sum_{k \in \{1, 2, 3, 4, 5\}} \alpha_{ZY}^{(k)}Z_{ik} + A_i \alpha_{AZY}^{(5)}Z_{i5} \end{aligned}$$

To better illustrate the use of our methods for settings where missingness is high, we conducted two analyses: one based on the original data, and another after inducing additional missingness. The missingness indicators were generated under Parzen's method with the following coefficients:

β_{0R}	$= 0.7,$	β_{XR}^T	$= (0, 0.3, 0.003, 0, 0.3, 0.3, 0.3, 0.3, 0, 0.3),$
β_{ZR}^T	$= (-0.00048, -0.0014, -0.407, -0.555),$	β_{AR}	$= -0.7,$
β_{AXR}^T	$= (0, 0, 0, 0, -0.6, -0.6, 0, -0.6, 0, 0),$	β_{AZR}^T	$= (0, 0.022, 0.904, 1.11),$
α_{0R}	$= 0.35,$	α_{ZR}^T	$= (0.0000645, 0.000190, 0.0543, 0.074),$
α_{AR}	$= -0.10,$	α_{AZR}^T	$= (0, -0.00291, -0.120, -0.148).$

Covariates not selected in the stepwise procedure were assigned coefficients 0. The values of nonzero entries of β_R were chosen so that $|\beta_{R,p}C_{ijp}| \leq 0.7$, where $\beta_{R,p}$ is the p th element of β_R , and C_{ijp} is the p th element of $(1, A_i, \mathbf{X}_{ij}, \mathbf{Z}_i)$. This bound ensures that the missingness probabilities do not wildly fluctuate to 0 or 1. Finally, the values of α_R were chosen to ensure a sizable correlation among the missingness indicators (~ 0.30) and to satisfy the constraints of Parzen's method. Under this scheme, the overall missingness was 26%.

For both analyses based on the original data and the data with induced missingness, we had difficulty obtaining convergent fits for the OM and PSM, and therefore applied stochastic GEE2 for these models. The estimates of these nuisance parameters were then used in DR-GEE2 to estimate the TM parameters in Eq 2. We had no difficulty fitting the TM, and therefore performed the fitting deterministically. The corresponding GEE2 analyses are included in Figure 5. We also ran the TM stochastically and the results were similar.

Based on the original data, DR-GEE2 yielded a significant intervention effect (OR = 1.58, p -value < 0.01), while results from complete-case GEE2 and IPW-GEE2's suggested no significant intervention effect (p -value > 0.15). This drastic change in the significance of the intervention effect is peculiar, especially given the low proportion of missingness. AUG-GEE2 yielded similar estimates as DR-GEE2 did, which suggests that the augmentation portion heavily influenced the resulting estimates. This implies that, while the sanitation data were obtained from a randomized study, there existed substantial residual imbalance in covariates that were associated with outcome. In such settings, inverse-probability of missingness weighting would not adjust this residual confounding. Methods commonly used

for observational studies such as inverse-probability of receiving treatment weighting (Rosenbaum and Rubin, 1983) provide one way to adjust for this; alternatively, augmenting the estimating equations with the outcome model serves the same purpose. The combination of both yields doubly robust estimators for observational studies (Bang and Robins, 2005). For the association parameters, inferences were the same using all methods; the ICC within the control arm was statistically significantly different from 0 (p -value < 0.01 for $H_0: \alpha_0^* = 0$ vs $H_1: \alpha_0^* \neq 0$), while the intervention-specific ICC was not (p -value > 0.60 for $H_0: \alpha_1^* = 0$ vs $H_1: \alpha_1^* \neq 0$).

Using the induced-missingness dataset, inference for the intervention effect was similar to that for the original dataset: complete-case GEE2 suggested no significant intervention effect (p -value = 0.34). But there was a significant difference in intervention-specific ICCs (p -value = 0.046). Using the IPW-GEE2 methods, the magnitude of intervention effect estimates increased but remained non-significant (p -value > 0.1 for both), and the differences in ICCs between the two arms decreased and became non-significant (p -value > 0.15 for both). Compared to the IPW methods, both AUG-GEE2 and DR-GEE2 led to a further increased intervention effect (p -value < 0.01) and a further decreased difference in intervention-specific ICCs (p -value > 0.15).

7 Discussion

In this paper, we propose 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 stochastic 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 their 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 work.

We also proposed DR-GEE2 for estimating the marginal treatment effect and treatment-specific ICCs in cluster randomized trials in the presence of informative missing outcome data. We note that, in general, GEE2 is less robust than GEE1 due to the need to correctly specify the second moments. In some settings, for example, a binary outcome with a focus on estimating treatment-specific ICCs, the GEE2 portion represents a saturated model and is correctly specified.

Our estimators are most useful in the settings where estimation of ICCs is the focus. If the interest is solely on marginal treatment effect, 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 2 and are regular asymptotically linear (RAL) must be of the form $\mathbf{0} = \sum_{i=1}^I \mathbf{h}(A_i) \mathbf{E}_i$.

The choice of index function $\mathbf{h}(A_i) = \mathbf{D}_i^T \mathbf{V}_i^{-1}$, which reduces back to GEE2, results in the efficient score for the canonical TM, hence leading to the minimum asymptotic variance RAL estimator for (β_Y^*, α_Y^*) (Chamberlain, 1986). However, in the case of IPW-GEE2 or DR-GEE2, this choice is no longer optimal and the actual $\mathbf{h}_{\text{opt}}(A_i)$ 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 $\mathbf{h}_{\text{opt}}(A_i)$ are modest and very sensitive to the correct specification of all components that comprise $\mathbf{h}_{\text{opt}}(A_i)$, which in practice is nearly impossible to achieve. Our simulation studies in Section 5 also provide corroborative evidence supporting the use of an independence covariance structure when estimating the first-order effects.

Although our work was motivated by the need for better estimation of ICCs in 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 and variance functions for the corresponding likelihood. When outcomes within clusters are not equicorrelated, our ICC estimators marginalize out factors which contribute to the non-exchangeable structure and return an estimate which can be construed as an “average” correlation.

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge NIH grants T32 ES007142, R37 AI51164 and R01 AI136947.

References

- Anderson D, and Aitkin M (1988), “Variance component models with binary response: Interviewer variability,” *JRSS B*, 47, 203–210.
- Bang H, and Robins JM (2005), “Doubly robust estimation in missing data and causal inference models,” *Biometrics*, 61(4), 962–973. [PubMed: 16401269]
- Bates D (2010), *Mixed-effects modeling with R*, New York: Springer.

- Blum JR (1954), "Multidimensional stochastic approximation methods," *The Annals of Mathematical Statistics*, pp. 737–744.
- Bottou L (2012), "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*, New York: Springer, pp. 421–436.
- Braun TM, and Feng Z (2001), "Optimal permutation tests for the analysis of group randomized trials," *Journal of the American Statistical Association*, 96(456), 1424–1432.
- Byrd RH, Hansen SL, Nocedal J, and Singer Y (2016), "A stochastic quasi-Newton method for large-scale optimization," *SIAM Journal on Optimization*, 26(2), 1008–1031.
- Carey V, Zeger SL, and Diggle P (1993), "Modelling multivariate binary data with alternating logistic regressions," *Biometrika*, 80(3), 517–526.
- Carnegie NB, Wang R, and De Gruttola V (2016), "Estimation of the overall treatment effect in the presence of interference in cluster-randomized trials of infectious disease prevention," *Epidemiologic Methods*, 5, 57–68.
- Chamberlain G (1986), "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32, 189–218.
- Clmenon S, Bertail P, Chautru E, and Papa G (2015), "Survey schemes for stochastic gradient descent with applications to M-estimation," arXiv preprint arXiv:1501.02218,.
- Crespi CM, Wong WK, and Mishra SI (2009), "Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster-randomized trials," *Statistics in Medicine*, 285, 814–827. [PubMed: 19109804]
- Donner A, and Klar N (2000), *Design and analysis of cluster randomization trials in health research*, Vol. 1, New York: Wiley.
- Duchi J, Hazan E, and Singer Y (2011), "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, pp. 2121–2159.
- Fitzmaurice GM (1995), "A caveat concerning independence estimating equations with multivariate binary data," *Biometrics*, pp. 309–317. [PubMed: 7766784]
- Gaolathe T, Wirth KE, Holme MP, Makhema J, Moyo S, Chakalisa U, Yankinda EK, Lei Q, Mmalane M, Novitsky V et al. (2016), "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. [PubMed: 27126489]
- Guiteras R, Levinsohn J, and Mobarak AM (2015), "Encouraging Sanitation Investment in the Developing World: A Cluster-Randomized Trial,," *Science*, 348(6237), 903–906. [PubMed: 25883316]
- Hall DB, and Severini TA (1998), "Extended generalized estimating equations for clustered data," *Journal of the American Statistical Association*, 93(444), 1365–1375.
- Hayes RJ, and Bennett S (1999), "Simple sample size calculation for cluster-randomized trials," *International Journal of Epidemiology*, 282, 319–326. [PubMed: 10342698]
- Hayes R, and Moulton L (2009), *Cluster randomised trials*, Boca Raton: Chapman & Hall/CRC.
- Klar N, and Donner A (2001), "Current and future challenges in the design and analysis of cluster randomization trials," *Statistics in Medicine*, 2024, 3729–3740. [PubMed: 11782029]
- Laird NM, and Ware JH (1982), "Random-effects models for longitudinal data," *Biometrics*, pp. 963–974. [PubMed: 7168798]
- Le Cessie S, and Van Houwelingen J (1994), "Logistic regression for correlated binary data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1), 95–108.
- Leckie G, French R, Charlton C, and Browne W (2014), "Modeling heterogeneous variance-covariance components in two-level models," *Journal of Educational and Behavioral Statistics*, 39(5), 307–332.
- Liang KY, and Zeger SL (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 731, 13–22.
- Liang KY, and Zeger SL (1992), "Multivariate regression analyses for categorical data,," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 3–40.
- Lipsitz SR, and Fitzmaurice GM (1996), "Estimating equations for measures of association between repeated binary responses," *Biometrics*, pp. 903–912. [PubMed: 8805760]

- Lipsitz SR, Molenberghs G, Fitzmaurice GM, and Ibrahim J (2000), "GEE with Gaussian estimation of the correlations when data are incomplete," *Biometrics*, 56(2), 528–536. [PubMed: 10877313]
- Mancl LA, and DeRouen TA (2001), "A covariance estimator for GEE with improved small-sample properties," *Biometrics*, 57(1), 126–134. [PubMed: 11252587]
- McCulloch C, and Searle S (2001), *Generalized, linear, and mixed models*, New York: John Wiley & Sons.
- McDonald BW (1993), "Estimating logistic regression parameters for bivariate binary data," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 391–397.
- Morel JG, Bokossa M, and Neerchal NK (2003), "Small sample correction for the variance of GEE estimators," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(4), 395–409.
- Nesterov Y (1983), "A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$," *Doklady ANSSSR* (translated as *Soviet.Math.Docl.*), 269, 543–547.
- Paik MC (1992), "Parametric variance function estimation for nonnormal repeated measurement data," *Biometrics*, pp. 19–30. [PubMed: 1581484]
- Parzen M (2009), "Random effects model for simulating clustered binary data," unpublished, .
- Parzen M, Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Troxel A, and Molenberghs G (2007), "Pseudo-likelihood Methods for the Analysis of Longitudinal Binary Data Subject to Nonignorable Non-monotone Missingness," *Journal of Data Science*, 5, 1–21.
- Prague M, Wang R, Stephens A, Tchetgen Tchetgen E, and DeGruttola V (2016), "Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes," *Biometrics*, 72(4), 1066–1077. [PubMed: 27060877]
- Robbins H, and Monro S (1951), "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407.
- Robins JM, Rotnitzky A, and Zhao LP (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866.
- Rosenbaum PR, and Rubin DB (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.
- Rubin DB (1976), "Inference and missing data," *Journal of the American Statistical Association*, 63(3), 581–592.
- Stephens AJ, Tchetgen EJT, and De Gruttola V (2012), "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. [PubMed: 22359361]
- Stephens AJ, Tchetgen Tchetgen EJ, and DeGruttola VD (2014), "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. [PubMed: 24566369]
- Stratelli R, Laird N, and Ware J (1984), "Random-effects models for serial observations with binary response," *Biometrics*, 40, 961–971. [PubMed: 6534418]
- Tsiatis A (2007), *Semiparametric theory and missing data*, New York: Springer Science & Business Media.
- Ugray Z, Lasdon L, Plummer JC, Glover F, Kelly J, and Marti R (2007), "Scatter Search and Local NLP Solvers: A Multistart Framework for Global Optimization," *INFORMS Journal on Computing*, 19(3), 328–340.
- Van der Laan MJ, and Robins JM (2003), *Unified methods for censored longitudinal data and causality*, New York: Springer Science & Business Media.
- Van der Vaart AW (2000), *Asymptotic statistics*, Vol. 3, Cambridge: Cambridge University Press.
- Wang R, and De Gruttola V (2017), "The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials," *Statistics in medicine*, 36(18), 2831–2843. [PubMed: 28464567]
- Wang R, Goyal R, Lei Q, Essex M, and De Gruttola V (2014), "Sample size considerations in the design of cluster randomized trials of combination HIV prevention," *Clinical trials*, 11(3), 309–318. [PubMed: 24651566]

- West BT, Conrad FG, Kreuter F, and Mittereder F (2018), “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.
- Wu S, Crespi CM, and Wong WK (2012), “Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials,” *Contemporary Clinical Trials*, 335, 869–880. [PubMed: 22627076]
- Zeger SL (1988), “A regression model for time series of counts,” *Biometrika*, 75(4), 621–629.
- Zeger SL, Liang KY, and Albert PS (1988), “Models for longitudinal data: a generalized estimating equation approach,” *Biometrics*, pp. 1049–1060. [PubMed: 3233245]
- Zeiler MD (2012), “ADADELTA: an adaptive learning rate method,” *arXiv preprint, arXiv*, 1212.5701
- Zhao LP, and Prentice RL (1990), “Correlated binary regression using a quadratic exponential model,” *Biometrika*, 773, 642–648.
- Ziegler A, Kastner C, and Blettner M (1998), “The Generalised Estimating Equations: An Annotated Bibliography,” *Biometrical Journal*, 402, 115–139.
- Ziegler A, Kastner C, and Blettner M (2000), “Familial associations of lipid profiles: A generalised estimating equations approach,” *Statistics in Medicine*, 1924, 3345–3357. [PubMed: 11122500]
- Zinkevich M, Weimer M, Li L, and Smola AJ (2010), Parallelized stochastic gradient descent,, in *Advances in Neural Information Processing Systems*, pp. 2595–2603.

Fig. 1.

Forest plots for intervals $\bar{\theta} \pm 1.96 \cdot \widehat{SE}(\bar{\theta})$, where $\bar{\theta}$ is the sample average of the $\mathcal{R} = 1000$ replicate simulations. Both R_{ij} 's and Y_{ij} 's are simulated using Parzen's method. Dotted vertical lines represent true parameter values.

Fig. 2.

Forest plots for intervals $\bar{\theta} \pm 1.96 \cdot \widehat{SE}(\bar{\theta})$, where $\bar{\theta}$ was the sample average of the $\mathcal{R} = 1000$ simulation experiments. R_{ij} was simulated using the Parzen's method and Y_{ij} was simulated using the random-intercept method.

Fig. 3.

Forest plots for standard, stochastic, and parallel stochastic DR-GEE2 with intervals $\bar{\theta} \pm 1.96 \cdot \widehat{SE}(\bar{\theta})$, where $\bar{\theta}$ is the sample average of θ across the $\mathcal{R} = 1000$ replicate simulations.

Fig. 4. 25th-50th-75th percentile boxplots of runtimes for standard, stochastic, and parallel stochastic DR-GEE2. $\mathcal{R} = 1000$ replicate simulations. Boxplots 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.

Fig. 5.

Effects of the supply side-market ($A=1$) vs. control ($A=0$) 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 correlation among missingness indicators), 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

Table 1

Time complexities for stochastic GEE2 under various working covariance structures.

	Standard		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 2

Parameter values in the data generation process

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