Published in final edited form as:

Stat Med. 2010 March 15; 29(6): 659–670. doi:10.1002/sim.3853.

A U-statistics based Approach for Modeling Cronbach Coefficient Alpha within a Longitudinal Data Setting

Y. Ma¹, A. Gonzalez Della Valle², H. Zhang³, and X.M. Tu^{3,4}

¹ Hospital for Special Surgery, Department of Public Health, Weill Medical College of Cornell University, New York, NY 10021

² Department of Orthopedics, Hospital for Special Surgery, New York, NY 10021

³ Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642

⁴ Department of Psychiatry, University of Rochester, Rochester, NY 14642

Summary

Cronbach Coefficient Alpha (CCA) is a classic measure of item internal consistency of an instrument and is used in a wide range of behavioral, biomedical, psychosocial, and health-care related research. Methods are available for making inference about one CCA or multiple CCAs from correlated outcomes. However, none of the existing approaches effectively address missing data. As longitudinal study designs become increasingly popular and complex in modern-day clinical studies, missing data has become a serious issue, and the lack of methods to systematically address this problem has hampered the progress of research in the aforementioned fields. In this paper, we develop a novel approach to tackle the complexities involved in addressing missing data (at the instrument level due to subject dropout) within a longitudinal data setting. The approach is illustrated with both clinical and simulated data.

Keywords

Cronbach Coefficient Alpha; Inverse probability weighting; Missing data; Monotone missing data pattern; U-statistics

1 Introduction

Measurement error models play a particularly important role in psychosocial research, as many quantities of interest are latent constructs and not directly observable. Outcomes for measuring such constructs are often derived based on instruments (questionnaires) that delve into multifaceted dimensions of an individual's mental and physical health. For most instruments, items are scored on a Likert scale such as 0, 1, 2, 3, 4 and then totaled either across all items or a subset of items from a subscale to provide an assessment of the latent construct or a sub-domain of interest. An elegant measure to ensure that the different items in an instrument or a sub-scale capture a large amount of variability of either a uni-dimensional construct or multi-dimensional constructs sharing a common source is the Cronbach Coefficient Alpha (CCA). CCA ranges between 0 and 1, with larger values indicating higher degrees of consistency across the different items in terms of capturing a common source of variation [1,2].

Methods are available for making inference about a single CCA and multiple dependent CCAs, with the latter based on data from the same subjects [3]–[10]. However, as these methods are derived based on joint multivariate normal distributions of item scores, they are

sensitive to departures from such parametric distributional assumptions. Since item scores from most instruments in psychosocial research are based on Likert-type scales, they are intrinsically discrete and modeling them using parametric distributions for continuous outcomes such as the multivariate normal is theoretically flawed, especially when modeling subscales of an instrument with a very limited numerical range. Further, as these methods for comparing multiple CCAs from either several instruments when applied to a common group of subjects or subscales of an instrument are developed based on statistics for omnibus tests of overall homogeneity involving all the CCAs, they do not provide explicit asymptotic joint distributions of estimates of individual CCAs. As a result, they cannot be applied to test the equality of some combination of the CCAs, which is particularly common in longitudinal studies. For example, we may want to know whether the CCA for an instrument has dropped by 30% from baseline to some follow-up visit in a longitudinal study. None of these methods can be applied to address such a hypothesis.

As longitudinal study designs become increasingly popular and complex in modern-day studies, missing data are inevitable due to loss to follow up, making inference more challenging for analysis of longitudinal CCAs. Although a variety of methods are available for addressing missing data, none applies directly to the current setting involving inference for comparing multiple CCAs within a longitudinal data setting.

In this paper, we develop a novel approach to tackle the complexities involved in addressing missing data within a longitudinal and clustered data setting. The approach is illustrated with both clinical and simulated data.

2 Modeling CCA for Longitudinal Data

2.1 Complete Data

Consider an instrument consisting of K items, which is administered to n subjects in a longitudinal study with T assessments. Let y_{itk} denote the item score to the k th item from the i th subject at time t and let

$$\mathbf{y}_{it} = (y_{it1}, y_{it2}, \dots, y_{itK})^\top, \quad \mathbf{y}_i = (\mathbf{y}_{i1}^\top, \mathbf{y}_{i2}^\top, \dots, \mathbf{y}_{iT}^\top)^\top, \quad 1 \leq i \leq n.$$

Also, let

$$\mu_t = E(\mathbf{y}_{it}) = (\mu_{t1}, \mu_{t2}, \dots, \mu_{tK})^\top, \quad \theta_{t1} = E[(\mathbf{y}_{it} - \mu_t)^\top (\mathbf{y}_{it} - \mu_t)], \\ \theta_{t2} = \mathbf{1}^\top E[(\mathbf{y}_{it} - \mu_t)(\mathbf{y}_{it} - \mu_t)^\top] \mathbf{1}, \quad \theta_t = (\theta_{t1}, \theta_{t2})^\top, \quad \theta = (\theta_1^\top, \theta_2^\top, \dots, \theta_T^\top)^\top,$$

where $\mathbf{1}$ denotes a $K \times 1$ column vector of 1's. The Cronbach Coefficient Alpha α can be expressed as:

$$\alpha_t = \frac{K}{K-1} \left(1 - \frac{\theta_{t1}}{\theta_{t2}} \right) = f_t(\theta_t), \quad \alpha = (\alpha_1, \dots, \alpha_T)^\top = (f_1(\theta_1), \dots, f_T(\theta_T))^\top = \mathbf{f}(\theta).$$

Let

$$\begin{aligned} \mathbf{h}_{ijt} &= (h_{ijt1}, h_{ijt2})^\top = \frac{1}{2} \begin{pmatrix} (\mathbf{y}_{it} - \mathbf{y}_{jt})^\top (\mathbf{y}_{it} - \mathbf{y}_{jt}) \\ \mathbf{1}^\top (\mathbf{y}_{it} - \mathbf{y}_{jt}) (\mathbf{y}_{it} - \mathbf{y}_{jt})^\top \mathbf{1} \end{pmatrix}, \\ \mathbf{h}_{ij} &= (\mathbf{h}_{ij1}^\top, \mathbf{h}_{ij2}^\top, \dots, \mathbf{h}_{ijt}^\top)^\top, \quad (i, j) \in C_2^n, 1 \leq t \leq T. \end{aligned} \quad (1)$$

Then, it is readily checked that

$$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1^\top, \dots, \widehat{\theta}_r^\top)^\top = \left(\begin{matrix} n \\ 2 \end{matrix} \right)^{-1} \sum_{(i,j) \in C_2^n} \mathbf{h}_{ij} \quad (2)$$

is a one-sample, vector-valued U-statistic and $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ [11, chap. 3]. By applying the theory of multivariate U-statistics [12, chap. 5], and the Delta method, $\widehat{\boldsymbol{\alpha}} = \mathbf{f}(\widehat{\boldsymbol{\theta}})$ is a consistent and asymptotically normal estimate of $\boldsymbol{\alpha}$. We summarize the asymptotic properties of $\widehat{\boldsymbol{\alpha}}$ in a theorem below.

Theorem 1—Let $\boldsymbol{\Sigma}_\theta = 4V ar(E(\mathbf{h}_{ij} | \mathbf{y}_i))$. Then, under mild regularity conditions, we have:

- a. $\widehat{\boldsymbol{\alpha}}$ is consistent and asymptotically normal:

$$\widehat{\boldsymbol{\alpha}} \rightarrow_p \boldsymbol{\alpha}, \quad \sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \rightarrow_d N\left(\mathbf{0}, \sum_{\alpha} = \frac{\partial^\top}{\partial \theta} \mathbf{f}(\theta) \sum_{\theta} \frac{\partial}{\partial \theta} \mathbf{f}(\theta)\right), \quad (3)$$

where $\rightarrow_p$ and $\rightarrow_d$ denote convergence in probability and distribution, respectively.

- b. A consistent estimate of the asymptotic variance is obtained by substituting consistent estimates of the respective components of $\boldsymbol{\Sigma}_\alpha$; if $\widehat{\boldsymbol{\Sigma}}_\theta$ is a consistent

estimate of $\boldsymbol{\Sigma}_\theta$, a consistent estimate of $\boldsymbol{\Sigma}_\alpha$ is given by: $\widehat{\sum}_{\alpha} = \frac{\partial^\top}{\partial \theta} \mathbf{f}(\widehat{\boldsymbol{\theta}}) \widehat{\sum}_{\theta} \frac{\partial}{\partial \theta} \mathbf{f}(\widehat{\boldsymbol{\theta}})$.

The proof of Theorem 1 is not provided since these conclusions follow as a special case from the more general results concerning the missing data case to be discussed in Section 2.2.

To find a consistent estimate of $\boldsymbol{\Sigma}_\theta$, first note that:

$$\begin{aligned} E(h_{ijt1} | \mathbf{y}_i) &= \frac{1}{2} [(\mathbf{y}_{it} - \boldsymbol{\mu}_t)^\top (\mathbf{y}_{it} - \boldsymbol{\mu}_t) + \theta_{t1}], \quad E(E(h_{ijt1} | \mathbf{y}_i)) = \theta_{t1}, \\ E(h_{ijt2} | \mathbf{y}_i) &= \frac{1}{2} [\mathbf{1}^\top [(\mathbf{y}_{it} - \boldsymbol{\mu}_t) (\mathbf{y}_{it} - \boldsymbol{\mu}_t)^\top] \mathbf{1} + \theta_{t2}], \quad E(E(h_{ijt2} | \mathbf{y}_i)) = \theta_{t2}, \\ E(\mathbf{h}_{ijt} | \mathbf{y}_i) &= (E(h_{ijt1} | \mathbf{y}_i), E(h_{ijt2} | \mathbf{y}_i))^\top, \\ E(\mathbf{h}_{ij} | \mathbf{y}_i) &= (E(\mathbf{h}_{ij1}^\top | \mathbf{y}_i), \dots, E(\mathbf{h}_{ijt}^\top | \mathbf{y}_i))^\top, \quad E(E(\mathbf{h}_{ij} | \mathbf{y}_i)) = \boldsymbol{\theta}. \end{aligned} \quad (4)$$

Thus, a consistent estimate of $\boldsymbol{\Sigma}_\theta$ is given by:

$$\widehat{\sum}_{\theta} = \frac{4}{n-1} \sum_{i=1}^n \left[(\widehat{E}(\mathbf{h}_{ij} | \mathbf{y}_i) - \widehat{\boldsymbol{\theta}}) (\widehat{E}(\mathbf{h}_{ij} | \mathbf{y}_i) - \widehat{\boldsymbol{\theta}})^\top \right], \quad (5)$$

where $\hat{E}(\mathbf{h}_{ij} | \mathbf{y}_i)$ denotes $E(\mathbf{h}_{ij} | \mathbf{y}_i)$ given in (4) with estimated $\hat{\theta}$ and $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$ substituting in place of θ and μ .

We can use Theorem 1 to test linear contrasts of the form:

$$H_0: K\alpha = \mathbf{0}, \quad \text{vs.} \quad H_a: K\alpha \neq \mathbf{0},$$

where K is some $p \times T$ full rank matrix of known constants. For inference, we can use the

Wald statistic, $Q_n^2 = n\hat{\alpha}^\top K^\top \left(K \sum_{\alpha} K^\top \right)^{-1} K\hat{\alpha}$, which has an asymptotic central χ_p^2 distribution with p degrees of freedom under the null H_0 .

2.2 Missing Data

In the presence of missing data, one approach is to apply the complete-data approach above to the subsample consisting of those subjects with complete data. However, in most longitudinal studies, missing data do not occur completely at random, as they often result from subjects's deteriorated/improved health and other related conditions in response to the treatment. In such cases, missing data are predicted by observed responses and as a result, the missing completely at random (MCAR) assumption does not apply [13]. Thus, such a complete-data approach not only reduces power, but also likely yields biased estimates. A better alternative is to include all available data and to address the inherent missing data problem.

To this end, we assume that missing data only occurs at the instrument level and define a vector of binary variables for indicating missing (or rather observed) instrument-level response as follows:

$$r_{it} = \begin{cases} 1 & \text{if } \mathbf{y}_{it} \text{ is observed} \\ 0 & \text{if } \mathbf{y}_{it} \text{ is unobserved} \end{cases}, \quad \mathbf{r}_i = (r_{i1}, r_{i2}, \dots, r_{iT})^\top, \quad 1 \leq t \leq T. \quad (6)$$

Let

$$\Delta_{it} = \Pr(r_{it}=1 | \mathbf{y}_i), \quad \Delta_i = (\Delta_{i1}, \Delta_{i2}, \dots, \Delta_{iT})^\top.$$

Define a new estimate of α with the form as before $\hat{\alpha} = \mathbf{f}(\hat{\theta})$, but with $\hat{\theta} = (\hat{\theta}_1^\top, \dots, \hat{\theta}_T^\top)^\top$ redefined by:

$$\hat{\theta}_t = \left(\begin{matrix} n \\ 2 \end{matrix} \right)^{-1} \sum_{(i,j) \in C_2^n} \frac{r_{it} r_{jt}}{\Delta_{it} \Delta_{jt}} \mathbf{h}_{ijt}, \quad 1 \leq t \leq T. \quad (7)$$

Note that although $\mathbf{h}_{ijt}$ in (7) is not defined if one of the $\mathbf{y}_{it}$ and $\mathbf{y}_{jt}$ is missing, $\hat{\theta}_t$ is still well-defined since $\mathbf{h}_{ijt}$ can be assigned any value in such cases without affecting $\hat{\theta}_t$. The estimate $\hat{\theta}_t$ in (7) may be viewed as a generalization to a U-statistics setting of the classic *inverse probability weighted* (IPW) estimate for distribution-free inference about standard statistical models such as the weighted generalized estimating equations (WGEE) [14].

First, assume that Δ_i is known. Then, it can be shown that $\hat{\theta}$ is both consistent and asymptotically normal. Thus, by the Delta method, $\hat{\alpha} = \mathbf{f}(\hat{\theta})$ is a consistent and asymptotically normal estimate of α . We summarize these results in a theorem below, with a proof sketched in Appendix A1.

Theorem 2—Let

$$\mathbf{v}_{ijt} = \frac{r_{it}r_{jt}}{\Delta_{it}\Delta_{jt}}(\mathbf{h}_{ijt} - \theta_t), \quad \mathbf{v}_{ij} = (\mathbf{v}_{ij1}^\top, \mathbf{v}_{ij2}^\top, \dots, \mathbf{v}_{ijt}^\top)^\top, \\ \tilde{\mathbf{v}}_{it} = E(\mathbf{v}_{ijt}|\mathbf{y}_i, \mathbf{r}_i), \quad \tilde{\mathbf{v}}_i = (\tilde{\mathbf{v}}_{i1}^\top, \tilde{\mathbf{v}}_{i2}^\top, \dots, \tilde{\mathbf{v}}_{iT}^\top)^\top, \quad \Sigma_\theta = 4\text{Var}(\tilde{\mathbf{v}}_i). \quad (8)$$

Then,

- a. $\hat{\alpha}$ is consistent and asymptotically normal,

$$\hat{\alpha} \rightarrow_p \alpha, \quad \sqrt{n}(\hat{\alpha} - \alpha) \rightarrow_d N\left(0, \sum_{\alpha} = \frac{\partial^\top}{\partial \theta} \mathbf{f}(\theta) \sum_{\theta} \frac{\partial}{\partial \theta} \mathbf{f}(\theta)\right), \quad (9)$$

- b. The asymptotic variance Σ_α can be estimated by the Delta method with a consistent estimate $\hat{\Sigma}_\theta$ given by:

$$\widehat{\text{Cov}}(\tilde{\mathbf{v}}_{it}, \tilde{\mathbf{v}}_{it'}) = \frac{1}{n-1} \sum_{i=1}^n \frac{r_{it}r_{it'}}{\Delta_{it}\Delta_{it'}} \hat{\mathbf{e}}_{it} \hat{\mathbf{e}}_{it'}^\top, \quad 1 \leq t, t' \leq T, \\ \hat{\mathbf{e}}_{it} = (\hat{e}_{it1}, \hat{e}_{it2})^\top, \quad \hat{e}_{it1} = \frac{1}{2} [(\mathbf{y}_{it} - \hat{\mu}_t)(\mathbf{y}_{it} - \hat{\mu}_t) - \hat{\theta}_{t1}], \\ \hat{e}_{it2} = \frac{1}{2} [\mathbf{1}^\top [(\mathbf{y}_{it} - \hat{\mu}_t)(\mathbf{y}_{it} - \hat{\mu}_t)^\top] \mathbf{1} - \hat{\theta}_{t2}], \quad \hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n \frac{r_{it}}{\Delta_{it}} y_{it}. \quad (10)$$

This case with known Δ_i may arise if data are missing by designs, as in some multi-stage trials in which we may want to over-sample those subjects with larger outcomes at a prior stage to participate in the current stage. For example, in a two-stage study with n subjects at stage one, we can model the selection probability Δ_{i2} for stage two using the following logistic regression:

$$\text{logit}(\Delta_{i2}) = \text{logit}(\Pr(r_{i2}=1|y_{i1})) = \beta_0 + \beta_1 y_{i1}, \quad 1 \leq i \leq n, \quad (11)$$

where r_{i2} denotes the selection indicator (1 for selected) and y_{i1} the outcome from the i th subject in the first stage. In the above model, β_0 and β_1 (≥ 0) determines the selection probability for subjects with outcomes y_{i1} to participate in the second stage.

In most applications, Δ_i is unknown and must be estimated. Under MCAR, $\mathbf{r}_i$ is independent of $\mathbf{y}_i$. Consequently, Δ_i is functionally independent of $\mathbf{y}_i$ and is readily estimated by the

respective sample moments: $\hat{\Delta}_t = \frac{1}{n} \sum_{i=1}^n r_{it}$ ($1 \leq t \leq T$). When Δ_i becomes dependent on $\mathbf{y}_i$, it is necessary to model Δ_i as a function of $\mathbf{y}_i$. However, it is difficult to model such a relationship without imposing some additional assumptions on the relationship between the occurrence of missing data and outcome [13]. As in the literature, we focus on the missing at random (MAR) assumption, in which case the occurrence of missing data only depends on the observed response.

Let $\mathbf{y}_i^{obs}$ denote the observed responses for the i th subject. Then, under MAR,

$$\Delta_{it} = \Pr(r_{it} = 1 | \mathbf{y}_i) = \Pr(r_{it} = 1 | \mathbf{y}_i^{obs}). \quad (12)$$

As $\mathbf{y}_i^{obs}$ is the sub-vector of $\mathbf{y}_i$ containing the non-missing responses, Δ_{it} above is essentially a function of the missing data patterns across the subjects. Since there are potentially a total of 2^T different patterns, it is generally not feasible to model and estimate Δ_{it} in most real studies unless there is a certain structure in the patterns. Fortunately, in most longitudinal trials, missing data often occur as the result of subject dropout due to deteriorated/improved health and other related conditions, exhibiting the so-called monotone missing data pattern (MMDP). The structured MMDP reduces the number of missing data patterns to T , making it possible to model such a dependence with limited data in most studies.

We assume no missing data at baseline $t = 1$ so that $r_{i1} = 1$. Under MMDP, if $\mathbf{y}_{it}$ is observed at t , $\mathbf{y}_{is}$ is then observed at all earlier times $s < t$. Let $\tilde{\mathbf{y}}_{it} = (\mathbf{y}_{i1}^\top, \mathbf{y}_{i2}^\top, \dots, \mathbf{y}_{i(t-1)}^\top)^\top$ denote the sub-vector of $\mathbf{y}_i$ containing responses up to and including time $t - 1$ ($2 \leq t \leq T$). Then, under MAR, it follows from (12) that

$$\Delta_{it} = \Pr(r_{it} = 1 | \mathbf{y}_i) = \Pr(r_{it} = 1 | \mathbf{y}_i^{obs}) = \Pr(r_{it} = 1 | \tilde{\mathbf{y}}_{it}). \quad (13)$$

To estimate Δ_{it} , we first model the one-step transition probability of the occurrence of missing data using logistic regression:

$$\begin{aligned} \text{logit}(p_{it}) &= \text{logit}(E(r_{it} = 1 | r_{i(t-1)} = 1, \tilde{\mathbf{y}}_{it})) = \beta_{t0} + \beta_{t1}^\top \tilde{\mathbf{y}}_{it} = \beta_{t0} + \sum_{s=1}^{t-1} \beta_{t1s}^\top \mathbf{y}_{is}, \\ \beta_{t1s} &= (\beta_{t1s1}, \beta_{t1s2}, \dots, \beta_{t1sK})^\top, \quad \beta_{t1} = (\beta_{t11}^\top, \beta_{t12}^\top, \dots, \beta_{t1(t-1)}^\top)^\top, \\ \beta_{t0} &= (\beta_{t0}, \beta_{t1}^\top)^\top, \quad \beta = (\beta_2^\top, \dots, \beta_T^\top)^\top. \end{aligned} \quad (14)$$

In many applications, the limited data may prevent us from reliably estimating β , especially with a large K . In such cases, we may simplify the above model by substituting the mean

item score $\bar{y}_{is} = \frac{1}{K} \sum_{k=1}^K y_{isk}$ in place of $\mathbf{y}_{is}$, yielding:

$$\begin{aligned} \text{logit}(p_{it}) &= \beta_{t0} + \sum_{s=1}^{t-1} \beta_{ts} \bar{y}_{is} = \beta_{t0} + \beta_{t1}^\top \tilde{\mathbf{y}}_{it}, \quad \tilde{\mathbf{y}}_{it} = (\bar{y}_{i1}, \bar{y}_{i2}, \dots, \bar{y}_{i(t-1)})^\top, \\ \beta_{t1} &= (\beta_{t11}, \beta_{t12}, \dots, \beta_{t1(t-1)})^\top, \quad \beta_{t0} = (\beta_{t0}, \beta_{t1}^\top)^\top, \quad \beta = (\beta_2^\top, \dots, \beta_T^\top)^\top. \end{aligned} \quad (15)$$

Now by invoking MMDP and the assumption of no missing data at $t = 1$, we obtain:

$$\Delta_{it} = \Pr(r_{it} = 1, r_{i(t-1)} = 1 | \tilde{\mathbf{y}}_{it}) = p_{it} \Pr(r_{i(t-1)} = 1 | \tilde{\mathbf{y}}_{i(t-1)}) = \prod_{s=2}^t p_{is}. \quad (16)$$

Note that by setting $\beta_{t1} = \mathbf{0}$ in (14) or (15), we can use (16) to estimate Δ_{it} under the MCAR assumption, which is equivalent to estimating Δ_{it} using the sample moments,

$$\widehat{\Delta}_{it} = \frac{1}{n} \sum_{i=1}^n r_{it} (1 \leq t \leq T), \text{ as discussed earlier.}$$

When Δ_i is estimated by the logistic regression model in (14), $\Delta_i(\beta)$ becomes subject to the sampling variability of β . Since Theorem 2 treats Δ_i as a known constant, we must account for this extra variability for correct inference about α .

Suppose that β in the logistic model in (14) is estimated by maximum likelihood. Then, $\hat{\beta}$ is the solution to the following score equations:

$$\begin{aligned} \sum_{i=1}^n \mathbf{w}_i(\beta) &= \sum_{i=1}^n (\mathbf{w}_{i2}^\top, \dots, \mathbf{w}_{iT}^\top)^\top = \mathbf{0}, \\ \mathbf{w}_{it} &= \frac{\partial}{\partial \beta_i} [I(r_{it}=1) \log(p_{it}) + I(r_{it}=0) \log(1-p_{it})], \end{aligned} \quad (17)$$

When using the estimated $\hat{\beta}$, the estimate defined in (7) becomes

$$\widehat{\theta}_t(\widehat{\beta}_t) = \left(\frac{n}{2} \right)^{-1} \sum_{(i,j) \in C_2^n} \frac{r_{it} r_{jt}}{\widehat{\Delta}_{it}(\widehat{\beta}_t) \widehat{\Delta}_{jt}(\widehat{\beta}_t)} \mathbf{h}_{ijt}, \quad 1 \leq t \leq T. \quad (18)$$

Thus, we must also include the variability of β in the asymptotic variance of $\widehat{\theta}(\beta)$. By utilizing the properties of score equations and the projection-based U-statistics asymptotic expansion, we can derive the asymptotic variance to take into account this extra variability. We summarize the results below with a justification given in Appendix A2.

Theorem 3—Let

$$C = 2E \left(\frac{\partial^\top}{\partial \beta} \tilde{\mathbf{v}}_i(\beta) \right), \quad H = E \left(\frac{\partial}{\partial \beta} \mathbf{w}_i(\beta) \right). \quad (19)$$

Then, under the assumptions of Theorem 2,

- a. $\hat{\alpha}$ is consistent and asymptotically normal, with the asymptotic variance,

$$\begin{aligned} \Sigma_{\alpha\Phi} &= \frac{\partial^\top \mathbf{f}}{\partial \theta} (\Sigma_\theta + \Phi) \frac{\partial \mathbf{f}}{\partial \theta}, \\ \Phi &= -4 \left[CH^{-1}C^\top + E(\tilde{\mathbf{v}}_i \mathbf{w}_i^\top H^{-1} C^\top) + (E(\tilde{\mathbf{v}}_i \mathbf{w}_i^\top H^{-1} C^\top))^\top \right], \end{aligned} \quad (20)$$

where Σ_θ is given in (8).

- b. A consistent estimate of $\Sigma_{\alpha\Phi}$ is $\sum_{\alpha\Phi} = \frac{\partial^\top \mathbf{f}}{\partial \theta}(\widehat{\theta}) \left(\widehat{\Sigma}_\theta + \widehat{\Phi} \right) \frac{\partial \mathbf{f}}{\partial \theta}(\widehat{\theta})$, where $\widehat{\Sigma}_\theta$ is given in (10) and $\widehat{\Phi}$ is obtained by substituting moment estimates in place of H , C , and $E(\tilde{\mathbf{v}}_i \mathbf{w}_i^\top H^{-1} C^\top)$.

It is seen from (20) that when the weight function is estimated, the asymptotic variance of $\hat{\alpha}$ has an additional component Φ , which reflects the sampling variability in $\hat{\beta}$.

Although derived from a longitudinal data setting, the results in the theorems can also be applied to cross-sectional study data. For example, most of the available methods have been developed to compare whether several CCAs either from different instruments or from multiple domains, or subscales of an instrument are identical [3]–[10]. By identifying T as the number of instruments or domains of an instrument, y_{itk} then represents the score from the i th subject in response to the k th item within the t th domain. We can use Theorem 1 for inference about such an omnibus hypothesis if there is no missing data. If missing data do occur at the domain level, we can still model the missing data using the above approaches. Although MCAR presents no additional problem, applications under MAR to the current context are generally more difficult, since it may not be possible to identify a simple structure to model the missing data indicators in (6) such as the MMDP for longitudinal data.

3 Application

We illustrate the approach with both clinical and simulated data. We first present a clinical application and then follow up with investigations of the performance of the approach under small to moderate samples by simulation. In all of the examples, we set the statistical significance at $\alpha = 0.05$. All analyses are carried out using a package we have developed for implementing the proposed approach using the R software platform (R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>)

3.1 Clinical Study

A study on the benefits of total knee arthroplasty (TKA) was recently conducted in the Hospital for Special Surgery in 2008 (*“Early clinical and radiographic results of the standard posterior stabilized and high flex polyethylene inserts in the genesis II total knee replacement system: A pilot study”* was approved by the Institutional Review Board of the Hospital for Special Surgery). A modified version of the Knee Society rating System (KSS) [15] was used to report results for patients who underwent TKA at three visits including pre-operation, 6 weeks, and 4 months post-operation. For illustration purposes, we focused on a total of 117 patients who had observed data at the pre-op in the domain of the patient reported portion of the KSS that consists of 5 items measuring knee functions in walking (a 6-level ordinal scale), going up stairs (a 4-level ordinal scale), going down stairs (a 4-level ordinal scale), getting out of chair (a 4-level ordinal scale), and necessity of support for walking (a 3-level ordinal scale), respectively.

In this sample, 20 of the 117 patients were lost to follow-up, yielding a monotone missing data pattern. Let $\mathbf{y}_{it} = (y_{it1}, y_{it2}, \dots, y_{it5})^T$ denote the 5 item scores at time t with $t = 1$ denoting pre-op, and $t = 2, 3$ corresponding to the two post-op assessments at 6 weeks, and 4 months, respectively. The missingness at time t was modeled using the logistic regression model in (15), with the predictor, $\tilde{\mathbf{y}}_{it} = \bar{\mathbf{y}}_{i(t-1)}$, the mean score of $\mathbf{y}_{i(t-1)}$, and baseline covariates including age, sex, and body mass index (BMI):

$$\text{logit}(p_{it}) = \beta_{0t} + \beta_{1t}\tilde{\mathbf{y}}_{it} + \beta_{2t}I(\text{sex}=\text{female}) + \beta_{3t}\text{age} + \beta_{4t}\text{BMI}. \quad (21)$$

Shown in Table 1 are the estimates of parameters from the above model, their standard errors, and corresponding p-values. Missingness at 4 months post-op was dependent on the observed KSS at 6 weeks post-op (p-value=0.043). There was no evidence indicating the dependence of missingness at 6 weeks on the observed KSS at pre-op. The occurrence of missing data was not highly associated with any of the baseline covariates. Given these

results, we proceeded with all subsequent analyses under MAR with inference based on Theorem 3.

Shown in Table 2 are the estimated α_t 's and their standard errors over time, and the Wald test statistic and associated p-value under H_0 of equal CCAs over all assessments computed under MAR. The estimates seem to show a drop in CCA at the first 6 weeks postoperative visit, but the Wald test indicates that this drop is not statistically significant (p-value=0.54). The early postoperative course of patients who underwent a TKA is generally different across patients, with considerable variations in symptoms, range of motion, and functions. These clinical differences are likely responsible for the drop of CCA at the first postoperative visit. As most patients are fully recovered 4 months after the surgery, the variability in such outcomes will reduce substantially, making the CCA bounce back to the level at the pre-operative visit. Thus, despite the large variation in patients' clinical outcomes over the course of surgery, the KSS questionnaire seems to remain internally consistent so that the total score can be used to provide a reliable measure to assist objective evaluation of postoperative total knee arthroplasty.

3.2 Simulation Study

3.2.1 Simulation Method—A limited simulation study was conducted to examine the empirical type I error rate for testing the null of equal CCAs over time based on a 5-item, Likert-scale questionnaire under a longitudinal study design with three assessments for four sample sizes—30, 50, 75, and 100—under complete, and missing data with MCAR and MAR. For each sample size, we generated observers' ratings over time $\mathbf{y}_{it} = (y_{it1}, y_{it2}, \dots, y_{it5})^T$ under the null, $H_0: \alpha_t = \alpha$, with α_t denoting the CCA at time t ($1 \leq t \leq 3$). We tested the null using the Wald statistic Q_n^2 and estimated the type I error rate based on the empirical distribution of the test statistic obtained from 1,000 Monte Carlo (MC) replications.

We created the data $\mathbf{y}_{it}$ over three assessments in two steps. First, we generated continuous outcomes over time, $\mathbf{z}_{it} = (z_{it1}, z_{it2}, \dots, z_{it5})^T$, by simulating a 15×1 vector, $\mathbf{z}_i = (\mathbf{z}_{i1}^T, \mathbf{z}_{i2}^T, \mathbf{z}_{i3}^T)^T$, from a 15-variate normal distribution with mean $\mathbf{0}$ and variance matrix $\Sigma = \Sigma_2 \otimes \Sigma_1$, where $\otimes$ denotes the Kronecker product and Σ_k are defined by:

$$\Sigma_1 = AR_5(1, \rho_{bi}), \quad \Sigma_2 = C_3(\rho_{bt}).$$

In the above, Σ_1 denotes the within-visit, between-item correlation matrix, $AR_m(1, \rho_{bi})$ represents a $m \times m$ first-order autoregressive correlation structure with auto-correlation ρ_{bi} , Σ_2 is the within-subject correlation matrix, and $C_m(\rho_{bt})$ denotes a $m \times m$ compound symmetry correlation matrix with correlation ρ_{bt} . For the simulation study, we set $\rho_{bi} = 0.6$ and $\rho_{bt} = 0.5$, yielding the true CCAs over time, $\alpha_t = \alpha = 0.742$ ($1 \leq t \leq 3$). Next, we transformed each of the continuous observer score z_{itk} into a 5-level categorical response by:

$$y_{itk} = I(z_{itk} \leq \Phi^{-1}(0.1)) + 2I(\Phi^{-1}(0.1) < z_{itk} \leq \Phi^{-1}(0.35)) + \\ + 3I(\Phi^{-1}(0.35) < z_{itk} \leq \Phi^{-1}(0.5)) + 4I(\Phi^{-1}(0.5) < z_{itk} \leq \Phi^{-1}(0.8)) + \\ + 5(I(\Phi^{-1}(0.8) < z_{itk})), \quad 1 \leq t \leq 3, \quad 1 \leq k \leq 5,$$

where Φ^{-1} denotes the inverse of the CDF of the standard normal.

For a given sample size n , let Q_{nj}^2 be the Wald statistic from the j th MC replication and $G_{0.95}$ the 95th percentile of the χ_2^2 distribution with 2 degrees of freedom. The empirical size is calculated as $\widehat{\delta} = \frac{1}{1000} \sum_{j=1}^{1000} I_{\{Q_{nj}^2 \geq G_{0.95}\}}$.

For the missing data case, we assumed no missing data at baseline $t = 1$ and simulated the missing response according to a MCAR and MAR model. For MAR, we considered the MMDP model, and simulated the missing data indicators at each time t , r_{it} , from a Bernoulli distribution with the probability of success equal to p_{it} ($t = 2, 3$). We specified the one-step transition probabilities p_{it} at time t according to the logistic model in (15) under the Markov assumption, which in this case had the following form:

$$\begin{aligned} \text{logit}(p_{it}) &= \text{logit}(E(r_{it}=1 | r_{i(t-1)}=1, \tilde{\mathbf{y}}_{it})) = \beta_{t0} + \beta_{t1} \bar{y}_{i(t-1)}, \\ \beta_t &= (\beta_{t0}, \beta_{t1})^\top, \quad \beta = (\beta_2^\top, \beta_3^\top)^\top, \end{aligned}$$

where $\bar{y}_{i(t-1)}$ denotes the mean of $\mathbf{y}_{it}$. We fixed $\beta_{t1} = 0.3$ ($t = 1, 2$) and solved the following equations for β_{20} and β_{30} to generate missing response rates of approximately 10% and 15% for times 2 and 3, respectively:

$$0.9n = \sum_{i=1}^n p_{i2}, \quad 0.85n = \sum_{i=1}^n p_{i2} p_{i3}. \quad (22)$$

To ensure that the missing data indicators r_{it} followed the MMDP, we further imposed the following restrictions, $r_{i3} = r_{i3} \times r_{i2}$ ($1 \leq i \leq n$). For MCAR, the same approach above was used except for setting $\beta_{t1} = 0$ in (22) ($t = 1, 2$).

In addition to examining type I error rates, power analysis for testing the null of equal CCAs over time was conducted in a second simulation study within the same longitudinal data setting under complete data, MCAR, and MAR. The true alpha was set at $\alpha_1 = 0.7$, $\alpha_2 = 0.803$, and $\alpha_3 = 0.898$ for times 1, 2, and 3, respectively. The power was estimated by

$$\widehat{\Lambda} = \frac{1}{1000} \sum_{j=1}^{1000} I_{\{Q_{nj}^2 \geq G_{0.95}\}}.$$

3.2.2 Simulation Result—We computed the estimate $\hat{\alpha}$ of $\alpha = (\alpha_1, \alpha_2, \alpha_3)^\top$ and estimated its asymptotic variance $\hat{\Sigma}_{\alpha\Phi}$ using Theorems 3. Shown in Table 3 are the averaged $\hat{\alpha}$ over 1,000 Monte Carlo (MC) replications, along with the empirical variance $Var(\hat{\alpha})$. The type I

error rates were obtained from $\widehat{\delta} = \frac{1}{1000} \sum_{j=1}^{1000} I_{\{Q_{nj}^2 \geq G_{0.95}\}}$. Under complete data, MCAR, and MAR, the averaged $\hat{\alpha}$ seemed to have some downward bias, especially for small samples $n = 30, 50$ under missing data. The asymptotic standard errors based on $\hat{\Sigma}_{\alpha\Phi}$ were almost identical to their empirical counterparts, $Var(\hat{\alpha})$, across samples under both complete and missing data. Although the empirical type I error rates $\hat{\delta}$ were slightly higher than the nominal value 0.05 at $n = 30$, the upward bias seemed to have rapidly disappeared as the sample size increased.

Shown in Table 4 are the averaged $\hat{\alpha}$ over 1,000 Monte Carlo (MC) replications, the empirical variance $Var(\hat{\alpha})$, the model based asymptotic variance $\hat{\Sigma}_{\alpha\Phi}$, and estimated power $\hat{\Lambda}$. Similar patterns in power were found across complete data, MCAR, and MAR. Although the estimated power was low at $n = 30$, it increased rapidly as the sample size increased.

We repeated the simulation for a larger number of items, $k = 10$. Compared to the results for $k = 5$ items, type I error rates are closer to the nominal 0.05 even for the small sample size $n = 30$. Power analysis was also improved significantly by achieving 60% power for sample size $n = 30$, which jumped up to 85% for complete, and MCAR, and 83% for MAR at $n = 50$.

4 Discussion

We developed a U-statistics based approach to concurrently address missing data and stringent distribution assumptions when modeling the Cronbach Coefficient Alpha (CCA) within a longitudinal data setting. Unlike available methods, this approach requires no parametric distribution assumption on the item score and thus provides robust inference regardless of the data distribution. As longitudinal studies have become increasing popular in today's research, the proposed approach fills a critical gap on the extant literature to address this timely data-analytic issue.

To demonstrate the robustness feature, we also conducted a simulation study to compare the U-statistic based approach with a popular method developed by Feldt et al [6] in a cross-sectional setting by considering the hypothesis $H_0: \alpha = \alpha_0$ vs. $H_0: \alpha \neq \alpha_0$ under various sample sizes ($n = 30, 50, 100$, and 200). Feldt et al [6] showed that the transformed alpha $\frac{1-\alpha}{1-\alpha_0}$ follows a F distribution under the assumed ANOVA model where items by subjects interaction effects and the residual errors are independent and normally distributed with homogeneous variance. To examine its performance under likert-type item scores, we generated skewed likert-type item scores with true alpha $\alpha_0 = 0.693$ and $k = 5$ items. Shown in Table 5 are the estimates of CCA and type I error rates for the proposed U-statistic based approach and their F -distribution based counterpart. Although the estimates of alpha, $\hat{\alpha}_U$ (U-statistics based) and $\hat{\alpha}_F$ (F -distribution based), are identical and close to the true alpha, the type I error rates are quite different between the two approaches. The type I error rate for the U-statistic based approach had some upward bias when the sample size was small, but rapidly decreased to the nominal level 0.05 as the sample size increased. In contrast, the type I error rate for the F -distribution based method remained upwardly biased despite the increase of the sample size. In fact, the type I error will remain upwardly biased even if the sample size approaches infinity, since the F -distribution is not the asymptotic distribution of the estimate. The observed difference in the behavior of type I error rates reflects the fact that the U-statistic based approach is a distribution-free procedure, yielding not only consistent estimates of CCA but also valid inference regardless of the distribution of the item score, while the inference of the F -distribution based method is based on the normal distribution and is not robust against departures from this underlying assumption.

Modeling CCA presents a major challenge under the popular mean response based regression paradigm as typified by the weighted generalized estimating equations (WGEE) due to the complexity in the analytic expression of this coefficient involving second-order moments. Although GEE II may be used to model the required second-order moments [12,16,17], it leads to more cumbersome algebra. Further, as it requires modeling more parameters, GEE II is generally not as efficient as the proposed U-statistics based approach.

The proposed IPW-based estimate may not be fully efficient since information from the subjects with missing data is only used to estimate the weight function. In many situations, reliable models may also exist for directly modeling the relationship between (missing) response and observed covariates or other auxiliary information. In such cases, we may combine the proposed IPW with this extra model to develop a *double robust* estimate, which not only ensures consistency if only one of the models—the model for missing data or the one for the auxiliary information—is correct, but also improves efficiency [14,18].

In addition to CCA, the proposed approach is similarly applied to modeling other measures of agreement for continuous outcome such as the closely related intraclass correlation coefficient (ICC) [19]–[22]. In addition, missing data may also occur at the item level which can become complicated depending on the nature of the question and on the design of the study. For example, person-to-person interviews often result in more missing data and/or unreliable responses compared to questionnaires as interviewees may feel uncomfortable and refuse to answer certain types of questions in a face to face meeting. As missing responses arising from such sensitive questions likely follow the non-ignorable missing data rather than the missing at random (MAR) mechanism as in longitudinal studies, they are best addressed up front by study designs, since the former missingness mechanism is more difficult to model as compared to the MAR mechanism. Work is currently underway to generalize the proposed approach to address these limitations as well as to develop double robust estimates within a longitudinal setting.

Acknowledgments

This research is supported in part by NIH grant U54 RR023480. Dr. Ma was partially supported by the following grants: Center for Education and Research in Therapeutics (CERTs) (AHRQ RFA-HS-05-14) and Clinical Translational Science Center (CTSC) (UL1-RR024996). We sincerely thank Ms. Cheryl Bliss-Clark at the University of Rochester, an Editor and two anonymous reviewers for their constructive comments to improve the presentation of the manuscript.

References

1. Cronbach, L.J.; Gleser, G.C.; Nanda, H.; Rajaratnam, N. The Dependability of Behavioral Measurements. Wiley; New York: 1972.
2. Nunnally, J.C.; Bernstein, I.H. Psychometric Theory. McGraw Hill; New York: 1994.
3. Duhachek A, Iacobucci D. Alpha's Standard Error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*. 2004; 89:792–808. [PubMed: 15506861]
4. Feldt LS. The approximate sampling distribution of the Kuder-Richardson reliability coefficient twenty. *Psychometrika*. 1965; 34:363–373.
5. Feldt LS. A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*. 1980; 45:99–105.
6. Feldt LS, Woodruff DJ, Salih FA. Statistical inference for coefficient alpha. *Applied Psychological Measurement*. 1987; 11:93–103.
7. Feldt LS, Kim S. Testing the difference between two alpha coefficients with small samples of subjects and raters. *Educational and Psychological Measurement*. 2006; 66:589–600.
8. Woodruff DJ, Feldt LS. Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*. 1986; 51:393–413.
9. Alsawalmeh YM, Feldt LS. Testing the equality of independent alpha coefficients adjusted for test length. *Educational and Psychological Measurement*. 1999; 59:373–383.
10. van Zyl JM, Neudecker H, Nel DG. On the distribution of the maximum likelihood estimator of Cronbach's Alpha. *Psychometrika*. 2000; 65:271–280.
11. Randels, R.H.; Wolfe, D.A. Introduction to the Theory of Nonparametric Statistics. Wiley; New York: 1979.
12. Kowalski, J.; Tu, X.M. Modern Applied U Statistics. Wiley; New York: 2007.
13. Little, R.J.A.; Rubin, D.B. Statistical Analysis with Missing Data. Wiley; New York: 1987.
14. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *JASA*. 1995; 90:106–121.
15. Insall JN, Dorr LD, Scott RD, Scott WN. Rationale of the knee society clinical rating system. *Clinical Orthopaedics and Related Research*. 1989; 248:13–14. [PubMed: 2805470]
16. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics*. 1988; 44:321–327.

17. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*. 1991; 47:825–839. [PubMed: 1742441]
18. Tsiatis, AA. *Semiparametric Theory and Missing Data*. Springer; New York: 2006.
19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
20. McGraw KO, Wong SP. Forming inferences about some Intraclass Correlation Coefficients. *Psychological Methods*. 1996; 1:30–46.
21. Fleiss, JL.; Levin, B.; Paik, MC. *Statistical Methods for Rates and Proportions*. 3. Wiley; New York: 2003.
22. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*. 1977; 33:363–374. [PubMed: 884196]

Appendix

A1. Proof of Theorem 2

We first establish a lemma.

Lemma

Let

$$\mathbf{U}_n = \left(\frac{n}{2} \right)^{-1} \sum_{(i,j) \in C_2^n} \mathbf{v}_{ij} = (\mathbf{U}_{1,n}^\top, \dots, \mathbf{U}_{T,n}^\top)^\top, \mathbf{U}_{t,n} = \left(\frac{n}{2} \right)^{-1} \sum_{(i,j) \in C_2^n} \mathbf{v}_{ijt},$$

where $\mathbf{v}_{ijt}$ are defined in (8). Then, $E(\mathbf{U}_n) = \mathbf{0}$ and

$$\sqrt{n}\mathbf{U}_n = \sqrt{n} \left(\frac{n}{2} \right)^{-1} \sum_{(i,j) \in C_2^n} \mathbf{v}_{ij} \rightarrow_d N(\mathbf{0}, \sum_U = 4\text{Var}(\tilde{\mathbf{v}}_i)) \quad (23)$$

Proof of Lemma

It is readily checked by the iterated conditional expectation that [12, chap. 1]

$$\begin{aligned} E(\mathbf{U}_{t,n}) &= E \left[E \left(\frac{r_{it}r_{jt}}{\Delta_{it}\Delta_{jt}} (\mathbf{h}_{ijt} - \theta_t) | \mathbf{y}_i, \mathbf{y}_j \right) \right] \\ &= E \left[\Delta_{it}^{-1} \Delta_{jt}^{-1} (\mathbf{h}_{ijt} - \theta_t) E(r_{it} | \mathbf{y}_i) E(r_{jt} | \mathbf{y}_j) \right] \\ &= \mathbf{0}. \end{aligned} \quad (24)$$

It follows that $E(\mathbf{U}_n) = \mathbf{0}$. Since $\mathbf{v}_{kjt} = \mathbf{v}_{jkt}$ ($j \neq k$), we have

$$\tilde{\mathbf{v}}_{it} = E(\mathbf{v}_{jkt} | \mathbf{y}_i, \mathbf{r}_i) = \begin{cases} 0 & \text{if } j \neq i, k \neq i \\ E(\mathbf{v}_{ikt} | \mathbf{y}_i, \mathbf{r}_i) & \text{if } j=i \\ E(\mathbf{v}_{ijt} | \mathbf{y}_i, \mathbf{r}_i) & \text{if } k=i \end{cases}.$$

Let $\Upsilon(i) = \{(j, k) \in C_2^n; j \neq i, k \neq i\}$. It then follows that

$$\begin{aligned} E(\mathbf{U}_{t,n}|\mathbf{y}_i, \mathbf{r}_i) &= \left(\frac{n}{2}\right)^{-1} \left[\sum_{(j,k) \notin \Upsilon(i)} E(\mathbf{v}_{jkt}|\mathbf{y}_i, \mathbf{r}_i) + \sum_{(j,k) \in \Upsilon(i)} E(\mathbf{v}_{jkt}|\mathbf{y}_i, \mathbf{r}_i) \right] \\ &= \left(\frac{n}{2}\right)^{-1} \left[\sum_{(j,k) \notin \Upsilon(i)} E(\mathbf{v}_{jkt}|\mathbf{y}_i, \mathbf{r}_i) \right] \\ &= \left(\frac{n}{2}\right)^{-1} \left[\sum_{j=i, (j,k) \notin \Upsilon(i)} E(\mathbf{v}_{jkt}|\mathbf{y}_i, \mathbf{r}_i) + \sum_{k=i, (j,k) \notin \Upsilon(i)} E(\mathbf{v}_{jkt}|\mathbf{y}_i, \mathbf{r}_i) \right] \\ &= \left(\frac{n}{2}\right)^{-1} \left[\sum_{k=i+1}^n E(\mathbf{v}_{jkt}|\mathbf{y}_i, \mathbf{r}_i) + \sum_{j=1}^{i-1} E(\mathbf{v}_{ijt}|\mathbf{y}_i, \mathbf{r}_i) \right] \\ &= \frac{2}{n} \tilde{\mathbf{v}}_{it}. \end{aligned}$$

Thus, the projection of $\mathbf{U}_{t,n}$ is given by [12, chap. 5]

$$\tilde{\mathbf{U}}_{t,n} = \sum_{i=1}^n E(\mathbf{U}_{t,n}|\mathbf{y}_i, \mathbf{r}_i) = \frac{1}{n} \sum_{i=1}^n 2\tilde{\mathbf{v}}_{it}.$$

Since $\tilde{\mathbf{U}}_{t,n}$ is a sum of independently and identically distributed random variables, it follows from the central limit theorem (CLT) that

$$\sqrt{n}\tilde{\mathbf{U}}_{t,n} = \frac{\sqrt{n}}{n} \sum_{i=1}^n 2\tilde{\mathbf{v}}_{it} \rightarrow_d N(\mathbf{0}, \sum_t = 4\text{Var}(\tilde{\mathbf{v}}_{it})). \quad (25)$$

By the theory of U-statistics (e.g. [12, chap. 5], $\tilde{\mathbf{U}}_{t,n}$ and $\mathbf{U}_{t,n}$ have the same asymptotic distribution and thus

$$\sqrt{n}\mathbf{U}_{t,n} \rightarrow_d N(\mathbf{0}, \sum_{\gamma_t} = 4\text{Var}(\tilde{\mathbf{v}}_{it})).$$

The lemma follows by applying a similar argument to the vector $\mathbf{U}_n$.

Proof of Theorem 2

Let $\mathbf{g}_{ijt} = \frac{r_{it}r_{jt}}{\Delta_{it}\Delta_{jt}} \mathbf{h}_{ijt}$. By an argument similar to (24), we have, $E(\mathbf{g}_{ijt}) = \boldsymbol{\theta}_t$. It then follows from the theory of U-statistics that

$$\widehat{\boldsymbol{\theta}}_t = \left(\frac{n}{2}\right)^{-1} \sum_{(i,j) \in C_2^n} \mathbf{g}_{ijt} \rightarrow_p \boldsymbol{\theta}_t.$$

Thus, by Slutsky's theorem, $\widehat{\boldsymbol{\alpha}}_t = \mathbf{f}(\widehat{\boldsymbol{\theta}}_t)$ is consistent. Further, by applying the Lemma and Slutsky's theorem, we obtain the asymptotic distribution of $\widehat{\boldsymbol{\alpha}}_t$:

$$\sqrt{n}(\widehat{\theta}_t - \theta_t) = \sqrt{n} \left(\frac{n}{2} \right)^{-1} \sum_{(i,j) \in C_2^n} \tilde{\mathbf{U}}_{t,n} \rightarrow_d N(0, 4\text{Var}(\tilde{\mathbf{v}}_{it})).$$

Similarly, by considering the vector $\widehat{\boldsymbol{\theta}}$, we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{n} \left(\frac{n}{2} \right)^{-1} \sum_{(i,j) \in C_2^n} \tilde{\mathbf{U}}_n \rightarrow_d N(\mathbf{0}, \sum_{\boldsymbol{\theta}} = 4\text{Var}(\tilde{\mathbf{v}}_i)).$$

Theorem 2/(a) follows by applying the Delta method to $\widehat{\boldsymbol{\rho}} = \mathbf{f}(\widehat{\boldsymbol{\theta}})$.

To show (10), first note that

$$\begin{aligned} E(h_{ijt1}|\mathbf{y}_i) &= \frac{1}{2} [(\mathbf{y}_{it} - \mu_t)^\top (\mathbf{y}_{it} - \mu_t) + \theta_{t1}], \\ E(h_{ijt2}|\mathbf{y}_i) &= \frac{1}{2} [\mathbf{1}^\top [(\mathbf{y}_{it} - \mu_t) (\mathbf{y}_{it} - \mu_t)^\top] \mathbf{1} + \theta_{t2}], \end{aligned}$$

Further, we have

$$\begin{aligned} \tilde{\mathbf{v}}_{it} &= E(\mathbf{v}_{ijt}|\mathbf{y}_i, \mathbf{r}_i) = E \left[\frac{r_{it}r_{jt}}{\Delta_{it}\Delta_{jt}} (\mathbf{h}_{ijt} - \theta_t) | \mathbf{y}_i, \mathbf{r}_i \right] = \frac{r_{it}}{\Delta_{it}} E \left[\frac{r_{jt}}{\Delta_{jt}} (\mathbf{h}_{ijt} - \theta_t) | \mathbf{y}_i, \mathbf{r}_i \right] \\ &= \frac{r_{it}}{\Delta_{it}} E \left\{ E \left[\frac{r_{jt}}{\Delta_{jt}} (\mathbf{h}_{ijt} - \theta_t) | \mathbf{y}_i, \mathbf{y}_j, \mathbf{r}_i \right] | \mathbf{y}_i, \mathbf{r}_i \right\} = \frac{r_{it}}{\Delta_{it}} E \left\{ \Delta_{jt}^{-1} (\mathbf{h}_{ijt} - \theta_t) E[r_{jt} | \mathbf{y}_i, \mathbf{y}_j, \mathbf{r}_i] | \mathbf{y}_i, \mathbf{r}_i \right\} \\ &= \frac{r_{it}}{\Delta_{it}} [E(\mathbf{h}_{ijt}|\mathbf{y}_i) - \theta_t] = \frac{r_{it}}{\Delta_{it}} \mathbf{e}_{it}. \end{aligned}$$

Thus, Theorem 2/(b) follows by taking the covariance between $\tilde{\mathbf{v}}_{it}$ and $\tilde{\mathbf{v}}_{it'}$ with μ_t substituted by its consistent estimate in (10).

A2. Proof of Theorem 3

As noted in Section 2.2, $\widehat{\boldsymbol{\beta}}$ is the solution to the score equations in (17). From the properties of score equations, we have:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = -H^{-1} \frac{\sqrt{n}}{n} \sum_{i=1}^n \mathbf{w}_i + \mathbf{o}_p(1), \quad (26)$$

where H is given in (19) and $\mathbf{o}_p(1)$ denotes the stochastic $\mathbf{o}(1)$ [12, chap. 1]. It follows from (25), (26), and the projection-based U-statistics asymptotic expansion that [12, chap. 5]

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}) &= \sqrt{n}[\widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}) - \boldsymbol{\theta} + \widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{\theta}}(\boldsymbol{\beta})] = \sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}) + C \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{o}_p(1) \\ &= \sqrt{n}\tilde{\mathbf{U}}_n(\boldsymbol{\beta}) + C \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{o}_p(1) = \sqrt{n} \sum_{i=1}^n (\tilde{\mathbf{v}}_i - CH^{-1}\mathbf{w}_i) + \mathbf{o}_p(1), \end{aligned} \quad (27)$$

where C is given in (19). Theorem 3/(a) follows from (27), while (b) is obtained by an application of Slutsky's theorem.

Table 1

Estimates of logistic regression model for missingness under MMDP for the Total Knee Arthroplasty Study.

Assessment time	Predictors	Estimates	Standard errors	P-values
6 weeks post-op ($t = 2$)	Intercept	-0.725	6.337	0.909
	$\tilde{y}_{i2}$	1.951	1.251	0.119
	female	1.411	1.073	0.189
	age	-0.019	0.06	0.749
	BMI	0.004	0.094	0.967
4 months post-op ($t = 3$)	Intercept	-5.974	3.363	0.076
	$\tilde{y}_{i3}$	1.699	0.842	0.043
	female	-0.136	0.729	0.853
	age	0.028	0.03	0.346
	BMI	0.034	0.046	0.464

Table 2

Estimates of CCA over time, test statistics and p-values for testing the null of equal CCA over time under MAR for the Total Knee Arthroplasty Study.

Estimates of CCA (standard error)		
Pre-OP ($t = 1$)	6 weeks Post-OP ($t = 2$)	4 months Post-OP ($t = 3$)
0.721(0.043)	0.646(0.087)	0.748(0.034)
Hypothesis testing $H_0: \alpha_1 = \alpha_2 = \alpha_3$		
Q_n^2		p-value
1.233		0.54

Averaged $\hat{\alpha}$ and $\hat{\Sigma}_{\alpha\phi}$ over 1,000 MC replications along with estimated empirical variance of $\hat{\alpha}$ and empirical type I error rate δ under complete data, MCAR, and MAR with the true $\alpha_t = \alpha = 0.742$ ($t = 1, 2, 3$).

Table 3

Sample size	30			50			75			100		
	Assessment time											
	1	2	3	1	2	3	1	2	3	1	2	3
Complete data												
$\hat{\alpha}$	0.725	0.729	0.725	0.73	0.732	0.731	0.735	0.736	0.734	0.696	0.800	0.891
$Var(\hat{\alpha})$	0.008	0.007	0.008	0.004	0.004	0.004	0.003	0.002	0.003	0.003	0.001	0.001
$\hat{\Sigma}_{\alpha\phi}$	0.007	0.006	0.007	0.004	0.004	0.004	0.002	0.002	0.002	0.002	0.001	0.001
δ	0.071				0.055		0.052				0.049	
MCAR												
$\hat{\alpha}$	0.725	0.726	0.722	0.73	0.731	0.729	0.735	0.736	0.733	0.738	0.736	0.736
$Var(\hat{\alpha})$	0.008	0.009	0.009	0.004	0.004	0.005	0.003	0.003	0.003	0.002	0.002	0.002
$\hat{\Sigma}_{\alpha\phi}$	0.007	0.009	0.01	0.004	0.005	0.005	0.002	0.003	0.003	0.002	0.002	0.002
δ	0.059				0.052		0.047				0.051	
MAR												
$\hat{\alpha}$	0.725	0.725	0.721	0.73	0.731	0.729	0.735	0.736	0.732	0.738	0.736	0.735
$Var(\hat{\alpha})$	0.008	0.009	0.01	0.004	0.004	0.005	0.003	0.003	0.003	0.002	0.002	0.002
$\hat{\Sigma}_{\alpha\phi}$	0.007	0.009	0.01	0.004	0.005	0.005	0.002	0.003	0.003	0.002	0.002	0.002
δ	0.061				0.056		0.056				0.052	

Table 4

Averaged $\hat{\alpha}$ and $\hat{\Sigma}_{\alpha\phi}$ over 1,000 MC replications along with estimated empirical variance of $\hat{\alpha}$ and power $\hat{I}$ under complete data, MCAR, and MAR with the true $\alpha_1 = 0.7$, $\alpha_2 = 0.803$, and $\alpha_3 = 0.898$, at $t = 1, 2, 3$, respectively.

Sample size	Assessment time								
	30			50			75		
	1	2	3	1	2	3	1	2	3
Complete data									
$\hat{\alpha}$	0.681	0.791	0.897	0.686	0.797	0.885	0.691	0.799	0.891
$Var(\hat{\alpha})$	0.011	0.005	0.006	0.006	0.002	0.002	0.004	0.001	0.001
$\hat{\Sigma}_{\alpha\phi}$	0.009	0.004	0.003	0.005	0.002	0.002	0.003	0.001	0.001
$\hat{I}$	0.464			0.662			0.839		0.919
MCAR									
$\hat{\alpha}$	0.681	0.789	0.866	0.686	0.796	0.884	0.691	0.799	0.889
$Var(\hat{\alpha})$	0.011	0.006	0.008	0.006	0.002	0.003	0.004	0.002	0.002
$\hat{\Sigma}_{\alpha\phi}$	0.009	0.005	0.006	0.005	0.003	0.002	0.003	0.002	0.002
$\hat{I}$	0.42			0.628			0.786		0.908
MAR									
$\hat{\alpha}$	0.681	0.789	0.868	0.686	0.796	0.884	0.691	0.799	0.889
$Var(\hat{\alpha})$	0.011	0.006	0.007	0.006	0.003	0.003	0.004	0.002	0.002
$\hat{\Sigma}_{\alpha\phi}$	0.009	0.005	0.006	0.005	0.003	0.002	0.003	0.002	0.001
$\hat{I}$	0.413			0.624			0.795		0.902

Table 5

Estimates of type I error rate and CCA for the U-statistic based and F-distribution based approaches.

Sample size	Estimates of CCA	Type I error
N=30	$\hat{\alpha}_U = 0.664$	$\hat{\delta}_U = 0.105$
	$\hat{\alpha}_F = 0.664$	$\hat{\delta}_F = 0.108$
N=50	$\hat{\alpha}_U = 0.675$	$\hat{\delta}_U = 0.09$
	$\hat{\alpha}_F = 0.675$	$\hat{\delta}_F = 0.11$
N=100	$\hat{\alpha}_U = 0.687$	$\hat{\delta}_U = 0.054$
	$\hat{\alpha}_F = 0.687$	$\hat{\delta}_F = 0.099$
N=200	$\hat{\alpha}_U = 0.689$	$\hat{\delta}_U = 0.052$
	$\hat{\alpha}_F = 0.689$	$\hat{\delta}_F = 0.116$