Jiwei He*, Alisa Stephens-Shields and Marshall Joffe

Structural Nested Mean Models to Estimate the Effects of Time-Varying Treatments on Clustered Outcomes

DOI 10.1515/ijb-2014-0055

Abstract: In assessing the efficacy of a time-varying treatment structural nested models (SNMs) are useful in dealing with confounding by variables affected by earlier treatments. These models often consider treatment allocation and repeated measures at the individual level. We extend SNMMs to clustered observations with time-varying confounding and treatments. We demonstrate how to formulate models with both cluster- and unit-level treatments and show how to derive semiparametric estimators of parameters in such models. For unit-level treatments, we consider interference, namely the effect of treatment on outcomes in other units of the same cluster. The properties of estimators are evaluated through simulations and compared with the conventional GEE regression method for clustered outcomes. To illustrate our method, we use data from the treatment arm of a glaucoma clinical trial to compare the effectiveness of two commonly used ocular hypertension medications.

Keywords: clustered observations, time-varying confounding, structural nested mean models

1 Introduction

Structural nested models (SNMs) are a class of models that is useful for estimating the causal effect of a time-varying treatment in the presence of time-varying confounding [1]. In some studies, treatments are affected by covariates at earlier time points and subsequently affect covariates at later time points. Confounding that arises in such situations cannot be properly controlled for by conventional regression methods. However, such problems can be solved by several other methods, including marginal structural models (MSMs) and SNMs. Compared to MSMs, SNMs have the advantage of being able to model effect modification by time-varying covariates, whereas MSMs can only model effect modification by baseline covariates. SNMs are also useful for modeling the effect of dynamic treatment regimes, in which the treatment plan changes according to observed patient characteristics during the treatment course, in addition to static treatment regimes, in which the entire treatment schedule is known at the time of initiation [2]. Another advantage of SNMs is that they can work under conditions where certain assumptions fail. For example, SNMs may be fit under instrumental variable assumptions when there is unmeasured confounding [3, 4]. There are two types of SNMs: (1) structural nested mean models (SNMMs) model the joint effect of treatments on the mean of the outcome conditional on the treatment and covariate history, and (2) structural nested distribution models (SNDMs) model the joint effect of treatments on the entire distribution of the outcome conditional on the treatment and covariate history. So far, SNMs have seldom been used for clustered observations. Brumback et al. [5] used SNMs and an instrumental variable approach to estimate the effect of cluster-level treatment on unit-level outcomes in the presence of unmeasured confounding [5]. This paper generalizes previous work by formulating SNMMs for clustered observations with time-varying confounding and treatments. The formulation applies to both unit-level and cluster-level treatments.

A motivating example is the study of eye disease treatment, where two eyes of a given subject form a cluster. Eye-specific outcomes from a given subject tend to be correlated, because they share subject-specific characteristics. Rubin [6] calls a situation where “the observation on one unit is unaffected by the

*Corresponding author: Jiwei He, Department of Biostatistics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA, E-mail: jiweihe@mail.med.upenn.edu

Alisa Stephens-Shields, Marshall Joffe, Department of Biostatistics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, USA

particular treatment assignment to the other units” the “stable unit-treatment value assumption (SUTVA)” [6]. If we consider each eye as a separate unit for treatment, the SUTVA assumption may be violated. Under eye-specific treatment, treatment received by one eye may affect outcomes in the other eye of the same subject. If the treatments are binary, each eye will have not only two but four possible treatments, depending on the treatment received by both eyes. VanderWeele [7] has discussed the SUTVA in the context of neighborhood effects [7]. In the ophthalmology study, treatments are eye-specific drops. Treatments, clinical covariates, and outcomes are measured repeatedly over an extended period of time; treatments are therefore subject to time-dependent confounding. We are interested in modeling the joint effect of a series of treatments on a series of clustered outcomes. Other examples of clustered observations include repeated measurements of disease symptoms on the same individual or an infectious disease among people in the same household.

The structure of the paper is as the following: we first describe the setup for exploring causal effects among clustered observations. We then formulate SNMMs for clustered observations in a single time point setting and describe an efficient estimating equation. We then extend the model and estimation to time-varying treatments and repeatedly measured outcomes. Last, the method is evaluated using simulation studies and applied to an ophthalmology dataset.

2 Notation for clustered observations

We first describe the general structure of data with repeatedly measured treatments and repeatedly measured clustered outcomes and the definition of a causal effect in such a context. The notation is based on Robins [1] with extension for clustered data [1]. Suppose that data is collected on N clusters, $i = 1, 2, \dots, N$ at $K + 1$ time points, $k = 0, 1, \dots, K$. At each time point k , outcomes are measured from J units in each cluster, $j = 1, 2, \dots, J$. Let A_{imj} denote unit-specific treatment in the j th unit at time m for cluster i , and $A_{im}^\dagger$ denote systemic treatment received at time m for cluster i . Let L_{imj} denote unit-specific covariates in the j th unit, and $L_{im}^\dagger$ denote cluster-specific covariates measured at time m for cluster i . Y_{imj} denotes the observed outcome in j th unit measured at time m for cluster i . Y_{imj} is considered as part of L_{imj} . $\mathbf{A}_{im} = \{A_{im1}, \dots, A_{imJ}, A_{im}^\dagger\}$ is all treatments received at time m by cluster i . $\mathbf{L}_{im} = \{L_{im1}, \dots, L_{imJ}, L_{im}^\dagger\}$ is all covariates measured at time m in cluster i . $\mathbf{Y}_{im} = \{Y_{im1}, \dots, Y_{imJ}\}$ is all observed outcomes measured at time m in cluster i . We use overbars to denote the history of a variable; thus, $\bar{\mathbf{A}}_{im} = \{\mathbf{A}_{i0}, \mathbf{A}_{i1}, \dots, \mathbf{A}_{im}\}$ is treatment history through time m in cluster i , and $\bar{\mathbf{L}}_{im} = \{\mathbf{L}_{i0}, \mathbf{L}_{i1}, \dots, \mathbf{L}_{im}\}$ is covariate history through time m in cluster i . We use underbars to denote the future of a variable; thus, $\underline{\mathbf{Y}}_{im} = \{\mathbf{Y}_{im}, \mathbf{Y}_{i,m+1}, \dots, \mathbf{Y}_{iK}\}$ is outcomes observed since time m in cluster i . If unstated, the notation in formulation and figures is for one cluster and therefore omits the subscript i for cluster. The subscript m for time is omitted in examples with single time point. Unbold A_{im} , L_{im} , Y_{im} represent scalar treatment, covariate and outcome, respectively for subject i at time m in situations without clustered outcomes, which we use to reference previous literature on SNMMs.

To define causal effects and structural models, we will use potential or counterfactual outcomes that are outcomes that would be observed were the subject to receive a certain treatment [8]. We presume that treatments at time m do not affect outcomes prior to time m . Potential outcomes are considered under a regime, which is defined as a sequential set of treatments $\bar{\mathbf{a}}_K = \{\mathbf{a}_1, \dots, \mathbf{a}_K\}$. In formulating SNMMs, the regime $(\bar{\mathbf{a}}_m, \mathbf{0})$, the counterfactual treatment regime that agrees with $\bar{\mathbf{a}}_m$ through time m and is $\mathbf{0}$ thereafter, is of particular interest. Let $\mathbf{Y}_{ik}^{\bar{\mathbf{a}}_m, \mathbf{0}} = \{Y_{ik1}^{\bar{\mathbf{a}}_m, \mathbf{0}}, \dots, Y_{ikJ}^{\bar{\mathbf{a}}_m, \mathbf{0}}\}$ denote the potential outcome in cluster i that would be observed at time k where $k > m$ were the cluster to receive treatment history $\bar{\mathbf{a}}_m$ through time m and $\mathbf{0}$ thereafter. Let $\underline{\mathbf{Y}}_{i,m+1}^{\bar{\mathbf{a}}_m, \mathbf{0}} = \{\mathbf{Y}_{i,m+1}^{\bar{\mathbf{a}}_m, \mathbf{0}}, \dots, \mathbf{Y}_{iK}^{\bar{\mathbf{a}}_m, \mathbf{0}}\}$ denote the series of potential outcomes in cluster i that would be observed after time m . Causal effects can be defined as contrasts of potential outcomes for the same group of subjects under different treatment regimes. For instance, $\mathbf{Y}_{im}^{\bar{\mathbf{a}}_K} - \mathbf{Y}_{im}^{\bar{\mathbf{a}}'_K}$, where $\bar{\mathbf{a}}_K \neq \bar{\mathbf{a}}'_K$, denotes the causal effect of treatment $\bar{\mathbf{a}}_K$ versus $\bar{\mathbf{a}}'_K$ on a subject's outcome at time m .

3 Structural models and estimation for point treatments in clustered observations

Prior to describing the formulation of SNMMs for clustered observations, we review the general concepts of SNMMs for nonclustered outcomes. SNMMs model the contrast between conditional means of potential outcomes $\gamma_m^*(\bar{l}_m, \bar{a}_m) = g\left(E\left(Y_{m+1}^{a_m, 0} | \bar{L}_m = \bar{l}_m, \bar{A}_m = \bar{a}_m\right)\right) - g\left(E\left(Y_{m+1}^{a_{m-1}, 0} | \bar{L}_m = \bar{l}_m, \bar{A}_m = \bar{a}_m\right)\right)$, where g can be any monotone link function. Frequently used link functions are the identity ($g(x) = x$), loglinear ($g(x) = \log(x)$), or logistic ($g(x) = \text{logit}(x)$) [1, 9]. This contrast represents the sequential removal of the effect of a treatment a_m at time m on the means of subsequent potential outcomes after having removed the effect of all later treatments, a process called “blipping down” [1]. A function $\gamma_m^*(\bar{l}_m, \bar{a}_m)$ that defines such a contrast is called a blip function. It can be parameterized with finite dimensional parameter ψ .

We first formulate SNMMs for clustered outcomes in a single time point example, where treatments, covariates and outcomes are measured at one time point only. In this example, we assume that each cluster consists of two units, and consider unit-level treatment only. The results can be easily generalized to situations with more than two units or with cluster-level treatment. A directed acyclic graph (DAG) is shown in Figure 1. The subscript here refers to unit number. $\mathbf{L} = \{L_1, L_2, L^\dagger\}$ is the set of cluster- and unit-level measured confounder for the effect of joint treatment $\mathbf{A} = \{A_1, A_2\}$ on clustered outcomes $\mathbf{Y} = \{Y_1, Y_2\}$. We assume that there is no unmeasured confounding between $\mathbf{A}$ and $\mathbf{Y}$. Dashed double arrows represent possible unmeasured confounding within $\mathbf{L}$, $\mathbf{A}$ or $\mathbf{Y}$ between units. U and dotted arrows represent possible unmeasured confounding between $\mathbf{L}$ and $\mathbf{Y}$. Let $Y_j^{\mathbf{a}} = Y_j^{a_j, a_{j'}}$ denote the potential outcome in j th unit, where a_j and $a_{j'}$ refer to unit-specific treatment to the same unit and the other within-cluster unit respectively, $\{j, j'\} = \{1, 2\}, \{2, 1\}$.

Figure 1: Causal diagram for an example with single time point.

3.1 Structural nested mean models

For binary treatments, we may consider the joint treatment $\{A_1, A_2\}$ across units within a cluster as one combined treatment at the cluster level with 4 levels, namely $\{0, 0\}$, $\{1, 0\}$, $\{0, 1\}$, $\{1, 1\}$. Then a SNMM with joint treatment can be formulated in the same way as one with a single treatment as a contrast between means of potential outcomes under joint treatment versus means of potential outcomes under no treatment as seen in eq. (1), where $g(\cdot)$ is a monotone link function. If there is an additional binary systemic treatment $A^\dagger$, the combined treatment $\{A^\dagger, A_1, A_2\}$ will contain eight levels. In the rest of the paper, we focus on the SNMM with identity link such that $g(x) = x$. The method introduced, however, can be applied to other link functions, which will be discussed in the discussion section:

$$g\left\{E\left(\begin{matrix} Y_1^{a_1, a_2} \\ Y_2^{a_2, a_1} \end{matrix} \mid \mathbf{L} = \mathbf{l}, \mathbf{A} = \mathbf{a}\right)\right\} - g\left\{E\left(\begin{matrix} Y_1^{0,0} \\ Y_2^{0,0} \end{matrix} \mid \mathbf{L} = \mathbf{l}, \mathbf{A} = \mathbf{a}\right)\right\} = \begin{pmatrix} \gamma_1^*(a_1, a_2, \mathbf{l}) \\ \gamma_2^*(a_2, a_1, \mathbf{l}) \end{pmatrix}. \quad (1)$$

The vector-valued blip function $\gamma^*(\cdot)$ models the effect of removing joint treatment $\mathbf{a}$ from both unit-level outcomes within the cluster. It can be parameterized with the finite dimensional parameter ψ to form

$$\begin{pmatrix} \gamma_1^*(a_1, a_2, \mathbf{l}; \psi) \\ \gamma_2^*(a_2, a_1, \mathbf{l}; \psi) \end{pmatrix}. \quad (2)$$

As an example, let

$$\begin{pmatrix} \gamma_1^*(a_1, a_2, \mathbf{l}; \psi) \\ \gamma_2^*(a_2, a_1, \mathbf{l}; \psi) \end{pmatrix} = \begin{pmatrix} \psi_1 a_1 + \psi_2 a_2 + \psi_3 a_1 a_2 \\ \psi_{1'} a_2 + \psi_{2'} a_1 + \psi_{3'} a_2 a_1 \end{pmatrix}. \quad (3)$$

In eq. (3), $\psi \in R^6$; ψ_1 and $\psi_{1'}$ represent the effect of unit-specific treatment on the outcome in the same unit. ψ_2 and $\psi_{2'}$ represent the effect of treatment on the outcome in the opposite unit, ψ_3 and $\psi_{3'}$ represent the interaction between treatments in different units. We can also interpret ψ_1 and $\psi_{1'}$ as the direct effect of treatment on the same unit. ψ_2 represents the spillover effect of unit 2's treatment on unit 1 when $a_1 = 0$, and $\psi_2 + \psi_3$ represents the spillover effect of unit 2's treatment on unit 1 when $a_1 = 1$. Similarly, $\psi_{2'}$ represents the spillover effect of unit 1's treatment on unit 2 when $a_2 = 0$, and $\psi_{2'} + \psi_{3'}$ represents the spillover effect of unit 1's treatment on unit 2 when $a_2 = 1$. This example imposes the assumption that effect modification by covariates is absent.

The dimension of causal parameters will increase sharply with increase in the number of units per cluster. Some additional model assumptions are considered to reduce the dimension of causal parameters. For the ophthalmology example, it is reasonable to assume symmetry between two eyes, namely the effect of the joint treatment is the same for each unit. In eq. (3), the symmetry assumption implies $\psi_1 = \psi_{1'}$, $\psi_2 = \psi_{2'}$, $\psi_3 = \psi_{3'}$. In some cases, we may want to assume no interference, namely unit-specific treatment on one unit has no effect on outcomes in the other units of the same cluster. With two units per cluster, it refers to

$$Y_j^{a_j, a_{j'}} = Y_j^{a_j, 0}.$$

In eq. (3), the no interference assumption implies $\psi_2 = \psi_3 = \psi_{2'} = \psi_{3'} = 0$.

3.2 Identifiability

A1–A4 below are sufficient assumptions to identify the causal effect of joint treatment $\mathbf{a}$ relative to no treatment from observed data and therefore the causal parameter ψ .

A1. Consistency. If $\mathbf{A} = \mathbf{a}$ for a given cluster, then $\mathbf{Y}^{\mathbf{a}} = \mathbf{Y}$ for that cluster.

A2. Conditional exchangeability. $\mathbf{Y}^{0,0} \perp \mathbf{A} \mid \mathbf{L} = \mathbf{l}$.

A3. Positivity. If $f_{\mathbf{L}}(\mathbf{l}) > 0$, then $f_{\mathbf{A}|\mathbf{L}}(\mathbf{a} \mid \mathbf{l}) > 0$ for all treatments $\mathbf{a}$.

A4. Cluster-level SUTVA. The observation on one cluster is unaffected by the particular treatment assignment to the other clusters.

In the presence of interference between units, potential outcomes like $Y_j^{a_j}$ are not well-defined. However, under no interference, it is reasonable to relax A1 and define consistency under the standard assumption for non-clustered data: if $A_j = a_j$ for a given unit j , then $Y_j^{a_j} = Y_j$ for that unit. A2 can be viewed as no unmeasured confounding. It states the independence between treatment-free potential outcomes and the joint treatment to a cluster conditional on the covariates within the cluster. A3 can be relaxed in SNMMs such that the assumption does not need to hold for all $\mathbf{L}$ with $f_{\mathbf{L}}(\mathbf{l}) > 0$. A4 is similar to the partial interference assumption described in Sobel [10]. Note that our method allows for interference between units within the same cluster, but the SUTVA assumption needs to hold for clusters.

3.3 Alternative modeling

An alternative way of modeling the joint effect of $\{a_1, a_2\}$ is to use the concepts in SNMMs with time-varying treatments and sequentially remove the effect of each unit-specific treatment. In the following example, the effect of treatment on the opposite unit is removed first, followed by the effect of treatment on the same unit.

1. Remove the effect of treatment on the opposite unit $a_{j'}$:

$$E(Y_j^{a_j, a_{j'}} | \mathbf{L} = \mathbf{l}, \mathbf{A} = \mathbf{a}) - E(Y_j^{a_j, 0} | \mathbf{L} = \mathbf{l}, \mathbf{A} = \mathbf{a}) = \gamma_{1j}^*(a_{j'}, \mathbf{l}, a_j) \quad (4a)$$

2. Remove the effect of treatment on the same unit a_j :

$$E(Y_j^{a_j, 0} | \mathbf{L} = \mathbf{l}, A_j = a_j) - E(Y_j^{0, 0} | \mathbf{L} = \mathbf{l}, A_j = a_j) = \gamma_{2j}^*(a_j, \mathbf{l}) \quad (4b)$$

For example, the blip functions in eqs (4a) and (4b) can be parameterized as $\gamma_{1j}^*(a_{j'}, \mathbf{l}, a_j; \psi) = \psi_2 a_{j'} + \psi_3 a_{j'} a_j$ and $\gamma_{2j}^*(a_j, \mathbf{l}; \psi) = \psi_1 a_j$, respectively.

Assumptions A1 and A3 plus a sequential ignorability assumption are sufficient to identify the causal effect in eqs (4a) and (4b). Specifically, the sequential ignorability assumption states that $Y_j^{a_j, 0} \perp A_{j'} | \mathbf{L}, A_j$ and $Y_j^{0, 0} \perp A_j | \mathbf{L}$. Testing for no interference under this model is equivalent to testing whether the blip function in eq. (4a) equals 0. In this case, since a_j and $a_{j'}$ are simultaneous treatments, the order of blipping down is arbitrary. Parameters from different orders of blipping down have different interpretations. For this reason, the approach in eq. (1) is used for the rest of the paper. However, the alternative approach may be useful in other settings where the order of blipping down is meaningful. For instance, if clusters are defined by families, it may be reasonable to blip down the treatment on parents first and then that on children or vice versa.

3.4 Estimation

Robins [1] has derived a class of estimating equations for parameters of SNMMs based on semiparametric theory [1]. He defined $H_k(\psi) = Y_k - \sum_{l=0}^{k-1} \gamma_{l,k}^*(\bar{L}_l, \bar{A}_l; \psi)$ as the fully blipped down outcome at time k , and $H^{(m)} = [H_{m+1}(\psi), H_{m+2}(\psi), \dots, H_K(\psi)]^T$ as the vector of blipped down outcomes after time m , where $m = 0, \dots, K-1$. Under the sequential ignorability assumption, $E(H^{(m)} | \bar{L}_m, \bar{A}_m) = E(H^{(m)} | \bar{L}_m, \bar{A}_{m-1})$. $\dot{H}^{(m)}$ is defined as $\dot{H}^{(m)} = H^{(m)} - E(H^{(m)} | \bar{L}_m, \bar{A}_{m-1})$. The efficient estimating function, known as the efficient score [11], under general settings involves complex recursive expressions, but under the assumption that

$$E(H_k H_j | \bar{L}_m, \bar{A}_m) = E(H_k H_j | \bar{L}_m, \bar{A}_{m-1}), \quad (5)$$

for $k > m, j > m, m = 0, \dots, K-1$, the efficient score can be simplified as:

$$S_\psi^{eff} = \sum_{i=1}^N \sum_{m=0}^{K-1} \left[E \left\{ \frac{\partial \dot{H}_i^{(m)}}{\partial \psi^T} | \bar{L}_{im}, \bar{A}_{im} \right\} - E \left\{ \frac{\partial \dot{H}_i^{(m)}}{\partial \psi^T} | \bar{L}_{im}, \bar{A}_{i,m-1} \right\} \right] \cdot \text{Var} \left(\dot{H}_i^{(m)} | \bar{L}_{im}, \bar{A}_{i,m-1} \right)^{-1} \cdot \dot{H}_i^{(m)}. \quad (6)$$

By setting eq. (6) to 0 and solving the estimating equation, one can obtain an efficient estimator for the causal parameter ψ . If the assumption in eq. (5) does not hold, the estimator from eq. (6) will not achieve semiparametric efficiency but will nonetheless be consistent. Equivalently, we can express eq. (6) in terms of the partially blipped down outcome U_m with removal of the effect of treatments from time m onwards [4]. Specifically, U_m is a vector with elements $U_{m,k} = Y_k - \sum_{l=m}^{k-1} \gamma_{l,k}^*(\bar{L}_l, \bar{A}_l)$, for $k = m+1, \dots, K$. U_m mimics the potential outcomes in the sense that $E(U_m | \bar{L}_m, \bar{A}_m) = E(Y_{m+1}^{\bar{a}_{m-1}, 0} | \bar{L}_m, \bar{A}_m)$. The alternative expression is

$$S_{\psi}^{\text{eff}} = \sum_{i=1}^N \sum_{m=0}^{K-1} \left[E \left\{ \frac{\partial U_{im}}{\partial \psi^T} \mid \bar{L}_{im}, \bar{A}_{im} \right\} - E \left\{ \frac{\partial U_{im}}{\partial \psi^T} \mid \bar{L}_{im}, \bar{A}_{i,m-1} \right\} \right] \cdot \text{Var}(U_{im} \mid \bar{L}_{im}, \bar{A}_{i,m-1})^{-1} \cdot \{U_{im} - E(U_{im} \mid \bar{L}_{im}, \bar{A}_{i,m-1})\}. \quad (7)$$

The theory for SNMMs can be directly applied to situations with clustered observations. We first construct blipped down outcomes $\mathbf{U}$. For a single time point example with bivariate outcomes, they can be written as

$$\mathbf{U} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \gamma_1^*(A_1, A_2, \mathbf{L}; \psi) \\ \gamma_2^*(A_2, A_1, \mathbf{L}; \psi) \end{pmatrix}.$$

The following is a parameterized example for blip function under the symmetry assumption:

$$\begin{pmatrix} \gamma_1^*(a_1, a_2, \mathbf{L}; \psi) \\ \gamma_2^*(a_2, a_1, \mathbf{L}; \psi) \end{pmatrix} = \begin{pmatrix} \psi_1 a_1 + \psi_2 a_2 + \psi_3 a_1 a_2 \\ \psi_1 a_2 + \psi_2 a_1 + \psi_3 a_2 a_1 \end{pmatrix}. \quad (8)$$

Based on eq. (7), the efficient score for a single time point example can be written as

$$S_{\psi}^{\text{eff}} = \sum_{i=1}^N \left\{ \frac{\partial \mathbf{U}_i}{\partial \psi^T} - E \left(\frac{\partial \mathbf{U}_i}{\partial \psi^T} \mid \mathbf{L}_i \right) \right\} \cdot \text{Var}(\mathbf{U}_i \mid \mathbf{L}_i)^{-1} \cdot \{\mathbf{U}_i - E(\mathbf{U}_i \mid \mathbf{L}_i)\}. \quad (9)$$

For the example in eq. (8), $\frac{\partial \mathbf{U}}{\partial \psi^T}$ has the following expression:

$$\frac{\partial \mathbf{U}}{\partial \psi^T} = \begin{pmatrix} A_1 & A_2 \\ A_2 & A_1 \\ A_1 A_2 & A_2 A_1 \end{pmatrix}.$$

$E \left(\frac{\partial \mathbf{U}}{\partial \psi^T} \mid \mathbf{L} \right)$ is based on a model for treatment. If treatment probabilities are unknown, a multinomial logistic regression model can be used to estimate the treatment probability in each category of the joint treatment. The estimating equation in eq. (9) is doubly robust, because the estimator for ψ will be consistent if either the model for $E \left(\frac{\partial \mathbf{U}}{\partial \psi^T} \mid \mathbf{L} \right)$ or the model for $E(\mathbf{U} \mid \mathbf{L})$ is correctly specified. When both are correctly specified and the conditional variance $\text{Var}(\mathbf{U} \mid \mathbf{L})$ is also correctly specified, the estimator achieves the semiparametric efficiency bound for model (2) if assumption (5) is true. However, correct specification of $\text{Var}(\mathbf{U} \mid \mathbf{L})$ is not required for the consistency of the estimator.

4 Structural models and estimation for time-varying treatments in clustered observations

We next extend the method to time-varying treatments. Suppose that treatments and covariates are measured repeatedly at times $m = 0, \dots, K-1$, and outcomes are measured only at the last time point K . Figure 2 is a causal diagram that shows the effect of a joint treatment from two arbitrary time points $m-1$ and m on the clustered outcomes at time $m+1$. The cluster-specific covariate $L_m^\dagger$ is omitted from the graph for simplicity. In this example, $\mathbf{L}_m = \{L_{m,1}, L_{m,2}\}$ is a time-varying confounder for the effect of $\mathbf{A}_m = \{A_{m,1}, A_{m,2}\}$ on $\mathbf{Y} = \{Y_1, Y_2\}$. Therefore, they need to be controlled for in analysis. $\mathbf{L}_m$, however, also mediates the effect of earlier treatments $\mathbf{A}_{m-1}$ on $\mathbf{Y}$, which suggests that analysis should not control for $\mathbf{L}_m$. In addition, $\mathbf{L}_m$ is a collider on the pathway from $\mathbf{A}_{m-1}$ to $\mathbf{Y}$. Controlling for $\mathbf{L}_m$ will thus induce bias in estimating the causal effect. For these reasons, conventional regression methods will fail in the presence of time-varying confounding [12].

Figure 2: Causal diagram for an example with time-varying treatments.

4.1 Structural nested mean models

In the SNMM below, the blip function vector $\gamma_m^*(\cdot)$ models the effect of removing joint treatment $\mathbf{a}_m$ at time m from both outcomes, after having removed the effect of all subsequent treatments.

$$E\left(\begin{matrix} Y_{1,m}^{\mathbf{a}_m,0} \\ Y_{2,m}^{\mathbf{a}_m,0} \end{matrix} \middle| \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m = \bar{\mathbf{a}}_m\right) - E\left(\begin{matrix} Y_{1,m-1}^{\mathbf{a}_m,0} \\ Y_{2,m-1}^{\mathbf{a}_m,0} \end{matrix} \middle| \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m = \bar{\mathbf{a}}_m\right) = \begin{pmatrix} \gamma_{m,1}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \\ \gamma_{m,2}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \end{pmatrix}. \quad (10)$$

For repeatedly measured outcomes, eq. (10) can be further generalized to

$$E\left(\mathbf{Y}_{m+1}^{\bar{\mathbf{a}}_m,0} \middle| \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m = \bar{\mathbf{a}}_m\right) - E\left(\mathbf{Y}_{m+1}^{\bar{\mathbf{a}}_m-1,0} \middle| \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m = \bar{\mathbf{a}}_m\right) = \gamma_m^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi), \quad (11)$$

where γ_m^* is a vector with components $\gamma_{m,k}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi)$ that models the effect of removing joint treatment at time m from both outcomes at time k , for $k = m+1, \dots, K$. A parameterized example for eq. (10) is shown below:

$$\begin{pmatrix} \gamma_{m,1}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \\ \gamma_{m,2}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \end{pmatrix} = \begin{pmatrix} \psi_{m,1}a_{m,1} + \psi_{m,2}a_{m,2} + \psi_{m,3}a_{m,1}a_{m,2} \\ \psi_{m,1}'a_{m,2} + \psi_{m,2}'a_{m,1} + \psi_{m,3}'a_{m,2}a_{m,1} \end{pmatrix}. \quad (12)$$

The ψ parameters in eq. (12) have similar interpretation as those in example (3), except that they contain subscript m to indicate the effect of joint treatment at time m is time-specific. A more parsimonious example is shown below:

$$\begin{pmatrix} \gamma_{m,1}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \\ \gamma_{m,2}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \end{pmatrix} = \begin{pmatrix} \psi_1a_{m,1} + \psi_2a_{m,2} + \psi_3a_{m,1}a_{m,2} \\ \psi_1a_{m,2} + \psi_2a_{m,1} + \psi_3a_{m,2}a_{m,1} \end{pmatrix}. \quad (13)$$

Model (13) makes the symmetry assumption and also assumes the effect of treatments is the same for each time point. Notice that subscript m is omitted for ψ . The ψ parameters can be interpreted as effect of treatment per unit time. For example, ψ_1 represents the effect of unit-specific treatment on the outcome in the same unit per unit time.

The sequential ignorability assumption is required to identify the contrast in eqs (10) and (11). It is a generalization of assumption A2 to a series of time-varying treatments as shown below:

$$\mathbf{Y}_{m+1}^{\bar{\mathbf{a}}_m-1,0} \perp \mathbf{A}_m \mid \bar{\mathbf{A}}_{m-1} = \bar{\mathbf{a}}_{m-1}, \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m.$$

It states that there is no unmeasured confounding between the partially blipped down potential outcomes and the joint treatment to a cluster at time m , conditional on the covariate and treatment history within the cluster.

4.2 Estimation

Similar to the single time point case, we first construct a series of partially blipped down outcomes $\mathbf{U}_m$ for each time m :

$$\mathbf{U}_m = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \sum_{l=m}^{K-1} \begin{pmatrix} \gamma_{l,1}^*(\bar{\mathbf{a}}_l, \bar{\mathbf{l}}_l; \psi) \\ \gamma_{l,2}^*(\bar{\mathbf{a}}_l, \bar{\mathbf{l}}_l; \psi) \end{pmatrix}.$$

The partially blipped down outcomes mimic the potential outcomes $\mathbf{Y}^{\bar{\mathbf{a}}_{m-1}, \mathbf{0}}$ in the sense that $E(\mathbf{U}_m | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m) = E(\mathbf{Y}^{\bar{\mathbf{a}}_{m-1}, \mathbf{0}} | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m)$. For repeatedly measured outcomes, $\mathbf{U}_m$ is a vector with components

$$\mathbf{U}_{m,k} = \begin{pmatrix} Y_{k,1} \\ Y_{k,2} \end{pmatrix} - \sum_{l=m}^{k-1} \begin{pmatrix} \gamma_{l,k,1}^*(\bar{\mathbf{a}}_l, \bar{\mathbf{l}}_l; \psi) \\ \gamma_{l,k,2}^*(\bar{\mathbf{a}}_l, \bar{\mathbf{l}}_l; \psi) \end{pmatrix},$$

for $k = m + 1, \dots, K$.

For time-varying treatments, we suggest using the following score:

$$S_\psi = \sum_{i=1}^N \sum_{m=0}^{K-1} \left[\frac{\partial \gamma_m^*(\bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{im}; \psi)}{\partial \psi^T} - E \left\{ \frac{\partial \gamma_m^*(\bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{im}; \psi)}{\partial \psi^T} \middle| \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1} \right\} \right] \cdot \text{Var} \left(\mathbf{U}_{im} \middle| \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1} \right)^{-1} \cdot \left\{ \mathbf{U}_{im} - E \left(\mathbf{U}_{im} \middle| \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1} \right) \right\}. \quad (14)$$

The efficient score in eq. (6) is usually too complicated to be used for situations with time-varying treatments. The estimator from the score in eq. (14) is not optimal but is reasonably efficient and doubly robust. It is straightforward to show that the estimator is consistent, if for each time m , either the model for treatment, $E \left\{ \frac{\partial \gamma_m^*(\bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{im}; \psi)}{\partial \psi^T} \middle| \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1} \right\}$, or the model for outcome, $E(\mathbf{U}_m | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$, is correctly specified.

A simpler non-doubly robust estimating equation is

$$S_\psi = \sum_{i=1}^N \sum_{m=0}^{K-1} \left[\frac{\partial \gamma_m^*(\bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{im}; \psi)}{\partial \psi^T} - E \left\{ \frac{\partial \gamma_m^*(\bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{im}; \psi)}{\partial \psi^T} \middle| \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1} \right\} \right] \cdot \mathbf{U}_{im}. \quad (15)$$

It does not involve modeling $E(\mathbf{U}_m | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$ and assumes it to be 0. Therefore, the estimator from eq. (15) is consistent if and only if the model for treatment is correct.

For the parameterized example in eq. (13), we have the following expression for $\frac{\partial \gamma_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \psi)}{\partial \psi^T}$:

$$\frac{\partial \gamma_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \psi)}{\partial \psi^T} = \begin{pmatrix} A_{m,1} & A_{m,2} \\ A_{m,2} & A_{m,1} \\ A_{m,1}A_{m,2} & A_{m,2}A_{m,1} \end{pmatrix}.$$

Similarly, $E \left\{ \frac{\partial \gamma_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \psi)}{\partial \psi^T} \middle| \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1} \right\}$ is based on a model for treatment. If treatment probabilities are unknown, a multinomial logistic regression model can be used to estimate treatment probability in each category of the combined treatment $\{A_{m,1}, A_{m,2}\}$ received at time m . We can either fit a separate model for each time m , or fit a single model to pooled data from all person times. For the pooled treatment model, we may allow intercepts to be time-varying with the use of a spline.

4.3 Closed form estimator

If $\text{Var}\left(\mathbf{U}_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1}\right)$ is given, and a linear working model $\mathbf{D}_m \delta = \mathbf{d}_m(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1}) \delta$ is assumed for $E\left(\mathbf{U}_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1}\right)$, where δ is the q -dimensional nuisance parameter in the outcome model and $\mathbf{d}_m$ is a $J \times q$ matrix-valued function of $\{\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1}\}$, and the blip function is linear in ψ with $\gamma_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m) = \mathbf{R}_m \psi = \mathbf{r}_m(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m)$, where ψ is p -dimensional and $\mathbf{r}_m$ is a $J \times p$ matrix-valued function of $\{\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m\}$, the estimator for ψ will have the following closed form (based on Robins and Hernán [12]):

$$\begin{pmatrix} \hat{\psi} \\ \hat{\delta} \end{pmatrix} = \left\{ \sum_{i=1}^N \sum_{m=0}^{K-1} \begin{pmatrix} \mathbf{C}_{im} \\ \mathbf{D}_{im} \end{pmatrix}^T \cdot \text{Var}\left(\mathbf{U}_{im} \mid \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1}\right)^{-1} \cdot \begin{pmatrix} \sum_{l=m}^{K-1} \mathbf{R}_{il} & \mathbf{D}_{im} \end{pmatrix} \right\}^{-1} \cdot \left\{ \sum_{i=1}^N \sum_{m=0}^{K-1} \begin{pmatrix} \mathbf{C}_{im} \\ \mathbf{D}_{im} \end{pmatrix}^T \cdot \text{Var}\left(\mathbf{U}_{im} \mid \bar{\mathbf{L}}_{im}, \bar{\mathbf{A}}_{i,m-1}\right)^{-1} \cdot \mathbf{Y}_i \right\}, \quad (16)$$

where $\mathbf{C}_m = \sum_{l=m}^{K-1} \left[\mathbf{R}_l - E\left\{ \mathbf{R}_l \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1} \right\} \right]$.

A simple working model for $\text{Var}(\mathbf{U}_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$ is $\sigma^2 I_{J \times J}$ for some constant variance σ^2 at all times, where $I_{J \times J}$ is an identity matrix of dimension J . However, the independence structure and constant variance over time are often far from truth. A simulation study described in the next section will examine the impact of such misspecification on the efficiency of $\hat{\psi}$. If one assumes a working model other than independence structure and constant variance, estimation will then require iterative procedures similar to those in generalized estimating equations (GEEs). We can start by assuming working model $\sigma^2 I_{J \times J}$ to obtain an initial estimate for ψ , and then iterate between estimating parameters in $\text{Var}(\mathbf{U}_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$ and estimating ψ until convergence is achieved. We adapted the method of moments approach by Liang and Zeger [13] in estimating variance parameters [13]. An asymptotic variance estimator for $\hat{\psi}$ may be derived using the sandwich method, with the exact form found in Appendix A. Like GEE, $\text{Var}(\mathbf{U}_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$ does not need to be correctly specified for the estimator for ψ to be consistent.

5 Simulation study

5.1 Algorithm

The data generating process is based on the likelihood for SNMMs described in Robins et al. [9]. Details about the likelihood can be found in Appendix B. A simulation study is conducted based on the setting where treatments and covariates are measured repeatedly at times $m = 0, \dots, K-1$, and outcomes are measured at the last time point K only. The number of time points is $K = 5$, and the cluster size is $J = 2$. The total number of clusters is varied as $N = 100$ or $1,000$. The algorithm for generating data for one cluster is shown below. It includes two examples with different covariate effects in the conditional mean of outcomes. Variables with subscript $m = -1$ assume value 0.

For $m = 0, \dots, K-1$, we generated covariates $\mathbf{L}_m$ from a multivariate normal distribution:

$$\begin{pmatrix} L_{m,1} \\ L_{m,2} \\ L_m^\dagger \end{pmatrix} \mid \bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1} \sim N_3 \left(\begin{pmatrix} 0.3L_{m-1,1} + 0.2A_{m-1,1} \\ 0.3L_{m-1,2} + 0.2A_{m-1,2} \\ 0.3L_{m-1} + 0.1A_{m-1,1} + 0.1A_{m-1,2} \end{pmatrix}, \begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix} \right).$$

We generated joint treatment $\mathbf{A}_m$ from a multinomial logistic regression model with probabilities $p_{m,00}$, $p_{m,10}$, $p_{m,01}$, $p_{m,11}$:

$$\begin{pmatrix} A_{m,1} \\ A_{m,2} \end{pmatrix} \middle| \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1} \sim \text{multinomial}(p_{m,00}, p_{m,10}, p_{m,01}, p_{m,11}),$$

where

$$p_{m,10} = \text{mlogit}^{-1}(0.5L_{m,1} - 0.5L_m^\dagger + 0.2A_{m-1,1}),$$

$$p_{m,01} = \text{mlogit}^{-1}(0.5L_{m,2} - 0.5L_m^\dagger + 0.2A_{m-1,2}),$$

$$p_{m,11} = \text{mlogit}^{-1}(0.5L_{m,1} + 0.5L_{m,2} - L_1^\dagger + 0.2A_{m-1,1} + 0.2A_{m-1,2}),$$

$$p_{m,00} = 1 - p_{m,01} - p_{m,10} - p_{m,11}.$$

Last, observed outcomes $\mathbf{Y}_K$ were generated. We first generated the conditional mean of $\mathbf{Y}_K$ under the sequential ignorability assumption:

$$E(\mathbf{Y}_K \mid \bar{\mathbf{l}}_{K-1}, \bar{\mathbf{a}}_{K-1}) = \sum_{m=0}^{K-1} \gamma_m^*(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m) + \sum_{m=1}^{K-1} \left[E(\mathbf{Y}_K^{\mathbf{a}_{m-1},0} \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1}) - E(\mathbf{Y}_K^{\mathbf{a}_{m-1},0} \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}) \right] + E(\mathbf{Y}_K^{\mathbf{0}} \mid \mathbf{l}_0).$$

For both simulation examples,

$$\sum_{m=0}^{K-1} \gamma_m^*(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m) = \sum_{m=0}^{K-1} \left(\psi_1 a_{m,1} + \psi_2 a_{m,2} + \psi_3 a_{m,1} a_{m,2} + \psi_4 a_{m,1} l_{m,1} + \psi_5 a_{m,1} l_m^\dagger \right),$$

where $\psi_1 = 2$, $\psi_2 = 0.5$, $\psi_3 = 0.2$, $\psi_4 = 0.5$, $\psi_5 = -0.2$, and

$$E(\mathbf{Y}_K^{\mathbf{0}} \mid \mathbf{l}_0) = \begin{pmatrix} 2l_{0,1} - l_0^\dagger \\ 2l_{0,2} - l_0^\dagger \end{pmatrix}.$$

The two examples differ in the association between covariates and the potential outcomes. In Example 1, the association is linear:

$$\begin{aligned} & \sum_{m=1}^{K-1} \left[E(\mathbf{Y}_K^{\mathbf{a}_{m-1},0} \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1}) - E(\mathbf{Y}_K^{\mathbf{a}_{m-1},0} \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}) \right] \\ &= \sum_{m=1}^{K-1} \left\{ \begin{pmatrix} l_{m,1} - l_m^\dagger \\ l_{m,2} - l_m^\dagger \end{pmatrix} - E \left[\begin{pmatrix} l_{m,1} - l_m^\dagger \\ l_{m,2} - l_m^\dagger \end{pmatrix} \middle| \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1} \right] \right\}. \end{aligned} \quad (17)$$

In Example 2, the association is quadratic:

$$\begin{aligned} & \sum_{m=1}^{K-1} \left[E(\mathbf{Y}_K^{\mathbf{a}_{m-1},0} \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1}) - E(\mathbf{Y}_K^{\mathbf{a}_{m-1},0} \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}) \right] \\ &= \sum_{m=1}^{K-1} \left\{ \begin{pmatrix} (l_{m,1} - l_m^\dagger)^2 \\ (l_{m,2} - l_m^\dagger)^2 \end{pmatrix} - E \left[\begin{pmatrix} (l_{m,1} - l_m^\dagger)^2 \\ (l_{m,2} - l_m^\dagger)^2 \end{pmatrix} \middle| \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1} \right] \right\}. \end{aligned} \quad (18)$$

$\mathbf{Y}_K$ was generated from its conditional mean plus a bivariate normal error term:

$$\mathbf{Y}_K = E(\mathbf{Y}_K \mid \bar{\mathbf{l}}_{K-1}, \bar{\mathbf{a}}_{K-1}) + \varepsilon,$$

where the error term also follows multivariate normal distribution:

$$\varepsilon \left| \bar{\mathbf{L}}_{K-1}, \bar{\mathbf{A}}_{K-1} \sim N_2 \left(\mathbf{0}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \right).$$

In addition, we explored the impact of varying the correlation between unit-specific treatments in the same cluster on the estimates. The simulation procedure is the same as Example 1 with the number of clusters as $N = 1,000$, except that simpler SNMMs with fewer parameters are used, as shown in eqs (19) and (20), and the correlation between unit-specific treatments in the same cluster is varied between 0, 0.9 and -0.9 . We tested one example without interference and one with interference respectively. In the example without interference, the SNMM contains the effect of treatment in the same unit and the effect modification by unit-specific covariate in the same unit:

$$\gamma_m^*(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m) = \begin{pmatrix} \psi_1 a_{m,1} + \psi_2 a_{m,1} l_{m,1} \\ \psi_1 a_{m,2} + \psi_2 a_{m,2} l_{m,2} \end{pmatrix}, \quad (19)$$

where $\psi_1 = 2, \psi_2 = 0.5$. In the example with interference, the SNMM contains the effect of treatment in the same unit, the effect of treatment in the opposite unit, and the interaction between treatments:

$$\gamma_m^*(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m) = \begin{pmatrix} \psi_1 a_{m,1} + \psi_2 a_{m,2} + \psi_3 a_{m,1} a_{m,2} \\ \psi_1 a_{m,2} + \psi_2 a_{m,1} + \psi_3 a_{m,2} a_{m,1} \end{pmatrix}, \quad (20)$$

where $\psi_1 = 2, \psi_2 = 0.5, \psi_3 = 0.2$.

Last, we explored the impact of a misspecified SNMM on the estimates. For this simulation, we used the SNMM with interference in eq. (20) as the true model, and fit a misspecified SNMM containing the effect of treatment in the same unit only.

5.2 Results

Two doubly robust (DR) estimators eq. (14) and one without the doubly robust property eq. (15) were computed using the simulated data (Example 1). For both DR estimators, we assume a linear working model for $E(\mathbf{U}_m | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$. For $Var(\mathbf{U}_m | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$, we either assume an exchangeable working model $\Sigma_m = \sigma_m^2 \begin{pmatrix} 1 & \rho_m \\ \rho_m & 1 \end{pmatrix}$ for each time m and denote the estimator as DR1, or a working model with independence structure and constant variance $\Sigma = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for all time points and denote the estimator as DR2. The DR2 estimator has the closed form described in eq. (16), but not the DR1 estimator. The non-DR estimator is based on the score in eq. (15).

To estimate treatment probabilities, a multinomial logistic regression model is fit to pooled data from all time points. Simulation and estimation procedures were repeated for 1,000 replicates. The results are shown in Table 1. Averaged estimates (Ave EST), empirical standard errors (Emp SE), and averaged asymptotic standard errors (Ave SE) are provided. The square root of mean squared error ($\sqrt{\text{MSE}}$) is computed from Ave EST and Emp SE. All three estimators have negligible bias. Coverage of all 95% confidence intervals (CI) at $N = 1000$ show the expected range of 93% – 96%. At $N = 100$, all 95% CIs show under-coverage, likely due to the small sample size. The DR1 estimator has the worst coverage probabilities at $N = 100$, but it is the most efficient among the three estimators. The relative $\sqrt{\text{MSE}}$ of DR2 and non-DR estimators relative to DR1 estimator ranges from 1.3 to 1.4 and 2.4 to 3.4, respectively. As seen in Table 2, for the DR1 estimator, the estimates of the variance parameter σ_m^2 and correlation parameter ρ_m in the conditional variance of $\mathbf{U}_m$ decrease over time, since more of the variance is accounted for by the increased treatment and covariate history in the conditioning. Therefore, a flexible working model for $Var(\mathbf{U}_m | \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1})$ used in DR1 estimator tends to improve efficiency.

Table 1: Small and large sample performance of g-estimators in a simulation study based on Example 1.

N	Param	True	DR1 estimator					DR2 estimator					Non-DR estimator				
			Ave EST	Emp SE	Ave SE	95% COV	$\sqrt{\text{MSE}}$	Ave EST	Emp SE	Ave SE	95% COV	Rel $\sqrt{\text{MSE}}$	Ave EST	Emp SE	Ave SE	95% COV	Rel $\sqrt{\text{MSE}}$
100	ψ_1	2.0	2.00	0.18	0.13	0.83	0.18	1.98	0.24	0.19	0.88	1.36	1.99	0.45	0.39	0.91	2.53
	ψ_2	0.5	0.50	0.17	0.13	0.86	0.17	0.48	0.24	0.19	0.88	1.36	0.50	0.44	0.38	0.92	2.52
	ψ_3	0.2	0.20	0.26	0.19	0.84	0.26	0.21	0.36	0.28	0.87	1.37	0.18	0.66	0.58	0.92	2.53
	ψ_4	0.5	0.50	0.11	0.08	0.84	0.11	0.49	0.16	0.12	0.86	1.41	0.48	0.35	0.31	0.93	3.06
	ψ_5	-0.2	-0.20	0.11	0.08	0.85	0.11	-0.18	0.16	0.12	0.88	1.41	-0.17	0.36	0.31	0.90	3.17
1,000	ψ_1	2.0	2.00	0.05	0.05	0.95	0.05	2.00	0.07	0.07	0.95	1.35	2.00	0.14	0.13	0.95	2.64
	ψ_2	0.5	0.50	0.05	0.05	0.95	0.05	0.50	0.07	0.07	0.95	1.32	0.50	0.13	0.13	0.95	2.64
	ψ_3	0.2	0.20	0.07	0.07	0.95	0.07	0.20	0.10	0.10	0.96	1.34	0.20	0.20	0.20	0.94	2.74
	ψ_4	0.5	0.50	0.03	0.03	0.94	0.03	0.50	0.04	0.04	0.93	1.32	0.50	0.11	0.11	0.94	3.34
	ψ_5	-0.2	-0.20	0.03	0.03	0.95	0.03	-0.20	0.05	0.04	0.94	1.41	-0.20	0.11	0.11	0.94	3.38

Notes: DR1 estimator and DR2 estimator are doubly robust, whereas non-DR estimator is not. DR1 uses a flexible working model for the covariance, whereas DR2 estimator assumes independence and constant variance over time. Ave EST refers to average estimates, Emp SE refers to empirical standard error, Ave SE refers to average asymptotic standard error, 95% COV refers to 95% coverage probability, $\sqrt{\text{MSE}}$ refers to square root of mean squared error, and Rel $\sqrt{\text{MSE}}$ refers to the ratio of square root of mean squared error relative to DR1 estimator.

Table 2: Estimates for variance parameter σ_m^2 and correlation parameter ρ_m at each time point for DR1 estimator from *Example 1*.

Time (m)	θ_m^2		$\hat{\rho}_m$	
	Ave EST	Emp SE	Ave EST	Emp SE
0	7.76	0.38	0.51	0.03
1	6.06	0.29	0.50	0.03
2	4.37	0.19	0.47	0.03
3	2.69	0.11	0.42	0.03
4	1.03	0.03	0.19	0.03

Notes: Sample size $N = 1,000$. σ_m^2 and ρ_m are estimated by the method of moments approach by Liang and Zeger [13]. AVE EST refers to average estimates, and EMP SE refers to empirical standard error.

We varied the correlation between unit-specific treatments in the same cluster given predictors in the treatment model and explored its impact on the estimates. Table 3(A) shows the results from the example with no interference. A large positive correlation of 0.9 between unit-specific treatments increased the standard errors of the estimates slightly. The relative $\sqrt{\text{MSE}}$ compared to no correlation ranges from 1.06 to 1.27. In contrast, a large negative correlation of -0.9 decreased the standard errors of the estimates slightly. The relative $\sqrt{\text{MSE}}$ compared to no correlation ranges from 0.67 to 0.87. Table 3(B) shows the results from the example with interference. In this case, both a large positive correlation of 0.9 and a large negative correlation of -0.9 between unit-specific treatments greatly increased the standard errors of the estimates. The relative $\sqrt{\text{MSE}}$ compared to no correlation ranges from 2.00 to 2.65. Therefore, varying the correlation between unit-specific treatments has a greater impact on the efficiency of the estimators in settings with interference. In such cases, when the unit-specific treatments in the same cluster are highly correlated

Table 3: Examine the impact of the correlation between unit-specific treatments in the same cluster sample size $N = 1,000$.

		$\rho_A = 0$				$\rho_A \approx 0.9$			$\rho_A \approx -0.9$		
	Param	True	Ave EST	Emp SE	$\sqrt{\text{MSE}}$	Ave EST	Emp SE	Rel $\sqrt{\text{MSE}}$	Ave EST	Emp SE	Rel $\sqrt{\text{MSE}}$
A: An example with no interference											
DR1	ψ_1	2.0	2.00	0.03	0.03	2.00	0.04	1.27	2.00	0.03	0.83
	ψ_2	0.5	0.50	0.03	0.03	0.50	0.03	1.06	0.50	0.03	0.85
DR2	ψ_1	2.0	2.00	0.04	0.04	2.00	0.05	1.21	2.00	0.03	0.67
	ψ_2	0.5	0.50	0.04	0.04	0.50	0.05	1.10	0.50	0.04	0.86
Non-DR	ψ_1	2.0	2.00	0.08	0.08	2.00	0.10	1.26	2.00	0.06	0.82
	ψ_2	0.5	0.50	0.10	0.10	0.50	0.11	1.09	0.50	0.09	0.87
B: An example with interference											
DR1	ψ_1	2.0	2.00	0.05	0.05	2.00	0.12	2.40	2.00	0.13	2.55
	ψ_2	0.5	0.50	0.05	0.05	0.50	0.12	2.34	0.50	0.13	2.56
	ψ_3	0.2	0.20	0.07	0.07	0.20	0.19	2.60	0.20	0.17	2.34
DR2	ψ_1	2.0	2.00	0.07	0.07	2.00	0.15	2.16	2.00	0.16	2.44
	ψ_2	0.5	0.50	0.07	0.07	0.50	0.15	2.23	0.51	0.16	2.53
	ψ_3	0.2	0.20	0.10	0.10	0.21	0.25	2.52	0.19	0.22	2.24
Non-DR	ψ_1	2.0	2.00	0.12	0.12	2.00	0.27	2.30	2.02	0.26	2.19
	ψ_2	0.5	0.50	0.12	0.12	0.50	0.27	2.22	0.52	0.26	2.18
	ψ_3	0.2	0.20	0.19	0.19	0.20	0.49	2.65	0.18	0.37	2.00

Notes: Ave EST refers to average estimates, Emp SE refers to empirical standard error, $\sqrt{\text{MSE}}$ refers to square root of mean squared error, and Rel $\sqrt{\text{MSE}}$ refers to the ratio of square root of mean squared error relative to no correlation.

either negatively or positively, the standard errors of the estimates will be inflated because of the multicollinearity problem. This will affect the power of the study.

For the example with interference in Table 3(B), we also fit a misspecified SNMM with no interference and examined its impact on the estimates. The results are shown in Table 4. The estimates from fitting a misspecified SNMM are biased. In this case, the estimates from the DR1 approach that assumes non-independence working correlation have larger bias than the estimates from the DR2 approach that assume independence working correlation and the non-DR estimates. Therefore, it is not recommended to use the DR1 approach when one speculates there is interference within the cluster but cannot fit a SNMM with interference for practical reasons. The $\sqrt{\text{MSE}}$ of these estimates are also larger than those in Table 3(B) because of bias.

Table 4: Examine the impact of a misspecified SNMM on the estimates.

	Param	True	Ave EST	Emp SE	Ave SE	95% COV	$\sqrt{\text{MSE}}$
DR1	ψ_1	2.00	1.85	0.03	0.03	0.01	0.16
DR2	ψ_1	2.00	2.09	0.04	0.04	0.41	0.10
Non-DR	ψ_1	2.00	2.09	0.07	0.07	0.73	0.12

Notes: The true SNMM contains interference, and the estimates from fitting a true model are shown in Table 3(B). The misspecified SNMM only contains the effect of treatment in the same unit. Sample size $N = 1,000$. Ave EST refers to average estimates, Emp SE refers to empirical standard error, Ave SE refers to average asymptotic standard error, 95% COV refers to 95% coverage probability, and $\sqrt{\text{MSE}}$ refers to square root of mean squared error.

We did a comparison between g-estimation of SNMMs and standard GEE regression, a conventional method to deal with clustered outcomes. We considered two simulation examples with different covariate effects as shown in eqs (17) and (18). Both SNMM and GEE regression approaches were applied to the simulated data. Predictors in restricted mean models for GEE include all factors in the blip functions as well as all covariates used in generating the conditional mean of outcomes. An exchangeable working model is assumed for the correlation structure. The conditional mean models used in the GEE approach in the two examples are

$$Y_j \sim \sum_{m=0}^{K-1} A_{mj} + \sum_{m=0}^{K-1} A_{mj'} + \sum_{m=0}^{K-1} A_{mj} A_{mj'} + \sum_{m=0}^{K-1} A_{mj} L_{mj} + \sum_{m=0}^{K-1} A_{mj} L_m + \sum_{m=0}^{K-1} L_{mj} + \sum_{m=0}^{K-1} L_m,$$

and

$$Y_j \sim \sum_{m=0}^{K-1} A_{mj} + \sum_{m=0}^{K-1} A_{mj'} + \sum_{m=0}^{K-1} A_{mj} A_{mj'} + \sum_{m=0}^{K-1} A_{mj} L_{mj} + \sum_{m=0}^{K-1} A_{mj} L_m + L_{0j} + L_0 + \sum_{m=1}^{K-1} L_{mj}^2 + \sum_{m=1}^{K-1} L_m^2 + \sum_{m=1}^{K-1} L_{mj} \cdot L_m$$

respectively. The models included correctly specified predictors, but did not properly control for time-varying confounding. The coefficients corresponding to treatment and effect modification are shown in Table 5. As expected, the DR1 estimators from SNMMs and g-estimation are unbiased, whereas bias is induced in GEE estimators by controlling for covariates affected by earlier treatments. In the first example, the GEE estimators are in fact more efficient than the DR1 estimators. However, the first two GEE estimators are biased. In the second example, the GEE estimators have larger standard errors compared to the DR1 estimators, and all of them are biased.

Table 5: Compare the estimators obtained from SNMM and g-estimation to those from GEE linear regression using two simulation examples.

Example	Param	True	DR1 estimator		GEE estimator		
			Bias	Emp SE	Bias	Emp SE	REL $\sqrt{\text{MSE}}$
1	ψ_1	2.0	−0.00	0.05	−0.08	0.03	1.69
	ψ_2	0.5	−0.00	0.05	0.08	0.03	1.71
	ψ_3	0.2	0.00	0.07	0.00	0.04	0.61
	ψ_4	0.5	0.00	0.03	0.00	0.02	0.56
	ψ_5	−0.2	−0.00	0.03	0.00	0.02	0.62
2	ψ_1	2.0	−0.00	0.06	−0.17	0.10	3.07
	ψ_2	0.5	−0.00	0.07	−0.02	0.09	1.46
	ψ_3	0.2	0.00	0.09	0.07	0.13	1.58
	ψ_4	0.5	0.00	0.04	0.14	0.06	3.56
	ψ_5	−0.2	0.00	0.04	−0.31	0.06	7.25

Notes: Sample size $N = 1,000$. In *Example 1*, the association between covariates and potential outcomes is linear, whereas in *Example 2*, the association is quadratic. EMP SE refers to empirical standard error, and REL $\sqrt{\text{MSE}}$ refers to the ratio of square root of mean squared error relative to DR1 estimator.

6 Analysis of glaucoma data

The Ocular Hypertension Treatment Study (OHTS) is a multi-center randomized trial designed to evaluate the safety and efficacy of topical ocular hypertensive medication in delaying or preventing primary open-angle glaucoma in subjects with elevated intra-ocular pressure (IOP) [14]. In the first part of the study, a total of 1,636 subjects with an IOP between 24 and 32 mm in one eye and between 21 and 32 mm in the other eye were randomized to either observation or treatment with commercially available topical ocular hypertension medication. In the second part of the study, both the observation arm and the treatment arm received medication. The subjects were followed up every 6 months up to 7.5 years median time in the first part of the study and up to 13 years median time in the second part.

We use data from 817 subjects in the treatment arm to illustrate our method. In the treatment arm, the subjects received several classes of topical ocular hypertension medications, including beta blockers, prostaglandin analogues, topical carbonic anhydrase inhibitors and alpha 2 agonists. Sometimes more than one type of medication was prescribed for one eye. We are interested in comparing the effectiveness of the two most common classes, beta blockers and prostaglandin analogues in lowering IOP. The treatment is eye-specific drops, and the outcome is IOP in each eye. Within the treatment arm, different medications were not randomized. At each visit, clinicians made decisions about the type and dose of medication given to each subject, although dose information is not available in the data provided. Treatment decisions at each time may be influenced by the subject's treatment and covariate history. Therefore, the treatment received by these subjects is time-varying. The subject's IOPs at each visit are affected by earlier treatment and IOPs, and in turn affect later treatment and IOPs. Therefore, they are likely to be time-varying confounders.

We created a categorical treatment variable containing three categories: beta blocker only (BB), prostaglandin analogue only (PROST) and beta blocker plus prostaglandin analogue (BB + PROST). Since almost all subjects received medication only to one eye at baseline, we did not model baseline treatment but still included relevant baseline covariates in our analysis. In the follow-up visits, once a subject received no medication or medications other than a beta blocker and prostaglandin analogue in one of his eyes, the subsequent person-times were excluded from analysis. Since all remaining subjects received a beta blocker at the first follow-up visit, treatment at the first follow-up visit was also not modeled. In this study, most subjects did not receive different treatments in each eye. Therefore, we decide not to model the effect of treatment on

the opposite eye. For each eye, we pooled all subject-visits and fit a pooled multinomial logistic model after performing model selection based on AIC. The resulting treatment model was

$$\text{mlogitPr}[A_{tj} = a \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{L}}_t] = \alpha_{0t} + \alpha_1 A_{t-1j} + \alpha_2 A_{t-1j'} + \alpha_3 \text{IOP}_{tj} + \alpha_4 \text{IOP}_{tj'} + \alpha_5 \text{IOP}_{t-1j} + \\ \alpha_6 A_{t-1j} \cdot \text{IOP}_{t-1j} + \alpha_7 \text{RACE} + \alpha_8 \text{AGE} + \alpha_9 \text{IOP}_{0j},$$

where $a = 1$ or 2 refers to “PROST” or “BB + PROST” versus the reference category “BB”. The significant predictors include IOP in the current time in each eye, IOP in the previous time in the same eye, treatment in the previous time in each eye, interaction between IOP and treatment in the previous time in the same eye, as well as baseline covariates race, age and baseline IOP in the same eye. The time-varying intercept is modeled by a spline to allow more flexibility. The causal effect is modeled with a blip function with two parameters

$$E\left(Y_{k,j}^{\bar{\mathbf{a}}_m, \mathbf{0}} \mid \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m = \bar{\mathbf{a}}_m\right) - E\left(Y_{k,j}^{\bar{\mathbf{a}}_{m-1}, \mathbf{0}} \mid \bar{\mathbf{L}}_m = \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m\right) = \gamma_{m,k,j}^*(\bar{\mathbf{a}}_m, \bar{\mathbf{L}}_m; \psi) \\ = \psi_1 I(a_{m,j} = 1) + \psi_2 I(a_{m,j} = 2),$$

where ψ_1 represents the effect of “Prost” versus “BB”, and ψ_2 represents the effect of “BB + Prost” versus “BB”. Negative values refer to lowering IOP more and therefore indicate that a treatment is beneficial relative to “BB”. This model makes the symmetry assumption and also assumes the effect of treatment at each time is the same on each of the later outcomes. Both non-doubly robust (non-DR) and doubly robust (DR) estimators were computed for ψ , and the results are shown in Table 6.

The DR estimators have smaller standard errors compared to the non-DR estimators. Results from the DR approach show that $\hat{\psi}_1 = -0.15$ with p -value = 0.22, suggesting the effect of “PROST” is not significantly different from that of “BB”, and $\hat{\psi}_2 = 0.19$ with p -value = 0.047, suggesting the combined treatment “BB + PROST” is significantly worse than “BB” in lowering IOP. The fact that $\hat{\psi}_2$ is positive is counter-intuitive, since we would expect a combined treatment to work better than any of the single treatments. One possible explanation for this effect is residual confounding not adjusted for in the model.

Table 6: Results of glaucoma data analysis.

	Non-DR estimator			DR estimator		
	EST	SE	p -Value	EST	SE	p -Value
ψ_1	−0.27	0.34	0.21	−0.15	0.20	0.22
ψ_2	0.26	0.23	0.12	0.19	0.12	0.05

Note: Estimators from doubly robust (DR) and non-doubly robust (non-DR) approaches are shown. EST refers to estimate, and SE refers to asymptotic standard error.

7 Discussion

In this work, we have extended SNMMs to settings with clustered observations and time-varying confounding. We formulate models that are generalized to both cluster- and unit-level treatments, with consideration of the effect of interference. We also derive semiparametric estimators of parameters in those models. Our simulation study demonstrated that the estimators are consistent and achieve reasonable efficiency in contrast to those from the conventional GEE regression approach.

We have focused on examples with two units per cluster in all clusters. It is interesting to explore the performance of the estimators under larger cluster sizes, especially in the presence of interference. The

symmetry assumption is more important in such cases for reducing the dimension of parameters. One may also consider cases where the cluster size varies between clusters in the study population.

We have applied the method to compare the effectiveness of two commonly used topical ocular hypertension medication in lowering IOP using data from the treatment arm of OHTS. In the glaucoma study, two eyes of a given subject form a cluster. At each visit, different medications were not randomized in the treatment arm. Therefore, it is a setting with clustered observations and time-varying confounding. The effect of prostaglandin relative to beta-blocker is not shown to be statistically significant. The fact that the effect of the combined treatment is shown to be worse than that of the single treatments is counterintuitive. One likely explanation is residual confounding in the analysis, since clinicians usually prescribe combined treatment if the single ones do not work. Some of the residual confounding may also be due to the lack of dose information, which limits the capacity to fairly compare the three treatment categories. The analysis can potentially be improved by adjusting for loss to follow-up and exclusion of patients on other medications through inverse probability weighting (IPW). In this particular dataset, the treatment did not differ between two eyes in most subjects. Therefore, there is not enough power to analyze the effect of treatment on the opposite eye. However, the method introduced can be used to model interference among units within the same cluster when treatment assignments within the cluster are more variable.

We have demonstrated and implemented our method with continuous outcomes and an identity link in SNMMs. Our method can potentially be generalized to discrete outcomes and other link functions [9]. The extension to the log link will be straightforward. A single time point example of SNMM with a log link is

$$\log \left\{ E \left(\begin{matrix} Y_1^{a_1, a_2} \\ Y_2^{a_2, a_1} \end{matrix} \middle| \mathbf{L} = \mathbf{l}, \mathbf{A} = \mathbf{a} \right) \right\} - \log \left\{ E \left(\begin{matrix} Y_1^{0,0} \\ Y_2^{0,0} \end{matrix} \middle| \mathbf{L} = \mathbf{l}, \mathbf{A} = \mathbf{a} \right) \right\} = \begin{pmatrix} \gamma_1^*(a_1, a_2, \mathbf{l}) \\ \gamma_2^*(a_2, a_1, \mathbf{l}) \end{pmatrix}.$$

A blipped down outcome

$$\mathbf{U} = \begin{pmatrix} Y_1 \exp \{ -\gamma_1^*(A_1, A_2, \mathbf{L}; \psi) \} \\ Y_2 \exp \{ -\gamma_2^*(A_2, A_1, \mathbf{L}; \psi) \} \end{pmatrix}$$

can be used for estimation. For binary outcomes and the logit link, the causal parameter in SNMMs cannot be estimated with G-estimation [2]. Vansteelandt and Goetghebuer [15] proposed the generalized structural mean model to overcome this limitation [15]. The two-stage model includes a structural model to model causal effect and a traditional association model to model the mean of observed outcome in the treatment arm. They also proposed an estimator that is robust to misspecification of the nuisance association model. Robins [16] proposed a similar approach that uses a different parametric modeling restriction on the distribution of the observed data [16]. The causal model in these approaches can be extended to clustered observations in a similar way described in our method. Extensions of the association models to clustered observations will be straightforward.

One limitation of our approach is that when the causal model is misspecified we cannot obtain a meaningful interpretation of the causal parameters. For example, consider the case in which spillover effects exist but are omitted from the SNMM. As we showed via simulation, such misspecification results in biased estimates. Unless the investigator is certain that such effects do not exist or cannot be estimated relatively precisely, the SNMMs should include spillover effects. Neugebauer and van der Laan [17] proposed a nonparametric MSM that has meaningful interpretation for parameters even when the causal model is misspecified [17]. However, an extension of this approach to SNMMs does not give a straightforward interpretation of the parameters. Investigators also want to carefully consider specification of the treatment model, particularly when not using the doubly robust estimator. To improve the likelihood of correctly specifying this model, nonlinear regression techniques such as splines may be employed to inform covariates' functional form.

Acknowledgement: The authors are grateful to Vision Research Coordinating Center at the University of Washington in St. Louis for providing the glaucoma data, and to Dr. John Kempen for providing subject matter knowledge in analyzing the glaucoma data.

Funding: This work was funded by the National Institute of the Diabetes and Digestive and Kidney Diseases (NIDDK) (grant/award number: “R01-DK09004”).

Appendix

A. Asymptotic variance estimator

We adapted the approach for calculating asymptotic variance of $\hat{\psi}$ from Robins [18]. Let α denote nuisance parameters in the treatment model, S_α the score for α , S_ψ the score for ψ , α_0 and ψ_0 the true parameters for α and ψ , $S_{(\alpha,\psi)} = (S_\alpha^T, S_\psi^T)^T$, $S_{\alpha,i}$ and $S_{\psi,i}$ the scores for subject i , $S_i = (S_{\alpha,i}^T, S_{\psi,i}^T)^T$. Since S_ψ is the score from a semiparametric model, $n^{-1/2}S_\psi | \psi = \psi_0$ is asymptotically normal with mean 0 and asymptotic variance $I(\psi_0) = E(S_{\psi,i}S_{\psi,i}^T | \psi = \psi_0)$. If α in the treatment model is known, $I(\psi_0)$ can be consistently estimated by $\hat{I}(\hat{\psi}) = \frac{1}{n} \sum_i S_{\psi,i}S_{\psi,i}^T | \psi = \hat{\psi}$. If α needs to be estimated, we first calculate $\hat{I}(\hat{\alpha}, \hat{\psi}) = \begin{pmatrix} \hat{I}_{\alpha\alpha} & \hat{I}_{\alpha\psi} \\ \hat{I}_{\psi\alpha} & \hat{I}_{\psi\psi} \end{pmatrix} = \frac{1}{n} \sum_i S_i S_i^T | \alpha = \hat{\alpha}, \psi = \hat{\psi}$. By Taylor's expansion and Slutsky's theorem, $I(\psi_0)$ is consistently estimated by $\hat{I}(\hat{\psi}) = \hat{I}_{\psi\psi} - \hat{I}_{\psi\alpha} \hat{I}_{\alpha\alpha}^{-1} \hat{I}_{\alpha\psi}$. Again by Taylor's expansion, $n^{1/2}(\hat{\psi} - \psi_0)$ will be asymptotically normal with mean 0 and asymptotic variance $nB^{-1}I(\psi_0)B^{-1}$ which will be consistently estimated by $n\hat{B}^{-1}\hat{I}(\hat{\psi})\hat{B}^{-1}$, where $B^{-1} = E\left(\frac{\partial S_{\psi,i}}{\partial \psi^T} | \alpha = \alpha_0, \psi = \psi_0\right)$, $\hat{B}^{-1} = n^{-1} \frac{\partial S_\psi}{\partial \psi^T} | \alpha = \hat{\alpha}, \psi = \hat{\psi}$.

B. Likelihood of SNMMs for clustered observations

For a single time point example, if we assume ignorable treatment, likelihood of a SNMM can be written as the following:

$$\begin{aligned} g\{E(\mathbf{Y} | \mathbf{l}_0, \mathbf{a}_0)\} &= \gamma^*(\mathbf{l}_0, \mathbf{a}_0) + [g\{E(\mathbf{Y}^0 | \mathbf{l}_0)\} - g\{E(\mathbf{Y}^0)\}] + g\{E(\mathbf{Y}^0)\} \\ &= \gamma^*(\mathbf{l}_0, \mathbf{a}_0) + g\{E(\mathbf{Y}^0 | \mathbf{l}_0)\} \\ \varepsilon &= \mathbf{Y} - E(\mathbf{Y} | \mathbf{L}_0, \mathbf{A}_0) \end{aligned}$$

$$f(O; \rho) = f(\varepsilon(\rho) | \mathbf{L}_0, \mathbf{A}_0; \eta_1) \cdot f(\mathbf{L}_0; \eta_2) \cdot f(\mathbf{A}_0 | \mathbf{L}_0; \eta_3),$$

where $O = (\mathbf{Y}, \mathbf{L}_0, \mathbf{A}_0)$ represents observed data, ω represents nuisance parameters in the model for $E(\mathbf{Y}^0 | \mathbf{l}_0)$, $\rho = \psi \times \eta \times \omega$ represents the parameter space.

The above example can be extended to time-varying treatments where treatments and covariates are measured repeatedly at times 0 to $K - 1$. Outcome is measured only at the last time point K . First define

$$\begin{aligned} v_m^*(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1}) &= g\{E(\mathbf{Y}^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1})\} - g\{E(\mathbf{Y}^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{l}}_{m-1}, \mathbf{l}_m = \mathbf{0}, \bar{\mathbf{a}}_{m-1})\} \\ \Gamma_m^*(\bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}) &= g \int g^{-1}\{v_m^*(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1})\} dF(\mathbf{l}_m | \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}) \\ &= -g\{E(\mathbf{Y}^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{l}}_{m-1}, \mathbf{l}_m = \mathbf{0}, \bar{\mathbf{a}}_{m-1})\} + g\{E(\mathbf{Y}^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1})\} \end{aligned}$$

The functions v_m^* and Γ_m^* are related to the association between covariates $\mathbf{L}_m$ and counterfactual outcomes $\mathbf{Y}^{\bar{\mathbf{a}}_{m-1},0}$. Likelihood of a SNMM with identity or log link can be written as the following:

$$g[E(\mathbf{Y} | \bar{\mathbf{L}}_{K-1}, \bar{\mathbf{A}}_{K-1})] = \sum_{m=0}^{K-1} \gamma_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_m) + \sum_{m=1}^{K-1} \{v_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1}) - \Gamma_m^*(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{a}}_{m-1})\} + g\{E(\mathbf{Y}^0 | \mathbf{l}_0)\}$$

$$\varepsilon = \mathbf{Y} - E(\mathbf{Y} | \bar{\mathbf{L}}_{K-1}, \bar{\mathbf{A}}_{K-1})$$

$$f(O; \rho) = f(\varepsilon(\rho) | \bar{\mathbf{L}}_{K-1}, \bar{\mathbf{A}}_{K-1}; \eta_1) \cdot \prod_{k=0}^{K-1} f(\mathbf{L}_k | \bar{\mathbf{L}}_{k-1}, \bar{\mathbf{A}}_{k-1}; \eta_2) \cdot f(\mathbf{A}_k | \bar{\mathbf{L}}_k, \bar{\mathbf{A}}_{k-1}; \eta_3),$$

where $O = (\mathbf{Y}, \bar{\mathbf{L}}_{K-1}, \bar{\mathbf{A}}_{K-1})$ represents observed data, ω represents nuisance parameters in the model for $E(\mathbf{Y}^0 | \mathbf{l}_0)$, μ represents nuisance parameters in the models for $v_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1})$, $\rho = \psi \times \eta \times \mu \times \omega$ represents the parameter space. Choosing a linear working model for $v_m^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1})$ will facilitate writing an analytical form for the full likelihood.

Lastly, the second example can be further generalized to repeatedly measured outcomes at times $k = 1, \dots, K$. Similarly, we define

$$v_{m,k}^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1}) = g\left\{E\left(\mathbf{Y}_k^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1}\right)\right\} - g\left\{E\left(\mathbf{Y}_k^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{L}}_{m-1}, \mathbf{l}_m = \mathbf{0}, \bar{\mathbf{a}}_{m-1}\right)\right\}$$

$$\Gamma_{m,k}^*(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{a}}_{m-1}) = -g\left\{E\left(\mathbf{Y}_k^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{L}}_{m-1}, \mathbf{l}_m = \mathbf{0}, \bar{\mathbf{a}}_{m-1}\right)\right\} + g\left\{E\left(\mathbf{Y}_k^{\bar{\mathbf{a}}_{m-1},0} | \bar{\mathbf{L}}_{m-1}, \bar{\mathbf{a}}_{m-1}\right)\right\}$$

Likelihood of a SNMM with identity or log link can be written as the following:

$$g(E(\mathbf{Y}_k | \bar{\mathbf{L}}_{k-1}, \bar{\mathbf{A}}_{k-1})) = \sum_{m=0}^{k-1} \gamma_{m,k}^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_m) + \sum_{m=1}^{k-1} \{v_{m,k}^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1}) - \Gamma_{m,k}^*(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{a}}_{m-1})\} + g\{E(\mathbf{Y}_k^0 | \mathbf{l}_0)\}$$

$$\varepsilon_k = \mathbf{Y}_k - E(\mathbf{Y}_k | \bar{\mathbf{L}}_{k-1}, \bar{\mathbf{A}}_{k-1})$$

$$f(O; \rho) = \prod_{k=1}^K f(\varepsilon_k(\rho) | \bar{\mathbf{L}}_{k-1}, \bar{\mathbf{A}}_{k-1}; \eta_1) \cdot \prod_{k=0}^{K-1} f(\mathbf{V}_k | \bar{\mathbf{L}}_{k-1}, \bar{\mathbf{A}}_{k-1}, \mathbf{Y}_k; \eta_2) \cdot f(\mathbf{A}_k | \bar{\mathbf{L}}_k, \bar{\mathbf{A}}_{k-1}; \eta_3),$$

where $\mathbf{L}_k = (\mathbf{V}_k, \mathbf{Y}_k)$, $O = (\bar{\mathbf{Y}}_K, \bar{\mathbf{L}}_{K-1}, \bar{\mathbf{A}}_{K-1})$, ω represents nuisance parameters in the models for $E(\mathbf{Y}_k^0 | \mathbf{l}_0)$, μ represents nuisance parameters in the models for $v_{m,k}^*(\bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1})$, $\rho = \psi \times \eta \times \mu \times \omega$ represents the parameter space.

References

1. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat Theor Met* 1994;23:2379–412.
2. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME and Berry D, editors. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer, 2000:95–133.
3. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996;91:444–55.
4. Vansteelandt S, Joffe M. Structural nested models and g-estimation: the partially realized promise. *Stat Sci* 2014;29:707–31.
5. Brumback BA, He Z, Prasad M, Freeman MC, Rheingans R. Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial. *Stat Med* 2014;33:1490–502.
6. Rubin DB. Comment: which ifs have causal answers. *J Am Stat Assoc* 1986;81:961–2.
7. VanderWeele TJ. Ignorability and stability assumptions in neighborhood effects research. *Stat Med* 2008;27:1934–43.
8. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688.

9. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME and Berry D, editors. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer, 2000:1–94.
10. Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J Am Stat Assoc* 2006;101:1398–407.
11. Tsiatis A. *Semiparametric theory and missing data*. Chicago: Springer, 2007.
12. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. *Longitud Data Anal* 2009:553–99.
13. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:13–22.
14. Kass MA, Heuer DK, Higginbotham EJ, Johnson CA, Keltner JL, Miller JP, et al.. The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol* 2002;120:701–13.
15. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *J R Stat Soc Ser B Stat Methodol* 2003;65:817–35.
16. Robins J, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* 2004;91:763–83.
17. Neugebauer R, van der Laan M. Nonparametric causal effects based on marginal structural models. *J Stat Plan Inference* 2007;137:419–34.
18. Robins J. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 1992;79:321–34.