1997 Proceedings of the American Statistical Association,
Section on Bayesian Statistical Science, 1998, pp. 1-10.

Marginal Structural Models

James M. Robins, Harvard School of Public Health
677 Huntington Avenue, Boston, MA 02115, USA

Key Words: causal inference; semiparametric models; longitudinal data.

1. Introduction

Robins (1993, 1994, 1997, 1998) has developed a set of causal or counterfactual models, the structural nested models (SNMs). The purpose of this paper is to introduce a simpler class of causal models – the (non-nested) marginal structural models (MSMs). We will then describe a class of semiparametric estimators for the parameters of these new models under a sequential randomization assumption. We then compare the strengths and weaknesses of MSMs versus SNMs for causal inference from complex longitudinal data. Our results provide an extension to continuous treatments of the semiparametric efficient propensity score estimator of an average treatment effect.

2. The Data

Consider a study where we observe n iid copies of data $O = \left(\overline{A}(C), \overline{L}(C)\right)$, where C is an administrative end of follow-up time, $\overline{A}(C)$ is a treatment process, $\overline{L}(C)$ is an outcome or response process and, for any $Z(u), \overline{Z}(t) \equiv \{Z(u); 0 \leq u \leq t\}$. We assume C is an element of $L(0)$ since it is assumed known at time 0.

For purposes of causal inference, we assume the existence of an underlying treatment process $\overline{A} = \{A(u); 0 \leq u < \infty\}$ with $A(u)$ taking values in a set $\mathcal{A}(u)$ and the existence of underlying counterfactual random variables

$$\left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\} \tag{1}$$

where $\overline{L}_{\overline{a}} = \{L_{\overline{a}}(u); 0 \leq u < \infty\}$, $\overline{a} = a(\cdot) = \{a(t); 0 \leq t < \infty \text{ and } a(t) \in \mathcal{A}(t)\}$ is a treatment plan (equivalently, regime or function) lying in a set of functions $\overline{\mathcal{A}}$. Given a regime $\overline{a}$, let $\overline{L}_{\overline{a}(u),\mathbf{0}}$ be counterfactual history under a regime $\overline{a}^*$ that agrees with $\overline{a}$ through time u and is 0 thereafter, where 0 is the baseline value of $a(t)$. Then we assume that the $\overline{L}_{\overline{a}}$ satisfy the following consistency assumption with probability 1:

$$\overline{L}_{\overline{a}(u),\mathbf{0}}(u) = \overline{L}_{\overline{a}(t),\mathbf{0}}(u) = \overline{L}_{\overline{a}}(u) = \overline{L}_{\overline{a}^\dagger}(u) \tag{2}$$

for all $t > u$ and all $\overline{a}^\dagger$ with $\overline{a}^\dagger(u) = \overline{a}(u)$. This assumption essentially says that the future does not determine the past. The observed data are linked to the counterfactual data by

$$\overline{L}(C) = \overline{L}_{\overline{A}(C),0}(C) \ . \tag{3}$$

We assume $\overline{L}_{\overline{a}} = \left(\overline{Y}_{\overline{a}}, \overline{V}_{\overline{a}}\right)$ where $\overline{Y}_{\overline{a}}$ is an outcome process of interest and $\overline{V}_{\overline{a}}$ is the process of other recorded variables. Further, we shall make the sequential randomization (i.e., ignorable treatment

assignment) assumption that for all t and $\overline{a} \in \overline{\mathcal{A}}$,

$$\underline{Y_{\overline{a}}}(t) \coprod A(t) \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right) \tag{4}$$

where for any variable $\underline{Z}(t) = \{Z(u); u \geq t\}$. Because of measurability issues, (4) is not well-defined. If the $A(t)$ process can only jump at discrete non-random times $t_1, t_2, \ldots$ and the $\overline{L}(t)$ process has left-hand limits, i.e., $\overline{L}(t^{-}) \equiv \lim_{u \uparrow t} \overline{L}(u)$, (4) is formally, for each t_k,

$$f\left[A(t_k) \mid \overline{L}\left(t_k^{-}\right), \overline{A}\left(t_k^{-}\right), \underline{Y_{\overline{a}}}(t_k)\right] = f\left[A(t_k) \mid \overline{L}\left(t_k^{-}\right), \overline{A}\left(t_k^{-}\right)\right] \ . \tag{5}$$

where $f(\cdot \mid \cdot)$ is the conditional density of $A(t_k)$ with respect to a dominating measure. If $A(t)$ is a marked point process that can jump in continuous time with CADLAG (continuous from the right with left-hand limits) step-function sample paths, then Eq. (4) is formally that

$$\lambda_A\left[t \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right), \underline{Y_{\overline{a}}}(t)\right] = \lambda_A\left[t \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right)\right] \tag{6a}$$

and

$$\begin{aligned} &f[A(t) \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right), A(t) \neq A\left(t^{-}\right), \underline{Y_{\overline{a}}}(t)\,] = \\ &f\left[A(t) \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right), A(t) \neq A\left(t^{-}\right)\right] \ . \end{aligned} \tag{6b}$$

Here, the intensity process $\lambda_A(t \mid \cdot)$ is $\lim_{\delta\, t \to 0} pr[A(t+\delta t) \neq A(t^{-}) \mid A(t^{-}), \cdot]/\delta t$.

Following Heitjan and Rubin (1991), we say the data are coarsened at random (CAR) if

$$\begin{aligned} &f\left[\overline{A}(C) \mid \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}\right] \text{ depends only on} \\ &O = \left(\overline{A}(C), \overline{L}(C)\right) \ . \end{aligned} \tag{7}$$

We assume the $\left\{L_{\overline{a}}(u); \overline{a} \in \overline{\mathcal{A}}^{\dagger} \subseteq \overline{\mathcal{A}}\right\}$ have a non-degenerate joint distribution whenever $\overline{a}_1(u) \neq \overline{a}_2(u)$ for all $\overline{a}_1, \overline{a}_2 \in \overline{\mathcal{A}}^{\dagger}$. Then CAR implies sequential randomization (4) but the converse is not true. Robins (1997, pg. 83) gives examples where one would expect (4) to be true even when (7) is false. In this paper, we shall only need (4). However, if (7) is the sole restriction imposed, this essentially places no restrictions on the joint distribution of the observable random variables (Gill, van der Laan, Robins, 1997) and, thus, is not subject to empirical test.

2.1. MSMs

A MSM for $\left\{\overline{Y}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}$ places restrictions on the marginal distribution of the $\overline{Y}_{\overline{a}}$ possibly conditional on a baseline variable $V^{\dagger}$ in $V(0)$ (with $C \in V^{\dagger}$ if C is random). Examples of MSMs follow.

Model 1: Suppose $C = K+1$ w.p.1., the $\overline{A}(C)$ process jumps only at times $0, 1, 2, \ldots, K$ and the $\overline{L}_{\overline{a}}$ process jumps only at times $0^{-}, 1^{-}, 2^{-}, \ldots, K^{-}, K+1^{-}$. In models 1a-1c, we are only concerned with an outcome measured at end of follow-up. Hence, we set $Y_{\overline{a}}(m) \equiv 0$ with probability 1 *for* $m \leq K$ and define $Y_{\overline{a}} = Y_{\overline{a}}(K+1)$. Then we have

Model 1a – non-linear least squares: $E\left[Y_{\overline{a}} \mid V^{\dagger}\right] = g\left[\overline{a}(K), V^{\dagger}, \beta_0\right]$ where $g(\cdot,\cdot,\cdot)$ is a known function.

Model 1b – semiparametric regression: $\eta\left\{E\left[Y_{\overline{a}} \mid V^{\dagger}\right]\right\} = g\left[\overline{a}(K), V^{\dagger}, \beta_0\right] + g^{\dagger}\left(V^{\dagger}\right)$ where $\eta(\cdot)$ is a known monotone link function, $g^{\dagger}(\cdot)$ is unknown and unrestricted and $g(\cdot,\cdot,\cdot)$ is a known function

satisfying $g\left(\mathbf{0}, V^{\dagger}, \beta\right) = 0$. The requirement that $g\left(\mathbf{0}, V^{\dagger}, \beta\right) = 0$ implies that $g^{\dagger}\left(V^{\dagger}\right)$ is the "main effect of $V^{\dagger}$."

Model 1c – stratified transformation model: $pr\left[R\left(\overline{a}, \beta_0\right) < t \mid V^{\dagger}\right] = F_0\left(t \mid V^{\dagger}\right), F_0\left(t \mid V^{\dagger}\right)$ an unknown distribution function, $R\left(\overline{a}, \beta\right) = r\left(Y_{\overline{a}}, \overline{a}, V^{\dagger}, \beta\right)$ is a known increasing function of $Y_{\overline{a}}$ satisfying $r\left(y, \overline{a}, V^{\dagger}, \beta\right) = y$ if $\overline{a} \equiv \mathbf{0}$ or $\beta = 0$.

In the following model, we are interested in the outcome at each $m \geq 1$ so we no longer assume that $Y_{\overline{a}}(m) \equiv 0$ with probability 1.

Model 1d – multivariate non-linear least squares: $E\left[Y_{\overline{a}}(m) \mid V^{\dagger}\right] = g_m\left[\overline{a}\left(m-1\right), V^{\dagger}, \beta_0\right]$, $m = 1, \ldots, K+1$ where the $g_m\left(\cdot, \cdot, \cdot\right)$ are known.

Model 2: $C = \infty$, $Y_{\overline{a}}$ is a failure time process, i.e., $Y_{\overline{a}}$ jumps from 0 to 1 at some particular time and stays at 1. Then define the failure time $T_{\overline{a}}$ by the equation $Y_{\overline{a}}\left(T_{\overline{a}}\right) = 1$ and $Y_{\overline{a}}\left(T_{\overline{a}}^{-}\right) = 0$. Let $\lambda_0\left(t\right)$ and $\lambda_0\left(t \mid V^{\dagger}\right)$ be unknown non-negative functions of t and $\left(t, V^{\dagger}\right)$ respectively and, for any $Z, \lambda_Z\left(u\right)$ is the hazard of Z.

Model 2a – Cox proportional hazards model: $\lambda_{T_{\overline{a}}}\left[t \mid V^{\dagger}\right] = \lambda_0\left(t\right) \exp\left[r\left(\overline{a}\left(t^{-}\right), t, V^{\dagger}; \beta_0\right)\right]$ where $r\left(\right)$ is a known function satisfying $r\left(\mathbf{0}, t, 0; \beta\right) = 0$.

Model 2b – stratified Cox proportional hazards model: $\lambda_{T_{\overline{a}}}\left(t \mid V^{\dagger}\right) = \lambda_0\left(t \mid V^{\dagger}\right) \exp\left[r\left(\overline{a}\left(t^{-}\right) t, V^{\dagger}; \beta_0\right)\right]$ where, now, $r\left(\mathbf{0}, t, V^{\dagger}; \beta\right) = 0$.

Model 2c – stratified time-dependent accelerated failure time model: $pr\left[r\left(T_{\overline{a}}, \overline{a}, V^{\dagger}, \beta_0\right) < t \mid V^{\dagger}\right] = F_0\left(t \mid V^{\dagger}\right)$ where $r\left(u, \overline{a}, V^{\dagger}, \beta\right) = r\left(u, \overline{a}\left(u\right), V^{\dagger}, \beta\right)$ is a known function increasing in its first argument satisfying $r\left(u, \mathbf{0}, V^{\dagger}, \beta\right) = u$. This model can also be written as

$$\lambda_{R(\overline{a}, \beta_0)}\left(t \mid V^{\dagger}\right) = \lambda_0\left(t \mid V^{\dagger}\right)$$

for $\overline{a} \in \overline{\mathcal{A}}$, where $R\left(\overline{a}, \beta\right) = r\left(T_{\overline{a}}, \overline{a}, V^{\dagger}, \beta\right)$.

3. Estimation

3.1. Ancillary treatment process

In this section, we consider estimation of the parameter β_0 of our marginal structural models. In this subsection, we will suppose that $\overline{A}$ is an ancillary (i.e., exogenous) covariate process, i.e.,

$$\overline{A} \coprod \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\} \mid V^{\dagger} . \tag{8}$$

The often unrealistic assumption (8) implies CAR but, in contrast to CAR, places restrictions on the joint distribution of the data. Specifically (8) implies

$$A\left(t\right) \coprod \overline{L}\left(t^{-}\right) \mid \overline{A}\left(t^{-}\right), V^{\dagger} \tag{9}$$

and thus (8) is subject to an empirical test.

Given (8), the restrictions on the observables O implied by any MSM are (9) and that the restrictions on the distribution of $\overline{Y}_{\overline{a}}$ given $V^{\dagger}$ specified by the MSM hold for the conditional distribution of the observable $\overline{Y}\left(C\right)$ conditional on $\left(\overline{A}\left(C\right), V^{\dagger}\right)$.

For reasons that will become clear below, we indicate with a "*" any expectations, probabilities or hazard functions computed under the assumption that (8) and (9) hold. For convenience, denote $\overline{A}\left(C\right)$ as $\overline{A}$. Thus, for our MSM models (1a) – (2c), (8) implies

Model 1a: $E^*\left[Y \mid V^\dagger, \overline{A}\right] = g\left(\overline{A}, V^\dagger, \beta_0\right)$
Model 1b: $\eta\left\{E^*\left[Y \mid V^\dagger, \overline{A}\right]\right\} = g\left(\overline{A}, V^\dagger, \beta_0\right) + g^\dagger\left(V^\dagger\right)$.
Model 1c: $R\left(\beta_0\right) \coprod^* \overline{A} \mid V^\dagger$ where $R\left(\beta_0\right) \equiv R\left(\overline{A}, \beta_0\right)$.
Model 1d: $E^*\left[Y\left(m\right) \mid V^\dagger, \overline{A}\right] = g_m\left[\overline{A}\left(m-1\right), V^\dagger; \beta_0\right], m = 1, \ldots, K+1$.
Model 2a: $\lambda_T^*\left[t \mid V^\dagger, \overline{A}\right] = \lambda_T^*\left[t \mid V^\dagger, \overline{A}\left(t^-\right)\right] = \lambda_0\left(t\right)\exp\left[r\left(\overline{A}\left(t^-\right), t, V^\dagger, \beta_0\right)\right]$.
Model 2b: $\lambda_T^*\left[t \mid V^\dagger, \overline{A}\right] = \lambda_T^*\left[t \mid V^\dagger, \overline{A}\left(t^-\right)\right] = \lambda_0\left(t \mid V^\dagger\right)\exp\left[r\left(\overline{A}\left(t^-\right), t, V^\dagger, \beta_0\right)\right]$.
Model 2c: $\lambda_{R(\beta_0)}^*\left[u \mid V^\dagger, \overline{A}\right] = \lambda_{R(\beta_0)}^*\left[u \mid \overline{A}\left[r^{-1}\left(u, \overline{A}, V^\dagger, \beta_0\right)\right], V^\dagger\right] = \lambda_0\left(u \mid V^\dagger\right)$ where $R\left(\beta_0\right) \equiv R\left(\overline{A}, \beta_0\right)$ and $r^{-1}\left(u, \overline{A}, V^\dagger, \beta\right) \equiv t$ if $r\left(t, \overline{a}, V^\dagger, \beta\right) = u$.

We shall now consider estimation of these models for the observables, under assumption (9), and the further assumption that

$$\overline{A}\left(C\right) \text{ has a known conditional distribution given } V^\dagger \,. \tag{10}$$

Semiparametric inference in models 1a - 2c without (9) and (10) imposed has been examined previously by many authors. Below we use their results to solve the estimation problem in our semiparametric model.

We will show that associated with each MSM model with (9) and (10) imposed is a class of regular asymptotically linear (RAL) estimators $\left\{\widehat{\beta}^*\left(h, \phi\right)\right\}$ for β_0, indexed by vector functions $h \in \mathcal{H}$ and $\phi \in \Phi$ such that the set $\mathcal{IF}^* = \left\{IF^*\left(h, \phi\right)\right\}$ of influence functions of the $\widehat{\beta}^*\left(h, \phi\right)$ constitute all the influence functions for the model, in the sense that if $\widetilde{\beta}^*$ is any other RAL estimator, then the influence function of $\widetilde{\beta}^*$ equals $IF^*\left(h, \phi\right)$ for some functions $h \in \mathcal{H}, \phi \in \Phi$. We obtain $\widehat{\beta}^*\left(h, \phi\right)$ by solving the estimating equations $n^{-\frac{1}{2}}\sum_i \widehat{D}_i^*\left(\beta, h, \phi\right) = o_p\left(1\right)$ described below. The solution $\widehat{\beta}^*\left(h, \phi\right)$ has influence function $IF^*\left(h, \phi\right) = \left\{\kappa^*\left(h\right)\right\}^{-1} U^*\left(\beta_0, h, \phi\right)$ where $U_i^*\left(\beta_0, h, \phi\right)$ depends only on subject i's data, $\kappa^*\left(h\right) = -\partial E^*\left[U^*\left(\beta, h, \phi\right)\right]/\partial\beta_{|\beta=\beta_0}$ does not depend on ϕ, and $n^{-\frac{1}{2}}\sum_i \widehat{D}_i^*\left(\beta_0, h, \phi\right) = n^{-\frac{1}{2}}\sum_i U_i^*\left(\beta_0, h, \phi\right) + o_p\left(1\right)$. Furthermore, $\Lambda^\perp = \left\{U^*\left(\beta_0, h, \phi\right)\right\}$ is the linear span of $\mathcal{IF}^*$ and thus is the orthogonal complement to the nuisance tangent space for the model in the Hilbert space induced by the covariance norm. We refer to $U^*\left(\beta_0, h, \phi\right)$ as the influence function for the estimating function $\widehat{D}\left(\beta_0, h, \phi\right)$. More specifically, $U^*\left(\beta, h, \phi\right)$ and $\widehat{D}^*\left(\beta, h, \phi\right)$ are each expressed as the sum of the two components, one of which $U_{tp}^*\left(\phi\right) = D_{tp}^*\left(\phi\right)$ is independent of the choice of the MSM and follows from the fact that, for the "treatment process (tp)," (9) and (10) are assumed. Specifically, if the $A\left(t\right)$ can jump only at times $0, 1, 2, \ldots, U_{tp}^*\left(\phi\right) = \sum_{k=0}^{int(C)} \phi\left(k, \overline{A}\left(k\right), \overline{L}\left(k^-\right)\right) - E^*\left[\phi\left(k, \overline{A}\left(k\right), \overline{L}\left(k^-\right)\right) \mid \overline{L}\left(k^-\right), \overline{A}\left(k^-\right)\right]$ where $int\left(C\right)$ is the greatest integer less that or equal to C. It is easy to see that $\left\{U_{tp}^*\left(\phi\right)\right\}$ is, as ϕ varies, the sum over k of functions of the observed data $\left(\overline{A}\left(k\right), \overline{L}\left(k^-\right)\right)$ with mean zero given $\left(\overline{A}\left(k^-\right), \overline{L}\left(k^-\right)\right)$. If $A\left(t\right)$ is a continuous time marked point process, then $U_{tp}^*\left(\phi\right) = \int dM_A^*\left(u\right)\phi_1\left(u, \overline{A}\left(u^-\right), \overline{L}\left(u^-\right)\right) + \int dN_A\left(u\right)\{\phi_2\left(u, \overline{A}\left(u\right), \overline{L}\left(u^-\right)\right) - E^*[\phi_2\left(u, \overline{A}\left(u\right), \overline{L}\left(u^-\right)\right) \mid A\left(u\right) \neq A\left(u^-\right), \overline{L}\left(u^-\right), \overline{A}\left(u^-\right)]\}$ where $dM_A^*\left(u\right) = dN_A\left(u\right) - \lambda_A^*\left[u \mid \overline{A}\left(u^-\right), \overline{L}\left(u^-\right)\right]du$ and $dN_A\left(u\right) = I\left\{A\left(u\right) \neq A\left(u^-\right)\right\}$ counts jumps in the $\overline{A}$ process.

The other structural model-specific component $\widehat{D}_{sm}^*\left(\beta, h\right)$ and $U_{sm}^*\left(\beta, h\right)$ of $\widehat{D}^*\left(\beta, h, \phi\right)$ and $U^*\left(\beta, h, \phi\right)$ are the well-known estimating functions and their associated influence functions for models 1a - 2c with neither (9) nor (10) imposed.
Model 1a: $\widehat{D}_{sm}^*\left(\beta, h\right) = U_{sm}^*\left(\beta, h\right) = h\left(\overline{A}, V^\dagger\right)\varepsilon\left(\beta\right)$ with $\varepsilon\left(\beta\right) = Y - g\left(\overline{A}, V^\dagger, \beta\right)$ and $h\left(\overline{A}, V^\dagger\right)$ is any $\dim\left(\beta\right)$ vector function.
Model 1b: $\eta\left(x\right) = x : \widehat{D}_{sm}^*\left(\beta, h\right) = U_{sm}^*\left(\beta, h\right) = \left\{\varepsilon\left(\beta\right) - h_1\left(\overline{A}, V^\dagger\right)\right\}$

$\left\{h_2\left(\overline{A},V^\dagger\right)-E^*\left[h_2\left(\overline{A},V^\dagger\right)\mid V^\dagger\right]\right\}$ where h_1 is any real valued function, $\varepsilon(\beta)$ is as just defined and the range of h_2 is of $\dim(\beta)$.

$\eta(x)=\ln[x/(1-x)]$: $U^*_{sm}(\beta,h)=U^\dagger(h,P(\beta))$ and $\widehat{D}^*_{sm}(\beta,h)\equiv U^\dagger_{sm}\left(h,\widehat{P}(\beta)\right)$, where $P(\beta)=expit\left[g\left(\overline{A},V^\dagger,\beta\right)+g^\dagger\left(V^\dagger\right)\right]$, $\widehat{P}(\beta)=expit\left[g\left(\overline{A},V^\dagger,\beta\right)+\widehat{g}^\dagger\left(V^\dagger\right)\right]$, $expit(x)=e^x/(1+e^x)$, $\widehat{g}^\dagger\left(V^\dagger\right)$ is a $n^{\frac{1}{4}}$ - consistent estimate of $g^\dagger\left(V^\dagger\right)$, and $U^\dagger(h,P(\beta))\equiv$
$\{Y-P(\beta)\}\{h\left(\overline{A},V^\dagger\right)-E^*\left[h\left(\overline{A},V^\dagger\right)P(\beta)\{1-P(\beta)\}\mid V^\dagger\right]/E^*\left[P(\beta)\{1-P(\beta)\}\mid V^\dagger\right]\}$.
Model 1c: $\widehat{D}^*_{sm}(\beta,h)=U^*(\beta,h)=h\left[R(\beta),\overline{A},V^\dagger\right]-\int h\left[R(\beta),\overline{a},V^\dagger\right]dF^*\left[\overline{a}\mid V^\dagger\right]$.
Model 1d: Let $\varepsilon(\beta)=\{\varepsilon_1(\beta),\ldots\varepsilon_{K+1}(\beta)\}'$, $\varepsilon_m(\beta)=Y(m)-g_m\left[\overline{A}(m-1),V^\dagger;\beta\right]$. Then $\widehat{D}^*_{sm}(\beta,h)=U^*_{sm}(\beta,h)=h\left(\overline{A},V^\dagger\right)\varepsilon(\beta)$ where $h\left(\overline{A},V^\dagger\right)$ is now any $\dim(\beta)\times(K+1)$ matrix of real valued functions.
Model 2a: $\widehat{D}^*_{sm}(\beta,h)=\int_0^\infty dN_T(u)\left\{h\left(u,\overline{A}(u),V^\dagger\right)-\widetilde{\mathcal{L}}(h,u,\beta)\right\}$, where $\widetilde{\mathcal{L}}(h,u,\beta)=\widetilde{J}[h,\beta]/\widetilde{J}[\mathbf{1},\beta]$; for any $h\left(u,\overline{A}(u),V^\dagger\right)$, $\widetilde{J}(h,\beta)=\widetilde{E}[h\left(u,\overline{A}(u),V^\dagger\right)I(T>u)\exp\left\{r\left[\overline{A}(u),u,V^\dagger,\beta\right]\right\}]$; for any H_i, $\widetilde{E}(H)=\sum_{i=1}^n H_i/n$; $\mathbf{1}$ is the constant function equal to one; and $N_T(u)=I(T\le u)$. $U^*_{sm}(\beta,h)=\int_0^\infty dM_T(u)\left\{h\left(u,\overline{A}(u),V^\dagger\right)-\mathcal{L}^*\left(h,u,V^\dagger,\beta\right)\right\}$ where $\mathcal{L}^*(h,u,\beta)=J^*[h,\beta]/J^*[\mathbf{1},\beta]$; $J^*[h,\beta]$ is defined like $\widetilde{J}(h,\beta)$ but with E^* replacing $\widetilde{E}$; and $dM_T(u)=dN_T(u)-\lambda_T\left(u\mid\overline{A},V^\dagger\right)I(T>u)\,du$.
Model 2b: $U^*_{sm}(\beta,h)$ and $\widehat{D}^*_{sm}(\beta,h)$ are as above except $J^*(h,\beta)\equiv E^*[h\left(u,\overline{A}(u),V^\dagger\right)I(T>u)\exp\left\{r\left[\overline{A}(u),u,V^\dagger,\beta\right]\right\}\mid V^\dagger]$ *and* $\widetilde{J}(h,\beta)$ replaces $E^*\left(\cdot\mid V^\dagger\right)$ in $J^*(h,\beta)$ by a $n^{\frac{1}{4}}$ - consistent estimator $\widehat{E}\left(\cdot\mid V^\dagger\right)$.
Model 2c: $\widehat{D}^*_{sm}(\beta,h)=\int_0^\infty duI[R(\beta)>u]\left\{H_2(u,\beta)-E^*\left[H_2(u,\beta)\mid V^\dagger\right]\right\}+$
$\int_0^\infty dN_{R(\beta)}(u)\left[H_1(u,\beta)-E^*\left[H_1(u,\beta)\mid V^\dagger\right]\right]$ and, for $j=0,1$, $H_j(u,\beta)=h_j\left[u,\overline{A}\left\{r^{-1}\left(u,\overline{A},V^\dagger,\beta\right)\right\},V^\dagger\right]$.
$U^*_{sm}(\beta,h)=\widehat{D}^*_{sm}(\beta,h)-E^*\left[D^*_{sm}(\beta,h)\mid\ \overline{A},V^\dagger\right]$.

Remark: Note that in model 2b and in model 1b with $\eta(x)=\ln[x/(1-x)]$, smooths are necessary if $V^\dagger$ has continuous components. In particular, due to the curse of dimensionality, it is not possible to obtain a reasonable $n^{\frac{1}{2}}$ – consistent estimator of β_0 in these models when $V^\dagger$ has multiple continuous components.

3.2. Non-ancillary treatment process

In this section, we no longer assume (8) is true. The essential idea of this section (requiring some minor modification) is to reweight $\widehat{D}^*_{sm}(\beta,h)$ by the inverse of a subject's probability of having had his observed treatment history. We continue to assume that

$$\begin{aligned}&f\left[a(t)\mid\overline{L}\left(t^-\right),\overline{A}\left(t^-\right),V^\dagger\right]\\&\text{is known for } t\le C\end{aligned}\tag{11}$$

which implies that if $A(t)$ jumps at non-random times $0,\ldots,K$, $W(k)=f\left[A(k)\mid\overline{L}(k^-),\overline{A}(k^-)\right]$ and $\overline{W}(k)=\prod_{m=0}^k W(k)$ are known. If $A(t)$ jumps in continuous time, $\overline{W}(t)=$
$\exp\left[-\int_0^t\lambda_A\left[u\mid\overline{L}(u^-),\overline{A}(u^-)\right]du\right]\prod_{\{u;A(u)\neq A(u^-),u<t\}}f[A(u)\mid\overline{A}(u^-),\overline{L}(u^-),A(u)\neq A(u^-)]$ is known.
Now if $A(t)$ jumps at non-random times, let $\overset{\circ}{\mathcal{A}}\left(k,\overline{a}(k^-),v^\dagger\right)=$
$\{a(k);f\left[a(k)\mid\overline{L}(k^-),\overline{A}(k^-)=\overline{a}(k^-),V^\dagger=v^\dagger\right]\neq 0\ w.p.1\}$ and set $C^\dagger=\min\left\{k;A(k)\notin\overset{\circ}{\mathcal{A}}\left(k,\overline{A}(k^-),V^\dagger\right)\right\}$.
If $A(t)$ jumps in continuous time, let $\overset{\circ}{\mathcal{A}}\left(t,\overline{a}(t^-),v^\dagger\right)=\{a(t);f[a(t)\mid\overline{L}(t^-),\overline{A}(t^-)=\overline{a}(t^-),A(t)\neq a(t^-),V^\dagger=v^\dagger]\neq 0\ w.p.1$ or $a(t)=a(t^-)\}$ and set $C^\dagger=\inf\left\{t;A(t)\notin\overset{\circ}{\mathcal{A}}\left(t,\overline{A}(t^-),V^\dagger\right)\right\}$. The variable $C^\dagger$ is crucial because, as indicated in the remark following Lemma 3.1 below, one can only unbiasedly

reweight a function of $A(t)$ for $t \in \overset{\circ}{\mathcal{A}}\left(t, v^{\dagger}\right)$.

Let $f^{*}\left(\overline{a} \mid V^{\dagger}\right)$ be a density (chosen by the analyst). Let F^{*} denote the joint distribution which differs from the true distribution F of O only in that $f\left[a(t) \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right), V^{\dagger}\right]$ is replaced by the ancillary density $f^{*}\left[a(t) \mid \overline{A}\left(t^{-}\right), V^{\dagger}\right]$. Further, define $\mathcal{W}(t) \equiv \overline{W}(t) / f^{*}\left[\overline{A}(t) \mid V^{\dagger}\right]$ and $O^{\dagger} = \left(\overline{L}\left(C^{\dagger}\right), \overline{A}\left(C^{\dagger}\right)\right)$. A key result, which follows from direct calculation, is

Lemma 3.1: For any $z\left(O^{\dagger}\right)$, $E\left[z\left(O^{\dagger}\right) / \mathcal{W}\left(C^{\dagger}\right) \mid V^{\dagger}\right] = E^{*}\left[z\left(O^{\dagger}\right) \mid V^{\dagger}\right]$.

Remark: It is false that $E\left[z(O) / \mathcal{W}(C) \mid V^{\dagger}\right] = E^{*}\left[z(O) \mid V^{\dagger}\right]$. Let $D^{*}_{sm}(\beta, h)$ be the probability limit under F^{*} of $\widehat{D}^{*}_{sm}(\beta, h)$ and let $\{U_{sm}(\beta, h)\}$ and $\{D_{sm}(\beta, h)\}$ be the subsets of $\{U^{*}_{sm}(\beta, h)\}$ and $\{D^{*}_{sm}(\beta, h)\}$, respectively, that depend on the data only through $O^{\dagger}$. Set $\mathcal{W} = \mathcal{W}\left(C^{\dagger}\right)$ and note $D_{sm}(\beta, h)$ and $U_{sm}(\beta, h)$ often depend on $E^{*}\left[\cdot \mid V^{\dagger}\right] = E\left[\cdot / \mathcal{W} \mid V^{\dagger}\right]$ or $E^{*}[\cdot] = E[\cdot / \mathcal{W}]$. Define $\widehat{D}_{sm}(\beta, h)$ and $\widehat{U}_{sm}(\beta, h)$ like $D_{sm}(\beta, h)$ and $U_{sm}(\beta, h)$ except replace any unknown expectations $E\left[\cdot / \mathcal{W} \mid V^{\dagger}\right]$ and $E[\cdot / \mathcal{W}]$ with appropriate estimates $\widehat{E}\left[\cdot / \mathcal{W} \mid V^{\dagger}\right]$ and $\widehat{E}[\cdot / \mathcal{W}]\cdot$.

Examples: In model 1b, with $\eta(x) = x$, $U_{sm}(\beta, h) = \widehat{D}_{sm}(\beta, h) = U^{*}_{sm}(\beta, h)$ has $h_1\left(\overline{A}, V^{\dagger}\right)$ and $h_2\left(\overline{A}, V^{\dagger}\right)$ being functions only of $\left\{\overline{A}\left(C^{\dagger}\right), V^{\dagger}\right\}$. Note $E^{*}\left[h_2\left(\overline{A}, V^{\dagger}\right) \mid V^{\dagger}\right]$ is known and need not be estimated.

In contrast, in models 2a and 2b, $\widehat{D}_{sm}(\beta, h)$ will be defined like $\widehat{D}^{*}_{sm}(\beta, h)$ except in defining $\widetilde{J}(h, \beta)$ we replace $I(T > u)$ by $I(T > u) / \mathcal{W}$ in order to estimate the unknown expectations.

In model 1b with $\eta(x) = \ln[x/(1-x)]$, $\widehat{g}\left(V^{\dagger}\right) \equiv \widehat{g}\left(V^{\dagger}, \beta\right)$ could be chosen to minimize $\widetilde{E}\left[\left(Y - expit\left\{g\left(\overline{A}, V^{\dagger}, \beta\right) + g^{\dagger}\left(V^{\dagger}\right)\right\}\right)^{2} / \mathcal{W}\right]$ over $g^{\dagger}\left(V^{\dagger}\right)$ in some class (e.g., splines), whose dimension may increase with sample size. In the Appendix, we sketch a proof of the following.

Theorem 3.1: Subject to regularity conditions, in the semiparametric model *(i)* characterized by (4), (11), the data O, and a MSM, $\left\{\widehat{\beta}(h, \phi)\right\}$ solving $0 = \sum_i \widehat{D}_i(\beta, h, \phi)$ with $\left\{\widehat{D}(\beta, h, \phi)\right\} = \left\{\widehat{D}_{sm}(\beta, h) / \mathcal{W} + D_{tp}(\phi)\right\}$ is a class of RAL estimators with influence functions $\mathcal{IF} = \{IF(h, \phi)\}$, $IF(h, \phi) = \{\kappa(h)\}^{-1} U(\beta_0, h, \phi)$, $\kappa(h) = -\partial E[U(\beta, h, \phi)] / \partial \beta_{|\beta=\beta_0}$, $U(\beta_0, h, \phi) = U_{sm}(\beta_0, h) / \mathcal{W} + U_{tp}(\phi)$, where $U_{tp}(\phi) = D_{tp}(\phi)$ is defined like $U^{*}_{tp}(\phi)$ except with the true law F replacing F^{*}. Furthermore, $\mathcal{IF}$ is the set of all influence functions.

3.3. Efficiency for fixed h

By a projection argument similar to that given in Robins et al. (1994), we have

Theorem 3.2: For a given h, among all estimators $\widehat{\beta}(h, \phi)$, the most efficient has ϕ equal to $\phi_{opt} \equiv \phi_{opt}(h)$: if $A(t)$ only jumps at non-random times $0, 1, 2, \ldots$ then $\phi_{opt} \equiv 0$ if $k > C^{\dagger}$, and if $k \leq C^{\dagger}$, $\phi_{opt}\left[k, \overline{a}(k), \overline{\ell}\left(k^{-}\right)\right] = E[U_{sm}(h) / \mathcal{W} \mid \overline{A}(k) = \overline{a}(k), \overline{L}\left(k^{-}\right) = \overline{\ell}\left(k^{-}\right)] =$
$\left\{\overline{W}(k)\right\}^{-1} \int\int d\mu\left(\underline{a}_{k+1}\right) f^{*}\left(\overline{a} \mid v^{\dagger}\right) E[u_{sm}\left\{\overline{a}\left(C^{\dagger}\right), \overline{Y}_{\overline{a}}\left(C^{\dagger}\right), V^{\dagger}, h\right\} \mid \overline{L}_{\overline{a}}\left(k^{-}\right) = \overline{\ell}\left(k^{-}\right), \overline{A}(k) = \overline{a}(k)]$ and $U_{sm}(h) \equiv U_{sm}(\beta_0, h)$. Furthermore, if CAR holds, $\overline{A}(k) = \overline{a}(k)$ can be removed from the last conditioning event above. If $A(t)$ jumps in continuous time, $\phi_{1,opt} = E[U_{sm}(h) / \mathcal{W} \mid \overline{L}(u^{-}), \overline{A}(u^{-}), A(u) \neq A(u^{-})] - E\left[U_{sm}(h) / \mathcal{W} \mid \overline{L}(u^{-}), \overline{A}(u^{-})\right]$ since $E\left[U_{sm}(h) / \mathcal{W} \mid \overline{L}(u^{-}), \overline{A}(u^{-})\right] =$
$E\left[U_{sm}(h) / \mathcal{W} \mid \overline{L}(u^{-}), \overline{A}(u^{-}), A(u) = \overline{A}(u^{-})\right]$, and $\phi_{2,opt} = E\left[U_{sm}(h) / \mathcal{W} \mid \overline{A}(u), \overline{L}(u^{-})\right]$.

Theorem 3.2: a) The semiparametric model *(ii)* characterized by (4), data O, and a MSM (with (11) not imposed), has the set of influence functions $\left\{IF\left(h, \phi_{opt}(h)\right)\right\}$.

b): In model *(iii)* characterized by (4), data O, a MSM, and a parametric model indexed by parameter α for $f\left[a(t) \mid \overline{L}\left(t^{-}\right), \overline{A}\left(t^{-}\right)\right]$, the set of influence functions for the model is the set $\left\{\kappa(h)^{-1}\left[U(h, \phi) - E\left[U(h, \phi) S'_{\alpha}\right]\left\{E\left[S_{\alpha} S'_{\alpha}\right]\right\}^{-1} S_{\alpha}\right]\right\}$ of influence functions of $\left\{\widehat{\beta}(h, \phi, \widehat{\alpha})\right\}$ solving $o_p(1) = n^{-\frac{1}{2}} \sum_i \widehat{D}_i(\beta, h, \phi, \widehat{\alpha})$ where $\widehat{\alpha}$ is the MLE of α, S_{α} is the subject-specific score for α, and $D(\beta, h, \phi, \widehat{\alpha})$

is $D(\beta, h, \phi)$ evaluated at $\widehat{\alpha}$.

Theorem 3.2b can be extended to semiparametric models for $f\left(a(t) \mid \overline{L}(t^-), \overline{A}(t^-)\right)$ such as a Cox proportional hazard model as in Robins (1993).

4. Semiparametric Efficiency

4.1. The Efficient Score

In any semiparametric model, the semiparametric variance bound is the inverse of the variance of the efficient score S_{eff}. The efficient score in models *(i) - (iii)* of Theorems 3.1 and 3.2 are the same and, by Theorem 5.3 in Newey and McFadden (1993), equal $S_{eff} = U\left(\beta_0, h_{eff}, \phi_{eff}\right)$ where $\phi_{eff} = \phi_{opt}(h_{eff})$ and h_{eff} is uniquely characterized by the requirement that for all $U(\beta_0, h, \phi)$

$$E\left[U(\beta_0, h, \phi) U\left(\beta_0, h_{eff}, \phi_{opt}(h_{eff})\right)'\right] = \kappa(h)$$

which is equal to

$$E\left[U_{sm}(\beta_0, h) U\left(\beta_0, h_{eff}, \phi_{opt}(h_{eff})\right)'\right] = \kappa(h). \tag{12}$$

To show how to use (12) to calculate h_{eff}, we consider the following simple example.

Model 1a: Consider MSM 1a with $C^\dagger = C = K + 1 = 1$ w.p.1 so $K = 0$ and the $\overline{A}(C)$ process only jumps at time zero. So $\overline{A} = A(0)$ and $\mathcal{W}^{-1} = f^*\left[\overline{A} \mid V^\dagger\right] / f\left[\overline{A} \mid L(0)\right]$ where $V^\dagger \subset L(0) = V(0)$. As the proof of the following theorem, available from the author, shows, we can let $f^*\left(\overline{A} \mid V^\dagger\right) = 1$ w.p.1 without worrying that it is not a density, because it can be absorbed into $h_{eff}\left(\overline{A}, V^\dagger\right)$.

Theorem 4.1: With $f^*\left(\overline{A} \mid V^\dagger\right) = 1$ w.p.1, Eq. 12 implies $h_{eff}\left(\overline{A}, V^\dagger\right)$ is the unique solution to the type two Fredholm equation $h_{eff}\left(\overline{A}, V^\dagger\right)\left[\int var\left[\varepsilon \mid \overline{A}, L(0)\right]\left\{f\left(\overline{A} \mid L(0)\right)\right\}^{-1} f\left(V^\bullet \mid V^\dagger\right) d\mu\left(V^\bullet\right)\right] + \int h_{eff}\left(\overline{a}, V^\dagger\right) \omega\left(\overline{a}, \overline{A}, V^\dagger\right) d\mu(\overline{a}) = \partial g\left(\overline{A}, V^\dagger, \beta_0\right) / \partial\beta$ where $V^\bullet = L(0) \backslash V^\dagger$ and $\omega\left(\overline{a}, \overline{A}, V^\dagger\right) = \left[\int E\left[\varepsilon \mid \overline{a}, L(0)\right] E\left[\varepsilon \mid \overline{A}, L(0)\right] f\left(V^\bullet \mid V^\dagger\right) d\mu\left(V^\bullet\right)\right]$. Note that if $\overline{A}$ has finite support, this is a finite dimensional matrix equation. Our estimators, specialized to this example, are continuous treatment extensions of efficient propensity score estimators of an average treatment effect.

4.2. Efficiency calculations using missing data theory

Given (4), imposing CAR cannot change the efficient score. Thus, it is of interest to rederive the efficient score using the Hilbert space results of van der Vaart (1991) and of Robins et al. (1994) for missing data models under CAR. For convenience, assume C is non-random and write $\overline{A} \equiv \overline{A}(C)$. The full data are $\overline{L}^F = \left\{\overline{L}_{\overline{a}}; \overline{a} \in \overline{\mathcal{A}}\right\}$. Given any $B = b\left(\overline{L}^F\right)$, the score operator $\mathbf{s}(B) = E[B \mid O], O = \left(\overline{A}, \overline{L}_{\overline{A}}\right)$. For any $Q = q(O)$, the non-parametric adjoint operator $\mathbf{s}^\dagger$ under CAR is $\mathbf{s}^\dagger(Q) = E\left[Q \mid \overline{L}^F\right] = \int d\mu(\overline{a}) q\left(\overline{a}, \overline{L}_{\overline{a}}\right) f\left[\overline{a} \mid \overline{L}_{\overline{a}}\right]$. Suppose for the remainder of this subsection that the $\overline{A}$ jumps only at times $0, 1, \ldots, K$ and $\overline{L}$ jumps at $0^-, \ldots, K + 1^-$ and $C = K + 1$. We then have by CAR

$$f\left[\overline{a}(k) \mid \overline{L}_{\overline{a}}\right] = \prod_{m=0}^{k} f\left[a(m) \mid \overline{a}(m-1), \overline{L}_{\overline{a}}(m)\right] \tag{13}$$

and $f\left[\overline{a} \mid \overline{L}_{\overline{a}}\right] = f\left[\overline{a}(K) \mid \overline{L}_{\overline{a}}\right]$. It is then easy to check the null space of $\mathbf{s}^\dagger$, $N\left(\mathbf{s}^\dagger\right) = \{U_{tp}(\phi)\}$. Now define the non-parametric information operator, $\mathbf{m} = \mathbf{s}^\dagger \mathbf{s} : \overline{L}^F \to R\left(\mathbf{s}^\dagger\right)$ where $R\left(\mathbf{s}^\dagger\right)$ is the range of $\mathbf{s}^\dagger$. Note that $R\left(\mathbf{s}^\dagger\right) = \left\{B = \int d\mu(\overline{a}) b\left(\overline{a}, \overline{L}_{\overline{a}}\right)\right\}$. Let $\mathbf{m}^{-1} : R\left(\mathbf{s}^\dagger\right) \to R\left(\mathbf{s}^\dagger\right)$ be the inverse of $\mathbf{m}$ on $R\left(\mathbf{s}^\dagger\right)$. Given $\overline{a}_1, \overline{a}_2$, let u_{12} be the smallest u with $a_1(u) \neq a_2(u)$. We then have by a direct calculation

Theorem 4.2: If for all $\overline{a}_1, \overline{a}_2$

$$\overline{L}_{\overline{a}_1} \coprod \overline{L}_{\overline{a}_2} \mid \overline{L}_{\overline{a}_1}\left(u_{12}^{-}\right) \tag{14}$$

then

$$\begin{aligned}&\mathbf{m}^{-1}\left[\int d\mu\left(\overline{a}\right) b\left(\overline{a}, \overline{L}_{\overline{a}}\right)\right] = \int d\mu\left(\overline{a}\right)\left\{\sum_{m=1}^{K+1}\left\{f\left[\overline{a}\left(m-1\right) \mid \overline{L}_{\overline{a}}\right]\right\}^{-1}\right. \\ &\left.\left\{E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \overline{L}_{\overline{a}}\left(m\right)\right] - E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \overline{L}_{\overline{a}}\left(m-1\right)\right]\right\} + E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \overline{L}_{\overline{a}}\left(0\right)\right]\right\} .\end{aligned} \tag{15}$$

Remark: Since (14) places no restriction on the law of the observed data O, we can and do always assume that (14) holds.

Now let S_{eff}^{F} and $\Lambda^{F,\perp}$ be the efficient score and the orthogonal complement to the nuisance tangent space for the parameter β of our marginal structural model when we have data on $\overline{L}^{F}$. Then the efficient score S_{eff} based on data O under CAR is $g\left[m^{-1}\left(D_{eff}\right)\right]$ where D_{eff} is the unique member of $\Lambda^{F,\perp} \cap R\left(\mathbf{s}^{\dagger}\right)$ satisfying

$$\Pi\left[\mathbf{m}^{-1}\left(D_{eff}\right) \mid \Lambda^{F,\perp}\right] = S_{eff}^{F} , \tag{16}$$

where Π is the Hilbert space projection operator. To show how to use this result to calculate S_{eff}, we revisit the example given in the last subsection.

Example: Model 1a: Consider MSM 1a as in Sec. 4.1. Then, by an extension of Theorem 8.3 of Robins et al. (1994)

$$\Lambda^{F,\perp} = \left\{\int d\mu\left(\overline{a}\right) h\left(\overline{a}\right) \varepsilon\left(\overline{a}\right)\right\} \tag{17}$$

where *(i)* $\varepsilon\left(\overline{a}\right) = \varepsilon\left(\overline{a}, \beta_0\right)$, *(ii)* $h\left(\overline{a}\right)$ is a vector valued function of the dimension of β_0. Note that $\Lambda^{F,\perp}$ is contained in $R\left(\mathbf{s}^{\dagger}\right)$ as will be the case for MSMs with positive information.

Remark:

$$\text{If } \overline{\mathcal{A}} = \left\{\overline{a}_1, \ldots, \overline{a}_S\right\} \text{ is finite,} \tag{18}$$

then $\varepsilon\left(\overline{a}\right)$ can be identified with the S vector that has components $\varepsilon_s\left(\overline{a}_s\right) = Y_{\overline{a}_s} - g\left(\overline{a}_s, V^{\dagger}, \beta_0\right)$. For arbitrary $\overline{\mathcal{A}}$, $\varepsilon\left(\overline{a}\right)$ is a stochastic process with index set $\overline{\mathcal{A}}$. For $\overline{a}, \overline{a}^{*} \in \overline{\mathcal{A}}$, let $\mathbf{cv}\left(\overline{a}, \overline{a}^{*}\right) = cov\left(\varepsilon\left(\overline{a}\right), \varepsilon\left(\overline{a}^{*}\right)\right)$. If $\overline{\mathcal{A}}$ is given by (18), $\mathbf{cv}\left(\overline{a}, \overline{a}^{*}\right)$ corresponds to the $S \times S$ matrix with j,k entry $\mathbf{cv}\left(\overline{a}_j, \overline{a}_k\right)$. Let $\mathbf{cv}^{-1}\left(\overline{a}^{**}, \overline{a}^{*}\right)$ be a (generalized) inverse of $\mathbf{cv}\left(\overline{a}, \overline{a}^{*}\right)$, i.e., by definition, for any function $q\left(\overline{a}^{*}\right)$,
$\int\left[\int \mathbf{cv}^{-1}\left(\overline{a}^{**}, \overline{a}\right) \mathbf{cv}\left(\overline{a}, \overline{a}^{*}\right) d\mu\left(\overline{a}\right)\right] q\left(\overline{a}^{*}\right) d\mu\left(\overline{a}^{*}\right) = q\left(\overline{a}^{**}\right)$. In particular, if (18) holds, $\mathbf{cv}^{-1}\left(\overline{a}^{*}, \overline{a}\right)$ is just the inverse of the matrix identified with $\mathbf{cv}\left(\overline{a}, \overline{a}^{*}\right)$. Then, generalizing Chamberlain (1987)

$$S_{eff}^{F} = \int d\mu\left(\overline{a}\right)\left\{\partial g\left(\overline{a}, V^{\dagger}; \beta_0\right) / \partial\beta\right\}\left[\int \mathbf{cv}^{-1}\left(\overline{a}, \overline{a}^{*}\right) \varepsilon\left(\overline{a}^{*}\right) d\mu\left(\overline{a}^{*}\right)\right] . \tag{19}$$

If $\overline{\mathcal{A}}$ is given by (18), $\partial g\left(\overline{a}, V^{\dagger}; \beta_0\right) / \partial\beta$ can be identified with the $\dim\beta \times S$ matrix with j,k entry $\partial g\left(\overline{a}_k, V^{\dagger}; \beta_0\right) / \partial\beta_j$. Again, generalizing Theorem 8.3 in Robins et al. (1994),

$$\begin{aligned}&\Pi\left[\int d\mu\left(\overline{a}\right) b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \mid \Lambda^{F,\perp}\right] = \\ &\int d\mu\left(\overline{a}\right) E\left[b\left(\overline{a}, \overline{L}_{\overline{a}}\right) \varepsilon\left(\overline{a}\right) \mid V^{\dagger}\right]\left[\int \mathbf{cv}^{-1}\left(\overline{a}, \overline{a}^{*}\right) \varepsilon\left(\overline{a}^{*}\right) d\mu\left(\overline{a}^{*}\right)\right] .\end{aligned} \tag{20}$$

Hence to solve (16), we need to find the solution $h_{eff}\left(\overline{a}, V^{\dagger}\right)$ to the equation

$$E[\mathbf{m}^{-1}\left\{\int d\mu\left(\overline{a}^{*}\right) h\left(\overline{a}^{*}, V^{\dagger}\right) \varepsilon\left(\overline{a}^{*}\right)\right\} \varepsilon\left(\overline{a}\right) \mid V^{\dagger}] = \partial g\left(\overline{a}, V^{\dagger}; \beta_{0}\right) / \partial \beta \ . \tag{21}$$

By (15) with $K = 0$ and the fact that, by CAR, $f\left(\overline{a} \mid \overline{L}_{\overline{a}}\right) = f\left(\overline{a} \mid L\left(0\right)\right)$, the LHS of (21) can be written

$$E\{\int d\mu\left(\overline{a}^{*}\right) h\left(\overline{a}^{*}, V^{\dagger}\right)\left\{\left[\varepsilon\left(\overline{a}^{*}\right) - E\left[\varepsilon\left(\overline{a}^{*}\right) \mid L\left(0\right)\right]\right]\left\{f\left(\overline{a}^{*} \mid L\left(0\right)\right)\right\}^{-1} + \left[\varepsilon\left(\overline{a}^{*}\right) \mid L\left(0\right)\right]\right\} \varepsilon\left(\overline{a}\right) \mid V^{\dagger}\} =$$
$$E\{\int d\mu\left(\overline{a}^{*}\right) h\left(\overline{a}^{*}, V^{\dagger}\right)\left[cov\left[\varepsilon\left(\overline{a}^{*}\right), \varepsilon\left(\overline{a}\right) \mid L\left(0\right)\right]\left\{f\left(\overline{a}^{*} \mid L\left(0\right)\right)\right\}^{-1} + E\left[\varepsilon\left(\overline{a}^{*}\right) \mid L\left(0\right)\right] E\left[\varepsilon\left(\overline{a}\right) \mid L\left(0\right)\right]\right] | V^{\dagger}\} \ .$$

However, since by assumption (14), $Y_{\overline{a}_j} \coprod Y_{\overline{a}_k} \mid \bar{L}\left(0\right)$ for $k \neq j$, (21) reduces to

$$h\left(\overline{a}, V^{\dagger}\right) E\{var\left[\varepsilon\left(\overline{a}\right) \mid L\left(0\right)\right] f\left(\overline{a} \mid L\left(0\right)\right)^{-1} \mid V^{\dagger}\} +$$
$$\int d\mu\left(\overline{a}^{*}\right) h\left(\overline{a}^{*}, V^{\dagger}\right) E\{E\left[\varepsilon\left(\overline{a}^{*}\right) \mid L\left(0\right)\right] E\left[\varepsilon\left(\overline{a}\right) \mid L\left(0\right)\right] \mid V^{\dagger}\} = \partial g\left(\overline{a}, V^{\dagger}, \beta_{0}\right) / \partial \beta.$$

Upon noting that, by CAR, $f\left[\varepsilon \mid \overline{A} = \overline{a}, L\left(0\right)\right] = f\left[\varepsilon\left(\overline{a}\right) \mid L\left(0\right)\right], w$. We see that this is the same expression for $h_{eff}\left(\overline{a}, V^{\dagger}\right)$ as obtained in Theorem 4.1.

4.3. A practical approach to obtaining reasonable efficiency

Estimation of h_{eff} is computationally difficult because of the need to solve integral equations without closed form solutions. A practical approach to choosing h and $f^{*}\left(\overline{a} \mid V^{\dagger}\right)$ is important. Given a model for $f\left[a\left(t\right) \mid \overline{a}\left(t^{-}\right), \overline{\ell}\left(t^{-}\right)\right]$ depending on parameter $\alpha' = (\alpha_1', \alpha_2')$ such that $\alpha_1 = 0 \Leftrightarrow f\left[a\left(t\right) \mid \overline{a}\left(t^{-}\right), \overline{\ell}\left(t^{-}\right)\right] = f\left[a\left(t\right) \mid \overline{a}\left(t^{-}\right), v^{\dagger}\right]$, rather than choosing $f^{*}\left[\overline{a} \mid V^{\dagger}\right]$, we use $f^{*}\left[\overline{a} \mid V^{\dagger}; \widetilde{\alpha}_2\right]$ where $\widetilde{\alpha}_2$ is the MLE of α_2 with α_1 set to zero. [The fact that $f^{*}\left[\overline{a} \mid V^{\dagger}\right]$ is estimated does not influence the asymptotic distribution of $\widehat{\beta}\left(h, \phi\right)$.] It follows that if (8) holds [i.e., $\overline{A}$ is an ancillary process], $\mathcal{W}$ will converge to 1. When (8) is false, any variation in $\mathcal{W}$ will be unavoidable variation. Further, in each of the models 1a – 2c of Sec. 3.1, the efficient choice of h, say, h_{opt}, for solving $\sum_i \widehat{D}^{*}_{sm,i}\left(h\right) = 0$ when (9) is not imposed is well known. We suggest choosing h to be h_{opt} or an estimate $\widehat{h}_{opt}$ thereof, and choosing ϕ to be an estimate of $\phi_{opt}\left(\widehat{h}_{opt}\right)$. Such a choice guarantees that if $\overline{A}$ is an ancillary process, our estimate of β_0 will be more efficient than the estimate based on solving $0 = \sum_i \widehat{D}_{sm,i}\left(h_{opt}\right)$. Specifically, in MSM 1a, $h_{opt} = \left\{\partial \varepsilon\left(\beta_0\right) / \partial \beta\right\}\left\{var\left(\varepsilon\left(\beta_0\right) \mid \overline{A}, V^{\dagger}\right)\right\}^{-1}$. For model 1b, Chamberlain (1988) gives $h_{1,opt}$ and $h_{2,opt}$. In model 1c, $h_{opt} = \left[\partial / \partial \beta\right]\left[ln\left\{\left[\partial R\left(\beta_0\right) / \partial Y\right] f\left[R\left(\beta_0\right) \mid V^{\dagger}\right]\right\}\right]$. In model 1d, h_{opt} is as in model 1a with $\varepsilon\left(\beta_0\right)$ now a vector. In model 2a, $h_{opt} = \partial ln r\left[\overline{A}\left(u^{-}\right), u, V^{\dagger}, \beta_0\right] / \partial \beta$. For model 2b, h_{opt} is given by Sasieni (1992). In model 2c, $h_{opt,2} = h_{opt,1} \lambda_0\left(u \mid V^{\dagger}\right)$ and $h_{opt,1} = \partial \ln \lambda_{R(\beta)}\left(u \mid V^{\dagger}\right) / \partial \beta_{\mid \beta = \beta_0}$.

5. Comparison of MSMs and SNMs

5.1. Structural Nested Distribution Models

For concreteness, we consider the setting of the MSM model 1a - 1c with $Y_{\overline{a}} = Y_{\overline{a}}\left(K+1\right)$ and the A process and L process jumping at non-random times $0, \ldots, K$ and $0^{-}, \ldots, K+1^{-}$ respectively. Henceforth, we take $V^{\dagger} = \emptyset$. Suppose Y is a continuous variable. Then let $\gamma\left(y, \overline{\ell}\left(m\right), \overline{a}\left(m\right)\right)$ be the unique function mapping quantiles of $Y_{\overline{a}(m),0}$ into those of $Y_{\overline{a}(m-1),0}$ conditional on $\overline{L}\left(m\right) = \overline{\ell}\left(m\right), \overline{A}\left(m\right) = \overline{a}\left(m\right)$. A structural nested distribution model (SNDM) specifies that $\gamma\left(y, \overline{\ell}\left(m\right), \overline{a}\left(m\right)\right) = \gamma\left(y, \overline{\ell}\left(m\right), \overline{a}\left(m\right), \beta_0\right)$

where $\gamma\left(y,\overline{\ell}\left(m\right),\overline{a}\left(m\right),\beta\right)$ is a known increasing function of y satisfying $\gamma\left(y,\overline{\ell}\left(m\right),\overline{a}\left(m\right),\beta\right)=y$ if $a\left(m\right)=0$ or $\beta=0$. Recursively define random variables $\overset{\bullet}{R}_K\left(\beta\right),\ldots,\overset{\bullet}{R}_0\left(\beta\right)$ by $\overset{\bullet}{R}_K\left(\beta\right)=\gamma\left(Y,\overline{L}\left(K\right),\overline{A}\left(K\right),\beta\right)$ and $\overset{\bullet}{R}_m\left(\beta\right)=\overset{\bullet}{r}_m\left(Y,\overline{L}\left(K\right),\overline{A}\left(K\right),\beta\right)=\gamma\left(\overset{\bullet}{R}_{m+1}\left(\beta\right),\overline{L}\left(m\right),\overline{A}\left(m\right),\beta\right)$ and set $\overset{\bullet}{R}\left(\beta\right)=\overset{\bullet}{r}\left(Y,\overline{L}\left(K\right),\overline{A}\left(K\right),\beta\right)\equiv\overset{\bullet}{R}_0\left(\beta\right)$. Also let $\overset{\bullet}{r}^{-1}\left(y,\overline{\ell}\left(K\right),\overline{a}\left(K\right),\beta\right)$ be the inverse of the function $\overset{\bullet}{r}$ with respect to its first argument. If $\gamma\left(y,\overline{\ell}\left(m\right),\overline{a}\left(m\right),\beta\right)=\gamma\left(y,\overline{a}\left(m\right),\beta\right)$ does not depend on $\overline{\ell}\left(m\right)$, we say that the SNDM model has no interaction.

Theorem 5.1: Under (4), a no interaction SNDM model is a stratified transformation model (STM), i.e., MSM 1c, with $R\left(\overline{a},\beta\right)=\overset{\bullet}{r}\left(Y,\overline{a},\beta\right)$ and $R\left(\beta\right)=\overset{\bullet}{R}\left(\beta\right)$. However, the converse is not true.

The semiparametric information bound for β is greater if we impose a no-interaction SNDM than if we only imposed the corresponding STM. Theorem 5.1 indicates that a STM is the natural MSM analog of a SNDM. If $\gamma\left(y,\overline{\ell}\left(m\right),\overline{a}\left(m\right)\right)$ depends on $\overline{\ell}\left(m\right)$, we must choose between analyzing the data under a SNDM versus a STM. To understand the advantages and disadvantages of each, we need some additional background. Define a regime $g=\left(g_0,\ldots,g_K\right)\in\mathcal{G}$ to be a collection of functions $g_m:\overline{\mathcal{L}}_m\rightarrow\mathcal{A}_m$. Define $g\left(\overline{\ell}\left(m\right)\right)=\left\{g_0\left(\bar{\ell}_0\right),\ldots,g_m\left(\overline{\ell}\left(m\right)\right)\right\}$. Let Y_g be the counterfactual value of Y if regime g were followed. If $g\left(\overline{\ell}_K\right)=\overline{a}_K\equiv\overline{a}$ does not depend on $\overline{\ell}_K$, then $Y_g=Y_{\overline{a}}$ and we say g is non-dynamic; otherwise, g is dynamic. Let $g\left[\overline{\ell}\left(k\right)\right]$ denote a realization of $\overline{A}\left(k\right)$. If

$$Y_g\coprod A\left(t\right)\mid\overline{L}\left(t^-\right),\overline{A}\left(t^-\right)\ , \tag{22}$$

then, by Theorem 3.2 of Robins (1997), the law of Y_g is given by the G-computation algorithm formula

$$F_{Y_g}\left(y\mid\overline{L}\left(k\right),g\left[\overline{\ell}\left(k-1\right)\right]\right)=$$
$$\iint F_Y\left(y\mid\overline{\ell}\left(K\right),g\left(\overline{\ell}\left(K\right)\right)\right)\prod_{m=k+1}^{K}dF\left[\ell\left(m\right)\mid\overline{\ell}\left(m-1\right),g\left(\overline{\ell}\left(m-1\right)\right)\right] \tag{23}$$

which, for continuous Y and $k=-1$, equals

$$F_{Y_g}\left(y\right)=\iint I[\overset{\bullet}{r}^{-1}\left\{u,\overline{\ell}\left(K\right),g\left(\overline{\ell}\left(K\right)\right),\beta_0\right\}>y]$$
$$\prod_{m=0}^{K}dF[\ell\left(m\right)\mid\overline{\ell}\left(m-1\right),g\left(\overline{\ell}\left(m-1\right)\right),\overset{\bullet}{R}\left(\beta_0\right)=u]dF_{\overset{\bullet}{R}\left(\beta_0\right)}\left(u\right). \tag{24}$$

In many settings, the g-null hypothesis that

$$F_{Y_{g1}}\left(y\right)=F_{Y_{g2}}\left(y\right)\ \text{ for all } g1,g2\in\mathcal{G} \tag{25}$$

will be of interest. Robins (1997) proves the following.

Theorem 5.2: Given (22), (25) holds $\Leftrightarrow$ (23) is the same for all $g\Leftrightarrow\gamma\left(y,\overline{\ell}_m,\overline{a}_m\right)=y\Leftrightarrow$

$$Y\coprod A\left(k\right)\mid\overline{L}\left(k\right),\overline{A}\left(k-1\right),k=0,\ldots,K. \tag{26}$$

5.2. Advantages of SNDMs with a continuous Y

1. Although (25) implies $\beta_0=0$ for both a SNDM and a STM, only for a SNDM is (25) equivalent to $\beta_0=0$.

2. If the $L\left(k\right)$ are discrete with only a moderate number of levels, then, even with $f\left[a\left(k\right)\mid\overline{\ell}\left(k-1\right),\overline{a}\left(k-1\right)\right]$

totally unrestricted, an asymptotically distribution-free g-null test of $\beta_0 = 0$ (and thus of (25)) exists for a SNDM but, because of the curse of dimensionality, not for a STM. Specifically, a non-parametric g-null test is equivalent to a test of lack of correlation of Y with $A(k)$ within strata defined jointly by $\overline{L}(k), \overline{A}(k-1)$ (Robins, 1997). Thus, even if $A(k)$ is continuous, a test of lack of correlation of $A(0)$ with Y within levels of $L(0)$ will be an asymptotic α - level test under (25). In contrast, a test of $\beta_0 = 0$ in a STM (without (25) additionally imposed) requires, by Theorem 3.2a that $\mathcal{W}$ can be consistently non-parametrically estimated which will not be possible due to the curse of dimensionality.

3. Henceforth, assume a correct model for $f\left[a(t) \mid \overline{\ell}(t^-), \overline{a}(t^-)\right]$ is available. Given a SNDM, with some difficulty the law of Y_g for dynamic g can be estimated using (24). In contrast, the law of Y_g for dynamic g is very hard to estimate given a STM. Specifically, given a SNDM, we estimate the law of Y_g as follows: *(i)* obtain an estimate $\widehat{\beta}$ by g-estimation (Robins, 1997), *(ii)* estimate $F_{\overset{\bullet}{R}(\beta_0)}(u)$ by the empirical law of the $\overset{\bullet}{R}_i\left(\widehat{\beta}\right)$, *(iii)* specify and estimate a parametric model for $f\left[L(m) \mid \overline{L}(m-1), \overline{A}(m-1), \overset{\bullet}{R}\left(\widehat{\beta}\right)\right]$, *(iv)* and then evaluate the estimated version of the integral (24) by Monte carlo.

In contrast, given a STM, we must, as discussed in Robins (1997, pg. 114; 1998, Sec. 11), specify a parametric model for $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ where $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right) = \nu\left(y, \overline{\ell}(m), \overline{a}\right) - \nu\left(y, \left\{\overline{\ell}(m-1), \ell(m) = 0\right\}, \overline{a}\right)$ and $\nu\left(y, \overline{\ell}(m), \overline{a}\right)$ maps quantiles of $Y_{\overline{a}}$ given $\overline{\ell}_m, \overline{a}_{m-1}$ into quantiles of $Y_{\overline{a}}$ given $\overline{\ell}_{m-1}, \overline{a}_{m-1}$. As discussed in Robins (1997, pg. 114-116; 1998, Sec. 11), estimation of $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ is a computational nightmare; indeed, fully parametric Bayesian or likelihood-based inference for a MSM is computationally extremely burdensome.

4. As discussed in Robins (1997, Sec. 9), for SNDMs it is easy to perform a sensitivity analysis in which the fundamental assumption (22) of ignorable treatment assignment is no longer imposed. For a STM such a sensitivity analysis is much less straightforward and cannot be based on estimating equations similar to those in Theorem 3.1.

5. A parameter β_0 of a SNDM, in contrast to that of a STM, can often still be consistently estimated if (22) is false but data are available on an instrumental variable. Specifically, suppose $A(t) = (A_1(t), A_2(t))$ with $A_1(t)$ recording a physician's prescribed treatment and $A_2(t)$ recording treatment actually received. We might suppose (22) is false, but $A_1(t) \coprod Y_g \mid \overline{L}(t^-), \overline{A}(t^-)$ is true if a predictor of Y_g and of $A_2(t)$ was not recorded in $\overline{L}(t^-)$. $A_1(t)$ is often then referred to as an instrumental variable process, particularly when $A_1(t)$ has no direct causal effect, i.e., $Y_{\overline{a}} = Y_{\overline{a}_2}$ w.p.1. In this setting, the parameter of a STM is not identified but the parameter of a SNDM can still in general be consistently estimated by g-estimation (Robins, 1993).

5.3. Advantages of MSMs with continuous Y

1. Even in the presence of interaction [i.e., $\gamma\left(y, \overline{\ell}_m, \overline{a}_m\right)$ depends on $\overline{\ell}_m$], given a STM, $F_{Y_{\overline{a}}}(y)$ can be estimated by $n^{-1}\sum_i I\left\{r^{-1}\left[R_i\left(\widehat{\beta}\right), \overline{a}, \widehat{\beta}\right] > y\right\}$ without requiring either integration or modelling of the conditional law of $L(m)$. In contrast, as described in point 3 of Sec. 5.2 above, for a SNDM, both integration and modelling are required.

2. Any MSM that can be easily estimated when (8) holds (i.e., $\overline{A}$ is an ancillary process) can be easily estimated when (8) is false. For example, we can use the Cox proportional hazards MSM 2a for a continuous failure time outcome $T_{\overline{a}}$. In contrast, a structural nested Cox model would model the ratio of the conditional hazard given $\overline{\ell}_m, \overline{a}_m$ of $Y_{\overline{a}(m),0}$ to that of $Y_{\overline{a}(m-1),0}$. Unfortunately, a structural nested Cox model does not admit any simple semiparametric estimators, and even complex estimators will fail due to the curse of dimensionality.

A hybrid approach is to impose a MSM model and then specify a model for $\nu^*\left(y, \overline{\ell}(m), \overline{a}\right)$ whose

parameter is the product of the parameter β of the MSM model with another parameter ψ, so that ψ is identified only if $\beta \neq 0$. Such a model can overcome objections 1 and 2 of Sec. 5.2 (but not objections 3-5) while retaining the advantages 1-2 of Sec. 5.3.

5.4. Structural Nested Mean Models (SNMMs)

SNMMs are applicable to discrete and continuous outcomes. Let $\gamma\left(\overline{\ell}(m), \overline{a}(m)\right) = E\left[Y_{\overline{a}(m),0} - Y_{\overline{a}(m-1),0} \mid \overline{\ell}(m), \overline{a}(m)\right]$. Let $\gamma^{\dagger}\left(\overline{\ell}(m), \overline{a}(m)\right) = ln\{E\left[Y_{\overline{a}(m),0} \mid \overline{\ell}(m), \overline{a}(m)\right] / E\left[Y_{\overline{a}(m-1),0} \mid \overline{\ell}(m), \overline{a}(m)\right]\}$. A structural nested mean model (SNMM) specifies $\gamma\left(\overline{\ell}(m), \overline{a}(m)\right) = \gamma\left(\overline{\ell}(m), \overline{a}(m), \beta_0\right)$ with $\gamma\left(\overline{\ell}_m, \overline{a}_m, \beta\right)$ a known function satisfying $\gamma\left(\overline{\ell}_m, \overline{a}_m, \beta\right) = 0$ if $a_m = 0$ or $\beta = 0$. A multiplicative SNMM specifies $\gamma^{\dagger}\left(\overline{\ell}(m), \overline{a}(m)\right) = \gamma\left(\overline{\ell}(m), \overline{a}(m), \beta_0\right)$. The g-null mean hypothesis is the hypothesis

$$E[Y_{g1}] = E[Y_{g2}], g_1, g_2 \in \mathcal{G}. \tag{27}$$

Robins (1997) proves the following.

Theorem 5.2: Given (22), (27) holds if and only if $\gamma\left(\overline{\ell}(m), \overline{a}(m)\right) = 0 \Leftrightarrow \gamma^{\dagger}\left(\overline{\ell}(m), \overline{a}(m)\right) = 0 \Leftrightarrow E\left[Y \mid \overline{A}(k), \overline{L}(k)\right] = E\left[Y \mid \overline{A}(k-1), \overline{L}(k)\right], k = 0, \ldots, K$:

Advantages *(1) - (4)* of Sec. 5.2 of a SNM over a MSM for continuous Y also will hold (appropriately modified) for discrete Y when considering the g-null mean hypothesis or when estimating $E[Y_g]$. Advantage *(5)* is no longer so clear, since estimation of a SNMM when (22) is false is not as straightforward as for a SNDM. Advantages *(1) - (2)* in Sec. 5.3 of a MSM over a SNDM for continuous outcome also hold in the discrete case.

An important advantage of MSMs over SNMs with Y dichotomous (or, more generally, when Y has finite support) is that neither a SNMM or multiplicative SNMM naturally imposes the fact that, for dichotomous Y, $E[Y_g] \in [0, 1]$ (and logistic SNMMs have not been developed). In contrast, using the MSM model 1a with $g(\overline{a}, \beta)$ a logistic function, the above restriction is naturally imposed. Analogously, in the setting of MSM model 1d, we can use standard marginal logistic models for the $Y_{\overline{a}}(m)$.

5.5. Direct effect models

In this section we show that some of the advantages of MSMs described in Secs. 5.3 and 5.4 are not retained in semiparametric models for the effect of a treatment a_1 when a second treatment a_2 is held fixed (set). In such a setting, both MSMs and direct effect SNMs (Robins, 1998) have important limitations due to the curse of dimensionality. Let $a(u) = (a_1(u), a_2(u))$ and, in a slight abuse of notation, set $\overline{a}(u) = (\bar{a}_1(u), \bar{a}_2(u))$ and $\overline{a} = (\bar{a}_1, \bar{a}_2)$. Continue to assume $V^{\dagger} = \emptyset$. Consider the following.

Model 3a – direct effect semiparametric regression: Consider the set-up of MSM 1b, with

$$\eta\{E[Y_{\overline{a}}]\} = g[\overline{a}, \beta_0] + g^{\dagger}(\overline{a}_2)$$

where $g[\overline{a}, \beta_0] = 0$ if $\overline{a}_1 \equiv 0$ and $g^{\dagger}(\cdot, \cdot)$ is unknown and unrestricted. Since, according to the model, $E[Y_{\overline{a}_1, \overline{a}_2} - Y_{\overline{a}_1 \equiv 0, \overline{a}_2}] = g(\overline{a}, \beta_0)$, it follows we are modelling the direct effect of treatment $\overline{a}_1$. Furthermore, the main effect of the second treatment $g^{\dagger}(\overline{a}_2) = E[Y_{\overline{a}_1 \equiv 0, \overline{a}_2} - Y_{\overline{a}_1 \equiv 0, \overline{a}_2 \equiv 0}]$ is completely unrestricted. The model for the observables O induced by MSM 3a is isomorphic to that induced by MSM 1b with $\overline{A}_2 \equiv \overline{A}_2(K)$ playing the role of $V^{\dagger}$ (under sequential randomization assumption (4)). In particular, if $\eta(x) = x$ [or $\ln(x)$], $\widehat{\beta}(h, \phi)$ will perform well in moderate size samples, provided $f\left[a(t) \mid \overline{L}(t^-), \overline{A}(t^-)\right]$

is known or can be parametrically modelled. However, as discussed in the final remark of Sec. 3.1, if $\eta(x) = ln[x/(1-x)]$, reasonable estimators of β_0 are unavailable because $\overline{A}_2(K)$ will be high-dimensional. Indeed, any choice of $\eta(x)$ that guarantees that $E[Y_{\overline{a}}] \in [0,1]$ will fail to provide reasonable estimators of β_0, negating the advantage of this MSM for dichotomous Y.

Model 3b – direct effect semiparametric Cox proportional hazard model: Consider the set up of MSM 2b, with

$$\lambda_{T_{\overline{a}}}(t) = \lambda_{T_{\overline{a}_1=0,\overline{a}_2}}(t) \exp\left[r\left\{\overline{a}\left(t^-\right), t; \beta_0\right\}\right]$$

with $r(\overline{a}(t^-), t; \beta_0) = 0$ if $\overline{a}_1(t^-) = \mathbf{0}$. This is a model for the direct effect of treatment $\overline{a}_1$ on the hazard of T with the main effect of $\overline{a}_2$ left unrestricted. Given (4), MSM 3b induces a model for the observables isomorphic to that induced by MSM 2b. This implies that, as discussed in the remark in Sec. 3.1, due to the curse of dimensionality, it will not be possible to obtain reasonable estimators of β_0 negating advantage 2 of Sec. 5.3.

Model 3c – direct effect semiparametric time-dependent accelerated failure time model: Consider the model

$$\lambda_{R(\overline{a},\beta_0)}(t) = \lambda_{T_{\overline{a}_1=0,\overline{a}_2}}(t)$$

where $R(\overline{a}, \beta) = r(T_{\overline{a}}, \overline{a}, \beta)$ satisfies $r(t, \overline{a}, \beta) = t$ if $\overline{a}_1 = \mathbf{0}$ or $\beta = 0$. The model for the observables induced by MSM 3c is isomorphic to that induced by MSM 2c with $\overline{A}_2$ in the role of $V^\dagger$. Hence, model 3c can be used to estimate the direct effect of $\overline{a}_1$ on T with the main effect of $\overline{a}_2$ unrestricted. MSM 3c is the natural MSM associated with a structural nested failure time model (SNFTM) (Robins, 1993, App. 1, 1998) since a direct-effect SNFTM without interaction is a MSM 3c. In the presence of interaction, the MSM 3c retains advantage 1 of Sec. 5.3.

6. Censoring

No new idea is required to account for and adjust for right censoring. Specifically, let Q be censoring time. Define a censoring process $A_2(u)$ by $A_2(u) = 0$ if $Q > u$ and $A_2(u) = 1$ otherwise. Let the treatment of interest be $a_1(u)$ and define $a(u) = (a_1(u), a_2(u))$. To want to adjust for censoring is only to say that interest is in the direct effect of $\overline{a}_1$ when $\overline{a}_2 \equiv 0$, i.e., when censoring is abolished. As a concrete example, the Cox model MSM 2a in the presence of censoring would become

$$\lambda_{T_{\overline{a}_1,\overline{a}_2\equiv 0}}\left(t \mid V^\dagger\right) = \lambda_0(t) \exp\left\{r\left[\overline{a}_1\left(t^-\right), t, V^\dagger; \beta_0\right]\right\}.$$

If $\overline{A}$ is ancillary (now including the censoring process), $\widehat{D}^*_{sm}(\beta, h)$ and $U^*_{sm}(\beta, h)$ are as above except that now $N_T(u) = I[T \le u] I[T < Q]$ and $I[T > u]$ is everywhere replaced by $I[T > u] I[Q > u]$. Of course $\overline{W}(t)$ is now the probability that a subject would have his observed treatment and censoring history.

Appendix 1:

By arguments as in Robins et al. (1994), Theorem 3.1 and 3.2b are easy corollaries of Theorem 3.2a.

Sketch of Proof of Theorem 3.2a: For convenience, assume the L process and A process jump at times $0^-, 1^-, \ldots$ and $0, 1, \ldots$ respectively. Then by Theorem 3.2 of Robins (1997), Eq. (4) implies the G-computation algorithm formula

$$f_{\overline{Y}_{\overline{a}}(k)}\left[\overline{y}(k) \mid v^\dagger\right] = \int \prod_{m=0}^{k} f[y(m), v(m) \mid \overline{y}(m-1), \overline{v}(m-1), \overline{a}(m-1), v^\dagger] \qquad \text{(A1)}$$

$$\prod_{m=1}^{k} d\mu[v(m)]\, d\mu\left(\overset{\bullet}{v}\right),$$

with $\overset{\bullet}{v} \equiv v(0) \backslash v^{\dagger}$. Thus, if for some $j < k$, the proposition $f\left[a(j) \mid \overline{a}(j-1), \overline{L}(j), v^{\dagger}\right] \neq 0$ *w.p.*1 given $V^{\dagger}$ is false, then (A1) is not identified. Hence, a MSM model places no (local) restrictions on $f\left[L(m) \mid \overline{L}(m-1), \overline{A}(m-1)\right]$ for $m > C^{\dagger}$. Hence, in semiparametric model *(ii)*, every function of O with mean zero given $O^{\dagger}$ is in the nuisance tangent space for the model. It follows that all members of $\Lambda^{\perp}$ in model *(ii)* depend on the data only through $O^{\dagger}$.

Because of our assumed knowledge of $\Lambda^{\perp,*}$ (the orthogonal complement to the nuisance tangent space under F^{*}), it is sufficient to show that $U \in \Lambda^{\perp} \Leftrightarrow U\mathcal{W} \in \Lambda^{\perp,*}$ when F^{*} is chosen such that $C^{\dagger,*}$ is equal to $C^{\dagger}$. This follows from the fact that $\overset{\circ}{\Lambda} = \overset{\circ}{\Lambda}^{*}$ where $\overset{\circ}{\Lambda} \equiv \Lambda \bigcap \{U_{tp}(\phi)\}^{\perp} \bigcap \left\{z\left(O^{\dagger}\right)\right\}$ and the fact that $E[UB] = E^{*}[U\mathcal{W}B]$ for any $B \in \overset{\circ}{\Lambda}$ by Lemma 3.1.

References

Chamberlain, G. (1987), Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, *J. Econometrics*, 34, 305-324.

Chamberlain, G. (1988). Efficiency bounds for semiparametric regression. Technical Report, Department of Statistics, University of Wisconsin.

Gill, R.D., van der Laan, M.J., & Robins, J.M. (1996). Coarsening at random: characterizations, conjectures and counterexamples. **Proceedings of the First Seattle Symposium on Survival Analysis**, to appear.

Heitjan, D.F., & Rubin, D.B. (1991), Ignorability and Coarse Data, **The Annals of Statistics**, 19, 2244-2253.

Newey, W.K. & McFadden, D. (1993). Estimation in large samples. **Handbook of Econometrics**, Vol. **4**. Eds. McFadden, D., Engler, R. Amsterdam: North Holland.

Robins, J.M. (1993). Analytic methods for estimating HIV treatment and cofactor effects. **Methodological Issues of AIDS Mental Health Research**. Eds: Ostrow D.G., Kessler R. New York: Plenum Publishing. pp. 213-290.

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Stat.*, 23:2379-2412.

Robins, J.M. (1997). Causal inference from complex longitudinal data. In: **Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)**, M. Berkane, Editor. NY: Springer Verlag, 69-117.

Robins, J.M., Rotnitzky, A., Zhao LP. (1994). Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89:846-866.

Robins, J.M. (1998). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. **Causation and Computation**, Ed. G. Cooper and C. Glymour, MIT/AAAI Press (to appear).

Sasieni, P. (1992). Information bounds for the conditional hazard ratio in a nested family of regression models. *J. Royal Stat. Soc. B*, 54:617-635.

van der Vaart, A.W. (1991). On differentiable functionals. *Ann. Stat.,* 19:178-204.