G-computation Estimation of Nonparametric Causal Effects on Time-Dependent Mean Outcomes in Longitudinal Studies

Romain Neugebauer[*] Mark J. van der Laan[†]

[*] Division of Biostatistics, School of Public Health, University of California, Berkeley, romain.s.neugebauer@kp.org

[†] Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

G-computation Estimation of Nonparametric Causal Effects on Time-Dependent Mean Outcomes in Longitudinal Studies

Romain Neugebauer and Mark J. van der Laan

Abstract

Two approaches to Causal Inference based on Marginal Structural Models (MSM) have been proposed. They provide different representations of causal effects with distinct causal parameters. Initially, a parametric MSM approach to Causal Inference was developed: it relies on correct specification of a parametric MSM. Recently, a new approach based on nonparametric MSM was introduced. This later approach does not require the assumption of a correctly specified MSM and thus is more realistic if one believes that correct specification of a parametric MSM is unlikely in practice. However, this approach was described only for investigating causal effects on mean outcomes collected at the end of longitudinal studies. In this paper we first generalize the nonparametric MSM approach to the investigation of causal effects on time-dependent outcomes, i.e. for outcomes collected throughout longitudinal studies. This article then develops the G-computation estimation of the corresponding nonparametric MSM parameters and compares its implementation to its analogue in the parametric MSM approach. Finally, we propose new algorithms to address an important computing limitation independent of the MSM approach chosen but inherent to the implementation of the G-computation estimator a) with continuous treatment and/or b) in longitudinal studies with long follow-up and time dependent outcomes. These new algorithms for the implementation of the G-computation estimator lead to a generalization of nonparametric causal effects and should allow broader application of these methodologies in real life studies. Results are illustrated with two simulation studies.

1 Introduction

1.1 Data structure and question of interest

For more details on the statistical framework used in this paper, see chapter 6 of van der Laan and Robins (2002) [5].

For all experimental units in a random population sample of size n, we observe a treatment regimen $(A(0), \ldots, A(K))$ over time $t = 0, \ldots, K$ and a covariate process $(L(0), \ldots, L(K+1))$ measured at baseline and after a new treatment is assigned. The covariate $L(t)$ is measured after $A(t-1)$ and before $A(t)$. Note that $K+1$ represents the length of the treatment regimen in the appropriate unit of time and n the sample size.

In the formal counterfactual framework for longitudinal study [5], the data are represented as n independent and identically distributed (i.i.d.) realizations of:

$$O = (L(0), A(0), L(1), A(1), \ldots, L(K), A(K), L(K+1)) = (\bar{A}(K), \bar{L}(K+1)) \sim P,$$

where P represents the distribution of the stochastic process O, referred to as the observed data, and the general notation $\bar{\cdot}(t)$ represents the history of the variable $\cdot$ between time 0 and t: a) $\bar{\cdot}(t) = (\cdot(0), \ldots, \cdot(t))$ if $t \geq 0$ and b) $\bar{\cdot}(t) = \varnothing$ (empty) if $t < 0$.

We define W as the smallest subset of baseline covariates, $W \subset L(0)$, such that the support of $g(\bar{a}(K) \mid L(0))$ only depends on W and we denote this support with $\mathcal{A}_W(K)$.

We define V as a subset of the baseline covariates containing W, $V \subset L(0)$ and $W \subset V$. We denote the time-dependent outcome with $Y(t)$. We have $Y(t) \in L(t)$ for $t \in \mathcal{T}$, where $\mathcal{T}$ denotes the set of time points t such that the outcome, $Y(t+1)$, is of interest. We have $\mathcal{T} \subset \{0, \ldots, K\}$. Typically $\mathcal{T} = \{0, \ldots, K\}$ except when one is interested in the outcome collected at the end of the study only, i.e. when $\mathcal{T} = \{K\}$. This later case has been treated in previous work from which this manuscript is inspired [2]. However, the focus is now on cases where $\text{Card}(\mathcal{T}) > 1$, where $\text{Card}(\cdot)$ denotes the cardinal of a set $\cdot$. In addition, we adopt the following conventional notation, $\cdot_i$, to represent any random variable $\cdot$ associated with a given experimental unit i.

The question of interest is to investigate the causal effect of the treatment on the time-dependent outcomes of interest. In the literature, this problem has been addressed with MSMs. In MSM-based Causal Inference, the investigation of the causal relationship of interest relies on a representation of the effects of the treatment history $\bar{A}(t)$ on the time-dependent outcome, $Y(t+1) \in L(t+1)$, for all $t \in \mathcal{T}$.

1.2 Assumptions

Existence of counterfactuals: we assume the existence of the following treatment-specific processes, also referred to as counterfactual processes, $\bar{L}_{\bar{a}(K)}(K+1)$ for every treatment regimen $\bar{a}(K) = (a(0), \ldots, a(K)) \in \mathcal{A}_V(K)$ where $\mathcal{A}_V(K)$ designates all possible treatment regimens between time points 0 and K for an experimental unit characterized by the baseline covariate V, i.e. the support of the conditional distribution of $\bar{A}(K)$ given V, $g(\bar{A}(K) \mid V)$. Note that we have $\mathcal{A}_V(K) = \mathcal{A}_W(K)$. See Rubin (1976) [4] for details on the concept of counterfactuals. We denote the so-called full data process with $X = \Big(V, (\bar{L}_{\bar{a}(K)}(K+1))_{\bar{a}(K)\in\mathcal{A}_V(K)}\Big)$ and its distribution with F_X. Note that $W \subset V$ implies that X is indeed well defined.

Note that the existence of the counterfactual process $\bar{L}_{\bar{a}(K)}(K+1)$ for every treatment regimen $\bar{a}(K) \in \mathcal{A}_V(K)$ implies the existence of the counterfactual processes $\bar{L}_{\bar{a}(t)}(t+1) \equiv \bar{L}_{\bar{a}(t),A(t+1),\ldots,A(K)}(t+1) \subset X$ for every $t = 0, \ldots, K-1$ and every treatment regimen $\bar{a}(t) = (a(0), \ldots, a(t)) \in \mathcal{A}_V(t)$ where $\mathcal{A}_V(t)$ designates all possible treatment regimens between time points 0 and t, i.e. the support of the conditional distribution of $\bar{A}(t)$ given V, $g(\bar{A}(t) \mid V)$. We have $\mathcal{A}(t) = \{\bar{a}(t) : \exists\ \bar{a}'(K) \in \mathcal{A}_V(K) \quad \bar{a}(t) = \bar{a}'(t)\}$ for $t = 0, \ldots, K-1$ and $\mathcal{A}(t)$ is thus entirely defined by $\mathcal{A}(K)$.

Consistency assumption: at any time point t, we assume the following link between the observed data and the counterfactuals: $L(t) = L_{\bar{A}(K)}(t)$. Under this assumption, we have: $O = (\bar{A}(K), \bar{L}_{\bar{A}(K)}(K+1)) \equiv \phi(\bar{A}(K), X)$, where ϕ is a specified function of the full data process X. This notation indicates that the problem can be treated as a missing data problem. Only the counterfactual associated with the observed treatment $\bar{A}(K)$ is observed; the others are missing.

Temporal Ordering assumption: at any time point t, we assume that any treatment specific variable can only be affected by past treatments: $L_{\bar{a}(K)}(t) = L_{\bar{a}(t-1)}(t)$ for $t = 0, \ldots, K+1$, where $L_{\bar{a}(-1)}(0) = L(0)$. This assumption is typically implied by the data collection procedure: the covariate $L(t)$ is measured after $A(t-1)$ and before $A(t)$.

Sequential Randomization Assumption (SRA): at any time point t, we assume that the observed treatment is independent of the full data given the data observed up to time point t: $A(t) \perp X \mid \bar{A}(t-1), \bar{L}(t)$. Under the SRA, the treatment mechanism, i.e. the conditional density or probability of $\bar{A}(K)$ given X: $g(\bar{A}(K) \mid X)$, becomes:

$$g(\bar{A}(K) \mid X) = \prod_{t=0}^{K} g(A(t) \mid \bar{A}(t-1), X) \overset{SRA}{=} \prod_{t=0}^{K} g(A(t) \mid \bar{A}(t-1), \bar{L}(t)).$$

The SRA implies coarsening at random [1] and thus the likelihood of the observed data factorizes into two parts: a so-called F_X and g part. The F_X part of the likelihood only depends on the full data process distribution, and the g part of the likelihood only depends on the treatment mechanism. As a consequence of this factorization of the likelihood under the SRA, we now denote the distribution of the observed data with $P_{F_X,g}$ and the likelihood of O is:

$$\mathcal{L}(O) \overset{SRA}{=} \overbrace{f(L(0)) \underbrace{\prod_{t=1}^{K+1} f(L(t) \mid \bar{L}(t-1), \bar{A}(t-1))}_{Q_{F_X}}}^{F_X \text{ part}} \overbrace{g(\bar{A}(K) \mid X)}^{g \text{ part}}.$$

In addition, we denote the set of conditional densities or probabilities that define the F_X part of the likelihood, except for $f(L(0))$ with Q_{F_X}.

1.3 Marginal structural model and parameter of interest

An MSM is a full data model which defines a parameter of interest based on a feature of the marginal distribution of the counterfactual outcomes of interest, $Y_{\bar{a}(t)}(t+1)$ conditionally on V. This parameter of interest represents the causal effect of interest and can thus be interpreted causally. Typically and specifically in this paper, one is interested in average causal effects which can be represented by causal parameters defined by MSMs of $E_{F_X}(Y_{\bar{a}(t)}(t+1) \mid V)$ for $t \in \mathcal{T}$. We denote a causal parameter defined by an MSM with $\beta_t(F_X \mid \cdot)$ to indicate that it is a mapping from the space of full data distribution F_X to the space of real numbers and that this mapping is a function of modeling assumptions represented for now by $\cdot$. Under the assumptions presented in the previous section, the causal parameter, $\beta_t(F_X \mid \cdot)$, can be consistently estimated with the observed data based on three estimators: the Inverse Probability of Treatment Weighted (IPTW), the G-computation (G-comp) and Double Robust (DR) estimators.

Two approaches to Causal Inference based on MSM have been proposed. They provide different representations of causal effects with distinct causal parameters. Initially, a parametric MSM approach to Causal Inference was developed: it relies on correct specification of a parametric MSM. Recently, a new approach based on nonparametric MSM was introduced. This later approach does not require the assumption of a correctly specified MSM and thus is more realistic if one believes that correct specification of a parametric MSM is unlikely in practice.

Unlike the parametric MSM approach, the nonparametric MSM approach was only developed for longitudinal studies where the outcome of interest is collected at the end of the study, i.e. for investigation of the causal effect defined by an MSM of $E(Y_{\bar{a}(K)}(K+1) \mid V)$. We first address the generalization of the nonparametric MSM approach to longitudinal studies with time-dependent outcomes. We then define the G-computation estimators of the corresponding nonparametric MSM parameters and compare them to their analogues in the parametric MSM approach. Finally, we propose new algorithms to address an important computing limitation independent of the MSM approach chosen but inherent to the implementation of the G-computation estimator a) with continuous treatment and/or b) in longitudinal studies with long follow-up and time-dependent outcomes. These new algorithms for the implementation of the G-computation estimator lead us to further generalize the definition of nonparametric causal effects before illustrating the results and concepts introduced in this manuscript with two simulation studies.

Note that all results and procedures are developed for the case where the treatment is discrete and $\mathcal{A}_V(K)$ is finite (sections 2 through 4). In section 5, we extend these results and procedures to problems with bounded continuous treatments.

2 Generalization of the nonparametric MSM approach to time-dependent outcomes

The nonparametric MSM approach relies on a working model, also referred to as the causal model (CM), which is the analog of the parametric MSM in the parametric MSM approach. Thus, we will adopt a common notation to represent a CM and a parametric MSM of $E(Y_{\bar{a}(t)}(t+1) \mid V)$ and use the general word "model" to designate either the CM or the parametric MSM interchangeably or when the context leaves no ambiguity about which object is considered.

Independently of the two alternative MSM approaches available, the investigation of causal effects on time-dependent outcomes can be based on two different analyses, each corresponding with a different model: a stratified or a pooled analysis. We generalize the nonparametric MSM approach to time-dependent outcomes for both of these analyses.

2.1 Stratified analysis

In this analysis, **causal effects are modelled separately for each time point** $t \in \mathcal{T}$, i.e., one **separately** investigates the causal effects on the outcomes of interest, $Y(t+1)$ for $t \in \mathcal{T}$, through the estimation of **distinct causal parameters** $\beta_t(F_X \mid \cdot)$ for $t \in \mathcal{T}$ defined based on $l = \mathrm{Card}(\mathcal{T})$ distinct models m_t for $t \in \mathcal{T}$.

Under this model, the investigation of the causal effects of interest based on the parametric MSM approach to causal inference corresponds to the estimation of the l distinct causal parameters $\beta_t \equiv \beta_t(F_X \mid m_t)$ defined such that:

$$E_{F_X}(Y_{\bar{a}(t)}(t+1) \mid V) = m_t(\bar{a}(t), V \mid \beta_t) \text{ for } t \in \mathcal{T}.$$

The generalization of the nonparametric MSM approach to time-dependent outcomes in such a stratified analysis is derived directly from the application of the nonparametric MSM methodology proposed in [2] to each outcome, $Y(t+1)$ for $t \in \mathcal{T}$, **separately** since each $Y(t+1)$ can be viewed as the outcome at the end of a longitudinal study where the corresponding observed data are: $O(t) = (\bar{A}(t), \bar{L}(t+1))$. The investigation of the causal effects on the time-dependent outcomes can thus be conducted as proposed in [2] through the estimation of the following l distinct causal parameters

$$\begin{aligned} \beta_t(F_X \mid m_t, \lambda_t) \equiv \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}\, E_{F_X}\Big[\sum_{\bar{a}(t) \in \mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m_t(\bar{a}(t), V \mid \beta))^2 \times \\ \lambda_t(\bar{a}(t), V)\Big], \end{aligned} \tag{1}$$

for $t \in \mathcal{T}$, where λ_t are user-specified weighting functions referred to as causal kernel smoothers (CKSs). The definition and interpretation of $\beta_t(F_X \mid m_t, \lambda_t)$ are detailed in [2]. In short, these parameters are analogous to smoothing parameters in conventional analyses. They represent smoothed versions of the causal effects of $\bar{A}(t)$ on $Y(t+1)$ for

$t \in \mathcal{T}$, i.e. approximations of the true causal effects, where each smoothing is controlled by the choice of 1) a CM, m_t, and 2) a CKS, λ_t, such that the CM approximates the true causal effect best over ranges of the data where $\lambda_t(\bar{a}(t), V)$ are larger.

Note that for each $t \in \mathcal{T}$, the parameters of interest, β_t, defined in both the parametric and nonparametric MSM approach are identical if m_t is correctly specified (i.e. if there exists a value for β such that: $E_{F_X}(Y_{\bar{a}(t)}(t+1) \mid V) = m_t(\bar{a}(t), V \mid \beta)$. If m_t is misspecified, the parameter of interest is not defined in the parametric MSM approach. Thus, the nonparametric MSM approach is more general that the parametric MSM approach.

In practice and regardless of the MSM approach chosen, it may be reasonable to represent all causal effects of $\bar{A}(t)$ on $Y(t+1)$ for $t \in \mathcal{T}$ with a unique choice of model, m. The choice of a single model common to each outcome of interest motivates the following approach to more efficiency estimate a single parameter (representing all causal effects) by pooling data across time.

2.2 Pooled analysis

In this analysis, **causal effects are modelled simultaneously for each time point** $t \in \mathcal{T}$, i.e. the change of the causal effect on the outcome over time is represented by a smooth function of time: one **simultaneously** investigates the causal effects on $Y(t)$ for all $t \in \mathcal{T}$ through the estimation of a **single causal parameter** $\beta(F_X \mid \cdot)$ defined based on a single model $m(t, \bar{a}(t), V \mid \beta)$.

Under this model, the investigation of the causal effects of interest based on the parametric MSM approach to causal inference corresponds to the estimation of the single parameter $\beta \equiv \beta(F_X \mid m)$ defined such that:

$$E_{F_X}(Y_{\bar{a}(t)}(t+1) \mid V) = m(t, \bar{a}(t), V \mid \beta) \text{ for } \textbf{all } t \in \mathcal{T}.$$

The generalization of the nonparametric MSM approach to time-dependent outcomes in such a pooled analysis is derived from the application of the nonparametric MSM methodology proposed in [2] to each outcome, $Y(t+1)$ for $t \in \mathcal{T}$, **simultaneously**. The nonparametric MSM parameter can be defined such that it simultaneously minimizes $E_{F_X}\left[\sum_{\bar{a}(t) \in \mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2\right]$ for all $t \in \mathcal{T}$, where the importance of the minimization of this expectation for each t can be controlled by a use-specified weighting function γ. In other words, the parameter of interest can be defined as:

$$\beta(F_X \mid m, (\lambda_t)_{t \in \mathcal{T}}, \gamma) = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} E_{F_X}\Big[\sum_{t \in \mathcal{T}} \Big(\sum_{\bar{a}(t) \in \mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \times \lambda_t(\bar{a}(t), V)\Big)\gamma(t)\Big],$$

where 1) every function $\lambda_t(\bar{a}(t), V)$ can be interpreted as a weighting function which separately controls the importance of the goodness of fit of the CM to the causal curve $E(Y_{\bar{a}(t)}(t+1) \mid V)$ across the different regions $(\bar{a}(t), V)$ and 2) γ can be interpreted as a weighting function which controls the importance of the overall goodness of fit of the

CM to each causal curve of interest, $E(Y_{\bar{a}(t)}(t+1) \mid V)$ for $t \in \mathcal{T}$ across the different time points t. We now define $\xi(t, \bar{a}(t), V) \equiv \lambda_t(\bar{a}(t), V)\gamma(t)$ and refer to it as the CKS in this pooled analysis, it generalizes the definition of the CKSs, λ_t, since it controls not only the smoothing of the causal effects at each time point but also the smoothing of the causal effects over time. The investigation of the causal effects on the time-dependent outcomes can thus be conducted through the estimation of the following single causal parameter

$$\begin{aligned}\beta(F_X \mid m, \xi) \quad \equiv \quad & \underset{\beta \in I\!R^k}{\operatorname{argmin}}\, E_{F_X}\Big[\sum_{t \in \mathcal{T}} \sum_{\bar{a}(t) \in \mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \times \\ & \qquad \xi(t, \bar{a}(t), V)\Big]. \end{aligned} \tag{2}$$

In practice, a natural choice for the CKS, ξ, is $\xi(t, \bar{a}(t), V) = \dfrac{1}{\mathrm{Card}(\mathcal{A}_V(t))}$ since it corresponds to allocating the same importance to the goodness of fit of the CM, m, to all causal curves $E(Y_{\bar{a}(t)}(t+1) \mid V)$. Indeed, it seems natural that one would use the stratified approach described earlier if one is more specifically interested in some causal curves $E(Y_{\bar{a}(t)}(t+1) \mid V)$ among all causal curves $E(Y_{\bar{a}(t)}(t+1) \mid V)$ for $t \in \mathcal{T}$.

Note that 1) the parameters of interest, β, defined in both the parametric and nonparametric MSM approach in a pooled analysis are identical if m is correctly specified, and 2) the parametric MSM parameter is not defined when the CM, m, is misspecified.

Finally, it is essential to notice that definition (2) of the nonparameteric MSM parameter is typically different from the following definition:

$$\begin{aligned}\beta(F_X \mid m, \xi) \quad \equiv \quad & \underset{\beta \in I\!R^k}{\operatorname{argmin}}\, E_{F_X}\Big[\sum_{t \in \mathcal{T}} \sum_{\bar{a}(K) \in \mathcal{A}_V(K)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \times \\ & \qquad \xi(t, \bar{a}(t), V)\Big] \\ = \quad & \underset{\beta \in I\!R^k}{\operatorname{argmin}}\, E_{F_X}\Big[\sum_{t \in \mathcal{T}} \sum_{\bar{a}(t) \in \mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \times \\ & \qquad \xi(t, \bar{a}(t), V)\mathrm{Card}(\mathcal{A}^+_{V,\bar{a}(t)}(t+1))\Big], \end{aligned} \tag{3}$$

where $\mathcal{A}^+_{V,\bar{a}(t)}(t+1)$ represents the set of all possible treatment histories between time $t+1$ and K after assignment of treatment history $\bar{a}(t)$, i.e. the support of $g(A(t+1), \ldots, A(K) \mid \bar{A}(t) = \bar{a}(t), V)$. Note that: $\mathrm{Card}(\mathcal{A}^+_{V,\bar{a}(t)}(t+1)) = \mathrm{Card}(\{\bar{a}^{'}(K) \in \mathcal{A}_V(K) : \bar{a}^{'}(t) = \bar{a}(t)\})$ for $t = 0, \ldots, K-1$ and $\mathrm{Card}(\mathcal{A}^+_{V,\bar{a}(K)}(K+1)) \equiv 1$ by convention.

We now propose a procedure for implementing the G-computation estimator of the nonparametric MSM causal parameters as previously defined in both the stratified and pooled analysis. We compare it to the G-computation estimation procedure developed for the parametric MSM approach.

3 G-computation estimation

For a stratified analysis, the implementation of the G-computation estimator of nonparametric MSM parameter, β_t as defined by (1), can easily be derived from the G-computation estimation procedure for parametric MSM parameters. The later relies on a least squares regression whereas the former relies on a weighted least squares regression with weights defined by the CKS, λ_t. This straightforward generalization of the implementation of the G-computation estimator for nonparametric causal effects was introduced and developed in [2]. For a pooled analysis, the implementation of the G-computation estimator of nonparametric MSM parameter, β as defined by (2), is introduced in this section.

Before developing this later estimation procedure, we first remind the reader of the G-computation estimation procedure proposed for 1) MSM causal effects (parametric and nonparametric) in a stratified analysis and 2) parametric MSM causal effect in a pooled analysis. The description of the G-computation estimation procedures for parametric MSM causal effects and its generalization to nonparametric causal effect will allow a direct comparison of both implementations and underline the danger of a hasty generalization in a pooled analysis setting.

3.1 G-computation estimation of MSM causal effects in a stratified analysis

We summarize the G-computation estimation procedure for nonparametric causal effect in a stratified analysis (see [2] for details) as follows:

```
Estimate $Q_{F_X}$ with $Q_{F_X,n}$.
Repeat the following for each $t \in \mathcal{T}$ seperately
{
  Repeat the following $N$ times:
   {
    Randomly draw an observation of $L(0)$ based on its empirical
    distribution and store the observation of $V \subset L(0)$.
    for($\bar{a}(t) \in \mathcal{A}_V(t)$)
    {
       Based on the observation of $L(0)$ and $Q_{F_X,n}$, generate an observation of
       $\bar{L}_{\bar{a}(t)}(t+1)$ by Monte Carlo simulation.
       Store $\bar{a}(t)$ and the observation of $Y_{\bar{a}(t)}(t+1)$.
    }
   }
  Using weighted least squares regression and the data previously generated, regress $Y_{\bar{a}(t)}(t+$
  $1)$ on $\bar{a}(t)$ and $V$ with the CM $m_t$ and weights defined by $\lambda_t(\bar{a}(t), V)$.
}
```

Note that this estimation procedure is a direct generalization of the G-computation estimation procedure for parametric MSM causal effects [5, 3, 2] in which the least squares regression is simply replaced by a weighted least squares regression.

3.2 G-computation estimation of parametric MSM causal effects in a pooled analysis

Based on an estimate $Q_{F_X,n}$ of the Q_{F_X} part of the likelihood and Monte Carlo simulations, one simulates N (e.g. 10,000) observations of the counterfactual process $(\bar{L}_{\bar{a}(K)}(K+1))$ for each $\bar{a}(K) \in \mathcal{A}_V(K)$. All counterfactual processes are stored and used in a least squares regression to estimate the parameter of interest, $\beta(F_X \mid m)$, defined by the assumed MSM. In details, the following algorithm was proposed [5] :

```
Estimate $Q_{F_X}$ with $Q_{F_X,n}$.
Repeat the following $N$ times:
{
    Randomly draw an observation of $L(0)$ based on its empirical
    distribution and store the observation of $V \subset L(0)$.
    for($\bar{a}(K) \in \mathcal{A}_V(K)$)
    {
        Based on the observation of $L(0)$ and $Q_{F_X,n}$, generate an observation of
        $\bar{L}_{\bar{a}(K)}(K+1)$ by Monte Carlo simulation.
        For $t \in \mathcal{T}$, store $t$, $\bar{a}(t)$ and the observation of $Y_{\bar{a}(t)}(t+1) = Y_{\bar{a}(K)}(t+1)$.
    }
}
Using least squares regression and the data previously generated, regress $Y_{\bar{a}(t)}(t+1)$ on
$\bar{a}(t)$, $V$ and $t$ with the assumed MSM $m$.
```

Note that this procedure corresponds with minimizing for β an approximation of:

$$E_{F_X}\Big[\sum_{t\in\mathcal{T}} \sum_{\bar{a}(K)\in\mathcal{A}_V(K)} (Y_{\bar{a}(t)}(t) - m(\bar{a}(t), V, t \mid \beta))^2\Big].$$

The consistency of this estimation procedure relies on consistent estimation of the Q_{F_X} part of the likelihood.

Note that the number of observations in this regression is random:

$$N_g = \mathrm{Card}(\mathcal{T}) \sum_{i=1}^{N} \mathrm{Card}(\mathcal{A}_{V_i}(K)). \qquad (4)$$

For most longitudinal studies with long follow-up (i.e. K large), the number of observations (4) will be too large to be successfully handled with the computing resources commonly available today to most investigators . We propose new procedures in section 4 to overcome a similar computing limitation of the nonparametric MSM approach to causal inference. These procedures can easily be adapted for the parametric MSM approach to causal inference.

3.3 G-computation estimation of nonparametric MSM causal effects in a pooled analysis

Note first that a naive generalization of the previously described procedure where the least squares regression is simply replaced by a weighted least squares regression with weights defined as $\xi(t, \bar{a}(t), V)$ cannot provide consistent estimation of $\beta(F_X \mid m, \lambda_t, \gamma)$ as defined by equation (2). Instead, such a naive generalization would typically provide consistent estimation of a **different** parameter of interest that is defined by equation (3).

A correct generalization would be to utilize the procedure previously described where the least squares regression is replaced by a weighted least squares regression with weights defined as

$$\frac{\xi(t, \bar{a}(t), V)}{\text{Card}(\mathcal{A}^+_{V,\bar{a}(t)}(t+1))}.$$

Note that such weights will typically be constant and equal to $\frac{1}{\text{Card}(\mathcal{A}_V(K))}$ if $\xi(t, \bar{a}(t), V) = \frac{1}{\text{Card}(\mathcal{A}_V(t))}$. Typically, the weighted least squares regression will thus be equivalent to the original unweighted regression. In addition, note that this protocol suffers from the same computing limitation described above. The number of observations in the weighted regression is also:

$$N_g = \text{Card}(\mathcal{T}) \sum_{i=1}^{N} \text{Card}(\mathcal{A}_{V_i}(K)),$$

i.e. this procedure often requires an amount of computer memory that is not available to most investigators.

4 G-computation estimation procedures to overcome computing limitations

To decrease the computer memory requirements, a very similar estimation procedure could be used in which not all elements of the counterfactual processes, $(\bar{L}_{\bar{a}(K)}(K+1))$, are stored for all $\bar{a}(K) \in \mathcal{A}_V(K)$. Instead, for each of the N observations obtained by Monte Carlo simulation, $Y_{\bar{a}(t)}(t+1) = Y_{\bar{a}(K)}(t+1)$ will be stored only for unique $\bar{a}(t)$. In details, this algorithm goes as follows:

```
We denote the support of $g(A(t) \mid \bar{A}(t-1) = \bar{a}(t-1), V)$ with $\mathcal{A}_{V,\bar{a}(t-1)}(t)$ for $t = 0, \ldots, K$.
Estimate $Q_{F_X}$ with $Q_{F_X,n}$.
Repeat the following $N$ times:
{
  Randomly draw an observation of $L_{\bar{a}(K)}(0) = L(0)$ based on its empirical
  distribution and store the observation of $V \subset L(0)$.
  for($a(0) \in \mathcal{A}_V(0)$)
  {
    Based on the observation of $L_{\bar{a}(K)}(0)$ and $Q_{F_X,n}$, generate $L_{\bar{a}(K)}(1) = L_{\bar{a}(0)}(1)$
```

```
        by Monte Carlo simulation.
        ⋮
        ... for($a(i) \in \mathcal{A}_{V,\bar{a}(i-1)}(i)$)
            {
                Based on the observations of $\bar{L}_{\bar{a}(K)}(i)$ and $Q_{F_X,n}$, generate
                $L_{\bar{a}(K)}(i+1) = L_{\bar{a}(i)}(i+1)$ by Monte Carlo simulation.
                ⋮
                ... for($a(K) \in \mathcal{A}_{V,\bar{a}(K-1)}(K)$)
                    {
                        Based on the observations of $\bar{L}_{\bar{a}(K)}(K)$ and $Q_{F_X,n}$, generate
                        $L_{\bar{a}(K)}(K+1)$ by Monte Carlo simulation.
                        If $K \in \mathcal{T}$ then store $t = K$, $\bar{a}(K)$ and the observation of
                        $Y_{\bar{a}(K)}(K+1)$.
                    }
                ⋮
                If $i \in \mathcal{T}$ then store $t = i$, $\bar{a}(i)$ and the observation of
                $Y_{\bar{a}(i)}(i+1) = Y_{\bar{a}(K)}(i+1)$.
            }
        ⋮
        If $0 \in \mathcal{T}$ then store $t = 0$, $\bar{a}(0)$ and the observation of $Y_{\bar{a}(0)}(1) = Y_{\bar{a}(K)}(1)$.
    }
}
Using weighted least squares regression and the data previously generated, regress $Y_{\bar{a}(t)}(t+1)$ on $\bar{a}(t)$, $V$ and $t$ with the CM $m$ and weights defined by $\xi(t, \bar{a}(t), V)$.
```

Note that this procedure corresponds with minimizing for β an approximation of:

$$E_{F_X} \sum_{t \in \mathcal{T}} \sum_{\bar{a}(t) \in \mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \xi(t, \bar{a}(t), V).$$

The consistency of this estimation procedure relies on consistent estimation of the Q_{F_X} part of the likelihood.

Note that the number of observations in this weighted regression is also random:

$$\begin{aligned} N_g &= \sum_{i=1}^{N} \sum_{t \in \mathcal{T}} \mathrm{Card}(\mathcal{A}_{V_i}(t)) \qquad (5) \\ &< \mathrm{Card}(\mathcal{T}) \sum_{i=1}^{N} \mathrm{Card}(\mathcal{A}_{V_i}(K)), \end{aligned}$$

i.e. the number of observations needed in this procedure is typically much smaller than the number of observations required in the procedure previously described.

Computer memory limitations may nevertheless still not allow implementation of the latter procedure described above for the more complex real life applications. To allow implementation of the G-computation estimator for even the more complex applications, we propose below two new procedures to significantly decrease the amount of computer memory required in practice. More importantly, both procedures allow investigators to customize each of the proposed algorithms such that the corresponding memory requirements fit the computing resources available to them.

Proposed procedure 1. For each of the N experimental units, instead of simulating by Monte Carlo simulations all its counterfactual processes, i.e. $\bar{L}_{\bar{a}(K)}(K+1)$ for all $\bar{a}(K) \in \mathcal{A}_V(K)$, we propose to only simulate a random subset of all its counterfactual processes, i.e. $L_{\bar{a}(t)}(t+1) = L_{\bar{a}(K)}(t+1)$ for p elements $(t, \bar{a}(K)) \in \mathcal{T} \times \mathcal{A}_V(K)$.

One first simulates N (user-specified) independent observations of $L(0)$ before simulating by Monte Carlo simulation a subset of p elements of all its counterfactual processes, i.e. $(\bar{L}_{\bar{a}(t)}(t+1) = \bar{L}_{\bar{a}(K)}(t+1))$ for p observations, $(t, \bar{a}(K))$, of $(T, \bar{A}(K))$ where 1) T and $\bar{A}(K)$ are uniformly distributed over $\mathcal{T}$ and $\mathcal{A}_V(K)$ respectively, and 2) T is independent of $\bar{A}(K)$. For each experimental unit represented by a draw of $L(0)$, only p corresponding counterfactuals $Y_{\bar{a}(t)}(t+1)$ are stored and used in the weighted regression of $Y_{\bar{a}(t)}(t+1)$ on the treatment histories $\bar{a}(t)$, baseline covariate V, and time variable t with the CM m and the weights defined by $\frac{\xi(t,\bar{a}(t),V)}{\text{Card}(\mathcal{A}^+_{V,\bar{a}(t)}(t+1))}$. The estimator implemented in this procedure is asymptotically linear.

Note that the number of observations in this regression is user-specified (i.e, not random) $N_g = pN$ and that the G-computation estimate obtained only relies on subset of the full data, i.e. a limited number, p, of counterfactuals among the set of all possible counterfactuals for a given unit characterized by V. For a given V, the number of possible counterfactuals is $\sum_{t\in\mathcal{T}} \text{Card}(\mathcal{A}_V(t))$ and note that $p << \sum_{t\in\mathcal{T}} \text{Card}(\mathcal{A}_V(t))$. We can however use the asymptotic linearity of the estimator to obtain a more efficient estimator that relies on a larger set of counterfactuals. This is done by repeating the estimation procedure described above R times. The average of the corresponding R estimates obtained correspond with the G-computation estimate. In details, this algorithm goes as follows:

```
Estimate $Q_{F_X}$ with $Q_{F_X,n}$.
Repeat the following $R$ times:
{
    Repeat the following $N$ times:
    {
        Randomly draw an observation of $L(0)$ based on its empirical distribution and
        store the observation of $V \subset L(0)$.
        Repeat the following $p$ times:
        {
            Draw $(t, \bar{a}(K))$ from independent uniform distributions of $T$ and $\bar{A}(K)$
            over $\mathcal{T}$ and $\mathcal{A}_V(K)$ conditionally on $V$ respectively.
```

```
      Simulate by Monte Carlo simulation the process $(\bar{L}_{\bar{a}(t)}(t+1) = \bar{L}_{\bar{a}(K)}(t+1))$.
      Store $\bar{a}(t)$, $t$ and $Y_{\bar{a}(t)}(t+1) = Y_{\bar{a}(K)}(t+1)$.
    }
  }
  Using weighted least squares regression and the data previously generated,
  regress $Y_{\bar{a}(t)}(t+1)$ on $\bar{a}(t)$, $V$ and $t$ based on the CM $m$ and with weights
  $\frac{\xi(t,\bar{a}(t),V)}{Card(\mathcal{A}^{+}_{V,\bar{a}(t)}(t+1))}$ and store the estimate obtained.
}
The average of the $R$ estimates obtained is the G-computation estimate.
```

Note that the regressions in this procedure corresponds with minimizing for β an approximation of $E_{F_X,U}(D(T,\bar{A}(K),\bar{L}_{\bar{A}(K)}(T+1) \mid \beta))$, where:

$$D(T,\bar{A}(K),\bar{L}_{\bar{A}(K)}(T+1) \mid \beta) = (Y_{\bar{A}(T)}(T+1) - m(\bar{A}(T),V,T \mid \beta))^2 \frac{\xi(T,\bar{A}(T),V)}{\text{Card}(\mathcal{A}^{+}_{V,\bar{a}(T)}(T+1))},$$

and U is the joint distribution of $(T,\bar{A}(K))$ defined by the two independent marginal uniform distributions of T and $\bar{A}(K)$ conditional on V over $\mathcal{T}$ and $\mathcal{A}_V(K)$ respectively. We have:

$$P(T=t \mid V) = P(T=t) = \frac{I(t \in \mathcal{T})}{\text{Card}(\mathcal{T})} \text{ and } P(\bar{A}(K) = \bar{a}(K) \mid V) = \frac{I(\bar{a}(K) \in \mathcal{A}_V(K))}{\text{Card}(\mathcal{A}_V(K))} \tag{6}$$

The causal parameter minimizing this quantity is indeed $\beta(F_X \mid m,\xi,\gamma)$ as defined by (2) since we have:

$$\begin{aligned}
&E_{F_X,U}\left[D(T,\bar{A}(K),\bar{L}_{\bar{A}(K)}(T+1) \mid \beta)\right] \\
&\quad = E_{F_X} E_U \left[D(T,\bar{A}(K),\bar{L}_{\bar{A}(K)}(T+1)) \mid X\right] \\
&\quad = E_{F_X} \sum_{t\in\mathcal{T}} \sum_{\bar{a}(K)\in\mathcal{A}_V(K)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t),V,t \mid \beta))^2 \times \\
&\qquad\qquad \frac{\xi(t,\bar{a}(t),V)}{\text{Card}(\mathcal{A}^{+}_{V,\bar{a}(t)}(t+1))} \frac{1}{\text{Card}(\mathcal{T})} \frac{1}{\text{Card}(\mathcal{A}_V(K))} \\
&\quad = \frac{1}{\text{Card}(\mathcal{T})\text{Card}(\mathcal{A}_V(K))} E_{F_X} \sum_{t\in\mathcal{T}} \sum_{\bar{a}(t)\in\mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t),V,t \mid \beta))^2 \times \\
&\qquad\qquad \xi(t,\bar{a}(t),V),
\end{aligned}$$

Proposed procedure 2. The previous procedure overcomes limitations due to restricted computer memory however it does not optimize the computation time as for each draw of $(t,\bar{a}(K))$, only the counterfactual outcome, $Y_{\bar{a}(t)}(t+1)$, is used to derive the G-computation estimate in this procedure, i.e. all intermediate outcomes remain unused even though they are required to be simulated in the Monte Carlo simulation and could be used towards fitting the CM. That is why we propose the following procedure that should minimize not only the computer memory requirement but also improve the computation time by making use of every intermediate outcomes obtained from the Monte Carlo simulation:

```
Estimate $Q_{F_X}$ with $Q_{F_X,n}$.
Repeat the following $R$ times:
{
    Repeat the following $N$ times:
    {
        Randomly draw an observation of $L(0)$ based on its empirical distribution and
        store the observation of $V \subset L(0)$.
        Repeat the following $q$ times:
        {
            Draw $(t, \bar{a}(K))$ from independent uniform distributions of $T$ and $\bar{A}(K)$
            over $\mathcal{T}$ and $\mathcal{A}_V(K)$ conditionally on $V$ respectively.
            Simulate by Monte Carlo simulation the process $(\bar{L}_{\bar{a}(t)}(t+1) = \bar{L}_{\bar{a}(K)}(t+1))$.
            Store $\bar{a}(j)$, $j$ and $Y_{\bar{a}(j)}(j+1) = Y_{\bar{a}(K)}(j+1)$ for $j \in \mathcal{T}$ and $j \leq t$.
        }
    }
    Using weighted least squares regression and the data previously generated,
    regress $Y_{\bar{a}(j)}(j+1)$ on $\bar{a}(j)$, $V$ and $j$ based on the CM $m$ and with weights
    $\frac{\xi(j,\bar{a}(j),V)}{Card(\mathcal{A}^{+}_{V,\bar{a}(j)}(j+1))Card(\{j' \in \mathcal{T} : j' \geq j\})}$ and store the estimate obtained.
}
The average of the $R$ estimates obtained is the G-computation estimate.
```

Note that the weighted regressions in this procedure correspond with minimizing for β an approximation of $E_{F_X,U}(D(T, \bar{A}(K), \bar{L}_{\bar{A}(K)}(K+1) \mid \beta))$, where:

$$\begin{aligned} D(T, \bar{A}(K), \bar{L}_{\bar{A}(K)}(K+1) \mid \beta) &= \sum_{j \in \mathcal{T} : j \leq T} (Y_{\bar{A}(j)}(j+1) - m(\bar{A}(j), V, j \mid \beta))^2 \times \\ & \qquad \frac{\xi(j, \bar{A}(j), V)}{\text{Card}(\mathcal{A}^{+}_{V,\bar{a}(j)}(j+1))\text{Card}(\{j' \in \mathcal{T} : j' \geq j\})}, \end{aligned}$$

and U is the joint distribution of $(T, \bar{A}(K))$ defined by the two independent marginal uniform distributions (6) of T and $\bar{A}(K)$ conditional on V over $\mathcal{T}$ and $\mathcal{A}_V(K)$ respectively. The causal parameter minimizing this quantity is indeed $\beta(F_X \mid m, \xi, \gamma)$ as defined by (2) since we have:

$$\begin{aligned} & E_{F_X,U}\left[D(T, \bar{A}(K), \bar{L}_{\bar{A}(K)}(T+1) \mid \beta)\right] \\ &= E_{F_X} E_U \left[D(T, \bar{A}(K), \bar{L}_{\bar{A}(K)}(T+1)) \mid X\right] \\ &= E_{F_X} \sum_{\bar{a}(K) \in \mathcal{A}_V(K)} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{T} : j \leq t} (Y_{\bar{a}(j)}(j+1) - m(\bar{a}(j), V, j \mid \beta))^2 \times \end{aligned}$$

$$
\begin{aligned}
& \frac{\xi(j,\bar{a}(j),V)}{\mathrm{Card}(\mathcal{A}^{+}_{V,\bar{a}(j)}(j+1))\mathrm{Card}(\{j' \in \mathcal{T} : j' \geq j\})} \frac{1}{\mathrm{Card}(\mathcal{A}_V(K))} \frac{1}{\mathrm{Card}(\mathcal{T})} \\
= & \frac{1}{\mathrm{Card}(\mathcal{A}_V(K))\mathrm{Card}(\mathcal{T})} E_{F_X} \sum_{\bar{a}(K)\in\mathcal{A}_V(K)} \sum_{t\in\mathcal{T}} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t),V,t \mid \beta))^2 \times \\
& \frac{\xi(t,\bar{a}(t),V)}{\mathrm{Card}(\mathcal{A}^{+}_{V,\bar{a}(t)}(t+1))} \\
= & \frac{1}{\mathrm{Card}(\mathcal{A}_V(K))\mathrm{Card}(\mathcal{T})} E_{F_X} \sum_{t\in\mathcal{T}} \sum_{\bar{a}(t)\in\mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t),V,t \mid \beta))^2 \times \\
& \xi(t,\bar{a}(t),V),
\end{aligned}
$$

Note that the number of observations in each regression is random: $N_g = q\sum_{i=1}^{N}\mathrm{Card}(\{j \in \mathcal{T} : j \leq T_i\}) \leq qN\mathrm{Card}(\mathcal{T})$ and that the asymptotical linearity of the estimation procedure implemented is also used to obtain a more efficient estimator by iterating the estimation procedure R times.

5 Generalization

We now generalize the methodology proposed in this paper to bounded continuous treatments.

In a stratified analysis, the nonparametric MSM approach corresponds to the estimation of the following l causal parameters:

$$
\begin{aligned}
\beta_t(F_X \mid m_t, g_t) \equiv & \operatorname*{argmin}_{\beta\in I\!R^k} E_{F_X}\Big[\int_{\bar{a}(t)\in\mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m_t(\bar{a}(t),V \mid \beta))^2 \times \\
& g_t(\bar{a}(t) \mid V) d\mu(\bar{a}(t))\Big],
\end{aligned} \tag{7}
$$

where 1) g_t represents a conditional distribution of the treatment history $\bar{A}(t)$ conditional on the baseline covariate V and 2) $d\mu(\bar{a}(t))$ represents the denominating measure of the cumulative distribution function g_t which is typically the Lebesque measure when A is continuous, i.e. $d\mu(\bar{a}(t)) = da(0)\ldots, da(t)$. The function g_t is the analogue to λ_t in definition (1) of the causal parameter of interest in the discrete case. Definition (7) is thus the analogue to definition (1) for bounded continuous treatments.

In a pooled analysis, the nonparametric MSM approach corresponds to the estimation of the following causal parameters:

$$
\begin{aligned}
\beta(F_X \mid m, \xi) \equiv & \operatorname*{argmin}_{\beta\in I\!R^k} E_{F_X}\Big[\sum_{t\in\mathcal{T}} \int_{\bar{a}(t)\in\mathcal{A}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t),V,t \mid \beta))^2 \times \\
& \xi(t,\bar{a}(t),V) d\mu(\bar{a}(t))\Big],
\end{aligned} \tag{8}
$$

where 1) $\xi(t,\bar{a}(t),V) \equiv g_t(\bar{a}(t) \mid V)\gamma(t)$ is the analogue to the CKS introduced in section 2.2 and 2) γ is the analogue to the weighting function. Definition (8) is thus the analogue to definition (2) for bounded continuous treatments. In practice and similar to the discrete case, note that a natural choice for the CKS, ξ, is $\xi(t, \bar{a}(t), V) = \frac{1}{\int_{\bar{a}(t)\in\mathcal{A}_V(t)} d\mu(\bar{a}(t))}$ since it corresponds with allocating the same importance to the goodness of fit of the CM, m, to all causal curves $E(Y_{\bar{a}(t)}(t+1) \mid V)$.

The G-computation estimation procedures described in the discrete case can be generalized to the continuous case. However such procedures require the simulation of counterfactuals for all treatment histories in $\mathcal{A}_V(t)$, for $t \in \mathcal{T}$. Note that the sets of possible treatment histories are infinite in the continuous case and thus the G-computation estimation procedures described in the discrete case are not directly practicable. To overcome this practical issue, we propose to approximate the set of possible treatment histories by discretization. Based on treatment discretization, one can obtain G-computation estimates with any of the procedures described for the discrete case but applied to the discretized treatments. Note that the sets of possible treatments, $\mathcal{A}_V(t)$, being infinite, an acceptable approximation of these sets by discretization will typically involve finite sets, $\mathcal{A}_V^{d,u}(t)$ characterized by large cardinals such that only the proposed G-computation estimation procedure 1 and 2 will typically be practicable in the continuous case. That is why we only provide generalization of these two procedures to the continuous case in this manuscript.

We propose the following methodology for discretizing continuous treatments and using the resulting discretization for G-computation estimation of nonparametric MSM parameters with procedures 1 and 2 described in the previous section. Note that this methodology is easily generalizable to the estimation of parametric MSM parameters by replacing the weighted least squares regressions in the last steps of procedures 1 and 2 with simple least squares regressions.

Based on a consistent estimate, $g_n(\bar{A}(K) \mid V)$, of the conditional distribution of $\bar{A}(K)$ given V, $g(\bar{A}(K) \mid V)$, one can approximate the set of possible treatment histories $\mathcal{A}_V(K)$ with the finite set $\mathcal{A}_V^d(K) \equiv \{\bar{a}_1^d(K), \ldots, \bar{a}_{n_d}^d(K)\}$ consisting of n_d draws, $\bar{a}_i^d(K)$, from the conditional distribution $g_n(\bar{A}(K) \mid V)$. We denote the set of unique elements of $\mathcal{A}_V^d(K)$ with $\mathcal{A}_V^{d,u}(K)$. Note that this discretizing procedure implies that $\lim_{n_d \longrightarrow +\infty} \mathcal{A}_V^{d,u}(K) = \mathcal{A}_V(K)$ if $g_n(\bar{A}(K) \mid V)$ is a consistent estimator of $g(\bar{A}(K) \mid V)$. In practice, n_d should be chosen very large to ensure proper representation of $\mathcal{A}_V(K)$ with $\mathcal{A}_V^{d,u}(K)$, e.g. $n_d = 10,000$. Based on this discretization and lemma A.1, we can generalize the procedures 1 and 2 proposed in the previous section as follows:

Procecure 1

Estimate $g(\bar{A}(K) \mid V)$*with* $g_n(\bar{A}(K) \mid V)$.
Discretize $\mathcal{A}_V(K)$ *with* n_d *draws,* $\bar{a}_i^d(K)$*, from* $g_n(\bar{A}(K) \mid V)$:
$\mathcal{A}_V^d(K) \equiv \{\bar{a}_1^d(K), \ldots, \bar{a}_{n_d}^d(K)\}$.
$\mathcal{A}_V^{d,u}(K)$ *represents the set of unique elements of* $\mathcal{A}_V^d(K)$.
Estimate Q_{F_X} *with* $Q_{F_X,n}$.
Repeat the following R *times:*
{

```
    Repeat the following $N$ times:
    {
        Randomly draw an observation of $L(0)$ based on its empirical distribution and
        store the observation of $V \subset L(0)$.
        Repeat the following $p$ times:
        {
            Independently draw: 1) $t$ from the uniform distribution of $T$ over $\mathcal{T}$, and 2) $\bar{a}_d(K)$
            from the following discrete distribution of $\bar{A}^d(K)$ over $\mathcal{A}_V^{d,u}(K)$ conditionally
            on $V$:

            $P(\bar{A}^d(K) = \bar{a}^d(K) \mid V) = I(\bar{a}^d(K) \in \mathcal{A}_V^{d,u}(K)) \sum_{i=1}^{n_d} \frac{\frac{I(\bar{a}_i^d(K) = \bar{a}^d(K))}{g_n(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g_n(\bar{a}_i^d(K)|V)}}. \qquad (9)$

            Simulate by Monte Carlo simulation the process $(\bar{L}_{\bar{a}^d(t)}(t+1) = \bar{L}_{\bar{a}^d(K)}(t+1))$
            Store $Y_{\bar{a}^d(t)}(t+1) = Y_{\bar{a}^d(K)}(t+1)$.
        }
    }
    Using weighted least squares regression and the data previously generated,
    regress $Y_{\bar{a}^d(t)}(t+1)$ on $\bar{a}^d(t)$, $V$ and $t$ based on the CM $m$ and with weights
    $\frac{\xi(t,\bar{a}^d(t),V)}{Card(\mathcal{A}^{d,u+}_{V,\bar{a}^d(t)}(t+1))}$ , where $\mathcal{A}^{d,u+}_{V,\bar{a}^d(t)}(t+1)$ is defined as: $\{\bar{a}(t+1,K) \equiv (a(t+1),\ldots,$
    $a(K)) : \exists\ \bar{a}^{d'}(K) \in \mathcal{A}_V^{d,u}(K) \ \ \bar{a}^{d'}(t) = \bar{a}^d(t)\ \&\ \bar{a}^{d'}(t+1,K) = \bar{a}(t+1,K)\}$.
    Store the estimate obtained.
}
The average of the $R$ estimates obtained is the G-computation estimate.
```

Procecure 2

```
Estimate $g(\bar{A}(K) \mid V)$with $g_n(\bar{A}(K) \mid V)$.
Discretize $\mathcal{A}_V(K)$ with $n_d$ draws, $\bar{a}_i^d(K)$, from $g_n(\bar{A}(K) \mid V)$:
$\mathcal{A}_V^d(K) \equiv \{\bar{a}_1^d(K),\ldots,\bar{a}_{n_d}^d(K)\}$.
$\mathcal{A}_V^{d,u}(K)$ represents the set of unique elements of $\mathcal{A}_V^d(K)$.
Estimate $Q_{F_X}$ with $Q_{F_X,n}$.
Repeat the following $R$ times:
{
    Repeat the following $N$ times:
    {
        Randomly draw an observation of $L(0)$ based on its empirical distribution and
        store the observation of $V \subset L(0)$.
        Repeat the following $q$ times:
        {
            Independently draw: 1) $t$ from the uniform distribution of $T$ over $\mathcal{T}$, and 2) $\bar{a}_d(K)$
            from the following discrete distribution of $\bar{A}^d(K)$ over $\mathcal{A}_V^{d,u}(K)$ conditionally
            on $V$:

            $P(\bar{A}^d(K) = \bar{a}^d(K) \mid V) = I(\bar{a}^d(K) \in \mathcal{A}_V^{d,u}(K)) \sum_{i=1}^{n_d} \frac{\frac{I(\bar{a}_i^d(K)=\bar{a}^d(K))}{g_n(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g_n(\bar{a}_i^d(K)|V)}}.$

            Simulate by Monte Carlo simulation the process $(\bar{L}_{\bar{a}^d(t)}(t+1) = \bar{L}_{\bar{a}^d(K)}(t+1))$.
            Store $j$ and $Y_{\bar{a}^d(j)}(j+1) = Y_{\bar{a}^d(K)}(j+1)$ for $j \in \mathcal{T}$ and $j \leq t$.
        }
    }
    Using weighted least squares regression and the data previously generated,
    regress $Y_{\bar{a}^d(j)}(j+1)$ on $\bar{a}^d(j)$, $V$ and $j$ based on the CM $m$ and with weights
    $\frac{\xi(j,\bar{a}^d(j),V)}{Card(\mathcal{A}^{d,u+}_{V,\bar{a}^d(j)}(j+1))\, Card(\{j' \in \mathcal{T} : j' \geq j\})}$. Store the estimate obtained.
}
The average of the $R$ estimates obtained is the G-computation estimate.
```

Note that these two procedures rely on correct specification of a model for $g(\bar{A}(K) \mid V)$ to ensure that treatment discretization provides an accurate approximation of $\mathcal{A}_V(K)$. If the model is misspecified, it is possible that $\lim_{n_d \longrightarrow +\infty} \mathcal{A}_V^{d,u}(K) = \mathcal{A}'_V(K)$ where $\mathcal{A}'_V(K) \neq \mathcal{A}_V(K)$. In such a situation the G-computation estimator implemented in both procedures is inconsistent for the estimation of $\beta(F_X \mid m, \xi)$ as defined by (8). Instead the two procedures implement the consistent G-computation estimator of the following parameter:

$$\begin{aligned}\beta'(F_X \mid m, \xi) \quad &\equiv \quad \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}\; E_{F_X}\Big[\sum_{t \in \mathcal{T}} \int_{\bar{a}(t) \in \mathcal{A}'_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \times \\ &\qquad\qquad \xi(t, \bar{a}(t), V) d\mu(\bar{a}(t))\Big].\end{aligned}$$

In practice, $g(\bar{A}(K) \mid V)$ is unknown and thus proper interpretation of estimates obtained from the procedure described above relies on correct specification of an assumed model for $g(\bar{A}(K) \mid V)$. In light of this remark, we propose to generalize the parameter of interest developed in this manuscript to allow the investigator to better control the interpretation of the estimates obtained from the procedures described in this manuscript. In a stratified analysis, the proposed generalized nonparametric MSM parameter of interest is:

$$\begin{aligned}\beta_t(F_X \mid m_t, \lambda_t, (\mathcal{S}_V(K))) \quad &\equiv \quad \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}}\; E_{F_X}\Big[\sum_{\bar{a}(t) \in \mathcal{S}_V(t)} (Y_{\bar{a}(t)}(t+1) - m_t(\bar{a}(t), V \mid \beta))^2 \times \\ &\qquad\qquad \lambda_t(\bar{a}(t) \mid V)\Big], \qquad (10)\end{aligned}$$

where $(\mathcal{S}_V(K))$ is a user-specified set of V-specific sets of treatment histories $\bar{A}(K)$ (possibly infinite) for which the investigator wishes to study the causal effect on the outcome of interest. Note that $\mathcal{S}_V(t)$ for $t = 0, \ldots, K$ are defined from $\mathcal{S}_V(K)$: $\mathcal{S}_V(t) = \{\bar{a}(t) :$

$\exists\ \bar{a}^{'}(K) \in \mathcal{S}_V(K)\ \bar{a}^{'}(t) = \bar{a}(t)\}$. In a pooled analysis, the proposed generalized nonparametric MSM parameter of interest is:

$$\beta(F_X \mid m, \xi, (\mathcal{S}_V(K))) \equiv \underset{\beta \in I\!R^k}{\operatorname{argmin}}\ E_{F_X}\Big[\sum_{t \in \mathcal{T}} \sum_{\bar{a}(t) \in \mathcal{S}_V(t)} (Y_{\bar{a}(t)}(t+1) - m(\bar{a}(t), V, t \mid \beta))^2 \times \xi(t, \bar{a}(t), V)\Big]. \tag{11}$$

We now illustrate the concepts and results introduced in this manuscript with two simulation studies.

6 Illustration with simulations

6.1 Simulation study with longitudinal data

6.2 Aims and protocol

This simulation study illustrates the importance and validity of the two new algorithms proposed in section 4. It underlines the computing limitations inherent to G-computation estimation in longitudinal studies with long follow-up and illustrates how the two proposed procedures overcome these implementation problems. This simulation study also illustrates the validity of the two proposed procedures generalized to bounded continuous treatments as described in section 5.

The observed data in this simulation are n i.i.d. observations of the treatment process $\bar{A}(K) = (\bar{A}_1(K), \bar{A}_2(K))$ and the covariate process $\bar{L}(K)$, where A_1 is a categorical variables, A_2 and L are binary variables, and $L(0) = \varnothing$ (i.e. $V = \varnothing$). The observed data generating distribution is defined by the following g and F_X parts of the likelihood:

- $P(A_1(0) = k \mid X) = P(A_1(0) = k)$ as defined in table i.

k	1	2	3	4	5
$P(A_1(0) = k)$	0.2	0.2	0.4	0.2	0.2

Table i: Marginal distribution of $A_1(0)$.

- $P(A_2(0) = k \mid X, A_1(0)) = P(A_2(0) = k)$ as defined in table ii.

k	0	1
$P(A_2(0) = k)$	0.8	0.2

Table ii: Marginal distribution of $A_2(0)$.

- $P(L(1) = 1 \mid A_1(0), A_2(0)) = \frac{1}{1+\exp\left(-(\alpha_0+\alpha_1 A_1(0)+\alpha_2 A_2(0))\right)}$

- $P(A_1(t) = k \mid X, \bar{A}(t-1)) = P(A_1(t) = k \mid A_1(t-1))$ for $t = 1, \ldots, K$ as defined in table iii.

k	1	2	3	4	5
$P(A_1(t) = k \mid A_1(t-1) = 1)$	0.4	0.2	0.2	0.1	0.1
$P(A_1(t) = k \mid A_1(t-1) = 2)$	0.2	0.4	0.2	0.1	0.1
$P(A_1(t) = k \mid A_1(t-1) = 3)$	0.1	0.2	0.4	0.2	0.1
$P(A_1(t) = k \mid A_1(t-1) = 4)$	0.1	0.1	0.2	0.4	0.2
$P(A_1(t) = k \mid A_1(t-1) = 5)$	0.1	0.1	0.2	0.2	0.4

Table iii: Distribution of $A_1(t)$ conditional on $A_1(t-1)$ for $t = 0, \ldots, K$.

- $P(A_2(t) = k \mid X, \bar{A}(t-1), A_1(t)) = P(A_2(t) = k \mid L(t))$ as defined in table iv.

k	0	1
$P(A_2(t) = k \mid L(t) = 0)$	0.8	0.2
$P(A_2(t) = k \mid L(t) = 1)$	0.2	0.8

Table iv: Distribution of $A_2(t)$ conditional on $L(t)$ for $t = 0, \ldots, K$.

- For $t = 2, \ldots, K+1$:

$$\begin{aligned}
&P(L(t) = 1 \mid \bar{L}(t-1), \bar{A}(t-1)) \\
&= \quad P(L(t) = 1 \mid A_1(t-2), A_1(t-1), A_2(t-2), A_2(t-1), L(t-1)) \\
&= \quad \frac{1}{1+\exp\Big(-(\gamma_0+\gamma_1 MA_1(t-1)+\gamma_2 t MA_1(t-1)+\gamma_3 SA_2(t-1)+\gamma_4 L(t-1))\Big)},
\end{aligned}$$

where $MA_1(t-1) = \frac{A_1(t-2)+A_1(t-1)}{2}$ and $SA_2(t-1) = A(t-2) + A(t-1)$.

Note that this data generating distribution implies:

$$\mathcal{A}(t) = \Big(\{1;2;3;4;5\} \times \{0,1\}\Big)^{t+1}, \text{ for } t = 0, \ldots, K$$

For different values for K, $\alpha = (\alpha_0, \alpha_1, \alpha_2)$ and $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$, we implemented 5 algorithms of the G-computation estimator of $\beta(F_X \mid m, \xi)$ as defined by equality (2) with:

$$m(\bar{a}(t) \mid \beta) = \beta_0 + \beta_1 ma_1(t) + \beta_2 a_2(t-1) \text{ and } \xi(t, \bar{a}(t)) = \frac{1}{\mathrm{Card}(\mathcal{A}(t))} \text{ for } t = 0, \ldots, K,$$

where $ma_1(0) \equiv a_1(0)$. The first algorithm is described in section 3.3. We will refer to it as the conventional algorithm (Conv) since it corresponds to the algorithm that was first proposed for G-computation estimation of parametric MSM parameters of interest. The second and third algorithms correspond to the proposed procedures 1 and 2 described in section 4. We will refer to them as procedure 1 (Proc 1) and procedure 2 (Proc 2) respectively. The last two algorithms correspond to the generalization of procedures 1 and 2 to bounded continuous treatments and are described in section 5. We will refer to them as generalized procedure 1 (Gen proc 1) and generalized procedure 2 (Gen proc 2) respectively. Note that all algorithms depend on a user-specified value for N which was set to $10,000$ for all simulations. All algorithms except for the conventional algorithm depend on a user-specified value for R. Procedure 1 and generalized procedure 1 also depend on a user-specified value for p whereas procedure 2 and generalized procedure 2 also depend on a user-specified value for q. Generalized procedures 1 and 2 depend on an additional user-specified value for n_d.

Note that generalized procedures 1 and 2 were developed for problems involving bounded continuous treatments. Nevertheless, both algorithms can be used with discrete treatments thus providing a means for testing the validity of both procedures.

In addition, note that all estimates were obtained based on correct model specification for the Q_{F_X} part of the likelihood. The generalized procedures 1 and 2 were implemented based on the true distribution $P(\bar{A}(K))$, which was calculated from the known data generating distribution with a recursive algorithm.

7 Results and interpretation

The first set of results were obtained with $n = 1000$, $K = 0$, $\alpha = (-1; 0.1; -3)$, $R = 10$, $p = q = 5$ and $n_d = 10; 10,000$. Results are displayed in table v. The barplots in figure 1 represent the estimated uniform distribution over $\mathcal{A}(K)$ based on n_d draws from $P(\bar{A}(K))$, i.e. the barplots of $\bar{a}^d(K)$ against $P(\bar{A}^d(K) = \bar{a}^d(K))$ as defined by equality (9) where each $\bar{a}^d(K) \in \mathcal{A}(K)$ is represented on the x axis by a given integer between 1 and Card$(\mathcal{A}(K))$.

The second set of results were obtained with $n = 1000$, $K = 1$, $\alpha = (-1; 0.1; -3)$, $\gamma = (-1; 0.1; 0.8; -2; 3)$, $R = 10$, $p = q = 20$ and $n_d = 10; 10,000$. Results and corresponding barplots are displayed in table vi and figure 2 respectively.

The third set of results were obtained with $n = 1000$, $K = 3$, $\alpha = (-1; 0.2; -3)$, $\gamma = (-1; 0.2; 0.2; -3; 3)$, $R = 10$, $p = q = 50$ and $n_d = 10; 10,000; 100,000$. Results and barplots are displayed in table vii and figure 3 respectively.

The first two sets of results illustrate the validity of procedures 1 and 2 and of the generalized procedures 1 and 2 by comparison of the corresponding estimates with the estimate obtained from the conventional algorithm. Note that the first set of results illustrates the validity of the four proposed algorithms for point-treatment studies while the second set of results is truly concerned with longitudinal data. Note that for the third set of results, the G-computation estimate obtained with the conventional algorithm is not available. This is due to the computing limitations associated with this algorithm which

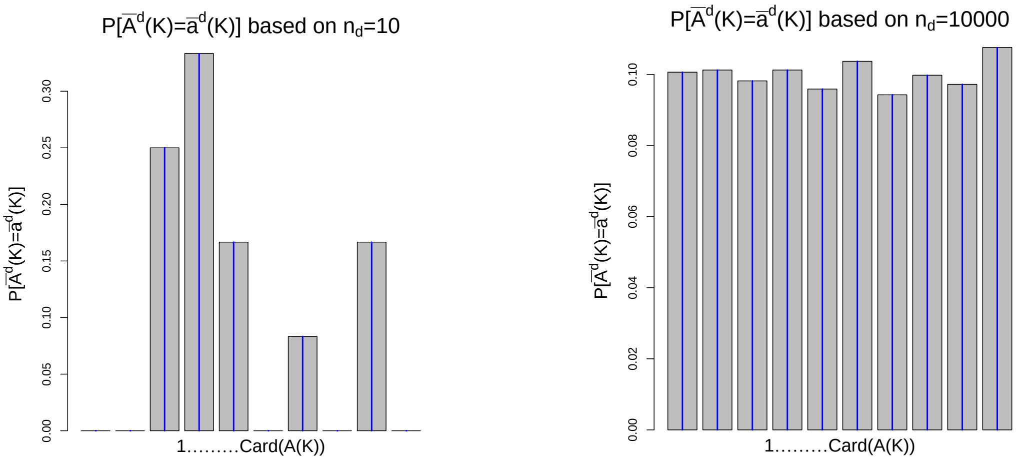


Figure 1: Estimated uniform distribution based on $n_d = 10; 10,000$ draws from $P(\bar{A}^d(K))$ as defined by equality (9).

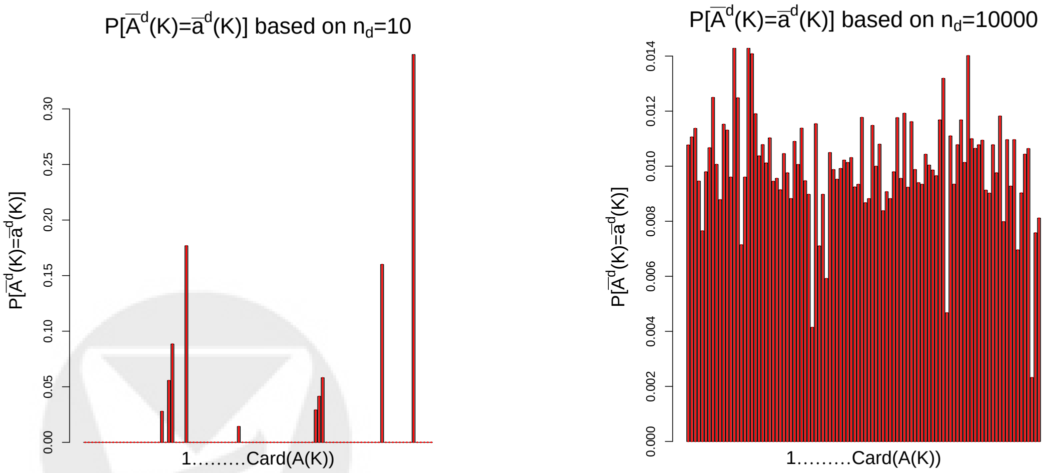


Figure 2: Estimated uniform distribution based on $n_d = 10; 10,000$ draws from $P(\bar{A}^d(K))$ as defined by equality (9).

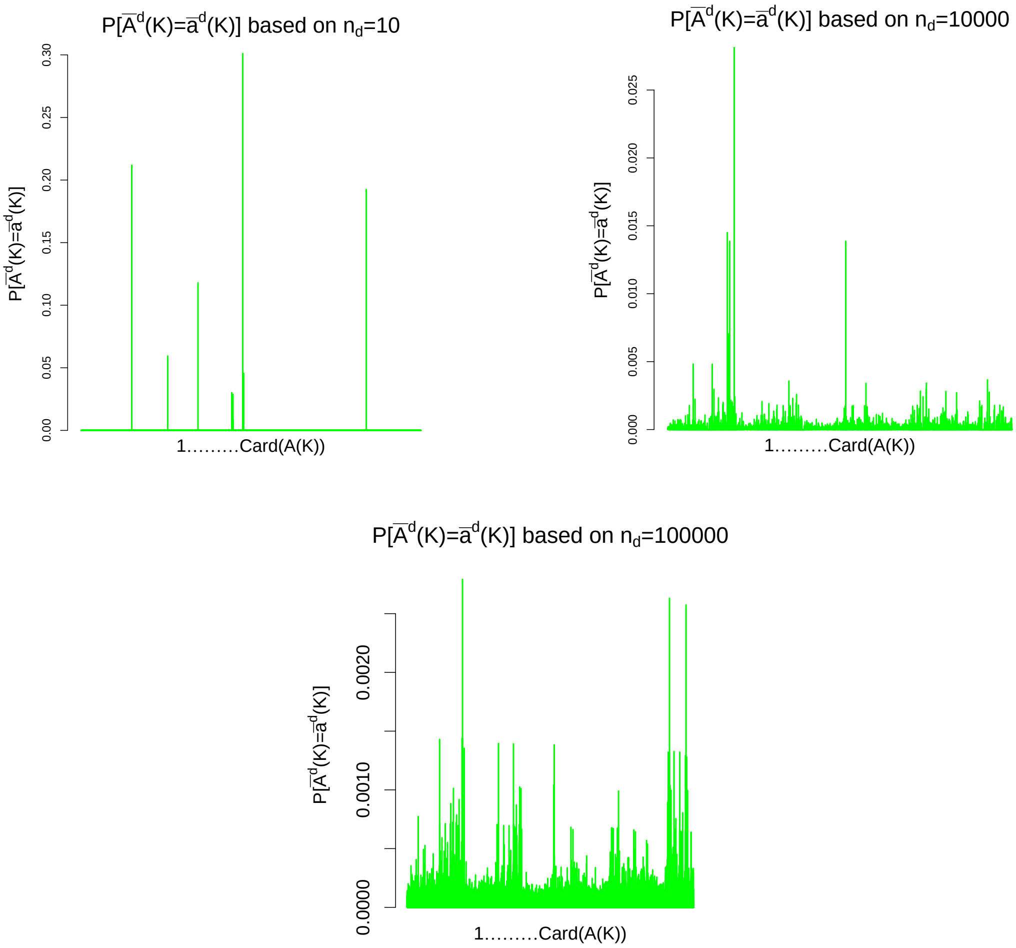


Figure 3: Estimated uniform distribution based on $n_d = 10; 10,000; 100,000$ draws from $P(\bar{A}^d(K))$ as defined by equality (9).

	β_0	β_1	β_2
Conv	-1.219	0.103	-3.018
Proc 1	-1.227	0.102	-2.979
Proc 2	-1.227	0.102	-2.979
Gen proc 1 - $n_d = 10$	-1.221	0.099	-2.969
Gen proc 2 - $n_d = 10$	-1.221	0.099	-2.969
Gen proc 1 - $n_d = 10,000$	-1.228	0.103	-2.978
Gen proc 2 - $n_d = 10,000$	-1.228	0.103	-2.978

Table v: G-computation estimates obtained based on 5 algorithms in the simulation study with longitudinal data where $n = 1000$, $K = 0$, $\alpha = (-1; 0.1; -3)$, $R = 10$, $p = q = 5$ and $n_d = 10; 10,000$.

	β_0	β_1	β_2
Conv	-0.818	0.289	-1.455
Proc 1	-0.827	0.292	-1.460
Proc 2	-0.827	0.292	-1.459
Gen proc 1 - $\mathbf{n_d} = 10$	**0.653**	**-0.139**	**-0.402**
Gen proc 2 - $\mathbf{n_d} = 10$	**0.649**	**-0.137**	**-0.403**
Gen proc 1 - $n_d = 10,000$	-0.859	0.296	-1.411
Gen proc 2 - $n_d = 10,000$	-0.859	0.297	-1.414

Table vi: G-computation estimates obtained based on 5 algorithms in the simulation study with longitudinal data where $n = 1000$, $K = 1$, $\alpha = (-1; 0.1; -3)$, $\gamma = (-1; 0.1; 0.8; -2; 3)$, $R = 10$, $p = q = 20$ and $n_d = 10; 10,000$.

prevent us from applying this algorithm to the corresponding simulated data. Indeed, since $N = 10,000$, $K = 3$, $\text{Card}(\mathcal{T}) = 3$ and $\text{Card}(A(K)) = 10^{(K+1)}$, the number of observations in the weighted regression involved in the conventional algorithm would be: $N_g = 400,000,000$ as defined by equality (4). The third set of results thus illustrates how the four proposed algorithms overcome the implementation problems inherent to the conventional algorithm. In addition, the results associated with the generalized procedures 1 and 2 illustrate the importance of using a large value for n_d to insure proper treatment discretization, i.e. discretization of $\mathcal{A}_V(K)$: Gen proc 1 and 2 with $n_d = 10$ indeed lead to bias estimation as shown in tables vi and vii.

	β_0	β_1	β_2
Conv	NA	NA	NA
Proc 1	-1.228	0.365	-1.609
Proc 2	-1.228	0.364	-1.609
Gen proc 1 - $\mathbf{n_d = 10}$	**-1.465**	**0.497**	**-1.197**
Gen proc 2 - $\mathbf{n_d = 10}$	**-1.468**	**0.498**	**-1.196**
Gen proc 1 - $n_d = 10,000$	-1.313	0.396	-1.642
Gen proc 2 - $n_d = 10,000$	-1.313	0.396	-1.644
Gen proc 1 - $n_d = 100,000$	-1.223	0.363	-1.620
Gen proc 2 - $n_d = 100,000$	-1.221	0.362	-1.623

Table vii: G-computation estimates obtained based on 5 algorithms in the simulation study with longitudinal data where $n = 1000$, $K = 3$, $\alpha = (-1; 0.2; -3)$, $\gamma = (-1; 0.2; 0.2; -3; 3)$, $R = 10$, $p = q = 50$ and $n_d = 10; 10,000; 100,000$. NA stands for not available.

7.1 Point-treatment simulation with continuous treatment

7.2 Aims and protocol

This simulation study illustrates the results and concepts developed in section 5 (treatment discretization, parameter interpretation, generalized nonparametric causal effects) with point-treatment data and a continuous treatment variable. Note that even though the results in section 5 are very general and apply to longitudinal data problems, we chose to present this point-treatment example for illustration clarity since point-treatment data facilitate graphical representations of the concepts and results also relevant to the more complex longitudinal data structure.

The observed data in this simulation are $n = 1,000$ i.i.d. observations of the treatment A, outcome Y, and baseline covariate W which confounds the effect of A on Y. The distribution of W is a mixture of two uniform distributions $\mathcal{U}_1$ and $\mathcal{U}_2$ over $[1; 80]$ and $[80; 100]$ respectively. The distribution of A conditional on W is gaussian with mean W and standard deviation 5, $\mathcal{N}(W, 5)$. The distribution of Y conditional on A and W is gaussian with mean $5 + (\frac{A}{15})^4 + 4W - \frac{10}{85.05}AW$ and standard deviation 1, $\mathcal{N}(5 + (\frac{A}{15})^4 + 4W - \frac{10}{85.05}AW, 1)$. Thus in this simulation, the observed data distribution implies correct model specification of the following causal model for $E(Y_a)$:

$$m(a \mid \beta) = 345.2 + \left(\frac{a}{15}\right)^4 - 10a, \text{ where } \beta = (345.2, \left(\frac{1}{15}\right)^4, -10)$$

since $E(W) = 0.1 \times \frac{81}{2} + 0.9 \times \frac{180}{2} = 85.05$. The causal curve of interest captured by this model is represented in figure 4. We wish to investigate this causal curve with a misspecified linear causal model, $m(a \mid \beta) = \beta_0 + \beta_1 a$ so as to obtain an overall summary of the causal curve of interest over the whole range of possible treatment levels. In other

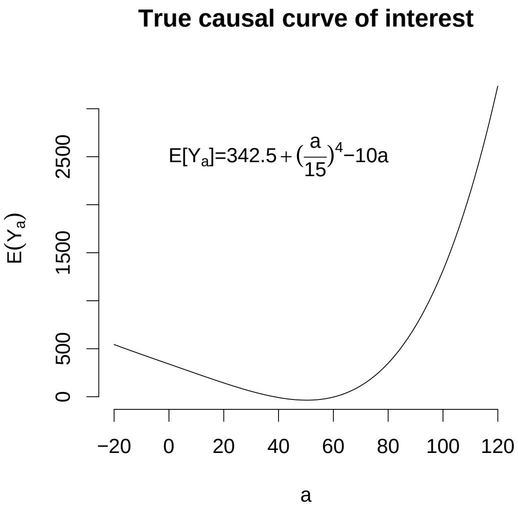


Figure 4: True causal curve of interest in the simulation study with point-treatment data and continuous outcome.

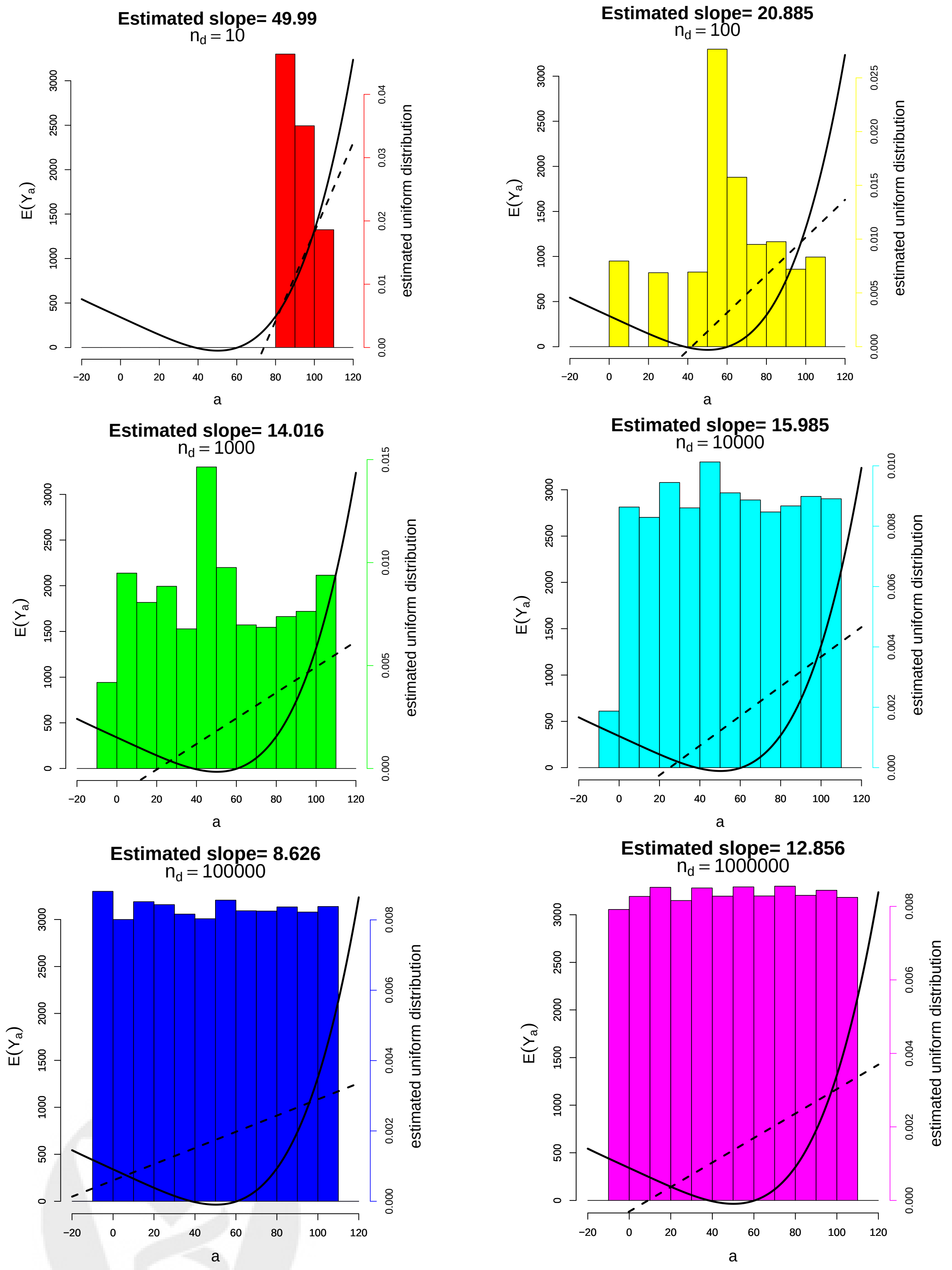
words, the initial causal parameter of interest is defined by equality (7) where g_t is the uniform distribution.

We used the procedures described in section 5 with $N = 10,000$, $R = 1,000$ and $p = q = 100$ to implement the G-computation estimator of the parameter of interest. Note that in point-treatment studies, the implementations of the G-computation estimator with procedures 1 and 2 are identical. The corresponding algorithm relies on proper discretization of the set of possible treatments, $\mathcal{A}_V(K)$, where $K = 0$ and $V = \varnothing$. Proper discretization depends on consistent estimation of the marginal distribution of A which is typically unknown in practice. In this simulation, we have a close form representation of this distribution:

$$g(A) = \frac{0.1}{79}\int_1^{80} g(A \mid w)dw + \frac{0.9}{20}\int_{80}^{100} g(A \mid w)dw, \tag{12}$$

where $g(A \mid W)$ is the density of the conditional gaussian distribution of A given W: $\mathcal{N}(W, 5)$. To illustrate the consequence of model misspecification for this distribution and the impact of the choice for n_d, we computed the G-computation estimates based on a) correct and incorrect model specification for $g(A)$, and b) $n_d = 10$; 100 ; $1,000$; $10,000$; $100,000$; $1,000,000$. Results are presented in figures 5 and 6 and are based on correct model specification for the Q_{F_X} part of the likelihood, i.e. for $E(Y \mid A, W)$. The incorrect model for the marginal distribution of A is the gaussian distribution with mean equal to the empirical mean of the observed treatment, A, and standard deviation equal to the empirical standard deviation of the observed treatment, A.

$$A \sim \mathcal{N}\Big(\frac{1}{n}\sum_{i=1}^{n} a_i, \Big[\frac{1}{n}\sum_{i=1}^{n}(a_i - \frac{1}{n}\sum_{i=1}^{n} a_i)^2\Big]^{\frac{1}{2}}\Big) \tag{13}$$



Estimated slope= 49.99
$n_d = 10$

Estimated slope= 20.885
$n_d = 100$

Estimated slope= 14.016
$n_d = 1000$

Estimated slope= 15.985
$n_d = 10000$

Estimated slope= 8.626
$n_d = 100000$

Estimated slope= 12.856
$n_d = 1000000$

Figure 5: G-computation estimates based on 6 discretizations of the set of possible treatments corresponding with increasing values for n_d and a correctly specified model for $g(A)$ as defined by equality (12). The plain curves represent the true causal curve of interest. The dashed curves represent the estimated nonparametric causal effects based on a linear causal model. The histograms represent the estimated uniform distribution over $-10 \leq a \leq 110$ based on 100,000 draws from (9).

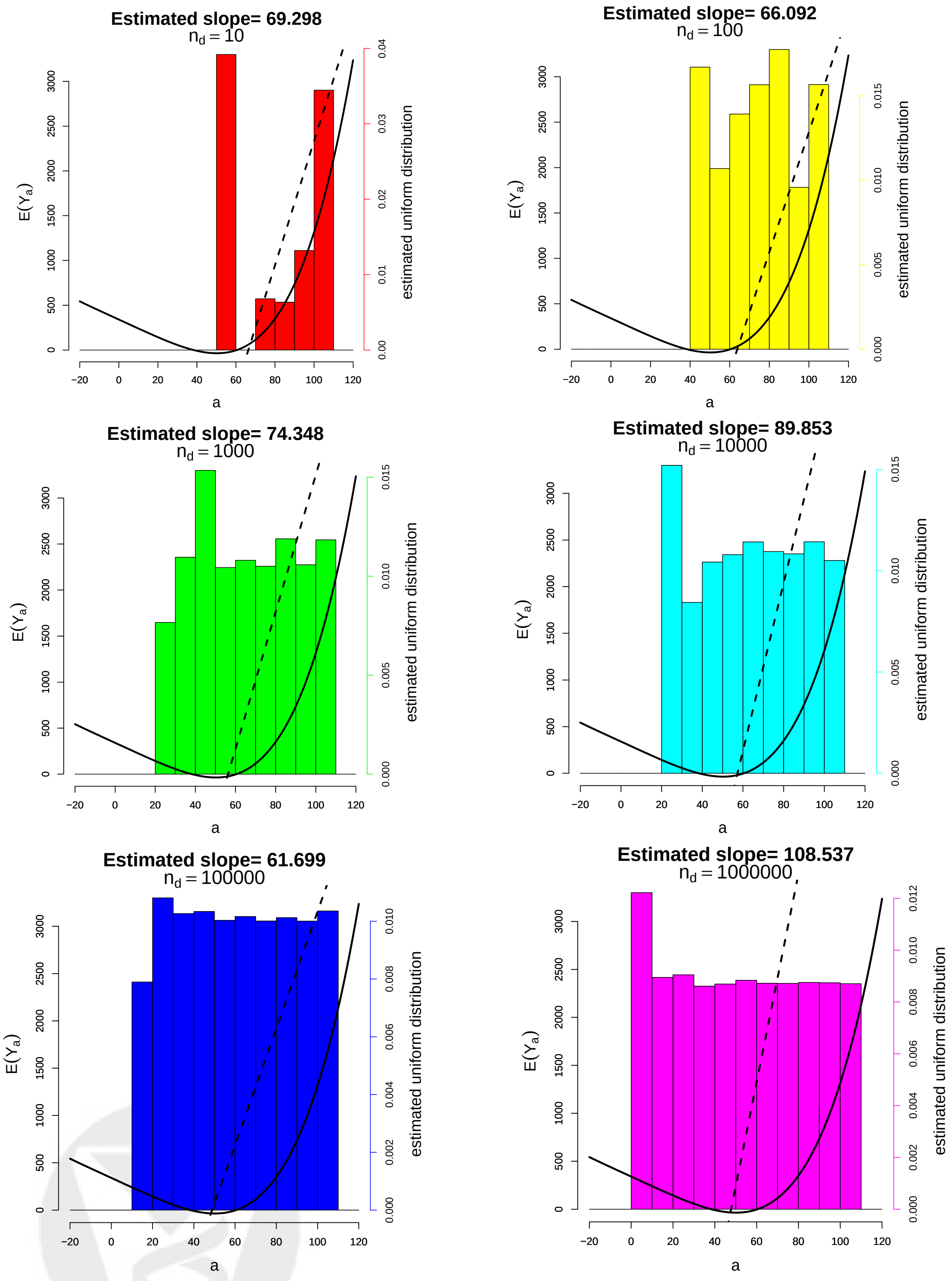


Figure 6: G-computation estimates based on 6 discretizations of the set of possible treatments corresponding with increasing values for n_d and a misspecified model for $g(A)$ as defined by (13). The plain curves represent the true causal curve of interest. The dashed curves represent the estimated nonparametric causal effects based on a linear causal model. The histograms represent the estimated uniform distribution over $-10 \leq a \leq 110$ based on 100,000 draws from (9).

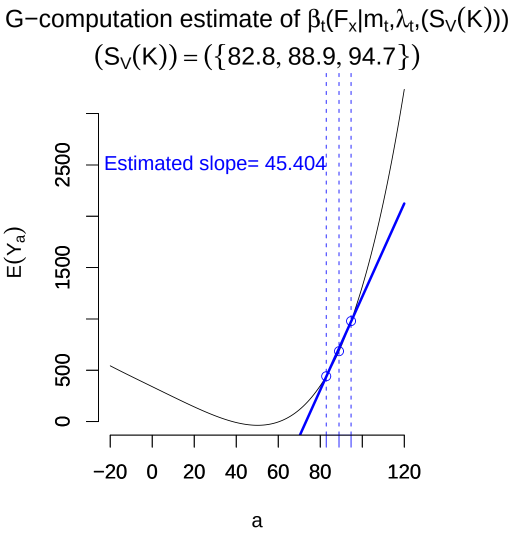


Figure 7: G-computation estimate of $\beta_t(F_X \mid m_t, \lambda_t, (\mathcal{S}_V(K)))$ based on point-treatment data ($t = 0$, $K = 0$ and $V = \varnothing$) with a continuous treatment where $\mathcal{S}(0)$ is defined based on the first, second and third quartiles of the observed treatment A.

Finally, we illustrate the generalization (10) of the definition of nonparametric causal effects with the following choices for $\mathcal{S}(0)$: a) $\mathcal{S}(0)$ is defined based on the first, second and third quartiles of the observed treatment A, b) $\mathcal{S}(0)$ is defined based on four treatment levels obtained by splitting the range of the observed treatment A in four equal segments, and c) $\mathcal{S}(0)$ is defined based on the 0.1^{th}, first and second percentiles of the observed treatment A. We computed the G-computation estimates of the three parameters of interest defined by the three choices for $\mathcal{S}(0)$ and a CKS which gives equal weights to all elements in $\mathcal{S}(0)$. The three estimates were obtained from the implementation procedure described in section 3.3 and adapted to the estimation of generalized nonparametric causal effects. They are represented on figures 7, 8, and 9 respectively. Note that these estimates are also based on correct model specification for the Q_{F_X} part of the likelihood.

7.3 Interpretation

This simulation study clearly underlines the importance of proper discretization of the infinite set of possible treatment regimens through 1) correct specification of a model for $g(\bar{A}(K) \mid V)$, and 2) a large value for n_d (e.g. $n_d = 100,000$). This simulation study demonstrates the impact on G-computation estimation when n_d is chosen too small or when the model for $g(\bar{A}(K) \mid V)$ is misspecified. In such cases, the G-computation estimates cannot be clearly interpreted and the investigator may thus prefer to lower his/her research ambition by estimating the generalized nonparametric causal effects as defined by equalities (10) and (11). The interpretation of such parameters is then entirely controlled by the investigator and a sensitivity-type analysis based on different choices

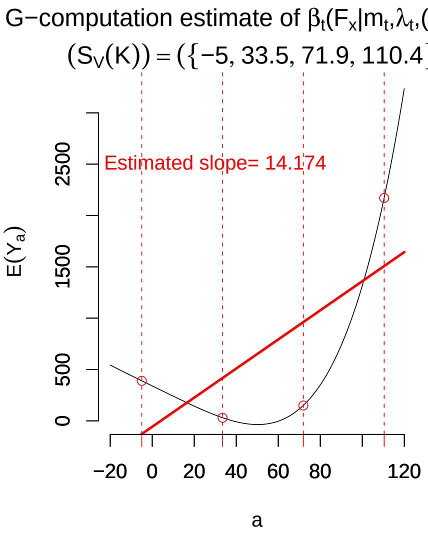


Figure 8: G-computation estimate of $\beta_t(F_X \mid m_t, \lambda_t, (\mathcal{S}_V(K)))$ based on point-treatment data ($t = 0$, $K = 0$ and $V = \varnothing$) with a continuous treatment where $\mathcal{S}(0)$ is defined based on four treatment levels obtained by splitting the range of the observed treatment A in four equal segments.

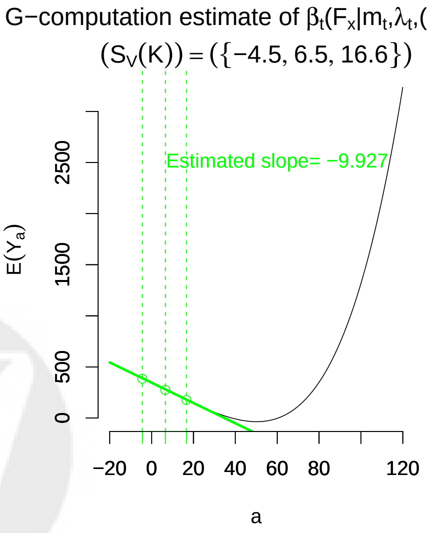


Figure 9: G-computation estimate of $\beta_t(F_X \mid m_t, \lambda_t, (\mathcal{S}_V(K)))$ based on point-treatment data ($t = 0$, $K = 0$ and $V = \varnothing$) with a continuous treatment where $\mathcal{S}(0)$ is defined based on the 0.1^{th}, first and second percentiles of the observed treatment A.

for the CKS and $(S_V(K))$ may enable the investigator to gain sufficient insight about the causal curve of interest even when the causal model is misspecified.

Appendix

A Important results

lemma A.1 *We adopt the notations previously introduced in this manuscript. In particular, $\mathcal{A}_V^d(K)$ represents the set of n_d i.i.d. observations of $\bar{A}(K)$ with distribution $g(\bar{A}(K) \mid V)$:*

$$\mathcal{A}_V^d(K) \equiv \{\bar{a}_1^d(K), \ldots, \bar{a}_{n_d}^d(K)\},$$

and $\mathcal{A}_V^{d,u}(K)$ represents the set of unique elements of $\mathcal{A}_V^d(K)$. Consider the discrete random process, $\bar{A}^d(K)$, with conditional probability over $\mathcal{A}_V^{d,u}(K)$:

$$P(\bar{A}^d(K) = \bar{a}^d(K) \mid V) = I(\bar{a}^d(K) \in \mathcal{A}_V^{d,u}(K)) \sum_{i=1}^{n_d} \frac{\frac{I(\bar{a}_i^d(K)=\bar{a}^d(K))}{g(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}}. \tag{14}$$

The limit distribution of $\bar{A}^d(K)$ is the uniform distribution of $\bar{A}(K)$ over the support $\mathcal{A}_V(K)$.

Proof:
Note first that equality (14) indeed defines a probability with support $\mathcal{A}_V^{d,u}(K)$:

$$\begin{aligned} P\Big(\cup_{\bar{a}^d(K)\in\mathcal{A}_V^{d,u}(K)} \bar{A}^d(K) = \bar{a}^d(K) \mid V \Big) &= \sum_{\bar{a}^d(K)\in\mathcal{A}_V^{d,u}(K)} P\Big(\bar{A}^d(K) = \bar{a}^d(K) \mid V\Big) \\ &= \sum_{\bar{a}^d(K)\in\mathcal{A}_V^{d,u}(K)} \sum_{i=1}^{n_d} \frac{\frac{I(\bar{a}_i^d(K)=\bar{a}^d(K))}{g(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}} \\ &= \frac{\sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n} \frac{1}{g(\bar{a}_i^d(K)|V)}} \\ &= 1 \end{aligned}$$

We obviously have:

$$\lim_{n_d \longrightarrow +\infty} \mathcal{A}_V^{d,u}(K) = \mathcal{A}_V(K)$$

and thus the support of the limit distribution of $\bar{A}^d(K)$ is indeed $\mathcal{A}_V(K)$. In addition we have for any set $\mathcal{A} \subset \mathcal{A}_V(K)$:

$$\begin{aligned}
\lim_{n_d \longrightarrow +\infty} P\Big(\bar{A}^d(K) \in \mathcal{A} \mid V\Big) &= \lim_{n_d \longrightarrow +\infty} \sum_{\bar{a}^d(K) \in \mathcal{A}} P\Big(\bar{A}^d(K) = \bar{a}^d(K) \mid V\Big) \\
&= \lim_{n_d \longrightarrow +\infty} \sum_{\bar{a}^d(K) \in \mathcal{A}} I(\bar{a}^d(K) \in \mathcal{A}_V^{d,u}(K)) \frac{\sum_{i=1}^{n_d} \frac{I(\bar{a}_i^d(K) = \bar{a}^d(K))}{g(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}} \\
&= \lim_{n_d \longrightarrow +\infty} \sum_{\bar{a}^d(K) \in \mathcal{A}_V^{d,u}(K)} \frac{\sum_{i=1}^{n_d} \frac{I(\bar{a}_i^d(K) = \bar{a}^d(K)) I(\bar{a}_i^d(K) \in \mathcal{A})}{g(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}} \\
&= \lim_{n_d \longrightarrow +\infty} \frac{\sum_{i=1}^{n_d} \frac{I(\bar{a}_i^d(K) \in \mathcal{A})}{g(\bar{a}_i^d(K)|V)}}{\sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}} \\
&= \lim_{n_d \longrightarrow +\infty} \frac{\frac{1}{n_d} \sum_{i=1}^{n_d} \frac{I(\bar{a}_i^d(K) \in \mathcal{A})}{g(\bar{a}_i^d(K)|V)}}{\frac{1}{n_d} \sum_{i=1}^{n_d} \frac{1}{g(\bar{a}_i^d(K)|V)}} \\
&= \lim_{n_d \longrightarrow +\infty} \frac{E_g\left(\frac{I(\bar{A}(K) \in \mathcal{A})}{g(\bar{A}(K)|V)} \mid V\right)}{E_g\left(\frac{1}{g(\bar{A}(K)|V)} \mid V\right)} \\
&= \lim_{n_d \longrightarrow +\infty} \frac{\int_{\bar{a}(K) \in \mathcal{A}_V(K)} \frac{I(\bar{a}(K) \in \mathcal{A})}{g(\bar{a}(K)|V)} g(\bar{a}(K) \mid V) d\mu(\bar{a}(K))}{\int_{\bar{a}(K) \in \mathcal{A}_V(K)} \frac{1}{g(\bar{a}(K)|V)} g(\bar{a}(K) \mid V) d\mu(\bar{a}(K))} \\
&= \lim_{n_d \longrightarrow +\infty} \frac{\int_{\bar{a}(K) \in \mathcal{A}_V(K)} I(\bar{a}(K) \in \mathcal{A}) d\mu(\bar{a}(K))}{\int_{\bar{a}(K) \in \mathcal{A}_V(K)} d\mu(\bar{a}(K))} \\
&= \lim_{n_d \longrightarrow +\infty} \frac{\int_{\bar{a}(K) \in \mathcal{A}} d\mu(\bar{a}(K))}{\int_{\bar{a}(K) \in \mathcal{A}_V(K)} d\mu(\bar{a}(K))} \text{ since } \mathcal{A} \subset \mathcal{A}_V(K).
\end{aligned}$$

References

[1] R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: Characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics, 1995*, Lecture Notes in Statistics, pages 255–294, New York, 1997. Springer.

[2] R. Neugebauer and M.J. van der Laan. Locally efficient estimation of nonparametric causal effects on mean outcomes in longitudinal studies. Working paper 134, U.C. Berkeley Division of Biostatistics Working Paper Series, 2003. [www http://www.bepress.com/ucbbiostat/paper134/].

[3] R. Neugebauer and M.J. van der Laan. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, 129:405–426, February 2005. [www http://www.sciencedirect.com/science/article/B6V0M-4D5X61S-1/2/b7d0d4635a72e22ee511cce500b33901].

[4] D.B. Rubin. Inference and missing data. *Biometrika*, pages 581–590, 1976.

[5] M.J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York, 2002.