University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 134

Locally Efficient Estimation of Nonparametric Causal Effects on 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

http://biostats.bepress.com/ucbbiostat/paper134

Locally Efficient Estimation of Nonparametric Causal Effects on Mean Outcomes in Longitudinal Studies

Romain Neugebauer and Mark J. van der Laan

Abstract

Marginal Structural Models (MSM) have been introduced by Robins (1998a) as a powerful tool for causal inference as they directly model causal curves of interest, i.e. mean treatment-specific outcomes possibly adjusted for baseline covariates. Two estimators of the corresponding MSM parameters of interest have been proposed, see van der Laan and Robins (2002): the Inverse Probability of Treatment Weighted (IPTW) and the Double Robust (DR) estimators. A parametric MSM approach to causal inference has been favored since the introduction of MSM. It relies on correct specification of a parametric MSM to consistently estimate the parameter of interest using the IPTW or DR estimator. In this paper, we develop an alternative nonparametric MSM approach to causal inference that extends the definition of causal parameters of interest. Such an approach is particularly suitable for investigating causal effects in practice as it does not require the assumption of a correctly specified MSM. We first propose a methodology to generate nonparametric parameters of interest for investigating causal curves in which the treatment is longitudinal. We provide insight on how to interpret these parameters in practice and choose the parameter of interest to best answer the causal question of interest. We also provide two estimators consistent with this approach, i.e. which do not entirely rely, even indirectly, on correct specification of a MSM: the unique IPTW and locally efficient DR estimators. All results are illustrated with a simulation study in which the practical performances of the DR estimators are evaluated for the first time using longitudinal non-survival data. In the last section, we compare the proposed nonparametric MSM approach to causal inference to the more typical parametric MSM approach and contribute to the general understanding of MSM estimation by addressing the issue of MSM misspecification.

1 Introduction

1.1 Data structure

The results presented in this paper apply to causal inference for both point treatment and longitudinal non-survival data structures. See Yu and van der Laan (2002b) for an application to longitudinal survival data. We will use notations to represent longitudinal non-survival data since it is more general and can also be used for the more simple structure of point treatment data. For more details on the statistical framework used in this paper, see chapter 6 of van der Laan and Robins (2002).

For every subject we observe a treatment regimen $(A(0), \ldots, A(k))$ over time $j = 0, \ldots, k$ and covariates $(L(0), \ldots, L(k+1))$ measured at baseline and when treatment changes. The covariates $L(j)$ are measured after $A(j-1)$ and before $A(j)$. The outcome of interest is defined as $Y = L(k+1)$. Thus the observed data is:

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

where the notation $\bar{W}(j)$ represents the history of the variable W between time $t = 0$ and $t = j \geq 0$: $\bar{W}(j) = \{W(0), \ldots, W(j)\}$. If $j < 0$, $\bar{W}(j)$ is defined as the empty set. For simplicity we will also denote $\bar{A}(k)$ and $\bar{L}(k+1)$ with $\bar{A}$ and $\bar{L}$ respectively.
We define V as a subset of the baseline covariates, $V \subset L(0)$. We denote the observed value for any random variable W and for individual i with w_i and denote the number of individuals in the observed data set with n: $i = 1, \ldots, n$.

1.2 Assumptions

- **Existence of counterfactuals:** we assume the existence of the counterfactuals or the treatment specific process $L_{\bar{a}}(j)$, for $j = 0, \ldots, k+1$ and every treatment regimen $\bar{a} = (a(0), \ldots, a(k)) \in \mathcal{A}$ where $\mathcal{A}$ designates every possible treatment regimen. See Rubin (1976) for details on the concept of counterfactuals. We denote the full data process with $X = (\bar{L}_{\bar{a}}(k+1))_{\bar{a} \in \mathcal{A}}$ and its distribution with F_X. We also denote the distribution of V in the full data with F_V and the support of F_V with S_V.

- **Consistency assumption:** at any time point j, we assume the following link between the observed data and the counterfactuals: $L(j) = L_{\bar{A}}(j)$. Under this assumption, we have:

$$O = (\bar{A}(k), \bar{L}_{\bar{A}(k)}(k+1)) \equiv \phi(\bar{A}, X),$$

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 j, we assume that the treatment specific process can only be affected by past treatments:

$$\bar{L}_{\bar{a}}(j) = \bar{L}_{\bar{a}(j-1)}(j).$$

- **Sequential Randomization Assumption (SRA):** at any time point j, we assume that the observed treatment is independent of the full data given the data observed before time point j:

$$A(j) \perp X \mid \bar{A}(j-1), \bar{L}(j).$$

Under the SRA, the treatment mechanism, i.e. the conditional density or probability of $\bar{A}$ given X: $g(\bar{A} \mid X)$, is such that:

$$g(\bar{A} \mid X) = \prod_{j=0}^{k} g(A(j) \mid \bar{A}(j-1), X) \stackrel{SRA}{=} \prod_{j=0}^{k} g(A(j) \mid \bar{A}(j-1), \bar{L}(j)).$$

The SRA implies coarsening at random, Gill, van der Laan and Robins (1997), 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. Thus, we denote the distribution of the observed data O with $P_{F_X,g}$ and the likelihood of O is:

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

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

1.3 Description of the problem and approach

We wish to use this statistical framework to investigate the adjusted causal effect of $\bar{A}$ on Y adjusted for V defined as the parameter function $m^* : \mathcal{A} \times S_V \longrightarrow \mathbb{R}$ where $m^*(\bar{a}, V) = E_{F_X}(Y_{\bar{a}} \mid V)$. This parameter function is defined for any $F_X \in \mathcal{M}^F_{NP}$ where $\mathcal{M}^F_{NP}$ is the set of all distributions F_X and we will refer to m^* as the causal curve.

This issue is addressed in this paper by developing estimators of euclidean nonparametric causal parameters of m^*, i.e. finite vectors of real valued summary measures of the adjusted causal effect of interest defined without further assumptions, in particular on the causal curve, i.e. on F_X. We will refer to these parameters of interest as nonparametric causal effects.

Parameters of interest are defined nonparametrically in this approach and that is why we wish to develop nonparametric estimators, i.e. estimators which do not rely on further assumptions. However due to the 'curse of dimensionality', see van der Laan and Robins (2002), this goal cannot be typically achieved in practice and one needs to make additional assumptions on g or Q_{F_X}. We thus favor two estimators which do not rely entirely on further assumption on the distribution defining the parameter of interest, i.e. F_X: the IPTW and DR estimators.

In section 2, we first present a methodology for generating nonparametric potential euclidean causal parameters of interest for investigating the causal curve, m^*. We provide the interpretation of such parameters and describe how to apply this methodology to define the actual parameter of interest in practice. In section 3, we describe two estimators of these potential parameters of interest, one of which is locally efficient: the IPTW and DR estimator. The so-called G-computation estimator is also described in that section as it is required in the procedure used in this paper to obtain the DR estimate in practice. In section 4, we illustrate the results presented in this paper by a simulation study. We finally discuss the importance of this nonparametric MSM approach to causal inference in section 5 and compare it to the typical parametric MSM approach found in the literature.

2 Methodology for generating nonparametric euclidean causal parameters of interest

2.1 Definition

We denote the set of functions from $\mathcal{A} \times S_V$ to $\mathbb{R}$ with $\mathcal{M}^*$. We have $m^* \in \mathcal{M}^*$. We define $m : \mathbb{R}^k \longrightarrow \mathcal{M}^*$ where $m(\beta) = m(\cdot, \cdot \mid \beta) \in \mathcal{M}^*$ and denote the image of m with $\mathcal{M}$. We have $\mathcal{M} \subset \mathcal{M}^*$ but note that m^* is not necessarily an element of $\mathcal{M}$. We define $\lambda : \mathcal{A} \times S_V \longrightarrow \mathbb{R}$ where λ is different from the null function. We will refer to m and λ as the causal model (CM) and causal kernel smoother (CKS) respectively and will justify these appellations in the next section.

For any CM, m, and CKS, λ, we define the following potential causal parameter of interest for investigating a causal curve:

$$\beta(\cdot \mid m, \lambda) : \mathcal{M}^F_{NP} \longrightarrow \mathbb{R}^k, \text{ where:}$$

$$\beta(F_X \mid m, \lambda) \equiv \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \, E_{F_X} \left[\sum_{\bar{a} \in \mathcal{A}} \left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta) \right)^2 \lambda(\bar{a}, V) \right]. \qquad (1)$$

Equality (1) defines a nonparametric causal effect, i.e. a euclidean parameter of the causal curve defined without further assumption on F_X. We denote the resulting parameter with $\beta_{m,\lambda} = \beta(F_X \mid m, \lambda)$. We do not state the conditions under which this parameter exists and is unique in this paper. We do, however, demonstrate in the next section why such a parameter is indeed of potential interest for investigating a causal curve.

2.2 Interpretation

We define F_λ such that $dF_\lambda(\bar{a}, V) = I(\bar{a} \in \mathcal{A})\lambda(\bar{a}, V)dF_V(V)$. The space of functions $\mathcal{M}^*$ endowed with the inner product:

$$< f, g >_{F_\lambda} = \sum_{\bar{a}} \int f(\bar{a}, V)g(\bar{a}, V)dF_\lambda$$

is a Hilbert space and we denote the distance defined in this Hilbert space with $||\cdot||_{F_\lambda}$ and we have $||f||_{F_\lambda} = \sqrt{< f, f >_{F_\lambda}}$.
Using these notations, we have:

$$\beta_{m,\lambda} = \underset{\beta \in I\!R^k}{\operatorname{argmin}} ||m^* - m(\cdot\,, \cdot \mid \beta)||_{F_\lambda}. \tag{2}$$

This is shown as follows:

$$\begin{aligned}
\beta_{m,\lambda} &\equiv \underset{\beta \in I\!R^k}{\operatorname{argmin}} E_{F_X} \sum_{\bar{a} \in \mathcal{A}} \left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta)\right)^2 \lambda(\bar{a}, V) \\
&= \underset{\beta \in I\!R^k}{\operatorname{argmin}} E_{F_X} \sum_{\bar{a} \in \mathcal{A}} \left[\left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta)\right)^2 - \left(Y_{\bar{a}} - m^*(\bar{a}, V)\right)^2\right] \lambda(\bar{a}, V) \\
&= \underset{\beta \in I\!R^k}{\operatorname{argmin}} E_{F_X} \sum_{\bar{a} \in \mathcal{A}} \left(2Y_{\bar{a}} - m(\bar{a}, V \mid \beta) - m^*(\bar{a}, V)\right) \left(m^*(\bar{a}, V) - \right. \\
&\qquad\qquad \left. m(\bar{a}, V \mid \beta)\right) \lambda(\bar{a}, V) \\
&= \underset{\beta \in I\!R^k}{\operatorname{argmin}} E_{F_V} \sum_{\bar{a} \in \mathcal{A}} \left(m^*(\bar{a}, V) - m(\bar{a}, V \mid \beta)\right)^2 \lambda(\bar{a}, V) \\
&= \underset{\beta \in I\!R^k}{\operatorname{argmin}} \sum_{\bar{a} \in \mathcal{A}} \int \left(m^*(\bar{a}, V) - m(\bar{a}, V \mid \beta)\right)^2 \lambda(\bar{a}, V) dF_V(V).
\end{aligned} \tag{3}$$

Equality (2) makes the interpretation of $\beta_{m,\lambda}$ more explicit. It is the parameter minimizing the distance between the causal curve m^* and m where the notion of distance is defined by the choice of function λ. This justifies why we refer to m as the causal model: it approximates the causal curve, m^*, best at the parameter of interest.
Equality (3) shows more explicitly how λ influences the definition of $\beta_{m,\lambda}$ by defining the notion of distance between the CM and the causal curve. The function λ attributes a weight to each region $(\bar{a}, V)$ of $\mathcal{A} \times S_V$ so as to define the parameter of interest such that the CM approximates the causal curve best in the regions given

the more weight. This justifies why we refer to λ as the causal kernel smoother: it defines the smoothing of the causal curve using the CM by weighting each region of the causal curve, the higher weights being in the regions in which we wish to focus the investigation of the causal curve. From equality (2), it is obvious that the choice for λ does not influence the definition of $\beta_{m,\lambda}$ when the CM is correctly specified, i.e. when $\mathcal{M} \ni m^*$ we have $\beta_{m,\lambda} = \beta_m$, where β_m is defined such that $m(\cdot\,, \cdot \mid \beta_m) = m^*$.

To summarize, $\beta_{\lambda,m}$ is the parameter for which the CM approximates the causal curve best in the regions defined by the CKS. Therefore, the parameter $\beta_{m,\lambda}$ is indeed of potential interest for investigating the causal curve. It provides a global and accurate summary measure of the causal curve when the CM is correctly specified and provides a localized and approximate summary measure of the causal curve when the CM is misspecified.

We described a method to generate potential parameters of interest for investigating a causal curve through the choice of a CM and a CKS. We now provide practical guidelines on how to choose m and λ in order to generate actual parameters of interest, i.e. parameters answering the causal questions of interest.

2.3 Application in practice

The parameters of interest to be generated using the methodology presented previously are dictated by the questions of interest to be answered, i.e. the choice for m and λ follows from the causal aims of the analysis. We can crudely classify causal aims into two categories:

- **Obtaining as global and accurate a representation of the causal curve as possible.** The CM, m, should be chosen such that: $\mathcal{M} \ni m^*$, i.e. the CM needs to be correctly specified. If the CM is correctly specified, the CKS, λ, can be chosen arbitrarily since the definition of the parameter of interest is then independent of the choice for λ. This analytical goal relies on the implicit belief that one knows or can select a correctly specified CM.

- **Obtaining an informative summary, possibly localized, representation of the causal curve.** The CM is chosen so as to extract the informative and summary trend of interest from the causal curve. The CM will typically be willingly misspecified. As an example, one will assume a linear CM if one is interested in capturing the general monotone trend of m^*. This will be assumed even if one suspects that the causal curve is more complex. The choice for the CKS is then essential to define the parameter of interest since the CM can be misspecified. The CKS should be specified such that the regions given the highest weights are the regions of the causal curve for which

one wants to obtain an informative summary representation.
Possible choices for λ:

- any known function of $\bar{a}$ and V. In particular, $\lambda^{uni}(\bar{a}, V) = 1$: this uniform CKS is recommended when one wants to obtain a global summary representation of the causal curve, i.e. when one wants the CM to approximate the causal curve equally well in all regions of the causal curve.
- $\lambda^{loc}_{var}(\bar{a}, V) = \frac{1}{V_{F_X}(\varepsilon(\beta)|\bar{a},V)}$: such a CKS is recommended when one wants the CM to approximate the causal curve best in the regions where the conditional observed residual variance is smallest.
- $\lambda^{loc}_{g}(\bar{a}, V) = g(\bar{a} \mid V)$: such a CKS is recommended when one wants the CM to approximate the causal curve best in the regions where $\bar{a}$ is more represented within strata of V in the observed data.
- $\lambda^{loc}_{var,g}(\bar{a}, V) = \frac{g(\bar{a}|V)}{V_{F_X}(\varepsilon(\beta)|\bar{a},V)}$: such a CKS is recommended when one wants the CM to approximate the causal curve best in the regions where both the conditional observed residual variance is smallest and $\bar{a}$ is more represented within strata of V in the observed data.

In the last section, we discuss in more details the practical implications of these two analytical goals and how they can relate more directly to the MSM parametric or MSM nonparametric approach to causal inference.

We described one method to generate potential nonparametric parameters of interest but there are however other nonparametric methods to generate such euclidean parameters of the causal curve without specification of a correct MSM. For instance one could define the following parameters of interest:

$$\beta(F_X \mid \lambda) = E_{F_X}\left[\sum_{\bar{a}\bar{\in}\mathcal{A}} \frac{d}{d\bar{a}} m^*(\bar{a}, V)\lambda(\bar{a}, V)\right] \text{ or}$$

$$\beta(F_X \mid \lambda) = E_{F_X}\left[\sum_{\bar{a}\bar{\in}\mathcal{A}} m^*(\bar{a}, V)\lambda(\bar{a}, V)\right],$$

where λ can be defined as $\frac{1}{h}K(\frac{v-v_0}{h})$ for K and h being a kernel smoother and a bandwidth respectively. We will focus on the estimation of the potential parameters of interest as defined in this section. The results presented for such parameters, however, can be extended to all nonparametric MSM parameters.

3 Estimators of the parameter of interest

3.1 Full data estimating function

For any CM, m, and any CKS, λ, we define $h_\lambda : \mathcal{A} \times S_V \longrightarrow I\!R$ where:

$$h_\lambda(\bar{a}, V) \equiv \lambda(\bar{a}, V) \frac{\partial}{\partial \beta} m(\bar{a}, V \mid \beta), \tag{4}$$

and

$$D_{h_\lambda}(X \mid \beta) \equiv \sum_{\bar{a} \in \mathcal{A}} h_\lambda(\bar{a}, V) \varepsilon_{\bar{a}}(\beta), \tag{5}$$

where $\varepsilon_{\bar{a}}(\beta) \equiv Y_{\bar{a}} - m(\bar{a}, V \mid \beta)$. Under regularity conditions we have:

$$\beta_{m,\lambda} = \underset{\beta \in I\!R^k}{\operatorname{argmin}} E_{F_X} \left[\sum_{\bar{a} \in \mathcal{A}} \left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta) \right)^2 \lambda(\bar{a}, V) \right]$$
$$\iff E_{F_X} \left[D_{h_\lambda}(X \mid \beta_{m,\lambda}) \right] = 0. \tag{6}$$

This can be shown as follows:

$$\beta_{m,\lambda} \equiv \underset{\beta \in I\!R^k}{\operatorname{argmin}} E_{F_X} \left[\sum_{\bar{a} \in \mathcal{A}} \left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta) \right)^2 \lambda(\bar{a}, V) \right]$$

Under regularity conditions:

$$\iff \frac{d}{d\beta} E_{F_X} \left[\sum_{\bar{a} \in \mathcal{A}} \left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta) \right)^2 \lambda(\bar{a}, V) \right] \Bigg|_{\beta = \beta_{m,\lambda}} = 0$$

Under regularity conditions:

$$\iff E_{F_X} \left[\sum_{\bar{a} \in \mathcal{A}} \lambda(\bar{a}, V) \frac{d}{d\beta} m(\bar{a}, V \mid \beta) \Big|_{\beta = \beta_{m,\lambda}} \left(Y_{\bar{a}} - m(\bar{a}, V \mid \beta_{m,\lambda}) \right) \right] = 0.$$

Note that we also have from the previous equation:

$$E_{F_V} \left[\sum_{\bar{a} \in \mathcal{A}} h_\lambda(\bar{a}, V) \left(m_0^*(\bar{a}, V) - m(\bar{a}, V \mid \beta_{m,\lambda}) \right) \right] = 0. \tag{7}$$

A corollary of equivalence (6) is that $\beta_{m,\lambda}$ can be estimated consistently using the estimating function of the full data, $D_{h_\lambda}(X \mid \beta)$, where h_λ is defined by (4). Note that this estimating function cannot be used directly with the observed data, O, and that it actually only depends on part of the full data: $Q(X) = (V, (Y_{\bar{a}})_{\bar{a} \in \mathcal{A}}) \subset X$.
In addition, it can be shown that this estimating function is the only full data estimating function defining a consistent, regular and asymptotically linear estimator of $\beta_{m,\lambda}$ since F_X is left nonparametric, see van der Laan and Robins (2002). As a result, there will be only one IPTW, G-computation and DR estimator of $\beta_{m,\lambda}$. See van der Laan and Robins (2002) for details on how these estimators are derived from the full data estimating functions defining consistent, regular and asymptotically linear estimators of the parameter of interest.

3.2 IPTW estimator

3.2.1 Definition

The unique IPTW estimating function for $\beta_{m,\lambda}$ with nuisance parameter g is defined as:

$$D_{h_\lambda}(O \mid g, \beta) = \frac{h_\lambda(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)} \text{ where } \varepsilon(\beta) = \varepsilon_{\bar{A}}(\beta),$$

Note that the IPTW estimating function is defined under the SRA since $g(\bar{A} \mid X)$ only depends on the observed data when the SRA holds. Furthermore, we will not refer to λ as a nuisance parameter as it defines the parameter of interest, is user-specified and will typically be a known function. However λ will be treated similar to a nuisance parameter when unknown as it will need to be estimated before estimating the parameter of interest.
We define the Experimental Treatment Assignment (ETA) assumption for estimating $\beta_{m,\lambda}$ as follows:

$$\max_{\bar{a} \in \mathcal{A}} \frac{h_\lambda(\bar{a}, V)}{g(\bar{a} \mid X)} < \infty \; F_X - a.e.$$

We denote the estimator of λ and g with λ_n and g_n respectively. If λ is known, we define $\lambda_n = \lambda$.

The IPTW estimator of $\beta_{m,\lambda}$ is defined as the solution of the estimating equation associated with the observed data O and the IPTW estimating function at g_n and λ_n:

$$\sum_{i=1}^{n} D_{h_{\lambda_n}}(o_i \mid g_n, \beta) = 0 \tag{8}$$

3.2.2 Property of the IPTW estimating function

Under the ETA assumption, the IPTW estimating function is unbiased at $\beta_{m,\lambda}$ under $P_{F_X,g}$:

$$E_{P_{F_X,g}}\left[D_{h_\lambda}(O \mid g, \beta_{m,\lambda})\right] = 0.$$

Proof:

$$\begin{aligned}
E_{P_{F_X,g}}\left[D_{h_\lambda}(O \mid g, \beta_{m,\lambda})\right] &= EE\left(\frac{h_\lambda(\bar{A}, V)\epsilon(\beta_1^\lambda)}{g(\bar{A} \mid X)} \mid X\right) \\
&= E_{F_X}\left(\sum_{\bar{a}: g(\bar{a}|X) \neq 0} \frac{h_\lambda(\bar{a}, V)\epsilon_{\bar{a}}(\beta_{m,\lambda})}{g(\bar{a} \mid X)} g(\bar{a} \mid X)\right) \\
&\overset{\text{ETA}}{=} E_{F_X}\left(\sum_{\bar{a}} h_\lambda(\bar{a}, V)\epsilon_{\bar{a}}(\beta_{m,\lambda})\right) \\
&= E_{F_V} E\left(\sum_{\bar{a}} h_\lambda(\bar{a}, V)(Y_{\bar{a}} - m(\bar{a}, V \mid \beta_{m,\lambda})) \mid V\right)
\end{aligned}$$

$$= E_{F_V}\left(\sum_{\bar{a}} h_\lambda(\bar{a}, V)(m_0^*(\bar{a}, V) - m(\bar{a}, V \mid \beta_{m,\lambda}))\right)$$
$$\stackrel{(7)}{=} 0.\square$$

3.2.3 Practical implications for the IPTW estimator and implementation

The previous result justifies the use of the IPTW estimator to estimate $\beta_{m,\lambda}$ consistently with the observed data when the ETA assumption holds. In practice, a corollary of the unbiasedness of the IPTW estimating function is that, under regularity conditions, the IPTW estimator of $\beta_{m,\lambda}$ is consistent and asymptotically linear if the ETA assumption holds and if g_n and λ_n are consistent estimators of g and λ respectively.

The IPTW estimate of $\beta_{m,\lambda}$ can be obtained in practice by performing a weighted least squares regression of Y on $\bar{A}$ and V using the CM and the following weights for each observation: $w(\bar{A}, V) = \frac{\lambda_n(\bar{A},V)}{g_n(\bar{A}|X)}$. It can indeed be shown that the resulting estimate is a solution of the estimating equation (8). Moreover, in order to obtain a consistent estimator of $\beta_{m,\lambda}$, λ_n should be a consistent estimator of λ. In addition, it is known, see van der Laan and Robins (2002), that one should always specify g_n as a consistent estimator of g even when g is known. As a result, the IPTW estimator may gain in efficiency by possibly capturing empirical confounding, without loss of consistency.

Note that the ETA assumption is a critical assumption for the consistency of the IPTW estimator. Since the choice for h_λ is dictated by the parameter of interest definition, the user cannot arrange for the ETA assumption to be holding by choosing an appropriate h_λ. That is why the ETA assumption is equivalent in practice to a stronger assumption that we designate as the Strong ETA (SETA) assumption:

$$\min_{\bar{a}\in\mathcal{A}} g(\bar{a} \mid X) > 0 \quad F_X - a.e.$$

Note that the SETA assumption implies the ETA assumption. Under the SRA, the SETA assumption is equivalent to:

$$\forall\, j \in \{0, \ldots, k\} \quad \forall\, \bar{a}(j-1) \in \mathcal{A}(j-1) \quad \forall\, a(j) \in \mathcal{A}_{\bar{a}(j-1)}(j)$$
$$g(A(j) = a(j) \mid \bar{A}(j-1) = \bar{a}(j-1), \bar{L}(j)) > 0 \quad P_{F_X,g} - a.e.,$$

where $\mathcal{A}(j-1) = \{\bar{a}^*(j-1) : \bar{a}^* \in \mathcal{A}\}$ and $\mathcal{A}_{\bar{a}(j-1)}(j) = \{a^*(j) : \bar{a}^* \in \mathcal{A},\ \bar{a}^*(j-1) = \bar{a}(j-1)\}$.

3.3 G-computation estimator

3.3.1 Definition

Under the SRA, we have seen that the likelihood of the observed data factorizes into two parts. We define the G-computation formula as the F_X part of the likelihood where $\bar{A}$ is set to $\bar{a}$ for every $\bar{a} \in \mathcal{A}$ and we denote it with $f_{\bar{a}}^{Q_{F_X}}(\bar{L}(k+1))$:

$$f_{\bar{a}}^{Q_{F_X}}(\bar{L}(k+1)) \equiv \prod_{j=0}^{k+1} f(L(j) \mid \bar{L}(j-1), \bar{A}(j-1) = \bar{a}(j-1)).$$

Note that this product, i.e. the G-computation formula, is only defined under $P_{F_X,g}$ for every $\bar{a} \in \mathcal{A}$ if the SETA assumption holds. We define $f_{\bar{a}}^{Q_n}(\bar{L}(k+1))$ for every $\bar{a} \in \mathcal{A}$ using the G-computation formula in which Q_{F_X} is replaced by its estimator, Q_n.

The G-computation estimator is defined as the solution of the approximation of $E_{Q_n} D_{h_{\lambda_n}}(X \mid \beta) = 0$ where E_{Q_n} is approximated by an empirical mean over draws $\hat{X}$ from Q_n using $f_{\bar{a}}^{Q_n}(\bar{L}(k+1))$ and $L(0)$:

$$\sum_{j=1}^{p} D_{h_{\lambda_n}}(\hat{X}_j \mid \beta) = 0, \tag{9}$$

where p is a large number of draws $\hat{X}_j$ from Q_n.

3.3.2 Property of the G-computation formula

Under the SETA assumption, for every $\bar{a} \in \mathcal{A}$ the G-computation formula calculated at $\bar{a}$ is defined and identifies the marginal probabilities or densities of $\bar{L}_{\bar{a}}(k+1)$ defined under F_X:

$$f_{\bar{a}}^{Q_{F_X}}(\bar{L}(k+1)) = f(\bar{L}_{\bar{a}}(k+1)).$$

See Robins (1997), Robins (1998b), Gill and Robins (2001) and Yu and van der Laan (2002a) for the proof.

This result establishes a link between the observed data distribution, $P_{F_X,g}$, and the full data distribution, F_X. More precisely it establishes a link between Q_{F_X} and the marginal distribution of the counterfactual process. This last distribution is sufficient to identify the expectation under F_X of $D_{h_\lambda}(X \mid \beta)$ and that is why the G-computation formula can be used to simulate full data, $\hat{X}$, in order to estimate parameters of m^* like $\beta_{m,\lambda}$ using the full data estimating function $D_{h_\lambda}(X \mid \beta)$.

A corollary of this result important in practice is:

$$f(Y_{\bar{a}} = l(k+1) \mid L(0) = l(0)) = \sum_{l(1)} \ldots \sum_{l(k)} \prod_{j=1}^{k+1} f(l(j) | \bar{l}(j-1), \bar{a}(j-1)). \tag{10}$$

3.3.3 Practical implications for the G-computation estimator and implementation

The previous result and equivalence (6) justify the use of the G-computation estimator to estimate $\beta_{m,\lambda}$ consistently using the observed data when the SETA assumption holds. In practice, a corollary of equivalence (6) and the previous result concerning the G-computation formula is that, under regularity conditions, the G-computation estimator of $\beta_{m,\lambda}$ is consistent and asymptotically linear if the SETA assumption holds and if Q_n and λ_n are consistent estimators of Q_{F_X} and λ respectively.

The G-computation estimate can be obtained in practice by carrying out the following two steps:

1. **Generate the simulated data $(\mathbf{Y}_{\bar{\mathbf{a}}})_{\bar{\mathbf{a}}\in\mathcal{A}}$.** We can use a Monte Carlo simulation procedure based on the user-specified distributions in Q_n, f_n, and the observed baseline covariates, $L(0)$, to simulate $(Y_{\bar{a}})_{\bar{a}\in\mathcal{A}}$. For every $\bar{a} \in \mathcal{A}$ and every individual $i = 1, \ldots, n$: starting at $j = 1$, keep simulating $l_i(j)$ using $f_n(L(j) \mid \bar{L}(j-1), \bar{A}(j-1))$, $\bar{a}_i(j-1)$ and $\bar{l}_i(j-1)$ until $j = k+1$. By doing so, equation (10) ensures that the simulated data $l_i(k+1)$ are generated using the conditional distribution of $Y_{\bar{a}}$ given $L(0) = l(0)$ defined under Q_n.

2. **Solve for β equation (9) using the simulated full data $(\mathbf{Y}_{\bar{\mathbf{a}}})_{\bar{\mathbf{a}}\in\mathcal{A}}$ and the observed data V.** This can be done by performing a weighted least square regression of $Y_{\bar{a}} = L(k+1)$ on $\bar{a}$ and V using the data obtained by pulling together all observations $(v_i, \bar{a}, l_i(k+1))$ obtained under every treatment regimen $\bar{a}$, treating these observations as independent and using $\lambda_n(\bar{a}, V)$ as weights.

In order to obtain a consistent estimator of $\beta_{m,\lambda}$ in practice, λ_n and Q_n should be consistent estimators of λ and Q_{F_X}. The observed data are used to estimate the parameters of the models for Q_{F_X} using the method of maximum likelihood. Indeed under the SRA, the likelihood is a product of the distributions in Q_{F_X}, $f(L(0))$ and the g part of the likelihood. Maximizing a product of functions over independent parameters is the same as maximizing each element of the product separately over its corresponding parameters. Therefore, Q_{F_X} can be estimated using the method of maximum likelihood but only using the Q_{F_X} part of the likelihood instead of the complete likelihood formula, $\mathcal{L}(0)$. Note that we do not need to estimate the g part of the likelihood in this approach.

The G-computation estimator as defined in this paper had not been previously tested using real or simulated longitudinal data. In section 4, we report the practical performances of this estimator which has now been implemented successfully using the procedure presented above for a simulation study.
Note however that the G-computation estimator is of interest in the nonparametric

MSM approach to causal inference not as an estimator of the parameter of interest but as a building block for the DR estimator. Indeed, in this approach, the estimators of interest do not entirely rely, even indirectly, on correct specification of a parametric MSM, i.e. assumptions on F_X.

3.4 DR estimator

3.4.1 Definition

The unique DR estimating function for $\beta_{m,\lambda}$ with nuisance parameters g and Q_{F_X} is defined as:

$$D_{h_\lambda}(O \mid g, Q_{F_X}, \beta) = \frac{h_\lambda(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)} - \sum_{j=0}^{k} \left[E_{Q_{F_X},g} \left(\frac{h_\lambda(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)} \mid \bar{A}(j), \bar{L}(j) \right) - E_{Q_{F_X},g} \left(\frac{h_\lambda(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)} \mid \bar{A}(j-1), \bar{L}(j) \right) \right],$$

Note that this estimating function does indeed depend on $P_{F_X,g}$ through (Q_{F_X}, g).

The DR estimator of $\beta_{m,\lambda}$ is defined as the solution of the estimating equation associated with the observed data O and the DR estimating function at g_n, Q_n and λ_n:

$$\sum_{i=1}^{n} D_{h_{\lambda_n}}(o_i \mid g_n, Q_n, \beta) = 0 \tag{11}$$

3.4.2 Property of the DR estimating function

Under the ETA assumption for g_1, we have

$$E_{P_{F_X,g}} \left[D_{h_\lambda}(O \mid g_1, Q_1, \beta_{m,\lambda}) \right] = 0 \text{ if } \begin{cases} g_1 = g \text{ or} \\ Q_1 = Q_{F_X} \end{cases}$$

This result is a corollary of the results in section 3.2.2 and lemma 1.9 in van der Laan and Robins (2002) applied to the IPTW estimating function for estimating $\beta_{m,\lambda}$ as defined in section 3.2.1.

3.4.3 Practical implications for the DR estimator and implementation

The previous result justifies the use of the the DR estimator to estimate $\beta_{m,\lambda}$ consistently with the observed data under the conditions for which either the IPTW or G-computation estimator is consistently estimating $\beta_{m,\lambda}$. In practice, a corollary of the previous result is that, under regularity conditions, the DR estimator of $\beta_{m,\lambda}$ is consistent and asymptotically linear if λ_n is a consistent estimator of λ and if either the ETA assumption holds for g and g_n is a consistent estimator of g, or if Q_n is a consistent estimator of Q_{F_X} and the DR estimating function is

defined at a g_1 for which the ETA assumption holds.

The DR estimator can be obtained in practice by directly solving for β equation (11) using the Newton-Raphson algorithm. This algorithm can indeed be used to minimize complex multivariate functions. It consists of an iterative procedure starting with an initial estimate of $\beta_{m,\lambda}$ that is iteratively updated until a user-specified convergence criteria is reached. The latest update of the estimate of $\beta_{m,\lambda}$ corresponds with the DR estimate. More details on how to use the Newton-Raphson algorithm in our problem can be found in van der Laan and Robins (2002). Another approach to solve the DR estimating equations is described in Robins (2000).
Prior to solving equation (11), g_n and Q_n need to be specified similar to the IPTW and G-computation approach respectively, i.e. in order to obtain a consistent DR estimator of $\beta_{m,\lambda}$, the estimators λ_n, g_n and Q_n should consistently estimate λ, g and Q_{F_X} respectively. Similar to the G-computation approach, Q_n needs to be specified and used to estimate the expectations defining $D_{\lambda_n}(O \mid g_n, Q_n, \beta)$ using Monte Carlo simulations, Yu and van der Laan (2002b). The conditional densities or probabilities in Q_n that need to be specified correspond exactly with the conditional densities or probabilities that need to be modelled for the G-computation estimator. This last observation and the result presented in the previous section suggest the following DR estimation procedure when the ETA assumption holds for g:

1. Estimate the treatment mechanism g with g_n and Q_{F_X} with Q_n.

2. Compute once and for all the expectations defining $D_{\lambda_n}(O \mid g_n, Q_n, \beta)$ using Monte Carlo simulations and the G-computation estimate, both relying on Q_n.

3. Use the Newton-Raphson algorithm to solve the DR estimating function in which the expectations are known from the previous step and remain unchanged throughout the algorithm procedure. The initial value for β is chosen as the G-computation estimate relying on Q_n.

In such a procedure, the expectations defining $D_{\lambda_n}(O \mid g_n, Q_n, \beta)$ do not need to be computed at each step of the iterative Newton-Raphson algorithm. Thus, the DR estimate is obtained for a low computing cost without losing its double robustness property. Solving the estimating equations associated with $D_{\lambda_n}(O \mid g_n, Q_n, \beta)$ is thus very fast in practice after the G-computation estimate is obtained and the expectations defining the DR estimator are computed. In section 4, we report the practical performances of the DR estimator which has been implemented successfully for the first time in a non-survival longitudinal data simulation study using the procedure described previously.

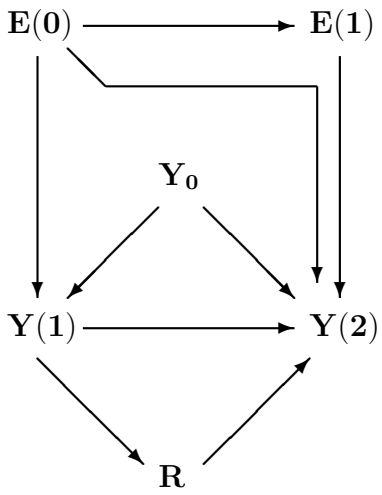

Figure 1: DAG of the simplified data structure considered in this paper

4 Illustration with a simulation study

4.1 Simulation overview

To illustrate the results presented in this paper, we perform a simulation study mimicking a real epidemiological study in which non-survival longitudinal data are collected to investigate the effect of air pollutant exposure on an asthma symptom, wheeze. This simulation study is inspired by the Fresno Asthmatic Children and Environmental Study (FACES).

In this simulation, we limit ourselves to a non-survival longitudinal data structure with only a few time points mimicking a simplified version of the data structure obtained in FACES during two-week monitoring periods. In this simplified data structure, the outcome of interest is $Y(2)$, wheezing indicator in the morning. We denote the indicator of rescue medication use the night prior to outcome report with R. As suggested by initial analyses using real data from FACES, the effect of rescue medication use on morning symptoms is confounded by prevalence of wheezes, $Y(1)$, reported the evening prior to outcome report. The exposure levels of a given pollutant (e.g $PM_{2.5}$) are measured on the two days prior to outcome report using appropriate metrics: $E(0)$ and $E(1)$. The observed data structure O can thus be ordered as follows:

$$O = (E(0), Y(1), E(1), R, Y(2)).$$

The data can also be represented by a diagram in which the assumed causal relationships between the variables can be represented by directed arrows (Directed Acyclic Graph or DAG, see figure 1). In this diagram, we can include unobserved covariates like Y_0 (e.g. asthma severity) which are assumed to have causal effects on observed variables.

We wish to use the simulated observed longitudinal data collected on a population of asthmatic children to investigate the joint causal effect of rescue medication use and a two-day $PM_{2.5}$ exposure on the prevalence of morning wheezes. In other words, the counterfactuals of interests are: $Y_{e_0,e_1,r}$, i.e. the outcome $Y(2)$ under the treatment history $(E(0) = e_0, E(1) = e_1, R = r)$.
Because of the simulation procedure used, the true causal relationship of interest: $m^*(e_0, e_1, r) \equiv E(Y_{e_0,e_1,r})$ is unknown as in most real life studies. We use a nonparametric MSM causal approach to causal inference to investigate the causal curve of interest, m^*, and we define four parameters of interest using two CKS: $\lambda_1 = \lambda_{var}^{loc}$ and $\lambda_2 = \lambda_{var,g}^{loc}$ and the following two CM:

- $m_1(e_0, e_1, r \mid \beta) = \dfrac{1}{1 + exp\left[-(\beta_0 + \beta_1\, ma_2 + \beta_2\, r)\right]}$, where ma_2 designates a two-day moving average of $PM_{2.5}$ levels: $ma_2 = \dfrac{e_0 + e_1}{2}$,
- $m_2(e_0, e_1, r \mid \beta) = \dfrac{1}{1 + exp\left[-(\beta_0 + \beta_1\, ma_2 + \beta_2\, r + \beta_3\, ma_2\, r)\right]}$.

The corresponding four parameters of interest: β_{m_1,λ_1}, β_{m_1,λ_2}, β_{m_2,λ_1} and β_{m_2,λ_2} are estimated using the IPTW, G-computation and DR estimators defined in section 3.

We repeat this simulation procedure N_r times, i.e. N_r observed data sets are simulated. We report the mean and standard deviation of the four parameter estimates obtained for each simulated data.

4.2 Simulation procedure

4.2.1 Data simulation

The diagram in figure 1 is used to generate every data set in this simulation study according to known data generating distributions and the following protocol in which every parameter is known and user-specified before simulating data:

1. Generate Y_0 using a Bernoulli distribution with a user-specified probability of event: $P(Y_0 = 1) = p^{Y_0}$. This variable is binary and represents the asthma severity status of the subject: $Y_0 = 1$ indicates "severe asthma".

2. Generate $E(0)$ such that $P(E(0) = e) = p_e^{E_0}$ which value is user-specified for $e = 1, \ldots, 5$. This variable is discrete and represents the air pollutant level on the first day of the study: the larger e, the higher the pollution level.

3. Generate $Y(1)$ using the value of Y_0 and $E(0)$: if $Y_0 = 1$ then $Y(1) = 1$ else $P(Y(1) = 1 \mid Y_0 = 0, E(0) = e) = p_e^{Y(1)}$ which value is user-specified for $e = 1, \ldots, 5$. This variable is binary and represents the prevalence of wheeze reported the evening of the second day of the study: $Y(1) = 1$ indicates asthma symptom prevalence.

4. Generate $E(1)$ such that $P(E(1) = e_1 \mid E(0) = e_0) = p^{E_1}_{e_1,e_0}$ which value is user-specified for e_0 and e_1 in $\{1, \ldots, 5\}$. This variable is discrete and represents the air pollutant level the second day of the study: the larger e_1, the higher the pollution level.

5. Generate R using the value of $Y(1)$: $P(R = 1 \mid Y(1) = i) = p^R_i$ which value is user-specified for $i = 0$ and $i = 1$. This variable is binary and represents the use of rescue medication the night prior to outcome report (second day of the study): $R = 1$ indicates rescue medication use.

6. Generate $Y(2)$ using the value of Y_0, $E(0)$, $E(1)$, $Y(1)$ and R: if $Y_0 = 1$ then $Y(2) = 1$ else:

$$P(Y(2) = 1 \mid Y_0 = 0, E(0) = e_0, E(1) = e_1, Y(1) = y(1), R = r) = \frac{1}{1 + exp\left[-(\alpha_0 + \alpha_1\, y(1) + \alpha_2\, ma_2 + \alpha_3\, r + \alpha_4\, y(1)\, ma_2 + \alpha_5\, r\, ma_2)\right]},$$

where the value of $\alpha = (\alpha_0, \ldots, \alpha_5)$ is user-specified. This variable is binary and represents the outcome in the morning of the third day of the study: $Y(2) = 1$ indicates asthma symptom prevalence.

To generate a data set with n observations, we repeat the previous protocol n times and thus obtain n independent and identically distributed observations $o_i = (e_i(0), y_i(1), e_i(1), r_i, y_i(2))$ of $O = (E(0), Y(1), E(1), R, Y(2))$. Note that Y_0 is not part of the observed data but needs to be simulated to obtain the rest of the data.

4.2.2 Values of the parameters of interest and their IPTW, G-computation and DR estimates

In this simulation study, the data generating distributions in Q_{F_X} as defined in the previous sections are known and can be used to simulate a full data set X containing a large number of observations (e.g. N=10000). The values of the four parameters of interest can be calculated using such data by solving the estimating equation associated with the full data estimating function, $D_{h_\lambda}(X \mid \beta)$, where $\lambda = \lambda_1$ or $\lambda = \lambda_2$ and $m = m_1$ or $m = m_2$. According to equivalence (6), $D_{h_\lambda}(X \mid \beta)$ will provide four consistent estimators of what we defined earlier as the four parameters of interest. Because N is very large, we will consider the obtained four estimates as the true values for the four parameters of interest. If for a given CM, the parameters of interest defined using two different CKS are different then we can conclude that the corresponding CM is misspecified and we have two different parameters of interest defined using the same CM but different CKS. There is only one parameter of interest if the CM is correctly specified whatever the choice for the CKS, i.e. the two parameters of interest reported for a given CM using two different CKS will be very close if the CM is correctly specified.

The IPTW, G-computation and DR estimates of the four parameters of interest are obtained as described in section 3. Models are used to consistently estimate λ, g and Q_{F_X} using the simulated observed data, i.e. these models are always correctly specified.

4.3 Simulation results

The user-specified parameters of the simulation procedure are first set as follows:

- $N_r = 100$
- $n = 1000$ and $N = 10000$
- $p^{Y_0} = 0.8$
- $p_i^{E(0)} = 0.2$ for $i = 1, \ldots, 5$
- $p_e^{Y(1)}$ value for $e = 1, \ldots, 5$:

e	1	2	3	4	5
$p_e^{Y(1)}$	0.2	0.4	0.5	0.6	0.8

- $p_{e_1,e_0}^{E(1)}$ value for $e_0 = 1, \ldots, 5$ and $e_1 = 1, \ldots, 5$:

		e_0				
		1	2	3	4	5
e_1	1	0.4	0.3	0.1	0.1	0.1
	2	0.2	0.4	0.2	0.1	0.1
	3	0.1	0.2	0.4	0.2	0.1
	4	0.1	0.1	0.2	0.4	0.2
	5	0.1	0.1	0.1	0.3	0.4

- p_i^R for $i = 0, 1$:

i	0	1
p_i^R	0.3	0.7

- α_i for $i = 0, \ldots, 5$:

α_0	α_1	α_2	α_3	α_4	α_5
-3	2	0.1	-0.1	1	-2

The corresponding simulation results are given in table 1. A second simulation study was performed using the same parameters except for p_i^R. The new values for p_i^R were set to:

i	0	1
p_i^R	0.01	0.99

In this last simulation study, the ETA assumption is practically violated. The corresponding simulation results are given in table 2.

		β_0	β_1	β_2	β_3
CM 1	β_{m_1,λ_1}	1.78	0.13	-0.75	NA
	β_{m_1,λ_2}	1.90	0.08	-0.73	NA
	IPTW 1	1.79 [0.25]	0.13 [0.07]	-0.75 [0.10]	NA
	IPTW 2	1.92 [0.26]	0.08 [0.08]	-0.74 [0.10]	NA
	G-comp 1	1.82 [0.22]	0.12 [0.06]	-0.75 [0.10]	NA
	G-comp 2	1.93 [0.22]	0.08 [0.06]	-0.73 [0.10]	NA
	DR 1	1.80 [0.24]	0.13 [0.07]	-0.75 [0.10]	NA
	DR 2	1.89 [0.24]	0.08 [0.07]	-0.72 [0.10]	NA
CM 2	β_{m_2,λ_1}	1.14	0.36	0.28	-0.36
	β_{m_2,λ_2}	1.15	0.36	0.26	-0.36
	IPTW 1	1.18 [0.32]	0.35 [0.11]	0.23 [0.43]	-0.35 [0.15]
	IPTW 2	1.19 [0.33]	0.35 [0.11]	0.23 [0.44]	-0.35 [0.16]
	G-comp 1	1.19 [0.22]	0.34 [0.07]	0.26 [0.24]	-0.35 [0.08]
	G-comp 2	1.18 [0.23]	0.35 [0.08]	0.25 [0.25]	-0.35 [0.09]
	DR 1	1.17 [0.25]	0.35 [0.08]	0.26 [0.25]	-0.35 [0.09]
	DR 2	1.15 [0.26]	0.36 [0.09]	0.28 [0.26]	-0.36 [0.09]

Table 1: Simulation results when the ETA assumption is holding. NA stands for "Not Applicable". The values for the four parameters of interest are obtained using a simulated full data set X of $N = 10000$ observations. The values reported for the parameter of interest estimates are averages of parameter estimates obtained on $N_r = 100$ simulated data sets of $n = 1000$ observations each. The empirical standard deviations for the parameter of interest estimates are presented between squared brackets. The ones and twos after IPTW, G-comp and DR denotes the use of λ_1 and λ_2 respectively.

4.4 Result interpretation

From the results in table 1 and 2, we can conclude that CM 1 is misspecified since we clearly have $\beta_{m_1,\lambda_1} \neq \beta_{m_1,\lambda_2}$. In table 2, when choosing λ_1 instead of λ_2, the parameter of interest defined using CM 1 becomes 0.01 versus 0.12, i.e. the non-parametric causal effect of $PM_{2.5}$ on wheeze becomes almost null when one tries to approximate m^* using CM 1 over the regions (e_0, e_1, r) that are the most commonly observed versus over all regions (e_0, e_1, r). These results illustrate the impact of the choice for λ on the parameter of interest definition. Misspecification of CM 2 is not obvious since the values for β_{m_2,λ_1} and β_{m_2,λ_2} are very similar in both tables.

In addition, this simulation also illustrates the importance of the ETA assumption when using the IPTW estimator. In table 2, the IPTW estimator of the parameters of interest are clearly biased.

		β_0	β_1	β_2	β_3
CM 1	β_{m_1,λ_1}	1.79	0.12	-0.75	NA
	β_{m_1,λ_2}	1.96	0.01	-0.60	NA
	IPTW 1	1.90 [0.90]	0.09 [0.26]	-0.46 [0.44]	NA
	IPTW 2	2.31 [0.95]	-0.09 [0.29]	-0.31 [0.52]	NA
	G-comp 1	1.84 [0.52]	0.12 [0.11]	-0.78 [0.28]	NA
	G-comp 2	2.02 [0.39]	0.00 [0.07]	-0.62 [0.35]	NA
	DR 1	1.85 [0.57]	0.12 [0.12]	-0.79 [0.29]	NA
	DR 2	1.90 [0.42]	0.00 [0.08]	-0.50 [0.36]	NA
CM 2	β_{m_2,λ_1}	1.15	0.35	0.28	-0.36
	β_{m_2,λ_2}	1.22	0.31	0.20	-0.32
	IPTW 1	0.71 [1.77]	0.49 [0.57]	1.47 [2.18]	-0.63 [0.67]
	IPTW 2	0.56 [2.08]	0.52 [0.67]	1.60 [2.46]	-0.66 [0.77]
	G-comp 1	1.15 [0.88]	0.37 [0.27]	0.33 [1.01]	-0.39 [0.30]
	G-comp 2	1.18 [1.04]	0.34 [0.33]	0.30 [1.19]	-0.37 [0.37]
	DR 1	1.18 [0.94]	0.36 [0.29]	0.30 [1.03]	-0.39 [0.31]
	DR 2	1.12 [1.09]	0.31 [0.35]	0.35 [1.21]	-0.33 [0.38]

Table 2: Simulation results when the ETA assumption is practically violated. NA stands for "Not Applicable". The values for the four parameters of interest are obtained using a simulated full data set X of $N = 10000$ observations. The values reported for the parameter of interest estimates are averages of parameter estimates obtained on $N_r = 100$ simulated data sets of $n = 1000$ observations each. The empirical standard deviations for the parameter of interest estimates are presented between squared brackets. The ones and twos after IPTW, G-comp and DR denotes the use of λ_1 and λ_2 respectively.

5 Discussion

The proposed nonparametric MSM approach to causal inference should be compared to the typical parametric MSM approach found in the literature.

Both approaches rely on the same statistical framework as introduced in section 1. However the parametric MSM approach to causal inference relies on an additional assumption on the full data distribution F_X by assuming a parametric MSM, $m(\cdot,\cdot \mid \beta)$, i.e. a model for the causal curve of interest, m^*. The parameter of interest is thus defined as β_m such that: $m^*(\bar{a}, V) = m(\bar{a}, V \mid \beta_m)$ and its definition entirely relies on correctness of the assumed MSM. The model for F_X is not saturated when assuming a parametric MSM and as a consequence there exists a class of full data estimating functions. It follows that there exists a class of IPTW, G-computation and DR estimators. All these classes of estimators are defined using the same definitions used for the nonparametric MSM approach to causal inference introduced in this paper but extended to any choice for h instead

of constraining h to be equal to h_λ. For instance, the class of full data estimating function for estimating β_m is defined as follows:

$$\{D_h(X \mid \beta) = \sum_{\bar{a} \in \mathcal{A}} h(\bar{a}, V)\varepsilon_{\bar{a}}(\beta) : \text{ for any } h : \mathcal{A} \times S_V \longrightarrow I\!R\},$$

Therefore when using a parametric MSM, the IPTW, G-computation and DR estimators of β_m can be defined using any function h. In previous work by Robins, Hernán and Brumback (2000), the following choice for h was proposed to estimate β_m when the outcome is continuous:

$$h(\bar{A}, V) = g(\bar{A} \mid V)\frac{\partial}{\partial \beta} m(\bar{A}, V \mid \beta), \tag{12}$$

This is a choice equivalent to h_λ where $\lambda = \lambda_g^{loc}$ as defined in section 2.3. Similarly when the outcome is binary and the MSM is a logistic function, the recommended choice for h is h_λ where $\lambda = \lambda_{var,g}^{loc}$. These choices are motivated by the following observations:

- Using the IPTW estimating function corresponding with these choices for h and when the effect of $\bar{A}$ on Y is not confounded beyond V, becomes equivalent to performing a weighted regression where every observation is given the same weight of 1. Thus, this IPTW estimator is the least squares adjusted regression estimator commonly used when there is no confounder beyond V.

- In addition, the weights associated with these choices for h are more stable than the simple inverse of treatment mechanism probability weights corresponding with $h = h_{\lambda^{uni}}$, i.e. the corresponding IPTW estimator is also more efficient.

The nonparametric MSM approach to causal inference can thus be viewed as a more general tool for causal inference than the parametric MSM approach. It is not only an extension of the causal objects introduced by the parametric MSM approach but also an explicit extension of the causal questions that can be answered.

Indeed, the nonparametric MSM approach requires fewer assumptions as no MSM is assumed, i.e. the MSM is left nonparametric. In addition, the parameters of interest, $\beta_{m,\lambda}$, generalizes the unique parameter of interest, β_m, since $\beta_{m,\lambda} = \beta_m$ for all λ when $m(\cdot, \cdot \mid \beta)$ is correctly specified. The new parameters of interest extend the definition of β_m as they do not require the assumption of a MSM but only rely on its analog the CM and the CKS, both user-specified. Such extended parameters of interest can thus provide the answer to the typical aim of the parametric MSM causal analyses which is to provide a global and accurate representation of the causal curve using the CM and $\beta_{m,\lambda} = \beta_m$ when the CM is correctly specified. Furthermore, these parameters can now also explicitly answer less ambitious causal

questions of interest by providing an informative, possibly localized, representation of the causal curve using both an educated choice for the CM and CKS when the CM is misspecified willingly or not. Using this nonparametric MSM approach to causal inference, the CM can now be willingly misspecified in order to describe a complex causal curve by a summary measure, e.g. assuming a linear model when the true causal relationship is linear followed by a plateau. Such concepts are similar to the one introduced by misspecified models in association analysis.

The nonparametric MSM approach to causal inference also provides a better understanding of MSM estimation when the assumed MSM is misspecified. The parametric MSM approach to causal inference relies on the critical assumption of correctness of the MSM. When the MSM is however misspecified, the parameter of interest, β_m is not defined and it is fundamental to understand what are the behaviors of the IPTW, G-computation and DR estimators in such a scenario. It is important to understand if the estimates obtained are still of interest and how to causally interpret them if they are of interest.

If h is chosen equal to $h_{\lambda_g^{loc}}$ or $h_{\lambda_{var,g}^{loc}}$ according to the recommendation given by Robins, Hernán and Brumback (2000) then the IPTW, G-computation and DR estimator of β_m are implicitly the corresponding estimators of $\beta_{m,\lambda_g^{loc}}$ or $\beta_{m,\lambda_{var,g}^{loc}}$ respectively when the MSM is misspecified. As a result, the estimates obtained can be potentially of interest, although they are not guaranteed to be estimates of the more appropriate parameter to answer the causal question of interest when the MSM is misspecified: for instance, one might have preferred to define the parameter of interest using another choice for λ like λ^{uni} which would provide a global summary representation of the causal curve. If h is different from these previous recommended choices but are still of the type h_λ then the estimates are still of potential interest as estimates of the causal parameter $\beta_{m,\lambda}$. For any other choice for h, the behaviors of the IPTW, G-computation and DR estimators of β_m when the MSM is misspecified are unknown but it is likely that the obtained estimates will be of limited interest to investigate the causal curve of interest.

From the observations above, we can argue that the nonparametric MSM approach to causal inference developed in this paper is particularly suitable for investigating causal effects in practice.

Such a statement relies on the belief that correct MSM misspecification is very unlikely in practice. If admitted, this statement also reduces the importance of searching for the locally efficient estimator of β_m as one does not assume the CM to be correctly specified and therefore the unique locally efficient estimator of $\beta_{m,\lambda}$ is the known DR estimator. The bases for favoring a nonparametric MSM approach to causal inference are:

1. it extends the possibility of causal effect investigation even when the MSM are misspecified by defining new parameters of interest if the CM is misspecified. These new parameters of interest are less ambitious and thus appealing in

practice as they do not try to extract the global and accurate representation of the causal curve from a limited amount of data.

2. it allows the CM to be willingly misspecified in an effort to provide an informative summary, possibly localized, representation of the causal curve.

It could be argued that the parametric MSM approach to causal inference can provide implicitly the same type of causal effect investigation even when the MSM is misspecified as long as h is chosen of the type h_λ and λ is given the same estimation effort as g and Q_{F_X}. However, such an implicit extension of the parametric MSM approach is dangerous and confusing as this approach theoretically allows any choice for h and even if h is chosen equal to h_λ, the effort to model λ has previously been recommended to be minimal according to theoretical considerations. In addition, the assumed MSM cannot be willingly misspecified in theory. We might thus favor the nonparametric MSM approach to causal inference which provides an explicit control on the causal summary measure to be estimated.

In practice a correct MSM for the causal curve of interest is typically unknown. Therefore if one wishes to obtain a global and accurate representation of the causal curve, it would be natural to assume a nonparametric MSM to avoid model misspecification and to define the parameter of interest as $\beta_m = \beta_{m,\lambda}$ for every possible λ, where m is a nonparametric MSM. Estimators of such a parameter of interest, however, would most likely suffer from the curse of dimensionality (See van der Laan and Robins (2002)). Thus one typically needs in practice to assume a parametric MSM. One can use previous subject-matter knowledge or model selection methods to choose the assumed MSM. In both cases, the risk of model misspecification remains important. Selecting a MSM can be done using a cross-validation methodology, Brookhart and van der Laan (2003) and van der Laan and Dudoit (2003), however the true causal relationship m^* can be very complex and the observed data might not provide enough information (curse of dimensionality) to select a correct MSM. This is assuming that one correct MSM model is even considered in the MSM selection process. That is why we believe that correct parametric MSM misspecification is very unlikely in most practical applications and that is why the nonparametric MSM to causal inference is very appealing to most real life applications.

If one knows or believes one can select a correctly specified parametric MSM, the consequence of using the nonparametric MSM approach to causal inference instead of the parametric MSM approach could only be the loss in efficiency estimation.

To finish this discussion, we would like to address two issues concerning the G-computation estimator of $\beta_{m,\lambda}$. First, as underlined throughout this paper, the G-computation estimator is of interest in the nonparametric MSM approach to causal inference not as an estimator of the parameter of interest, but as a building

block for the DR estimator. Indeed in this approach, we wish that causal inferences be as nonparametric as possible and that they should, minimally, not rely entirely on correct specification of a parametric MSM as it is often unknown in practice. Because of the curse of dimensionality, a truly nonparametric approach to causal inference is typically not possible in practice. However, a nonparametric MSM approach to causal inference can be developed such that causal effect investigation does not entirely rely on prior assumptions on the causal curve. In such an approach, estimators which only rely on assumptions on the causal curve are thus not of interest. The G-computation estimator entirely relies on correct specification of Q_{F_X}, i.e. indirectly on a model for the causal curve. That is why this estimator is only of interest in this approach as a building block for the DR estimator which does not entirely rely on prior assumption on the causal curve. The DR estimator does, however, use such additional assumptions to make the DR estimator truly more nonparametric than the IPTW estimator.

Secondly, we would like to point out the importance of the SETA assumption when using the G-computation estimator to estimate $\beta_{m,\lambda}$ in practice. In theory however, the SETA assumption does not necessarily need to hold for the G-computation estimator to be consistent. However, even though not necessary for consistency, the SETA assumption asymptotically ensures that the G-computation estimator does not rely on guessed models or model extensions that will most likely be incorrect. That is why we included the SETA assumption in the results presented in section 3.3 as it is practically necessary.

If the SETA assumption is violated, there is at least one conditional density or probability defining the G-computation formula, $f_{\bar{a}}^{Q_{F_X}}(\bar{L}(k+1))$, that is not identifiable with the observed data, i.e. not defined under $P_{F_X^0,g_0}$ but only under $P_{F_X^0,g_1}$ where the SETA assumption holds for g_1. Correct models for conditional densities that are identifiable under $P_{F_X^0,g_0}$ could however be extended to model conditional densities only identifiable under $P_{F_X^0,g_1}$. If the SETA assumption is violated such that no model extensions can be used, models for the conditional densities only identifiable under $P_{F_X^0,g_1}$ can be guessed. The conditional densities non-identifiable under $P_{F_X^0,g_0}$ will then be consistently estimated depending on the correctness of these model extensions or guessed models. In practice, it is highly probable that such model extensions or guessed models will lead to misspecified models for the G-computation formula. Therefore for the G-computation estimator not to depend on these model extensions or guessed models asymptotically, i.e. for n large enough in practice, the SETA assumption is required from the treatment mechanism, Yu. Z and van der Laan (2002a).

References

Brookhart, M.A., van der Laan, J., 2003. A semiparametric model selection criterion with applications to the marginal structural model. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 129.
http://www.bepress.com/ucbbiostat/paper129

Fahrmeir, L., 1990. Maximum likelihood estimation in misspecified generalized linear models. Statistics. 21, 487–502.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer Verlag, New York.

Gill, R.D., Robins, J.M, 2001. Causal inference for complex longitudinal data: the continuous case. Annals of Statistics, 29, 1785-1811.

Gill, R., van der Laan, M., Robins, J., 1997. Coarsening at random: characterizations, conjectures and counterexamples. In: Lin, D.Y., Fleming, T.R. (Eds.), Proceedings of the First Seattle Symposium on Survival Analysis, Springer Verlag, New York, 255–294.

van der Laan, M.J., Dudoit, S., 2003. Unified cross-validation methodology for selection among estimators: finite sample results, asymptotic optimality, and applications. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130.
http://www.bepress.com/ucbbiostat/paper130

van der Laan, M.J., Robins, J.M., 2002. Unified methods for censored longitudinal data and causality. Springer Verlag, New York.

Neugebauer, R., van der Laan, M., 2002. Why prefer double robust estimates? Illustration with causal point treatment studies. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 115.
http://www.bepress.com/ucbbiostat/paper115

Robins, J.M., 1997. Causal inference from complex longitudinal data, latent variable modelling and applications to causality. Lecture notes in statistics, 120, Springer Verlag, New York.

Robins, J.M., 1998a. Marginal Structural Models. Proceedings of the American Statistical Association 1997, 1–10.

Robins, J.M., 1998b. Structural nested failure time models, in: survival analysis.

The Encyclopedia of Biostatistics, J. Willey, Chichester, New York.

Robins, J.M., 2000. Robust estimation in sequentially ignorable missing data and causal inference models. Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999, 6–10.

Robins, J.M., Hernán, M.A., Brumback, B., 2000. Marginal structural models and causal inference in epidemiology. Epidemiology.11(5), 550-560.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–590.

Scharfstein, D.O., Rotnitzky, A., Robins, J.M., 1999. Comments and rejoinder. Journal of the American Statistical Association.94(448), 1121–1146.

Yu, Z., van der Laan, M.J., 2002a. Construction of counterfactuals and the G-computation formula. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 122.
http://www.bepress.com/ucbbiostat/paper122

Yu, Z., van der Laan, M.J., 2002b. Double robust estimation in longitudinal marginal structural models. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 132.
http://www.bepress.com/ucbbiostat/paper132