Robust Estimation of Inverse Probability Weights for Marginal Structural Models

Kosuke Imai and Marc Ratkovic

Marginal structural models (MSMs) are becoming increasingly popular as a tool for causal inference from longitudinal data. Unlike standard regression models, MSMs can adjust for time-dependent observed confounders while avoiding the bias due to the direct adjustment for covariates affected by the treatment. Despite their theoretical appeal, a main practical difficulty of MSMs is the required estimation of inverse probability weights. Previous studies have found that MSMs can be highly sensitive to misspecification of treatment assignment model even when the number of time periods is moderate. To address this problem, we generalize the covariate balancing propensity score (CBPS) methodology of Imai and Ratkovic to longitudinal analysis settings. The CBPS estimates the inverse probability weights such that the resulting covariate balance is improved. Unlike the standard approach, the proposed methodology incorporates all covariate balancing conditions across multiple time periods. Since the number of these conditions grows exponentially as the number of time period increases, we also propose a low-rank approximation to ease the computational burden. Our simulation and empirical studies suggest that the CBPS significantly improves the empirical performance of MSMs by making the treatment assignment model more robust to misspecification.

Open-source software is available for implementing the proposed methods.

KEY WORDS: Causal inference; Covariate balancing propensity score; Inverse propensity score weighting; Observational studies; Sequential ignorability; Time-dependent treatments.

1. INTRODUCTION

Since its introduction by Robins (1999), marginal structural models (MSMs) have quickly gained popularity among applied researchers in biomedical and other fields as a tool for causal inference from longitudinal data in observational studies. The article that popularized MSMs in the field of epidemiology has more than 1000 Google citations as of March 2014 (Robins, Hernán, and Brumback 2000) and the method has been introduced to other disciplines (e.g., Blackwell 2013). As explained by Robins, Hernán, and Brumback (2000), when estimating the causal effects of time-varying treatments, standard regression models fail to appropriately adjust for time-dependent observed confounders that are affected by previous treatments. In contrast, MSMs allow one to estimate the causal effects of different treatment sequences while avoiding this post-treatment bias. Despite their theoretical appeal, a main practical difficulty of MSMs is the required estimation of inverse probability weights. Using simulation and empirical studies, a number of previous studies have found that MSMs can be highly sensitive to misspecification of treatment assignment model even when the number of time periods is moderate (e.g., Mortimer et al. 2005; Kang and Schafer 2007; Cole and Hernán 2008; Lefebvre, Delaney, and Platt 2008; Howe et al. 2011). The effect of misspecification can propagate across time periods because the inverse probability weights used for MSMs are typically based on the product of a propensity score estimated separately at each time period.

To address this problem, we introduce the covariate balancing propensity score (CBPS) methodology as an alternative estimation method for inverse probability weights of MSMs. The CBPS was first introduced by Imai and Ratkovic (2014) to improve the estimation of propensity score in the cross-section setting. In this article, we generalize the CBPS methodology to longitudinal data. In the cross-sectional case, inverse probability weights reduce confounding bias through balancing pretreatment covariates between treated and untreated observations. We extend this logic to the longitudinal setting. We show that, at every time period, MSM weights must balance across all potential future treatment assignments, conditional on the past treatment assignment. Therefore, unlike the standard approach, the proposed methodology incorporates all covariate balancing conditions when estimating inverse probability weights. We then use these balance conditions as estimating equations. The resultant weights are robust in the sense that they improve covariate balance even when the treatment assignment model is misspecified.

After briefly reviewing MSMs and their assumptions (Section 2), we describe the proposed CBPS methodology (Section 3). We then conduct simulation studies to show that the CBPS can dramatically improve the empirical performance of MSMs when the treatment assignment model is misspecified (Section 4). In addition, we present an empirical application to show that the CBPS achieves a greater degree of covariate balance than the standard approach and yields substantively different results (Section 5). The final section gives concluding remarks and discusses future research agenda.

2. A REVIEW OF MARGINAL STRUCTURAL MODELS

In this section, we briefly review the marginal structural models (MSMs) of Robins (1999). See Robins, Hernán, and
Brumback (2000) and Blackwell (2013) for more detailed introduction of MSMs. Suppose that we have a simple random sample of size $n$ from a population. For each unit $i = 1, 2, \ldots, n$, repeated measurements are taken at each of $J$ time periods. Specifically, at each time period $j = 1, 2, \ldots, J$, we observe the time-dependent treatment variable $T_{ij}$ as well as the time-dependent confounders $X_{ij}$ that are possibly affected by previous treatments. We assume that $X_{ij}$ is realized before the treatment at time $j$ and therefore is not affected by $T_{ij}$. We further assume that the treatment variable is binary where $T_{ij} = 1$ ($T_{ij} = 0$) implies unit $i$ receives (does not receive) the treatment at time $j$. Next, for each unit, we denote the treatment and covariate history up to time $j$ by $T_{ij} = (T_{i1}, T_{i2}, \ldots, T_{ij})$ and $X_{ij} = (X_{i1}, X_{i2}, \ldots, X_{ij})$, respectively. We also denote the set of possible treatment and covariate values at time $j$ as $T_j$ and $X_j$. Finally, we observe the outcome variable $Y_i$ for unit $i$ at the end of the study, that is, $Y_i$, after the treatment for the same time period, that is, $T_{ij}$, is administered.

The potential outcome framework of causal inference was originally developed by Neyman (1923) and Rubin (1973) in the cross-section setting, but Robins (1986) generalized it to longitudinal analysis. Under this framework, we use $Y_i(\bar{T}_j)$ to represent the potential outcome value of the eventual outcome variable for unit $i$ measured at time $j$ under the entire treatment history $\bar{T}_j$, where $\bar{T}_j \in T_j$. Thus, the observed outcome is given by $Y_i = Y_i(\bar{T}_j)$. Similarly, $X_{ij}(\bar{T}_{ij-1})$ denotes the potential values of covariates for unit $i$ at each time period $j$ under the treatment history up to time $j - 1$, that is, $\bar{T}_{ij-1}$. Therefore, the observed values of covariates can be written as $X_{ij} = X_{ij}(\bar{T}_{ij-1})$ for unit $i$ at time $j$. This setup relies upon the consistency assumption that the potential values of outcome and covariates for each unit are only functions of its own treatment history up to that point in time. The assumption excludes the possible interference between units (but not between time periods), implying that the potential values of outcome and covariates are not influenced by the treatment history of other units.

MSMs are based on the assumption of sequential ignorability, which states that the treatment assignment of unit $i$ at time $j$ is exogenous given the treatment and covariate history of the same unit up to that point in time. In other words, MSMs assume no unmeasured confounding at each time period. This sequential ignorability assumption can be formally written as

$$Y_i(\bar{T}_j) \perp T_{ij} \mid \bar{T}_{ij-1} = \bar{T}_{ij-1}, \bar{X}_{ij} = \bar{x}_j$$

at any time period $j$ for a given treatment history $\bar{T}_j = \{\bar{T}_{j1}, \bar{T}_{j2}, \ldots, \bar{T}_{jj}\} \in T_j$ and covariate history $\bar{x}_j \in X_j$. We also assume that the conditional probability of treatment assignment is bounded away from zero and one at each time period. That is,

$$0 < \Pr(T_{ij} = 1 \mid \bar{T}_{ij-1} = \bar{T}_{ij-1}, \bar{X}_{ij} = \bar{x}_j) < 1$$

at any time period $j$ for a given treatment history $\bar{T}_{ij-1} \in T_{j-1}$ and covariate history $\bar{x}_j \in X_j$.

Under these assumptions, Robins (1999) showed that the inverse-probability-of-treatment weighting can be used to consistently estimate the marginal mean of any potential outcome, that is, $\mathbb{E}[Y_i(\bar{T}_j)]$ for any treatment sequence $\bar{T}_j$. For the reason that will become clear later, we first define the potential value of this weight for unit $i$ under treatment history $\bar{T}_j$ as

$$w_i(\bar{T}_j, \bar{X}_{ij}(\bar{T}_{j-1})) = \frac{1}{P(T_{ij} = \bar{T}_{ij} \mid \bar{X}_{ij}(\bar{T}_{j-1}))} = \prod_{j=1}^{J} \frac{1}{P(T_{ij} = \bar{T}_{ij} = \bar{T}_{j-1} = \bar{T}_{j-1}, \bar{X}_{ij}(\bar{T}_{j-1}))}.$$

(3)

This weight is typically small and therefore the estimates become highly variable. Therefore, researchers commonly follow the suggestion given in the literature and use the stabilized weights of the form, $w_i^*(\bar{T}_j, \bar{X}_{ij}(\bar{T}_{j-1})) = P(\bar{T}_j = \bar{T}_{ij})/P(\bar{T}_j = \bar{T}_{ij} \mid \bar{X}_{ij}(\bar{T}_{j-1}))$, when fitting the outcome model. We denote the observed values of these weights as $w_i = w_i(\bar{T}_{ij}, \bar{X}_{ij})$ and $w_i^* = w_i^*(\bar{T}_{ij}, \bar{X}_{ij})$.

In an observational study, these weights are unknown and must be estimated. Typically, a parametric model is used to estimate the conditional probability of treatment assignments given the set of covariates,

$$w_i^{-1} = \pi_{\beta}(\bar{T}_j, \bar{X}_{ij}),$$

(4)

where $\beta$ is a finite-dimensional vector of unknown parameters. A common choice of parametric model is the logistic regression independently applied to each time period,

$$\pi_{\beta}(\bar{T}_j, \bar{X}_{ij}) = \prod_{j=1}^{J} \expit(2T_{ij} - 1)\beta_j^\top \bar{V}_{ij}^\ast,$$

(5)

where $\bar{V}_{ij} = [\bar{T}_{ij-1} \bar{X}_{ij}]$, $\expit(z) = (1 + \exp(-z))^{-1}$, and $\beta_j$ is a vector of unknown coefficients. The numerator of the stabilized weight is typically estimated using the sample proportion for each treatment sequence.

Once the (stabilized) weights are estimated, the conditional expectation of outcome is modeled as a function of treatment history alone without covariates, that is, $\mathbb{E}(Y_i \mid \bar{T}_j)$. For example, researchers may regress the outcome on the treatment indicators from all periods. Robins (1999) showed that this weighting approach yields a consistent estimate of the mean potential outcome, that is, $\mathbb{E}(Y_i(\bar{T}_j))$ thereby allowing researchers to compute the average outcome under any sequence of treatment assignments over time.

3. THE PROPOSED METHODOLOGY

In this section, we propose an alternative estimation procedure for the inverse-probability-treatment weight for MSMs. Unlike the standard approach, we estimate the weight such that time-dependent covariates are balanced across all appropriate subpopulations. The proposed methodology generalizes the CBPS of Imai and Ratkovic (2014) to the longitudinal data settings. The key idea of the CBPS is to estimate the propensity score of Imai and Ratkovic (2014) to the longitudinal data settings. Populations. The proposed methodology generalizes the CBPS of Imai and Ratkovic (2014) to the longitudinal data settings. The key idea of the CBPS is to estimate the propensity score given in the literature and use the stabilized weights of the form, $w_i^*(\bar{T}_j, \bar{X}_{ij}(\bar{T}_{j-1})) = P(\bar{T}_j = \bar{T}_{ij})/P(\bar{T}_j = \bar{T}_{ij} \mid \bar{X}_{ij}(\bar{T}_{j-1}))$, when fitting the outcome model. We denote the observed values of these weights as $w_i = w_i(\bar{T}_{ij}, \bar{X}_{ij})$ and $w_i^* = w_i^*(\bar{T}_{ij}, \bar{X}_{ij})$.

In an observational study, these weights are unknown and must be estimated. Typically, a parametric model is used to estimate the conditional probability of treatment assignments given the set of covariates,

$$w_i^{-1} = \pi_{\beta}(\bar{T}_j, \bar{X}_{ij}),$$

(4)

where $\beta$ is a finite-dimensional vector of unknown parameters. A common choice of parametric model is the logistic regression independently applied to each time period,

$$\pi_{\beta}(\bar{T}_j, \bar{X}_{ij}) = \prod_{j=1}^{J} \expit(2T_{ij} - 1)\beta_j^\top \bar{V}_{ij}^\ast,$$

(5)

where $\bar{V}_{ij} = [\bar{T}_{ij-1} \bar{X}_{ij}]$, $\expit(z) = (1 + \exp(-z))^{-1}$, and $\beta_j$ is a vector of unknown coefficients. The numerator of the stabilized weight is typically estimated using the sample proportion for each treatment sequence.

Once the (stabilized) weights are estimated, the conditional expectation of outcome is modeled as a function of treatment history alone without covariates, that is, $\mathbb{E}(Y_i \mid \bar{T}_j)$. For example, researchers may regress the outcome on the treatment indicators from all periods. Robins (1999) showed that this weighting approach yields a consistent estimate of the mean potential outcome, that is, $\mathbb{E}(Y_i(\bar{T}_j))$ thereby allowing researchers to compute the average outcome under any sequence of treatment assignments over time.
Table 1. Orthogonal representation of covariate balancing moment conditions in the two time period case

<table>
<thead>
<tr>
<th>Time period</th>
<th>Treatment history: (t₁, t₂)</th>
<th>Moment condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0,0)</td>
<td>(0,1)</td>
</tr>
<tr>
<td>Time 1</td>
<td>+ +</td>
<td>( E {(-1)^{t_1} w_i X_{i1} } = 0 )</td>
</tr>
<tr>
<td></td>
<td>+ +</td>
<td>( E {(-1)^{t_1} w_i X_{i1} } = 0 )</td>
</tr>
<tr>
<td></td>
<td>+ +</td>
<td>( E {(-1)^{t_1} w_i X_{i1} } = 0 )</td>
</tr>
<tr>
<td>Time 2</td>
<td>+ +</td>
<td>( E {(-1)^{t_1} w_i X_{i1} } = 0 )</td>
</tr>
<tr>
<td></td>
<td>+ +</td>
<td>( E {(-1)^{t_1} w_i X_{i1} } = 0 )</td>
</tr>
<tr>
<td></td>
<td>+ +</td>
<td>( E {(-1)^{t_1} w_i X_{i1} } = 0 )</td>
</tr>
</tbody>
</table>

NOTES: The first and second time periods have three and two moment conditions, respectively. There are four distinct values of treatment history with \( t_j \) representing the value of the treatment variable at time \( j \). The symbols, “+” and “−”, in these four treatment history columns show whether the weighted average of covariates among units with a certain treatment history is added or subtracted when formulating the moment condition. Within each time period, row vectors of +’s and −’s for the treatment history combinations are orthogonal to one another. The last column represents the corresponding moment condition.

This orthogonal representation of covariate balancing conditions is summarized in the first three rows of Table 1. In the table, if we treat + and − as +1 and −1, row vectors for each time period are orthogonal to each other.

The covariate balancing conditions at the second time period are similar to those at time 1, except that the covariates measured at time 2 are possibly functions of the treatment at time 1, that is, \( X_{i2} = X_{i2}(T_{i1}) \). This means that the covariate balancing conditions will be conditional on the observed treatment value at time 1. Using the potential outcomes notation, for all \( \vec{t}_2 = (t_1, t_2) \), we can write these covariate balancing conditions as follows:

\[
E \{ X_{i2}(t_1) \} = E \left[ I(T_{i1} = t_1, T_{i2} = t_2) w_i(\vec{t}_2, \bar{X}_{i2}(t_1))X_{i2}(t_1) \right].
\]

(11)

Because \( X_{i2}(t_1) \) is observed only when \( T_{i1} = t_1 \), the above covariate balancing equation implies that \( X_{i2} \) should be balanced across treatment values at time 2 conditional on the treatment value realized at time 1.

Similar to the baseline covariate case, we use the orthogonal representation, which in this case yields the following two moment conditions,

\[
E \{(-1)^{t_1} w_i X_{i2} \} = 0
\]

(12)

\[
E \{(-1)^{t_1} w_i X_{i2} \} = 0
\]

(13)

The bottom two rows of Table 1 summarize this result. While at time 1 both \( T_{i1} \) and \( T_{i2} \) are varied to generate three moment conditions, only \( T_{i2} \) is varied at time 2. By now, readers may realize the benefit of our orthogonal representation: as shown in Section 3.3, its compact notation allows one to easily extend the proposed methodology to the general case of more than two time periods.

**Estimation.** Since the number of moment conditions exceeds the number of parameters to be estimated, we use the generalized method of moments (GMM; Hansen 1982) estimation to combine the covariate balancing conditions derived above. Our optimal GMM estimator is given by

\[
\hat{\beta} = \arg \min_{\beta \in \Theta} \operatorname{vec}(G)^T W^{-1} \operatorname{vec}(G),
\]

(14)
where the sample moment conditions are given by

\[
G = \frac{1}{n} \sum_{i=1}^{n} \left[ \begin{array}{ccc}
(-1)^{T_i} w_i T_{i1} & (-1)^{T_i} w_i T_{i1} & (-1)^{T_i+1} w_i T_{i1} \\
0 & (-1)^{T_i} w_i T_{i1} & (-1)^{T_i} w_i T_{i1}
\end{array} \right],
\]

(15)

and their covariance \( W \) is given by

\[
W = \frac{1}{n} \sum_{i=1}^{n} E \left[ \begin{array}{ccc}
1 & 1 & (-1)^{T_i+1} \\
(-1)^{T_i+1} & 1 & (-1)^{T_i} \\
(-1)^{T_i} & (-1)^{T_i+1} & 1
\end{array} \right] \odot w^2_i \left( \begin{array}{c} X_{1i} T_{i1}^X \\
X_{12} T_{i1}^X \\
X_{12} T_{i2}^X
\end{array} \right) \left( \begin{array}{c} X_{11} \\
X_{12}
\end{array} \right) .
\]

(16)

The expectation in Equation (16) can be calculated analytically, for example, for the logistic regression case (Imai and Ratkovic 2014).

### 3.3 The General Longitudinal Case

We now extend the above formulation to the general case with more than two time periods, that is, \( j = 1, 2, \ldots, J \). We first generalize the covariate balancing conditions derived above and then propose the optimal GMM estimator. We also consider a low-rank approximation to address a computational challenge when the number of time periods is large.

**Covariate Balancing Conditions.** We characterize the covariate balancing conditions in the general case with an arbitrary number of time periods \( J \geq 2 \). Recall that in the two time period case, the weight for MSMs balances the covariates at the first time period across all potential combinations of the current and future treatment conditions. At the second time period, however, the weight only balances covariates across the treatment values at that time period among the units who receive the same treatment value in the first time period. In general, the weight balances covariates at a given time period across all potential combinations of the current and future treatment conditions given the past treatment sequence.

Formally, for a given time period \( j \) and fixed past treatment sequence up to that point \( \bar{t}_{j-1} \), we can write the covariate balancing conditions across all treatment sequences of the current and future time periods \( T_j = \{t_j, t_{j+1}, \ldots, t_J\} \) as

\[
E\{X_{ij}(\bar{t}_{j-1})\} = E\{1_{T_{j-1} = \bar{t}_{j-1}, T_{j}}\} = E\{1_{T_{j-1} = \bar{t}_{j-1}, T_{j}} \} w_i(\bar{t}_j, X_{ij}(\bar{t}_{j-1})) X_{ij}(\bar{t}_{j-1}^{-1}).
\]

(17)

where \( T_{j} = \{T_{i1}, T_{i,j+1}, \ldots, T_{ij}\} \) represents a vector of observed current and future treatment conditions.

In the two time period case, the balance conditions are characterized in terms of the sums and differences of \( w_i T_{ij} \) across all groups defined by a distinct value of the entire treatment sequence. We generalize that formulation here. Specifically, for each time period, we use the orthogonal representation of the covariate balancing conditions given in Equation (17) by aliasing the past treatment effects on the covariates at time \( j \). Since there exists a total of \( 2^{J-j+1} \) potential current and future treatments, Equation (17) implies \( 2^{J-j+1} - 1 \) orthogonal constraints given a particular history of treatment up to time \( j - 1 \), that is, \( \bar{t}_{j-1} \). There are a total of \( 2^{J-1} \) possible treatment histories and hence all together we have \( (2^{J} - 2^{j-1}) \) covariate balancing conditions for each time period \( j \).

To formalize this idea, we use the theoretical framework developed for analyzing and designing randomized experiments based on the \( 2^J \) full factorial design (see, e.g., Box, Hunter, and Hunter 2005). In Table 2, we present a running example of the case with \( J = 3 \) where the first three columns present the design matrix in Yates order with +’s and −’s indicating the presence and absence of the treatment at each time period, respectively. It is well recognized that the full \( 2^J \) factorial design can be represented by Hadamard matrix of order \( 2^J \). Recall that Hadamard matrix of order \( n \), denoted by \( H_\text{H}_n \), is an \( n \times n \) matrix of +1’s and −1’s whose rows are orthogonal to one another, implying that \( H_\text{H}_n' H_\text{H}_n = n I_n \).

To construct a Hadamard matrix that corresponds to the full \( 2^J \) factorial design, let \( D \) be the \( 2^J \times J \) “negative” design matrix of +1’s and −1’s sorted in Yates order,

\[
D = [d_0, d_1, d_2, d_3, d_4, d_5, \ldots]'.
\]

where \( d_0 \) is a \( J \)-dimensional column vector of 1’s and \( d_j \) is a column vector of length \( J \), where the elements of set \( j \) indicate the indexes of the elements of the vector with −1 and the other elements of the vector are 1’s. For example, when \( J = 3 \), we have \( d_{12} = (-1, -1, 1)' \). Thus, +’s and −’s in Table 2 correspond to −1’s and +1’s in \( D \), respectively. Let \( c_j \) be the \( j \)th column of \( D \) so that \( D = [c_1, c_2, \ldots, c_{J}] \).

Further, denote the common component of \( d_j \) and \( c_k \) by \( d_{jk} \). For a subset \( t \) of \( N_j = \{1, \ldots, J\} \), let the Hadamard product, denoted by \( H_t \), of columns \( c_k \) with \( k \in t \) be a \( 2^{|t|} \)-dimensional column vector with its \( j \)th element being \( \prod_{k \in t} d_{jk} \). Then, the Hadamard matrix of order \( 2^J \) can be constructed by collecting in Yates order all the Hadamard products of the columns of \( D \). The result is given by the following \( 2^J \times 2^J \) matrix,

\[
H_{2^J} = [h_0, h_1, h_2, h_3, \ldots]' \]

(19)

where \( h_0 \) is a column vector of +1’s. This matrix in the case of \( J = 3 \) is given in the middle columns of Table 2.

The Hadamard matrix representation allows us to enumerate all the covariate balancing moment conditions in a systematic way regardless of the number of time periods. Moreover, the successive multiplication procedure used for the construction of this Hadamard matrix directly justifies the notation used in Equations (12) and (13). In fact, it has long been known that this Hadamard matrix representation can be used to compute the mod 2 discrete Fourier transform (Good 1958). For example, the second and sixth rows of Table 2 correspond to the following covariate moment conditions,

\[
E\{(-1)^{T_{i1}} w_i T_{ij}\} = 0 \quad (20)
\]

\[
E\{(-1)^{T_{i1}+T_{i2}} w_i T_{ij}\} = 0. \quad (21)
\]

That is, one can use the design matrix to form the treatment variables that enter the exponent of −1 in the compact expression of the covariate balancing moment conditions. In sum, the \( 2^J \) factorial experiment framework allows us to directly generalize the orthogonal representation of the covariate moment conditions given in Section 3.2 to the general case with more than two time periods.

Therefore, this full \( 2^J \) factorial design framework clearly shows which covariate balancing moment conditions are binding at any given time period for the estimation of the weight
for MSMs. As noted above, the stabilized weight balances covariates measured at time $j$ across all possible current and future treatments but it does not balance across past treatments. The covariate balancing moment conditions, which correspond to the effects of past treatments and their interactions, are not binding. These conditions can be easily identified by the design matrix. For example, in Table 2, we see that the second row, corresponding to the main effect of time 1 treatment, that is, $T_{11}$, is not binding for time 2 covariates $X_{22}$. Similarly, for time 3 covariates, the moment conditions corresponding to the effects of $T_{11}$ and $T_{12}$ as well as their interaction are not binding. In general, for covariates measured at time $j$, the first $2^{j-1}$ rows of Hadamard matrix $H_{2^{j}}$ can be ignored when constructing the covariate balancing moment conditions.

Estimation. As in the two time period case, we use the GMM to combine all the covariate balancing conditions. We begin by defining the following matrices for covariates:

$$\tilde{X} = [\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n]^\top,$$  \hspace{1cm} (22)

where $\tilde{X}_i = [w_1X_{i1}, w_1X_{i2}, \ldots, w_1X_{ij}]^\top$ is a $(K \times J)$-dimensional column vector of covariates for unit $i$. Next, we construct the $n \times (2^J - 1)$ model matrix based on the design matrix $D$ arranged in Yates’ order as

$$M = [M_1, M_2, \ldots, M_n],$$ \hspace{1cm} (23)

where $M_j = [m_{10}, m_{11}, m_{12}, m_{112}, m_{13}, m_{113}, m_{123}, m_{1123}, m_{114}, m_{1124}, m_{1134}, m_{1234}, \ldots]^\top$ is a $(2^J - 1)$-dimensional column vector with $m_{ij} = 1$ and $m_{ij} = (-1)\sum_{s=0}^{T_*} c_s$ for $t \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, \ldots\}$. For example, $m_{123}$ equals $(-1)T_{11}T_{12}$ and $m_{1123}$ equals $(-1)T_{11}T_{12}T_{13}$. In fact, the $i$th row of $M$ is given by the row of the Hadamard matrix in Table 2 that corresponds to the treatment sequence of the $i$th observation.

Given this notation, our optimal GMM estimator is given by Equation (14) with the following generalized definitions of the sample balancing condition and their conditional covariance,

$$G = \frac{1}{n} \sum_{i=1}^{n} (M_i^\top \otimes \tilde{X}_i) R \hspace{1cm} (24)$$

$$W = \frac{1}{n} \sum_{i=1}^{n} E (M_iM_i^\top \otimes \tilde{X}_i\tilde{X}_i^\top | X_i) \hspace{1cm} (25)$$

Figure 1. Treatment variable data-generating process in simulation studies. In the first set of simulations summarized by the diagram of panel (a), a relatively simple treatment assignment model is used and we only misspecify the functional form while maintaining the correct lag structure. In the second set of simulations summarized by the diagram of panel (b), a more complex data-generating process is used and we examine the impact of incorrectly specifying the lag structure. The results of these simulations are given in Figures 2 and 3, respectively.
where $\mathbf{R}$ represents the “selection” matrix, which identifies the binding covariate balancing conditions for each time period and “zeros” them out. This matrix is formally defined as

$$
\mathbf{R} = [\mathbf{R}_1 \mathbf{R}_2 \ldots \mathbf{R}_J], \quad \text{where} \quad \mathbf{R}_j = \begin{bmatrix}
0_{2^{j-1} \times 2^{j-1}} & 0_{2^{j-1} \times (2^j - 2^{j-1})} \\
0_{(2^j - 2^{j-1}) \times 2^{j-1}} & \mathbf{I}_{2^j - 2^{j-1}}
\end{bmatrix}
$$

for each $j = 1, \ldots, J$. As mentioned earlier, the expectation in Equation (25) can be evaluated analytically for the logistic regression case.

When the number of time periods is large, the inversion of $\mathbf{W}$ can be computationally expensive because its dimension, which is $(2^J - 1) \times JK$, increases exponentially as a function of $J$. Here, we derive a low-rank approximation to the full covariance matrix as a way to overcome this computational difficulty. To do this, we assume that the correlation across balance conditions is zero. Note that when this assumption does not hold, the resulting GMM estimator is still consistent but no longer efficient. In our simulation and empirical studies (see Sections 4 and 5), we find that the empirical performance is not greatly affected by this approximation especially in a large sample size.

Specifically, our low-rank approximation to the covariance matrix is given by

$$
\tilde{\mathbf{W}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I} \otimes \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top = \mathbf{I} \otimes \mathbf{X}^\top \mathbf{X},
$$

where the variances in this new matrix are identical to those in the original $\mathbf{W}$ matrix of Equation (25) but certain covariances are zero. Then, our GMM estimator is given by

$$
\hat{\beta} = \arg\min_{\beta \in \Theta} \text{vec}(\mathbf{G})^\top \left( \mathbf{I} \otimes \tilde{\mathbf{x}}^\top \tilde{\mathbf{x}} \right)^{-1} \text{vec}(\mathbf{G})
= \arg\min_{\beta \in \Theta} \text{trace} \left( \mathbf{R}^\top \mathbf{M}^\top \tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top \left( \tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top \right)^{-1} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^\top \mathbf{M} \right).
$$

Thus, this approximation approach avoids the Kronecker product and the inversion of a large matrix.

### 3.4 Extension to Multiple Binary Treatments

The method described above naturally extends to the setting where there exist multiple binary treatments. Indeed, dynamic treatment regimes considered in this article is essentially a special case of $J$ multiple binary treatments. The only difference is that for dynamic treatment regimes some of the covariate
balancing conditions are not binding as indicated by zero elements of \( G \) matrix in Equations (15) and (24). In contrast, for multiple binary treatments, all of these covariate balancing conditions are binding. However, aside from this difference, the estimation for the case of multiple binary treatments proceeds in an identical manner.

### 4. SIMULATION STUDIES

We conduct four sets of simulation studies to assess the empirical performance of the proposed CBPS estimation. First, we show that when the treatment assignment model is correctly specified, the proposed methodology does as well as the standard maximum likelihood estimation. Second, we also examine several scenarios where the treatment assignment model is misspecified in terms of either the lag structure or the functional form while maintaining the correct lag structure. The treatment-generating process here is a functional forms using a nonlinear transformation. Specifically, the treatment assignment in the second set of simulations is given by \( \Pr(T_{ij} = 1) = \expit(\sum_{j=1}^{j-1} (T_{ij} - 1)/3 + (-1/2)) \). The true treatment assignment model is a function of the entire covariate and treatment history for each observation, but each method is applied using the most immediate covariates and treatment. To generate our covariates for this set of simulations, we adjust \( U_{ij} \) such that \( U_{ij} = \prod_{j=1}^{j-1} \{2 + (2T_{ij} - 1)/3\} \) for \( j = 2, 3 \) and \( U_{ij} = 1 \) for \( j = 1 \). The new set of covariates are then constructed as \( X_{ij} = (Z_{ij1}U_{ij}, Z_{ij2}U_{ij}, |Z_{ij3}U_{ij}|, |Z_{ij4}U_{ij}|)^T \) so that they are a function of all past treatments. The outcome model is the same as the one used for the first set of simulations except that the definition of \( X_{ij} \) is different. As before, we assess each methods’ performance when using the correct covariates, \( X_{ij} \), and the covariates after a mild nonlinear transformation, \( X_{ij} = (Z_{ij1}U_{ij})^3, 6 \cdot Z_{ij2}U_{ij}, \log |Z_{ij3}U_{ij}|, 1/|Z_{ij4}U_{ij}|)^T \).

To evaluate the performance of our proposed CBPS methodology, we simulate 2500 datasets using the aforementioned datagenerating processes. We then fit a logistic regression model (GLM) as the treatment assignment model independently for each time period using correct and incorrect model specifications as discussed above. We also fit the same exact logistic model using the proposed CBPS estimation but in two ways: first with the fully efficient covariance matrix (CBPS) and with its low-rank approximation (CBPS-Approximate). Finally, the marginal structural model (MSM) weights are constructed from each of the fitted models and then we regress the outcome variable on all three treatment variables using the stabilized MSM weights. The resulting regression coefficients are then compared with the numerical estimates of true regression coefficients obtained from a large number of simulations with the true treatment assignment probabilities.

Figure 2 presents the results from the first set of simulations where the misspecification of treatment assignment model is confined to the functional form and the correct lag structure is maintained. The first three columns show that the bias (upper two rows) and root mean squared error or RMSE (bottom two rows) for the estimated regression coefficients of the three treatment variables (one for each of the three time periods, i.e., \( \beta_1, \beta_2, \) and \( \beta_3 \), respectively) from the MSM. That is, we use a weighted linear regression where the outcome is regressed on three treatments using the MSM weights. The final column presents the bias and RMSE for the estimated mean potential outcome, \( E(Y_i(t_1, t_2, t_3)) \), averaged across eight unique treatment sequences. These estimates are obtained by calculating the weighted average of the outcome using the subset of data for each treatment sequence.

When the treatment assignment model is correctly specified, all methods have small bias (the first row) and small RMSE (third row) for all quantities of interest. For one parameter in a small sample size, CBPS with the low-rank approximation (thick dash line) has a greater bias than other methods. It is also interesting to note that CBPS is slightly more efficient than the GLM. This
finding is consistent with the theoretical result of Hirano, Imbens, and Ridder (2003), which implies that overparameterizing the propensity score model by adding moment conditions can sometimes lead to efficiency gains. However, when the model is misspecified, the bias and RMSE are large and even grow in sample size for GLM (thin dot-dashed lines). In contrast, CBPS with the fully efficient covariance matrix (thick solid line) and CBPS with the low-rank approximation have much smaller bias and RMSE across parameters. Unlike the GLM, both the bias and RMSE do not grow in sample size, thereby outperforming the standard estimation technique.

In the first row of Figure 3, the misspecified lag structure induces noticeable bias across all methods, with the CBPS methods showing modest gains in bias (first row) and RMSE (third row). When the lag structure is misspecified and additionally the covariates are transformed (second and fourth rows), the standard GLM estimation leads to much larger bias and RMSE and this bias increases in the sample size. In contrast, the CBPS methods minimize the impact of model misspecification and stays within a reasonable range for bias and RMSE across all quantities of interest.

5. AN EMPirical APplication

We now illustrate the proposed methodology through an empirical application. Blackwell (2013) had applied the MSM to the data from political science to estimate the impact of negative advertisements on election outcomes. Here, we analyze
were dropped so as to make the common support assumption leading to the elections, using a total of 58 U.S. Senate and 56 gubernatorial candidates from 114 races that were held during the years 2000, 2002, 2004, and 2006. During this time, there were 126 races total in which ads were aired during the last five weeks. Five races are dropped due to missing data, and, following the original author, seven additional noncompetitive races were dropped so as to make the common support assumption more credible.

In each week \( t = 1, \ldots, 5 \) leading up to the election, candidate \( i \) may run negative campaign \( (T_{it} = 1) \) or remain positive \( (T_{it} = 0) \). The time-varying covariates \( X_{it} \) include the Democratic share of the polls, proportion of voters undecided, campaign length, and the lagged and twice lagged treatment variables for each week. In addition, we use the time-invariant covariates including baseline Democratic votes, baseline proportion undecided, and indicators for election year, incumbency status, and type of office. The original study fit a single logistic regression model to all time periods, including a linear time trend. In contrast, we allow the coefficients in the model to be different for each time period. We find that the added flexibility yields significantly better covariate balance.

We consider three approaches: the logistic regression, the CBPS based on the optimal covariance matrix (CBPS), and the CBPS based on the low-rank approximation (CBPS-Approximate). For the CBPS and CBPS-approx approaches, we use the two-step and continuously updating GMM estimators, respectively. Note that the computational time for the CBPS was seven times as long as that for the CBPS-Approximate.

We begin by assessing the degree of covariate balance achieved by the logistic regression and the CBPS. Since there are 12 covariates per time period, we have a total of \( 1548 = \sum_{j=1}^{5} 12 \times (2^{5-j} - 1) \) different covariate balancing conditions. As shown in Section 3, these moment conditions are implied by the fact that at each time period, conditional on the past treatment history, the MSM weights should balance covariates across all future potential treatment sequences. We characterize the imbalance as the absolute value of the balance conditions for each balance condition \( G \).

**Figure 4** presents the absolute imbalance for each covariate balancing condition based on the logistic regression (vertical axis) and the CBPS (horizontal axis). The imbalance of all balance conditions appear together in the top left plot, and they are broken out by time period in the remaining five plots. Points above the 45° line indicate that a better balance is achieved for the CBPS than the logistic regression. Overall, the covariate balance achieved by the logistic regression tends to be worse than the CBPS.
81.3% in time 4. Relative to the CBPS, the logistic regression has both a greater average absolute imbalance (0.84 versus 0.24) and a larger spread in absolute imbalance (2.05 versus 0.65).

Table 3 presents the estimated impact of negative advertising on candidate’s voteshare. The left three columns present the estimated average causal effects of the time-specific decision to engage in negative advertising. The right three columns contain the estimated causal effects of the cumulative number of periods that the candidate has gone negative. All results are based on weighted linear regressions. The weights are estimated using the logistic regression (GLM), the CBPS with the optimal covariance matrix, and the CBPS with low-rank approximation (CBPS-Approximate). While these results are somewhat similar across the methods, there are some differences. In particular, when using the CBPS and CBPS-Approximate, the effects of negative advertising are estimated appear to be more strongly negative than the GLM: the magnitude of estimates effects is larger and standard errors tend to be smaller. The \( R^2 \) and \( F \) statistics are also greater when the weights are estimated using the CBPS and CBPS-Approximate. Finally, the low-rank approximation approach for the CBPS does not alter the results significantly.

### 6. CONCLUDING REMARKS

In this article, we have extended the CBPS methodology of Imai and Ratkovic (2014) to the estimation of inverse probability weights for marginal structural models (MSMs), a popular tool in the analysis of longitudinal data. The proposed methodology estimates these weights by improving the resulting covariate balance. This is an important advantage because checking covariate balance, after fitting the treatment assignment models, is a difficult task even when the number of time periods is moderate. As a result, detecting model misspecification is much more challenging in longitudinal data settings than simple cross-sectional data settings.

In addition, because the MSM weights are constructed by multiplying the inverse of the estimated propensity scores from each time period, MSMs can be highly sensitive to the misspecification of treatment assignment models. In contrast, the CBPS methodology provides a robust estimation method for inverse probability weights by maximizing covariate balance. Our simulation and empirical studies illustrate the effectiveness of the proposed method over the standard maximum likelihood estimation.

One possible future research agenda is the extension of the proposed methodology to nonparametric estimation using empirical likelihood. The MSM weights are often estimated using a parametric model but better covariate balance might be achieved by using a more flexible estimation approach. Another important and yet unresolved question concerns the selection of covariate balancing conditions when there exist many such conditions. As we have shown, the number of covariate balancing conditions grow exponentially as the number of time periods increases. Under this scenario, the data will become sparse and some treatment sequences have extremely small number of observations. Here, the application of moment selection methods may be useful. We plan to investigate how the proposed CBPS methodology performs in such a situation and develop effective strategies for addressing this issue.

### REFERENCES


