2-2015

Optimal Restricted Estimation for More Efficient Longitudinal Causal Inference

Edward H. Kennedy

Marshall M. Joffe
University of Pennsylvania

Dylan S. Small
University of Pennsylvania

Recommended Citation

Kennedy, E. H., Joffe, M. M., & Small, D. S. (2015). Optimal Restricted Estimation for More Efficient Longitudinal Causal Inference. *Statistics & Probability Letters,* 97 185-191. http://dx.doi.org/10.1016/j.spl.2014.11.022

Optimal Restricted Estimation for More Efficient Longitudinal Causal Inference

Abstract
Efficient semiparametric estimation of longitudinal causal effects is often analytically or computationally intractable. We propose a novel restricted estimation approach for increasing efficiency, which can be used with other techniques, is straightforward to implement, and requires no additional modeling assumptions.

Keywords
doubly robust, generalized method of moments, marginal structural model, semiparametric efficiency, structural nested model, time-varying confounding

Disciplines
Business | Business Analytics | Health Services Administration | Health Services Research | Medical Humanities | Statistics and Probability

Published in final edited form as:
Stat Probab Lett. 2015 February 1; 97: 185–191. doi:10.1016/j.spl.2014.11.022.

Optimal restricted estimation for more efficient longitudinal causal inference

Edward H. Kennedy[a,*], **Marshall M. Joffe**[a], and **Dylan S. Small**[b]

[a]Department of Biostatistics and Epidemiology, University of Pennsylvania

[b]Department of Statistics, The Wharton School, University of Pennsylvania

Abstract

Efficient semiparametric estimation of longitudinal causal effects is often analytically or computationally intractable. We propose a novel restricted estimation approach for increasing efficiency, which can be used with other techniques, is straightforward to implement, and requires no additional modeling assumptions.

Keywords

Doubly robust; Generalized method of moments; Marginal structural model; Semiparametric efficiency; Structural nested model; Time-varying confounding

1. Introduction

Locally efficient semiparametric estimation of causal effects in longitudinal studies can be analytically or computationally intractable; however, more simple and straightforward estimation techniques can be very imprecise. In this work we develop an approach for deriving more efficient estimators of parameters in such settings based on the idea of optimal restricted estimation, i.e., finding estimators that are optimally efficient among all those within some restricted class. In essence our approach amounts to finding optimal linear combinations of estimating functions, using constant coefficient matrices. The proposed approach can be used in conjunction with other techniques (such as those based on local efficiency derivations), is straightforward to implement, requires neither extra modeling assumptions nor extra model fitting, and comes with guarantees of better (or at least no worse) asymptotic efficiency. It can be viewed as a way to give analysts extra chances at attaining the semiparametric efficiency bound. We explore finite sample properties of our approach using simulated data.

[*]Corresponding author at: Room 507, Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, kennedye@mail.med.upenn.edu (Edward H. Kennedy).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

2. Setup

Many important models in longitudinal causal inference, including structural nested models (Robins, 1989, 1994) and marginal structural models (Robins, 2000; Hernán et al., 2002), lead to estimators that solve (at least up to asymptotic equivalence) estimating equations of the form

$$\mathbb{P}_n\left\{\sum_{t=1}^{K} m_t(\psi;\hat{\eta},h)\right\}=0$$

where $\mathbb{P}_n$ is the empirical measure so that $\mathbb{P}_n(W) = n^{-1}\sum_i W_i$ denotes a usual sample average, m_t is an estimating function of the same dimension as the parameter of interest $\psi \in \mathbb{R}^q$, η is a nuisance function taking values in some metric space, and h is an arbitrary function that affects the efficiency but not consistency of the estimator.

For example, in many settings the observed data consist of sequences of time-varying measurements of covariates L, treatment A, and outcome Y for each of n subjects. Let an overbar denote the past history of a variable so that $\bar{W}_t = (W_1, W_2, \dots, W_t)$, and let $X_t = (\bar{L}_t, \bar{Y}_t, \bar{A}_{t-1})$ represent the observed data available just prior to treatment at time t. Also for simplicity assume no censoring and discrete measurement times $t = 1, \dots, K$. Then a standard longitudinal study would yield an independent and identically distributed sample of observations $(Z_1, \dots, Z_n)$, with $Z = (\bar{L}_K, \bar{A}_K, \bar{Y}_{K+1})$. Figure 1 shows a directed acyclic graph illustrating this data structure, allowing for the presence of unmeasured variables U and only incorporating the assumed time ordering.

Let $Y_{t+1}^{\bar{a}_t}$ denote the potential outcome that would have been observed for a particular subject had that subject taken treatment sequence $\bar{a}_t$ up to time t. Then a standard repeated measures marginal structural mean model (MSMM) (Robins, 1989, 1994) assumes

$$E\left(Y_{t+1}^{\bar{a}_t}|V=v\right)=g_t(\bar{a}_t,v;\psi)$$

for $t = 1, \dots, K$ and g_t specified functions known up to the parameter of interest ψ, where $V \subseteq L_1$ is an arbitrary subset of baseline covariates whose modification of the effect of treatment is of particular interest. Similarly a standard structural nested mean model (SNMM) (Robins, 2000; Hernán et al., 2002) assumes that

$$E\left(Y_{K+1}^{\bar{a}_t,0}-Y_{K+1}^{\bar{a}_{t-1},0}|X_t=x_t,A_t=a_t\right)=\gamma_t(x_t,a_t;\psi)$$

for $t = 1, \dots, K$, where the specified functions γ_t (also known up to ψ) are restricted so that $\gamma_t(x_t, 0; \psi) = 0$ since $Y_{K+1}^{\bar{a}_t,0}-Y_{K+1}^{\bar{a}_{t-1},0}=0$ if $a_t = 0$. We consider linear SNMMs for effects on the last outcome for ease of notation, but one could similarly use a log link or repeated measures models for effects on all outcomes. One could also consider versions of the above models that contrast functionals other than the mean (e.g., percentiles).

As discussed by van der Laan and Robins (2003), Tsiatis (2006), and others, under standard ‘no unmeasured confounding’ identifying assumptions (e.g., sequential ignorability, or $Y_{t+s}^{\overline{a}_K} \perp\!\!\!\perp A_t | X_t$ for t = 1, …, K and s = 1, …, K+1−t), estimating functions m_t under the above MSMMs and SNMMs are given by $m_t(\psi; \eta, h) = \varphi_t(\psi; \eta_a, h) - E\{\varphi_t(\psi; \eta_a, h) \mid X_t, A_t\} + E\{\varphi_t(\psi; \eta_a, h) \mid X_t\}$ where

$$\phi_t(\psi;\eta_a,h)=\begin{cases} h_t(\overline{A}_t,V)\left\{\frac{Y_{t+1}-g_t(\overline{A}_t,V;\psi)}{\prod_{s=1}^{t}p(A_s|X_s)}\right\} & \text{for MSMMs} \\ \{h_t(X_t,A_t)-\int h_t(X_t,a_t)p(a_t|X_t)\,d\nu(a_t)\}\left\{Y_{K+1}-\sum_{s=t}^{K}\gamma_s(X_s,A_s;\psi)\right\} & \text{for SNMMs,} \end{cases}$$

with the functions $p(a_t \mid x_t)$ denoting the conditional density of treatment given observed history, and ν a dominating measure for the distribution of treatment. In this setting the nuisance function $\eta = (\eta_a, \eta_y)$ consists of two variation independent components; η_a denotes the conditional treatment densities $p(a_t \mid x_t)$ and η_y denotes the conditional outcome/covariate densities $p(l_t, y_t \mid x_{t-1}, a_{t-1})$. Importantly, the functions $h_t : D_t \to \mathbb{R}^q$ (where $D_t = (\bar{A}_t, V)$ for MSMMs and $D_t = (X_t, A_t)$ for MSMMs) are arbitrary but of the same dimension as ψ; they lie in q-replicating linear spaces $\mathscr{H}_t^q=\{(h_{t1},\ldots,h_{tq}):h_{tj}(D_t)\in\mathbb{R}\}$ of stacked one-dimensional functions (Tsiatis, 2006).

The standard approach for estimating ψ is to construct estimating functions based on the above using a simple choice h^* of h, for example $h_t^*=\partial g_t/\partial\psi$ for MSMMs or $h_t^*=\partial\gamma_t/\partial\psi$ for SNMMs. Under usual Glivenko-Cantelli and Donsker-type regularity conditions, standard Z-estimator (i.e., estimating equation) theory indicates that $\hat{\psi}$ solving $\Sigma_t\, \mathbb{P}_n\{m_t(\psi; \hat{\eta}, h^*)\} = 0$ will be consistent as long as at least one of the two nuisance functions η_a or η_y is estimated consistently; thus, letting η_0 denote the probability limit of $\hat{\eta}$, we only need to assume one of $\hat{\eta}_a$ or $\hat{\eta}_y$ converge to a corresponding true value. Further $\hat{\psi}$ will be root-n consistent and asymptotically normal as long as at least one of the two nuisance functions is estimated at a fast enough rate of convergence. Thus estimating functions of the above form have the property of double robustness (van der Laan and Robins, 2003; Tsiatis, 2006). In practice, especially in longitudinal settings, one often chooses η_y so that $m_t = \varphi_t$; for MSMMs, for example, this yields the class of popular inverse-probability-weighted estimators. Such estimators are often easier to construct with standard software, but are less robust since they require estimating η_a well and yield bias otherwise.

3. Restricted estimation

For given choices of the nuisance estimator $\hat{\eta} = (\hat{\eta}_a, \hat{\eta}_y)$, the efficiency of estimators $\hat{\psi}$ solving $\Sigma_t\, \mathbb{P}_n\{m_t(\psi; \hat{\eta}, h^*)\} = 0$ will in general vary greatly depending on the choice of the functions h^*. Let $\phi(\psi; \eta, h) = D(\psi; \eta, h)\, \Sigma_t\, m_t(\psi; \eta, h)$ denote the influence function of the estimator $\hat{\psi}$, where $D \in \mathbb{R}^{q\times q}$ is a scaling matrix ensuring that $E(\varphi S_{\psi}^{\mathrm{T}}) = I_q$, where S_ψ is the score function for ψ and I_q is the $(q \times q)$ identity matrix. The optimal choice of h is therefore

$$h^{\mathrm{opt}} = \underset{h_t \in \mathscr{H}_t^q}{\arg\min} E\left\{\varphi(\psi_0;\eta_0,h)^{\otimes 2}\right\}$$

so that $\varphi(\psi;\ \eta_0,\ h^{\mathrm{opt}}) = \varphi_{\mathrm{eff}}$ corresponds to the efficient influence function. Unfortunately, the optimal choice of h is often prohibitively complicated. For MSMMs h^{opt} is defined as the solution to a Fredholm integral equation of the second kind, and does not have a closed-form expression (Robins, 2000; van der Laan and Robins, 2003). Similarly, for SNMMs h^{opt} follows a lengthy recursive expression and requires extensive modeling (Robins, 1994).

For this reason, Tsiatis (2006) and Tan (2011) proposed approximate methods for increasing the efficiency of estimators in a context akin to that of a point-treatment or cross-sectional MSMM (i.e., $t = K = 1$). Specifically, instead of optimizing h over the infinite-dimensional spaces H_t^q, their approach adapted to our context involves choosing some fixed r-dimensional $h_t^* \in \mathscr{H}_t^r$ with $r > q$, and optimizing over the restricted finite-dimensional spaces

$$H_t^{\mathrm{res}}(h_t^*) = \left\{W h_t^* : W \in \mathbb{R}^{q\times r}, h_t^* \in \mathscr{H}_t^r \text{ fixed with } r > q\right\}. \quad (1)$$

This optimization thus finds the optimal way to combine or weight the estimating functions that make up h_t^*, using a constant “weight” matrix W. The dimension of the function h_t^* must be strictly greater than that of the parameter of interest (i.e., $r > q$) because otherwise any nonsingular matrix W would lead to the same estimator. More specifically if $r = q$ then the solution to $\mathbb{P}_n\{\sum_t m_t(\psi;\hat{\eta}, W h_t^*)\} = W\mathbb{P}_n\{\sum_t m_t(\psi;\hat{\eta}, h_t^*)\} = 0$ would also solve $\mathbb{P}_n\{\sum_t m_t(\psi;\hat{\eta}, h_t^*)\} = 0$ for any nonsingular W.

We now discuss the above approach in more detail, adapting it to the longitudinal causal MSMM and SNMM setting. Suppose that the nuisance functions are estimated with parametric models, so that η_a and η_y are known up to finite-dimensional α and β, respectively, with estimators $\hat{\eta} = (\hat{\alpha}, \hat{\beta})$ solving $\mathbb{P}_n\{S_a(\alpha)\} = \mathbb{P}_n\{S_y(\beta)\} = 0$, for $S_a(\alpha)$ and $S_y(\beta)$ appropriate estimating functions. Let η_0 denote the probability limit of $\hat{\eta}$, where it is assumed that either α_0 or β_0 corresponds to the true value of α or β. Then it is easily seen that influence functions under the restricted space H_t^{res} are given by $\varphi(\psi_0;\ \eta_0,\ W)$ with

$$\varphi(\psi;\eta,W) = \left[W \sum_{t=1}^{K} E\left\{\frac{\partial m_t(\psi;\eta,h^*)}{\partial \psi^{\mathrm{T}}}\right\}\right]^{-1} W \sum_{t=1}^{K} \tilde{m}_t(\psi;\eta,h^*),$$

where $\tilde{m_t}(\psi;\ \eta,\ h) = m_t(\psi;\ \eta,\ h) - E\{\partial m_t(\psi;\ \eta,\ h)/\partial\alpha^{\mathrm{T}}\}E\{\partial S_a(\alpha)/\partial\alpha^{\mathrm{T}}\}^{-1}S_a(\alpha) - E\{\partial m_t(\psi;\ \eta,\ h)/\partial\beta^{\mathrm{T}}\}E\{\partial S_y(\beta)/\partial\beta^{\mathrm{T}}\}^{-1}S_y(\beta)$. Thus for $\hat{\psi}_W$ solving $\mathbb{P}_n\{\sum_t m_t(\psi;\hat{\eta}, W h_t^*)\} = 0$ we have that

$\hat{\psi}_W - \psi_0 = \mathbb{P}_n\{\varphi(\psi_0;\eta_0, W)\} + o_p(1/\sqrt{n})$, and the asymptotic variance of estimator $\hat{\psi_W}$ is given by $E\{\phi(\psi_0;\ \eta_0,\ W)^{\otimes 2}\}$, which clearly depends on the choice of matrix *W*. In the next theorems we give the efficiency bound over the restricted class H_t^{res}, i.e., the asymptotic variance of the most efficient estimator for any choice of *W*.

Theorem 1

Consider the restricted class of functions $H_t^{res}(h_t^*) = \{W h_t^* : W \in \mathbb{R}^{q \times r}\}$ with $h_t^* \in \mathscr{H}_t^r$ fixed and $r > q$, and the corresponding class of restricted estimators solving $\mathbb{P}_n\{\sum_t m_t(\psi;\hat{\eta}, W h_t^*)\} = 0$ with $\eta = (\alpha, \beta) \in \mathbb{R}^d$ and $\hat{\eta}$ converging to probability limit η_0. The efficiency bound for estimators in this restricted class is $\Sigma^{\mathrm{res}}(\psi_0,\ \eta_0)$, where

$$\sum\nolimits^{res}(\psi,\eta) = \left(E\left\{ \sum_{t=1}^{K} \frac{\partial m_t(\psi;\eta,h^*)}{\partial \psi} \right\} E\left[\left\{ \sum_{t=1}^{K} \tilde{m}_t(\psi;\eta,h^*) \right\}^{\otimes 2} \right]^{-1} E\left\{ \sum_{t=1}^{K} \frac{\partial m_t(\psi;\eta,h^*)}{\partial \psi^{\mathrm{T}}} \right\} \right)^{-1}.$$

The (q×r) matrix W_{opt}^{res} that minimizes the asymptotic variance across all restricted estimators is given by $W_{opt}^{res}(\psi_0, \eta_0)$, where

$$W_{opt}^{res}(\psi,\eta) = \sum_{t=1}^{K} E\left\{ \frac{\partial m_t(\psi;\eta,h^*)}{\partial \psi} \right\} E\left[\left\{ \sum_{t=1}^{K} \tilde{m}_t(\psi;\eta,h^*) \right\}^{\otimes 2} \right]^{-1}.$$

In practice the optimal choice of *W* can be estimated with $\hat{W}_{\mathrm{opt}}^{\mathrm{res}} = \mathbb{P}_n\{\sum_t \partial m_t(\hat{\psi};\hat{\eta},h^*)/\partial\psi\}\mathbb{P}_n[\{\sum_t \tilde{m}_t(\hat{\psi};\hat{\eta},h^*)\}^{\otimes 2}]^{-1}$, based on an initial estimator $\hat{\psi}$ solving, for example, $\mathbb{P}_n\{\Sigma_t\, m_t(\psi;\ \eta,\ \hat{h}^*)\} = 0$ for some $h^* \in \mathscr{H}_t^q$. Estimators based on this optimal choice of *W* can be viewed as generalized method of moments estimators, combining estimating functions based on functions h_t^* of dimension $r > q$ (Hansen, 1982; Imbens, 2002).

4. Extended restricted estimation

In this section we propose an extension of the previous adapted estimation approach by optimizing *h* over larger restricted finite-dimensional spaces. Specifically we consider restricted estimation over the extended spaces

$$H_t^{\mathrm{ext}}(h_t^*) = \left\{ W_t h_t^* : W_t \in \mathbb{R}^{q \times r}, h_t^* \in \mathscr{H}_t^r \text{ fixed with } r \geq q \right\}. \quad (2)$$

We will see that these spaces have benefits both in terms of yielding efficiency gains and simplifying practical construction and implementation.

First, note that restricted estimation based on the space H_t^{ext} generalizes the approach from Section 2 based on the space H_t^{res} since the weighting matrices W_t can change with time, allowing more adaptation to the longitudinal data structure. Thus, as before, optimization over this space amounts to finding optimal combinations of estimating functions, but now the combinations are more flexible since they can change with time. When based on the same function h^*, if we take $W_s = W_t$ for all s and t then $H_t^{\text{res}}(h_t^*)=H_t^{\text{ext}}(h_t^*)$ and the above extended restricted space reduces to the previous restricted space H_t^{res}. Thus the restricted space $H_t^{\text{res}}(h_t^*)$ is contained in the extended space $H_t^{\text{ext}}(h_t^*)$ when based on the same function h_t^*, i.e., $H_t^{\text{res}}(h_t^*) \subseteq H_t^{\text{ext}}(h_t^*)$, and the extended space H_t^{ext} thus allows for extra efficiency gains over the restricted space H_t^{res}. Note however that for different choices of h_t^* this nesting may not occur, i.e., if $h_t^* \neq h_t^{**}$ then it may be possible that $H_t^{\text{res}}(h_t^*) \not\subseteq H_t^{\text{ext}}(h_t^{**})$.

The extended restricted space H_t^{ext} can also often be easier to construct in practice than the space H_t^{res}. This is because for the space H_t^{ext} the function h_t^* can be chosen to have the same dimension as ψ (i.e., it is only required that $r \geq q$, not $r > q$ as with H_t^{res}), which means one can use the same function $h_t^* \in \mathscr{H}_t^q$ that is required to compute a standard estimator (e.g., $h_t^*=\partial g_t/\partial\psi$ or $h_t^*=\partial\gamma_t/\partial\psi$). We can have $r = q$ in the extended setting because even then the matrices cannot be factored out of the estimating equations; thus we still obtain different estimators with different choices of W_t. This is due to the fact that the dimension of all the matrices taken together $(W_1, \ldots, W_K)$ is $\mathbb{R}^{q\times qK}$, which is larger than $\mathbb{R}^{q\times q}$ as long as we have longitudinal data so that $K > 1$. In contrast, as discussed earlier, when constructing the space H_t^{res} from Section 3, the analyst needs to augment the function h_t^* with an additional function of dimension $r - q > 0$.

Now we will consider some theoretical properties of the extended space H_t^{ext}. As in the previous section, assume that the nuisance functions are estimated with $\hat{\eta} = (\hat{\alpha}, \hat{\beta}) \in \mathbb{R}^d$ solving $\mathbb{P}_n\{S_a(\alpha)\} = \mathbb{P}_n\{S_y(\beta)\} = 0$, for $S_a(\alpha)$ and $S_y(\beta)$ appropriate estimating functions. Also let $W = (W_1, \ldots, W_K) \in \mathbb{R}^{q\times rK}$, $m=(m_1^{\text{T}}, \ldots, m_K^{\text{T}})^{\text{T}}$, and $\tilde{m}=(\tilde{m}_1^{\text{T}}, \ldots, \tilde{m}_K^{\text{T}})^{\text{T}}$ with $m_t(\psi;\ \eta, h)$ as defined earlier, so that m and $\tilde{m}$ are rK-dimensional vectors. Then influence functions under the extended restricted space H_t^{ext} are given by

$$\varphi(\psi;\eta,W)=\left[\sum_{t=1}^{K} W_t E\left\{\frac{\partial m_t(\psi;\eta,h^*)}{\partial\psi^{\text{T}}}\right\}\right]^{-1}\sum_{t=1}^{K} W_t\tilde{m}_t(\psi;\eta,h^*)=\left[WE\left\{\frac{\partial m(\psi;\eta,h^*)}{\partial\psi^{\text{T}}}\right\}\right]^{-1}W\tilde{m}(\psi;\eta,h^*).$$

In the next theorem we give the efficiency bound for estimators with influence functions of the above form, along with the optimal choice of the matrix W that yields an estimator that attains the efficiency bound.

Theorem 2

Consider the restricted class of functions $H_t^{ext}(h_t^*)=\{W_t h_t^*: W_t \in \mathbb{R}^{q\times r}\}$ with $h_t^* \in \mathscr{H}_t^r$ fixed and $r \geq q$, and the corresponding class of restricted estimators solving

$\mathbb{P}_n\{\sum_t m_t(\psi;\hat{\eta}, W_t h_t^*)\}=0$ with $\eta = (\alpha, \beta) \in \mathbb{R}^d$ and $\hat{\eta}$ converging to probability limit η_0. The efficiency bound for estimators in this restricted class is $\Sigma^{\text{ext}}(\psi_0, \eta_0)$, where

$$\sum\nolimits^{ext}(\psi,\eta)=\left[E\left\{\frac{\partial m(\psi;\eta,h^*)}{\partial\psi}\right\}E\left\{\tilde{m}(\psi;\eta,h^*)^{\otimes 2}\right\}^{-1}E\left\{\frac{\partial m(\psi;\eta,h^*)}{\partial\psi^{\mathrm{T}}}\right\}\right]^{-1}.$$

The (q×rK) matrix W_{opt}^{ext} that minimizes the asymptotic variance across all restricted estimators is given by $W_{opt}^{ext}(\psi_0,\eta_0)$, where

$$W_{opt}^{ext}(\psi,\eta)=E\left\{\frac{\partial m(\psi;\eta,h^*)}{\partial\psi}\right\}E\left\{\tilde{m}(\psi;\eta,h^*)^{\otimes 2}\right\}^{-1}.$$

Note that since $H_t^{\text{res}}(h_t^*) \subseteq H_t^{\text{ext}}(h_t^*)$ it immediately follows that the efficiency bound from Theorem 1 is no less than that from Theorem 2, i.e., $\Sigma^{\text{res}}(\psi_0, \eta_0) \geq \Sigma^{\text{ext}}(\psi_0, \eta_0)$. However, when the spaces are based on different functions h_t^*, this inequality may not necessarily hold.

As before, the optimal choice of W can be estimated with

$\hat{W}_{\text{opt}}^{\text{ext}}=\mathbb{P}_n\{\partial m(\hat{\psi};\hat{\eta},h^*)/\partial\psi\}\mathbb{P}_n\{\tilde{m}(\hat{\psi};\hat{\eta},h^*)^{\otimes 2}\}^{-1}$, based on an initial estimator $\hat{\psi}$ solving, for example, $\mathbb{P}_n\{\Sigma_t m_t(\psi; \eta, \hat{h}^*)\} = 0$ for some $h^* \in \mathscr{H}_t^q$. And again this estimator can be viewed as a generalized method of moments estimator, now additionally combining estimating functions across timepoints. The above estimator based on $\hat{W}_{\text{opt}}^{\text{ext}}$ is straightforward to use in practice because it only depends on simple sample averages of the estimating functions m and $\tilde{m}$; however, alternative estimators are available as well. These alternatives include iterated versions of the above and empirical likelihood estimators, which can have advantageous finite sample properties (Imbens, 2002). These could be particularly valuable for optimal restricted estimators based on the space H_t^{ext}, since such estimators optimize over larger matrices $W = (W_1, \ldots, W_K)$.

5. Simulation study

To investigate finite-sample properties of our proposed approach, we simulated data from the structural nested model given by

$$\begin{gathered}Y_{K+1}^0 \sim N(\theta,\sigma^2), \quad Y_t|X_{t-1},A_{t-1},Y_{K+1}^0 \sim N(\beta_0+\beta_1A_{t-1}+\beta_2Y_{t-1}+\beta_3A_{t-2}+\beta_4Y_{K+1}^0,\nu^2), \\ A_t|X_t \sim N(\alpha_0+\alpha_1Y_t+\alpha_2A_{t-1}+\alpha_3Y_{t-1},\tau^2), \quad \text{and} \quad Y_{K+1}=Y_{K+1}^0+\sum_{t=0}^{K}\frac{A_t\{\psi_0+\psi_1(Y_t-c)\}}{K+1-t}\end{gathered}$$

for $t = 1, \ldots, K$. Under this model we have $\gamma_t(X_t, A_t; \psi) = A_t\{\psi_0 + \psi_1(Y_t - c)\}/(K + 1 - t)$; this means that the effect of treatment on the outcome is inversely proportional to how long before the end of the study the treatment was given, and that this effect is modified by current outcome values. We chose parameter values $\theta = 29$, $\sigma = 4.5$, $\beta = (2, 1, 0.75, -0.75, 0.2)$, $\nu = 3$, $\alpha = (11, -0.15, 0.5, -0.05)$, $\tau = 1$, $\psi = (1, -0.1)$, and $c = 40$, with $Y_t \sim N(36, 4.5^2)$ and $A_t \sim N(7.5, 1)$ for $t < 1$. These parameters yield data that approximately match real claims data from the United States Renal Data System, where A_t is the log-dose of erythropoietin at time t and Y_t is hematocrit level at time t. We varied the sample size n and the number of time points K in simulations, and generated 1000 datasets at each setting.

We considered two methods for estimating ψ: a standard approach and restricted estimation. The standard approach used $h_t(X_t, A_t) = \partial\gamma_t(X_t, A_t; \psi)/\partial\psi = A_t(1, Y_t - c)^{\mathrm{T}}/(K + 1 - t)$, while restricted estimation used H_t^{ext} as in Section 4 with $h_t^*(X_t, A_t) = \partial\gamma_t(X_t, A_t; \psi)/\partial\psi$. The matrix W^{opt} was estimated as discussed in the previous section, using the standard approach to compute the initial estimator of ψ. Both approaches rely on correctly-specified models for $p(a_t \mid x_t; \alpha)$, and are doubly robust by modeling the quantity $E\left\{Y_{K+1} - \sum_{s=t}^{K} \gamma_s(X_s, A_s; \psi) | X_t\right\}$ with simple working models given by $\beta_{0t} + \beta_1 Y_t + \beta_2 A_{t-1} + \beta_3 Y_{t-1}$. Results are given in Table 1.

Restricted estimation gave better efficiency for every combination of n and K that we explored, with gains in RMSE relative to the standard estimator ranging from 1% to over 60%. In this simulation, gains were larger for the effect modification parameter ψ_1 than for the main effect parameter ψ_0. Further, in terms of RMSE, restricted estimation was most beneficial in studies with more timepoints and when sample sizes were not too large. For illustration, the estimated optimal weight matrix at the median simulation setting ($n = 1000$ and $K = 4$), averaged across simulations, was given by $\hat{W}^{\mathrm{opt}} = (\hat{W}_1, \ldots, \hat{W}_4)^{\mathrm{T}}$ with

$$\hat{W}_1 = \begin{pmatrix} 0.084 & 0.000 \\ 0.732 & 0.047 \end{pmatrix}, \hat{W}_2 = \begin{pmatrix} 0.096 & 0.000 \\ 0.874 & 0.050 \end{pmatrix}, \hat{W}_3 = \begin{pmatrix} 0.100 & 0.000 \\ 0.723 & 0.058 \end{pmatrix}, \hat{W}_4 = \begin{pmatrix} 0.063 & 0.000 \\ 0.004 & 0.061 \end{pmatrix}.$$

The average weight matrices at the first three times were similar, each having one large off-diagonal element and somewhat dissimilar diagonal elements, while the weight matrix at the fourth time had small off-diagonal elements and similar diagonal elements. This suggests that the efficiency benefits do not necessarily come from simply reweighting across time with scalar matrices (i.e., using $W_t = c_t I$ for I the identity matrix); rather, there appears to be additional restructuring of the estimating functions. This phenomenon may account for the larger increase in efficiency for the effect modification parameter.

Restricted estimation gave some finite-sample bias in studies with six timepoints, but in terms of RMSE these biases were more than offset by decreases in variance. Such biases were expected based on results from the generalized method of moments literature, and could potentially be mitigated with empirical likelihood or bias-corrected approaches (Imbens, 2002).

6. Discussion

In this paper we have discussed restricted estimation approaches for developing more efficient estimators of causal effects in longitudinal studies, where local efficiency can be a prohibitively difficult goal. We adapted the approaches developed by Tsiatis (2006) and Tan (2011) to novel longitudinal causal settings involving both marginal structural models and structural nested models. We also developed an extended approach that allows for more adaptation to the longitudinal data structure, and derived efficiency bounds and simple but optimal estimators. We illustrated our methods in a simulation experiment, which showed potential for large gains in efficiency, albeit sometimes at the cost of some finite sample bias. Based on simulations as well as higher-order asymptotic theory from the generalized method of moments literature (Imbens, 2002), our hypothesis is that finite-sample biases are more likely to arise in settings with smaller sample sizes and larger numbers of timepoints (or in general when there are more parameters in the weight matrices).

Future work will explore approaches for alleviating finite-sample bias, for example based on alternatives to simple two-step estimators (e.g., empirical likelihood). We also hope to develop computationally feasible but accurate confidence interval estimators, particularly for settings with many timepoints. Finally, it will be very useful to apply restricted estimation methodology to MSMMs, which are more popular and widely used than SNMMs, but which can have more severe issues with low efficiency.

Acknowledgments

The authors thank an anonymous referee for helpful comments and suggestions. Edward Kennedy and Marshall Joffe were supported by NIH grant R01-DK090385, and Dylan Small was supported by NSF grant SES-1260782.

References

Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982; 50 (4):1029–1054.

Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. Statistics in Medicine. 2002; 21 (12):1689–1709. [PubMed: 12111906]

Imbens GW. Generalized method of moments and empirical likelihood. Journal of Business & Economic Statistics. 2002; 20 (4):493–506.

Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. Health Service Research Methodology: A Focus on AIDS. 1989:113–159.

Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics - Theory and Methods. 1994; 23 (8):2379–2412.

Robins, JM. Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer; 2000. Marginal structuralmodels versus structural nestedmodels as tools for causal inference; p. 95-133.

Tan Z. Efficient restricted estimators for conditional mean models with missing data. Biometrika. 2011; 98 (3):663–684.

Tsiatis, AA. Semiparametric Theory and Missing Data. Springer; 2006.

van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. Springer; 2003.

Appendix A

The following proof uses the same logic as that in Hansen (1982). For Theorem 1 (restricted estimation using the space H_t^{res}), let $\Delta(\psi, \eta) = E\{\Sigma_t\ \partial m_t(\psi; \eta, h^*)/\partial \psi^{\mathrm{T}}\}$ and $\Omega(\psi, \eta) = E[\{\Sigma_t\ m_t(\tilde{\psi}; \eta, h^*)\}^{\otimes 2}]$. For Theorem 2 (restricted estimation using the space H_t), let $\Delta(\psi, \eta) = E\{\partial m(\psi; \eta, h^*)/\partial \psi^{\mathrm{T}}\}$ and $\Omega(\psi, \eta) = E\{m(\tilde{\psi}; \eta, h^*)^{\otimes 2}\}$, with m and m ãs defined in the main text. For both theorems, we let $\Delta = \Delta(\psi_0, \eta_0)$ and $\Omega = \Omega(\psi_0, \eta_0)$.

For restricted estimators $\hat{\psi}_W$ based on a general W, with influence functions as in the main text, we have

$$\sqrt{n}(\hat{\psi}_W - \psi_0) \xrightarrow{d} N\left(0, (W\Delta)^{-1} W \Omega W^{\mathrm{T}} (\Delta^{\mathrm{T}} W^{\mathrm{T}})^{-1}\right).$$

We proceed by considering the difference between the above asymptotic variance for a general restricted estimator and the proposed efficiency bound given by $(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1}$. This difference can be written as QQ^{T}, where

$$Q = (W\Delta)^{-1} W \Gamma - (\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1} \Delta^{\mathrm{T}} (\Gamma^{\mathrm{T}})^{-1}$$

with $\Gamma\Gamma^{\mathrm{T}} = \Omega$ the Cholesky decomposition of the symmetric variance matrix Ω. Since QQ^{T} is positive semi-definite by construction, the matrix $(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1}$ corresponds to the minimum possible variance for any choice of W.

To prove that $W^{\text{opt}} = \Delta^{\mathrm{T}}\Omega^{-1}$, we will show that $QQ^{\mathrm{T}} = 0$ if and only if $W = \Delta^{\mathrm{T}}\Omega^{-1}$. First assume $QQ^{\mathrm{T}} = 0$. Then

$$0 = (W\Delta) Q \Gamma^{-1} = (W\Delta)\left\{(W\Delta)^{-1} W\Gamma - (\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1}\Delta^{\mathrm{T}}(\Gamma^{\mathrm{T}})^{-1}\right\}\Gamma^{-1} = W - W\Delta(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1}\Delta^{\mathrm{T}}\Omega^{-1},$$

which implies that $W = \Delta^{\mathrm{T}}\Omega^{-1}$ up to a scaling constant. Now assume $W = \Delta^{\mathrm{T}}\Omega^{-1}$. Then QQ^{T} can be written as

$$\begin{aligned}
&(W\Delta)^{-1} W\Omega W^{\mathrm{T}} (\Delta^{\mathrm{T}} W^{\mathrm{T}})^{-1} \\
&\quad -(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1} \\
&\quad =(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1}\Delta^{\mathrm{T}}\Omega^{-1}\Omega\Omega^{-1}\Delta(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1} \\
&\quad -(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1} \\
&\quad =(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1} \\
&\quad -(\Delta^{\mathrm{T}}\Omega^{-1}\Delta)^{-1},
\end{aligned}$$

which equals zero. Therefore the minimum variance is in fact achieved when $W = W^{\text{opt}} = \Delta^{T}\Omega^{-1}$.

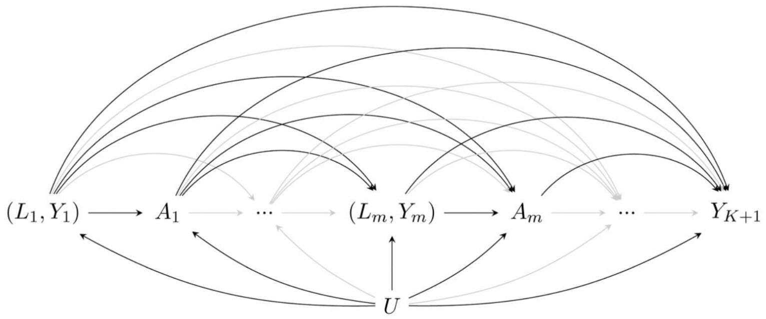

Figure 1.

Directed acyclic graph of data structure assuming only time ordering.

Table 1

Simulation results (across 1000 simulations)

		Standard Approach				**Restricted Estimation**					
		Main effect ψ_0		**Interaction ψ_1**		**Main effect ψ_0**		**Interaction ψ_1**		**% Decrease in RMSE**	
K	***n***	**% Bias**	**SE**	**% Bias**	**SE**	**% Bias**	**SE**	**% Bias**	**SE**	ψ_0	ψ_1
2	500	−0.6	3.49	−1.3	1.04	−0.3	3.38	0.1	0.62	3.1	40.8
	1000	0.4	3.38	−1.0	0.95	−0.1	3.30	−0.6	0.59	2.3	38.0
	5000	−0.1	3.25	0.6	0.93	−0.2	3.21	0.5	0.59	1.3	37.1
4	500	1.5	3.65	−3.0	1.28	−1.7	2.81	5.7	0.45	22.5	63.4
	1000	0.3	3.30	−1.8	0.94	−1.4	3.04	3.0	0.47	6.8	49.1
	5000	0.1	2.99	−0.3	0.86	−0.5	2.89	1.1	0.48	2.5	43.5
6	500	1.1	4.34	−6.7	1.42	−5.2	2.59	10.6	0.39	34.8	68.0
	1000	0.5	3.17	−1.6	0.86	−2.8	2.72	6.6	0.41	10.0	47.0
	5000	0.1	2.96	−0.1	0.77	−0.7	2.76	2.0	0.43	5.2	40.7

Note: SE = standard error (scaled by $n^{1/2}$), RMSE = root mean squared error