Optimal balancing of time-dependent confounders for marginal structural models

European Causal Inference Meeting

Michele Santacatterina[1]
michele.santacatterina@ki.se

Nathan Kallus[2]
kallus@cornell.edu

[1]Unit of Biostatistics, Karolinska Institutet
[2]Operations Research and Information Engineering and Cornell Tech, Cornell University

April 11, 2018

Marginal structural models I

Marginal structural models (MSM) have been used to estimate the causal effect of a time-varying treatment on an outcome of interest with longitudinal data in observational studies.

- MSM control for time-dependent confounders, which are confounders that are affected by previous treatment and affect future ones.
- MSM consistently estimate the causal effect of a time-varying treatment via inverse probability of treatment weighting (IPTW).

Marginal structural models II

Despite their theoretical appeal, these methods have two main limitations:

- ✗ Highly sensitive to the misspecification of the treatment assignment/propensity score model.
- ✗ Practical violations of the positivity assumption: extreme weights, erroneous inference, low precision.

Kernel optimal weighting

Kernel Optimal Weighting (KOW) simultaneously balances time-dependent confounders and control for the precision of the resulting MSM.

- Define imbalance as the sum of absolute empirical discrepancies between the weighted observed data and the counterfactuals of interest.
- Minimize imbalance over all possible realizations of some unknown functions.
- Regularize the weights in such a way that the precision of the resulting MSM is controlled.
- Use kernels and quadratic programming to compute weights that optimally balance time-dependent confounders and control for precision.

Defining imbalance I

One time period: assuming consistency, positivity and ignorability, we can show that

$$\begin{aligned}
\mathbb{E}\left[W\mathbb{1}[A_1 = a_1]Y\right] &= \mathbb{E}\left[W\mathbb{1}[A_1 = a_1]Y(a_1)\right] \\
&= \mathbb{E}\left[W\mathbb{E}\left[\mathbb{1}[A_1 = a_1]Y(a)|X_1\right]\right] \\
&= \mathbb{E}\left[W\mathbb{E}\left[\mathbb{1}[A_1 = a_1]|X_1\right]\mathbb{E}\left[Y(a_1)|X_1\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y(a_1)|X_1\right]\right] + \delta_{a_1}^{(1)}(W, g_{a_1}^{(1)}) \\
&= \mathbb{E}\left[Y(a_1)\right] + \delta_{a_1}^{(1)}(W, g_{a_1}^{(1)})
\end{aligned}$$

where,

$$\delta_{a_1}^{(1)}(W, g_{a_1}^{(1)}) = \mathbb{E}\left[W\mathbb{1}[A_1 = a_1]g_{a_1}^{(1)}(X_1)\right] - \mathbb{E}\left[g_{a_1}^{(1)}(X_1)\right]$$

Define imbalance as

$$\mathrm{IMB}(W; (g_{a_1}^{(1)})_{a_1 \in \{0,1\}}) = \sum_{a_1 \in \{0,1\}} \left|\hat{\delta}_{a_1}^{(1)}(W, g_{a_1}^{(1)})\right|.$$

Defining imbalance II

$T > 1$ time periods: assuming consistency, positivity and sequential ignorability, we can show that

$$\mathbb{E}\left[W\mathbb{1}[\overline{A}=\overline{a}]Y\right] - \mathbb{E}\left[Y(\overline{a})\right] = \sum_{t=1}^{T} \delta_{a_t}^{(t)}(W, g_{\overline{a}}^{(t)}).$$

where,

$$\begin{aligned}
\delta_{a_t}^{(t)}(W, g_{\overline{a}}^{(t)}) &= \mathbb{E}\left[W\mathbb{1}[A_t = a_t] g_{\overline{a}}^{(t)}(\overline{A}_{t-1}, \overline{X}_t)\right] - \mathbb{E}\left[W g_{\overline{a}}^{(t)}(\overline{A}_{t-1}, \overline{X}_t)\right] \\
g_{\overline{a}}^{(t)}(\overline{A}_{t-1}, \overline{X}_t) &= \mathbb{1}[\overline{A}_{t-1} = \overline{a}_{t-1}]\mathbb{E}\left[Y(\overline{a})|\overline{X}_t\right] \\
\delta_{a_1}^{(1)}(W, g_{\overline{a}}^{(1)}) &= \mathbb{E}\left[W\mathbb{1}[A_1 = a_1] g_{\overline{a}}^{(1)}(X_1)\right] - \mathbb{E}\left[g_{\overline{a}}^{(1)}(X_1)\right] \\
g_{\overline{a}}^{(1)}(X_1) &= \mathbb{E}\left[Y(\overline{a})|X_1\right].
\end{aligned}$$

Define imbalance as

$$\mathrm{IMB}(W; (g_{\overline{a}}^{(t)})_{t\in\{1,\dots,T\},\overline{a}\in\mathcal{A}}) = \sum_{\overline{a}\in\mathcal{A}} \left|\sum_{t=1}^{T} \hat{\delta}_{a_t}^{(t)}(W, g_{\overline{a}}^{(t)})\right|,$$

Squared worst case imbalance

We find weights that minimize imbalance over all possible realizations of the unknown functions $g_{\overline{a}}^{(t)}$ to which this quantity depends on. Since unknown, we want to limit the "size" of these functions by guarding against any of their possible realizations. We define,

$$\Delta_{a_t}^{(t)}(W) = \sup_{\|g_{\overline{a}}^{(t)}\|_t^2 \leq 1} \hat{\delta}_{a_t}^{(t)}(W, g_{\overline{a}}^{(t)})$$

Then the normalized squared worst case imbalance is

$$\begin{aligned}\mathcal{B}^2(W) &= \sup_{\sum_{t \in \{1,\ldots,T\}, \overline{a} \in \mathcal{A}} \|g_{\overline{a}}^{(t)}\|_t^2 \leq 1} \frac{1}{|\mathcal{A}|} \mathrm{IMB}^2(W; (g_{\overline{a}}^{(t)})_{t \in \{1,\ldots,T\}, \overline{a} \in \mathcal{A}}) \\ &= \frac{1}{2} \sum_{t=1}^{T} (\Delta_0^{(t)}(W)^2 + \Delta_1^{(t)}(W)^2).\end{aligned}$$

This highlights that are only linearly-many imbalances that we need to account for: *two* for each time period.

Minimizing imbalance while controlling precision

We can obtain minimal imbalance by minimizing $\mathcal{B}^2(W)$. However, the resulting weights can be highly variable, leading to extreme weights which in turn yield erratic inferences and low precision.
We propose to find weights that minimizes $\mathcal{B}^2(W)$ plus a penalty.

$$\begin{aligned}\min_{w \in \mathcal{W}} \quad \mathcal{C}(W, \lambda) &= \frac{1}{2} \sum_{t=1}^{T} \left(\Delta_0^{(t)}(W)^2 + \Delta_1^{(t)}(W)^2 + \lambda_t \| W - e \|_2^2 \right) \\ &= \mathcal{B}^2(W) + \lambda \| W - e \|_2^2,\end{aligned} \tag{1}$$

RKHS and quadratic programming I

Recall that, $g_{\overline{a}}^{(t)}(\overline{A}_{t-1}, \overline{X}_t) = \mathbb{1}[\overline{A}_{t-1} = \overline{a}_{t-1}]\mathbb{E}[Y(\overline{a})|\overline{X}_t]$.

- If $\|\cdot\|_t$ is a reproducing kernel Hilbert space norm given by the kernel $\mathcal{K}_t$, then we can express $\Delta_{a_t}^{(t)}(W)$ as a convex quadratic function in W

Let us define the matrix $K_t \in \mathbb{R}^{n \times n}$ as $K_{tij} = \mathcal{K}_t((\overline{A}_{i,t-1}, \overline{X}_{it}), (\overline{A}_{j,t-1}, \overline{X}_{it}))$ and note that it is positive semidefinite. Then,

$$
\begin{aligned}
\Delta_{a_t}^{(t)}(W)^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(W_i\mathbb{I}[A_{it} = a_t] - W_i)(W_j\mathbb{I}[A_{it} = a_t] - W_j)K_{tij} \\
&= \frac{1}{n^2}(I_{a_t}^{(t)}W - W)^T K_t(I_{a_t}^{(t)}W - W) \\
&= \frac{1}{n^2}W^T((I_{a_t}^{(t)} - I)K_t(I_{a_t}^{(t)} - I))W,
\end{aligned}
$$

RKHS and quadratic programming II

Let $K_t^\circ = I_0^{(t)} K_t I_0^{(1)} + I_1^{(t)} K_t I_1^{(t)}$ for $t = 1$, and $K_t^\circ = (I_0^{(t)} - I) K_t (I_0^{(t)} - I) + (I_1^{(t)} - I) K_t (I_1^{(t)} - I)$ for $t \geq 2$, which are given by setting every entry i, j of K_t to 0 whenever $A_{it} \neq A_{jt}$, and let $K = \sum_{t=2}^T K_t$, and $K^\circ = K_1^\circ + \sum_{t=2}^T K_t^\circ$. We then get that

$$\begin{aligned}\mathcal{B}^2(W) &= \frac{1}{n^2} \frac{1}{2} \sum_{t=1}^{T} (\Delta_0^{(t)}(W)^2 + \Delta_1^{(t)}(W)^2) \\ &= \frac{1}{n^2} (\frac{1}{2} W^T K^\circ W - e^T K_1 W + e^T K_1 e).\end{aligned}$$

Finally, we obtain weights that optimally balance covariates to control for time-dependent confounding while controlling precision by solving the following quadratic optimization problem,

$$\min_{w \in \mathcal{W}} \quad \frac{1}{n^2} (\frac{1}{2} W^T K_\lambda^\circ W - e^T K_\lambda W + e^T K_\lambda e) \qquad (2)$$

where $K_\lambda^\circ = K^\circ + \lambda I$, $K_\lambda = K_1 + \lambda I$ and $\lambda = \sum_{t=1}^T \lambda_t$.

Practical guidelines

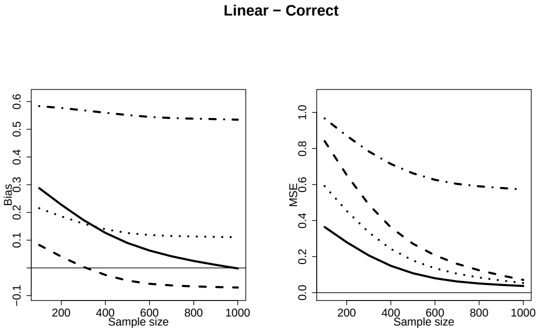
Solutions to the quadratic problem (2) depend on several factors. They depend on

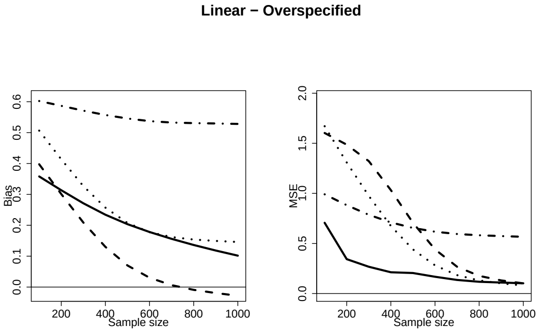
1) *the chosen kernel for the treatment history and that for the time-invariant confounders and the history of time-dependent confounders.*
2) *the estimated values for the kernels' hyperparameters, and, those for the penalization parameters λ_t for all $t = 1, \ldots, T$.*
3) *the chosen set of time-invariant and time-dependent confounders.*

We suggest to

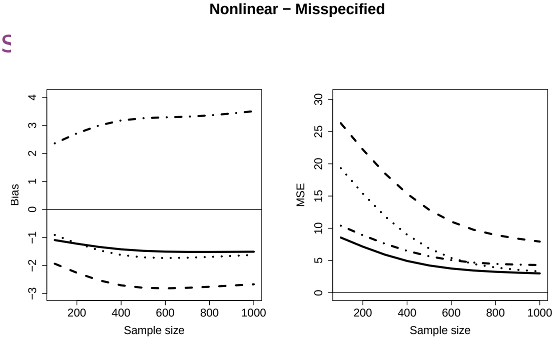
1) use a linear kernel for the treatment history and a polynomial kernel of degree $d > 1$ for the time-invariant confounders and the history of time-dependent confounders.
2) obtain the kernels' hyperparameters and the penalization parameter λ by postulating a Gaussian process and minimizing the negative log marginal likelihood.
3) include all possible time-dependent and time-invariant confounders when specifying the kernels

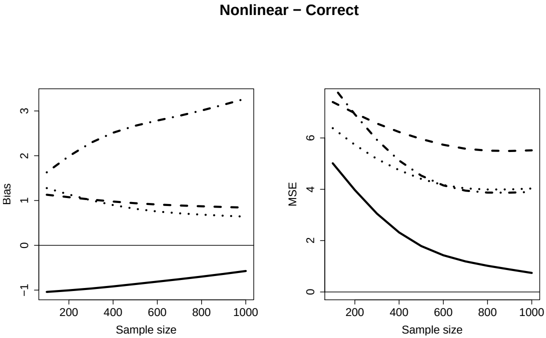
Simulations





Simulations





KOW with informative censoring

Assuming consistency, positivity, sequential ignorability, and ignorable censoring, we can show that

$$\mathbb{E}\left[W\mathbb{1}[\overline{A}=\overline{a}]\mathbb{1}[\overline{C}=\overline{0}]Y\right]-\mathbb{E}\left[Y(\overline{a})\right] \quad =\sum_{t=1}^{T}\delta_{a_t,c_t}^{(t)}(W,g_{\overline{a}}^{(t)}),$$

We therefore define imbalance as

$$\mathrm{IMB}(W;(g_{\overline{a}}^{(t)})_{t\in\{1,\ldots,T\},\overline{a}\in\mathcal{A},c_t\in\{0,1\}})=\sum_{\overline{a}\in\mathcal{A}}\left|\sum_{t=1}^{T}\sum_{c_t\in\{0,1\}}\hat{\delta}_{a_t,c_t}^{(t)}(W,g_{\overline{a}}^{(t)})\right|,$$

and, the normalized squared worst case imbalance becomes

$$\mathcal{B}^2(W)=\frac{1}{2}\sum_{t=1}^{T}(\Delta_{0,0}^{(t)}(W)^2+\Delta_{1,0}^{(t)}(W)^2+\Delta_{0,1}^{(t)}(W)^2+\Delta_{1,1}^{(t)}(W)^2).$$

The effect of HIV treatment on time to death

MSM have been used to estimate the causal effect of a time-varying treatment on time to death among people who live with HIV in the presence of time-dependent confounding.

- Using real-world data from the Multicenter AIDS Cohort Study (MACS), we estimated the parameters of the MSM by KOW.
- We compared KOW with IPTCW and stable IPTCW (sIPTCW).

Table: Estimated hazard ratio of the effect of HIV treatment initiation on time to death.

	KOW		Logistic	
	$\mathcal{K}_1$	$\mathcal{K}_2$	IPTCW	sIPTCW
$\hat{HR}$	0.38*	0.50*	0.14	1.25
SE	(0.33)	(0.30)	(1.15)	(0.30)

Note: $\hat{HR}$ is the estimated hazard ratio of the effect of HIV treatment initiation on time to death. SE is the estimated robust standard error. Weights were obtained as: KOW ($\mathcal{K}_1$) a product of two linear kernels, one for the treatment history and one for the time-invariant and the history of time-dependent confounders; KOW ($\mathcal{K}_2$) a product between a linear kernel for the treatment history and a polynomial kernel of degree 2 for the time-invariant and the history of time-dependent confounders. IPTCW: by using a logistic regression inverse probability of treatment and censoring weighting. sIPTCW: by using stable logistic regression IPTCW. * indicates statistical significance at the 0.05 level.

Conclusions

Unlike other methods, KOW simultaneously

- ✓ improves covariate balance, which provides more robust estimates of the causal effect, and
- ✓ controls for extreme weights, which provides more precise inferences.

KOW has several attractive characteristics.

- Mitigates the effects of possible misspecification of the treatment model by directly balancing covariates and control for precision by penalizing extreme weights.
- Allows balancing non-additive covariate relationships by using kernels to generalize the structure of conditional expectation functions.
- Can be easily generalized to other settings, such as informative censoring.
- Needs to minimizes a number of imbalances that grows linearly (and not exponentially) in the number of time periods.