Kernel Methods for Unobserved Confounding

Algorithm Derivation

Presenter: Yibin Xiong

June 21, 2022

Parameter Identification Results [1]

Theorem 4.1 (Representation via kernel mean embedding)**.** *Suppose the conditions of Theorem 3.1 hold. Further suppose Assumption 4.1 holds, $\gamma_0 \in \mathcal{H}_{RF}$, and $h_0 \in \mathcal{H}$. Then*

$$\gamma_0(d,x,z) = \langle h_0, \phi(d) \otimes \phi(x) \otimes \mu_w(d,x,z) \rangle_{\mathcal{H}} \text{ where } \mu_w(d,x,z) := \int \phi(w)\mathbb{P}(w|d,x,z)$$

Moreover

1. $\theta_0^{ATE}(d) = \langle h_0, \phi(d) \otimes \mu \rangle_{\mathcal{H}}$ *where* $\mu := \int [\phi(x) \otimes \phi(w)]\mathbb{P}(x,w)$
2. $\theta_0^{DS}(d, \tilde{\mathbb{P}}) = \langle h_0, \phi(d) \otimes \nu \rangle_{\mathcal{H}}$ *where* $\nu := \int [\phi(x) \otimes \phi(w)]\tilde{\mathbb{P}}(x,w)$
3. $\theta_0^{ATT}(d,d') = \langle h_0, \phi(d') \otimes \mu(d) \rangle_{\mathcal{H}}$ *where* $\mu(d) := \int [\phi(x) \otimes \phi(w)]\mathbb{P}(x,w|d)$
4. $\theta_0^{CATE}(d,v) = \langle h_0, \phi(d) \otimes \phi(v) \otimes \mu(v) \rangle_{\mathcal{H}}$ *where* $\mu(v) := \int [\phi(x) \otimes \phi(w)]\mathbb{P}(x,w|v)$

Step 1: Estimate Conditional Mean Embedding μ

By Singh's KIV paper [2] Algorithm 1,

$$\hat{\mu}_w(d,x,z) = \sum_{i=1}^{n} \beta_i(d,x,z)\phi(w_i)$$

where

$$\beta(d,x,z) = (K_{DD} \odot K_{XX} \odot K_{ZZ} + n\lambda I)^{-1}[K_{Dd} \odot K_{Xx} \odot K_{Zz}] \in \mathbb{R}^n$$

Step 2: Estimate Confounding Bridge h by regression

- To estimate h, consider the kernel ridge regression problem

$$h = \arg\min_{h \in \mathcal{H}} \mathcal{E}_\xi^n(h) \text{ (recall } \mathcal{H} := \mathcal{H}_D \otimes \mathcal{H}_X \otimes \mathcal{H}_W)$$

$$\mathcal{E}_\xi^n(h) = \frac{1}{n} \sum_{i=1}^{n} \| y_i - \langle h, \phi(d_i) \otimes \phi(x_i) \otimes \hat{\mu}(d_i, x_i, z_i) \rangle_{\mathcal{H}} \|_{\mathcal{Y}}^2 + \xi \|h\|_{\mathcal{H}}^2$$

Step 2: Estimate Confounding Bridge h by regression

▷ To estimate h, consider the kernel ridge regression problem

$$h = \arg\min_{h \in \mathcal{H}} \mathcal{E}_{\xi}^{n}(h) \text{ (recall } \mathcal{H} := \mathcal{H}_D \otimes \mathcal{H}_X \otimes \mathcal{H}_W)$$

$$\mathcal{E}_{\xi}^{n}(h) = \frac{1}{n} \sum_{i=1}^{n} \| y_i - \langle h, \phi(d_i) \otimes \phi(x_i) \otimes \hat{\mu}(d_i, x_i, z_i) \rangle_{\mathcal{H}} \|_{\mathcal{Y}}^2 + \xi \|h\|_{\mathcal{H}}^2$$

- Due to the ridge penalty, the stated objective is **coercive** and strongly convex with respect to h. Hence it has a unique minimizer that **obtains** the minimum.
- write $\hat{h} = \hat{h}_n + \hat{h}_n^{\perp}$, where $\hat{h}_n \in \text{span}\{\phi(d_i) \otimes \phi(x_i) \otimes \phi(w_i)\}$
- Note that $\hat{\mu}(d_i, x_i, z_i) \in \text{span}\{\phi(w_i)\}$

Step 2: Estimate Confounding Bridge h by regression

$$
\begin{aligned}
\mathcal{E}_\xi^n(\hat{h}) &= \frac{1}{n}\sum_{i=1}^{n} \|y_i - \langle \hat{h}_n + \hat{h}_n^\perp, \phi(d_i)\otimes\phi(x_i)\otimes\hat{\mu}(d_i,x_i,z_i)\rangle_{\mathcal{H}}\|_{\mathcal{Y}}^2 + \xi\|\hat{h}_n + \hat{h}_n^\perp\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|y_i - \langle \hat{h}_n, \phi(d_i)\otimes\phi(x_i)\otimes\hat{\mu}(d_i,x_i,z_i)\rangle_{\mathcal{H}} - 0\|_{\mathcal{Y}}^2 + \xi\|\hat{h}_n\|_{\mathcal{H}}^2 + \xi\|\hat{h}_n^\perp\|_{\mathcal{H}}^2 \\
&= \mathcal{E}_\xi^n(\hat{h}_n) + \xi\|\hat{h}_n^\perp\|_{\mathcal{H}}^2
\end{aligned}
$$

which implies $\mathcal{E}_\xi^n(\hat{h}) \geq \mathcal{E}_\xi^n(\hat{h}_n)$.

Step 2: Estimate Confounding Bridge h by regression

$$
\begin{aligned}
\mathcal{E}_\xi^n(\hat{h}) &= \frac{1}{n}\sum_{i=1}^{n} \|y_i - \langle \hat{h}_n + \hat{h}_n^\perp, \phi(d_i)\otimes\phi(x_i)\otimes\hat{\mu}(d_i,x_i,z_i)\rangle_{\mathcal{H}}\|_{\mathcal{Y}}^2 + \xi\|\hat{h}_n + \hat{h}_n^\perp\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|y_i - \langle \hat{h}_n, \phi(d_i)\otimes\phi(x_i)\otimes\hat{\mu}(d_i,x_i,z_i)\rangle_{\mathcal{H}} - 0\|_{\mathcal{Y}}^2 + \xi\|\hat{h}_n\|_{\mathcal{H}}^2 + \xi\|\hat{h}_n^\perp\|_{\mathcal{H}}^2 \\
&= \mathcal{E}_\xi^n(\hat{h}_n) + \xi\|\hat{h}_n^\perp\|_{\mathcal{H}}^2
\end{aligned}
$$

which implies $\mathcal{E}_\xi^n(\hat{h}) \geq \mathcal{E}_\xi^n(\hat{h}_n)$.

$\triangleright$ Since the minimizer is unique, then

$$
\hat{h} = \hat{h}_n = \sum_{i=1}^{n} \alpha_i[\phi(d_i)\otimes\phi(x_i)\otimes\phi(w_i)] \text{ for some } \alpha_i \tag{1}
$$

Step 2: Estimate Confounding Bridge h by regression

▷ Substitute the functional form of $\hat{h}$ into the loss

$$\begin{aligned}
&\langle h, \phi(d) \otimes \phi(x) \otimes \hat{\mu}(d,x,z)\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i [\phi(d_i) \otimes \phi(x_i) \otimes \phi(w_i)], \phi(d) \otimes \phi(x) \otimes \sum_{j=1}^{n} \beta_j(d,x,z)\phi(w_j) \right\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i [\phi(d_i) \otimes \phi(x_i) \otimes \phi(w_i)], \sum_{j=1}^{n} \beta_j(d,x,z)[\phi(d) \otimes \phi(x) \otimes \phi(w_j)] \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \beta_j(d,x,z) \Big\langle \phi(d_i) \otimes \phi(x_i) \otimes \phi(w_i), \phi(d) \otimes \phi(x) \otimes \phi(w_j) \Big\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \beta_j(d,x,z) k(d_i,d) k(x_i,x) k(w_i,w_j) \\
&= \alpha^T K(d,x) \beta(d,x,z)
\end{aligned}$$

where $K(d,x) := K_{Dd}\mathbb{1}_n^T \odot K_{Xx}\mathbb{1}_n^T \odot K_{WW} = K_{WW} \odot [(K_{Dd} \odot K_{Xx})\mathbb{1}_n^T] \in \mathbb{R}^n$

Step 2: Estimate Confounding Bridge h by regression

By Singh et al., 2020 [3] Algorithm C.1, $[A \odot b\mathbb{1}_n^T]a = [Aa] \odot b$, so

$$
\begin{aligned}
& K(d,x)\beta(d,x,z) \\
& = \{K_{WW} \odot [K_{Dd} \odot K_{Xx}]\mathbb{1}_n^T\}\beta(d,x,z) \\
& = [K_{WW}\beta(d,x,z)] \odot [K_{Dd} \odot K_{Xx}] \\
& = [K_{WW}(K_{DD} \odot K_{XX} \odot K_{ZZ} + n\lambda I)^{-1}\{K_{Dd} \odot K_{Xx} \odot K_{Zz}\}] \odot [K_{Dd} \odot K_{Xx}]
\end{aligned}
$$

Step 2: Estimate Confounding Bridge h by regression

By Singh et al., 2020 [3] Algorithm C.1, $[A \odot b\mathbb{1}_n^T]a = [Aa] \odot b$, so

$$
\begin{aligned}
&K(d,x)\beta(d,x,z) \\
&= \{K_{WW} \odot [K_{Dd} \odot K_{Xx}]\mathbb{1}_n^T\}\beta(d,x,z) \\
&= [K_{WW}\beta(d,x,z)] \odot [K_{Dd} \odot K_{Xx}] \\
&= [K_{WW}(K_{DD} \odot K_{XX} \odot K_{ZZ} + n\lambda I)^{-1}\{K_{Dd} \odot K_{Xx} \odot K_{Zz}\}] \odot [K_{Dd} \odot K_{Xx}]
\end{aligned}
$$

We need $\langle h, \phi(d_i) \otimes \phi(x_i) \otimes \hat{\mu}(d_i,x_i,z_i)\rangle_{\mathcal{H}} = \alpha^T K(d_i,x_i)\beta(d_i,x_i,z_i)$

$\triangleright$ Now we define a matrix M s.t. the i-th column is $K(d_i,x_i)\beta(d_i,x_i,z_i)$.

Explicitly,

$$M = K_{WW}(K_{DD} \odot K_{XX} \odot K_{ZZ} + n\lambda I)^{-1}\{K_{DD} \odot K_{XX} \odot K_{ZZ}\}] \odot [K_{DD} \odot K_{XX}]$$

$\triangleright$ Then the first term of $\mathcal{E}_\xi^n(\hat{h})$ can be written as $\frac{1}{n}\|Y - M^T\alpha\|_2^2$

Step 2: Estimate Confounding Bridge h by regression

▷ Regularization Term:

$$\begin{aligned}
\|\hat{h}\|_{\mathcal{H}}^2 &= \langle \hat{h}, \hat{h} \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i [\phi(d_i) \otimes \phi(x_i) \otimes \phi(w_i)], \sum_{j=1}^{n} \alpha_j [\phi(d_j) \otimes \phi(x_j) \otimes \phi(w_j)] \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(d_i, d_j) k(x_i, x_j) k(w_i, w_j) \\
&= \alpha^T [K_{DD} \odot K_{XX} \odot K_{WW}] \alpha
\end{aligned}$$

Step 2: Estimate Confounding Bridge h by regression

▷ Finally, do optimization on the matrix form!

$$\mathcal{E}_\xi^n(\hat{h}) = \frac{1}{n}\|Y - M^T\alpha\|_2^2 + \xi\alpha^T[K_{DD} \odot K_{XX} \odot K_{WW}]\alpha$$

$$\nabla_\alpha \mathcal{E}_\xi^n(\hat{h}) = \frac{1}{n} \cdot 2M(M^T\alpha - Y) + 2\xi[K_{DD} \odot K_{XX} \odot K_{WW}]\alpha = 0$$

$$\hat{\alpha} = (MM^T + n\xi[K_{DD} \odot K_{XX} \odot K_{WW}])^{-1}MY$$

Step 2: Estimate Confounding Bridge h by regression

▷ Finally, do optimization on the matrix form!

$$\mathcal{E}_\xi^n(\hat{h}) = \frac{1}{n}\|Y - M^T\alpha\|_2^2 + \xi\alpha^T[K_{DD} \odot K_{XX} \odot K_{WW}]\alpha$$
$$\nabla_\alpha \mathcal{E}_\xi^n(\hat{h}) = \frac{1}{n} \cdot 2M(M^T\alpha - Y) + 2\xi[K_{DD} \odot K_{XX} \odot K_{WW}]\alpha = 0$$
$$\hat{\alpha} = (MM^T + n\xi[K_{DD} \odot K_{XX} \odot K_{WW}])^{-1}MY$$

▷ The dual form solution $\hat{\alpha}$ gives us $\hat{h}$. Now given a new example,

$$\begin{aligned}
\hat{h}(d, x, w) &= \langle \hat{h}, \phi(d) \otimes \phi(x) \otimes \phi(w) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} \hat{\alpha}_i[\phi(d_i) \otimes \phi(x_i) \otimes \phi(w_i)], \phi(d) \otimes \phi(x) \otimes \phi(w) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \hat{\alpha}_i k(d_i, d)k(x_i, x)k(w_i, w) \\
&= \hat{\alpha}^T[K_{Dd} \odot K_{Xx} \odot K_{Ww}]
\end{aligned}$$

Step 3: Estimate Treatment Effects

▷ Recall that $\theta_0^{ATE}(d) = \langle \hat{h}, \phi(d) \otimes \mu \rangle_{\mathcal{H}}$, where $\mu = \int [\phi(x) \otimes \phi(w)] \mathbb{P}(x, w)$

▷ Estimate the mean embedding: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} [\phi(x_i) \otimes \phi(w_i)]$

* If other treatment effects involve *conditional* mean embedding, still we need to resort to Singh's KIV paper [2] for estimators (see next slide).

▷ Substitution

$$\begin{aligned}
\langle \hat{h}, \phi(d) \otimes \mu \rangle_{\mathcal{H}} &= \left\langle \hat{h}, \phi(d) \otimes \frac{1}{n} \sum_{i=1}^{n} [\phi(x_i) \otimes \phi(w_i)] \right\rangle_{\mathcal{H}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\langle \hat{h}, \phi(d) \otimes \phi(x_i) \otimes \phi(w_i) \right\rangle_{\mathcal{H}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}^T [K_{Dd} \odot K_{Xx_i} \odot K_{Ww_i}]
\end{aligned}$$

Step 3: Estimate Treatment Effects

Slightly more complicated: ATT

▷ Recall that

$$\theta_0^{ATT}(d, d') = \langle \hat{h}, \phi(d') \otimes \mu(d) \rangle_{\mathcal{H}}$$

where $\mu(d) = \int [\phi(x) \otimes \phi(w)] \mathbb{P}(x, w|d)$.

▷ Estimate the *conditional* mean embedding

$$\hat{\mu}(d) = \sum_{i=1}^{n} \beta_i(d)[\phi(x_i) \otimes \phi(w_i)]$$

where $\beta(d) = (K_{DD} + n\lambda_1 I)^{-1} K_{Dd}$

Step 3: Estimate Treatment Effects

▷ Substitution

$$
\begin{aligned}
\langle \hat{h}, \phi(d') \otimes \hat{\mu}(d) \rangle_{\mathcal{H}} &= \left\langle \hat{h}, \phi(d') \otimes \sum_{i=1}^{n} \beta_i(d)[\phi(x_i) \otimes \phi(w_i)] \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \beta_i(d) \left\langle \hat{h}, \phi(d') \otimes \phi(x_i) \otimes \phi(w_i) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \beta_i(d) \hat{\alpha}^T [K_{Dd'} \odot K_{Xx_i} \odot K_{Ww_i}] \\
&= \hat{\alpha}^T K_{Dd'} \odot \left\{ \sum_{i=1}^{n} \beta_i(d)[K_{Xx_i} \odot K_{Ww_i}] \right\} \\
&= \hat{\alpha}^T [K_{Dd'} \odot \{[K_{XX} \odot K_{WW}]\beta(d)\}] \\
&= \hat{\alpha}^T [K_{Dd'} \odot \{[K_{XX} \odot K_{WW}](K_{DD} + n\lambda_1 I)^{-1} K_{Dd}\}]
\end{aligned}
$$

References

[1] Rahul Singh. “Kernel methods for unobserved confounding: Negative controls, proxies, and instruments”. In: *arXiv preprint arXiv:2012.10315* (2020).

[2] Rahul Singh, Maneesh Sahani, and Arthur Gretton. “Kernel instrumental variable regression”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[3] Rahul Singh, Liyuan Xu, and Arthur Gretton. “Kernel Methods for Policy Evaluation: Treatment Effects, Mediation Analysis, and Off-Policy Planning”. In: *CoRR* abs/2010.04855 (2020). URL: https://arxiv.org/abs/2010.04855.