Reading Group: Causal Isotonic Regression

Ted Westling, Peter Gilbert, Marco Carone (JRSSB, 2020)

Agenda

1. Proposed approach

2. Theoretical properties

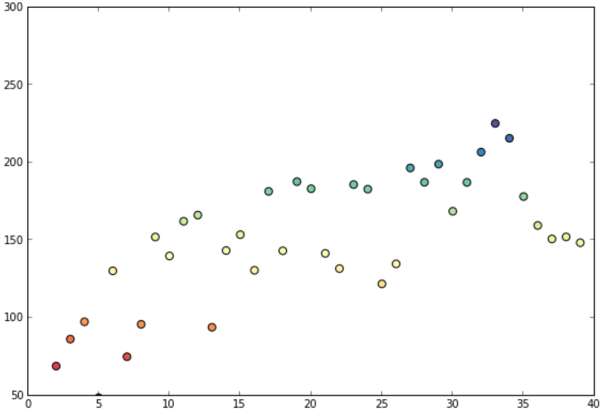
Classical least square isotonic regression

Linear regression: find $\beta = \hat{\beta}$ s.t. $\sum_{i=1}^{n} \left(Y_i - \beta A_i\right)^2$ is minimized

Isotonic regression: find $r = r_n$ s.t. $\sum_{i=1}^{n} \left[Y_i - r(A_i)\right]^2$ is minimized

- $Y_{1:n}$: responses
- $A_{1:n}$: continuous exposures
- r: any monotone non-decreasing function
- r_n can be obtained via pool adjacent violators algorithm (PAVA)
  - Not true without assuming piecewise linearity of r?
  - PAVA can be used to find best monotone fit $\hat{Y}_i$ only
- r_n can also be represented by greatest convex minorants (GCMs)
  - Probably because isotonic regression can be formulated as a convex programming problem
  - See section 2.3 of the R package *isotone*'s vignette

Pool adjacent violators algorithm



Source: Pedregosa, Fabian (2013)

Pool adjacent violators algorithm

Target: find best monotone fit $\hat{Y}_i$ of response Y_i

- Exposure A_i are ordered in i first, i.e. $A_1 \leq A_2 \leq \cdots \leq A_n$
- Response Y_i may not be monotone as the sorting is done on A_i
- Fit $\hat{Y}_i = r_n(A_i)$ must be monotone in i
  - That's way $r = r_n$ should be a monotone non-decreasing function
  - Identifiable without assuming piecewise linearity of r?

Algorithm (sketch):

1. Initialize $l := 0$, $B^{(0)} := n$, $\hat{Y}_r^{(0)} := Y_r$ for $r = 1, \ldots, n$
2. Merge $\hat{Y}^{(l)}$-values into blocks if $\hat{Y}_{r+1}^{(l)} < \hat{Y}_r^{(l)}$ for $r = 1, \ldots, B^{(l)}$
3. Minimize the loss function for each block r, which gives $\hat{Y}_r^{(l+1)}$
4. If $\hat{Y}_{r+1}^{(l)} < \hat{Y}_r^{(l)}$ for some r, set $l = l + 1$ and go back to step 2
5. Expand the block values w.r.t. to $i = 1, \ldots, n$

Greatest convex minorant

GCM of a function f bounded on $[a, b]$: supremum over all convex functions g such that $g \leq f$

Let F_n be the empirical distribution function of $A_{1:n}$. It can be shown that the isotonic regression estimator $r_n(a)$ is

- the left derivative, evaluated at $F_n(a)$,
- of the GCM over the interval $[0, 1]$ of the linear interpolation,
- of the cumulative sum diagram $\left\{\frac{1}{n}\left[i, \sum_{j=0}^{i} Y_{(i)}^{*}\right] : i = 0, 1, \ldots, n\right\}$,
- where $Y_{(0)}^{*} := 0$ and $Y_{(i)}^{*}$ is the response Y sorted by value of exposure A

Attractive properties of isotonic regression estimator

No need to choose kernel, bandwidth or any other tuning parameter

- The monotone fit restriction is kind of like a kernel already
- but it is true that no choice of tuning parameter is needed

Invariant to strictly increasing transformations of A

Uniform consistency on any strict subinterval of A

Limit distribution available

- $n^{\frac{1}{3}}\left[r_n(a) - r_0(a)\right] \xrightarrow{d} \left[4r_0'(a)\sigma_0^2(a)/f_0(a)\right]^{\frac{1}{3}}\mathbb{W}$ for any interior point $a \in \mathcal{A}$ at which $r_0'(a)$, $f_0(a) := F_0'(a)$ and $\sigma_0^2(a) := E_0\left[\{Y - r_0(a)\}^2 | A = a\right]$ exist and are positive and continuous in a neighbourhood of a
- $\mathbb{W}$ follows Chernoff's distribution, which often appears in the limit distribution of monotonicity-constrained estimators

Definition of proposed estimator

Definition: pointwise outcome

$$\mu_P(a, w) := E_P(Y|A = a, W = w)$$

for any given $P \in \mathcal{M}$

Definition: normalized exposure density

$$g_P(a, w) := \pi_P(a|w)/f_P(a)$$

where $\pi_P(a|w)$ is the conditional density evaluated at a given $W = w$, f_P is the marginal density of A under P

Definition: pseudo-outcome

$$\xi_{\mu,g,Q}(y, a, w) := \frac{y - \mu(a, w)}{g(a, w)} + \int \mu(a, z)Q(dz)$$

Monotonicity of proposed estimator

Kennedy *et al.* (2017) used pseudo-outcome to develop local linear regression for inference of $\theta_0(a)$. In the setting of this paper, $\theta_0(a)$ is known to be monotone

- I think the monotonicity of $\theta_0(a)$ is an assumption
- Yet it seems reasonable as continuous treatment usually has monotone causal effect (if effective) within certain range
- Example: daily exercise time (0-2 hours) on life expectancy
- Counterexample: daily exercise time (0-12 hours)
- So the reasonability of monotonicity may depend on the range of treatment A (experiemental) or data exploration (observational)

Under monotonicity, it is natural to consider the isotonic regression of the pseudo-outcomes on $A_{1:n}$

Proposed estimation procedure

Estimation of $\theta_n(a)$

1. Construct estimators μ_n, g_n of μ_0, g_0 respectively
2. For each a in the unique values of $A_{1:n}$, compute and set

$$\begin{aligned}\Gamma_n(a) := \tfrac{1}{n}\textstyle\sum_{i=1}^n I_{(-\infty,a]}(A_i)\tfrac{Y_i-\mu_n(A_i,W_i)}{g_n(A_i,W_i)} \\ + \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \sum_{j=1}^n I_{(-\infty,a]}(A_i)\mu_n(A_i,W_j)\end{aligned} \tag{1}$$

3. Compute the GCM $\bar{\Psi}_n$ of the set of points $\{(0,0)\} \cup \{\left(F_n(A_i), \Gamma_n(A_i)\right) : i = 1, 2, \ldots, n\}$ over $[0, 1]$
4. Define $\theta_n(a)$ as the left derivative of $\bar{\Psi}_n$ evaluated at $F_n(a)$

Asymptotic framework for the proposed estimator

As in Kennedy *et al.* (2017), $\theta_n(a)$ deviate from classical results

- Pseudo-outcomes $\xi_{\mu,g,Q}(y,a,w)$ are dependent because they depend on the estimator μ_n, g_n, Q_n estimated with all observations
- Hence classical results from isotonic regression do not apply
- However, θ_n is of generalized Grenander type
- Asymptotic results of Westling and Carone (2020) can be used

We skip the proof of θ_n to be Grenander type here, which is to show θ_n falls in the class of estimator discussed in Westling and Carone (2020)

Some remarks on monotonicity and generality

Monotonicity: if $\theta_0(a)$ were only known to be monotone on a fixed subinterval $\mathcal{A}_0 \subset \mathcal{A}$

- We discuss this assumption on page 9
- The estimation procedure is still valid by first defining $F_p(a) := P(A \leq a | A \in \mathcal{A}_0)$ and F_n as its empirical counterpart
- Then replace $I_{(-\infty,a]}(A_i)$ in equation 1 by $I_{(-\infty,a]\cap\mathcal{A}_0}(A_i)$

Generality: the proposed estimator θ_n generalizes the classical r_n

- Condition 1: $A \perp\!\!\!\perp W \implies g_0(a,w) = 1$, so we may take $g_n = 1$
- Condition 2: $Y|A \perp\!\!\!\perp W|A \implies \mu_n(a,w) = \mu_n(a)$
- Under these conditions, equation 1 becomes

$$\Gamma_n(a) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,a]}(A_i)Y_i - \mu_n(A_i)$$

- As a result, $\theta_n(a) = r_n(a)$ for each a

Agenda

1 Proposed approach

2 Theoretical properties

Invariance to strictly increasing transform of exposure

$\theta_n(a)$ is invariant to strictly increasing transform $H(\cdot)$ of exposure A

- Intuition: composition preserve monotonicity
- Desirable property since scale of exposure is often arbitrary
- Example: temperature in degrees Fahrenheit or Celsius or in kelvins
- Change of scale does not affect available information

We skip the proof because the intuition is simple (composition of monotone functions is also monotone)

Condtions for consistency

Notations:

- $\mathcal{F}$: a uniformly bounded class of functions
- Q: a finite discrete probability measure
- $N\{\epsilon, \mathcal{F}, L_2(Q)\}$: the ϵ-covering-number, i.e. the smallest number of $L_2(Q)$ balls of radius less than or equal to ϵ needed to cover $\mathcal{F}$
- $\log\left[\sup_Q N\{\epsilon, \mathcal{F}, L_2(Q)\}\right]$: the uniform ϵ-entropy of $\mathcal{F}$

Condition 1

There exist constants $C, \delta, K_0, K_1, K_2 \in (0, \infty)$ and $V \in [0, 2)$ s.t., almost surely as $n \to \infty$, μ_n and g_n are contained in classes of functions $\mathcal{F}_0$ and $\mathcal{F}_1$ respectively, satisfying

1. $|\mu| \leq K_0, \forall \mu \in \mathcal{F}_0$, and $K_1 \leq g \leq K_2, \forall g \in \mathcal{F}_1$
2. $\log\left[\sup_Q N\{\epsilon, \mathcal{F}_0, L_2(Q)\}\right] \leq C\epsilon^{-V/2}$ and $\log\left[\sup_Q N\{\epsilon, \mathcal{F}_1, L_2(Q)\}\right] \leq C\epsilon^{-V}, \forall \epsilon \leq \delta$

Condtions for consistency

P_0: the true data-generating distribution but not projection

Condition 2

There exists $\mu_\infty \in \mathcal{F}_0$ and $g_\infty \in \mathcal{F}_1$ s.t. $P_0(\mu_n - \mu_\infty)^2 \xrightarrow{p} 0$ and $P_0(g_n - g_\infty)^2 \xrightarrow{p} 0$

Condition 3

There exist subsets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of $\mathcal{A} \times \mathcal{W}$ s.t. $P_0(\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3) = 1$ and

1. $\mu_\infty(a, w) = \mu_0(a, w), \forall (a, w) \in \mathcal{S}_1$
2. $g_\infty(a, w) = g_0(a, w), \forall (a, w) \in \mathcal{S}_2$
3. $\mu_\infty(a, w) = \mu_0(a, w)$ and $g_\infty(a, w) = g_0(a, w), \forall (a, w) \in \mathcal{S}_3$

These conditions control the uniform entropy of certain classes of functions, which is related to empirical process theory. A thorough treatment is provided in van der Vaart and Wellner (1996)

Consistency

Theorem 1

If Conditions 1-3 hold, then $\theta_n(a) \xrightarrow{p} \theta_0(a)$ for any value $a \in \mathcal{A}$ s.t. $F_0(a) \in (0,1)$, θ_0 is continuous at a and F_0 is strictly increasing in a neighbourhood of a.

If θ_0 is uniformly continuous and F_0 is strictly increasing on $\mathcal{A}$, then $\sup_{a \in \mathcal{A}_0} \left[\theta_n(a) - \theta_0(a)\right] \overset{p}{0}$ for any bounded strict subinterval $\mathcal{A}_0 \subset \mathcal{A}$.

(Well-known) boundary issues with Grenander-type estimators:

- In the pointwise statement, $F_0(a)$ is required to be in $[0,1]$
- Similarly, the uniform statement only covers strict subintervals of $\mathcal{A}$
- Various remedies have been proposed before to mitigate this
- Potential direction for future research

Remark on Condtions for consistency

Remark on Condition 1

Condition 1 requires that μ_n and g_n eventually be contained in uniformly bounded function classes that are sufficiently small for certain empirical process terms to be controlled. This is satisfied by parametric classes and many infinite dimensional function classes. See chapter 2.6 of van der Vaart and Wellner (1996).

There is also an asymmetry between the entropy requirements for $\mathcal{F}_0$ and $\mathcal{F}_1$ in part 2 of Condition 1. This is due to the term $\int \int_{-\infty}^{a} \mu_n(u,w) F_n(du) Q_n(dw)$ appearing in $\Gamma_n(a)$. To control this term, an upper bound of the form $\int_0^1 \log \left[\sup_Q N\{\epsilon, \mathcal{F}_0, L_2(Q)\}\right] d\epsilon$ from the theory of empirical U-process is used (Nolan and Pollard, 1987).

The later part of this paper (section 3.7) considers the use of cross-fitting to avoid these entropy conditions in Condition 1.

Remark on Condtions for consistency

Remark on Condition 2 and 3

Condition 2 requires that μ_n and g_n tend to limit functions μ_∞ and g_∞, and Condition 3 requires that requires that either $\mu_\infty(a, w) = \mu_0(a, w)$ or $g_\infty(a, w) = g_0(a, w)$ for $(F_0 \times Q_0)$ almost every (a, w).

- This is equivalent to saying that μ_n or g_n is consistent?
- Seems to be in line with Kennedy *et al.* (2017)

If either

1. $\mathcal{S}_1$ and $\mathcal{S}_3$ are null sets or
2. $\mathcal{S}_2$ and $\mathcal{S}_3$ are null sets,

then Condition 3 is known simply as double robustness of the estimator θ_n relative to the nuisance functions μ_0 and g_0: θ_n is consistent as long as $\mu_\infty = \mu_0$ or $g_\infty = g_0$. However, Condition 3 is more general than classical double robustness as at least one of μ_n or g_n tends to the truth for **only** almost every point in the domain.

Double robustness

Multiply robustness: preserve consistency even if a subset of the N nuisance models is mispecified in the procedure
Double robustness: $N = 2$, so one model can be mispecified

Example: inverse probability weighted (IPW) estimator

- Y_i: response; $A_i \in \{0, 1\}$: treatment; W: covariates
- Estimator: $\hat{\mu}^{i-IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\pi_0(W_i)}$ where $\pi_0(W) = P(A = 1|W)$
  - Often infeasible since functional form $\pi_0(W)$ is unknown
  - (Example) nuisance model 1: $\pi_0(W) = \pi(W; \alpha_0) = \frac{\exp(\alpha_0^T \tilde{W})}{1+\exp(\alpha_0^T \tilde{W})}$
  - $\hat{\mu}^{f-IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\pi(W_i; \hat{\alpha})}$
- Augmented IPW (AIPW) estimator:
  - $\hat{\mu}^{f-\phi-IPW} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{A_i Y_i}{\pi(W_i; \hat{\alpha})} + \left\{ 1 - \frac{A_i}{\pi(W_i; \hat{\alpha})} \right\} \phi(W_i) \right]$
  - Nuisance model 2: $\phi(W) = E(Y|W, A = 1)$ is the most efficient

See Daniel (2017) for a simple introduction to this topic

Conditions for convergence in distribution

Notations:

- $d(h_1, h_2; a, \epsilon, \mathcal{S})$: pseudodistance; $\sigma_0^2(a, w)$: conditional variance
- $d(h_1, h_2; a, \epsilon, \mathcal{S}) := \sqrt{\sup_{|u-a| \leq \epsilon} E_0 \left[I_{\mathcal{S}}(u,W) \{ h_1 u,W) - h_2(u,W) \}^2 \right]}$
- $\sigma_0^2(a, w) := E_0 \left[\{ Y - \mu_0(A, W) \}^2 | A = a, W = w \right]$

Condition 4

There exists $\epsilon_0 > 0$ s.t.

1. $\max \left[d(\mu_n, \mu_\infty; a, \epsilon_0, \mathcal{S}_1), d(g_n, g_\infty; a, \epsilon_0, \mathcal{S}_2) \right] = o_p(n^{-1/3})$
2. $\max \left[d(\mu_n, \mu_\infty; a, \epsilon_0, \mathcal{S}_2), d(g_n, g_\infty; a, \epsilon_0, \mathcal{S}_1) \right] = o_p(1)$
3. $d(\mu_n, \mu_\infty; a, \epsilon_0, \mathcal{S}_3) d(g_n, g_\infty; a, \epsilon_0, \mathcal{S}_3) = o_p(n^{-1/3})$

Condition 5

$F_0, \mu_0, \mu_\infty, g_0, g_\infty$ and σ_0^2 are continuously differentiable in a neighbourhood of a uniformly over $w \in \mathcal{W}$

Convergence in distribution

Theorem 2

If Conditions 1-5 hold, then

$$n^{1/3}\{\theta_n(a) - \theta_0(a)\} \xrightarrow{d} \left\{ \frac{4\theta_0'(a)\kappa_0(a)}{f_0(a)} \right\}^{1/3} \mathbb{W}$$

for any $a \in \mathcal{A}$ such that $F_0(a) \in (0, 1)$, where $\mathbb{W}$ follows the standard Chernoff distribution and

$$\kappa_0(a) := E_0\left(E_0\left[\left\{ \frac{Y - \mu_\infty(a,W)}{g_\infty(a,W)} + \theta_\infty(a) - \theta_0(a) \right\}^2 \middle| A = a, W \right] g_0(a,W) \right)$$

with $\theta_\infty(a)$ denoting $\int \mu_\infty(a, w) Q_0(dw)$.

We skip the comparison between the limit distributions of θ_n and r_n as it is paritally discussed in p.12. In short, their limit distributions only differ in concentration, which is analogous to findings in linear regression

Remark on Conditions for convergence in distribution

Remark on Condition 4 and 5

The requirements of Condition 4 is equivalent to

1. On $\mathcal{S}_1$ where μ_n is consistent but g_n is not, μ_n converges faster than $n^{-1/3}$ uniformly in a neighbourhood of a,
2. Similarly for g_n on $\mathcal{S}_2$ and
3. On $\mathcal{S}_3$ where both μ_n and g_n are consistent, only the product of their rates of convergence must be faster than $n^{-1/3}$

This suggests the possibility of performing doubly robust inference for $\theta_0(a)$, which is explored in section 4. Note that as discussed in p.19, these conditions are more general than the classical double robustness

We skip the discussion of plug-in estimator $\theta_{\mu_n}(a)$, which can achieve faster rate of convergence than $\theta_n(a)$ but hinges entirely on the consistency of μ_n and may not admit a tractable limit theory

Grenander-type estimation without domain transform

The proposed estimator $\theta_n(a)$ coincides with a generalized Grenander-type estimator for which the marginal exposure empirical distribution function is used as domain transformation

An alternative estimator $\bar{\theta}_n$ could be constructed via Grenander-type estimation **without** the use of any domain transformation. We skip its construction here but there are several points to note:

- $\bar{\theta}_n$ does not generalize the classical isotonic regression
- $\bar{\theta}_n$ is not invariant to strictly increasing transform of A
- Domain of $\mathcal{A}$ needs to be known/chosen in defining $\bar{\theta}_n$
- When $\mu_\infty = \mu_0$, $\theta_n(a)$ and $\bar{\theta}_n$ may have the same limit distribution
- When $\mu_\infty \neq \mu_0$, $\bar{\theta}_n$ is dominated by $\theta_n(a)$ in AMSE sense
  - The transformation improves statistical efficiency in this case
  - Relative gain in efficiency is directly related to the asymptotic bias

Discrete domains

When A is discrete, $\theta_n(a)$ is asymptotically equivalent to the AIPW estimator, which is paritally discussed in p.20

As a result, the large sample properties of $\theta_n(a)$ can be derived from the large sample properties of the AIPW estimator and asymptotically valid inference can be obtained by using standard influence-function-based techniques

We skip the proof here as it is like realizing the isotonic regression of pseudo-outcome under discrete exposure coincides with the AIPW estimator. Instead, we shall have a short discussion on influence function

Influence function

Definition of influence function (Hampel *et al.*, 1986)

Let $T(F)$ be a statistical functional where F is a distribution. The influence function of T at F is given by

$$IF(x; T, F) := \lim_{t \downarrow 0} \frac{T\big[(1-t)F + t\delta_x\big] - T(F)}{t}$$

in those $x \in \mathcal{X}$ where this limit exists.

A complete discussion of this definition usually requires Gâteaux differentiability. We cover some of its usage instead:

- An estimator $\hat{\theta} \approx \theta(P_0) + E_n\big[IF(X)\big]$ can be dominated by a single outlier **unless** IF is bounded
- Asymptotic efficiency bound (Bias, Variance)
- Distributional decomposition, partial identification etc.

See this note for a quick summary

Large sample results for causal effects

The result so far concerns about the causal dose-response $a \mapsto m_0(a)$, which may not hold for the causal effect $(a_1, a_2) \mapsto m_0(a_1) - m_0(a_2)$

If the identification conditions discussed in Section 1.2 applied to each of a_1 and a_2, such causal effects can be identified with the observed data parameter $\theta_0(a_1) - \theta_0(a_2)$

If the condtions of Theorem 1 hold for both a_1 and a_2, we can establish consistency via the use of continuous mapping theorem

However, Theorem 2 only provides marginal distributional results. Joint convergence result is thus required for inference of causal effect

(Joint) convergence for causal effects

Theorem 3

Define $Z_n(a_1, a_2) := \left(n^{1/3}\{\theta_n(a_1) - \theta_0(a_1)\}, n^{1/3}\{\theta_n(a_2) - \theta_0(a_2)\}\right)$. If Conditions 1-5 hold for $a \in \{a_1, a_2\} \subset \mathcal{A}$ and $F_0(a_1), F_0(a_2) \in (0, 1)$, then

$$Z_n(a_1, a_2) \overset{d}{\to} \left(\{4\tau_0(a_1)\}^{1/3}\mathbb{W}_1, \{4\tau_0(a_2)\}^{1/3}\mathbb{W}_2\right)$$

where $\mathbb{W}_1, \mathbb{W}_2$ are independent standard Chernoff distributions and the scale parameter $\tau_0 = \frac{\theta_0'(a)\kappa_0(a)}{f_0(a)}$ is as defined in theorem 2.

Note that Theorem 3 implies

$$n^{1/3}\left[\left\{\theta_n(a_1) - \theta_n(a_2)\right\} - \left\{\theta_0(a_1) - \theta_0(a_2)\right\}\right] \overset{d}{\to} \left\{4\tau_0(a_1)\right\}^{1/3}\mathbb{W}_1 - \left\{4\tau_0(a_2)\right\}^{1/3}\mathbb{W}_2$$

Cross-fitting to avoid empirical process conditions

In observational studies, researchers can rarely specify a *priori* correct parametric models for μ_0 and g_0. This motivates use of data-adaptive estimators to meet Conditions 2 and 3

However, such estimators often leads to violation of Condition 1, or it may be onerous to determine that they do not. See slide p.18

In the context of asymptotically linear estimators, it has been noted that cross-fitting nuisance estimators can resolve this challenge by eliminating empirical process conditions

Therefore, this paper proposes cross-fitting of μ_n and g_n to avoid entropy conditions in Theorem 1 and 2

Estimation with cross-fitting

Estimation procedure with cross-fitting

1. Fix $V \in \{2, 3, \ldots, n/2\}$
2. Randomly partition the indices $\{1, 2, \ldots, n\}$ into V sets $\mathcal{V}_{n,1}, \mathcal{V}_{n,2}, \ldots, \mathcal{V}_{n,V}$
3. Assume $N := n/V \in \mathbb{Z}^+$. For each $v \in \{1, 2, \ldots, V\}$:
    1. Define $\mathcal{T}_{n,v} := \{O_i : i \notin \mathcal{V}_{n,v}\}$ as the *training set* for fold v
    2. Construct $\mu_{n,v}$ and $g_{n,v}$ using only observations from $\mathcal{T}_{n,v}$
4. Define pointwise the cross-fitted estimator Γ_n° of Γ_0 as

$$\Gamma_n^\circ(a) := \frac{1}{V} \sum_{v=1}^{V} \left[\frac{1}{N} \sum_{i \in \mathcal{V}_{n,v}} I_{(-\infty,a]}(A_i) \frac{Y_i - \mu_{n,v}(A_i, W_i)}{g_{n,v}(A_i, W_i)} \right.$$
$$\left. + \frac{1}{N^2} \sum_{i,j \in \mathcal{V}_{n,v}} I_{(-\infty,a]}(A_i) \mu_{n,v}(A_i, W_j) \right]$$

5. Construct the cross-fitted estimator θ_n° as in p.10

Remark: all results hold as long as $\max_v n/|\mathcal{V}_{n,v}| = O_p(1)$

Conditions for convergence under cross-fitting

Condition 6

There exist constants $C', \delta', K_0', K_1', K_2', K_3' \in (0, \infty)$ s.t., almost surely as $n \to \infty$ and for all v, $\mu_{n,v}$ and $g_{n,v}$ are contained in classes of functions $\mathcal{F}_0'$ and $\mathcal{F}_1'$ respectively, satisfying

1. $|\mu| \leq K_0', \forall \mu \in \mathcal{F}_0'$, and $K_1' \leq g \leq K_2', \forall g \in \mathcal{F}_1'$, and
2. $\sigma_0^2(a, w) \leq K_3'$ for almost all a and w

Condition 7

There exist $\mu_\infty \in \mathcal{F}_0'$ and $g_\infty \in \mathcal{F}_1'$ s.t. $\max_v P_0(\mu_{n,v} - \mu_\infty)^2 \xrightarrow{p} 0$ and $\max_v P_0(g_{n,v} - g_\infty)^2 \xrightarrow{p} 0$

Conditions for convergence under cross-fitting

Condition 8

There exists $\epsilon_0 > 0$ s.t.

1. $\max\left[d(\mu_{n,v}, \mu_\infty; a, \epsilon_0, \mathcal{S}_1), d(g_{n,v}, g_\infty; a, \epsilon_0, \mathcal{S}_2)\right] = o_p(n^{-1/3})$
2. $\max\left[d(\mu_{n,v}, \mu_\infty; a, \epsilon_0, \mathcal{S}_2), d(g_{n,v}, g_\infty; a, \epsilon_0, \mathcal{S}_1)\right] = o_p(1)$
3. $d(\mu_{n,v}, \mu_\infty; a, \epsilon_0, \mathcal{S}_3) d(g_{n,v}, g_\infty; a, \epsilon_0, \mathcal{S}_3) = o_p(n^{-1/3})$

Remark: Conditions 6, 7 and 8 are analogue of Conditions 1, 2 and 4 respectively under cross-fitting

Convergence under cross-fitting

Theorem 4

If Conditions 6, 7 and 3 hold, then $\theta_n^\circ(a) \xrightarrow{p} \theta_0(a)$ for any value $a \in \mathcal{A}$ s.t. $F_0(a) \in (0,1)$, θ_0 is continuous at a and F_0 is strictly increasing in a neighbourhood of a.

If θ_0 is uniformly continuous and F_0 is strictly increasing on $\mathcal{A}$, then $\sup_{a \in \mathcal{A}_0} \left[\theta_n^\circ(a) - \theta_0(a)\right] \overset{p}{0}$ for any bounded strict subinterval $\mathcal{A}_0 \subset \mathcal{A}$.

Theorem 5

If Conditions 6, 7, 3, 8, 5 hold, then

$$n^{1/3}\{\theta_n^\circ(a) - \theta_0(a)\} \xrightarrow{d} \{4\tau_0(a)\}^{1/3}\,\mathbb{W}$$

for any $a \in \mathcal{A}$ such that $F_0(a) \in (0,1)$, where $\mathbb{W}$ follows the standard Chernoff distribution.