Semiparametric Counterfactual Density Estimation

Edward Kennedy & Siva Balakrishnan & Larry Wasserman

Department of Statistics & Data Science
Carnegie Mellon University

IU Biostatistics, 19 Mar 2021

Motivation

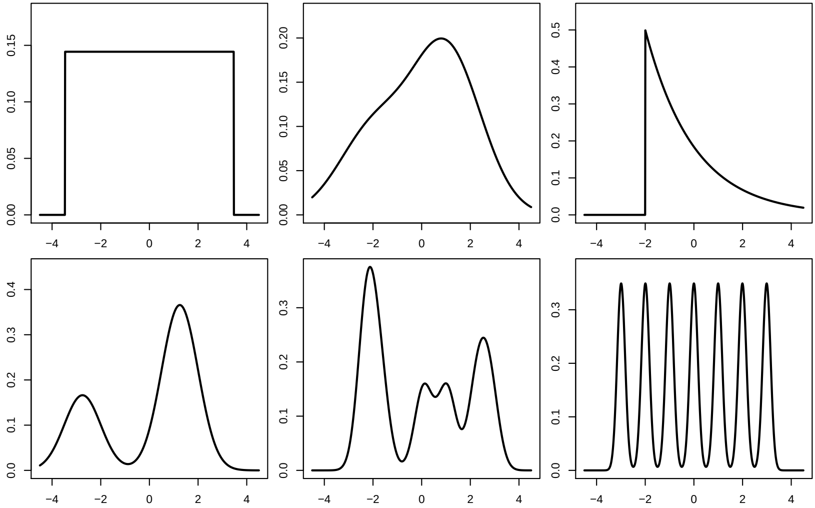
Let Y^a denote potential/counterfactual outcome that would have been observed under treatment $A = a$

- causal inference $\approx$ estimating features of distribution of Y^a

Very common to quantify effects with means, e.g., ATE = mean outcome if all versus none were treated

$$\mathbb{E}(Y^1 - Y^0)$$

Certainly a useful summary – but can miss important differences!


0.00
0.05
0.10
0.15
0.20
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4

Motivating application

What is effect on CD4 of combination antiretroviral therapy versus zidovudine alone in patients with HIV?

- mean effect $>$ median effect
- how is combination therapy affecting *distribution*?

Why do we care?

Knowing counterfactual densities can be very useful

- if densities differ at all, there is *some* treatment effect

Skew $\implies$ some subjects have extreme responses

- could try to find who they are, why responses are extreme

Multimodality $\implies$ may exist underlying heterogeneous subgroups

- could be useful for optimizing policy, understanding variation

Density shape can inform hypotheses about treatment mechanism

- maybe trt reduces variance, or drives up negative outcomes
- can help enhance future treatments, motivate new ones

Work on causal CDF estimation

Large literature on distributional effects via quantiles or CDFs

- Abadie ('02), Melly ('05), Chernozhukov et al. ('05, '13), Firpo ('07), Rothe ('10), Frolich & Melly ('13), Diaz ('17)

But challenges & methods are very different for densities

- $\mathbb{P}(Y \leq y) = \mathbb{E}\{\mathbb{1}(Y \leq y)\}$ so reduces to mean estimation
- CDF yields unbiased estimators, $\sqrt{n}$ rates; density requires bias/var trade-off (CDF pathwise differentiable, density not)
- CDFs easier to estimate, but densities easier to interpret

CDFs & densities should really be viewed as complementary

Work on causal density estimation

Counterfactual density estimation literature is much more sparse

- Dinardo et al. ('96) - IPW kernel estimator
- Robins & Rotnitzky ('01) - DR kernel estimator
- vdL & Dudoit ('03), Rubin & vdL ('06) - CV w/KL & L_2
- Westling & Carone ('20) - monotone densities
- Kim et al. ('18) - DR kernel estimator & L_1 distance

None uses semiparametric approach

- i.e., where density is approximated with d-dimensional model

Punchline

Our work aims to fill this gap in the literature

- also give data-driven model selection & aggregation tools

Separate contribution:

- generic density-based effects, which characterize the distance between counterfactual densities, using a generalized notion of distance that includes f-divergences as well as L_p norms

Setup

Given iid sample of $Z = (X, A, Y) \sim \mathbb{P}$ where

- $X \in \mathbb{R}^d$ = covariates, $A \in \{0, 1\}$ = trt, $Y \in \mathbb{R}$ = outcome

Some notation:

- $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$ = propensity score
- $\eta_a(y \mid x) = \frac{\partial}{\partial y}\mathbb{P}(Y \leq y \mid X = x, A = a)$ = outcome density

and covariate-adjusted density

$$p_a(y) = \int \eta_a(y \mid x)\, d\mathbb{P}(x)$$

= density of Y^a under consistency/positivity/exchangeability

Overview of target parameters

We consider two kinds of target parameters:

- approximation of the counterfactual density, defined via a projection in some distributional distance
- density-based causal effect, measuring difference between counterfactual densities in general f- or other divergences

Density effects give a more nuanced picture of how counterfactual densities differ, compared to the usual ATE

We also show how these two targets can be adapted for *model selection & aggregation*

Target 1: density functions

First: approximations of $p_a(y)$ based on model $\{g(y;\beta) : \beta \in \mathbb{R}^d\}$

- Exponential family: for basis $b(y) = \{b_1(y), ..., b_d(y)\}^{\mathrm{T}}$, let

$$g(y;\beta) = \exp\left\{\beta^{\mathrm{T}} b(y) - C(\beta)\right\}$$

where $C(\beta) = \log \int \exp\{\beta^{\mathrm{T}} b(y)\}\, dy$ so that $\int g(y;\beta)\, dy = 1$

- Truncated series: for base density $q(y)$ can use linear model

$$g(y;\beta) = q(y) + \sum_{j=1}^{d} \beta_j b_j(y)$$

e.g., for $Y \in [0,1]$ take $q(y) = 1$ and $b_j(y) = \sqrt{2}\cos(\pi j y)$

- Gaussian mixture: $g(y;\beta) = \sum_{j=1}^{k} \varpi_j \left(\frac{1}{\sigma_j}\right) \phi\left(\frac{y-\mu_j}{\sigma_j}\right)$

Projection parameter

We do not assume our model is correct! Instead just use it for defining approximations:

$$\beta_0 = \arg\min_{\beta \in \mathbb{R}^p} \; D_f\Big(p_a(y), g(y; \beta)\Big)$$

where D_f is a distributional distance

$$D_f(p, q) = \int f\Big(p(y), q(y)\Big) q(y) \; dy$$

for some given discrepancy function $f : \mathbb{R}^2 \to \mathbb{R}$

- generalization of f-divergence that includes L_p^p distances

Parameter interpretation

Mathematically $g(y;\beta_0)$ is the best-fitting model of this form

- if model is correct, $g(y;\beta_0) = p_a(y)$ is true density
- under misspecification, $g(y;\beta_0)$ is just best approximation

Actually assuming $p_a(y) = g(y;\beta_0)$ would be semiparametric

- all our results are formally nonparametric

Similar to best linear approximation in regression (White '80)

- long history in stats (Huber, Beran, White, Buja et al., etc.) & causal (vdL, Cuellar & Kennedy, Semenova & Chernoz.)

Statistical epistemology

Can imagine at least 3 approaches here:

1. modelist: assumes finite-dim model is *the* correct one
2. model-agnostic: *uses* finite-dim model, allows it to be wrong
3. anti-modelist: model's wrong, & don't want approximation

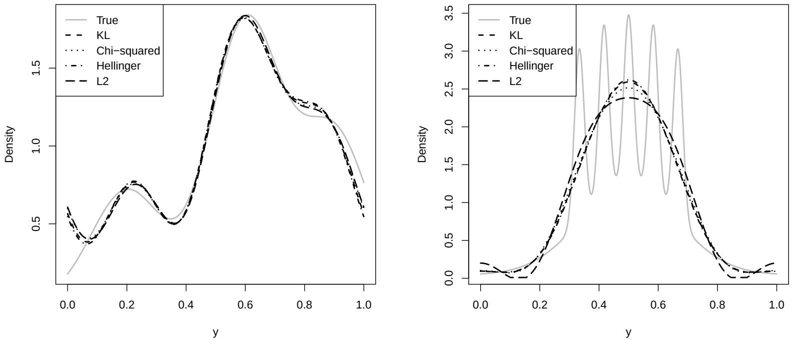
Each approach has trade-offs:

- modelist will do well if correct, otherwise biased
- anti-modelist doesn't need to worry about bias as much, but has to live with larger errors due to more ambitious target
- model-agnostic: if model is correct, can do nearly as well as modelist, otherwise inference still valid for approximation
  → but choosing model/distance be a challenge

Distances

- L_2^2: $f(p,q) = \frac{(p-q)^2}{q} \implies D_f(p,q) = \|p-q\|_2^2$
- KL: $f(p,q) = \frac{p}{q}\log\left(\frac{p}{q}\right) \implies D_f(p,q) = \mathrm{KL}(p,q)$
- χ^2: $f(p,q) = (p/q-1)^2 \implies D_f(p,q) = \chi^2(p,q)$
- Hellinger: $f(p,q) = (\sqrt{p/q}-1)^2 \implies D_f(p,q) = H^2(p,q)$
- TV: $f(p,q) = \frac{|p-q|}{2q} \implies D_f(p,q) = \mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$
- TV*: $f(p,q)\frac{\nu_t(p-q)}{2q}$ for ν_t smooth approximation of $|\cdot|$

Projection examples



Moment condition

For smooth distances, β_0 can be defined with moment condition

- links projection parameters to integral functionals of $p_a(y)$

Proposition

Assume smoothness conditions and let $f_2'(q_1, q_2) = \frac{\partial}{\partial q_2} f(q_1, q_2)$. *Then the projection parameter*

$$\beta_0 = \underset{\beta \in \mathbb{R}^p}{\arg\min}\ D_f\Big(p_a(y), g(y;\beta)\Big)$$

is a solution to the moment condition $m(\beta) = 0$, *where*

$$m(\beta) \equiv \int \frac{\partial g(y;\beta)}{\partial \beta} \left\{ f\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(p_a(y), g(y;\beta)\Big) \right\} dy.$$

Moment condition examples

If $D_f = L_2^2$, $Y \in [0,1]$, and $g(y;\beta) = 1 + \beta^{\mathrm{T}} b(y)$ then

$$\beta = \mathbb{E}\Big\{b(Y^a)\Big\}$$

if b is series w/ $\int b_j(y)\ dy = 0$ & $\int b_j(y) b_k(y)\ dy = \mathbb{1}(j = k)$

- closed form expression! just mean of transformed outcome

If $D_f = \mathrm{KL}$ and $g(y;\beta) \propto \exp\{\beta^{\mathrm{T}} b(y)\}$ then

$$m(\beta) = -\mathbb{E}\left\{\frac{\partial}{\partial\beta}\log g(Y^a;\beta)\right\} = \int b(y)\Big\{g(y;\beta) - p_a(y)\Big\}dy$$

- matches moments of b under $g(y;\beta)$ and $p_a(y)$

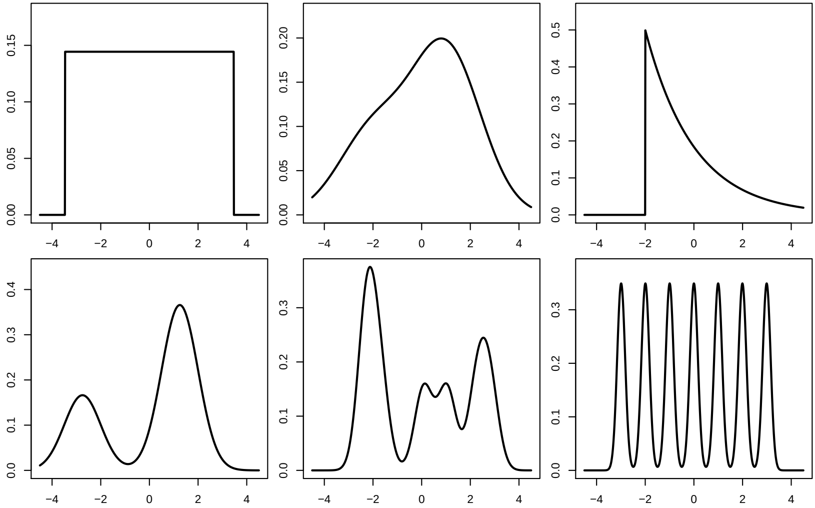
Target 2: density effects

In addition to density approximations we consider density effects

$$\psi_f = D_f\Big(p_1(y), p_0(y)\Big) = \int f\Big(p_1(y), p_0(y)\Big) p_0(y)\ dy$$

Note: here we do not require an approximating model!

Give more nuanced picture of how counterfactual densities differ, compared to the usual ATE


0.15
0.10
0.05
0.00
−4
−2
0
2
4
0.20
0.15
0.10
0.05
0.00
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.3
0.2
0.1
0.0
0.3
0.2
0.1
0.0

Model selection & aggregation

In practice may want to use data to choose among many models

- need to adapt CV/selection a la van der Laan & Dudoit ('03)

Given set of estimators $\{\widehat{g}_k(y) : k = 1, \ldots, K\}$ can define risk

$$R(\widehat{g}_k) = D_f\Big(p_a(y), \widehat{g}_k(y)\Big)$$

and oracle aggregator as $\widetilde{g}(y) = \sum_k \beta_{0k}\widehat{g}_k(y)$ where

$$\beta_0 = \arg\min_{\beta \in B} D_f\left(p_a(y), \sum_{k=1}^{K} \beta_k \widehat{g}_k(y)\right)$$

for some appropriate selection set, e.g., for convex aggregation the simplex $B = \{(\beta_1, \ldots, \beta_K) \in \mathbb{R}^K : \beta_k \geq 0, \sum_k \beta_k = 1\}$

Punchline

We give a crucial von Mises (i.e., distributional Taylor) expansion for generic density functionals, which yields EIFs

- so nonparametric efficiency bounds & local minimax lower bds
- also estimators that can be optimally efficient

Throughout we reference linear map $T \mapsto \phi_a(T; \mathbb{P})$ defined as

$$\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big\{T - \mathbb{E}(T \mid X, A=a)\Big\} + \mathbb{E}(T \mid X, A=a) - \mathbb{E}\{\mathbb{E}(T \mid X, A=a)\}$$

which outputs EIF for $\mathbb{E}\{\mathbb{E}(T \mid X, A=a)\}$. In our examples, $T = h(Y)$ will be non-trivial function of outcome Y, depending on model/distance

Master lemma

Lemma

Let $\psi = \psi(\mathbb{P}) = \int h(p_a(y))\, dy$ for some twice continuously differentiable function h. Then ψ satisfies the von Mises expansion

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \phi_a\left(h'\left(p_a(Y)\right); \overline{\mathbb{P}}\right)\, d(\overline{\mathbb{P}} - \mathbb{P}) + R_2(\overline{\mathbb{P}}, \mathbb{P})$$

where, for $p_a^*(y)$ between $p_a(y)$ and $\overline{p}_a(y)$, $R_2(\overline{\mathbb{P}}, \mathbb{P})$ is given by

$$\int\int h'(\overline{p}_a(y)) \left\{\frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1\right\} \left\{\eta_a(y \mid x) - \overline{\eta}_a(y \mid x)\right\} dy\ d\mathbb{P}(x)$$
$$+ \frac{1}{2} \int h''(p_a^*(y))\left\{\overline{p}_a(y) - p_a(y)\right\}^2 dy,$$

Density functions

Theorem

Let f be 2x differentiable & let $f_j'(q_1, q_2) = \frac{\partial}{\partial q_j} f(q_1, q_2)$ & $f_{jk}''(q_1, q_2) = \frac{\partial^2}{\partial q_j \partial q_k} f(q_1, q_2)$. The EIF for $m(\beta)$ is $\phi_a\Big(\gamma_f(Y;\beta)\Big)$ where

$$\gamma_f(y;\beta) = \frac{\partial g(y;\beta)}{\partial \beta} \left\{ f_1'\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_{21}''\Big(p_a(y), g(y;\beta)\Big) \right\}$$

The EIFs for β_0 and $g(y;\beta_0)$ are

$$-\frac{\partial m(\beta)}{\partial \beta}^{-1} \phi_a\Big(\gamma_f(Y;\beta)\Big) \Big|_{\beta=\beta_0}, \quad -\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}} \frac{\partial m(\beta)}{\partial \beta}^{-1} \phi_a\Big(\gamma_f(Y;\beta)\Big) \Big|_{\beta=\beta_0}$$

Density functions: L_2 & KL

Corollary

For L_2^2 and KL divergence the quantity γ_f reduces to

$$\gamma_f(y;\beta) = \begin{cases} -2\frac{\partial g(y;\beta)}{\partial \beta} & \text{if } D_f = L_2^2 \\ -\frac{\partial \log g(y;\beta)}{\partial \beta} & \text{if } D_f = KL. \end{cases}$$

Further, if either

1. $D_f = L_2^2$ *and* $g(y;\beta) = q(y) + \beta^{\mathrm{T}} b(y)$ *is truncated series*
2. $D_f = KL$ *and* $g(y;\beta) = \exp\{\beta^{\mathrm{T}} b(y) - C(\beta)\}$ *is exp fam*

then EIF for $m(\beta)$ *is proportional to*

$$\phi_a\Big(b(Y)\Big)$$

Density effects

Theorem

In an unrestricted nonparametric model, the efficient influence function for the density effect $\psi_f = \int f\left(p_1(y), p_0(y)\right) p_0(y)\, dy$ *is given by*

$$\phi_1\Big(\lambda_1(Y)\Big) + \phi_0\Big(\lambda_0(Y)\Big)$$

where

$$\lambda_1(y) = p_0(y) f_1'\Big(p_1(y), p_0(y)\Big)$$
$$\lambda_0(y) = f\Big(p_1(y), p_0(y)\Big) + p_0(y) f_2'\Big(p_1(y), p_0(y)\Big).$$

Density effects: L_2 & KL

Corollary

If $D_f = L_2^2$, *then the efficient influence function for* ψ_f *is*

$$2(\phi_1 - \phi_0)\Big(p_1(Y) - p_0(Y)\Big).$$

If $D_f = KL$, *then the efficient influence function for* ψ_f *is*

$$\phi_1 \left(\log\left(\frac{p_1(Y)}{p_0(Y)}\right)\right) - \phi_0 \left(\frac{p_1(Y)}{p_0(Y)}\right).$$

Proposed density estimator

A plug-in estimator is given by the solution to

$$\widehat{m}(\beta) \equiv \int \frac{\partial g(y;\beta)}{\partial \beta} \left\{ f\Big(\widehat{p}_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(\widehat{p}_a(y), g(y;\beta)\Big) \right\} dy$$

This will be suboptimal in general. Our proposed estimator solves

$$\widehat{m}(\beta) + \mathbb{P}_n \left\{ \widehat{\phi}_a\Big(\widehat{\gamma}_f(Y;\beta)\Big) \right\} = o_{\mathbb{P}}(1/\sqrt{n})$$

where $\widehat{\phi}_a(T) = \phi_a(T;\widehat{\mathbb{P}})$ is estimated EIF

- i.e., one-step bias-corrected estimators, which take the plug-in & add estimated bias, i.e., add average estimated IF

Proposed estimator: L_2 case

Proposition

If $D_f = L_2^2$, $Y \in [0,1]$, *and* $g(y;\beta) = 1 + \beta^{\mathrm{T}} b(y)$ *then plug-in is*

$$\widehat{\beta} = \mathbb{P}_n\{\widehat{\mu}_a(X;b)\},$$

where $\widehat{\mu}_a(x;b)$ *is estimate of* $\mu_a(x;b) = \mathbb{E}\{b(Y) \mid X = x, A = a)$. *In contrast, our proposed one-step estimator is given by*

$$\widehat{\beta} = \mathbb{P}_n\left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a(X)}\Big\{b(Y) - \widehat{\mu}_a(X;b)\Big\} + \widehat{\mu}_a(X;b)\right]$$

Proposed estimator: KL case

Proposition

If $D_f = KL$ *and* $g(y;\beta) = \exp\{\beta^{\mathrm{T}} b(y) - C(\beta)\}$, *then plug-in solves*

$$\int \Big[b(y) - \mathbb{P}_n\{\widehat{\mu}_a(X;b)\} \Big] \exp\Big\{ \beta^{\mathrm{T}} b(y) \Big\}\, dy = 0$$

where $\widehat{\mu}_a(x;b)$ *is estimate of* $\mu_a(x;b) = \mathbb{E}\{b(Y) \mid X = x, A = a)$.
In contrast, our proposed one-step estimator solves

$$\int \left(b(y) - \mathbb{P}_n \left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a(X)} \Big\{ b(Y) - \widehat{\mu}_a(X;b) \Big\} + \widehat{\mu}_a(X;b) \right] \right) \exp\Big\{ \beta^{\mathrm{T}} b(y) \Big\}\, dy = 0$$

Rates of convergence

Theorem

Let $\eta = (\pi_a, \eta_a)$, $\varphi(Z; \beta, \eta) = m(\beta; \eta) + \phi_a(\gamma_f(Y; \beta), \eta)$. *Assume:*

1. γ_f *and* $1/\widehat{\pi}_a$ *are bounded above, &* γ_f *is differentiable in* $p_a(y)$, *with derivative bounded above by* δ.
2. *The function class* $\{\varphi(z; \beta, \eta) : \beta \in \mathbb{R}^p\}$ *is Donsker in* β.
3. *Consistency, i.e.,* $\widehat{\beta} - \beta_0 = o_{\mathbb{P}}(1)$ *and* $\|\widehat{\eta} - \eta_0\| = o_{\mathbb{P}}(1)$.
4. *Map* $\beta \mapsto \mathbb{P}\{\varphi(Z; \beta, \eta)\}$ *is differentiable, with derivative matrix* $\frac{\partial}{\partial \beta}\mathbb{P}\{\varphi(Z; \beta, \widehat{\eta})\}|_{\beta=\beta_0} = V(\beta_0, \widehat{\eta}) \xrightarrow{p} V(\beta_0, \eta_0)$.

Then

$$\widehat{\beta} - \beta_0 = -V(\beta_0, \eta_0)^{-1}(\mathbb{P}_n - \mathbb{P})\left\{\phi_a\Big(\gamma_f(Y; \beta_0)\Big)\right\} + O_{\mathbb{P}}\left(\|\widehat{\pi}_a - \pi_a\|\|\widehat{\eta}_a - \eta_a\| + \delta\|\widehat{p}_a - p_a\|^2 + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\right).$$

Rates of convergence

Theorem shows $\widehat{\beta}$ attains faster rates than nuisance estimators $\widehat{\eta}$, & can be efficient under weak nonparametric conditions

- 1st condition ensures the IF is not too complex
- 2nd merely requires consistency of $(\widehat{\beta}, \widehat{\eta})$ at any rate
- 3rd requires some smoothness in β, to allow delta method

Rate is second-order in nuisance estimation error

- γ_f may not depend on $p_a(y)$, so derivative is zero & $\delta = 0$

Proposed effect estimator

The density effect estimators we propose are defined as

$$\widehat{\psi}_f = \int f\Big(\widehat{p}_1(y), \widehat{p}_0(y)\Big)\widehat{p}_0(y)\, dy + \mathbb{P}_n\left\{\widehat{\phi}_1\Big(\widehat{\lambda}_1(Y)\Big) + \widehat{\phi}_0\Big(\widehat{\lambda}_0(Y)\Big)\right\}$$

which can again be viewed as one-step bias-corrected estimators, w/plug-in bias estimated via an average of EIF

Note: rather than estimating the density η_a & integrating over its y argument, one could instead regress $\widehat{\lambda}_a$ on X for the integral terms in the EIF

Effect estimator: L_2

Proposition

If $D_f = L_2^2$ then proposed density effect estimator is

$$2\,\mathbb{P}_n\left(\frac{2A-1}{\widehat{\pi}_A(X)}\left[\left\{\widehat{p}_1(Y)-\widehat{p}_0(Y)\right\}-\int\left\{\widehat{p}_1(y)-\widehat{p}_0(y)\right\}\widehat{\eta}_A(y\mid X)\,dy\right]+\int\left\{\widehat{p}_1(y)-\widehat{p}_0(y)\right\}\left\{\widehat{\eta}_1(y\mid X)-\widehat{\eta}_0(y\mid X)\right\}\,dy\right)-\int\left\{\widehat{p}_1(y)-\widehat{p}_0(y)\right\}^2\,dy.$$

Rates of convergence

Theorem

Assume λ_a and $1/\widehat{\pi}_a$ are bounded above, and λ_a is differentiable in $p_a(y)$, with derivative bounded above by δ_a. Then

$$\widehat{\psi}_f - \psi_f = (\mathbb{P}_n - \mathbb{P})\left\{\phi_1\Big(\lambda_1(Y)\Big) + \phi_0\Big(\lambda_0(Y)\Big)\right\} + O_{\mathbb{P}}\left(\sum_{a=0}^{1}\left(\|\widehat{\pi}_a - \pi_a\|\|\widehat{\eta}_a - \eta_a\| + \delta_a\|\widehat{p}_a - p_a\|^2\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\right)$$

Inference

There is a special distinction in density effect estimation. Results suggest 95% CIs of the form

$$\widehat{\psi}_f \pm 1.96\sqrt{\widehat{\text{cov}}\left\{\widehat{\phi}_1\left(\widehat{\lambda}_1(Y)\right) + \widehat{\phi}_0\left(\widehat{\lambda}_0(Y)\right)\right\}/n}$$

These intervals are asymptotically valid as usual when $p_1 \neq p_0$, but not when $p_1 = p_0$, since then IF reduces to zero

- sample avg term no longer dominant
- similar to degenerate U-statistics

Simple fix is to use the interval $\widehat{\psi} \pm z_{\alpha/2}(s \vee 1/\sqrt{n})$ where $s = \sqrt{\widehat{\text{cov}}\{\widehat{\phi}_1(\widehat{\lambda}_1(Y)) + \widehat{\phi}_0(\widehat{\lambda}_0(Y))\}/n}$: valid but conservative

Model selection & aggregation

Consider linear aggregation, where our methods are straightforward. (Note f-divergences may not be well-defined.)

Our proposed approach is:

Step 1. Randomly split sample into training set D_n^0 and test set D_n^1.
Step 2. In training set D_n^0, estimate models $\widehat{g}_k(y) = g(y; \widehat{\beta}_k)$
Step 3. In test set D_n^1, estimate projection of linear span of $\widehat{g}_k$ onto basis to compute aggregated estimator $\widehat{g}(y) = \sum_k \widehat{\theta}_k \widehat{g}_k(y)$.
Step 4. Reverse roles of D_n^0 and D_n^1 and avg two resulting aggregates.

Model selection & aggregation

For model selection & convex aggregation, can estimate the distance between p_a & each of k candidates, & pick minimizer

For example, with L_2^2 can use

$$\widehat{\Delta}_f(g_k) = \int \Big(\widehat{p}_a(y) - g_k(y)\Big)^2 \, dy + 2\mathbb{P}_n \left\{ \widehat{\phi}_a \Big(\widehat{p}_a(Y) - g_k(Y) \Big) \right\}.$$

or pseudo-L_2^2 risk

$$\widehat{\Delta}_f^*(g_k) = -2 \, \mathbb{P}_n \left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a(X)} \left\{ g_k(Y) - \int g_k(y) \widehat{\eta}_a(y \mid X) \, dy \right\} + \int g_k(y) \widehat{\eta}_a(y \mid X) \, dy \right] + \int g_k(y)^2 \, dy,$$

since L_2^2 is this plus a term $\int p_a^2$ that does not depend on g_k

Data

We apply methods to study effects of combination antiretroviral therapy among $n = 2319$ patients with HIV

- Y = CD4 count at 96 weeks
- A = combination therapy vs zidovudine (& observed outcome)
- X = age, weight, Karnofsky score, race, gender, hemophilia, sexual orientation, drug use, symptoms, previous trt history

Data are freely available in speff2trial R package

Methods

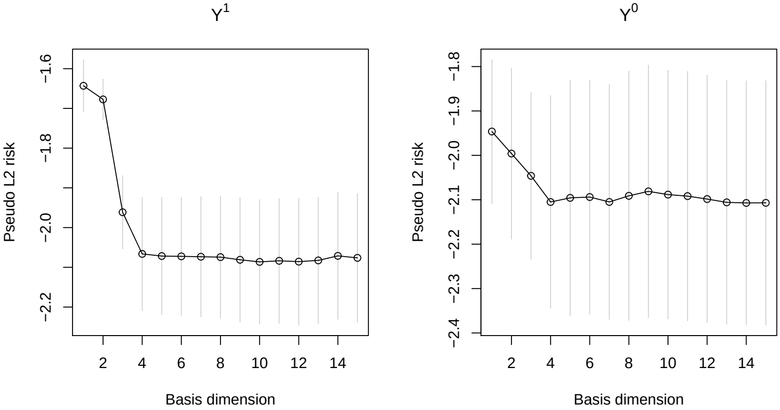
We used 5-fold cross-fitting with random forests

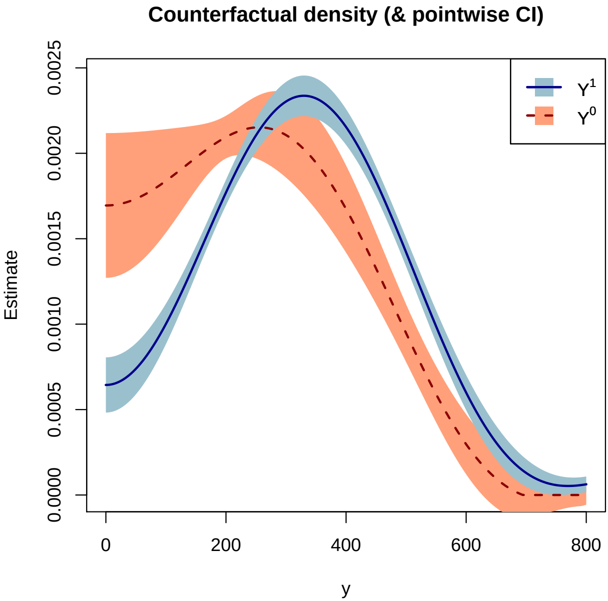
- $\widehat{\eta}_a$ constructed via RF regression: $\frac{1}{h}K\left(\frac{Y-y}{h}\right) \sim X, A$

Targets:

- L_2 distance between p_1 and p_0
- L_2 projections onto linear series with cosine basis
- L_2 risk for $k = 1, \ldots, 15$

Model selection




Counterfactual density (& pointwise CI)
Estimate
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0
200
400
600
800
y
Y1
Y0

Interpretation

The CD4 densities differ more substantially in the lowest CD4 range (e.g., 0-200)

- combination therapy may have increased CD4 count most for high-risk patients w/ lowest counts under control (zidovudine)

R code

```
# install npcausal package
install.packages("devtools"); library(devtools)
install_github("ehkennedy/npcausal"); library(npcausal)

# load data
library(speff2trial); data(ACTG175); dat <- ACTG175[,c(2:17,19,21,23)]
x <- dat[,!(colnames(dat) %in% c("treat","cd496"))]

# create treatment*missing indicator
a1 <- dat$treat*(!is.na(dat$cd496)); a0 <- (1-dat$treat)*(!is.na(dat$cd496))
a <- a1; a[a0==0 & a1==0] <- -1; y <- dat$cd496; y[is.na(dat$cd496)] <- 0

# estimate pseudo-l2 risk for k=1:15
cv.cdensity(y,a,x, kmax=15, gridlen=50,nsplits=5)

# estimate densities at k=4
res <- cdensity(y,a,x, kmax=4, kforplot=c(4,4), gridlen=50,nsplits=5,ylim=c(0,800))
```

Summary

We proposed methods for estimating counterfactual densities and corresponding distances and other functionals

- gave efficiency bounds & flexible optimal estimators for wide class of models & projection distances, & for new effects that quantify treatment impacts on the density scale

Also gave methods for data-driven model selection and aggregation

- illustrated in application studying effects of antiretroviral therapy on CD4 count

Discussion points

Lots of avenues for future work

- nonparametric version of the problem
- non-discrete treatments (where A is e.g., a continuous dose)
- computational aspects (require solving messy estimating eqs)
- time-varying trts, instrumental variables, conditional effects, density-optimal trt regimes, mediation, sensitivity analysis...

Paper is on arxiv:
https://arxiv.org/pdf/2102.12034.pdf

Feel free to email with any questions:
edward@stat.cmu.edu

Thank you!

Wrapping up

Corollary

The quantity $f\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(p_a(y), g(y;\beta)\Big)$ *in the integrand of the moment condition equals*

$$
\begin{cases}
2\Big\{g(y;\beta) - p_a(y)\Big\} & \text{if } D_f = L_2^2 \\
1 - \dfrac{p_a(y)}{g(y;\beta)} & \text{if } D_f = KL \\
1 - \left\{\dfrac{p_a(y)}{g(y;\beta)}\right\}^2 & \text{if } D_f = \chi^2 \\
1 - \sqrt{\dfrac{p_a(y)}{g(y;\beta)}} & \text{if } D_f = H^2 \\
-\nu_t'\Big\{p_a(y) - g(y;\beta)\Big\}/2 & \text{if } D_f = TV^*.
\end{cases}
$$

In a slight abuse of notation we define

$$\begin{aligned}\|\widehat{\eta}_a - \eta_a\|^2 &= \int \left\{ \int |\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x)| \; dy \right\}^2 \; d\mathbb{P}(x) \\ &\leq \int \{\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x)\}^2 \; d\mathbb{P}(y, x)\end{aligned}$$