One-Step Estimation of Differentiable Hilbert-Valued Parameters

Alex Luedtke

Joint work with Incheoul Chung

Department of Statistics
University of Washington

Acknowledgements

This research was supported by the:

- **NIH** through award numbers DP2-LM013340;
- **NSF** through award number DMS-2210216.

The work that I speak about today is based on the following preprint:

Luedtke, A., & Chung, I. (2023). One-Step Estimation of Differentiable Hilbert-Valued Parameters. arXiv preprint arXiv:2303.16711.

Objective: inference on unknown functions

Goal: infer an unknown function $\nu(P)$ that belongs to a real Hilbert space $\mathcal{H}$.

- Subgoal 1: **Estimate** $\nu(P)$ well, in a norm sense.
- Subgoal 2: Construct **confidence sets** for $\nu(P)$.

Working in a **nonparametric** statistical model.

Running example: counterfactual density function

Observe n iid draws of $Z = (X, A, Y) \sim P$

- X: Covariates
- A: Binary treatment
- Y: Outcome

Aim to estimate density of counterfactual outcome for $A = 1$:

$$\nu(P)(y) = \int p_{Y \mid A,X}(y \mid 1, x)\, P_X(dx)$$

Interpretation of $\nu(P)$ under causal conditions:

What would the density of Y be if, contrary to fact, everyone had received treatment $A = 1$?

Running example: counterfactual density function

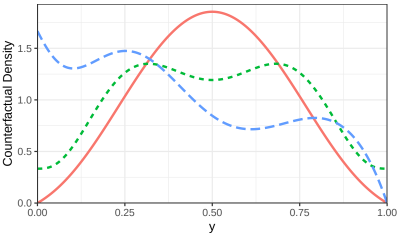


Figure: Three possible counterfactual density functions.

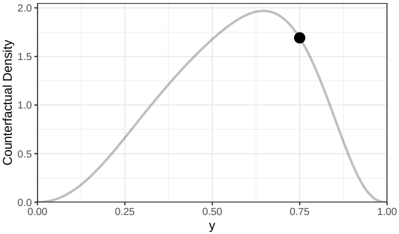
Kennedy et al. (2021) gave estimators of **finite-dimensional projections of $\nu(P)$**

We instead give estimators of the **actual function $\nu(P)$** as an element of $L^2(\lambda)$

In the paper, we show our framework also covers the following examples:

- **Conditional average treatment effect function**
  (Hill, 2011; Nie et al., 2021)
- **Causal dose-response function**
  (Díaz et al., 2013; Kennedy, Ma, et al., 2017)
- **(Counterfactual) kernel mean embedding**
  (Gretton et al., 2012; Muandet et al., 2021)

Existing strategies for estimating the function at a point



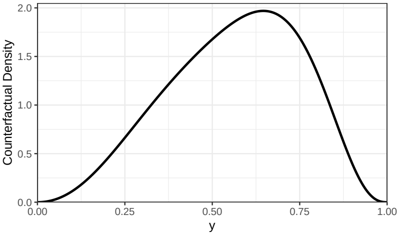
Debiased/targeted machine learning has been adapted to estimate $\nu(P)(y)$.

- e.g., Kennedy et al., 2017; van der Laan et al., 2018; Chernozhukov et al., 2018

Challenges:

- Because $P \mapsto \nu(P)(y)$ is not smooth, **local smoothing** is needed.
- Rate-optimally tuned estimator is **too biased** to facilitate inference.

Existing strategies for estimating the function



Estimation is often performed using tools from statistical learning

- e.g., Foster et al., 2019; van der Laan, 2006; Nie et al., 2021

Challenge: these strategies yield regret guarantees, but not confidence sets

Confidence sets are typically constructed using strategies that are distinct from those used to estimate the function

- e.g., Robins et al., 2008; Luedtke, Carone, et al., 2019; Hudson et al., 2021

Our question

Question: Can **turn-the-crank methods** be developed to infer about $\nu(P)$?

Our finding: Yes!

- And they resemble methods for estimating finite-dimensional quantities.

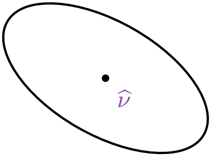
Summary of our approach for smooth parameters ν

One-step estimation:

$$\underbrace{\widehat{\nu}}_{\substack{\text{one-step}\\ \text{estimator}}} = \underbrace{\nu(\widehat{P})}_{\substack{\text{initial}\\ \text{estimator}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \phi_{\widehat{P}}(Z_i)}_{\substack{\text{function-valued}\\ \text{bias correction}}}$$

Wald-type confidence sets:

$$\text{confidence set} = \left\{ f \ : \ \underbrace{n \left\langle \Omega(\widehat{\nu} - f), \widehat{\nu} - f \right\rangle_{\mathcal{H}}}_{\text{quadratic form on } \mathcal{H}} \leq \zeta_n \right\}$$



Key regularity condition

Our approach applies whenever ν is **pathwise differentiable**.

- Surprising finding: many function-valued parameters satisfy this property!

Pathwise differentiability (van der Vaart, 1991; Bickel et al., 1993)

ν is **pathwise differentiable** if there is a bounded linear operator $\dot{\nu}_P : L^2(P) \to \mathcal{H}$ so that, for all smooth submodels $\{P_\epsilon : \epsilon \in \mathbb{R}\}$ with $P_0 = P$ and score s,

$$\frac{\nu(P_\epsilon) - \nu(P)}{\epsilon} \xrightarrow{\epsilon \to 0} \underbrace{\dot{\nu}_P(s)}_{\text{"local parameter"}}$$

Our paper gives **easy-to-check conditions** for verifying pathwise differentiability.

Existence of efficient influence functions

An **efficient influence function (EIF)** of ν is an $\mathcal{H}$-valued map ϕ_P that satisfies

$$\dot{\nu}_P(s) = \int \phi_P(z) s(z) P(dz) \text{ for all } s \in L^2(P)$$

An EIF may not exist.

But, if it does, then, we have the von Mises approximation

$$\nu(P) \;\approx\; \nu(\widehat{P}) + E_P[\phi_{\widehat{P}}(Z)]\,,$$

which suggests the one-step estimator

$$\widehat{\nu} \;=\; \nu(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \phi_{\widehat{P}}(Z_i)\,.$$

Two cases we'll consider

Does an EIF exist?

Yes

No

Two cases we'll consider

Does an EIF exist?

Yes

No

We'll focus on this case first.

When does an EIF exist?

In general:

$\mathcal{H}$ is a **reproducing kernel Hilbert space**

$$\Longrightarrow$$

an EIF exists

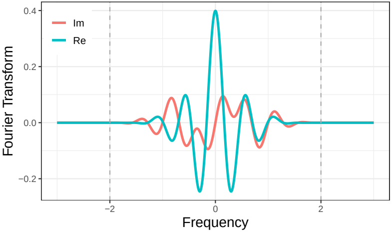
In our counterfactual density example:

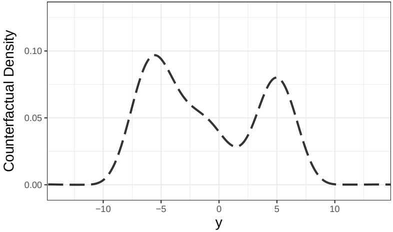
$\nu(P)$ is known to be bandlimited

$$\Longrightarrow$$

an EIF exists

Bandlimited densities (Ibragimov et al., 1983)



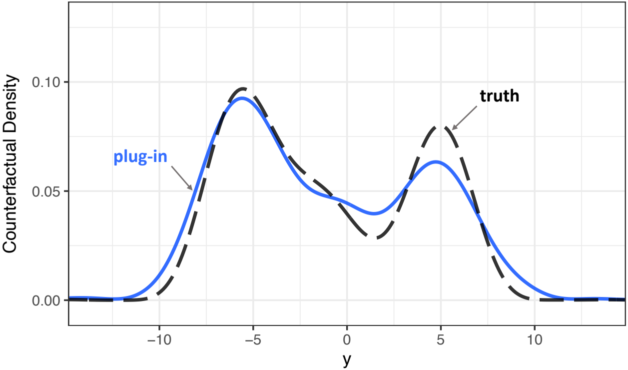


Assuming $\nu(P)$ is bandlimited is a **strong condition**!

- E.g., imposes that $\nu(P)$ must be **infinitely differentiable**.

This condition will be relaxed in the second half of the talk.

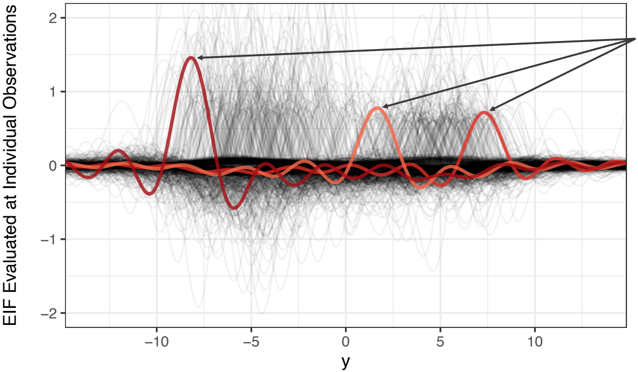
1 Use any statistical learning approach to construct a **plug-in estimator** of the **true counterfactual density**:



Recall: $\nu(P)(y) = \int p_{Y|A,X}(y \mid 1, x)\, P_X(dx)$

Illustration of one-step estimator

2 Evaluate the efficient influence function (EIF) of ν at each of the observations.



Highlighting **3 of the n total EIF evaluations** for illustration purposes.

The other EIF evaluations appear in light grey.

Illustration of one-step estimator

3 Approximate the **estimation error (= truth – plugin)** from plot ① with a **quantity computed using the data alone**.

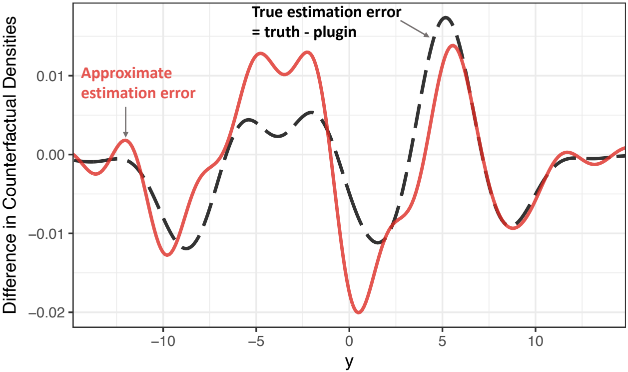
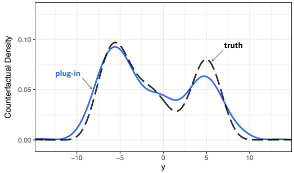
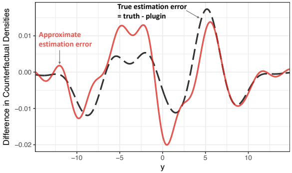


Illustration of one-step estimator

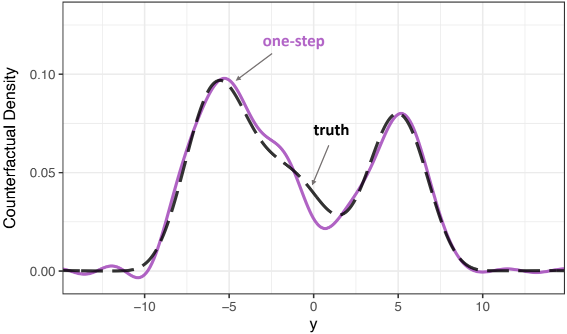
1



3



4 Improve the plug-in estimator from plot ① by adding the approximate estimation error from plot ③, yielding the **one-step estimator**.



Weak convergence

Under conditions, there's a **Gaussian random element** $\mathbb{H}$ such that

$$n^{1/2}[\widehat{\nu} - \nu(P)] \rightsquigarrow \mathbb{H}.$$

Key condition is that $\widehat{P}$ **estimates** P **well enough**.

In our counterfactual density example, this holds if

$$\underbrace{\|\widehat{p}_{A|X} - p_{A|X}\|_{L^2(P)}}_{\text{propensity estimation error}} \cdot \underbrace{\|\widehat{p}_{Y|A=1,X} - p_{Y|A=1,X}\|_{L^2(P)}}_{\text{outcome density estimation error}} = o_p(n^{-1/2}).$$

Weak convergence facilitates the construction of confidence sets

The continuous mapping theorem yields that

$$n\|\widehat{\nu} - \nu(P)\|_{\mathcal{H}}^2 \ \rightsquigarrow \ \|\mathbb{H}\|_{\mathcal{H}}^2.$$

This suggests determining a threshold ζ_n via the **bootstrap** and letting

$$\text{confidence set} \ = \ \left\{ f \ : \ n\|\widehat{\nu} - f\|_{\mathcal{H}}^2 \ \leq \ \zeta_n \right\}.$$

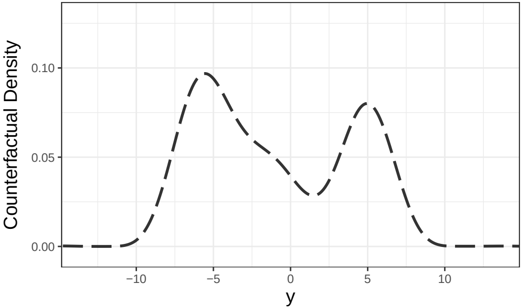
Other quadratic forms can also be used to construct the confidence set.

Simulation to evaluate our approach when an EIF exists

Estimating a **bandlimited counterfactual density function**

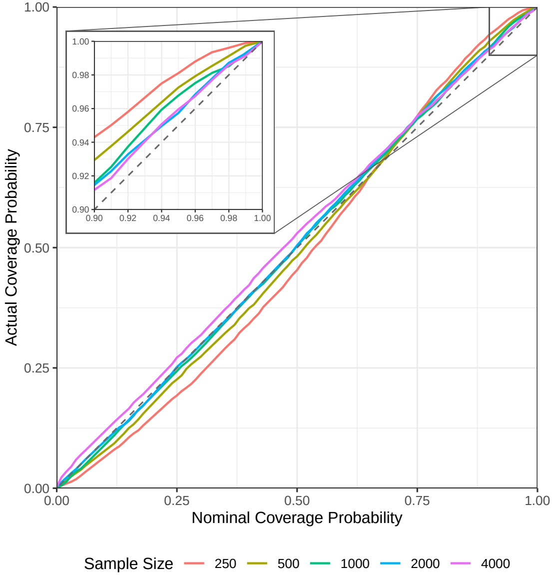
n iid draws of $Z = (X, A, Y) \sim P$ observed

- Covariate X is 5d
- Treatment A depends on X



Coverage of L^2-ball confidence sets

Diameters decay at $n^{-1/2}$ rate



Summary of case where ν has an EIF

Does an EIF exist?

Yes | No

Estimation and inference parallel the 1d case.

Examples of parameters with EIFs:

- Bandlimited counterfactual density function
- (Counterfactual) kernel mean embedding

Moving to case where ν does not have an EIF

Does an EIF exist?

Yes

No

Examples of parameters without EIFs:

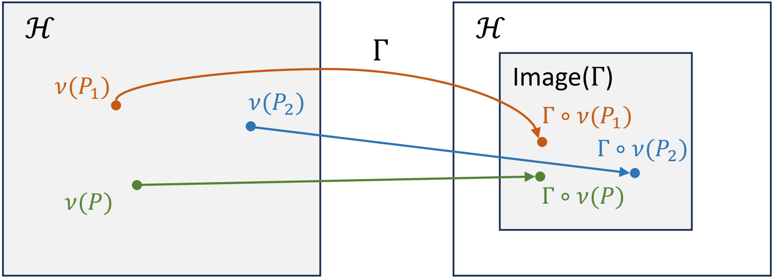
- Non-bandlimited counterfactual density function
- Conditional average treatment effect function
- Causal dose-response function

If there's no EIF, how can we infer about $\nu(P)$?

What we did when there was an EIF simply won't work:

$$\underbrace{\widehat{\nu}}_{\text{one-step estimator}} = \underbrace{\nu(\widehat{P})}_{\text{initial estimator}} + \underbrace{\cancel{\frac{1}{n}\sum_{i=1}^{n}\phi_{\widehat{P}}(Z_i)}}_{\text{how to compute a bias correction?}}$$

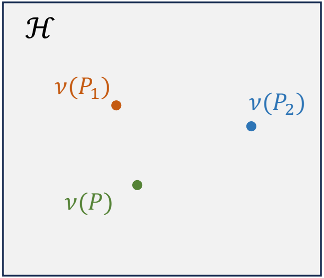
Our proposal: reduce the problem to one that has an EIF

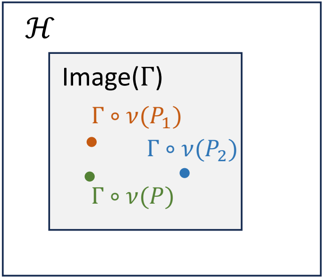


Key observation: There is an **injective** transformation $\Gamma : \mathcal{H} \to \mathcal{H}$ so that

$\Gamma \circ \nu$ **has an EIF**.

Our proposal: reduce the problem to one that has an EIF

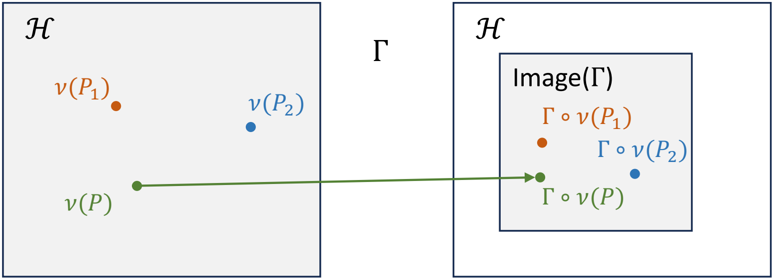




Our proposal:

1. Temporarily make $\Gamma \circ \nu(P)$, rather than $\nu(P)$, the target of inference.

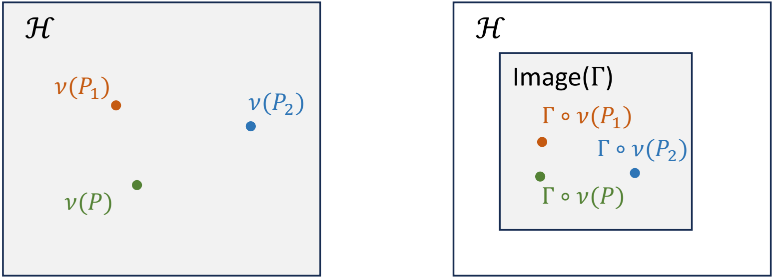
Our proposal: reduce the problem to one that has an EIF



Our proposal:

1. Temporarily make $\Gamma \circ \nu(P)$, rather than $\nu(P)$, the target of inference.

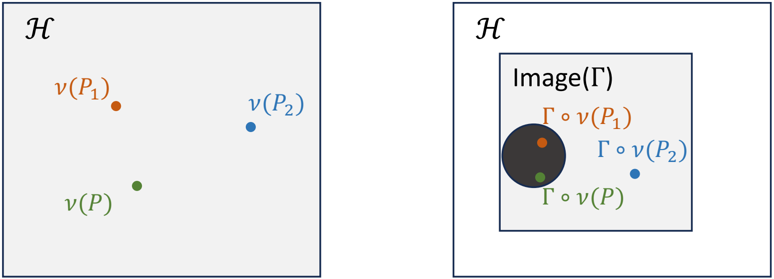
Our proposal: reduce the problem to one that has an EIF



Our proposal:

1. Temporarily make $\Gamma \circ \nu(P)$, rather than $\nu(P)$, the target of inference.

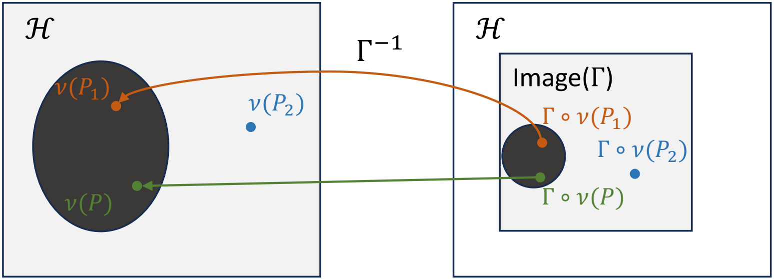
Our proposal: reduce the problem to one that has an EIF



Our proposal:

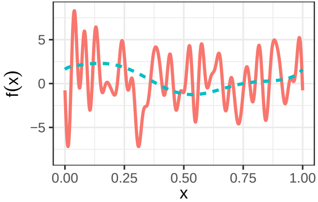
1. Temporarily make $\Gamma \circ \nu(P)$, rather than $\nu(P)$, the target of inference.
2. Construct confidence set $\mathcal{C}_n$ for $\Gamma \circ \nu(P)$.

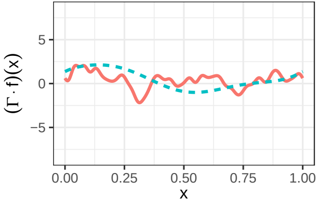
Our proposal: reduce the problem to one that has an EIF



Our proposal:

1. Temporarily make $\Gamma \circ \nu(P)$, rather than $\nu(P)$, the target of inference.
2. Construct confidence set $\mathcal{C}_n$ for $\Gamma \circ \nu(P)$.
3. Use $\Gamma^{-1}(\mathcal{C}_n)$ as a confidence set for $\nu(P)$.





In the paper, we study the confidence sets derived using

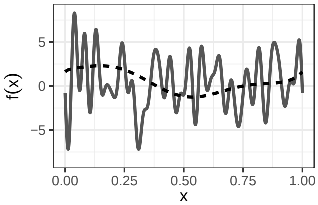
$$\Gamma(f) = \sum_{k=1}^{\infty} \beta_k \langle f, h_k \rangle_{\mathcal{H}} h_k,$$

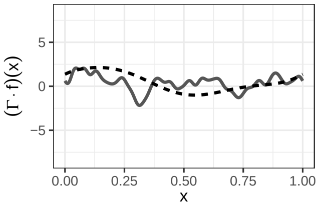
where

- $(h_k)_{k=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$;
- $(\beta_k)_{k=1}^{\infty}$ positive and square summable.

The left inverse of Γ is

$$\Gamma^{-1}(f) = \sum_{k=1}^{\infty} \frac{1}{\beta_k} \langle f, h_k \rangle_{\mathcal{H}} h_k.$$





In the paper, we study the confidence sets derived using

$$\Gamma(f) = \sum_{k=1}^{\infty} \beta_k \langle f, h_k \rangle_{\mathcal{H}} h_k,$$

where

- $(h_k)_{k=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$;
- $(\beta_k)_{k=1}^{\infty}$ positive and square summable.

The left inverse of Γ is

$$\Gamma^{-1}(f) = \sum_{k=1}^{\infty} \frac{1}{\beta_k} \langle f, h_k \rangle_{\mathcal{H}} h_k.$$

Illustration of confidence set

To visualize our confidence sets, we embed each $f \in \mathcal{H}$ into ℓ^2 as

$$\left(\langle f, h_1 \rangle, \langle f, h_2 \rangle, \quad \langle f, h_3 \rangle, \langle f, h_4 \rangle, \quad \langle f, h_5 \rangle, \langle f, h_6 \rangle, \quad \langle f, h_7 \rangle, \langle f, h_8 \rangle, \quad \langle f, h_9 \rangle, \langle f, h_{10} \rangle, \ \ldots \right)$$

Illustration of confidence set

To visualize our confidence sets, we embed each $f \in \mathcal{H}$ into ℓ^2 as

$$\left(\langle f, h_1\rangle, \langle f, h_2\rangle, \quad \langle f, h_3\rangle, \langle f, h_4\rangle, \quad \langle f, h_5\rangle, \langle f, h_6\rangle, \quad \langle f, h_7\rangle, \langle f, h_8\rangle, \quad \langle f, h_9\rangle, \langle f, h_{10}\rangle, \ \ldots\right)$$

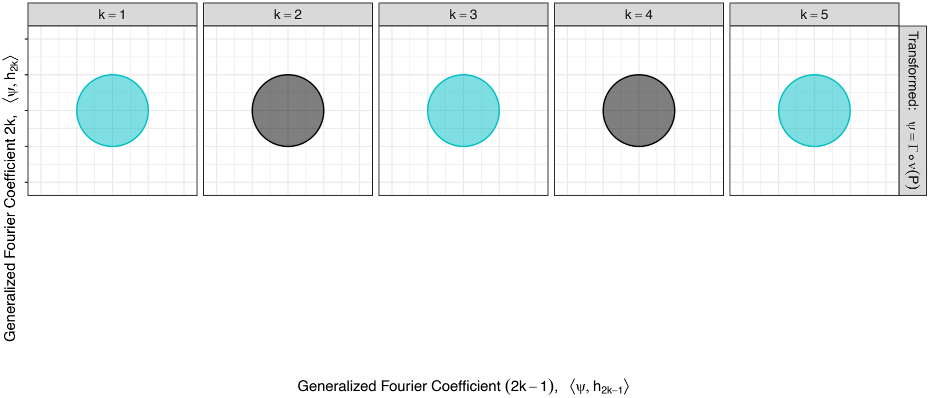


Illustration of confidence set

To visualize our confidence sets, we embed each $f \in \mathcal{H}$ into ℓ^2 as

$$(\langle f, h_1\rangle, \langle f, h_2\rangle, \quad \langle f, h_3\rangle, \langle f, h_4\rangle, \quad \langle f, h_5\rangle, \langle f, h_6\rangle, \quad \langle f, h_7\rangle, \langle f, h_8\rangle, \quad \langle f, h_9\rangle, \langle f, h_{10}\rangle, \ \ldots)$$

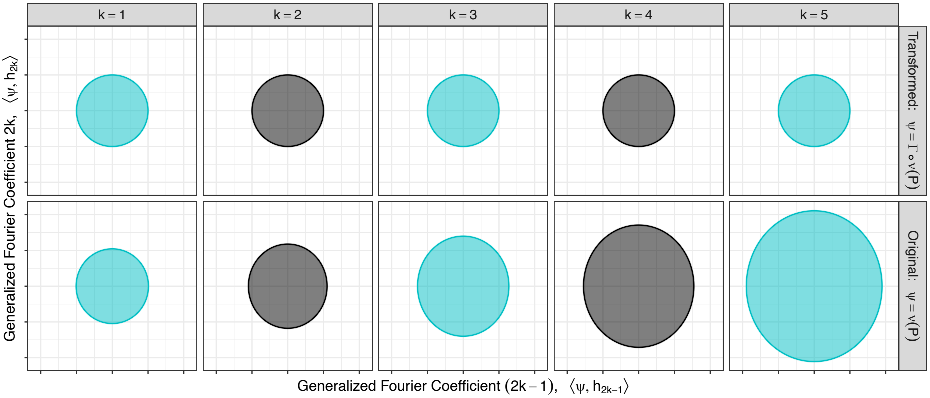
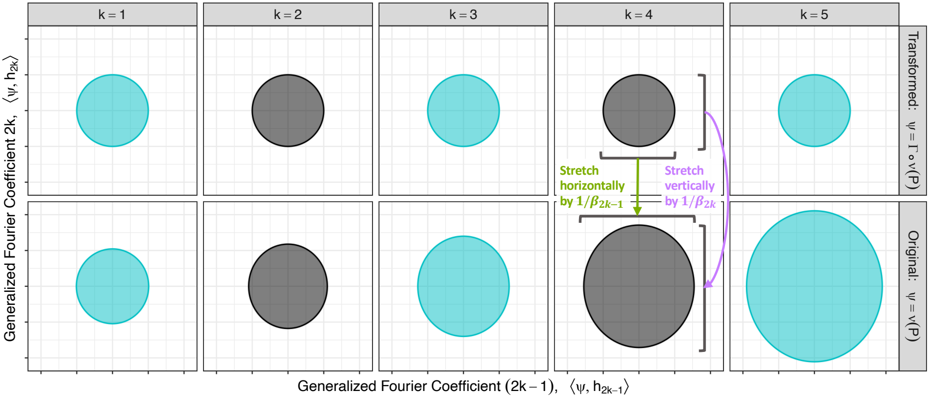


Illustration of confidence set

To visualize our confidence sets, we embed each $f \in \mathcal{H}$ into ℓ^2 as

$$\big(\langle f, h_1\rangle, \langle f, h_2\rangle, \quad \langle f, h_3\rangle, \langle f, h_4\rangle, \quad \langle f, h_5\rangle, \langle f, h_6\rangle, \quad \langle f, h_7\rangle, \langle f, h_8\rangle, \quad \langle f, h_9\rangle, \langle f, h_{10}\rangle, \ \ldots\big)$$



Simulation to evaluate our approach when an EIF does not exist

Test if **counterfactual density functions** under $A = 1$ and $A = 0$ are equal

Implement by checking whether 0 is in a confidence set for

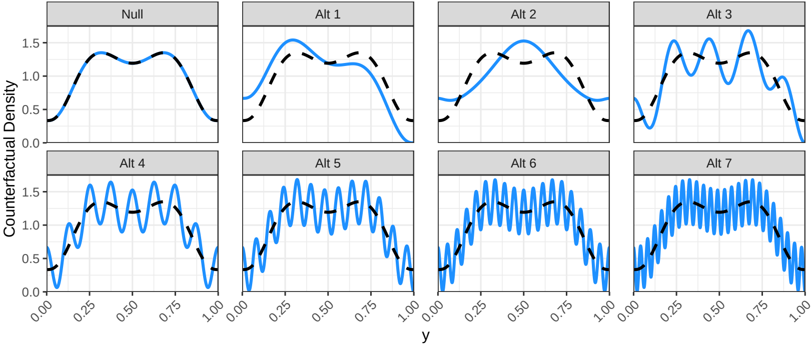
$$\nu(P)(\cdot) = \underbrace{\int p_{Y|A,X}(\,\cdot \mid 1, x)\, P_X(dx)}_{\text{density of } Y(1)} - \underbrace{\int p_{Y|A,X}(\,\cdot \mid 0, x)\, P_X(dx)}_{\text{density of } Y(0)}$$

Simulation to evaluate our approach when an EIF does not exist

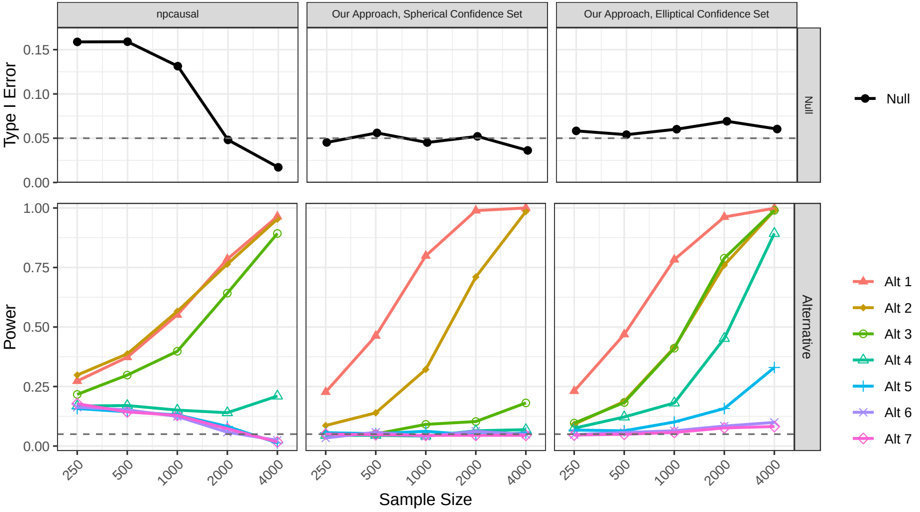
Test if **counterfactual density functions** under $A = 1$ and $A = 0$ are equal

Implement by checking whether 0 is in a confidence set for

$$\nu(P)(\cdot) = \underbrace{\int p_{Y|A,X}(\cdot \mid 1, x)\, P_X(dx)}_{\text{density of } Y(1)} - \underbrace{\int p_{Y|A,X}(\cdot \mid 0, x)\, P_X(dx)}_{\text{density of } Y(0)}$$



Type 1 error and power at different sample sizes



Future work

Do similar approaches enable **uncertainty quantification for other estimators**?

- e.g., orthogonal statistical learning (Foster et al., 2019)

Can we let the **Hilbert space depend on P**?

- e.g., $\mathcal{H} = L^2(P)$

Thank you!

Questions?

References I

Bickel, P. J. et al. (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore.

Chernozhukov, V., W. Newey, and R. Singh (2018). "De-biased machine learning of global and local parameters using regularized Riesz representers". In: *arXiv preprint arXiv:1802.08667*.

Díaz, I. and M. J. van der Laan (2013). "Targeted data adaptive estimation of the causal dose–response curve". In: *Journal of Causal Inference* 1.2, pp. 171–192.

Foster, D. J. and V. Syrgkanis (2019). "Orthogonal statistical learning". In: *arXiv preprint arXiv:1901.09036*.

Gretton, A. et al. (2012). "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1, pp. 723–773.

Hill, J. L. (2011). "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Hudson, A., M. Carone, and A. Shojaie (2021). "Inference on function-valued parameters using a restricted score test". In: *arXiv preprint arXiv:2105.06646*.

Ibragimov, I. and R. Khas'minskii (1983). "Estimation of distribution density belonging to a class of entire functions". In: *Theory of Probability & Its Applications* 27.3, pp. 551–562.

Kennedy, E. H., S. Balakrishnan, and L. Wasserman (2021). "Semiparametric counterfactual density estimation". In: *arXiv preprint arXiv:2102.12034*.

Kennedy, E. H., Z. Ma, et al. (2017). "Non-parametric methods for doubly robust estimation of continuous treatment effects". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 1229–1245.

Luedtke, A., M. Carone, and M. J. van der Laan (2019). "An omnibus non-parametric test of equality in distribution for unknown functions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.1, pp. 75–99.

Luedtke, A. and I. Chung (2023). "One-Step Estimation of Differentiable Hilbert-Valued Parameters". In: *arXiv preprint arXiv:2303.16711*.

Muandet, K. et al. (2021). "Counterfactual Mean Embeddings.". In: *J. Mach. Learn. Res.* 22, pp. 162–1.

Nie, X. and S. Wager (2021). "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2, pp. 299–319.

Robins, J. et al. (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and statistics: essays in honor of David A. Freedman* 2, pp. 335–421.

van der Laan, M. J. (2006). "Statistical inference for variable importance". In: *The International Journal of Biostatistics* 2.1.

van der Laan, M. J., A. Bibaut, and A. R. Luedtke (2018). "CV-TMLE for nonpathwise differentiable target parameters". In: *Targeted Learning in Data Science*. Springer, pp. 455–481.

van der Vaart, A. (1991). "On differentiable functionals". In: *The Annals of Statistics*, pp. 178–204.

Extra Slides

Adding and subtracting terms shows that

$$n^{1/2}[\widehat{\nu} - \nu(P)] = \underbrace{n^{-1/2}\sum_{i=1}^{n} \phi_P(Z_i)}_{\substack{\text{scaled sample mean}\\ \text{of mean-zero function}}} + \underbrace{n^{1/2}\int [\phi_{\widehat{P}}(z) - \phi_P(z)]d(P_n - P)(z)}_{\substack{\text{negligible if } \phi_{\widehat{P}} \to \phi_P\\ \text{in an appropriate sense}}}$$

$$+ \underbrace{n^{1/2}\left(\nu(\widehat{P}) - \nu(P) + E_P[\phi_{\widehat{P}}(Z)]\right)}_{\substack{\text{negligible under typical}\\ n^{-1/4}\text{-rate conditions on } \widehat{P}}}$$

Establishing weak convergence

Adding and subtracting terms shows that

$$n^{1/2}[\widehat{\nu} - \nu(P)] = \underbrace{n^{-1/2}\sum_{i=1}^{n}\phi_P(Z_i)}_{\substack{\text{scaled sample mean}\\ \text{of mean-zero function}}} + \underbrace{n^{1/2}\int[\phi_{\widehat{P}}(z) - \phi_P(z)]d(P_n - P)(z)}_{\substack{\text{negligible if } \phi_{\widehat{P}} \to \phi_P\\ \text{in an appropriate sense}}}$$

$$+ \underbrace{n^{1/2}\left(\nu(\widehat{P}) - \nu(P) + E_P[\phi_{\widehat{P}}(Z)]\right)}_{\substack{\text{negligible under typical}\\ n^{-1/4}\text{-rate conditions on } \widehat{P}}}$$

Key point: If terms 2 and 3 are $o_p(1)$, then, by the CLT:

$$n^{1/2}[\widehat{\nu} - \nu(P)] \rightsquigarrow \mathbb{H},$$

where $\mathbb{H}$ is a **Gaussian random element**.

Weak convergence facilitates the construction of confidence sets

For any continuous linear operator Ω,

$$n \left\langle \Omega \left[\widehat{\nu} - \nu(P)\right], \widehat{\nu} - \nu(P) \right\rangle_{\mathcal{H}} \rightsquigarrow \langle \Omega(\mathbb{H}), \mathbb{H} \rangle_{\mathcal{H}}.$$

This suggests determining a threshold ζ_n via the **bootstrap** and letting

$$\text{confidence set} = \Big\{ h \ : \ n \langle \Omega(\widehat{\nu} - h), \widehat{\nu} - h \rangle_{\mathcal{H}} \leq \zeta_n \Big\}.$$

Candidate choices of Ω:

- Identity operator
- (Regularized) inverse covariance operator of $\mathbb{H}$

EIF and form of estimator in counterfactual density example

The density of counterfactual outcome for $A = 1$ is given by:

$$\nu(P)(y) = \int p_{Y|A,X}(y \mid 1, x)\, P_X(dx)$$

We estimate its b-**bandlimiting**,

$$\underline{\nu}(P)(y) := \int_{-\infty}^{\infty} K_y(\tilde{y})\, \nu(P)(\tilde{y})\, d\tilde{y},$$

where $K_y(\tilde{y}) := \{\sin[b(\tilde{y} - y)]\}/[\pi(\tilde{y} - y)]$.

The EIF takes the form

$$\begin{aligned}\underline{\phi}_P(y, a, x) = &\frac{1\{a = 1\}}{g_P(a \mid x)} \{K_y - E_P[K_Y \mid A = a, X = x]\} \\ &+ E_P[K_Y \mid A = 1, X = x] - \underline{\nu}(P).\end{aligned}$$

Regularized one-step estimation

Though some pathwise differentiable parameters do not have EIFs, all of them have what we call **regularized EIFs**.

$$\underbrace{\widehat{\nu}}_{\substack{\text{regularized}\\ \text{one-step}\\ \text{estimator}}} = \underbrace{\nu(\widehat{P})}_{\substack{\text{initial}\\ \text{estimator}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\phi_{\mathrm{reg},\widehat{P}}(Z_i)}_{\substack{\text{empirical mean of}\\ \textit{regularized}\text{ EIF}}}$$

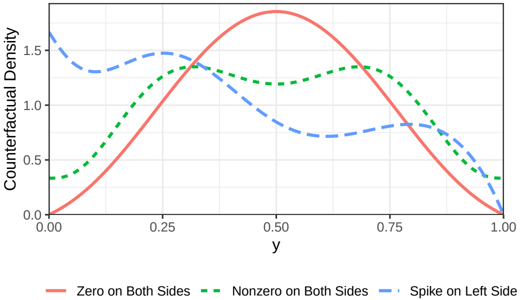
The level of regularization can be tuned via **cross-validation**.

Simulation to evaluate our approach when an EIF does not exist

Estimating a non-bandlimited **counterfactual density function**

n iid draws of $Z = (X, A, Y) \sim P$ observed

- Covariate X is 5d
- Treatment A depends on X



Mean integrated squared error vs. sample size in non-bandlimited settings