Introducing the Forster-Warmuth Nonparametric Counterfactual Regression

Eric J Tchetgen Tchetgen
University Professor
Department of Statistics and Data Science
The Wharton School
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine
University of Pennsylvania

Wayne A Fuller Distinguished Lecture, Iowa State U
09/30/2023

Outline

- Nonparametric Series Regression in Statistics

Outline

- Nonparametric Series Regression in Statistics
- Forster-Warmuth Nonparametric Series Regression

Outline

- Nonparametric Series Regression in Statistics
- Forster-Warmuth Nonparametric Series Regression
- Nonparametric Counterfactual Regression

Outline

- Nonparametric Series Regression in Statistics
- Forster-Warmuth Nonparametric Series Regression
- Nonparametric Counterfactual Regression
- Forster-Warmuth Nonparametric Counterfactual Regression

Outline

- Nonparametric Series Regression in Statistics
- Forster-Warmuth Nonparametric Series Regression
- Nonparametric Counterfactual Regression
- Forster-Warmuth Nonparametric Counterfactual Regression
- Simulation Studies and an Application

Outline

- Nonparametric Series Regression in Statistics
- Forster-Warmuth Nonparametric Series Regression
- Nonparametric Counterfactual Regression
- Forster-Warmuth Nonparametric Counterfactual Regression
- Simulation Studies and an Application
- Conclusion

Nonparametric Regression in Statistics

- Plays a central role in Statistics more broadly scientific inquiry where one seeks understanding of conditional mean function $E(Y|X)$ without a priori restriction on model.

Nonparametric Regression in Statistics

- Plays a central role in Statistics more broadly scientific inquiry where one seeks understanding of conditional mean function $E(Y|X)$ without a priori restriction on model.
  - Example: conditional cumulative distribution $E\left[1\left\{Y \leq t\right\}|X = x\right]$ $\Rightarrow$conditional quantiles.

Nonparametric Regression in Statistics

- Plays a central role in Statistics more broadly scientific inquiry where one seeks understanding of conditional mean function $E(Y|X)$ without a priori restriction on model.
  - Example: conditional cumulative distribution $E\left[1\left\{Y \leq t\right\}|X = x\right]$ $\Rightarrow$conditional quantiles.
  - Example: any conditional function defined by $\theta^{*}(x) = \arg\min_{\theta \in \mathbb{R}} E\left[\rho\left(Y, X; \theta\right)|X = x\right]$ for a given loss function leverages conditional means

Nonparametric Regression in Statistics

- Plays a central role in Statistics more broadly scientific inquiry where one seeks understanding of conditional mean function $E(Y|X)$ without a priori restriction on model.
  - Example: conditional cumulative distribution $E\left[1\left\{Y \leq t\right\}|X = x\right]$ $\Rightarrow$conditional quantiles.
  - Example: any conditional function defined by $\theta^{*}\left(x\right) = \arg\min_{\theta \in \mathbb{R}} E\left[\rho\left(Y, X; \theta\right)|X = x\right]$ for a given loss function leverages conditional means
- Series or Sieve estimation approximates unknown conditional mean based on linear combination of K basis functions $\overline{\phi}_{K}^{\mathsf{T}}\left(x\right) = \left[\phi_{1}\left(x\right), ...\phi_{K}\left(x\right)\right]$ and where K may grow with sample size n.

Nonparametric Regression in Statistics

- Plays a central role in Statistics more broadly scientific inquiry where one seeks understanding of conditional mean function $E(Y|X)$ without a priori restriction on model.
  - Example: conditional cumulative distribution $E\left[1\left\{Y \leq t\right\}|X = x\right]$ $\Rightarrow$conditional quantiles.
  - Example: any conditional function defined by $\theta^{*}(x) = \arg\min_{\theta \in \mathbb{R}} E\left[\rho\left(Y, X; \theta\right)|X = x\right]$ for a given loss function leverages conditional means
- Series or Sieve estimation approximates unknown conditional mean based on linear combination of K basis functions $\overline{\phi}_K^{\mathsf{T}}(x) = \left[\phi_1(x), ...\phi_K(x)\right]$ and where K may grow with sample size n.
  - Balancing bias and variance gives best rate for $k(n)$ in minimax sense.

Nonparametric Regression in Statistics

- One of oldest and most straightforward empirical approach to construct a series estimator is by method of least-squares (LS).

Nonparametric Regression in Statistics

- One of oldest and most straightforward empirical approach to construct a series estimator is by method of least-squares (LS).
- In this vein, given data $\{X_i, Y_i : i = 1, ..., n\}$, we aim to estimate $m^*(x) = E[Y|X = x]$ using LS series estimator defined by:

$$\widehat{m}(x) = \overline{\phi}_K^{\mathsf{T}}(x)\,\widehat{\beta} \text{ where } \widehat{\beta} = \left(\Phi_K^{\mathsf{T}}\Phi_K\right)^{-1}\Phi_k^{\mathsf{T}}\mathbf{Y}$$

where Φ_k is the $n \times k$ matrix $\left[\overline{\phi}_k(X_1), ..., \overline{\phi}_k(X_n)\right]^{\mathsf{T}}$ and $\mathbf{Y} = (Y_1, ..., Y_n)^{\mathsf{T}}$.

Nonparametric Regression in Statistics

- One of oldest and most straightforward empirical approach to construct a series estimator is by method of least-squares (LS).
- In this vein, given data $\{X_i, Y_i : i = 1, ..., n\}$, we aim to estimate $m^*(x) = E[Y|X = x]$ using LS series estimator defined by:

$$\widehat{m}(x) = \overline{\phi}_K^{\mathsf{T}}(x)\,\widehat{\beta} \text{ where } \widehat{\beta} = \left(\Phi_K^{\mathsf{T}}\Phi_K\right)^{-1}\Phi_k^{\mathsf{T}}\mathbf{Y}$$

  where Φ_k is the $n \times k$ matrix $\left[\overline{\phi}_k(X_1), ..., \overline{\phi}_k(X_n)\right]^{\mathsf{T}}$ and $\mathbf{Y} = (Y_1, ..., Y_n)^{\mathsf{T}}$.
- LS series estimator extensively studied in literature; several papers provide sufficient conditions for consistency, corresponding convergence rates, and asymptotic normality in case of polynomial series, regression splines, fourier series, Wavelets series, local polynomial partition series for example,see Chen (2007), Newey (1997), Györfi et al. (2002), Huang (2003), Cattaneo and Farrell (2013). Belloni et al. (2015).

Nonparametric Regression in Statistics

- These prior works established that under certain conditions on the basis functions $\overline{\phi}_k$,for X d-dimensional and m^* in Holder Space $\mathcal{H}\left(\beta_m, C_m\right)$ with $\lfloor\beta_m\rfloor$ bounded derivatives; setting $K_{opt} = n^{\frac{d}{2\beta_m+d}}$:

$$\|m^* - \widehat{m}_{LS}\|_2 = \sqrt{\int \left(m^*(x) - \widehat{m}_{LS}(x)\right)^2 dF(x)} \lesssim_P n^{-\frac{\beta_m}{2\beta_m+d}}$$

attains the L_2-minimax rate of convergence (Stone, 1980) in *probability*.

Nonparametric Regression in Statistics

- These prior works established that under certain conditions on the basis functions $\overline{\phi}_k$,for X d-dimensional and m^* in Holder Space $\mathcal{H}\left(\beta_m, C_m\right)$ with $\lfloor \beta_m \rfloor$ bounded derivatives; setting $K_{opt} = n^{\frac{d}{2\beta_m + d}}$:

$$\|m^* - \widehat{m}_{LS}\|_2 = \sqrt{\int \left(m^*(x) - \widehat{m}_{LS}(x)\right)^2 dF(x)} \lesssim_P n^{-\frac{\beta_m}{2\beta_m + d}}$$

  attains the L_2-minimax rate of convergence (Stone, 1980) in *probability*.
- Specifically, this is known to hold for local basis functions including Fourier, Wavelets, Splines (Newey, 1997, Huang, 2003, Belloni et al, 2015) and local polynomial partition series, Cattaneo and Farrell (2013) for all $\beta_m, d > 0$.

Nonparametric Regression in Statistics

- These prior works established that under certain conditions on the basis functions $\overline{\phi}_k$,for X d-dimensional and m^* in Holder Space $\mathcal{H}\left(\beta_m, C_m\right)$ with $\lfloor \beta_m \rfloor$ bounded derivatives; setting $K_{opt} = n^{\frac{d}{2\beta_m + d}}$:

$$\|m^* - \widehat{m}_{LS}\|_2 = \sqrt{\int \left(m^*(x) - \widehat{m}_{LS}(x)\right)^2 dF(x)} \lesssim_P n^{-\frac{\beta_m}{2\beta_m + d}}$$

  attains the L_2-minimax rate of convergence (Stone, 1980) in *probability*.
- Specifically, this is known to hold for local basis functions including Fourier, Wavelets, Splines (Newey, 1997, Huang, 2003, Belloni et al, 2015) and local polynomial partition series, Cattaneo and Farrell (2013) for all $\beta_m, d > 0$.
- In contrast, for polynomial series only known to hold only for $\beta_m \geq d/2$ (Belloni et al, 2015).

Nonparametric Regression in Statistics

- These results critically rely on "size" of basis functions, as well as its ability to approximate smooth functions.

Nonparametric Regression in Statistics

- These results critically rely on "size" of basis functions, as well as its ability to approximate smooth functions.
- Specifically, above optimality of LS estimator requires that

$$\sup_{x \in \mathcal{X}} ||\overline{\phi}_K(x)|| \lesssim \sqrt{K} \tag{1}$$

and that for any K, there exist a vector β such that

$$\left\| m^* - \overline{\phi}_K^{\mathsf{T}} \beta \right\|_{L_2(\mu)} \lesssim K^{-2\beta_m/d} \tag{2}$$

which essentially limits results to local basis functions.

Forster-Warmuth nonparametric Regression in Statistics

- We have developed a new type of series regression estimator that in principle can attain well-established minimax nonparametric rates of estimation in settings where covariates and outcomes are fully observed, under weaker conditions compared to existing literature (e.g. Belloni et al. (2015)) on the distribution of covariates and basis functions.

Forster-Warmuth nonparametric Regression in Statistics

- We have developed a new type of series regression estimator that in principle can attain well-established minimax nonparametric rates of estimation in settings where covariates and outcomes are fully observed, under weaker conditions compared to existing literature (e.g. Belloni et al. (2015)) on the distribution of covariates and basis functions.
- The approach builds on an estimator we refer to as Forster–Warmuth series estimator (FW-estimator) originating in the online learning literature, which is obtained via a careful modification of the renowned non-linear Vovk–Azoury–Warmuth forecaster (Vovk, 2001; Forster and Warmuth, 2002).

Forster-Warmuth Nonparametric Regression in Statistics

- We define the FW-regression as followed:

$$\widehat{\beta} = \underbrace{(1 - h_n)}_{(a)} \overline{\phi}_K^{\mathsf{T}}(x) \left(\Phi_K^{\mathsf{T}} \Phi_K + \underbrace{\overline{\phi}_K(x) \overline{\phi}_K^{\mathsf{T}}(x)}_{(b)} \right)^{-1} \Phi_k^{\mathsf{T}} \mathbf{Y}$$

where

$$h_n = \overline{\phi}_K^{\mathsf{T}}(x) \left(\Phi_K^{\mathsf{T}} \Phi_K + \overline{\phi}_K(x) \overline{\phi}_K^{\mathsf{T}}(x) \right)^{-1} \overline{\phi}_K(x) \in (0, 1)$$

Forster-Warmuth Nonparametric Regression in Statistics

- We define the FW-regression as followed:

$$\widehat{\beta} = \underbrace{(1 - h_n)}_{(a)} \overline{\phi}_K^{\mathsf{T}}(x) \left(\Phi_K^{\mathsf{T}} \Phi_K + \underbrace{\overline{\phi}_K(x) \overline{\phi}_K^{\mathsf{T}}(x)}_{(b)} \right)^{-1} \Phi_k^{\mathsf{T}} \mathbf{Y}$$

where

$$h_n = \overline{\phi}_K^{\mathsf{T}}(x) \left(\Phi_K^{\mathsf{T}} \Phi_K + \overline{\phi}_K(x) \overline{\phi}_K^{\mathsf{T}}(x) \right)^{-1} \overline{\phi}_K(x) \in (0, 1)$$

- (b) incorporates incorporates a fake observation at x to design matrix, while (a) adjusts for the leverage of this observation if extreme relative to observed basis functions. This effectively controls the size of the projection matrix at the value of interest x regardless of the basis function.

Forster-Warmuth Nonparametric Regression in Statistics

Theorem

(Yang, Kuchibhotla, TT, 2023) Suppose that $E(Y^2|X)$ is bounded almost surely and the basis functions $\overline{\phi}_K^{\mathsf{T}}$ satisfy the optimal approximation property that for each K, there exist a β such that

$$\left\| m^* - \overline{\phi}_K^{\mathsf{T}} \beta \right\|_{L_2(\mu)} \lesssim K^{-2\beta_m/d}$$

then letting $K_{opt} = n^{\frac{d}{2\beta_m + d}}$

$$E\left(m^*(X) - \widehat{m}_{FW}(X)\right)^2 \lesssim n^{-\frac{2\beta_m}{2\beta_m + d}}$$

regardless of the basis functions used.

Forster-Warmuth Nonparametric Regression in Statistics

- This theorem establishes that the FW-learner is minimax rate optimal (wrt MSE) for any basis system that constitutes a full approximation set in the sense of Lorentz (1966) and Yang and Barron (1999), including polynomial basis for all $\beta_m, d > 0$.

Forster-Warmuth Nonparametric Regression in Statistics

- This theorem establishes that the FW-learner is minimax rate optimal (wrt MSE) for any basis system that constitutes a full approximation set in the sense of Lorentz (1966) and Yang and Barron (1999), including polynomial basis for all $\beta_m, d > 0$.
- Corrections (a) and (b) in the FW-learner are key to this result and obviate the need to control the size of the basis function. Note also that (a) and (b) become negligible in large samples and $\widehat{m}_{FW}(X) \approx \widehat{m}_{LS}(X)$ essentially does away with the size control.

Nonparametric Counterfactual Nonparametric Regression

- In many practical applications in health and social sciences it is not unusual for an outcome to be missing on some subjects, either by design, say in two-stage sampling studies where the outcome can safely be assumed to be missing at random with known non-response mechanism, or by happenstance, in which case the outcome might be missing not at random.

Nonparametric Counterfactual Nonparametric Regression

- In many practical applications in health and social sciences it is not unusual for an outcome to be missing on some subjects, either by design, say in two-stage sampling studies where the outcome can safely be assumed to be missing at random with known non-response mechanism, or by happenstance, in which case the outcome might be missing not at random.
- An example of the former type might be a study (Cornelis et al., 2009) in which one aims to develop a polygenic risk prediction model for type-2 diabetes based on stage 1 fully observed covariate data on participants including high dimensional genotype (i.e., SNPs), age, and gender,

Nonparametric Counterfactual Nonparametric Regression

- In many practical applications in health and social sciences it is not unusual for an outcome to be missing on some subjects, either by design, say in two-stage sampling studies where the outcome can safely be assumed to be missing at random with known non-response mechanism, or by happenstance, in which case the outcome might be missing not at random.
- An example of the former type might be a study (Cornelis et al., 2009) in which one aims to develop a polygenic risk prediction model for type-2 diabetes based on stage 1 fully observed covariate data on participants including high dimensional genotype (i.e., SNPs), age, and gender,
- While costly manual chart review by a panel of physicians yield reliable type-2 diabetes labels on a subset of subjects with known selection probability based on stage-1 covariates.

Nonparametric Counterfactual Nonparametric Regression

- In contrast, an example of the latter type might be a household survey in Zambia (Marden et al., 2018) in which eligible household members are asked to test for HIV, however, nearly 30% decline the test and thus have missing HIV status.

Nonparametric Counterfactual Nonparametric Regression

- In contrast, an example of the latter type might be a household survey in Zambia (Marden et al., 2018) in which eligible household members are asked to test for HIV, however, nearly 30% decline the test and thus have missing HIV status.
- The concern here might be that participants who decline to test might not be a priori exchangeable with participants who agree to test for HIV with respect to key risk factors for HIV infection, even after adjusting for fully observed individual and household characteristics collected in the household survey.

Nonparametric Counterfactual Nonparametric Regression

- In contrast, an example of the latter type might be a household survey in Zambia (Marden et al., 2018) in which eligible household members are asked to test for HIV, however, nearly 30% decline the test and thus have missing HIV status.
- The concern here might be that participants who decline to test might not be a priori exchangeable with participants who agree to test for HIV with respect to key risk factors for HIV infection, even after adjusting for fully observed individual and household characteristics collected in the household survey.
- Any effort to build an HIV risk regression model that generalizes to the wider population of Zambia requires carefully accounting for HIV status possibly missing not at random for a non-negligible fraction of the sample.

Nonparametric Counterfactual Nonparametric Regression

- Beyond missing data, counterfactual regression also arises in causal inference where one might be interested in the CATE, the average causal effect experienced by a subset of the population defined in terms of observed covariates.

Nonparametric Counterfactual Nonparametric Regression

- Beyond missing data, counterfactual regression also arises in causal inference where one might be interested in the CATE, the average causal effect experienced by a subset of the population defined in terms of observed covariates.
- Missing data, in this case, arises as the causal effect defined at the individual level as a difference between two potential outcomes – one for each treatment value – can never be observed.

Nonparametric Counterfactual Nonparametric Regression

- Beyond missing data, counterfactual regression also arises in causal inference where one might be interested in the CATE, the average causal effect experienced by a subset of the population defined in terms of observed covariates.
- Missing data, in this case, arises as the causal effect defined at the individual level as a difference between two potential outcomes – one for each treatment value – can never be observed.
- This is because under the consistency assumption the observed outcome for subjects who actually received treatment matches their potential outcome under treatment, while their potential outcome under no treatment is missing, and vice-versa for the untreated.

Nonparametric Counterfactual Nonparametric Regression

- A major contribution of this paper is to propose a generic construction of a so-called pseudo-outcome which, as its name suggests, replaces the unobserved outcome with a carefully constructed response variable that

Nonparametric Counterfactual Nonparametric Regression

- A major contribution of this paper is to propose a generic construction of a so-called pseudo-outcome which, as its name suggests, replaces the unobserved outcome with a carefully constructed response variable that
  - (i) only depends on the observed data, possibly involving high dimensional nuisance functions that can nonetheless be identified from the observed data (e.g. propensity score), and therefore can be evaluated for all subjects in the sample and;

Nonparametric Counterfactual Nonparametric Regression

- A major contribution of this paper is to propose a generic construction of a so-called pseudo-outcome which, as its name suggests, replaces the unobserved outcome with a carefully constructed response variable that
  - (i) only depends on the observed data, possibly involving high dimensional nuisance functions that can nonetheless be identified from the observed data (e.g. propensity score), and therefore can be evaluated for all subjects in the sample and;
  - (ii) has conditional expectation given covariates that matches the counterfactual regression of interest if as for an oracle, nuisance functions were known.

Nonparametric Counterfactual Nonparametric Regression

- The proposed pseudo-outcome approach applies to a large class of counterfactual regression problems including the missing data and causal inference problems described above.

Nonparametric Counterfactual Nonparametric Regression

- The proposed pseudo-outcome approach applies to a large class of counterfactual regression problems including the missing data and causal inference problems described above.
- The proposed approach recovers in specific cases such as the CATE under unconfoundedness, previously proposed forms of pseudo-outcomes (Kennedy, 2020), while offering new pseudo-outcome constructions in several new, more challenging counterfactual prediction pbs.

Nonparametric Counterfactual Nonparametric Regression

- We describe FW-Counterfactual Regression for estimation of the CATE under unconfoundedness

$$m^{*}\left(x\right)=E\left(Y^{a=1}-Y^{a=0}|X=x\right)$$

where Y^{a} is a counterfactual outcome had a patient, possibly contrary to fact been treated with $a=0,1$.

Nonparametric Counterfactual Nonparametric Regression

- We describe FW-Counterfactual Regression for estimation of the CATE under unconfoundedness

$$m^{*}\left(x\right)=E\left(Y^{a=1}-Y^{a=0}|X=x\right)$$

  where Y^{a} is a counterfactual outcome had a patient, possibly contrary to fact been treated with $a=0,1$.
- By consistency $Y=AY^{a=1}+\left(1-A\right)Y^{a=0}$, that is $Y^{a=0}$ is missing for treated patients, and $Y^{a=1}$ is missing for untreated patients.

Nonparametric Counterfactual Nonparametric Regression

- We describe FW-Counterfactual Regression for estimation of the CATE under unconfoundedness

$$m^{*}\left(x\right)=E\left(Y^{a=1}-Y^{a=0}|X=x\right)$$

  where Y^{a} is a counterfactual outcome had a patient, possibly contrary to fact been treated with $a=0,1$.
- By consistency $Y=AY^{a=1}+\left(1-A\right)Y^{a=0}$, that is $Y^{a=0}$ is missing for treated patients, and $Y^{a=1}$ is missing for untreated patients.
- $m^{*}\left(x\right)$ is central to precision medicine as not only does it quantify effect heterogeneity by x, i.e. the extent to which X predicts a person's causal effect, it also can be used to decide whether or not a patient should be treated, i.e. if $m^{*}\left(x\right)>0$ if larger values of Y are desirable.

Nonparametric Counterfactual Nonparametric Regression

- An oracle with access to $Y^{a=1} - Y^{a=0}$ could in principle perform FW-regression to obtain a minimax rate optimal estimator of the CATE $m^*(x) = E\left(Y^{a=1} - Y^{a=0} | X = x\right)$.

Nonparametric Counterfactual Nonparametric Regression

- An oracle with access to $Y^{a=1} - Y^{a=0}$ could in principle perform FW-regression to obtain a minimax rate optimal estimator of the CATE $m^*(x) = E\left(Y^{a=1} - Y^{a=0} | X = x\right)$.
- Note that in a randomized experiment, where $A \sim$Bernoulli$(\frac{1}{2})$ is independent of $\left(Y^{a=1}, Y^{a=0}, X\right)$, one can match the oracle rate by replacing $Y^{a=1} - Y^{a=0}$ with the oracle pseudo-outcome $H = 2(-1)^{1-A} Y$ and running the FW-regression using H, upon noting that $E(H|X = x) = E\left(Y^{a=1} - Y^{a=0} | X = x\right)$.

Nonparametric Counterfactual Nonparametric Regression

- An oracle with access to $Y^{a=1} - Y^{a=0}$ could in principle perform FW-regression to obtain a minimax rate optimal estimator of the CATE $m^*(x) = E\left(Y^{a=1} - Y^{a=0} | X = x\right)$.
- Note that in a randomized experiment, where $A \sim$Bernoulli$(\frac{1}{2})$ is independent of $\left(Y^{a=1}, Y^{a=0}, X\right)$, one can match the oracle rate by replacing $Y^{a=1} - Y^{a=0}$ with the oracle pseudo-outcome $H = 2(-1)^{1-A} Y$ and running the FW-regression using H, upon noting that $E(H|X = x) = E\left(Y^{a=1} - Y^{a=0} | X = x\right)$.
- Neither approach is feasible in observational setting where A is not randomized. Therefore, the most one can hope for is to construct an estimator $\widehat{H}$ of the pseudo-outcome H with "small bias".

Nonparametric Counterfactual Nonparametric Regression

- Constructing a pseudo outcome requires a causal identification condition. The most common assumption is that of no unmeasured confounding given X, That is A is independent of $\left(Y^{a=1}, Y^{a=0}\right)$ conditional on X.

Nonparametric Counterfactual Nonparametric Regression

- Constructing a pseudo outcome requires a causal identification condition. The most common assumption is that of no unmeasured confounding given X, That is A is independent of $\left(Y^{a=1}, Y^{a=0}\right)$ conditional on X.
- In such case, candidate pseudo-oucomes include:

Nonparametric Counterfactual Nonparametric Regression

- Constructing a pseudo outcome requires a causal identification condition. The most common assumption is that of no unmeasured confounding given X, That is A is independent of $\left(Y^{a=1}, Y^{a=0}\right)$ conditional on X.
- In such case, candidate pseudo-oucomes include:
  - IPTW: $\widehat{H} = (-1)^{1-A} Y / \widehat{f}(A|X)$ where $\widehat{f}(A|X)$ is NP estimator in separate sample

Nonparametric Counterfactual Nonparametric Regression

- Constructing a pseudo outcome requires a causal identification condition. The most common assumption is that of no unmeasured confounding given X, That is A is independent of $\left(Y^{a=1}, Y^{a=0}\right)$ conditional on X.
- In such case, candidate pseudo-oucomes include:
  - IPTW: $\widehat{H} = (-1)^{1-A} Y / \widehat{f}(A|X)$ where $\widehat{f}(A|X)$ is NP estimator in separate sample
  - g-computation: $\widehat{H} = \widehat{E}(Y|A=1,X) - \widehat{E}(Y|A=0,X)$ where $\widehat{E}(Y|A=a,X)$ is NP estimator in separate sample

Nonparametric Counterfactual Nonparametric Regression

- Constructing a pseudo outcome requires a causal identification condition. The most common assumption is that of no unmeasured confounding given X, That is A is independent of $\left(Y^{a=1}, Y^{a=0}\right)$ conditional on X.
- In such case, candidate pseudo-oucomes include:
  - IPTW: $\widehat{H} = (-1)^{1-A} Y / \widehat{f}(A|X)$ where $\widehat{f}(A|X)$ is NP estimator in separate sample
  - g-computation: $\widehat{H} = \widehat{E}(Y|A=1,X) - \widehat{E}(Y|A=0,X)$ where $\widehat{E}(Y|A=a,X)$ is NP estimator in separate sample
- Unfortunately these natural choices do not lead to small bias, as $E\left(\widehat{H}|X\right) - E(H|X)$ dominated by bias of $\widehat{f}(A|X)$ for IPTW and by $\widehat{E}(Y|A=a,X)$ for g-computation, therefore Oracle minimax rate of the CATE may not be attainable. The next result gives a generic approach to debias any pseudo-outcome.

Nonparametric Counterfactual Nonparametric Regression

Theorem

(Yang, Kuchibhotla, TT, 2023) Let $n^*(x;\eta) = E\left(r\left(\eta\right)|x;\eta\right) = m^*\left(x\right)$ *for an observed data based initial pseudo-outcome* $r\left(\eta\right)$ *identifying a counterfactual regression* $m^*\left(x\right)$ *where* η *is a possibly infinite dimensional nuisance parameter* $\eta \in \mathbb{B}$ *(for a normed metric space* $\mathbb{B}$ *with norm* $\|\cdot\|_2$*), under a semiparametric model* $\mathcal{M}$*. Furthermore, suppose that there exists a function* $R\left(\eta, n^*(\eta)\right)$ *in* L_2 *such that for any regular parametric submodel* η_t *of* $\mathcal{M}$ *with score* S*, we have that*

$$\frac{\partial E\left(r\left(\eta_t\right)|x\right)}{\partial_t} = E\left[R\left(\eta, n^*(\eta)\right) S|X = x\right]$$

then $\left\|E\left\{R\left(\eta', n^*(\eta')\right) + r\left(\eta'\right) - n^*(X;\eta)|X\right\}\right\|_2 = O\left(\left\|\eta' - \eta\right\|_2^2\right)$*. Furthermore,* $R\left(\eta, n^*(\eta)\right) + r\left(\eta\right) - \psi$ *is an influence function of the marginal functional* $\psi = E\left\{n^*(x;\eta)\right\}$ *on* $\mathcal{M}$*.*

Nonparametric Counterfactual Nonparametric Regression

- Theorem 2 formally establishes that a pseudo-outcome for a given counterfactual regression $m^{*}(x)$ can be obtained by effectively deriving an influence function of the marginal functional $\psi = E\{n^{*}(x;\eta)\}$ under a given semiparametric model $\mathcal{M}$.

Nonparametric Counterfactual Nonparametric Regression

- Theorem 2 formally establishes that a pseudo-outcome for a given counterfactual regression $m^*(x)$ can be obtained by effectively deriving an influence function of the marginal functional $\psi = E\{n^*(x;\eta)\}$ under a given semiparametric model $\mathcal{M}$.
- The resulting influence function is given by $R(\eta, n^*(\eta)) + r(\eta) - \psi$ and the oracle pseudo-outcome may appropriately be defined as $R(\eta, n^*(\eta)) + r(\eta)$.

Nonparametric Counterfactual Nonparametric Regression

- Theorem 2 is quite general as it applies to the most comprehensive class of non-parametric counterfactual regressions studied to date. The result thus provides a unified solution to the problem of counterfactual regression, recovering several existing methods, and more importantly, providing a number of new results.

Nonparametric Counterfactual Nonparametric Regression

- Theorem 2 is quite general as it applies to the most comprehensive class of non-parametric counterfactual regressions studied to date. The result thus provides a unified solution to the problem of counterfactual regression, recovering several existing methods, and more importantly, providing a number of new results.
- Namely, the theorem provides a formal framework for deriving a pseudo-outcome which by construction is guaranteed to satisfy so-called "Neyman Orthogonality" property, i.e. that the bias incurred by estimating nuisance functions is at most of second order (Chernozhukov et al., 2017).

FW-Learner of the CATE under unconfoundedness

- Applying Theorem 2 to the CATE with initial pseudo outcome $r(\eta) = E(Y|A=1,X;\eta) - E(Y|A=0,X;\eta)$ produces the corrected pseudo outcome

$$\begin{aligned} & R(\eta, n^*(\eta)) + r(\eta) \\ = \quad & (-1)^{1-A} Y / f(A|X;\eta) + E(Y|A=1,X;\eta) - E(Y|A=0,X;\eta) \end{aligned}$$

FW-Learner of the CATE under unconfoundedness

- Applying Theorem 2 to the CATE with initial pseudo outcome $r(\eta) = E(Y|A=1,X;\eta) - E(Y|A=0,X;\eta)$ produces the corrected pseudo outcome

$$\begin{aligned} & R(\eta, n^*(\eta)) + r(\eta) \\ = \quad & (-1)^{1-A} Y / f(A|X;\eta) + E(Y|A=1,X;\eta) - E(Y|A=0,X;\eta) \end{aligned}$$

- The corresponding empirical pseudo-outcome

$$\widehat{H} = (-1)^{1-A} Y / \widehat{f}(A|X) + \widehat{E}(Y|A=1,X) - \widehat{E}(Y|A=0,X)$$

can then be used as an outcome to construct the FW-learner of the CATE.

FW-Learner of the CATE under unconfoundedness

- Note that the above results continue to apply if as often the case we aim to evaluate CATE with respect to a subset of covariate V of X.

$$m^{*}(v) = E\left(Y^{a=1} - Y^{a=0} | V = v\right)$$

FW-Learner of the CATE under unconfoundedness

- Note that the above results continue to apply if as often the case we aim to evaluate CATE with respect to a subset of covariate V of X.

$$m^{*}(v) = E\left(Y^{a=1} - Y^{a=0} | V = v\right)$$

- As confounding adjustment requires accounting for all of X, the empirical pseudo-outcome remains

$$\widehat{H} = (-1)^{1-A} Y / \widehat{f}(A|X) + \widehat{E}(Y|A=1,X) - \widehat{E}(Y|A=0,X)$$

however the FW-learner regresses $\widehat{H}$ on basis functions $\overline{\phi}_K(v)$.

Simulation study

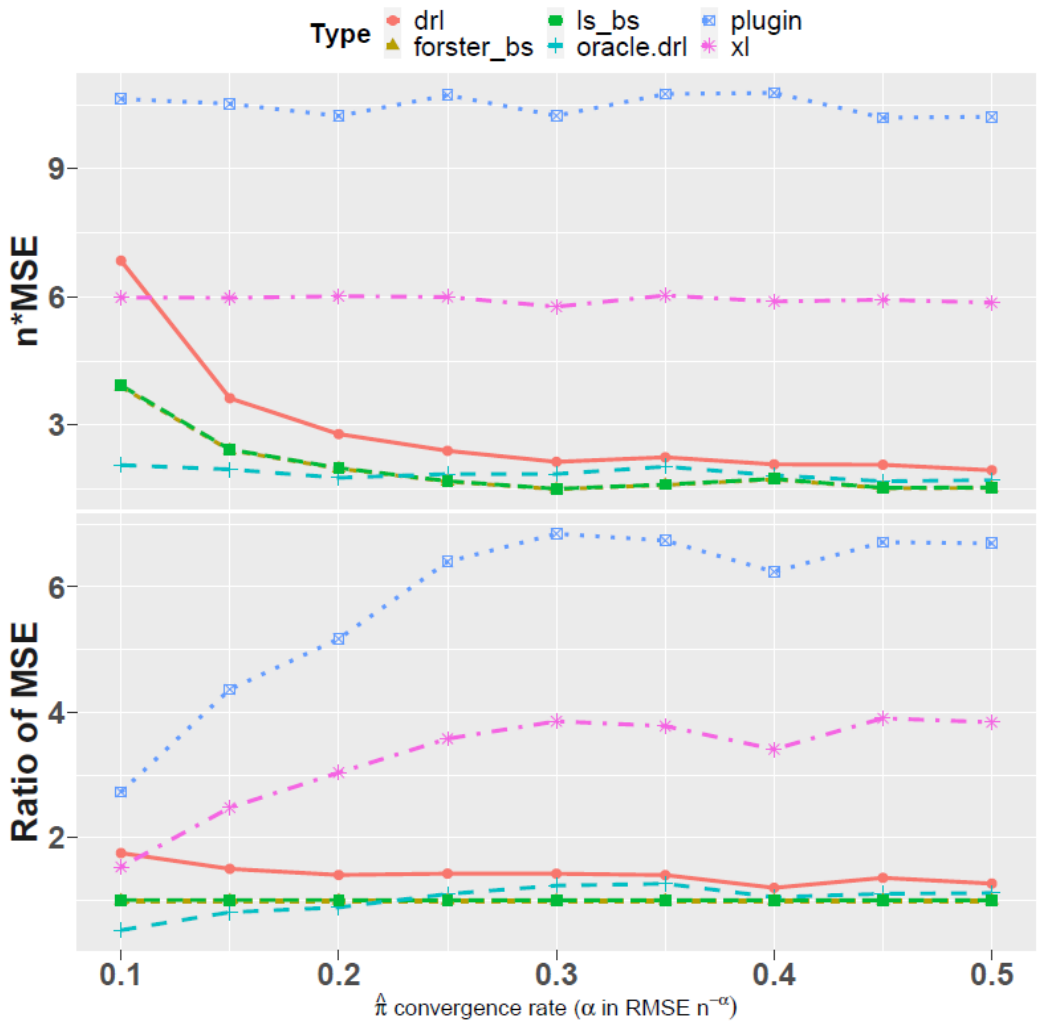


Figure 2: A comparison between different estimators, sample size $n = 2000$—-Top figure shows $n \times$ MSE of each estimator; The bottom plot shows the ratio of MSE of different estimators compared to the proposed Forster–Warmuth estimator with basic splines (baseline). The MSE is averaged over 500 simulations.

Simulation Study

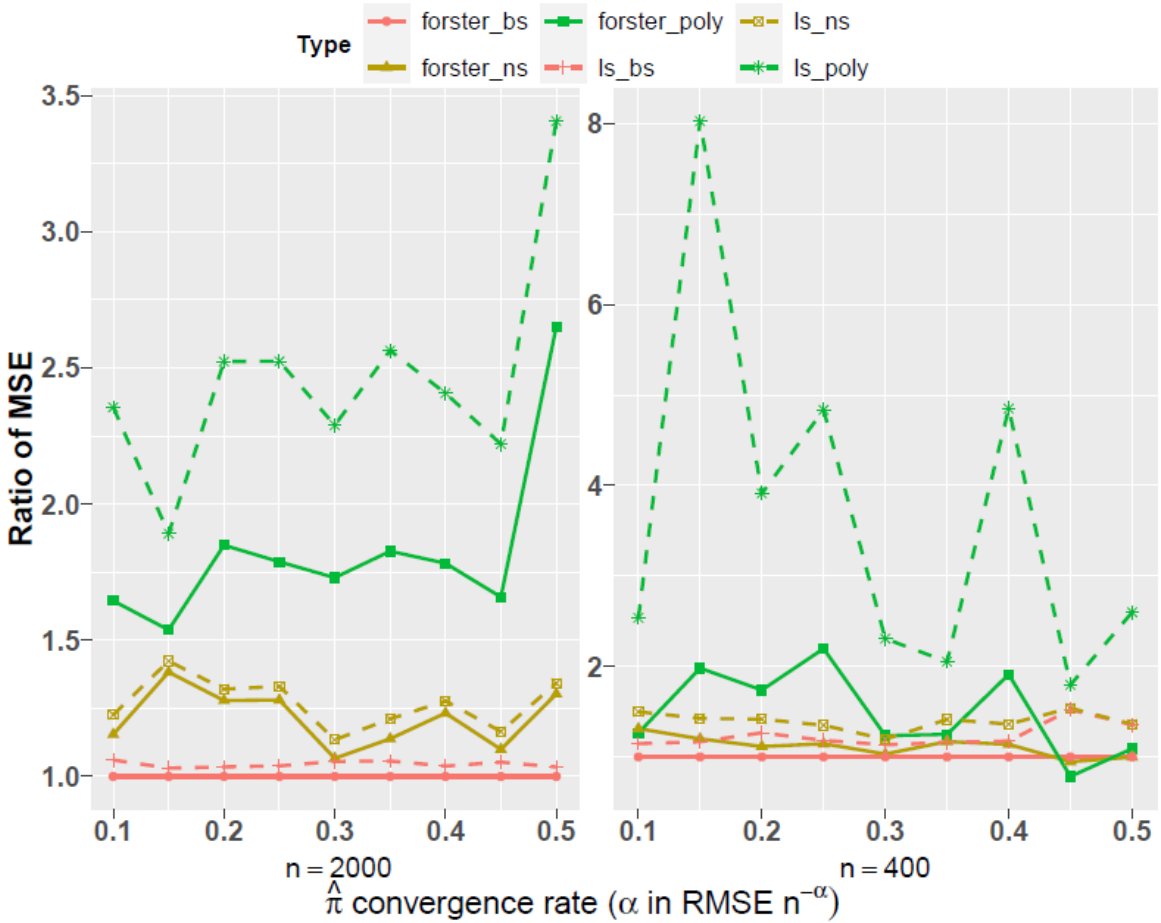


Figure 3: A comparison between FW and LS estimators with different basis for X with heavy-tailed distribution, baseline method is the FW-Learner with basic splines (FW_bs); Left: sample size $n = 2000$; Right: $n = 400$. The MSE is averaged over 500 simulations.

RHC Application

- We consider the analysis of SUPPORT study with the aim of evaluating the causal effect of right heart catheterization (RHC) during the initial care of critically ill patients in the intensive care unit (ICU) on survival time up to 30 days.

RHC Application

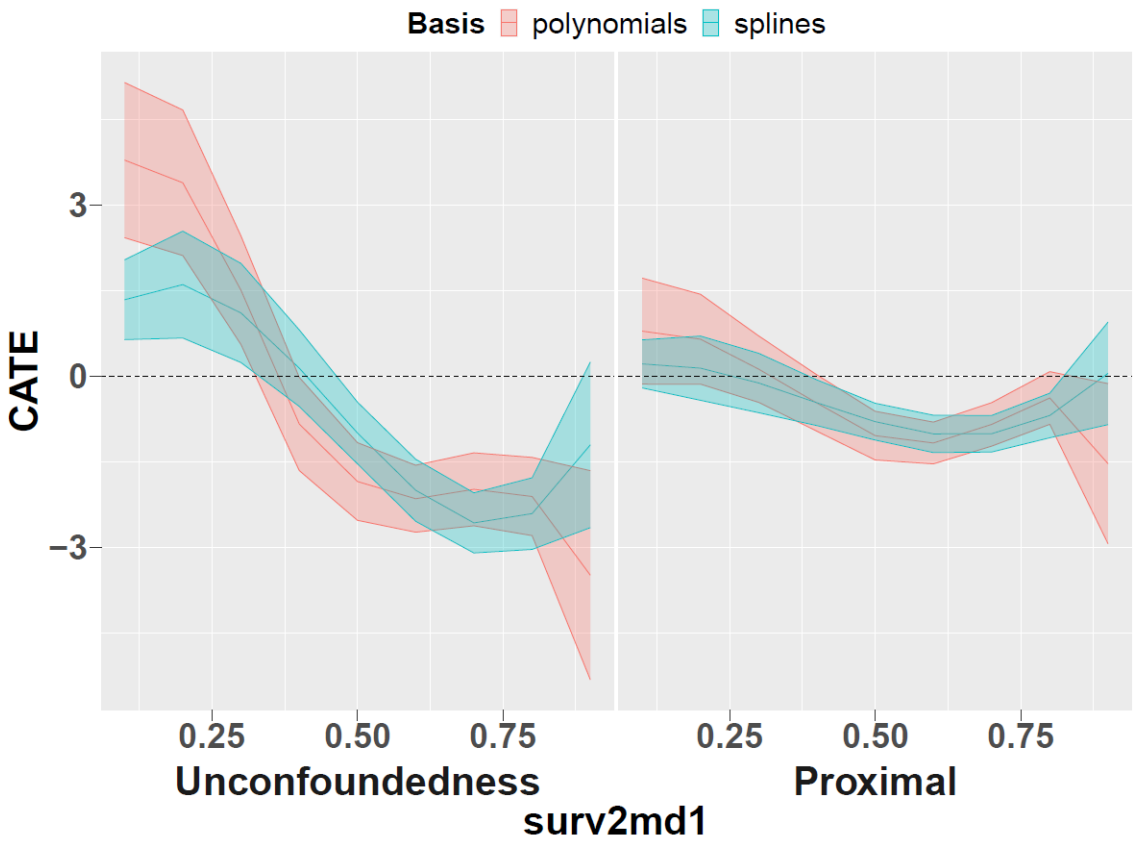
- We consider the analysis of SUPPORT study with the aim of evaluating the causal effect of right heart catheterization (RHC) during the initial care of critically ill patients in the intensive care unit (ICU) on survival time up to 30 days.
- RHC was performed in 2184 patients within the initial 24 hours of ICU stay, while 3551 patients were managed without RHC

RHC Application

- We consider the analysis of SUPPORT study with the aim of evaluating the causal effect of right heart catheterization (RHC) during the initial care of critically ill patients in the intensive care unit (ICU) on survival time up to 30 days.
- RHC was performed in 2184 patients within the initial 24 hours of ICU stay, while 3551 patients were managed without RHC
- Study collected rich patient information encoded in 73 covariates, including demographics (such as age, sex, race, education, income, and insurance status), baseline estimated probability of survival, comorbidity, vital signs, physiological status, and functional status. The outcome of interest is the number of days between admission and death or censoring at 30 days (Y).

RHC Application

- We aim to estimate the CATE conditional only on V =baseline estimated probability of survival.



Conclusion

- Some other counterfactual regression problems and their pseudo outcomes

Conclusion

- Some other counterfactual regression problems and their pseudo outcomes
  - Conditional Quantile Treatment Effect:

$$
\begin{aligned}
m^*(x) &= F^{-1}_{Y|A=1,X}\left(q|A=1,x\right) - F^{-1}_{Y|A=1,X}\left(q|A=0,x\right) \\
H &= \frac{(-1)^{1-A}\left[1\left(Y \leq F^{-1}_{Y|A,X}\left(q|A,x\right)\right) - q\right]}{f\left(A|X\right) f_{Y|A,X}\left(F^{-1}_{Y|A,X}\left(q|A,x\right)|A,x\right)} \\
&\quad + F^{-1}_{Y_{a=1|x}}\left(q|x\right) - F^{-1}_{Y_{a=0|x}}\left(q|x\right)
\end{aligned}
$$

Conclusion

- Some other counterfactual regression problems and their pseudo outcomes
  - Conditional Quantile Treatment Effect:

$$\begin{aligned} m^*(x) &= F^{-1}_{Y|A=1,X}\left(q|A=1,x\right) - F^{-1}_{Y|A=1,X}\left(q|A=0,x\right) \\ H &= \frac{(-1)^{1-A}\left[1\left(Y \leq F^{-1}_{Y|A,X}\left(q|A,x\right)\right) - q\right]}{f\left(A|X\right) f_{Y|A,X}\left(F^{-1}_{Y|A,X}\left(q|A,x\right)|A,x\right)} \\ &\quad + F^{-1}_{Y_{a=1|x}}\left(q|x\right) - F^{-1}_{Y_{a=0|x}}\left(q|x\right) \end{aligned}$$

  - Conditional Causal Effect for GLMs:

$$\begin{aligned} m^*(x) &= g^{-1}\left\{E\left(Y|A=1,X\right)\right\} - g^{-1}\left\{E\left(Y|A=1,X\right)\right\} \\ H &= \frac{(-1)^{1-A}\left[Y - E\left(Y|A,X\right)\right]}{f\left(A|X\right) g'\left(g^{-1}\left\{E\left(Y|A,X\right)\right\}\right)} \\ &\quad + F^{-1}_{Y_{a=1|x}}\left(q|x\right) - F^{-1}_{Y_{a=0|x}}\left(q|x\right) \end{aligned}$$

Conclusion

- Several more counterfactual regressions treated studied in the paper including:

Conclusion

- Several more counterfactual regressions treated studied in the paper including:
  - The CATE for the treated, the compliers, and for the overall population in the presence of unmeasured confounding identified by the conditional Wald estimand, by carefully leveraging a binary instrumental variable (Wang and Tchetgen Tchetgen, 2018);

Conclusion

- Several more counterfactual regressions treated studied in the paper including:
  - The CATE for the treated, the compliers, and for the overall population in the presence of unmeasured confounding identified by the conditional Wald estimand, by carefully leveraging a binary instrumental variable (Wang and Tchetgen Tchetgen, 2018);
  - the nonparametric counterfactual outcome mean for a continuous treatment both under unconfoundedness and proximal causal identification conditions.

Conclusion

- Several more counterfactual regressions treated studied in the paper including:
  - The CATE for the treated, the compliers, and for the overall population in the presence of unmeasured confounding identified by the conditional Wald estimand, by carefully leveraging a binary instrumental variable (Wang and Tchetgen Tchetgen, 2018);
  - the nonparametric counterfactual outcome mean for a continuous treatment both under unconfoundedness and proximal causal identification conditions.
  - The CATE using proxies.

Conclusion

- Several more counterfactual regressions treated studied in the paper including:
  - The CATE for the treated, the compliers, and for the overall population in the presence of unmeasured confounding identified by the conditional Wald estimand, by carefully leveraging a binary instrumental variable (Wang and Tchetgen Tchetgen, 2018);
  - the nonparametric counterfactual outcome mean for a continuous treatment both under unconfoundedness and proximal causal identification conditions.
  - The CATE using proxies.
  - Nonparametric regression under MAR and MNAR conditions using a shadow variable for the latter.

Acknowledgments

- Co-authors Elsa and Arun

Acknowledgments

- Co-authors Elsa and Arun
- This work is funded by NIH grants R01AG065276,R01CA222147 and R01AI27271. (PI:Tchetgen Tchetgen)

Acknowledgments

- Co-authors Elsa and Arun
- This work is funded by NIH grants R01AG065276,R01CA222147 and R01AI27271. (PI:Tchetgen Tchetgen)

- Yang, Y., Kuchibhotla, A.K. and Tchetgen, E.T., 2023. Forster-Warmuth Counterfactual Regression: A Unified Learning Approach. arXiv preprint arXiv:2307.16798.