

Distributional Treatment Effect with Kernels

Krikamol Muandet

Max Planck Institute for Intelligent Systems
Tübingen, Germany



Junhyung Park
MPI-IS



Uri Shalit
Technion



Bernhard Schölkopf
MPI-IS



Krikamol Muandet
MPI-IS

Hi! PARIS Summer School 2022
July 7, 2022

Kernel Methods

Distributional Treatment Effect

Discussion

Kernel Methods

Distributional Treatment Effect

Discussion

Kernelising Linear Methods

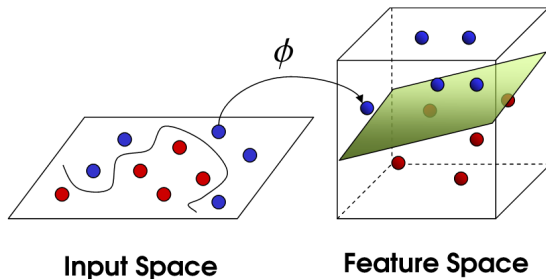
- ▶ Many popular methods in machine learning and statistics are linear, e.g. linear (ridge) regression, SVM, logistic regression, PCA, ...

Kernelising Linear Methods

- ▶ Many popular methods in machine learning and statistics are linear, e.g. linear (ridge) regression, SVM, logistic regression, PCA, ...
- ▶ However, the real world is often not linear.

Kernelising Linear Methods

- ▶ Many popular methods in machine learning and statistics are linear, e.g. linear (ridge) regression, SVM, logistic regression, PCA, ...
- ▶ However, the real world is often not linear.
- ▶ **Key idea:** Embed data into a high- (often infinite-)dimensional space, carry out a linear method there, and project back down to the original space to obtain a non-linear model.



Kernels and RKHS

Some Notations

- ▶ Denote the domain by \mathcal{X} .
- ▶ We call a symmetric, positive-definite function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a *kernel*. This means that
 - ▶ for any $x_1, x_2 \in \mathcal{X}$, $k(x_1, x_2) = k(x_2, x_1)$; and
 - ▶ for any $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$.

Kernels and RKHS

Some Notations

- ▶ Denote the domain by \mathcal{X} .
- ▶ We call a symmetric, positive-definite function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a *kernel*. This means that
 - ▶ for any $x_1, x_2 \in \mathcal{X}$, $k(x_1, x_2) = k(x_2, x_1)$; and
 - ▶ for any $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$.
- ▶ Each kernel k is associated to a unique *reproducing kernel Hilbert space* (RKHS) \mathcal{H} , a space of functions $\mathcal{X} \rightarrow \mathbb{R}$.

Kernels and RKHS

Some Notations

- ▶ Denote the domain by \mathcal{X} .
- ▶ We call a symmetric, positive-definite function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a *kernel*. This means that
 - ▶ for any $x_1, x_2 \in \mathcal{X}$, $k(x_1, x_2) = k(x_2, x_1)$; and
 - ▶ for any $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$.
- ▶ Each kernel k is associated to a unique *reproducing kernel Hilbert space* (RKHS) \mathcal{H} , a space of functions $\mathcal{X} \rightarrow \mathbb{R}$.
- ▶ For each $x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$.
 - ▶ In fact, \mathcal{H} is the closure of the linear span of $\{k(x, \cdot) : x \in \mathcal{X}\}$.
- ▶ For any $f \in \mathcal{H}$, $x \in \mathcal{X}$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$

Kernels and RKHS

Some Notations

- ▶ Denote the domain by \mathcal{X} .
- ▶ We call a symmetric, positive-definite function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a *kernel*. This means that
 - ▶ for any $x_1, x_2 \in \mathcal{X}$, $k(x_1, x_2) = k(x_2, x_1)$; and
 - ▶ for any $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$.
- ▶ Each kernel k is associated to a unique *reproducing kernel Hilbert space* (RKHS) \mathcal{H} , a space of functions $\mathcal{X} \rightarrow \mathbb{R}$.
- ▶ For each $x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$.
 - ▶ In fact, \mathcal{H} is the closure of the linear span of $\{k(x, \cdot) : x \in \mathcal{X}\}$.
- ▶ For any $f \in \mathcal{H}$, $x \in \mathcal{X}$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$

Examples of Kernels for $\mathcal{X} = \mathbb{R}^d$

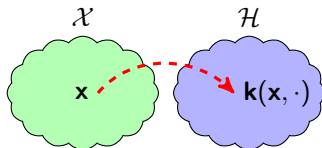
- ▶ **Linear kernel:** $k(x_1, x_2) = x_1 \cdot x_2$, $\mathcal{H} = \mathbb{R}^d$.
- ▶ **Polynomial kernel:** $k(x_1, x_2) = (x_1 \cdot x_2 + c)^m$, $\dim(\mathcal{H}) = \binom{d+m}{m}$.
- ▶ **Gaussian kernel:** $k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|_2^2}$, $\dim(\mathcal{H}) = \infty$.

Embedding Data-points

Recall,

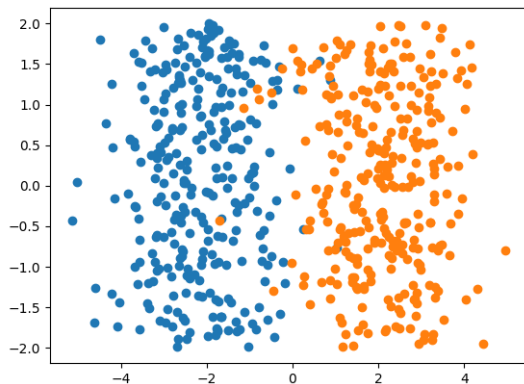
- ▶ **Key idea:** Embed data into a high- (often infinite-)dimensional space, carry out a linear method there, and project back down to the original space to obtain a non-linear model.

We can embed any $x \in \mathcal{X}$ into \mathcal{H} by $x \mapsto k(x, \cdot)$!



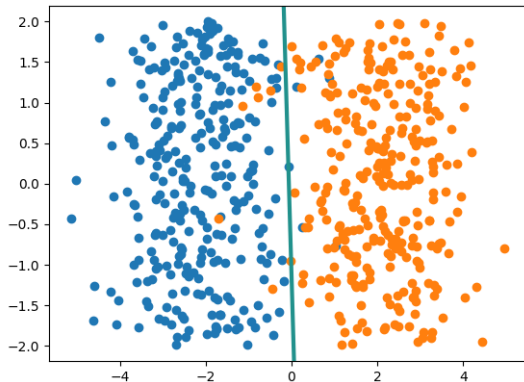
By the reproducing property, $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}} = k(x, x')$.

Example: Classification



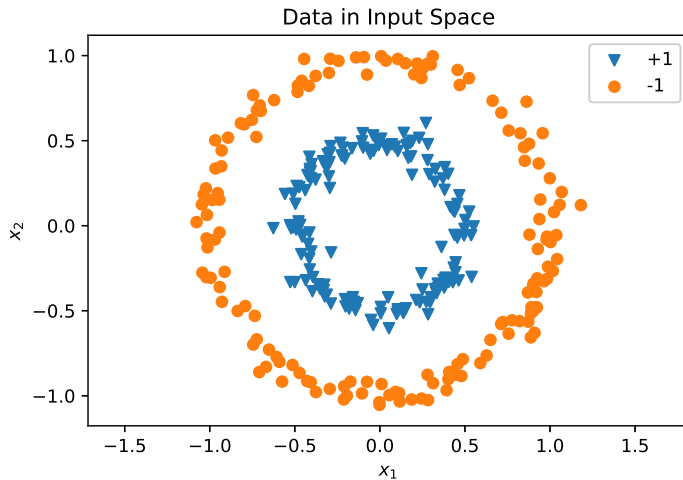
Data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, $x_i \in \mathbb{R}^2$, $y_i \in \{+1, -1\}$

Example: Classification



$$\text{Model: } f(x) = w^T x + b$$

Example: Classification

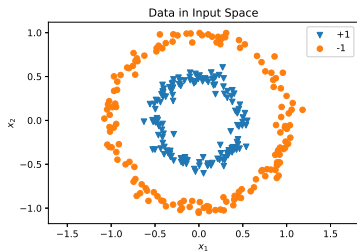


Kernel Feature Map

$$\phi : (x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

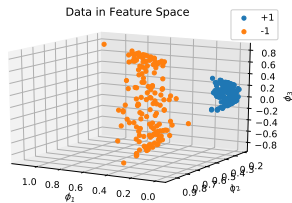
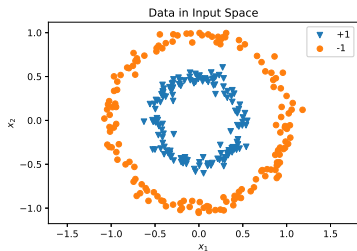
Kernel Feature Map

$$\phi : (x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$



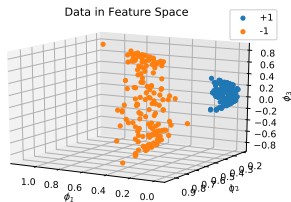
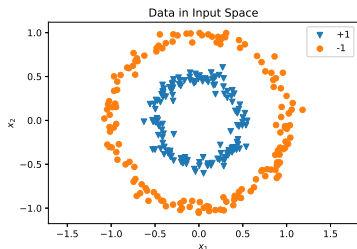
Kernel Feature Map

$$\phi : (x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$



Kernel Feature Map

$$\phi : (x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$



But observe that

$$\langle \phi(x), \phi(z) \rangle_{\mathbb{R}^3} = x_1^2 z_1^2 + x_2^2 z_2^2 + 2(x_1 x_2)(z_1 z_2) = (x_1 z_1 + x_2 z_2)^2 = (x \cdot z)^2$$

Model in primal and dual form:

$$f(x) = w^\top \phi(x) \quad \Rightarrow \quad f(x) = \sum_{i=1}^n \alpha_i \underbrace{\langle \phi(x_i), \phi(x) \rangle_{\mathbb{R}^3}}_{=(x_i \cdot x)^2 = k(x_i, x)}$$

Embedding Distributions

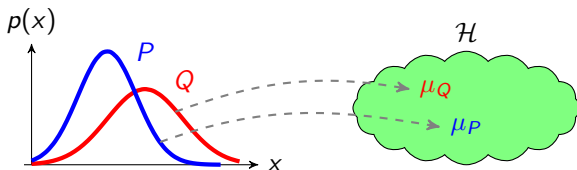
- ▶ Recall, point embeddings: $x \mapsto k(x, \cdot) : \mathcal{X} \rightarrow \mathcal{H}$.
- ▶ A random variable X taking values in \mathcal{X} , with *distribution* P .
- ▶ We can embed P into \mathcal{H} via “kernel mean embedding”:

$$P \mapsto \mu_P = \mathbb{E}[k(X, \cdot)] = \int_{\mathcal{X}} k(x, \cdot) dP(x).$$

- ▶ Given samples X_1, \dots, X_n from X , an empirical estimate of μ_P can easily be obtained:

$$\hat{\mu}_P = \frac{1}{n} \sum_{i=1}^n k(X_i, \cdot).$$

- ▶ $\mathbb{E}_{X \sim P}[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$.

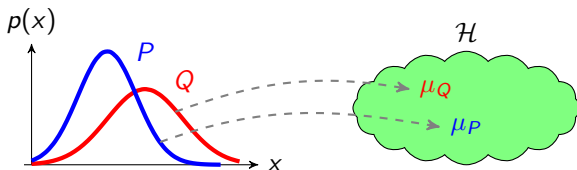


Characteristic Kernels

- ▶ **Def:** A kernel k is *characteristic* if the map $P \mapsto \mu_P$ is injective.
- ▶ A kernel k being characteristic means that the corresponding RKHS \mathcal{H} is “rich enough” to capture all information about a distribution.

Examples

- ▶ If $k(x_1, x_2) = x_1 \cdot x_2$, then $\mu_P = \mathbb{E}[k(X, \cdot)] = \mathbb{E}[X]$.
So μ_P only captures the *expectation* of P , and so k is not characteristic.
- ▶ If $k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}$, k can be shown to be characteristic.



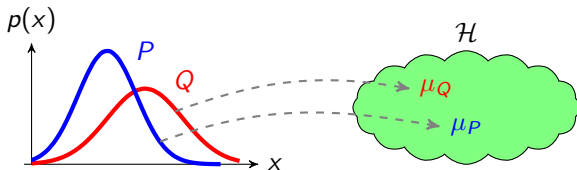
Maximum Mean Discrepancy

- ▶ The *maximum mean discrepancy* (MMD) is defined as

$$\text{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}.$$

- ▶ With a characteristic kernel, $\text{MMD}(P, Q) = 0 \iff P = Q$!
- ▶ With samples X_1, \dots, X_n from P and Y_1, \dots, Y_m from Q , we can use the MMD to carry out a two-sample test:

$$\widehat{\text{MMD}}(P, Q) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(X_i, X_j) + \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j)$$



Witness Functions

- ▶ Recall, $\text{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}$.
- ▶ The normand $\mu_P - \mu_Q$ belongs to \mathcal{H} , i.e. it is a *function* $\mathcal{X} \rightarrow \mathbb{R}$.

$$f(t) \propto \mu_P(t) - \mu_Q(t)$$

- ▶ This is called the (unnormalised) *witness function*, and by evaluating it at a particular point, we can identify regions in which the density of one distribution dominates the other.

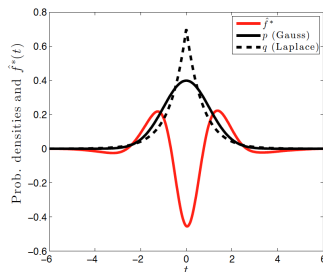


Figure: A Kernel Two-Sample Test, Gretton, Borgwardt, Rasch, Schölkopf and Smola, JMLR 2012.

Embedding Conditional Distributions

- ▶ Recall,
 - ▶ point embeddings $x \mapsto k(x, \cdot) : \mathcal{X} \rightarrow \mathcal{H}$ and
 - ▶ distribution embeddings $X \mapsto \mathbb{E}[k(X, \cdot)]$.
- ▶ What about conditional distributions?
- ▶ Let X and Z be random variables taking values in domains \mathcal{X} and \mathcal{Z} .
- ▶ Let $P_{X|Z}$ be the conditional distribution of X given Z .
- ▶ We define the *kernel conditional mean embedding* of X given Z as

$$\mu_{P_{X|Z}} = \mathbb{E}[k(X, \cdot) \mid Z].$$

- ▶ Unlike point and (unconditional) distribution embeddings, this is not a single element in \mathcal{H} , but a RV depending on the value of Z .

Further Information & References

- ▶ Schölkopf, B. and Smola, A., *Learning with Kernels*. MIT Press, 2002.
- ▶ Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B., *Kernel Mean Embedding of Distributions: A Review and Beyond*. Foundations and Trends in Machine Learning, 2017.
- ▶ **Today Talk:** Park, J., Shalit, U., Schölkopf B., and Muandet K. *Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression*. ICML 2021.
- ▶ **Recommended:** Kallus, N. and Oprescu, M. *Robust and Agnostic Learning of Conditional Distributional Treatment Effects*. ArXiv:2205.11486, 2022.

Kernel Methods

Distributional Treatment Effect

Discussion

Problem Set-Up: Potential Outcomes Framework

Notations

Probability Space (Ω, \mathcal{F}, P)

Treatment Assignment $Z : \Omega \rightarrow \{0, 1\}$

Covariate Variable $X : \Omega \rightarrow \mathcal{X}$

Potential Outcome under Control $Y_0 : \Omega \rightarrow \mathcal{Y}$

Potential Outcome under Treatment $Y_1 : \Omega \rightarrow \mathcal{Y}$

Observed Outcome $Y = (1 - Z)Y_0 + ZY_1$

Examples

- ▶ Drug administration
- ▶ Patient Characteristics
- ▶ Measurement without drug
- ▶ Measurement with drug
- ▶ Observed measurement

Problem Set-Up: Potential Outcomes Framework

Notations

Probability Space (Ω, \mathcal{F}, P)

Treatment Assignment $Z : \Omega \rightarrow \{0, 1\}$

Covariate Variable $X : \Omega \rightarrow \mathcal{X}$

Potential Outcome under Control $Y_0 : \Omega \rightarrow \mathcal{Y}$

Potential Outcome under Treatment $Y_1 : \Omega \rightarrow \mathcal{Y}$

Observed Outcome $Y = (1 - Z)Y_0 + ZY_1$

Examples

- ▶ Drug administration
- ▶ Patient Characteristics
- ▶ Measurement without drug
- ▶ Measurement with drug
- ▶ Observed measurement

- ▶ We assume either the *randomised control trial* setting, or *strong ignorability*:
 - ▶ **unconfoundedness** $Z \perp (Y_0, Y_1) \mid X$; and
 - ▶ **overlap** $0 < e(X) = P(Z = 1 \mid X) = \mathbb{E}[Z \mid X] < 1$.

Problem Set-Up: Potential Outcomes Framework

Notations

Probability Space (Ω, \mathcal{F}, P)

Treatment Assignment $Z : \Omega \rightarrow \{0, 1\}$

Covariate Variable $X : \Omega \rightarrow \mathcal{X}$

Potential Outcome under Control $Y_0 : \Omega \rightarrow \mathcal{Y}$

Potential Outcome under Treatment $Y_1 : \Omega \rightarrow \mathcal{Y}$

Observed Outcome $Y = (1 - Z)Y_0 + ZY_1$

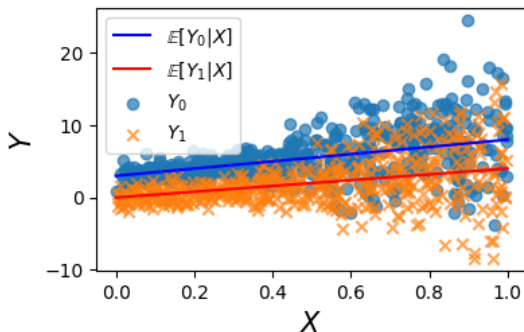
Examples

- ▶ Drug administration
- ▶ Patient Characteristics
- ▶ Measurement without drug
- ▶ Measurement with drug
- ▶ Observed measurement

- ▶ We assume either the *randomised control trial* setting, or *strong ignorability*:
 - ▶ **unconfoundedness** $Z \perp (Y_0, Y_1) \mid X$; and
 - ▶ **overlap** $0 < e(X) = P(Z = 1 \mid X) = \mathbb{E}[Z \mid X] < 1$.
- ▶ Quantities commonly used to measure treatment effect:
 - ▶ **Average Treatment Effect (ATE)** $\mathbb{E}[Y_1 - Y_0]$
 - ▶ **Conditional Average Treatment Effect (CATE)** $\mathbb{E}[Y_1 - Y_0 \mid X]$

Treatment Effect Quantification

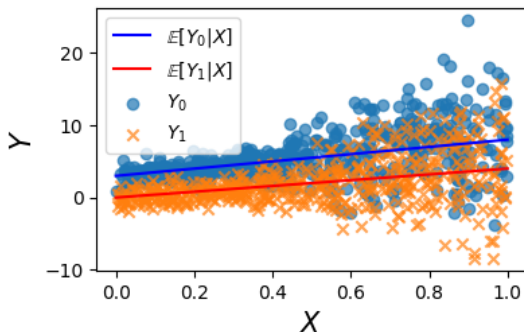
But...



- Estimating ATE and CATE is inherently a problem of *comparing two means*, and as such, is only meaningful if the corresponding variances are given.

Treatment Effect Quantification

But...



- ▶ Estimating ATE and CATE is inherently a problem of *comparing two means*, and as such, is only meaningful if the corresponding variances are given.
- ▶ Other distributional aspects could also be of interest, e.g. skewness.

Progresa: Conditional Cash Transfer programme in Mexico

- Investigation of how the programme affects the average food consumption.

Table 4. Treatment effects for all people living in treatment villages.

	Estimate	Lower Bound	Upper Bound
ME on mean	25.900	14.730	31.130

Notes: Shown are point estimates for marginal effects of the treatment at means (ME) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 14,740$.

Figure: Treatment effects beyond the mean using distributional regression:
Methods and guidance, Hohberg, Pütz and Kneib, 2020

Progresa: Conditional Cash Transfer programme in Mexico

- Investigation of how the programme affects the average food consumption.

Table 4. Treatment effects for all people living in treatment villages.

	Estimate	Lower Bound	Upper Bound
ME on mean	25.900	14.730	31.130
ME on variance	4828.316	820.841	7750.220

Notes: Shown are point estimates for marginal effects of the treatment at means (ME) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 14,740$.

Figure: Treatment effects beyond the mean using distributional regression:
Methods and guidance, Hohberg, Pütz and Kneib, 2020

Progresa: Conditional Cash Transfer programme in Mexico

- Investigation of how the programme affects the average food consumption.

Table 4. Treatment effects for all people living in treatment villages.

	Estimate	Lower Bound	Upper Bound
ME on mean	25.900	14.730	31.130
ME on variance	4828.316	820.841	7750.220
ME on Gini coefficient	0.007	-0.006	0.021

Notes: Shown are point estimates for marginal effects of the treatment at means (ME) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 14,740$.

Figure: Treatment effects beyond the mean using distributional regression:
Methods and guidance, Hohberg, Pütz and Kneib, 2020

Progresa: Conditional Cash Transfer programme in Mexico

- Investigation of how the programme affects the average food consumption.

Table 4. Treatment effects for all people living in treatment villages.

	Estimate	Lower Bound	Upper Bound
ME on mean	25.900	14.730	31.130
ME on variance	4828.316	820.841	7750.220
ME on Gini coefficient	0.007	-0.006	0.021
ME on Atkinson index ($e = 1$)	0.006	-0.005	0.018
ME on Atkinson index ($e = 2$)	0.012	-0.004	0.034
ME on Theil index	0.007	-0.012	0.026
ME on vulnerability	-0.056	-0.092	-0.040

Notes: Shown are point estimates for marginal effects of the treatment at means (ME) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 14,740$.

Figure: Treatment effects beyond the mean using distributional regression:
Methods and guidance, Hohberg, Pütz and Kneib, 2020

Conditional Distributional Treatment Effect

- **Definition:** Let D be some distance function between probability measures. We define the *conditional distributional treatment effect* (CoDiTE) associated with D as

$$U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x}).$$

Conditional Distributional Treatment Effect

- ▶ **Definition:** Let D be some distance function between probability measures. We define the *conditional distributional treatment effect* (CoDiTE) associated with D as

$$U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x}).$$

- ▶ Depending on the distance D chosen, we can extract and compare different aspects of the control and treatment distributions.

Conditional Distributional Treatment Effect

- ▶ **Definition:** Let D be some distance function between probability measures. We define the *conditional distributional treatment effect* (CoDiTE) associated with D as

$$U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x}).$$

- ▶ Depending on the distance D chosen, we can extract and compare different aspects of the control and treatment distributions.
- ▶ **Examples:**
 - ▶ We recover the CATE by setting $D(P_{Y_0|X}, P_{Y_1|X}) = \mathbb{E}[Y_0 - Y_1 | X]$.
 - ▶ Other works have considered D capturing quantiles, cumulative distribution functions or specific distributional parameters, such as mean, variance, skewness, etc.
 - ▶ In our work, we characterise distributions via kernel mean embeddings.

Conditional Distributional Treatment Effect

- ▶ **Definition:** Let D be some distance function between probability measures. We define the *conditional distributional treatment effect* (CoDiTE) associated with D as

$$U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x}).$$

- ▶ Depending on the distance D chosen, we can extract and compare different aspects of the control and treatment distributions.
- ▶ **Examples:**
 - ▶ We recover the CATE by setting $D(P_{Y_0|X}, P_{Y_1|X}) = \mathbb{E}[Y_0 - Y_1 | X]$.
 - ▶ Other works have considered D capturing quantiles, cumulative distribution functions or specific distributional parameters, such as mean, variance, skewness, etc.
 - ▶ In our work, we characterise distributions via kernel mean embeddings.
- ▶ The CoDiTE has a causal interpretation under the same assumptions, i.e. RCT or strong ignorability.

Testing Equality of Control and Treatment Groups

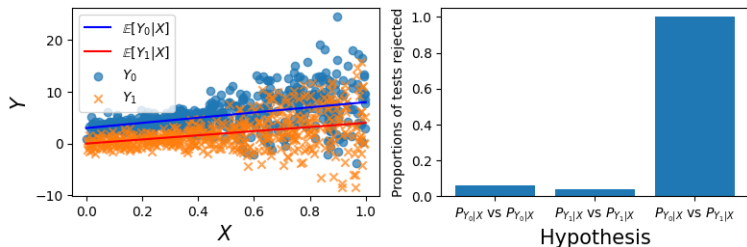
CoDiTE with MMD

- ▶ Recall, CoDiTE: $U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x})$.
- ▶ Let D be the MMD, so that $U_D(x)$ represents the MMD between the control and treatment distributions:

$$U_{\text{MMD}}(x) = \|\mu_{Y_1|X=x} - \mu_{Y_0|X=x}\|_{\mathcal{H}}.$$

- ▶ By integrating over X , we obtain a statistic to test for the equality between the distributions of control and treatment groups:

$$t = \mathbb{E} [\|\mu_{Y_1|X} - \mu_{Y_0|X}\|_{\mathcal{H}}^2].$$



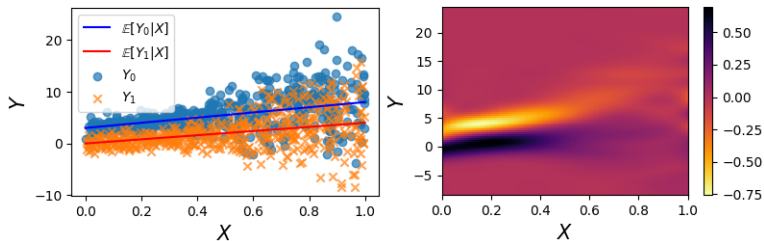
Exploratory Analysis of Conditional Densities

CoDiTE with MMD

- Recall, CoDiTE with MMD: $U_{\text{MMD}}(x) = \|\mu_{Y_1|X=x} - \mu_{Y_0|X=x}\|_{\mathcal{H}}$.
- Instead of taking the norm in \mathcal{H} , we can evaluate the normand (recall, it is the *witness function*) at each value of y .

$$f_x(y) \propto \mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y)$$

- We can then visually compare the conditional densities of the control and treatment groups.

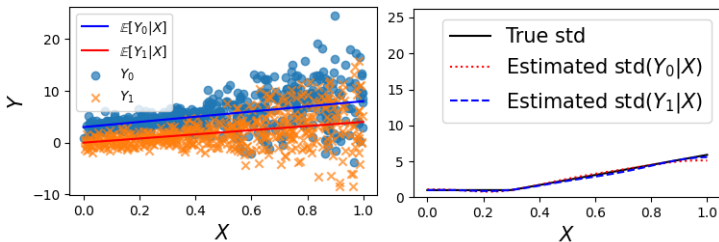


U-Statistic Regression

- ▶ Usual regression estimates the conditional mean, $\mathbb{E}[Y|X]$.
- ▶ But other quantities of interest, most prominently the conditional variance, can be expressed as a generalisation, as *conditional U-statistics*:

$$\mathbb{E}[h(Y_1, \dots, Y_r) | X_1, \dots, X_r].$$

- ▶ For example, $h(Y_1, Y_2) = \frac{1}{2}(Y_1 - Y_2)^2$ gives the conditional variance.
- ▶ We generalise kernel ridge regression accordingly, to estimate the *conditional variance*.



Semi-Synthetic IHDP Data

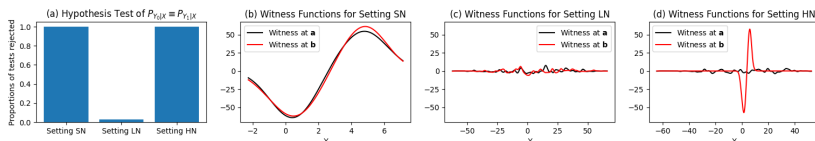
Bayesian Nonparametric Modeling for Causal Inference, Hill, 2011

- ▶ Real covariates, simulated outcome for control and treatment.

Semi-Synthetic IHDP Data

Bayesian Nonparametric Modeling for Causal Inference, Hill, 2011

- ▶ Real covariates, simulated outcome for control and treatment.
- ▶ Existing simulations in the context of CATE estimation: constant Gaussian noise across the entire covariate space.
- ▶ Using the same mean response surfaces, we simulate the data under three settings:
 - ▶ **Small Noise (SN)** Noise is small across the covariate space, so that CATE translates to a meaningful treatment effect.
 - ▶ **Large Noise (LN)** Noise is large such that the mean difference is negligible in comparison.
 - ▶ **Heterogeneous Noise (HN)** Noise is heterogeneous across the covariate space, resulting in meaningful treatment effect for some parts of the population and not others.



Kernel Methods

Distributional Treatment Effect

Discussion

Remarks & Limitations

- ▶ A policy intervention or a counterfactual change can have *non-trivial* effects on a population.
- ▶ Compared to the mean effects, the distributional effects can be harder to interpret (conditional witness function and U-statistic regression).
- ▶ Distributional considerations of treatment effect is pertinent also in the gold standard *randomised control* setting, as well as the more common *observational studies*.
- ▶ On the other hand, we made no effort to account for *selection bias*.
- ▶ Facilitate an algorithmic decision making, e.g., counterfactual fairness.
- ▶ Our methods are all based on variants of kernel ridge regression, and as such, are sensitive to the choice of hyperparameters.
- ▶ Also, when the covariate space becomes high-dimensional, performance deteriorates rapidly.