Reproducing Kernel Hilbert Spaces and Kernel Methods

Marcelo Ortiz-Villavicencio

EMORY UNIVERSITY

March 8, 2024

Econometrics Reading Group

Roadmap

Introduction

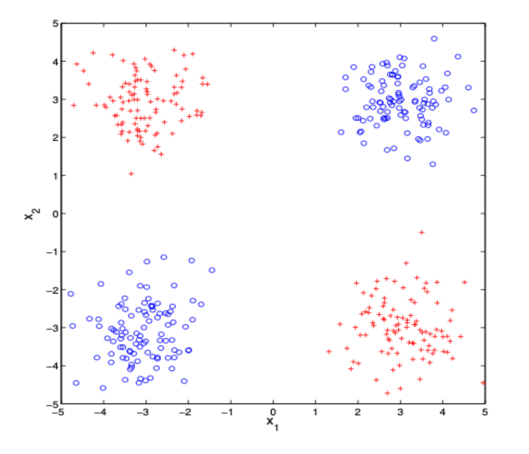
Introduction

EMORY UNIVERSITY

- Many problems in statistics like *nonparametric regression*, *density estimation*, *dimension reduction* and *testing* involve optimizing over function spaces.
- Why *Hilbert Spaces*? These include a broad function class and enjoy geometric properties similar to ordinary Euclidean space.
- We are going to focus on a particular class of function-based Hilbert Space which are defined by *reproducing kernels* (i.e., kernels with reproducing property).
- These spaces, known as *reproducing kernel Hilbert spaces* (**RKHS**), have attractive properties from both the computational and statistical points of view.
- **RKHS** provides a *mathematical framework* for understanding and leveraging the properties of *kernel methods*, allowing for flexible nonlinear modeling.
- My goal towards the end of this chapter is to present an application of these concepts in Causal Inference.

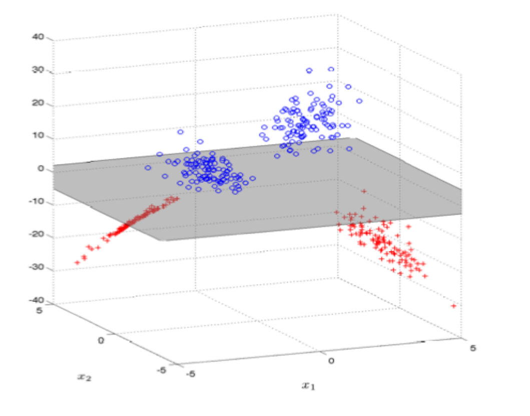
Motivating Example: Linear Classifiers

- Suppose that we want to separate (classify) the red points from the blue using a linear classifier.
- We have access to variables in two dimensions, $x \in \mathbb{R}^2$



Motivating Example: Linear Classifiers

- It's not possible to separate the points in the original space
- However if we map points to a *higher dimensional feature space* like $\phi(x) = \left[\begin{array}{ccc} x_1 & x_2 & x_1x_2 \end{array}\right] \in \mathbb{R}^3$ it is possible to use a linear classifier.



Motivating Example: Linear Classifiers

- Of course there is nothing new in doing a classifier via transformation of features, right?
- What distinguished kernel methods is that they can use *infinetely many features*
- We can use it as long as our algorithms are defined in terms of *dot products* between features, where these dot products can be computed in *closed form*.
- The term *kernel* simply refers to a *dot product* between possible infinitely many features.
- Alternatively, kernel methods can be used to control *smoothness* of a function used in regression or classification and avoid overfitting/underfitting.

Hilbert Space

Basics of Hilbert Spaces

- Hilbert Spaces are particular types of *vector spaces*, meaning that operations of *addition* and *scalar multiplication* are defined.

Definition (Inner Product)

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an inner product on $\mathcal{H}$ if

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$ for all $f_1, f_2, g \in \mathcal{H}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$ for all $f \in \mathcal{H}$.

- A vector space with an inner product is known as an *inner product space*
- We can define a *norm* using the inner product as $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.

EMORY UNIVERSITY

Definition (Cauchy Sequence)

A sequence $\{f_n\}_{n=1}^{\infty}$ of elements in a normed space $\mathcal{H}$ is said to be a Cauchy sequence if for every $\epsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$, such that for all $n, m \geq N$, $\|f_n - f_m\|_{\mathcal{H}} < \epsilon$

Definition (Hilbert Space)

A Hilbert space $\mathcal{H}$ is a space on which an inner product is defined and every Cauchy sequence $\{f_n\}_{n=1}^{\infty}$ in $\mathcal{H}$ converges to some element $f^* \in \mathcal{H}$.

- A metric space in which every *Cauchy sequence* converges to an element f^* of the space is known as *complete*
- In summary, a Hilbert space is a *complete inner product space*.

Riesz Representation Theorem

- The notion of a *linear functional* plays an important role in characterizing **RKHS**.
- A linear functional on a Hilbert space $\mathcal{H}$ is a mapping $L : \mathcal{H} \to \mathbb{R}$ that is linear, meaning that $L(f + \alpha g) = L(f) + \alpha L(g)$ for all $f, g \in \mathcal{H}$ and $\alpha \in \mathbb{R}$.
- A linear functional is said to be *bounded* if there exists some $M < \infty$ such that $|L(f)| \leq M\|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$.
- Given any $g \in \mathcal{H}$, the mapping $f \mapsto \langle f, g\rangle_{\mathcal{H}}$ defines a linear functional.
- It is bounded, since by the Cauchy-Schwarz inequality we have $|\langle f, g\rangle_{\mathcal{H}}| \leq M\|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$, where $M := \|g\|_{\mathcal{H}}$.
- The *Riesz representation theorem* guarantees that every bounded linear functional arises in exactly this way.

Theorem (Riesz Representation Theorem)

Let L be a bounded linear functional on a Hilbert Space. Then there exists a unique $g \in \mathcal{H}$ such that $L(f) = \langle f, g\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. We refer to g as the representer of the functional L.

Kernels and Operations

Definition (Kernel)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} .$$

- We generally don't require any conditions on $\mathcal{X}$
- A single kernel can correspond to several possible features. A trivial example for $\mathcal{X} := \mathbb{R}$:

$$\phi_1(x) = x \quad \text{and} \quad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

$$\langle \phi_1(x), \phi_1(x) \rangle = \langle \phi_2(x), \phi_2(x) \rangle = x^2$$

Theorem (Sum of kernels are kernels)

Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.

Theorem (Product of kernels are kernels)

Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$, then $k := k_1 \times k_2$ is a kernel on $\mathcal{X}$.

Sketch of the proof: Let us define two spaces $\mathcal{H}_1, \mathcal{H}_2$. $\mathcal{H}_1$ is the space of kernels between shapes with the following map,

$$\phi_1(x) = \begin{bmatrix} \mathbb{I}_{\square} \\ \mathbb{I}_{\triangle} \end{bmatrix} \quad \phi_1(\square) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad k_1(\square, \triangle) = \langle \phi_1(\square), \phi_1(\triangle) \rangle = 0.$$

$\mathcal{H}_2$ is the space of kernels between colors with the following map,

$$\phi_2(x) = \begin{bmatrix} \mathbb{I}_{\color{red}\bullet} \\ \mathbb{I}_{\color{blue}\bullet} \end{bmatrix} \quad \phi_2({\color{blue}\bullet}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad k_2({\color{blue}\bullet}, {\color{blue}\bullet}) = \langle \phi_2({\color{blue}\bullet}), \phi_2({\color{blue}\bullet}) \rangle = 1$$

Sum and Products

Sketch of the proof: Let's define a feature space for colors and shapes

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\color{red}\square} & \mathbb{I}_{\color{red}\triangle} \\ \mathbb{I}_{\color{blue}\square} & \mathbb{I}_{\color{blue}\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\color{red}\bullet} \\ \mathbb{I}_{\color{blue}\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Since inner product between 2 matrices $A \in \mathbb{R}^{mxn}$ and $B \in \mathbb{R}^{mxn}$ is $\langle A, B \rangle = tr(A^\top B)$, then the Kernel is:

$$k(x, x') = \sum_{i \in \{{\color{red}\bullet}, {\color{blue}\bullet}\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x') = \mathrm{tr}(\phi_1(x) \underbrace{\phi_2^\top(x)\phi_2(x')}_{k_2(x,x')} \phi_1^\top(x'))$$

$$= \mathrm{tr}(\underbrace{\phi_1^\top(x')\phi_1(x)}_{k_1(x,x')})k_2(x, x') = k_1(x, x')\, k_2(x, x')$$

- In simple words, the product of $k_1 k_2$ defines a valid inner product.
- The sum and product rules allow us to define a wide variety of kernels.

Lemma (Polynomial kernels)

Let $x, x' \in \mathbb{R}^d$ for $d \geq 1$, and let $m \geq 1$ be an integer and $c \geq 0$ be a positive real. Then

$$k(x, x') := (\langle x, x' \rangle + c)^m$$

is a valid kernel.

- We can also extend this combination of sum and product rules to sums with *infinitely many terms*.

Definition (p-summable sequences)

The space ℓ_p of the p-summable sequences is defined as all sequences $(a_i)_{i \geq 1}$ for which

$$\sum_{i=1}^{\infty} a_i^p < \infty.$$

p-summable sequences and kernels

- Kernels can be defined in terms of sequences in ℓ_2.

Lemma

Given a non-empty set $\mathcal{X}$, and a sequence of functions $(\phi_i(x))_{i\geq 1}$ in ℓ_2 where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the ith coordinate of the feature map $\phi(x)$. Then

$$k(x, x') := \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')$$

is a well-defined kernel in $\mathcal{X}$.

- So I can write a kernel even if I have infinitely many features.

Infinitely many features and Taylor Series

- Taylor series expansions can be used to define kernels that have *infinitely many features*.

Definition (Taylor series kernel)

Assume we can define the Taylor series $f(z) = \sum_{n=0}^{\infty} a_n z^n \quad |z| < r, z \in \mathbb{R}$, for $r \in (0, \infty]$, with $a_n \geq 0$ for all $n \geq 0$. Define $\mathcal{X}$ to be the $\sqrt{r}$-ball in $\mathbb{R}^d$. Then for $x, x' \in \mathbb{R}^d$ such that $\|x\| < \sqrt{r}$, we have the kernel

$$k(x, x') = f(\langle x, x' \rangle) = \sum_{n=0}^{\infty} a_n \langle x, x' \rangle^n$$

Example (Exponential Kernel)

The exponential kernel on $\mathbb{R}^d$ is defined as

$$k(x, x') := \exp(\langle x, x' \rangle)$$

A famous infinite feature space kernel

Example (Exponentiated quadratic kernel)

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \underbrace{\left(\sqrt{\lambda_\ell} e_\ell(x)\right)}_{\phi_\ell(x)} \underbrace{\left(\sqrt{\lambda_\ell} e_\ell(x')\right)}_{\phi_\ell(x')}$$

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx'$$

$$p(x) = \mathcal{N}\left(0, \sigma^2\right)$$

where

$$\lambda_\ell \propto b^\ell \quad b < 1$$

$$e_\ell(x) \propto \exp\left(-(c - a)x^2\right) H_\ell(x\sqrt{2c})$$

a, b, c are functions of σ, and H_ℓ is ℓ-th order Hermite polynomial (i.e., orthogonal polynomial sequence).

Given a function of two arguments, $k(x, x')$, how can we determine if it is a valid kernel?

1. Find a feature map?
   - Sometimes this is not obvious (e.g. if the feature vector is *infinite-dimensional*, e.g. the exponentiated quadratic kernel in the last slide)
   - The feature map is *not unique*.
2. A direct property of the function: positive definiteness.

Positive definiteness

- All kernel functions are **positive definite**
- In fact, if we have a *positive definite* function, we know there exist one (or more) feature spaces for which the kernel defines the inner product.
- We are not obliged to define the feature spaces explicitly!

Definition (Positive definite functions)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1, \forall (a_1, \ldots a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k\left(x_i, x_j\right) \geq 0$$

The function $k(\cdot, \cdot)$ is strictly positive definite if, for mutually distinct x_i, the equality holds only when all the a_i are zero.

Kernels are positive definite

Theorem

Let $\mathcal{H}$ be any Hilbert space (not necessarily an ***RKHS****), $\mathcal{X}$ a non-empty set, and $\phi : \mathcal{X} \to \mathcal{H}$. Then $k(x,y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a positive definite function.*

Proof:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k\left(x_i, x_j\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle a_i \phi\left(x_i\right), a_j \phi\left(x_j\right)\right\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} a_i \phi\left(x_i\right), \sum_{j=1}^{n} a_j \phi\left(x_j\right)\right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{n} a_i \phi\left(x_i\right) \right\|_{\mathcal{H}}^{2} \geq 0
\end{aligned}$$

The reproducing kernel Hilbert space

- So far, I have introduced some notation and properties on feature spaces and kernels.
- We conclude that these kernels are positive definite.
- In this section, we use these kernels to define functions on $\mathcal{X}$.

Constructing an RKHS from a Kernel

- In this section, we claim how any *positive definite* kernel function k defined in the Cartesian product space $\mathcal{X} \times \mathcal{X}$ can be used to construct a particular Hilbert *space of functions* on $\mathcal{X}$.
- This Hilbert space is *unique*, and has the following property

Lemma (Kernel Trick)

$\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}$,

$$\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$$

- Let us see an example!

Finite space, polynomial features

- From our motivating example we define a mapping $\phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mapsto \quad \phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}$$

  with kernel

$$k(x, y) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix}$$

- Let's now define a function of the features x_1, x_2 and $x_1 x_2$ of x, namely:

$$f(x) = a x_1 + b x_2 + c x_1 x_2$$

- The function f belongs to a *space of functions* mappings from $\mathcal{X} = \mathbb{R}^2$ to $\mathbb{R}$.

Finite space, polynomial features

- Defining an equivalent representation for f, we can define

$$f(\cdot) = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

- People sometimes write f rather than $f(\cdot)$. The notation $f(x) \in \mathbb{R}$ refers to the function evaluated at a particular point.
- Then, we can write

$$\begin{aligned} f(x) &= f(\cdot)^\top \phi(x) \\ &:= \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}} \end{aligned}$$

- Meaning that the evaluation of f at x can be written as an *inner product in feature space* and $\mathcal{H}$ is a *space of functions* mapping from $\mathbb{R}^2$ to $\mathbb{R}$

Functions of infinitely many features

- We can write a function of infinitely many features with an *exponentiated quadratic kernel*.

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} f_\ell \phi_\ell(x)$$

  where the expression is *bounded* in absolute value as long as $\sum_{\ell=1}^{\infty} f_\ell^2 < \infty$

$$f(x) = \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^{m} \alpha_i \phi_\ell(x_i) \right)}_{f_\ell} \phi_\ell(x)$$

$$= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}$$

$$= \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

- Nice! We got a function of *infinitely many features* expressed using *m coefficients*

Theorem

Given any positive definite kernel function k, there is a unique Hilbert space $\mathcal{H}$ in which the kernel satisfies reproducing property. It is known as the reproducing kernel Hilbert space associated with k.

- So there are 2 defining features of an **RKHS**:
  1. The feature map of every point is a function: $k(\cdot, x) = \phi(x) \in \mathcal{H}$ for any $x \in \mathcal{X}$, and

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}.$$

  2. The *reproducing property* : $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$

EMORY UNIVERSITY

Lemma (Tensor Products)

Suppose that $\mathcal{H}_1$ and $\mathcal{H}_2$ are reproducing kernel Hilbert spaces of real-valued functions with domains $\mathcal{X}_1$ and $\mathcal{X}_2$, and equipped with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the tensor product space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ is an ***RKHS*** *of real-valued functions with domain $X_1 \times \mathcal{X}_2$, and with kernel function*

$$\mathcal{K}\left(\left(x_1, x_2\right), \left(x_1', x_2'\right)\right) = \mathcal{K}_1\left(x_1, x_1'\right) \mathcal{K}_2\left(x_2, x_2'\right).$$

Fitting via kernel ridge regression

- The **RKHS** is a practical hypothesis space for *nonparametric regression*.
- Consider the output $Y \in \mathbb{R}$ and the input $W \in \mathcal{W}$.
- Our goal is to estimate the *conditional expectation function* $\gamma_0(w) = \mathbb{E}[Y \mid W = w]$

Definition

A *kernel ridge regression* estimator of γ_0 is

$$\hat{\gamma} = \arg\min_{\gamma \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \langle \gamma, \phi(W_i) \rangle_{\mathcal{H}}\}^2 + \lambda \|\gamma\|_{\mathcal{H}}^2,$$

where $\lambda > 0$ is a hyperparameter on the ridge penalty $\|\gamma\|_{\mathcal{H}}^2$, which imposes *smoothness* in estimation.

- The feature map takes a value in the original space $w \in \mathcal{W}$ and maps it to a feature in the **RKHS** $\phi(w) \in \mathcal{H}$.

- The solution to the optimization problem has a well-known closed form (Kimeldorf & Wahba, 1971), given by :

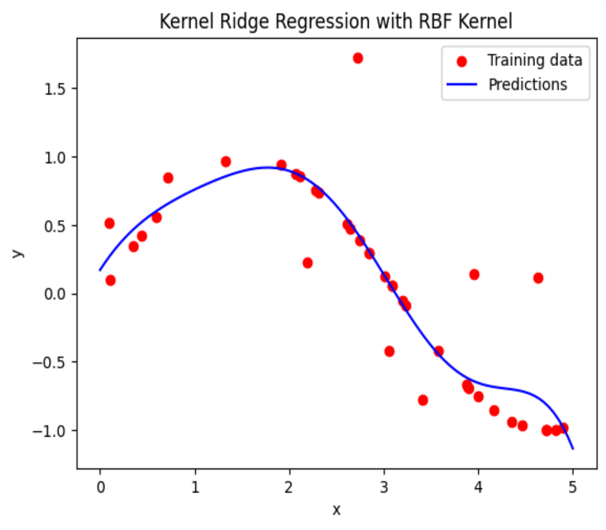
$$\hat{\gamma}(w) = Y^{\top} \left(\mathcal{K}_{WW} + n\lambda I\right)^{-1} \mathcal{K}_{Ww}.$$

- The closed-form solution involves the *kernel matrix* $\mathcal{K}_{WW} \in \mathbb{R}^{n \times n}$ with (i, j) -th entry $\mathcal{K}\left(W_i, W_j\right)$ and the *kernel vector* $\mathcal{K}_{Ww} \in \mathbb{R}^{n}$ with i th entry $\mathcal{K}\left(W_i, w\right)$.
- To tune the ridge hyperparameter λ, both *generalized cross-validation* and *leave-one-out cross-validation* have closed-form solutions, and the former is asymptotically optimal (Craven & Wahba, 1978; Li, 1986).

How to fit a KRR in python?

```
from sklearn.kernel_ridge import KernelRidge

# Generate synthetic data
X_train = np.sort(5 * np.random.rand(40, 1), axis=0)
y_train = np.sin(X_train).ravel()
# Add noise to every fifth data point
y_train[::5] += 3 * (0.5 - np.random.rand(8))
X_test = np.linspace(0, 5, 100)[:, np.newaxis]

# Fit Kernel Ridge Regression model
alpha = 1e-5  # Regularization parameter
kernel = 'rbf'  # Gaussian Radial Basis Function (RBF) kernel
gamma = 0.1  # Kernel coefficient for RBF kernel

kr = KernelRidge(alpha=alpha, kernel=kernel, gamma=gamma)
kr.fit(X_train, y_train)

# Predict
y_pred = kr.predict(X_test)
```

How to fit a KRR in python?



Application: Kernel Ridge Regression with Continuous Treatments

Kernel Methods for Causal Functions

- Singh et al. (2023) use **RKHS** and the Riesz representer theorem to propose a nonparametric estimator for *dose response curves* and other causal parameters that are inner products of kernel ridge regression.
- Treatments and covariates may be *discrete* or *continuous*
- The estimator has a closed-form solution due to the use of *kernel trick* specific to **RKHS**.
- Empirical Application: Using the Job Corps training experiment, they showed that different *intensities* of *job training* (e.g., hours of training) have smooth effects on counterfactual employment.

Kernel Methods for Causal Functions

- Let the treatment be a *continuous treatment* D and some covariates $X \in \mathcal{X}$.
- The *dose response*, a generalization of ATE, is given by $\theta_0^{\mathrm{ATE}}(d) = E\left\{Y^{(d)}\right\}$, which is the counterfactual mean outcome given the intervention $D = d$ for the entire population P.
- Under the selection on observables assumptions, we can identify the causal function of interest as an integral of the regression function γ_0 such that

$$\theta_0^{\mathrm{ATE}}(d) = \int \gamma_0(d, x) \mathrm{d}P(x)$$

  where $\gamma_0(d, x) = E(Y \mid D = d, X = x)$
- Estimation of nonparametric causal functions such as θ_0^{ATE} are *computationally demanding*.
- The *reproducing kernel Hilbert space* **RKSH** $\mathcal{H}$ solves the technical issues when estimating causal functions with a *continuous treatment*.

Riesz representation

- With continuous treatment, fix the values d and define the *linear functional* $F : \gamma_0 \mapsto \int \gamma_0(d, x)\mathrm{d}P(x)$ so that the dose response curve evaluated at d is $\theta_0(d) = F(\gamma_0)$.
- By the *Riesz representation theorem*, since F is a *bounded linear functional* over a Hilbert space, it admits an inner product representation within that Hilbert space: there exists some $\tilde{\alpha}_0 \in \mathcal{H}$ such that $F(\gamma) = \langle \gamma, \tilde{\alpha}_0 \rangle_{\mathcal{H}}$ for all $\gamma \in \mathcal{H}$.
- In particular, $\theta_0(d) = \langle \gamma_0, \tilde{\alpha}_0 \rangle_{\mathcal{H}}$.
- The Riesz representation separates the steps of nonparametric causal estimation in the **RKHS** into three simple steps:
  1. Estimate the regression γ_0
  2. Estimate $\tilde{\alpha}_0$, which embeds $P(x)$
  3. Computer their inner product.
- Based on this argument, they propose nonparametric estimators that are *inner products* of *kernel ridge regressions*, which therefore have *closed-form solutions*.

Decoupled representation and closed-form solution

- Let $k_{\mathcal{D}} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be measurable *positive definite kernels* corresponding to scalar-valued **RKHS**s $\mathcal{H}_{\mathcal{D}}$ and $\mathcal{H}_{\mathcal{X}}$.
- Define the *feature maps* $\phi_{\mathcal{D}} : \mathcal{D} \to \mathcal{H}_{\mathcal{D}}, d \mapsto k_{\mathcal{D}}(d, \cdot); \phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}, x \mapsto k_{\mathcal{X}}(x, \cdot)$.
- We assume that regression γ_0 is an element of **RKHS** $\mathcal{H}$ with *kernel* $k((d,x);(d',x')) = k_{\mathcal{D}}(d,d')\, k_{\mathcal{X}}(x,x')$.
- Under **RKHS** regularity conditions, $\theta_0^{\mathrm{ATE}}(d) = \langle \gamma_0, \phi(d) \otimes \mu_x \rangle_{\mathcal{H}}$, where $\mu_x = \int \phi(x) \mathrm{d}P(x)$;

Lemma (Estimation)

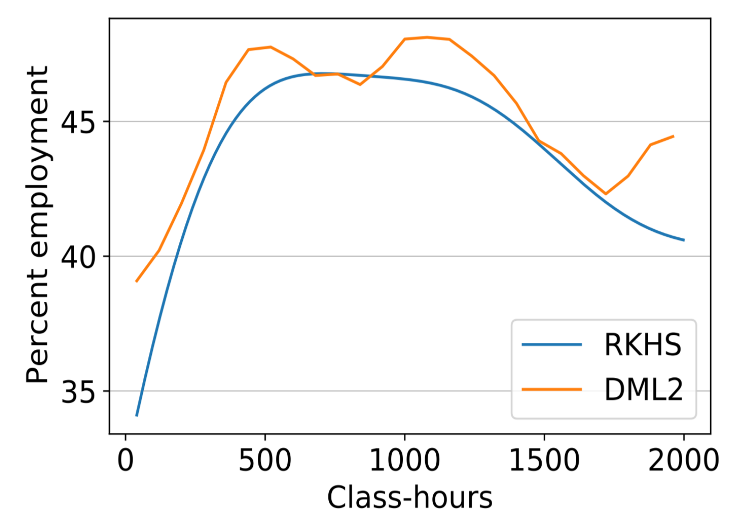
Denote by $K_{DD}, K_{XX} \in \mathbb{R}^{n \times n}$ the empirical kernel matrices calculated from observations drawn from population P. Denote by $\odot$ the elementwise product. The causal function estimator has closed-form solution

$$\hat{\theta}^{\mathrm{ATE}}(d) = n^{-1} \sum_{i=1}^{n} Y^{\top} \left(K_{DD} \odot K_{XX} + n\lambda I\right)^{-1} \left(K_{Dd} \odot K_{Xx_i}\right)$$

US Job Corps

- This paper estimates the dose response function of the *Job Corps*, the largest job training program for disadvantaged youth in the United States.
- Although access to the program was randomized, the participants could decide whether to participate and how many hours.
- The continuous treatment $D \in \mathbb{R}$ is the *total hours* spent in academic or vocational classes in the first year after randomization, and the continuous outcome $Y \in \mathbb{R}$ is the *proportion of weeks* employed in the second year after randomization.
- The covariates $X \in \mathbb{R}^{40}$ include age, gender, ethnicity, language competency, education, marital status, household size, household income, previous receipt of social aid, family background, health and health-related behavior at the base line.
- The dose response curve plateaus and reaches its maximum around $d = 500$, corresponding to 12.5 weeks of classes.

Results



Percent employment
45
40
35
0
500
1000
1500
2000
Class-hours
RKHS
DML2

Thanks!

marcelo.ortiz@emory.edu
marcelortiz.com
@marcelortizv

References

- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.
- Gretton, A. (2019). *Introduction to RKHS, and some simple kernel algorithms*.
- Singh, R., Xu, L., & Gretton, A. (2024). *Kernel methods for causal functions: Dose, heterogeneous and incremental response curves*. Biometrika, 111(2), 497–516. https://doi.org/10.1093/biomet/asad042