A rigorous theory of conditional mean embeddings

I. Klebanov[1] I. Schuster[2] **T. J. Sullivan**[1,3]

Mathematical Statistics Research Seminar
WIAS, Berlin, DE
5 February 2020

[1]**Zuse Institute Berlin, DE**
[2]Zalando Research, DE
[3]**Freie Universität Berlin, DE**

Outline

Motivation and context

Reproducing kernel Hilbert spaces

Kernel mean embedding

Conditional mean embedding

Rigorous settings for conditional mean embedding

Closing remarks

Motivation and context

- The MATH+ industrial transfer project TrU-2 *"Demand modelling and control for e-commerce using RKHS transfer operator approaches"* aims to transfer recent advances in kernel-based representations of operators and kernel-based machine learning into a specific commericial setting.
- We are aiming to — based on Zalando sales data — learn and then control the essential dynamics underlying time-dependent demand.

Aim = an empty warehouse and a full purse at season's end.

- Hence, we are interested in embedding probability distributions over time series into appropriate RKHS feature spaces, and performing conditioning.

Executive summary

- Parameters and data that live in nonlinear high- or infinite-dimensional spaces can be embedded "faithfully" into reproducing kernel Hilbert spaces, as can probability distributions over such parameters and data.
- Conditioning, e.g. for Bayesian inference, can be performed using surprisingly simple linear algebra in the RKHS.
- The current literature *almost* gets this right…but there are some typos, some mistaken or hard-to-check assumptions, and misapplied results.
- Our aim is to rigorously establish the conditional mean embedding formula under relaxed but also verifiable conditions.

Reproducing kernel Hilbert spaces

Theorem

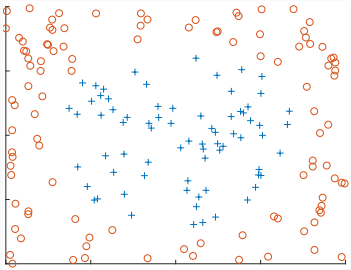
Let $\mathcal{H}$ be a Hilbert space of real-valued functions on a set $\mathcal{X}$. Then the following are equivalent:

- *For every $x \in \mathcal{X}$, the point evaluation functional $\delta_x \colon \mathcal{H} \to \mathbb{R}$, $\langle \delta_x | u \rangle := u(x)$, is bounded, i.e. in the dual space $\mathcal{H}'$.*
- *There exists a map $\phi \colon \mathcal{X} \to \mathcal{H}$ with the reproducing property: for all $u \in \mathcal{H}$ and all $x \in \mathcal{X}$, $u(x) = \langle \phi(x), u \rangle_{\mathcal{H}}$.*
- *There exists a symmetric and positive-definite function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and, for all $x \in \mathcal{X}$ and $u \in \mathcal{H}$, $u(x) = \langle k(x, \cdot), u \rangle_{\mathcal{H}}$.*

If one (and hence any) of these conditions hold, then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS), k its reproducing kernel, and ϕ its (canonical) feature map.

Reproducing kernel Hilbert spaces

- The Moore–Aronszajn theorem tells us that $\mathcal{H} = \overline{\operatorname{span}\{k(x, \cdot) \colon x \in \mathcal{X}\}}$ with inner product given by
$$\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}} := k(x, x').$$
- RKHSs are long-established and well-studied tools for machine learning.
- The successes of RKHSs are usually attributed to the kernel trick: many nonlinear statements about functions on $\mathcal{X}$ are turned into linear algebra in the RKHS $\mathcal{H}$.
- RKHS structure also allows for nice representation of linear operators on $\mathcal{H}$, singular value decompositions / principal component analysis etc. (Mollenhauer (2018))
- Important to bear in mind: the space $\mathcal{H}$ exists "on the blackboard" but is almost always only accessed indirectly via the kernel / feature map — ditto for linear operators on $\mathcal{H}$.



$$\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (x_1^2, x_1 x_2, x_2^2)$$

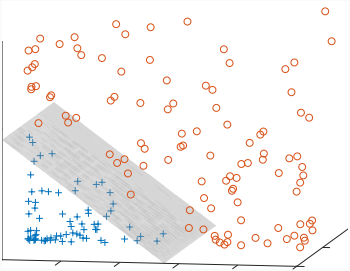


Figure 1: Example of the kernel trick: under the nonlinear feature map φ, the separation of the blue crosses from the orange circles becomes the problem of finding a plane in the feature space $\mathbb{R}^3$ that separates the images of these points.

Kernel mean embedding

Definition

Let X be a random variable with distribution $\mathbb{P}_X$ on $\mathcal{X}$, and let $\mathcal{H}$ be an RKHS over $\mathcal{X}$ with canonical feature map φ. The kernel mean embedding of X (or $\mathbb{P}_X$) is

$$\mu_X := \mathbb{E}[\varphi(X)] \equiv \int_{\mathcal{X}} \varphi(x)\, \mathbb{P}_X(\mathrm{d}x) \in \mathcal{H},$$

which is well defined if $\mathbb{E}[\|\varphi(X)\|_{\mathcal{H}}]$ is finite.

For any $h \in \mathcal{H}$, the reproducing property yields

$$\langle h, \mu_X \rangle_{\mathcal{H}} = \mathbb{E}[\langle h, \varphi(X) \rangle_{\mathcal{H}}] = \mathbb{E}[h(X)] \in \mathbb{R},$$

i.e. the function $\mu_X \colon \mathcal{X} \to \mathbb{R}$ is the embedding into $\mathcal{H}$ of the operation "integrate with respect to $\mathbb{P}_X$".

Definition

The RKHS $\mathcal{H}$ is called characteristic if the linear map

$$\mathbb{P}_X \mapsto \int_{\mathcal{X}} \varphi(x)\, \mathbb{P}_X(\mathrm{d}x) \in \mathcal{H}$$

is injective.

- This is, in some sense, a measure of expressivity of the kernel / feature map.
- It is known, for example, that the Gaussian kernel k on a bounded domain $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is characteristic.
- See Sriperumbudur et al. (2010) for necessary and sufficient conditions for a kernel to be characteristic, and for relationships to weak convergence.

Characteristic, universal, etc... (Sriperumbudur et al., 2010)

Sriperumbudur, Gretton, Fukumizu, Schölkopf and Lanckriet

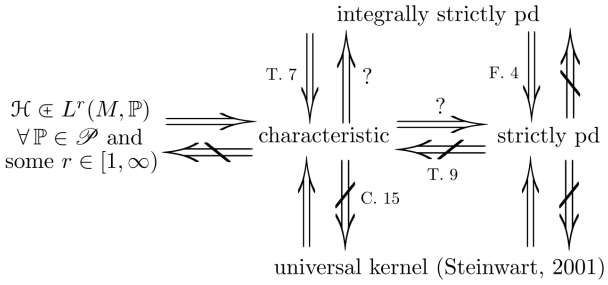


Figure 1: Summary of the relations between various families of kernels is shown along with the reference. The letters "C", "F", and "T" refer to Corollary, Footnote and Theorem respectively. For example, T. 7 refers to Theorem 7. The implications which are open problems are shown with "?". $A \Subset B$ indicates that A is a dense subset of B. Refer to Section 3.4 for details.

The outer product of $a, b \in \mathcal{H}$ is a linear operator from $\mathcal{H}$ into itself,

$$(a \otimes b)c := a\langle b, c\rangle_{\mathcal{H}}.$$

Definition

Let X be a random variable with distribution $\mathbb{P}_X$ on $\mathcal{X}$, and let $\mathcal{H}$ be an RKHS over $\mathcal{X}$ with canonical feature map φ. The uncentred kernel covariance operator of X (or $\mathbb{P}_X$) is

$${}^{u}C_X := \mathbb{E}[\varphi(X) \otimes \varphi(X)] \equiv \int_{\mathcal{X}} \varphi(x) \otimes \varphi(x)\, \mathbb{P}_X(\mathrm{d}x) \colon \mathcal{H} \to \mathcal{H},$$

which is well defined if $\mathbb{E}[\|\varphi(X)\|^2_{\mathcal{H}}]$ is finite. The centred kernel covariance operator of X (or $\mathbb{P}_X$) is

$$C_X := \mathbb{E}[(\varphi(X) - \mu_X) \otimes (\varphi(X) - \mu_X)] \equiv {}^{u}C_X - \mu_X \otimes \mu_X \colon \mathcal{H} \to \mathcal{H}.$$

Definition

Let (X, Y) be a random variable with distribution $\mathbb{P}_{(X,Y)}$ on $\mathcal{X} \times \mathcal{Y}$, and let $\mathcal{H}$ be an RKHS over $\mathcal{X}$ with canonical feature map φ, and $\mathcal{G}$ an RKHS over $\mathcal{Y}$ with canonical feature map ψ. The uncentred kernel cross-covariance operator of X and Y is

$$ {}^{u}C_{XY} := \mathbb{E}[\varphi(X) \otimes \psi(Y)] \equiv \int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) \otimes \psi(y)\, \mathbb{P}_{(X,Y)}(\mathrm{d}x, \mathrm{d}y) \colon \mathcal{G} \to \mathcal{H}, $$

which is well defined if $\mathbb{E}[\|\varphi(X)\|^2_{\mathcal{H}}]$ and $\mathbb{E}[\|\psi(Y)\|^2_{\mathcal{G}}]$ are finite. The centred kernel cross-covariance operator of X and Y is

$$ C_{XY} := \mathbb{E}[(\varphi(X) - \mu_X) \otimes (\psi(Y) - \mu_Y)] \equiv {}^{u}C_{XY} - \mu_X \otimes \mu_Y \colon \mathcal{G} \to \mathcal{H}. $$

- ${}^{u}C_X$ is the KME of $\mathbb{P}_X$ under the $\mathcal{H} \otimes \mathcal{H}$-valued tensor product feature map $\varphi \otimes \varphi$; it is a Hilbert–Schmidt operator on $\mathcal{H}$; similarly for C_X.
- The reproducing properties imply that, for all $h_1, h_2 \in \mathcal{H}$,

$$\langle h_1, {}^{u}C_X h_2 \rangle_{\mathcal{H}} := \mathbb{E}[\langle h_1, \varphi(X) \rangle_{\mathcal{H}} \langle h_2, \varphi(X) \rangle_{\mathcal{H}}] = \mathbb{E}[h_1(X) h_2(X)] \in \mathbb{R}$$

  and

$$\langle h_1, C_X h_2 \rangle_{\mathcal{H}} := \mathbb{E}[\langle h_1, \varphi(X) - \mu_X \rangle_{\mathcal{H}} \langle h_2, \varphi(X) - \mu_X \rangle_{\mathcal{H}}] = \mathrm{Cov}[h_1(X), h_2(X)] \in \mathbb{R}.$$

- Similarly, for all $h, h_1, h_2 \in \mathcal{H}$ and $g \in \mathcal{G}$,

$$\langle h, {}^{u}C_{XY} g \rangle_{\mathcal{H}} = \mathbb{E}[h(X), g(Y)] \in \mathbb{R},$$
$$\langle h, C_{XY} g \rangle_{\mathcal{H}} = \mathrm{Cov}[h(X), g(Y)] \in \mathbb{R}.$$

Conditional mean embedding

- As before, we consider random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, with joint distribution $\mathbb{P}_{(X,Y)}$, and RKHSs

$$(\mathcal{H}, \mathcal{X}, k, \varphi) \qquad\qquad (\mathcal{G}, \mathcal{Y}, \ell, \psi).$$

- We will think of Y as being parameters (with prior distribution $\mathbb{P}_X$) that we wish to condition on an observation $X = x$.
- Under mild assumptions[1] there is a $\mathbb{P}_X$-a.e. uniquely defined regular version of the conditional distribution $\mathbb{P}_{Y|X=x}$.
- **Question.** How are the KMEs μ_X, μ_Y, and $\mu_{Y|X=x}$ related?

[1] $\mathcal{X}$ needs measurable structure, and $\mathcal{Y}$ needs a measurable structure isomorphic to a Borel subset of $[0, 1]$; in particular, this holds if $\mathcal{Y}$ is Polish. See e.g. Kallenberg (2006).

Recall: Gaussian conditioning

- Suppose that we have a *bona fide* Gaussian random variable (U, V) taking values in $\mathcal{G} \oplus \mathcal{H}$, where the mean and covariance have the block structure

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_U \\ \mu_V \end{pmatrix}, \begin{pmatrix} C_U & C_{UV} \\ C_{VU} & C_V \end{pmatrix} \right).$$

- The well-known Sherman–Morrison–Woodbury / Schur complement formula for the Gaussian conditional random variable $U|V = v$ is

$$\begin{aligned} (U|V = v) &\sim \mathcal{N}(\mu_{U|V=v}, C_{U|V=v}) \\ \mu_{U|V=v} &= \mu_U + C_{UV} C_V^{-1} (v - \mu_V) \\ C_{U|V=v} &= C_U - C_{UV} C_V^{-1} C_{VU} \end{aligned}$$

  provided C_V is invertible and the data space $\mathcal{H}$ is finite-dimensional. (For the more general setting, see e.g. Owhadi and Scovel (2018).)

A shockingly naïve idea I

- Now...our embedded random variable $(\psi(Y), \varphi(X))$ is not Gaussian.
- Naïve proposal: the conditional mean embedding, i.e. the representative in $\mathcal{G}$ of $\mathbb{P}_{Y|X=x}$, is obtained by pretending that $(\psi(Y), \varphi(X))$ is Gaussian:

$$\mu_{Y|X=x} = \mu_Y + C_{YX}C_X^{-1}(\varphi(x) - \mu_X)???$$

- Does this formula, or anything like it, even stand a chance of being true? If so, then what are rigorous conditions for its validity?
- Must $\varphi(x) - \mu_X$ lie in $\operatorname{ran} C_X$? If not, how do we make sense of $C_{YX}C_X^{-1}(\varphi(x) - \mu_X)$?
- Can we handle non-invertible C_X and $\dim \mathcal{H} = \infty$?
- Why should $\mu_Y + C_{YX}C_X^{-1}(\varphi(x) - \mu_X)$ be the KME of anything at all, let alone of $Y|X = x$?

A shockingly naïve idea II

- Dare we apply the Gaussian conditioning formula with $U = \psi(Y)$, $V = \varphi(X)$?
- A priori, there seems to be no reason why $(\psi(Y), \varphi(X))$ should behave so much like a Gaussian with the same two moments that we may condition as if it were Gaussian.
- The data RKHS $\mathcal{H}$ is basically always ∞-dimensional (e.g. long time series in $\mathcal{X}$, furthermore mapped through φ).
- Slogans:
  - Life is not as simple as replacing inverses by pseudo-inverses.
  - Life is surprisingly good in RKHSs, but still one cannot be naïve.

- $(\mathcal{H}, \mathcal{X}, k, \varphi), (\mathcal{G}, \mathcal{Y}, \ell, \psi)$ RKHSs
- X, Y = random variables in $\mathcal{X}, \mathcal{Y}$ with joint distribution $\mathbb{P}_{XY}$
- Consider the random variable $(\psi(Y), \varphi(X))$ in $\mathcal{G} \oplus \mathcal{H}$ and define

kernel mean embeddings (KME) kernel (cross-) covariance operators

$$\mu := \mathbb{E}\left[\begin{pmatrix}\psi(Y)\\ \varphi(X)\end{pmatrix}\right] = \begin{pmatrix}\mu_Y\\ \mu_X\end{pmatrix}, \qquad C := \mathrm{Cov}\left[\begin{pmatrix}\psi(Y)\\ \varphi(X)\end{pmatrix}\right] = \begin{pmatrix}C_Y & C_{YX}\\ C_{XY} & C_X\end{pmatrix}.$$

- $\mu_{Y|X=x} = \mathbb{E}[\psi(Y)|X=x], \qquad C_{Y|X=x} = \mathrm{Cov}[\psi(Y)|X=x].$
- Basic assumptions:

$$\mathbb{E}[\|\varphi(X)\|^2_{\mathcal{H}}] < \infty, \qquad \mathbb{E}[\|\psi(Y)\|^2_{\mathcal{G}}] < \infty, \qquad \mathbb{E}[\|\psi(Y)\|^2_{\mathcal{G}}|X=x] < \infty,$$

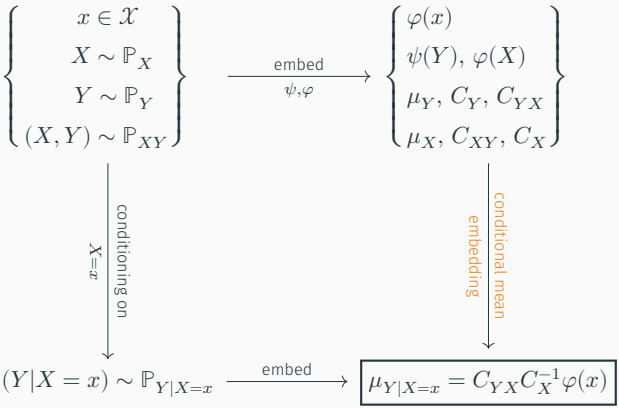
which imply existence of the KMEs and covariance operators and that

$$\mathcal{H} \subseteq \mathcal{L}^2(\mathbb{P}_X), \qquad \mathcal{G} \subseteq \mathcal{L}^2(\mathbb{P}_Y), \qquad \mathcal{G} \subseteq \mathcal{L}^2(\mathbb{P}_{Y|X=x}).$$

Conditional mean embeddings: idea

"observation" space $\mathcal{X}$
"parameter" space $\mathcal{Y}$

feature spaces $\mathcal{H}, \mathcal{G}$
(RKHS)



Conditional mean embeddings (Song et al., 2009, Theorem 4)

- For $g \in \mathcal{G}$ define $f_g(x) = \mathbb{E}[g(Y)|X = x]$ and assume $\boxed{f_g \in \mathcal{H}}$
- Note that $f_{\psi(y)}(x) = \mathbb{E}[\ell(y, Y)|X = x] = \mu_{Y|X=x}(y)$.
- Observe that, for $h \in \mathcal{H}$ and $g \in \mathcal{G}$,

$$\langle h, C_X f_g \rangle_{\mathcal{H}} = \mathrm{Cov}[h(X), f_g(X)] = \mathrm{Cov}[h(X), g(Y)] = \langle h, C_{XY} g \rangle_{\mathcal{H}}$$

$$\Rightarrow \; C_X f_g = C_{XY} g \qquad \Longrightarrow \qquad f_g = C_X^{-1} C_{XY} g$$

- Reproducing properties yield

$$\begin{aligned} \mu_{Y|X=x}(y) \qquad &= \langle f_{\psi(y)}, \varphi(x) \rangle_{\mathcal{H}} && = \langle C_X^{-1} C_{XY} \psi(y), \varphi(x) \rangle_{\mathcal{H}} \\ &= \langle \psi(y), C_{YX} C_X^{-1} \varphi(x) \rangle_{\mathcal{G}} && = (C_{YX} C_X^{-1} \varphi(x))(y) \end{aligned}$$

$$\Rightarrow \qquad \boxed{\mu_{Y|X=x} = C_{YX} C_X^{-1} \varphi(x)}$$

- This formula is *never* applicable for independent X, Y!

Problems with independence and constant functions

- In the trivial case where X and Y are independent, the CME should yield $\mu_{Y|X=x} = \mu_Y$. However, independence implies that $C_{XY} = 0$, and so the CME formula of Song et al. (2009) yields $\mu_{Y|X=x} = 0$, regardless of x.
- In order to understand what has gone wrong it is helpful to consider in turn the two cases in which the constant function $\mathbf{1}_{\mathcal{X}} \colon x \mapsto 1$ is, or is not, an element of $\mathcal{H}$.
  - If $\mathbf{1}_{\mathcal{X}} \in \mathcal{H}$, then C_X cannot be injective, since $C_X \mathbf{1}_{\mathcal{X}} = 0$, and the CME formula of Song et al. (2009) is not applicable.
  - If $\mathbf{1}_{\mathcal{X}} \notin \mathcal{H}$ and X and Y are independent, then the assumption $\mathbb{E}[g(Y)|X = \cdot\,] \in \mathcal{H}$ cannot hold (except for those special elements $g \in \mathcal{H}$ for which $\mathbb{E}[g(Y)] = 0$ or if $\mathbb{E}[\ell(y, Y)] = 0$ for all $y \in \mathcal{Y}$, respectively).
- In summary, the CME formula of Song et al. (2009) is never applicable for independent random variables except in certain degenerate cases.

Previous work & our contribution

	centered operators (Song et al., 2009)	uncentred operators (Fukumizu et al., 2013)
previous work	$\mu_{Y\|X=x} = C_{YX}C_X^{-1}\varphi(x)$ wrong Assumption $f_g \in \mathcal{H} \quad \forall g$ Assumption $\varphi(x) \in \operatorname{ran} C_X$	$\mu_{Y\|X=x} = {}^{u}C_{YX}\,{}^{u}C_X^{-1}\varphi(x)$ correct Assumption $f_g \in \mathcal{H} \quad \forall g$ Assumption $\varphi(x) \in \operatorname{ran} {}^{u}C_X$
our contribution	$\mu_{Y\|X=x} = \mu_Y + (C_X^{\dagger}C_{XY})^*(\varphi(x) - \mu_X)$ Assumption $f_g \in \mathbb{R} + \mathcal{H} \quad \forall g$ $C_X^{\dagger}C_{XY}$ bounded operator	$\mu_{Y\|X=x} = ({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^*\varphi(x)$ Assumption $f_g \in \mathcal{H} \quad \forall g$ ${}^{u}C_X^{\dagger}\,{}^{u}C_{XY}$ bounded operator

Some things are wrong or hard to verify in practice; other things are nice.

Rigorous settings for conditional mean embedding

Assumptions for CMEs I

- We establish the CME formula under various assumptions, for centred and uncentred covariance operators.
- The assumptions are all related to what we call Assumption A,

$$\mathsf{A}: \quad \text{For all } g \in \mathcal{G},\ f_g := \mathbb{E}[g(Y)|X = \cdot] \in \mathcal{H}.$$

- For centred operators we seek to establish that

$$\mu_{Y|X=x} = \mu_Y + (C_X^\dagger C_{XY})^*(\varphi(x) - \mu_X)$$

  and for uncentred operators we seek

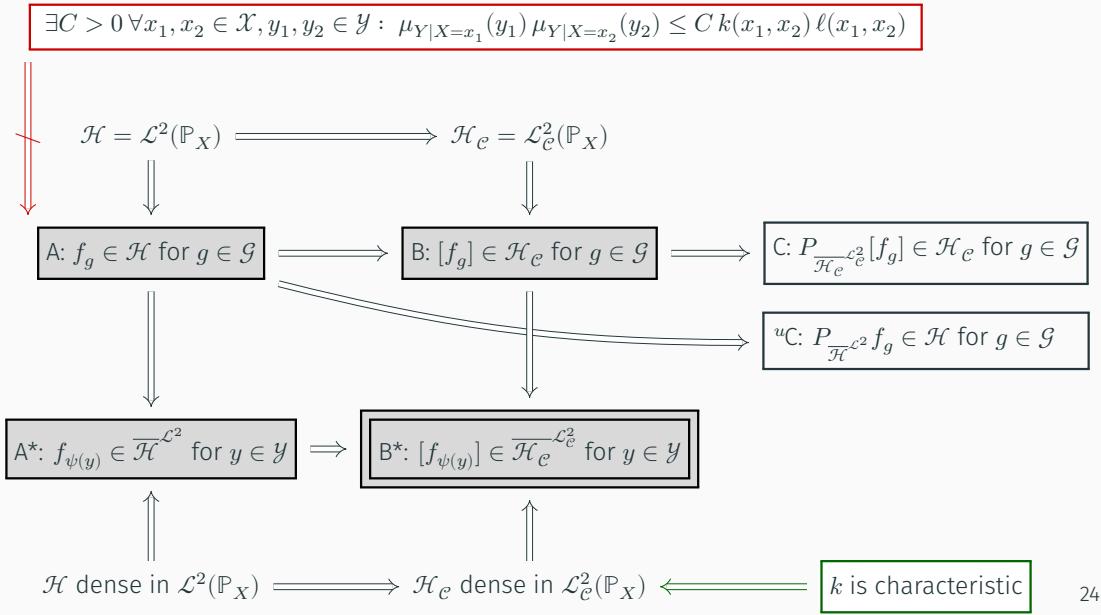
$$\mu_{Y|X=x} = ({}^uC_X^\dagger\, {}^uC_{XY})^*\varphi(x).$$

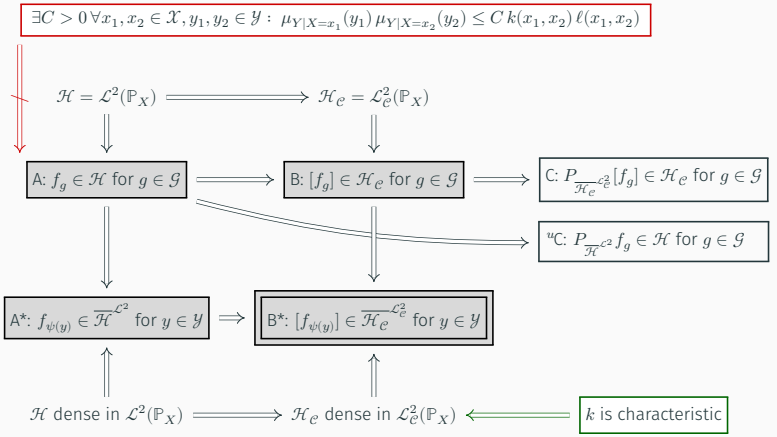
- We also seek to understand the effect of finite-dimensional approximation of the data space, and convergence as the approximation dimension tends to infinity.

- Replacement of C_X^{-1} by a Moore–Penrose pseudo-inverse $C_X^\dagger$ is a natural generalisation.
- Our principal aim is to replace Assumption A — that $\mathbb{E}[g(Y)|X = \cdot\,] \in \mathcal{H}$ for all $g \in \mathcal{G}$ — with something both weaker and practically verifiable.
- It turns out that k being characteristic is one such condition, but, unfortunately, the suggestion of Fukumizu et al. (2004) is flawed.
- To take care of issues related to constant functions we introduce

$$\begin{aligned} \mathcal{C} &:= \{f \in \mathcal{L}^2 \mid f \text{ is constant } \mathbb{P}_X\text{-a.s.}\}, \\ \mathcal{L}^2_{\mathcal{C}} &:= \mathcal{L}^2/\mathcal{C}, \\ \langle [f_1], [f_2] \rangle_{\mathcal{L}^2_{\mathcal{C}}} &:= \mathrm{Cov}[f_1(X), f_2(X)]. \end{aligned}$$

  and define $\mathcal{H}_{\mathcal{C}}$ similarly.
- Our (admittedly complex at first glance!) hierarchy of assumptions is...


$\exists C > 0\, \forall x_1, x_2 \in \mathcal{X}, y_1, y_2 \in \mathcal{Y}:\ \mu_{Y|X=x_1}(y_1)\, \mu_{Y|X=x_2}(y_2) \leq C\, k(x_1, x_2)\, \ell(x_1, x_2)$
$\mathcal{H} = \mathcal{L}^2(\mathbb{P}_X)$
$\mathcal{H}_\mathcal{C} = \mathcal{L}^2_\mathcal{C}(\mathbb{P}_X)$
A: $f_g \in \mathcal{H}$ for $g \in \mathcal{G}$
B: $[f_g] \in \mathcal{H}_\mathcal{C}$ for $g \in \mathcal{G}$
C: $P_{\overline{\mathcal{H}_\mathcal{C}}^{\mathcal{L}^2_\mathcal{C}}}[f_g] \in \mathcal{H}_\mathcal{C}$ for $g \in \mathcal{G}$
^{u}C: $P_{\overline{\mathcal{H}}^{\mathcal{L}^2}} f_g \in \mathcal{H}$ for $g \in \mathcal{G}$
A*: $f_{\psi(y)} \in \overline{\mathcal{H}}^{\mathcal{L}^2}$ for $y \in \mathcal{Y}$
B*: $[f_{\psi(y)}] \in \overline{\mathcal{H}_\mathcal{C}}^{\mathcal{L}^2_\mathcal{C}}$ for $y \in \mathcal{Y}$
$\mathcal{H}$ dense in $\mathcal{L}^2(\mathbb{P}_X)$
$\mathcal{H}_\mathcal{C}$ dense in $\mathcal{L}^2_\mathcal{C}(\mathbb{P}_X)$
k is characteristic



- Under A : $\mu_{Y|X=x} = ({}^uC_X^\dagger\, {}^uC_{XY})^* \varphi(x)$
- Under B : $\mu_{Y|X=x} = \mu_Y + (C_X^\dagger C_{XY})^* (\varphi(x) - \mu_X)$
- Under B*: $\mu_{Y|X=x} = \mu_Y + \lim_{n\to\infty} (C_X^{(n)\dagger} C_{XY}^{(n)})^* (\varphi(x) - \mu_X)$

Theorem (Centred CME; Klebanov et al., 2019)

Under Assumption C, $C_X^\dagger C_{XY} \colon \mathcal{G} \to \mathcal{H}$ is a bounded operator and, for all $y \in \mathcal{Y}$ and $h \in \mathcal{H}$,

$$\langle h, \mu_{Y|X=\cdot}(y) \rangle_{\mathcal{L}^2(\mathbb{P}_X)} = \left\langle h, (\mu_Y + (C_X^\dagger C_{XY})^* (\varphi(\cdot) - \mu_X))(y) \right\rangle_{\mathcal{L}^2(\mathbb{P}_X)}.$$

*If also k is characteristic **or** $\mathcal{H}_\mathcal{C}$ is dense in $\mathcal{L}^2_\mathcal{C}(\mathbb{P}_X)$, **or** Assumption B holds **or** $[f_{\psi(y)}] \in \mathcal{H}_\mathcal{C}$ for each $y \in \mathcal{Y}$, then, for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$,*

$$\mu_{Y|X=x} = \mu_Y + (C_X^\dagger C_{XY})^* (\varphi(x) - \mu_X).$$

- **Question:** How do CMEs behave as we see more and more of an in principle infinite-dimensional data object? (E.g. a long time series, or an image under increasing resolution.)
- Let $(h_n)_{n\in\mathbb{N}}$ be a complete orthonormal system of $\mathcal{H}$ that is an eigenbasis of C_X, let $\mathcal{H}^{(n)} := \operatorname{span}\{h_1, \dots, h_n\}$, let $\mathcal{F} := \mathcal{G} \oplus \mathcal{H}$, let $P^{(n)} \colon \mathcal{F} \to \mathcal{F}$ be the orthogonal projection onto $\mathcal{G} \oplus \mathcal{H}^{(n)}$, and let

$$C := \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix}, \qquad C^{(n)} := P^{(n)} C P^{(n)} = \begin{pmatrix} C_Y & C_{YX}^{(n)} \\ C_{XY}^{(n)} & C_X^{(n)} \end{pmatrix}.$$

Theorem (Centred CME with finite-rank approximation; Klebanov et al., 2019)

$\operatorname{ran} C_{XY}^{(n)} \subseteq \operatorname{ran} C_X^{(n)}$ and so $h_g^{(n)} := C_X^{(n)\dagger} C_{XY}^{(n)} g \in \mathcal{H}$ is well defined for each $g \in \mathcal{G}$. For each $y \in \mathcal{Y}$ and $h \in \mathcal{H}$,

$$\langle h, \mu_{Y|X=\cdot}(y) \rangle_{\mathcal{L}^2(\mathbb{P}_X)} = \lim_{n\to\infty} \left\langle h, (\mu_Y + (C_X^{(n)\dagger} C_{XY}^{(n)})^* (\varphi(\cdot) - \mu_X))(y) \right\rangle_{\mathcal{L}^2(\mathbb{P}_X)}. \tag{1}$$

*If also k is characteristic, **or** $\mathcal{H}_\mathcal{C}$ is dense in $\mathcal{L}^2_\mathcal{C}(\mathbb{P}_X)$, **or** B^* holds, then, for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$,*

$$\mu_{Y|X=x} = \mu_Y + \lim_{n\to\infty} (C_X^{(n)\dagger} C_{XY}^{(n)})^* (\varphi(x) - \mu_X). \tag{2}$$

Theorem (Uncentred CME: Klebanov et al., 2019)

Under Assumption ${}^{u}C$, the operator ${}^{u}C_X^{\dagger}\,{}^{u}C_{XY} : \mathcal{G} \to \mathcal{H}$ is bounded and, for all $y \in \mathcal{Y}$ and $h \in \mathcal{H}$,

$$\langle h, \mu_{Y|X=}\,.\,(y)\rangle_{\mathcal{L}^2(\mathbb{P}_X)} = \left\langle h, (({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^{*}\varphi(\,\cdot\,))(y)\right\rangle_{\mathcal{L}^2(\mathbb{P}_X)}.$$

*If also $\mathcal{H}$ is dense in $\mathcal{L}^2(\mathbb{P}_X)$, **or** Assumption A holds, **or** $f_{\psi(y)} \in \mathcal{H}$ for each $y \in \mathcal{Y}$, then, for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$,*

$$\mu_{Y|X=x} = ({}^{u}C_X^{\dagger}\,{}^{u}C_{XY})^{*}\,\varphi(x).$$

Connection to Gaussian conditioning in Hilbert spaces

$\mathcal{X}, \mathcal{Y}$ | $\mathcal{H}, \mathcal{G}$ | Gaussian r.v. on $\mathcal{H} \oplus \mathcal{G}$

$$\left\{\begin{array}{c} x \in \mathcal{X} \\ X \sim \mathbb{P}_X \\ Y \sim \mathbb{P}_Y \\ (X,Y) \sim \mathbb{P}_{XY} \end{array}\right\} \xrightarrow[\psi,\varphi]{\text{embed}} \left\{\begin{array}{l} \varphi(x) \\ \psi(Y), \varphi(X) \\ \mu_Y, C_Y, C_{YX} \\ \mu_X, C_{XY}, C_X \end{array}\right\} \xrightarrow[\text{on } \mathcal{H}\oplus\mathcal{G}]{\text{Gaussian}} \begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix}\right)$$

conditioning on $X=x$ | conditional mean embedding | conditioning on $V=v=\varphi(x)$

$$(Y|X=x) \sim \mathbb{P}_{Y|X=x} \xrightarrow[\psi,\varphi]{\text{embed}} \mu_{Y|X=x}, C_{Y|X=x} \xleftrightarrow[?]{\text{Gaussian on } \mathcal{G}} (U|V=v) \sim \mathcal{N}(\mu_{U|V=v}, C_{U|V=v})$$

Theorem (Klebanov et al., 2019)

Under B, for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$,*

$$\mu_{U|V=v} = \mu_{Y|X=x}, \qquad C_{U|V=v} = \mathbb{E}[C_{Y|X}] = \int_{\mathcal{X}} C_{Y|X=x} \, \mathbb{P}_X(\mathrm{d}x),$$

where $v = \varphi(x)$ and

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} C_Y & C_{YX} \\ C_{XY} & C_X \end{pmatrix} \right).$$

- Performing Gaussian conditioning on the kernel mean+covariance does yield the KME of the conditioned random variable!
- Interpretation of the conditioned covariance operator...?

Closing remarks

Summary

We now have:

- a rigorous derivation of conditional mean embeddings under weaker assumptions;
- an improved understanding of which assumptions have which implications, in particular,
  - stronger assumptions are needed for the uncentred case than the centred case, and
  - characteristic kernels are sufficient for the centered case;
- a connection to Gaussian conditioning — $(\psi(Y), \varphi(X))$ behaves astonishingly like a Gaussian random variable.

Summary

Work in progress:

- the "outbedding" that recovers $\mathbb{P}_{Y|X=x}$ from $\mu_{Y|X=x}$ — cf. Fukumizu et al. (2013);
- the effect on our results of empirical approximation using IID data $(x_m, y_m) \sim \mathbb{P}_{XY}$:

$$\mu_X \approx \hat{\mu}_X := \frac{1}{M} \sum_{m=1}^{M} \varphi(x_m),$$

$$C_X \approx \widehat{C}_X := \frac{1}{M-1} \sum_{m=1}^{M} (\varphi(x_m) - \hat{\mu}_X) \otimes (\varphi(x_m) - \hat{\mu}_X)$$

$$C_{XY} \approx \widehat{C}_{XY} := \frac{1}{M-1} \sum_{m=1}^{M} (\varphi(x_m) - \hat{\mu}_X) \otimes (\psi(y_m) - \hat{\mu}_Y).$$

(These are empirical RKHS operators in the sense of Mollenhauer (2018), so the linear algebra is nice, but the convergence is not yet clear.)

Summary

Work in progress:

- the "outbedding" that recovers $\mathbb{P}_{Y|X=x}$ from $\mu_{Y|X=x}$ — cf. Fukumizu et al. (2013);
- the effect on our results of empirical approximation using IID data $(x_m, y_m) \sim \mathbb{P}_{XY}$:

$$\mu_X \approx \hat{\mu}_X := \frac{1}{M} \sum_{m=1}^{M} \varphi(x_m),$$

$$C_X \approx \widehat{C}_X := \frac{1}{M-1} \sum_{m=1}^{M} (\varphi(x_m) - \hat{\mu}_X) \otimes (\varphi(x_m) - \hat{\mu}_X)$$

$$C_{XY} \approx \widehat{C}_{XY} := \frac{1}{M-1} \sum_{m=1}^{M} (\varphi(x_m) - \hat{\mu}_X) \otimes (\psi(y_m) - \hat{\mu}_Y).$$

(These are empirical RKHS operators in the sense of Mollenhauer (2018), so the linear algebra is nice, but the convergence is not yet clear.)

References

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 5(Jan):73–99, 2004. www.jmlr.org/papers/volume5/fukumizu04a/fukumizu04a.pdf.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *J. Mach. Learn. Res.*, 14(1):3753–3783, 2013. jmlr.org/papers/volume14/fukumizu13a/fukumizu13a.pdf.

O. Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, 2006. doi:doi.org/10.1007/978-1-4757-4015-8.

I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings, 2019. arXiv:1912.00671.

M. Mollenhauer. Singular value decomposition of operators on reproducing kernel Hilbert spaces. Master's thesis, Freie Universität Berlin, 2018.

H. Owhadi and C. Scovel. Conditioning Gaussian measure on Hilbert space. *J. Math. Stat. Anal.*, 1(1):1–15, 2018.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009. doi:10.1145/1553374.1553497.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010. www.jmlr.org/papers/volume11/sriperumbudur10a/sriperumbudur10a.pdf.