Reproducing Kernel Hilbert Spaces and their application in signal processing

P.O. Amblard

July 3, 2014

Contents

This document consists in notes concerning the use of RKHS in nonlinear signal processing. It is intended to be a brief introduction. We try however to give some precise mathematical facts when necessary.

1 Motivations

Escaping from linearity in science is not an easy task. However, it is necessary since Nature acts mostly in a nonlinear way. Linearity is interesting since it is quite simple, and yet it provides good models to understand many phenomena. However, it cannot grasps fine details and cannot explain several fundamental laws. Turbulence in fluids is a highly nonlinear phenomenon. Frequency doubling in optics is a nonlinear phenomenon. Chaos arises in nonlinear systems. Stochastic resonance (constructive role of noise) is present only in nonlinear systems.

But if linearity is well defined (superposition principle), nonlinearity is not: It is a nonproperty. There are infinitely many ways of performing nonlinearly!

Imposing linearity in signal processing leads to the powerful concept of linear filtering, a class of operation that can be handled nicely with the notion of correlation function, power spectral densities. . . When dealing with random signals, almost all linear operations can be performed and understood by using only second order statistics. This is not the case for nonlinear processing in general.

However, certain classes of nonlinear processing can be handled quite easily: Classes where the processing can be described in some linear spaces. This is the case for example of Volterra filters.

Volterra filters generalize Taylor expansion to functionals. In a finite dimensional setting, Volterra filters can even be precisely viewed as a Taylor expansion of a multivariate function. If two discrete time, jointly stationary, signals $x(t)$ and $y(t)$ are linked by a finite order Volterra filter, we can write

$$y(t) = h_0 + \sum_{1 \leq n \leq q} \sum_{k_1, \dots, k_n \leq M} h_n(k_1, \dots, k_n) x(t-k_1) \dots x(t-k_n)$$

The link between x and y is clearly nonlinear. However, the link is said to be linear in the parameters, since for a fixed input x, the weigthed sum of two finite order Volterra expansion is again a finite order Volterra expansion.Thus the space (indexed by x) of all signals that admit a Volterra expansion like above is a linear space, which can be given a Hilbert structure using an appropriate scalar product. To practically see this, it suffices to work in a a sufficiently high dimensional space, whose element will be 'vectors' containing as entries the monomials $x(t-k_1)\dots x(t-k_n)$ for all orders n, and all times k. Doing alike for the filter parameters h, $y(t)$ can be written as $\boldsymbol{h}^\top \boldsymbol{X}$, a linear relation!

Doing so we have embedded the input x, which for a memory m lies in a M dimensional space, into a larger linear space of very high dimension. Now, since the relation is linear in parameters, all the linear processing known can be performed. These linear processing will require evaluation of scalar product like $E[\boldsymbol{X}\boldsymbol{X}^\top]$, which can be very difficult to evaluate in practice, given the high dimensionality of the embedded vectors.

RKHS techniques elaborates on this idea of nonlinearly embedding data into a high (even infinite) dimensional linear space, but add the constraint of having a very efficient way of calculating the scalar products that will be needed for processing! Precisely, RKHS techniques will allow nonlinear processing without explicitly embed the data!

2 Kernels, RKHS

We will work extensively using the concept of Hilbert spaces. We just recall that a Hilbert space is a linear space endowed with a scalar product (bilinear, symmetric and definite positive form), complete with respect to the metric induced by the scalar product.

2.1 Kernels

We will consider an abstract space $\mathbb{X}$ and this space is embedded into a larger one $\mathcal{H}$, possibly of infinite dimension, of functions of $\mathbb{X}$ into $\mathbb{R}$. The embedding is supposed to be performed by a map that we denote as Φ. As discussed in the previous section, it will be easy to work with the embedded vectors if we can easily calculate an inner product in the feature space. This motivates the formal definition of a kernel :

Definition A function $k : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}$ is called a kernel on $\mathbb{X}$ if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathbb{X} \longrightarrow \mathcal{H}$ such that

$$\forall x, x' \in \mathbb{X}^2, \big\langle \Phi(x) \big| \Phi(x') \big\rangle_{\mathcal{H}} = k(x, x')$$

Let us present example of kernels.

Example Polynomial kernels.

Let $\mathbb{X} = \mathbb{R}^n$ equipped with the euclidean scalar product. Define $k(x,y) = (1+ < x|y >)^m$. To show that this bivariate function is a kernel on $\mathbb{R}^n$ we must exhibit the map Φ and the Hilbert space $\mathcal{H}$ such that $k(x,y) =< \Phi(x)|\Phi(y) >$. We do it first for the simpler kernel $k(x,y) =< x|y >^m$. We have

$$\begin{aligned}
< x|y >^m &= \Big(\sum_i x_i y_i\Big)^m \\
&= \sum_{i_1,\ldots,i_m} x_{i_1} y_{i_1} x_{i_2} y_{i_2} \ldots x_{i_m} y_{i_m} \\
&= \sum_{j_1+j_2+\ldots+j_n=m} \frac{m!}{j_1!\ldots j_n!} (x_1y_1)^{j_1} (x_2y_2)^{j_2} \ldots (x_n y_n)^{j_n} \\
&= \sum_{j_1+j_2+\ldots+j_n=m} \sqrt{\frac{m!}{j_1!\ldots j_n!}} x_1^{j_1} x_2^{j_2} \ldots x_n^{j_n} \times \sqrt{\frac{m!}{j_1!\ldots j_n!}} y_1^{j_1} y_2^{j_2} \ldots y_n^{j_n}
\end{aligned}$$

Let $\phi_j(x) = \sqrt{\frac{m!}{j_1!\ldots j_n!}} x_1^{j_1} x_2^{j_2} \ldots x_n^{j_n}$ where $\sum_{i=1}^n j_i = m$ and $j_i \in \{0,\ldots,m\}$.

Let Φ be a mapping from $\mathbb{X}$ to the Hilbert space of square integrable sequences of $\mathbb{N}^n$ defined by $\Phi(x) = (\phi_j(x))_{j\in\{0,\ldots,m\}^n}$ with $\sum_{i=1}^n j_i = m$. Then we have

$$< x|y >^m \quad = \quad < \Phi(x)|\Phi(y) >_{\ell^2(\mathbb{N}^n)}$$

Thus, $k(x,y) = (< x|y >)^m$ is a kernel.

An example using this map is provided in figure (1) where in the space $\mathbb{R}^2$ two sets of points are not linearly separable whereas they become linearly separable in the feature space associated with the map $\Phi(x,y) = (x^2, y^2, \sqrt{2}xy)$.

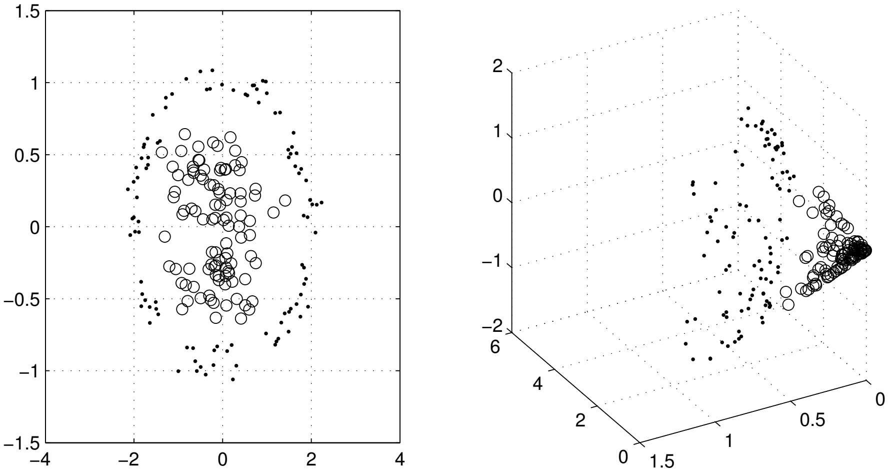

Figure 1: Illustration of linear separability in the feature space

If we consider $(1+ < x|y >)^m$ then we get the same result by considering all the monomials of all orders up to m. Thus, from a signal processing point of view, the Hilbert space $\mathcal{H}_x$ corresponds in that case to the Hilbert space of the Volterra filters of finite memory.

Note that the Hilbert spaces obtained using polynomial kernels are finite dimensional.

Standard operations on kernels can be used to define new kernels, or can be used to prove that some bilinear function is a kernel.

Proposition 2.1 *Building kernels.*

1. *Let $f_n : \mathbb{X} \longrightarrow \mathbb{R}$ be a series of functions such that $f_n(x) \in l_2(\mathbb{N}), \forall x \in \mathbb{X}$. Then $\sum_{n\geq 0} f_n(x)f_n(x')$ is a kernel on $\mathbb{X}$.*
2. *The sum of two kernels on $\mathbb{X}$ is a kernel on $\mathbb{X}$. The product between a positive number and a kernel on $\mathbb{X}$ is a kernel on $\mathbb{X}$.*
3. *If f is an arbitrary function on $\mathbb{X}$, $f(x)k(x,x')f(x')$ is a kernel on $\mathbb{X}$.*
4. *The product of two kernels on $\mathbb{X}_1$ and $\mathbb{X}_2$ is a kernel on $\mathbb{X}_1 \otimes \mathbb{X}_2$. In particular, the product of two kernels on a same space is a kernel on that space.*
5. *Let $a : \mathbb{X}_1 \longrightarrow \mathbb{X}_2$ be a map. If $k(x,y)$ is a kernel on $\mathbb{X}_2$, then $k(a(x), a(y))$ is a kernel on $\mathbb{X}_1$.*

Proof

1. Using Hlder inequality in the space of sequences $\ell_1(\mathbb{N})$ and $\ell_2(\mathbb{N})$ we have
$$\sum_n \left|f_n(x)f_n(x')\right| \leq \left|\left|f_n(x)\right|\right|_2 \left|\left|f_n(x')\right|\right|_2 < +\infty$$
and thus $k(x,x')$ is well defined since the sum defining it converges absolutely for all x, x'. If we set $\Phi(x) = (f_0(x), f_1(x), \ldots), \forall x \in \mathbb{X}$ and set $\mathcal{H} = \ell_2(\mathbb{N})$, then $k(x,x') =< \Phi(x)|\Phi(x') >_{\ell_2}$ and k is a kernel on $\mathbb{X}$.

2. Let $\alpha \geq 0$. We have $k(x,x') = k_1(x,x') + \alpha k_2(x,x') =< \Phi_1(x)|\Phi_1(x') >_{\mathcal{H}_1} + < \sqrt{\alpha}\Phi_2(x)|\sqrt{\alpha}\Phi_2(x') >_{\mathcal{H}_2}$. Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ with the scalar product $< (x,y)|(x',y') >_{\mathcal{H}} =< x|x' >_{\mathcal{H}_1} + < y|y' >_{\mathcal{H}_2}$. Then $\Phi(x) := (\Phi_1(x), \sqrt{\alpha}\Phi_2(x))$ from $\mathbb{X}$ to $\mathcal{H}$ satisfies $k(x,x') =< \Phi(x)|\Phi(x') >_{\mathcal{H}}$ and k is therefore a kernel on $\mathbb{X}$.

3. $f(x)k(x,x')f(x') =< f(x)\Phi(x)|f(x')\Phi(x') >$.

4. Let $\mathcal{H}_1 \otimes \mathcal{H}_2$ stands for the tensorial product of the Hilbert spaces.

   A formal definition is the following.

   There is one vector space $\mathcal{H}_1 \otimes \mathcal{H}_2$ and one bilinear application $\pi : \mathcal{H}_1 \times \mathcal{H}_2 \longrightarrow \mathcal{H}_1 \otimes \mathcal{H}_2$ such for any other other vector space and bilinear form $f : \mathcal{H}_1 \times \mathcal{H}_2 \longrightarrow F$, f is uniquely written as $\varphi \circ \pi$ where φ is linear from $\mathcal{H}_1 \otimes \mathcal{H}_2$ to F.

   A more constructive definition which provides more intuition (roughly, it is the linear space generated by the terms x_1x_2, $x_1 \in \mathcal{H}_1, x_2 \in \mathcal{H}_2$) is the following.

   For $x_1 \in \mathcal{H}_1, x_2 \in \mathcal{H}_2$, consider the bilinear form $x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ defined by $(x_1 \otimes x_2)(u_1, u_2) = \left\langle x_1|u_1\right\rangle_{\mathcal{H}_1} \left\langle x_2|u_2\right\rangle_{\mathcal{H}_2}$. The set of linear combination of such forms $\{\sum_i \alpha_i x_i \otimes y_i\}$ with the scalar product
$$\left\langle x_1 \otimes x_2 | y_1 \otimes y_2\right\rangle = \left\langle x_1|y_1\right\rangle_{\mathcal{H}_1} \left\langle x_2|y_2\right\rangle_{\mathcal{H}_2}$$
   is a preHilbert space, transformed into a complete Hilbert space by completion. The resulting space is the tensor product noted $\mathcal{H}_1 \otimes \mathcal{H}_2$.

   Coming back to the proposition, we have $k_1(x_1,x_1')k_2(x_2,x_2') =< \Phi_1(x_1)|\Phi_1(x_1') >_{\mathcal{H}_1} < \Phi_2(x_2)|\Phi_2(x_2') >_{\mathcal{H}_2} =< \Phi_1(x_1) \otimes \Phi_2(x_2)|\Phi_1(x_1') \otimes \Phi_2(x_2') >_{\mathcal{H}_1 \otimes \mathcal{H}_2}$. In particular the product of two kernels on a same set is a kernel on that set.

5. $k(a(x_1), a(y_1)) = k(x_2, y_2) =< \Phi_2(x_2)|\Phi_2(y_2) >=< \Phi_1(x_1)|\Phi_1(y_1) >$ where $\Phi_1 = \Phi_2 \circ a$ and the Hilbert space is the same. ∎

Example Gaussian kernel.

Consider $k(x, y) = \exp(-\|x - y\|^2/\eta)$ where $\eta > 0$ for x and y in some subspace of $\mathbb{R}^n$. The kernel can be written as

$$\begin{aligned}
k(x,y) &= \exp(-\frac{\|x\|^2}{\eta})\exp(-\frac{\|y\|^2}{\eta})\exp(2\frac{< x|y >}{\eta}) \\
&= \exp(-\frac{\|x\|^2}{\eta})\exp(-\frac{\|y\|^2}{\eta})\sum_{k\geq 0}\frac{2^k}{\eta^k k!} < x|y >^k \\
&= \exp(-\frac{\|x\|^2}{\eta})\exp(-\frac{\|y\|^2}{\eta})\sum_{k\geq 0}\frac{2^k}{\eta^k k!}\sum_{j_1+j_2+\ldots+j_n=k}\frac{k!}{j_1!\ldots j_n!}x_1^{j_1}x_2^{j_2}\ldots x_n^{j_n}y_1^{j_1}y_2^{j_2}\ldots y_n^{j_n} \\
&= \exp(-\frac{\|x\|^2}{\eta})\exp(-\frac{\|y\|^2}{\eta})\sum_{j_1,j_2,\ldots,j_n\geq 0}\frac{2^{j_1+\ldots+j_n}}{\eta^{j_1+\ldots+j_n}j_1!\ldots j_n!}x_1^{j_1}x_2^{j_2}\ldots x_n^{j_n}y_1^{j_1}y_2^{j_2}\ldots y_n^{j_n}
\end{aligned}$$

Let $\phi_j(x) = \sqrt{\frac{2^{j_1+\cdots+j_n}}{\eta^{j_1+\cdots+j_n}j_1!\ldots j_n!}}x_1^{j_1}x_2^{j_2}\ldots x_n^{j_n}$ where $j_i \geq 0$. Then $\Phi(x) = \exp(-\frac{\|x\|^2}{\eta})(\phi_1(x), \phi_2(x), \ldots)$ satisfies $k(x, y) =< \Phi(x)|\Phi(y) >$.

Note that the feature map is not unique. Indeed, it can be verified that

$$\Phi_\eta(x) = \frac{2^{\frac{n}{2}}}{\pi^{\frac{n}{4}}\eta^{\frac{n}{4}}}\exp(-\frac{2}{\eta}\|x - .\|^2)$$

is a feature map for the kernel, that is $\langle \Phi_\eta(x) \big| \Phi_\eta(y) \rangle = \exp(-\|x - y\|^2/\eta)$.

A fondamental property of kernel lies in their positive-definite character.

Definition A function $k : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}$ is positive definite if for all $n \in \mathbb{N}$, all $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ and all $(x_1, \ldots, x_n) \in \mathbb{X}^n$ we have $\sum_{i,j} \alpha_i\alpha_j k(x_i, x_j) \geq 0$. If there is equality for mutually distinct x_i only for $\alpha_1 = 0 = \alpha_2 = \ldots = \alpha_n$ then k is strictly definite positive.

In other words, k is positive definite if and only if all the gram matrices $K_{i,j} = k(x_i, x_j)$ are positive definite. Now we can state the link between kernels and positive-definite functions.

Theorem 2.2 *Kernel are symmetric positive-definite functions.*

A function k is a kernel on $\mathbb{X}$ if and only if it is a symmetric positive definite function.

Proof Let k be a kernel on $\mathbb{X}$. Then there is an Hilbert space $\mathcal{H}$ and a map $\Phi : \mathbb{X} \longrightarrow \mathcal{H}$ such that $k(x, y) =< \Phi(x)|\Phi(y) >_{\mathcal{H}}$. Consider an arbitrary $n \in \mathbb{N}$ and two arbitrary n-uplets α_i and x_i.

Then

$$\begin{aligned}\sum_{i,j}\alpha_i\alpha_j k(x_i,x_j) &= \sum_{i,j}\alpha_i\alpha_j < \Phi(x_i)|\Phi(x_j) >_{\mathcal{H}} \\ &= < \sum_i \alpha_i\Phi(x_i)|\sum_j \alpha_j\Phi(x_j) >_{\mathcal{H}} \\ &= \|\sum_i \alpha_i\Phi(x_i)\|^2_{\mathcal{H}} \geq 0\end{aligned}$$

Furthermore, as a scalar product in $\mathcal{H}$, k is symmetric.

For the converse, let $k(x,y)$ be a symmetric positive definite function on $\mathbb{X}\times\mathbb{X}$. Consider

$$\mathcal{H} = \Big\{ f : \mathbb{X} \longrightarrow \mathbb{R}/f(.) = \sum_{i=1}^{n}\alpha_i k(.,x_i), n\in\mathbb{N}, (\alpha_1,\dots,\alpha_n)\in\mathbb{R}^n, (x_1,\dots,x_n)\in\mathbb{X}^n \Big\}$$

For two functions $f(.) = \sum_{i=1}^n \alpha_i k(.,x_i)$ and $g(.) = \sum_{i=1}^n \beta_i k(.,y_i)$ of $\mathcal{H}$, define

$$< f|g >_{\mathcal{H}} = \sum_{i,j}\alpha_i\beta_j k(x_i,y_j)$$

We want to show that this function is a scalar product.

Note that using the symmetry of k, $< f|g >_{\mathcal{H}} = \sum_i \alpha_i g(x_i)$ which show that the scalar product does not depend on the representation of g. Likewise,$< f|g >_{\mathcal{H}} = \sum_j \beta_j f(y_j)$ does not depend on the particular representation of f. Furthermore, symmetry and bilinearity of $< f|g >_{\mathcal{H}}$ is evident. $< f|f >_H \geq 0$ since k is positive definite. It remains to show that $< f|f >_{\mathcal{H}} = 0$ implies $f = 0$. This uses the following lemma.

Lemma 2.3 *A positive (not necessarily definite) symmetric bilinear function satisfies the Cauchy-Schwartz inequality*

$$|k(x,y)|^2 \leq k(x,x)k(y,y)$$

The proof is immediate by noting that for x_1 and x_2 the matrix $K_{i,j} = k(x_i,x_j)$ is positive (definite) so that its determinant is positive. Note that the definite character is not necessary.

Coming back to the proof, remark that $f(x) = \sum_i \alpha_i k(x,x_i) =< f|k(.,x) >_{\mathcal{H}}$ since as seen above $< f|g >_{\mathcal{H}} = \sum_i \alpha_i g(x_i)$. Note that since it is positive $< | >_{\mathcal{H}}$ satisfies the Cauchy-Schwartz inequality (see the lemma) Thus,

$$|f(x)|^2 = | < f|k(.,x) >_{\mathcal{H}} |^2 \leq < k(.,x)|k(.,x) >_{\mathcal{H}} < f|f >_{\mathcal{H}}$$

Therefore, $< f|f >_{\mathcal{H}} = 0$ implies $f = 0$ and $< | >_{\mathcal{H}}$ is definite and defines a scalar product.

Finally, we have seen that $f(x) = \sum_i \alpha_i k(x,x_i) =< f|k(.,x) >_{\mathcal{H}}$ for any function in $\mathcal{H}$. In particular, $f(x) = k(y,x)$ satisfies $k(y,x) = f(x) =< k(y,.)|k(.,x) >_{\mathcal{H}} =< k(.,y)|k(.,x) >_{\mathcal{H}}$. Thus identifying $\Phi(x) = k(.,x)$ gives the feature map and finishes the proof that k is a kernel on $\mathbb{X}$. ■

Note that we have just used preHilbert spaces. Thus we should add in the proof the completion step in order to make the associated metric space complete.

We have noted in the proof above the strange property for function in H that $f(x) =< f|k(.,x) >_{\mathcal{H}}$. Usually this relation holds for the Dirac δ function. However, as proved above, it also holds in certain spaces, an example of which was constructed in the proof, which are called Reproducing Kernel Hilbert Spaces. The Reproducing term comes from the fact that the evaluation equation $f(x) =< f|k(.,x) >_{\mathcal{H}}$ is also true for the kernel and reads in that case $k(x,y) =< k(.,y)|k(.,x) >_{\mathcal{H}}$, a property know as the reproducing property.

2.2 Reproducing Kernel Hilbert Spaces

We begin with a formal definition of a Reproducing Kernel Hilbert Space, abbreviated RKHS in the sequel. The formal definition will highlight the importance of the Riesz representation theorem that we will recall, as it will be also useful later, when introducing covariance operators.

Definition Let $\mathbb{X}$ a set and $\mathcal{H}$ a Hilbert space of functions of $\mathbb{X}$ into $\mathbb{R}$.

1. $k : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ if $k(.,x) \in \mathcal{H}$ and for all $f \in \mathcal{H}$ and all $x \in \mathbb{X}$ the reproducing property $f(x) = \big\langle f \big| k(.,x) \big\rangle$ holds.
2. $\mathcal{H}$ is called a reproducing kernel Hilbert space if the evaluation functional

$$\begin{array}{rcl} \delta_x : \mathcal{H} & \longrightarrow & \mathbb{R} \\ f & \longmapsto & \delta_x(f) = f(x) \end{array}$$

is continuous.

The fact that the evaluation functional is continuous implies that the RKHS has a unique reproducing kernel. This comes from the Riesz representation for bounded linear functional. Recall that a linear functional $a : \mathcal{H} \to \mathbb{R}$ is bounded if $\|a\| = \sup_{f/\|f\|=1} |a(f)| < +\infty$. Recall also that a linear functional is bounded if and only if it is continuous. The Riesz representation theorem states that a bounded linear function can be represented as a scalar product. Precisely,

Theorem 2.4 *Let ℓ be a bounded linear functional from $\mathcal{H}$ to $\mathbb{R}$. Then there is a unique element g of $\mathcal{H}$ such that for all $f \in \mathcal{H}$, $\ell(f) = \big\langle f \big| g \big\rangle$*

Proof If $\ell = 0$, $g = 0$ satisfies the requirement. Let $\ell \neq 0$. Let $\tilde{g} \in (\ker \ell)^{\perp}$ such that $\|\tilde{g}\| = 1$. Then, $\tilde{g}\ell(f) - f\ell(\tilde{g}) \in \ker\ell$ since $\ell(\tilde{g}\ell(f) - f\ell(\tilde{g})) = \ell(\tilde{g})\ell(f) - \ell(f)\ell(\tilde{g}) = 0$. Thus writing this orthogonality using the scalar product we get $\big\langle \tilde{g}\ell(f) - f\ell(\tilde{g}) \big| \tilde{g} \big\rangle = 0$ or, solving for $\ell(f)$, $\ell(f) = \big\langle f \big| \ell(\tilde{g})\tilde{g} \big\rangle$. Then choosing $g = \tilde{g}\ell(\tilde{g})$ proves the existence. If two elements are solutions, then this implies $< (g_1 - g_2)|f >= 0$ for all $f \in \mathcal{H}$, and thus $g_1 - g_2 \in \mathcal{H}^{\perp} = \{0\}$. ■

Boundedness does not seem to be used. In fact it is. Its equivalence to continuity implies that the kernel of ℓ is a closed subspace ($\lim \ell(u_n) = \ell(\lim u_n) = 0$), and this ensures that there is an element in its orthogonal complement by application of the projection theorem (if a subspace is closed, any element of the space is uniquely written as the sum of an element of this subspace plus an element in its orthogonal complement).

Coming back to RKHS, since the evaluation functional is a linear continuous functional, it is bounded, and therefore the Riesz representation theorem state that there is a unique element of $\mathcal{H}$ such that $\delta_x(f) = \langle f | g_x \rangle$. Obviously here, the unique element is indexed by x. This equation is precisely the reproducing property, and therefore, g_x is a reproducing kernel. We denote it as $k(.,x)$. This shows that any reproducing kernel Hilbert space has a unique reproducing kernel. The last thing to show to close the loop is that a reproducing kernel is a kernel (in the sense of the previous section). It suffices to show the existence of a map $\Phi(x)$ from $\mathbb{X}$ to $\mathcal{H}$ such that $k(x,y) = \langle \Phi(x) | \Phi(y) \rangle$. But this is evident since using the reproducing property of the kernel we have $k(x,y) = \langle k(.,x) | k(.,y) \rangle$, such that choosing $\Phi(x) = k(.,x)$ solves the problem.

Finally, the completion of the space introduced in the previous section

$$\mathcal{H} = \Big\{ f : \mathbb{X} \longrightarrow \mathbb{R} / f(.) = \sum_{i=1}^{n} \alpha_i k(.,x_i), n \in \mathbb{N}, (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n, (x_1, \ldots, x_n) \in \mathbb{X}^n \Big\}$$

is the RKHS associated with the reproducing kernel k.

3 Application in ML

The aim of this section is to present the representer theorem which basically states that when using a RKHS of functions to optimize some empirical risk functions, even if of infinite dimension, it suffices to search for the solution in a finite dimensional subspace of the RKHS. We will not discuss the notion of empirical risk and of regularization. We refer to books on statistical learning theory for that.

We suppose that we have a learning set $\mathcal{L} = (x_i, y_i)$ of N examples where typically $x_i \in \mathbb{X}$ and $y_i \in \mathbb{Y}$, and we want to find a function f such that $\hat{y}_i = f(x_i)$ approximate correctly the observation y_i. If $\mathbb{Y}$ is a continuum we typically face a problem of regression whereas if $\mathbb{Y}$ is a finite set we deal with a classification problem. Obviously, f will be chosen by optimizing some criteria that depends on the application considered. We will use a cost function $c : \mathbb{X} \times \mathbb{Y} \times \mathbb{Y} \longrightarrow \mathbb{R}^+$ which assign the cost $c(x, y, f(x))$ to the particular example x, y when using function f. The average cost cannot be optimized in practice, so that the empirical risk is used and writes

$$R(f) = \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_i, f(x_i))$$

Furthermore, for problems of overfitting, the set over which f is search for has to be constrained. A usual approach is to add a penalty term to the empirical risk. This may for example limit the norm of the function f. We assume that if f is searched for in an Hilbert space, the penalty added to regularize is a strictly increasing function of the norm. Thus the risk reads $R(f)+\Omega(\|f\|_{\mathcal{H}})$ and has to be minimized over $\mathcal{H}$.

We can now state and prove the representer theorem.

Theorem 3.1 *Let $c : \mathbb{X}\times\mathbb{R}\times\mathbb{R}\longrightarrow\mathbb{R}^+$ a cost function and $\Omega : \mathbb{R}^+\longrightarrow\mathbb{R}$ a stricly increasing function. Let N couples $x_i, y_i \in \mathbb{X}\times\mathbb{R}$ be observed, and let $\mathcal{H}$ be a RKHS with reproducing kernel k on $\mathbb{X}$. Then, a solution of*

$$f = \arg\min_{f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} c(x_i, y_i, f(x_i)) + \Omega(\|f\|_{\mathcal{H}})$$

satisfies

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

Thus, even if we look for a function in an infinite dimensional space, it suffices to look for it in the space spanned by the features $k(., x_i)$ associated to the observed data.

Proof Let $f\in\mathcal{H}$. Then f can be decomposed into the sum of its part belonging to the subspace of $\mathcal{H}$ generated by the $k(., x_i)$, $i = 1, \ldots, N$ and its part belonging to the orthogonal complement. Thus

$$\begin{aligned} f(x) &= f_{\|}(x) + f_{\perp}(x) \\ &= \sum_{i=1}^{N}\alpha_i k(x, x_i) + f_{\perp}(x) \end{aligned}$$

To evaluate the empirical risk we need to evaluate $f(x_j), j1, \ldots, N$. We have using the reproducing property

$$\begin{aligned} f(x_j) &= \big\langle f \big| k(., x_j)\big\rangle \\ &= \big\langle f_{\|} + f_{\perp} \big| k(., x_j)\big\rangle \\ &= \big\langle f_{\|} \big| k(., x_j)\big\rangle \\ &= \big\langle \sum_{i=1}^{N}\alpha_i k(, x_i) \big| k(., x_j)\big\rangle \\ &= \sum_{i=1}^{N}\alpha_i \big\langle k(, x_i) \big| k(., x_j)\big\rangle \\ &= \sum_{i=1}^{N}\alpha_i k(x_i, x_j) \end{aligned}$$

from which we see that $f(x_j)$ does not depend on $f_{\perp}$. Then the empirical risk does not depend on $f_{\perp}$. Furthermore, since Ω is stricly increasing we have

$$\begin{aligned} \Omega(\|f\|_{\mathcal{H}}) &= \Omega\big(\sqrt{\|f_{\|}\|^2_{\mathcal{H}} + \|f_{\perp}\|^2_{\mathcal{H}}}\big) \\ &\geq \Omega(\|f_{\|}\|_{\mathcal{H}}) \end{aligned}$$

and the minimum will be obtained when $f_\perp = 0$. Since choosing $f_\perp = 0$ does not affect the empirical risk but strictly decreases the penalty term, any minimizer must have $f_\perp = 0$. ■

4 Regression

We present here the problem of regression in a simple setting. We will use the quadratic loss function and a power limitation on the function chosen. Thus, given a learning set of N examples (x_i, y_i) we want to find a regression function f in order to minimize

$$R(f) = \sum_i (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}}$$

where $\mathcal{H}$ is some RKHS of functions from $\mathbb{X}$ to $\mathbb{R}$, whith reproducing kernel k. Application of the representer theorem leads to the conclusion that a function minimizing the empirical risk will be written as

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

Thus if we introduce the vector $\boldsymbol{k}_x = (k(x, x_1), \ldots, k(x, x_N))^\top$ and the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^\top$, we will have $f(x) = \boldsymbol{k}_x^\top \boldsymbol{\alpha}$. Furthermore, the vector $\boldsymbol{f}$ containing the $f(x_i)$ can be written as

$$\boldsymbol{f} = \begin{pmatrix} \boldsymbol{k}_{x_1}^\top \\ \vdots \\ \boldsymbol{k}_{x_N}^\top \end{pmatrix} \boldsymbol{\alpha} = K\boldsymbol{\alpha} \text{ where } K \quad = \quad \begin{pmatrix} \boldsymbol{k}_{x_1}^\top \\ \vdots \\ \boldsymbol{k}_{x_N}^\top \end{pmatrix}$$

Note that due to symmetry, the matrix K is self adjoint, and furthermore $K_{ij} = k(x_i, x_j) = k(x_j, x_i)$. This matrix is called the Gram matrix.

From these notations, we get $\sum_i (y_i - f(x_i)) = (\boldsymbol{y} - K\boldsymbol{\alpha})^\top (\boldsymbol{y} - K\boldsymbol{\alpha})$. Furthemore, the square norm of f is evaluated as

$$\|f\|^2_{\mathcal{H}} = \langle f | f \rangle = \sum_{i,j} \alpha_i \alpha_j \langle k(., x_i) | k(., x_j) \rangle = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

Therefore solving $f = \arg\min R(f)$ is equivalent to solving

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha}} (\boldsymbol{y} - K\boldsymbol{\alpha})^\top (\boldsymbol{y} - K\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

Setting the gradient of the objective to zero we get $-2K\boldsymbol{y} + 2(K^2 + \lambda K)\boldsymbol{\alpha} = 0$ so that we obtain $\boldsymbol{\alpha}^\star = (K + \lambda I)^{-1} \boldsymbol{y}$. If we want to predict the value y at a test point x, we will then propose $y = {\boldsymbol{k}_x}^\top (K + \lambda I)^{-1} \boldsymbol{y}$.

5 Classification: SVM and their kernel counterparts

Suppose we are given N training samples $(x_i, y_i) \in \mathbb{X} \times \{-1, 1\}$. y_i is called the class to which the corresponding element x_i belongs. The aim of supervised classification is to learn a classification rule from the examples provided. The basic idea is to find a partition of the space $\mathbb{X}$ into two subsets such that each of the subset is a class. In linear classification, the set of observation is cut into two pieces by an hyperplane. We thus suppose that $\mathbb{X}$ is equipped with a scalar product, and we can define an hyperplane w as the set of points such that $\langle \boldsymbol{x} | \boldsymbol{w} \rangle + b = 0$ where b is a scalar. The hyperplane is the set of point that project orthogonally on the same point on the line defined by $\boldsymbol{w}$. $\boldsymbol{w}$ is orthogonal to the hyperplane. The set of training examples is said to be linearly separable is the classes can be separated by an hyperplane. In that case the decision function

$$\begin{aligned} f \quad : \quad & \mathbb{X} \longrightarrow \{-1,+1\} \\ & \boldsymbol{x} \longmapsto f_{w,b}(\boldsymbol{x}) = \text{Sign}(\langle \boldsymbol{x} | \boldsymbol{w} \rangle + b) \end{aligned}$$

will give the good decision for the training examples. Note however that multiplying $\boldsymbol{w}$ and b by the same scalar will give the same hyperplane. Thus to eliminate this degree of freedom, something has to be fixed. A way is to require that the points closest to the hyperplane give exactly $\pm 1 = \langle \boldsymbol{x} | \boldsymbol{w} \rangle + b$.

Choose two such points, each one being on each side of the hyperplane. Then we have

$$\langle \boldsymbol{x}_1 | \boldsymbol{w} \rangle + b = 1 = -(\langle \boldsymbol{x}_2 | \boldsymbol{w} \rangle + b)$$

and thus $\langle \boldsymbol{x}_1 - \boldsymbol{x}_2 | \boldsymbol{w} \rangle = 2$ meaning that the distance between those points and the hyperplane is $1/\|\boldsymbol{w}\|^2$. Since this is the distance between the closest points to the separating hyperplane, this quantity is called the margin.

Thus the ideas of linear classification is to find an hyperplane that correctly classifies the training examples and that is chosen in order to maximize the margin. A training example is correctly classified if $y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle) \geq 1$. Thus the classification approach can be stated as

$$\begin{aligned} & \max_{\boldsymbol{w} \in \mathbb{X}, b \in \mathbb{R}} \frac{1}{\|\boldsymbol{w}\|^2} \\ & \text{subject to } y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) \geq 1, \forall i = 1, \ldots, N \end{aligned}$$

or equivalently

$$\begin{aligned} & \min_{\boldsymbol{w} \in \mathbb{X}, b \in \mathbb{R}} \frac{1}{2} \|\boldsymbol{w}\|^2 \\ & \text{subject to } y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) \geq 1, \forall i = 1, \ldots, N \end{aligned}$$

To solve this problem, consider the Lagrangian

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + \sum_i \alpha_i \big(1 - y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) \big)$$

where the N dual variables α_i are positive. Recall that the KKT conditions for optimality are given by

$$\begin{aligned} \nabla_{\boldsymbol{w}} L &= 0 \\ \frac{\partial L}{\partial b} &= 0 \\ y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) &\geq 1, \forall i = 1, \ldots, N \\ \alpha_i &\geq 0, \forall i = 1, \ldots, N \\ \alpha_i \big(1 - y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) \big) &= 0, \forall i = 1, \ldots, N \end{aligned}$$

The two first lines leads to

$$\begin{aligned} \boldsymbol{w} &= \sum_i \alpha_i y_i \boldsymbol{x}_i \\ \sum_i \alpha_i y_i &= 0 \end{aligned}$$

Furthermore, from the last KKT conditions, for all $i = 1, \ldots, N$ we have either $\alpha_i = 0$ and $y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) \geq 1$, or $\alpha_i > 0$ and $y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) = 1$. The points $\boldsymbol{x}_i$ for which $\alpha_i > 0$ and $y_i(\langle \boldsymbol{x}_i | \boldsymbol{w} \rangle + b) = 1$ lies precisely on the plane defined by the margins and are the only one participating in $\boldsymbol{w}$ since the other are characterized by $\boldsymbol{\alpha}_i = 0$. The points for which $\alpha_i > 0$ are called the support vectors of the classifier.

The dual problem is obtained by injecting in the original Lagrangian the optimality conditions in $\boldsymbol{w}$ and b and maximizing the result (this come from the saddle-point sufficient condition on the Lagrangian for optimality).

Thus inserting $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$ and $\sum_i \alpha_i y_i = 0$ into L we obtain

$$L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i | \boldsymbol{x}_j \rangle$$

which has to be maximized subjected to the constraints $\alpha_i \geq 0, \forall i = 1, \ldots, N$ and $\sum_i \alpha_i y_i = 0$.

Now the decision function writes

$$\begin{aligned} f_{w,b}(\boldsymbol{x}) &= \text{Sign}(\langle \boldsymbol{x} | \boldsymbol{w} \rangle + b) \\ &= \text{Sign}(\sum_i \alpha_i y_i \langle \boldsymbol{x} | \boldsymbol{x}_i \rangle + b) \end{aligned}$$

and it only depends on the scalar product between the point to test and the data.

The dual formulation is very important if we want to generalize the approach using kernel. Indeed everything which has been written up to this point would be valid if we replace $\boldsymbol{x}_i$ by a nonlinear transform $\Phi(\boldsymbol{x}_i)$ of it, it is to say if we embed the data in another space, where hopefully linear separability would be attained. But this is precisely the setting we described to motivate RKHS and the use of kernels to evaluate scalar products between nonlinearly transformed data points. Therefore if we want to use the support vector classifier in a nonlinear setting, it will be easy if we use a description wich only uses scalar product. This is precisely obtained in the dual formulation. Thus, the nonlinear classifier decision function with kernel k will be

$$f_{w,b}(\boldsymbol{x}) = \text{Sign}(\sum_i \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b)$$

where the α solve

$$\begin{gathered} \max_{\boldsymbol{\alpha}} \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \forall i = 1, \ldots, N \end{gathered}$$

Furthermore, b is found from the KKT conditions $\alpha_i\big(1-y_i(\langle \boldsymbol{x}_i|\boldsymbol{w}\rangle+b)\big)=0$, since for the support vectors ($\alpha_i>0$)

$$\begin{aligned} y_i(\langle \boldsymbol{x}_i|\boldsymbol{w}\rangle+b) &= 1 \\ \Longleftrightarrow \sum_j \alpha_j y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)+b &= y_i \end{aligned}$$

Thus we can average $y_i-\sum_j \alpha_j y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ over the support vectors x_i to estimate b.

To illustrate, we generate 100 samples from a linear separable case and from the example of the first section. Data points from the first class are obtained as i.i.d. samples from a two dimensional Gaussian random variable with variance $\sigma^2 I=0.04I$. Their label is -1. The second class labelled by 1 is generated by equispacing the N points on the unit circle and addin g to each a two dimensional random Gaussian perturbation of variance $\sigma^2 I=0.04I$. To train the classifier, we select 50 data points at random. The results are plotted in figure 2. Circles are +1 labelled points. The big dots represents the support vectors obtained by the algorithm. For the kernel example on the right we chose the Gaussian radial kernel with variance 0.1. We only represent the result on the training data set, and therefore no study in generalization is provided here. The black line is an estimation of the separation "hyperplane".

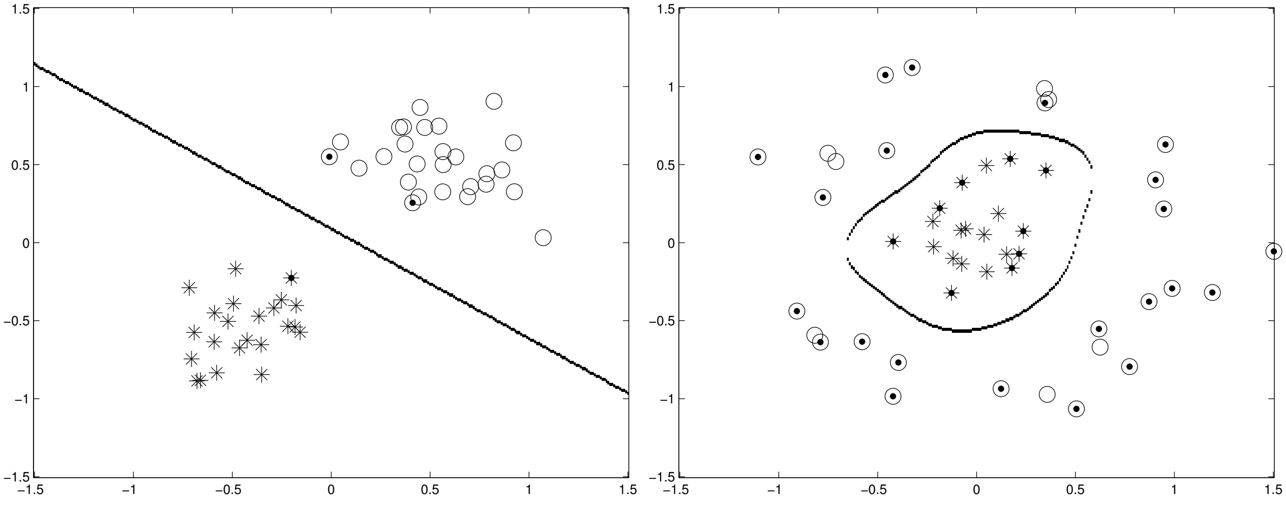

Figure 2: Illustration of linear separability in the feature space for SVM.

Note that generalization of the SVM to take into account ouliers or non separable case have been developed and should be used in place of the very elementary presentation made here.

6 Random variables and RKHS

We have seen that the usefulness of kernel methods is to deal with linear methods in spaces where original data are embedded in order to be more efficiently represented. What happened if the data are modeled as random variables? In this case, the image $k(.,X)$ of a random variable is a random function in the RKHS. Furthermore, we have seen that RKHS may be infinite dimensional, and then, studying the transformed data requires some care, and needs to deal with random variable taking values in Hilbert spaces.

Thus we will consider random variables taking values in some space $\mathbb{X}$ which is supposed to be equipped with some σ-algebra $\mathcal{B}$. Typically $\mathbb{X}$ will be $\mathbb{R}^n$ with the Borel σ-algebra. We consider a kernel k (symmetric, positive-definite function) measurable on $(\mathbb{X},\mathcal{B})$. We consider a random variable X on a probability space $(\Omega,\mathcal{F},P)$ that takes values in $\mathbb{X}$ and whose probability measure is denoted as P_X, where as usual, $P_X(B)=P(X^{-1}(B)),\forall B\in\mathcal{B}$.

The aim here is to describe the random variable when embedded in the RKHS $\mathcal{H}_x$ associated with k.

Prior studying the case of the embeding in a RKHS, let us consider the case of random variables which take values in a Hilbert space. A lemma states that a function from $(\Omega,\mathcal{F},P)$ to $\mathcal{H}$ is a random variable with values in $\mathcal{H}$ if and only if $x^*(X)$ is a real random variable for any $x^*\in\mathcal{H}^*$. Since the dual of a Hilbert space can be identified to itself, the linear form simply writes $x^*(X)=\langle x\big|X\rangle$ where $x\in\mathcal{H}$. Now, the linear form on $\mathcal{H}$ defined by $\ell_X(x)=E\langle x\big|X\rangle$ is bounded whenever $E\|X\|<+\infty$, and thus there exist a unique element m_X of $\mathcal{H}$ such that $E\langle x\big|X\rangle=\langle x\big|m_X\rangle$. m_X is the mean element and is denoted as $E[X]$. Now, the space $L^2_{\mathcal{H}}(P)$ of square integrable elements of $\mathcal{H}$, *i.e.* $E\|X\|^2<+\infty$ equipped with $\langle X\big|Y\rangle_{L^2}:=E\langle X\big|Y\rangle_{\mathcal{H}}$ is a Hilbert space.

The covariance operator is a linear operator from $\mathcal{H}$ to $\mathcal{H}$ defined by $\Sigma_X : x\longmapsto \Sigma_X(x) := E\big[\langle x\big|X-m_X\rangle(X-m_X)\big]$. It is bounded whenever $X\in L^2_{\mathcal{H}}(P)$. Likewise, we can define a cross-covariance operator between two elements X,Y of $L^2_{\mathcal{H}}(P)$ by the bounded linear operator from $\mathcal{H}$ to itself defined by $\Sigma_{XY}(x):=E[\langle x\big|X\rangle Y]$. The ajdoint operator defined by $\langle\Sigma^*_{XY}(y)\big|x\rangle$ is then Σ_{YX} since by definition $\Sigma_{YX}(y)=E[\langle y\big|Y\rangle X]$. Note that the two operators are completely defined by $\langle y\big|\Sigma_{XY}(x)\rangle=E\langle x\big|X\rangle\langle y\big|Y\rangle$. To conclude this rapid presentation, what happens if the space is $\mathbb{R}^n$. Then we work with usual vectors and the usual euclidean inner product to write that $\langle y\big|\Sigma_{XY}(x)\rangle=y^\top\Gamma_{YX}x=x^\top\Gamma_{XY}y$ where $\Gamma_{..}$ is the covariance matrix! Likewise, the mean element is the mean vector!

We now go back to the particular problem of reproducing kernel Hilbert spaces. The mean element m_X of X in $\mathcal{H}$ is the function defined by $m_X=E[k(.,X)]$. To exist, the kernel should be integrable. This requires

$$\begin{aligned} E[\|k(.,X)\|] &= E\Big[\big|\langle k(.,X)\big|k(.,X)\rangle\big|^{1/2}\Big] \\ &= E\Big[k(X,X)^{1/2}\Big]<+\infty \end{aligned}$$

a condition that we assume. We will in fact require for all the kernel we use that they are square integrable, meaning that $E[\|k(.,X)\|^2]=E[k(X,X)]<+\infty$. Then, note that by Jensen inequality for concave functions $E[f(X)]\leq f(E[X])$ that this implies the kernel is integrable.

Now, let $f \in \mathcal{H}_x$. Then the mean value of f at X

$$\begin{aligned} E[f(X)] &= E[\langle f | k(.,X)\rangle] \\ &= \langle f | E[k(.,X)]\rangle \\ &= \langle f | m_X \rangle \end{aligned}$$

We obtain thus the important result that knowing the mean element allows to evaluate the expectation of any function in $\mathcal{H}_x$. Furthermore, the function m_X can be found explicitly as

$$\begin{aligned} m_X(u) &= E[k(u,X)] \\ &= \int k(u,x) dP_X(x) \end{aligned}$$

Definition Characteristic kernel A kernel k is characteristic if the map $M_k : \mathcal{P} \to \mathcal{H}_k$, $P \mapsto m_P$ is injective, or equivalently if $\langle f | m_Q \rangle = \langle f | m_P \rangle \forall f \in \mathcal{H}_k \Longrightarrow P = Q$

Here, we have denoted as m_P the mean element in the RKHS of a random variable distributed under the probability $P \in \mathcal{P}$, the set of probability measures on the underlying space. $\mathcal{H}_k$ stands for the RKHS generated by k. The mean element of a characteristic kernel is thus a generalisation of the notion of characteristic functions.

We can now turn to the covariance operator definition. Let two random variables X and Y on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. The pair is assumed measurable on the product space as well. We call P_{XY} their joint probability measure. We embed them into RKHS $\mathcal{H}_x$ and $\mathcal{H}_y$ using kernels k_x and k_y. We would like to characterize the covariance between elements of the two RKHS, it is to say studying Cov $[f(X), g(Y)]$ where f and g are elements of $\mathcal{H}_x$ and $\mathcal{H}_y$ respectively. Since g belongs to $\mathcal{H}_y$ we can write

$$\begin{aligned} \text{Cov } [f(X), g(Y)] &= \text{Cov } \left[f(X), \langle g | k_y(.,Y)\rangle_{\mathcal{H}_y} \right] \\ &= \langle g | \text{Cov } [f(X), k_y(.,Y)]\rangle_{\mathcal{H}_y} \end{aligned}$$

Thus we see that the covariance may be expressed as a linear functional in $\mathcal{H}_y$. From a generalization to the Riesz representation theorem for operators between different spaces, there exists a unique operator $\Sigma_{YX} : \mathcal{H}_x \longrightarrow \mathcal{H}_y$ defined by $f \mapsto \Sigma_{YX} f = \text{Cov } [f(X), k_y(.,Y)]$ and called the cross-covariance operator. Its explicit expression is then

$$(\Sigma_{YX} f)(u) = \int \big(f(x) - E[f(X)]\big)\big(k_y(u,y) - m_Y(u)\big) dP_{XY}(x,y)$$

Obviously, this definition includes the definition of the covariance operator as the unique operator $\Sigma_{XX} : \mathcal{H}_x \longrightarrow \mathcal{H}_x$ defined by $f \mapsto \Sigma_{XX} f = \text{Cov } [f(X), k_x(.,X)]$ and called the cross-covariance operator. Its explicit expression is then

$$(\Sigma_{XX} f)(u) = \int \big(f(x) - E[f(X)]\big)\big(k_x(u,x) - m_X(u)\big) dP_X(x)$$

The key in defining these operators is the linear functional approach and the application of the Riesz representation theorem. This theorem shows unicity for bounded linear operators. Linearity is obvious when considering the covariance as a functional on $\mathcal{H}_y$. Boundedness is verified by the following set of inequalities, applying Schwartz inequality and Jensen inequality

$$\begin{aligned}
\big|\text{Cov}\,[f(X),g(Y)]\big| &\leq \big|E\Big[\big\langle f\big|k_x(.,X)\big\rangle_{\mathcal{H}_x}\big\langle g\big|k_y(.,Y)\big\rangle_{\mathcal{H}_y}\Big]\big| + \big|E\Big[\big\langle f\big|k_x(.,X)\big\rangle_{\mathcal{H}_x}\Big]\big|\big|E\Big[\big\langle g\big|k_y(.,Y)\big\rangle_{\mathcal{H}_y}\Big] \\
&\leq E[\|f\|_{\mathcal{H}_x}\|g\|_{\mathcal{H}_y}k_x(X,X)^{1/2}k_y(Y,Y)^{1/2}] + E[\|f\|_{\mathcal{H}_x}k_x(X,X)^{1/2}]E[\|g\|_{\mathcal{H}_y}k_y(Y,Y)^{1/2}] \\
&= \Big(E[k_x(X,X)^{1/2}k_y(Y,Y)^{1/2}] + E[k_x(X,X)^{1/2}]E[k_y(Y,Y)^{1/2}]\Big)\|f\|_{\mathcal{H}_x}\|g\|_{\mathcal{H}_y} \\
&\leq \Big(E[k_x(X,X)]^{1/2}E[k_y(Y,Y)]^{1/2} + E[k_x(X,X)^{1/2}]E[k_y(Y,Y)^{1/2}]\Big)\|f\|_{\mathcal{H}_x}\|g\|_{\mathcal{H}_y} \\
&\leq 2E[k_x(X,X)]^{1/2}E[k_y(Y,Y)]^{1/2}\|f\|_{\mathcal{H}_x}\|g\|_{\mathcal{H}_y}
\end{aligned}$$

The Adjoint operator, defined as $\big\langle \Sigma_{YX}^* g\big|f\big\rangle_{\mathcal{H}_x} = \big\langle g\big|\Sigma_{YX}f\big\rangle_{\mathcal{H}_y}$ is obviously Σ_{XY}.

Another interpretation of the covariance. As a mean element, the covariance operator can be interpreted as the mean of a kernel in an appropriate RKHS. It suffices to consider the tensor product of the kernels and its associated RKHS. Recall that given two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, their tensor product is roughly the set of linear combinations of cross-products of $\mathcal{H}_1$ and $\mathcal{H}_2$. The formal definition is the following. There is one vector space $\mathcal{H}_1 \otimes \mathcal{H}_2$ and one bilinear application $\pi : \mathcal{H}_1 \times \mathcal{H}_2 \longrightarrow \mathcal{H}_1 \otimes \mathcal{H}_2$ such for any other other vector space and bilinear form $f : \mathcal{H}_1 \times \mathcal{H}_2 \longrightarrow F$, f is uniquely written as $\varphi \circ \pi$ where φ is linear from $\mathcal{H}_1 \otimes \mathcal{H}_2$ to F. A natural scalar product on $\mathcal{H}_1 \otimes \mathcal{H}_2$ is

$$\big\langle x_1 \otimes x_2\big|y_1 \otimes y_2\big\rangle = \big\langle x_1\big|y_1\big\rangle_{\mathcal{H}_1}\big\langle x_2\big|y_2\big\rangle_{\mathcal{H}_2}$$

Then coming back to the covariance we have

$$\begin{aligned}
\text{Cov}\,[f(X),g(Y)] &= E[f(X)g(Y)] - E[f(x)]E[g(Y)] \\
&= E\Big[\big\langle f \otimes g\big|k_x(.,X) \otimes k_Y(.,Y)\big\rangle_{\mathcal{H}_x \otimes \mathcal{H}_y}\Big] - \big\langle f\big|m_X\big\rangle_{\mathcal{H}_x}\big\langle g\big|m_Y\big\rangle_{\mathcal{H}_y} \\
&= \big\langle f \otimes g\big|E\big[k_x(.,X) \otimes k_Y(.,Y)\big]\big\rangle_{\mathcal{H}_x \otimes \mathcal{H}_y} - \big\langle f \otimes g\big|m_X \otimes m_Y\big\rangle_{\mathcal{H}_x \otimes \mathcal{H}_y} \\
&= \big\langle f \otimes g\big|E\big[(k_x(.,X) - m_X) \otimes (k_Y(.,Y) - m_Y)\big]\big\rangle_{\mathcal{H}_x \otimes \mathcal{H}_y}
\end{aligned}$$

Thus the covariance can be seen as the mean kernel in the tensorial product of the RKHS. It is explicitly given by

$$m_{X,Y}(u,v) = \int (k_x(u,x) - m_X(u))(k_y(v,y) - m_Y(v))dP_{X,Y}(x,y)$$

Using the integral expression for Σ_{XY} we have

$$\begin{aligned}
(\Sigma_{YX}f)(v) &= \int \big(f(x) - E[f(X)]\big)\big(k_y(v,y) - m_Y(v)\big)dP_{XY}(x,y) \\
&= \int \big(\langle f(.)\big|k(.,x) - m_X(.)\big|\rangle_{\mathcal{H}_x}\big)\big(k_y(v,y) - m_Y(v)\big)dP_{XY}(x,y) \\
&= \big\langle f(.)\big|m_{X,Y}(.,v)\big\rangle_{\mathcal{H}_x}
\end{aligned}$$

Thus, Σ_{XY} and m_{XY} are identified by this relation.

6.1 Estimation

Even if operators between infinite dimensional space, covariance operators can be consistently estimated! Let (X_i, Y_i) N identically distributed realizations of the random variables X and Y.

The mean and the covariance are then estimated using their empirical estimators, of

$$\begin{aligned}\widehat{m}_X^N &= \frac{1}{N}\sum_{i=1}^{N} k_x(.,X_i) \\ \widehat{m}_{XY}^N &= \frac{1}{N}\sum_{i=1}^{N}(k_x(.,X_i)-\widehat{m}_X)\otimes(k_y(.,Y_i)-\widehat{m}_Y)\end{aligned}$$

If the realization are independent or satisfy some mixing conditions, then the law of large number may be applied to prove the almost sure convergence of these estimators. In [**?**], it is shown that $\sqrt{N}(\widehat{m}_X^N/N - E[k(.,X)])$ converges to (weakly) to a gaussian probability on $\mathcal{H}$ with covariance Cov $[f(X), g(X)]$, f and g in $\mathcal{H}$.

We can also show that the estimator converges strongly. Let us evaluate $\varepsilon_N^2 = E\|\widehat{m}_X^N - m_x\|^2$. We have

$$\begin{aligned}\varepsilon_N^2 &= \frac{1}{N^2}E\|\sum_i\left(k(.,X_i)-m_X\right)\|^2 \\ &= \frac{1}{N^2}\sum_{i,j}E\big[k(X_i,X_j)-m_X(X_i)-m_X(X_j)-\|m_X\|^2\big]\end{aligned}$$

Note that $\|m_X\|^2 = \int k(x,y)dP_X(x)dP_X(y)$. Furthermore by definition of the mean, $E[m_X(X_i)] = \langle m_X \big| m_X\rangle = \|m_X\|^2$. We also have for $i \neq j$ $Ek(X_i,X_j) = \int k(x,y)dP_X(x)dP_X(y) = \|m_X\|^2$. Therefore, the cross term cancelled, and the final result is

$$\begin{aligned}\varepsilon_N^2 &= \frac{1}{N}\big(Ek(X,X)-\|m_X\|^2\big) \\ &= \frac{1}{N}E\big\|k(.,X)-m_X\big\|^2\end{aligned}$$

Practically, these operators will be used by applying them to functions in the RKHS. But, when dealing with a finite number of observations, we have seen in the previous sections that most of optimizing function are finite combination of the kernel, if the optimizers are searched for in the RKHS. Thus, practically, these empirical operators will be applied to functions in the form

$$f(.) = \sum_{i=1}^{N}\alpha_i k_x(.,X_i)$$

Let us now explicitly write the result of these applications. First, consider $m = \langle f | \widehat{m}_X^N \rangle$. We have

$$\begin{aligned}
m &= \big\langle \sum_{i=1}^N \alpha_i k_x(.,X_i) \big| \frac{1}{N}\sum_{i=1}^N k_x(.,X_i) \big\rangle \\
&= \frac{1}{N}\sum_{i,j=1}^N \alpha_i \big\langle k_x(.,X_i) \big| k_x(.,X_j) \big\rangle \\
&= \frac{1}{N}\sum_{i,j=1}^N \alpha_i k_x(X_i,X_j)
\end{aligned}$$

Consider the matrix $\boldsymbol{K}_x$ with entries $\boldsymbol{K}_{x,ij} = k_x(X_i,X_j)$. We already have seen this matrix called the Gram matrix. Introduce also $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_N)^\top$ and $\mathbf{1}_N = (1/N,\ldots,1/N)^\top$ to finally write

$$m = \mathbf{1}_N^\top \boldsymbol{K}_x \boldsymbol{\alpha} = \boldsymbol{\alpha}^\top \boldsymbol{K}_x \mathbf{1}_N$$

Note that this formula leads to $\langle k(.,X_i) | \widehat{m}_X^N \rangle = \delta_i^\top \boldsymbol{K}_x \mathbf{1}_N$ where δ_i is a vector of zeros except a 1 at the ith position.

Note that once again, the mean can be calculating without embedding explicitely the data into the RKHS, but just by using the kernel evaluated at the data points.

Now for the covariance, let us first evaluate

$$\begin{aligned}
\big\langle f(.) \big| \widehat{m}_{XY}^N(.,v) \big\rangle_{\mathcal{H}_x} &= \big\langle \sum_{i=1}^N \alpha_i k_x(.,X_i) \big| \frac{1}{N}\sum_{i=1}^N (k_x(.,X_i) - \widehat{m}_X)(k_y(v,Y_i) - \widehat{m}_Y(v)) \big\rangle_{\mathcal{H}_x} \\
&= \frac{1}{N}\sum_{i,j=1}^N \alpha_i \big\langle k_x(.,X_i) \big| (k_x(.,X_j) - \widehat{m}_X) \big\rangle_{\mathcal{H}_x} (k_y(v,Y_j) - \widehat{m}_Y(v)) \\
&= \frac{1}{N}\sum_{i,j=1}^N \alpha_i \Big(\big\langle k_x(.,X_i) \big| (k_x(.,X_j) \big\rangle - \big\langle k_x(.,X_i) \big| \widehat{m}_X \big\rangle \Big) (k_y(v,Y_j) - \widehat{m}_Y(v)) \\
&= \frac{1}{N}\sum_{i,j=1}^N \alpha_i (K_{x,ij} - \delta_i^\top \boldsymbol{K}_x \mathbf{1}_N)(k_y(v,Y_j) - \widehat{m}_Y(v)) \\
&= \frac{1}{N}\sum_{j=1}^N \big((\boldsymbol{\alpha}^\top \boldsymbol{K}_x)_j - \boldsymbol{\alpha}^\top \boldsymbol{K}_x \mathbf{1}_N \big)(k_y(v,Y_j) - \widehat{m}_Y(v))
\end{aligned}$$

We can now apply this result to a function $g \in \mathcal{H}_y$ to obtain

$$\begin{aligned}
\langle g \big| \widehat{\Sigma}_{YX} f \rangle &= \text{Cov}\,[g(Y), f(X)] \\
&= \frac{1}{N}\sum_{j=1}^{N} \big((\boldsymbol{\alpha}^\top \boldsymbol{K}_x)_j - \boldsymbol{\alpha}^\top \boldsymbol{K}_x \mathbf{1}_N \big) \big\langle \sum_{i=1}^{N} \beta_i k_y(v, Y_i) \big| k_y(v, Y_j) - \widehat{m}_Y(v) \big\rangle \\
&= \frac{1}{N}\sum_{j=1}^{N} \big((\boldsymbol{\alpha}^\top \boldsymbol{K}_x)_j - \boldsymbol{\alpha}^\top \boldsymbol{K}_x \mathbf{1}_N \big) \sum_{i=1}^{N} \beta_i \big(K_{y,ij} - \delta_i^\top \boldsymbol{K}_y \mathbf{1}_N \big) \\
&= \frac{1}{N}\sum_{j=1}^{N} \big((\boldsymbol{\alpha}^\top \boldsymbol{K}_x)_j - \boldsymbol{\alpha}^\top \boldsymbol{K}_x \mathbf{1}_N \big) \big((\boldsymbol{\beta}^\top \boldsymbol{K}_y)_j - \boldsymbol{\beta}^\top \boldsymbol{K}_y \mathbf{1}_N \big) \\
&= \frac{1}{N} \boldsymbol{\alpha}^\top \boldsymbol{K}_x (\boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top) \boldsymbol{K}_y \boldsymbol{\beta}
\end{aligned}$$

The matrix $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ is the so-called centering matrix. If the mean operators are zero, then this matrix does not appear in these calculations. Note that $\boldsymbol{C}$ is idempotent, $\boldsymbol{C}^2 = \boldsymbol{C}$.

The main conclusion of this is the fact that the empirical mean and covariance operators are simply linked to the Gram matrices of the kernels.

7 Dependence measures

As application of the statistics in RKHS we present now recent development of their use to assess independence.

7.1 Independence measures

Measuring or assessing correlation is a very easy task from observations: we just need to estimate the correlation! Assessing independence is much more difficult to verify. A basic definition of independence is the fact that probability of independent events factorizes. By extension, two random variables are said independent if and only if $\Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$. This implies factorization of joint densities if they exist, of moments of separable functions, ... Thus, measures of independence are quite difficult to use. Approximate measures can exist, for example relying on cumulants of small orders. A popular way for assessing independence is to measure a divergence between a joint measure and the product of its marginal. If the divergence is well chosen, such as *e.g.* Kulback divergence, then independence is equivalent to zero divergence. However, divergences are difficult to estimate.

Another simple result may guide us in developing the intuition of using kernel for this problem. X and Y are independent if and only if $\text{Cov}\,[f(X), g(Y)] = 0$ for any continuous bounded function f and g. Evaluating these covariance for all functions may be cumbersome. However it is tempting to use this over a sufficiently rich class of function. This leads to the following measure.

Let $\mathcal{H}_x$ and $\mathcal{H}_y$ be RKHS. Then we measure the maximum correlation between transforms of X and Y using function from the RKHS. Let

$$COCO(X, Y; \mathcal{H}_x, \mathcal{H}_y) := \sup_{f \in \mathcal{H}_x, g \in \mathcal{H}_y} \text{Cov } [f(X), g(Y)]$$

The problem of course is to ensure that the spaces are rich enough so that the result for bounded continuous functions remains valid. Here, the concept of universal kernel enters as a fundamental concept.

A kernel is universal if its RKHS in dense into the space of continuous bounded functions for the sup norm. In this case, any continuous bounded functions can be approximately as closely as possible by a function in the RKHS.

We have the following result:

Theorem 7.1 *Suppose $\mathcal{H}_x, \mathcal{H}_y$ are RKHS of universal kernels. Then X and Y are independent if and only if $COCO(X, Y; \mathcal{H}_x, \mathcal{H}_y) = 0$.*

Proof If X and Y are independent, then for any f and g Cov $[f(X), g(Y)] = 0$. The sufficient condition is more tedious.

Suppose $COCO(X, Y; \mathcal{H}_x, \mathcal{H}_y) = 0$. To show independence it suffices to show that this implies $COCO(X, Y; \mathcal{C}(\mathcal{X}), \mathcal{C}(\mathcal{Y})) = 0$ where $\mathcal{C}(\mathcal{X})$ is the set of continuous bounded functions defined on $\mathcal{X}$.

The inverse will be shown: if $COCO(X, Y; \mathcal{C}(\mathcal{X}), \mathcal{C}(\mathcal{Y})) > 0$ then $COCO(X, Y; \mathcal{H}_x, \mathcal{H}_y) > 0$.

Let $f \in \mathcal{C}(\mathcal{X}), g \in \mathcal{C}(\mathcal{Y})$. For any f_x and g_y of respectively $\mathcal{H}_x, \mathcal{H}_y$ we have

$$\begin{aligned} \big|\text{Cov } [f_x, g_y] - \text{Cov } [f, g]\big| &= \big|\text{Cov } [f_x - f, g_y - g] + \text{Cov } [f, g_y - g] + \text{Cov } [f_x - f, g]\big| \\ &\leq \big|\text{Cov } [f_x - f, g_y - g]\big| + \big|\text{Cov } [f, g_y - g]\big| + \big|\text{Cov } [f_x - f, g]\big| \end{aligned}$$

Furthermore,

$$\begin{aligned} \big|\text{Cov } [f_x - f, g_y - g]\big| &\leq \big|E[(f_x - f)(g_y - g)]\big| + \big|E[(f_x - f)]E[(g_y - g)]\big| \\ &\leq 2\|f_x - f\|_\infty \|g_y - g\|_\infty \\ \big|\text{Cov } [f, g_y - g]\big| &\leq \big|E[f(g_y - g)]\big| + \big|E[f]E[(g_y - g)]\big| \\ &\leq 2\|g_y - g\|_\infty \|f\|_\infty \end{aligned}$$

Therefore

$$\big|\text{Cov } [f_x, g_y] - \text{Cov } [f, g]\big| \leq 2\|f_x - f\|_\infty \|g_y - g\|_\infty + 2\|g_y - g\|_\infty \|f\|_\infty + 2\|f_x - f\|_\infty \|g\|_\infty$$

and we obtain

$$\text{Cov } [f_x, g_y] \geq \text{Cov } [f, g] - 2\|f_x - f\|_\infty \|g_y - g\|_\infty - 2\|g_y - g\|_\infty \|f\|_\infty - 2\|f_x - f\|_\infty \|g\|_\infty$$

Now the argument goes as follows. Let $\mathcal{B}(\mathcal{X}) \subset \mathcal{C}(\mathcal{X})$ the subset of bounded continuous functions with sup norm less than or equal to 1. Then,

$$\sup_{f\in\mathcal{B}(\mathcal{X}),g\in\mathcal{B}(\mathcal{Y})} \text{Cov}\ [f(X), g(Y)] \leq \sup_{f\in\mathcal{C}(\mathcal{X}),g\in\mathcal{C}(\mathcal{Y})} \text{Cov}\ [f(X), g(Y)]$$

Thus it suffices to work on these unit ball sets. Call $c = \sup_{f\in\mathcal{B}(\mathcal{X}),g\in\mathcal{B}(\mathcal{Y})} \text{Cov}\ [f(X), g(Y)]$ and assume it is strictly positive. Note that is necessarily lower than 2 since $\big|\text{Cov}\ [f, g]\big| \leq 2\|f\|_\infty\|g\|_\infty$. Since c is a supremum, there exist f and g in $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$ such that $\text{Cov}\ [f(X), g(Y)] > c/2$. We have to show that we can find f_x and g_y such that their covariance will remain strictly positive.

Let ε_x such that $c\alpha \geq \varepsilon > 0$ for some positive α . Since $\mathcal{H}_x$ (resp. $\mathcal{H}_y$) is dense in $\mathcal{C}(\mathcal{X})$ (resp. $\mathcal{C}(\mathcal{Y})$), we can find $f_x \in \mathcal{H}_x$, (resp. $g_y \in \mathcal{H}_y$) such that $\|f_x - f\|_\infty \leq \varepsilon$ (resp. $\|g_y - g\|_\infty \leq \varepsilon$). Thus from the minorization above and the fact that f, g have norm less than 1, we obtain

$$\begin{aligned} \text{Cov}\ [f_x, g_y] &\geq \text{Cov}\ [f, g] - 2\varepsilon^2 - 2\varepsilon - 2\varepsilon \\ &\geq c/2.\big(1 - 8\alpha - 4\alpha^2 c\big) \end{aligned}$$

which is strictly positive if $c < 1/(4\alpha^2) - 2/\alpha$. Since c is lower than 2, choosing $\alpha = 1/10$ is sufficient to ensure that $\text{Cov}\ [f_x, g_y] > 0$. ■

Of course to be useable, the measure needs to be easily estimable. It turns out that if we restrict the functions f and g to be of norm less than one, a beautiful estimator can be exhibited, without loosing the previous result. Indeed the previous theorem remains true if we restrict the function to belong to the unit ball in their respective RKHS. It suffices to renormalize the covariance at the end of the proof by the norms of the functions, $\|f\|_{\mathcal{H}_x}$ and $\|y\|_{\mathcal{H}_y}$.

Therefore, the measure used is

$$COCO(X, Y; \mathcal{H}_x, \mathcal{H}_y) := \sup_{f\in\mathcal{U}_x,g\in\mathcal{U}_y} \text{Cov}\ [f(X), g(Y)]$$

where $\mathcal{U}_x = \{f \in \mathcal{H}_x / \|f\|_{\mathcal{H}_x} \leq 1\}$. Thus, if we write this using the covariance operator we end up with

$$COCO(X, Y; \mathcal{H}_x, \mathcal{H}_y) = \sup_{f\in\mathcal{U}_x,g\in\mathcal{U}_y} \big\langle g\big|\Sigma_{YX} f\big\rangle_{\mathcal{H}_y}$$

This is precisely the norm of the operator. Indeed, the operator norm is defined as

$$\|A\| \quad = \quad \sup_{f\in\mathcal{U}_x} \|Af\|_{\mathcal{H}_y}$$

But using Cauchy-Schwartz we have $|\big\langle g\big|f\big\rangle| \leq \|g\|_{\mathcal{H}_y}\|f\|_{\mathcal{H}_y}$ so that $\|g\|_{\mathcal{H}_y} = \sup_{f\in\mathcal{U}_y} |\big\langle g\big|f\big\rangle|$.

If we are given N realizations X_i, Y_i of the random variables, then the measure will be estimated as

$$COCO(X_i, Y_i; \mathcal{H}_x, \mathcal{H}_y) = \sup_{f\in\mathcal{U}_x,g\in\mathcal{U}_y} \big\langle g\big|\widehat{\Sigma}_{YX} f\big\rangle_{\mathcal{H}_y}$$

wich can be evaluated in close form! In fact we can apply the representer theorem to sow that a solution of the opimization problem here can be search for as $f = \sum_i \alpha_i k(., X_i)$ and $g = \sum_i \alpha_i k(., Y_i)$ and using the empirical form of the covariance operator obtained in the preceding paragraph, and the fact that $\|f\|^2 = \sum_{i,j} \alpha_i \alpha_j \big\langle k(., X_i) \big| k(., X_j) \big\rangle = \boldsymbol{\alpha}^\top \boldsymbol{K}_x \boldsymbol{\alpha}$, we immediately get

$$COCO(X_i, Y_i; \mathcal{H}_x, \mathcal{H}_y) = \sup_{\boldsymbol{\alpha}^\top \boldsymbol{K}_x \boldsymbol{\alpha} = \boldsymbol{\beta}^\top \boldsymbol{K}_y \boldsymbol{\beta} = 1} \frac{1}{N} \boldsymbol{\alpha}^\top \boldsymbol{K}_x \boldsymbol{C} \boldsymbol{K}_y \boldsymbol{\beta}$$

where recall that $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{N} \mathbf{1}\mathbf{1}^\top$.

A first way to handle that is to et $\boldsymbol{\alpha} \leftrightarrow \boldsymbol{K}_x^{1/2} \boldsymbol{\alpha}$ and $\boldsymbol{\beta} \leftrightarrow \boldsymbol{K}_y^{1/2} \boldsymbol{\beta}$ this reduces to

$$\begin{aligned} COCO(X_i, Y_i; \mathcal{H}_x, \mathcal{H}_y) &= \sup_{\boldsymbol{\alpha}^\top \boldsymbol{\alpha} = \boldsymbol{\beta}^\top \boldsymbol{\beta} = 1} \frac{1}{N} \boldsymbol{\alpha}^\top \boldsymbol{K}_x^{1/2} \boldsymbol{C} \boldsymbol{K}_y^{1/2} \boldsymbol{\beta} \\ &= \frac{1}{N} \left\| \boldsymbol{K}_x^{1/2} \boldsymbol{C} \boldsymbol{K}_y^{1/2} \right\|_2 \end{aligned}$$

where $\|A\|_2$ is the usual matrix spectral norm, $\|A\|_2 = \sqrt{\lambda_m(A^\top A)}$, where λ_m is the largest eigenvalue.

Noting that $\lambda_m(B) = \sqrt{\lambda_m(BB)}$ we have $\|AA^\top\|_2 = \|A^\top A\|_2 = \|A\|_2^2$. Further, if we denote $\tilde{\boldsymbol{K}}_x = \boldsymbol{C} \boldsymbol{K}_x \boldsymbol{C}$, if we recall that $\boldsymbol{C}\boldsymbol{C} = \boldsymbol{C}$, and that all the matrices are symmetric, then we have

$$\begin{aligned} COCO(X_i, Y_i; \mathcal{H}_x, \mathcal{H}_y) &= \frac{1}{N} \sqrt{\left\| \boldsymbol{K}_x^{1/2} \boldsymbol{C} \boldsymbol{K}_y^{1/2} \boldsymbol{K}_y^{1/2} \boldsymbol{C} \boldsymbol{K}_x^{1/2} \right\|_2} \\ &= \frac{1}{N} \sqrt{\left\| \boldsymbol{K}_x^{1/2} \boldsymbol{C} \tilde{\boldsymbol{K}}_y^{1/2} \tilde{\boldsymbol{K}}_y^{1/2} \boldsymbol{C} \boldsymbol{K}_x^{1/2} \right\|_2} \\ &= \frac{1}{N} \sqrt{\left\| \tilde{\boldsymbol{K}}_y^{1/2} \boldsymbol{C} \boldsymbol{K}_x^{1/2} \boldsymbol{K}_x^{1/2} \boldsymbol{C} \tilde{\boldsymbol{K}}_y^{1/2} \right\|_2} \\ &= \frac{1}{N} \sqrt{\left\| \tilde{\boldsymbol{K}}_y^{1/2} \tilde{\boldsymbol{K}}_x \tilde{\boldsymbol{K}}_y^{1/2} \right\|_2} \end{aligned}$$

But this requires the calculation of the square roots and of the eigenvalues which may by demanding in terms of calculation. For a sample of size N, the matrices are N dimensional. For large data sets, this approach may not be very practical. There is another way of using the covariance operators to create independence measures, and uses another operator norm called the Hilbert-Schmidt norm. This requires some insights into Hilbert-Schmidt operators.

An operator from $A : \mathcal{H}_x \longrightarrow \mathcal{H}_y$ is said to be Hilbert-Schmidt if for any complete orthonormal systems $\{\varphi_i\}$ of $\mathcal{H}_x$, the quantity

$$\|A\|_{HS}^2 = \sum_i \|A\varphi_i\|_{\mathcal{H}_y}^2$$

is finite. This is a norm on the space of HS operator. Its independent on the choice of the basis.

Indeed, let $\{\varphi_i'\}$ another ON basis of $\mathcal{H}_x$. Then

$$\begin{aligned}
\sum_i \|A\varphi_i\|^2_{\mathcal{H}_y} &= \sum_i \|(A^\top A)^{1/2}\varphi_i\|^2_{\mathcal{H}_x} \\
&= \sum_i \sum_k \left|\langle \varphi_k' \middle| (A^\top A)^{1/2}\varphi_i\rangle_{\mathcal{H}_x}\right|^2 \\
&= \sum_k \|(A^\top A)^{1/2}\varphi_k'\|^2_{\mathcal{H}_x}
\end{aligned}$$

where the second line is nothing but Parseval equality. The norm has also an expression involving a basis in $\mathcal{H}_y$. Let $\{\psi_i'\}$ a ON basis of $\mathcal{H}_y$. Then

$$\begin{aligned}
\|A\|^2_{HS} &= = \sum_i \langle A\varphi_i \big| A\varphi_i\rangle_{\mathcal{H}_y} \\
&= \sum_i \langle \sum_k < \psi_k \big| A\varphi_i >_{\mathcal{H}_y} \psi_k \big| A\varphi_i\rangle_{\mathcal{H}_y} \\
&= \sum_{i,k} \langle \psi_k \big| A\varphi_i\rangle^2_{\mathcal{H}_y}
\end{aligned}$$

Theorem VI-23 in Reed&Simon states that operators A defined on $L^2(\mathcal{X}, dP)$ are Hilbert-Schmidt if and only if there is a kernel $K \in L^2(\mathcal{X}\times\mathcal{X}, dP\otimes dP)$ such that

$$(Af)(x) = \int K(x,y)f(y)dP(y)$$

The Hilbert-Schmidt norm is then given by

$$\|A\|^2_{HS} = \int \big|K(x,y)\big|^2 dP(x)dP(y)$$

Finally it is known that $\|A\| \leq \|A\|_{HS}$.

We can now come back to the covariance operator. The HS norm of its estimator will be evaluated as follows. Pick any basis in $\mathcal{H}_x$. We have

$$\|\widehat{\Sigma}_{YX}\|^2_{HS} = \sum_i \langle \widehat{\Sigma}_{YX}\varphi_i \big| \widehat{\Sigma}_{YX}\varphi_i\rangle$$

and recall that $\widehat{\Sigma}_{YX}\varphi_i = N^{-1}\sum_k \tilde{k}_y(.,Y_k)\langle \tilde{k}_x(.,X_k)\big|\varphi_i\rangle$, where $\tilde{k}_x := k_x - \widehat{m}_X$. Thus,

$$\begin{aligned}
\|\widehat{\Sigma}_{YX}\|^2_{HS} &= \frac{1}{N}\sum_{i,k} \langle \tilde{k}_y(.,Y_k)\big|\widehat{\Sigma}_{YX}\varphi_i\rangle\langle \tilde{k}_x(.,X_k)\big|\varphi_i\rangle \\
\|\widehat{\Sigma}_{YX}\|^2_{HS} &= \frac{1}{N^2}\sum_{i,k,l} \langle \tilde{k}_y(.,Y_k)\big|\tilde{k}_y(.,Y_l)\rangle\langle \tilde{k}_x(.,X_l)\big|\varphi_i\rangle\langle \tilde{k}_x(.,X_k)\big|\varphi_i\rangle \\
&= \frac{1}{N^2}\sum_{k,l} (\widetilde{\boldsymbol{K}}_y)_{kl}\langle \sum_i \langle \tilde{k}_x(.,X_l)\big|\varphi_i\rangle\varphi_i\big|\tilde{k}_x(.,X_k)\rangle \\
&= \frac{1}{N^2}\sum_{k,l} (\widetilde{\boldsymbol{K}}_y)_{kl}\langle \tilde{k}_x(.,X_l)\big|\tilde{k}_x(.,X_k)\rangle \\
&= \frac{1}{N^2}\sum_{k,l} (\widetilde{\boldsymbol{K}}_y)_{kl}(\widetilde{\boldsymbol{K}}_x)_{kl} \\
&= \frac{1}{N^2}\mathrm{Tr}\big(\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}_x\big)
\end{aligned}$$

In these expressions, $\widetilde{\boldsymbol{K}}$ is the Gram matrix associated with the corresponding centered kernel, and is is easy to show that $\widetilde{\boldsymbol{K}} = \boldsymbol{CKC}$. $\|\hat{\Sigma}_{YX}\|_{HS}^2 = N^{-2}\mathrm{Tr}\big(\boldsymbol{K}_y\boldsymbol{C}\boldsymbol{K}_x\boldsymbol{C}\big)$. Then Comparing this with COCO, we see that calculating the Hilbert-Schmidt norm is much easier, since we just have to take the trace instead of calculating singular values. Furthermore the trace norm is bigger than the usual 2 norm. Thus if it is zero, then COCO is zero and the variables independent. Thus we have the same theorem with the Hilbert-Schmidt norm than with the usual 2 norm.

7.2 Conditional independence measures

Some recalls

Let X, Y, Z be three random variables in $\mathbb{R}^p$, and $\hat{X}(Z)$ and $\hat{Y}(Z)$ the best linear MMSE estimates of X and Y based on Z. It is well-known that these are given by $\hat{X}(Z) = \Sigma_{XZ}\Sigma_{ZZ}^{-1}Z$ and $\hat{Y}(Z) = \Sigma_{YZ}\Sigma_{ZZ}^{-1}Z$. The errors $X - \hat{X}(Z)$ are orthogonal to the linear subspace generated by Z, and this can be used to show the well-known relations

$$\begin{aligned}
\Sigma_{XX|Z} &:= \mathrm{Cov}\left[X - \hat{X}(Z), X - \hat{X}(Z)\right] \\
&= \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \\
\Sigma_{XY|Z} &:= \mathrm{Cov}\left[X - \hat{X}(Z), Y - \hat{Y}(Z)\right] \\
&= \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}
\end{aligned}$$

$\Sigma_{XX|Z}$ is the covariance of the error in estimating X linearly from Z. It is also called the partial covariance and it is equal to the conditional covariance in the Gaussian case. The second term measures the correlation remaining between X and Y once the effect of their possibly common observed cause Z has been linearly removed from them. $\Sigma_{XY|Z}$ is called the partial cross-covariance matrix and is equal to the conditional cross-covariance in the Gaussian case.

Therefore, in the Gaussian case, conditional independence can be assessed using linear prediction on the partial cross-covariance matrix. This has led to extensive development in the field of graphical modeling.

Using kernels

The approach above can be extended to assess conditional independence. It relies on the notion of conditional cross-covariance operators, a natural extension of the covariance operators. Having in mind that cross-covariance operators suffices to assess independence (as cross-covariance does in the finite dimensional Gaussian case), the idea is to study

$$\Sigma_{XY|Z} := \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$$

as a potential candidate to assess conditional independence. The first remark concerns the existence of this operator.

Σ_{ZZ} is an operator from $\mathcal{H}_z$ to itself. Let $\ker \Sigma_{ZZ}$ and $\mathrm{Im}\ \Sigma_{ZZ}$ be respectively its kernel and its range. We will suppose that it is invertible on its range and we will abusively denote the inverse as Σ_{ZZ}^{-1}. The inverse exits in full generality if and only if $\ker \Sigma_{ZZ} = \{0\}$ and $\mathrm{Im}\, \Sigma_{ZZ} = \mathcal{H}_z$, corresponding to injectivity and surjectivity. Thus in the sequel, when dealing with ensemble operators, covariance operator will be supposed invertible.

We could work with normalized covariance operators V, defined using $\Sigma_{XY} = \Sigma_{XX}^{1/2} V_{XY} \Sigma_{YY}^{1/2}$. Thus we could use

$$\Sigma_{XY|Z} := \Sigma_{XY} - \Sigma_{XX}^{1/2} V_{XZ} V_{ZY} \Sigma_{YY}^{1/2}$$

and the normalized version

$$V_{XY|Z} := V_{XY} - V_{XZ} V_{ZY}$$

Several theorems show the meaning of these operators, and how we can assess conditional independence with them. They are all mainly due to K. Fukumizu, F. Bach and M. Jordan in their publications.

The first result links conditional expectation to covariance and cross-covariance operators.

Theorem 7.2 *For all $g \in \mathcal{H}_y$,*

$$\big\langle g \big| \Sigma_{YY|X} g \big\rangle = \inf_{f \in \mathcal{H}_x} E\Big[\big((g(Y) - E[g(Y)]) - (f(X) - E[f(X)]\big)^2\Big]$$

If furthermore the direct sum $\mathcal{H}_x + \mathbb{R}$ is dense in $L^2(P_X)$, then

$$\big\langle g \big| \Sigma_{YY|X} g \big\rangle = E_X\left[Var[g(Y)\big|X]\right]$$

The density assumption means than any second order random variable function of X can be approximated as closely as desired by a function $\mathcal{H}_x$ plus a real. Note that the result of the theorem is an extension of what we recalled above, but stated in RKHS. The operator $\Sigma_{YY|X}$ measures the power of the error in approximating a function of a random variable in a RKHS by a function of another in its respective RKHS. The second result generalizes the Gaussian case since under the assumption of density the operator evaluates a conditional variance.

Proof Bidou's proof. Let $\mathcal{E}_g(f) = E\Big[\big((g(Y) - E[g(Y)]) - (f(X) - E[f(X)]\big)^2\Big]$. Then f_0 provides the infimum if $\mathcal{E}_g(f_0 + f) - \mathcal{E}_g(f_0) \geq 0$ for all $f \in \mathcal{H}_x$. But we have

$$\mathcal{E}_g(f_0 + f) - \mathcal{E}_g(f_0) \quad = \quad \big\langle \Sigma_{XX} f \big| f \big\rangle + 2\big\langle \Sigma_{XX} f_0 - \Sigma_{XY} g \big| f \big\rangle$$

Obviously, $\Sigma_{XX} f_0 - \Sigma_{XY} g = 0$ satisfies the condition. It is also necessary. Indeed, suppose $\Sigma_{XX} f_0 - \Sigma_{XY} g \neq 0$. Σ_{XX} is auto-ajoint and thus only has positive or null eigen values. Thus $\Sigma_{XX} f = -f$ has no solution and $\ker \Sigma_{XX} + I = 0$ and $\Sigma_{XX} + I$ is invertible. Therefore there is an non zero f such that $\Sigma_{XX} f + f = -2(\Sigma_{XX} f_0 - \Sigma_{XY} g)$, and this f satisfies $\mathcal{E}_g(f_0 + f) - \mathcal{E}_g(f_0) = -\big\langle f \big| f \big\rangle < 0$, giving a contradiction. Thus, this gives the result. Note we use the fact that Σ_{XX} is invertible, at least on its range.

Fukumizu's proof. The proof in Fukumizu's paper does not need to explicitely solve the problem and then does not require invertibility of the covariance.

If we write $\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$ then

$$\begin{aligned}
\mathcal{E}_g(f) &:= E\Big[\big((g(Y) - E[g(Y)]) - (f(X) - E[f(X)]\big)^2\Big] \\
&= \big\langle f \big| \Sigma_{XX} f \big\rangle + \big\langle g \big| \Sigma_{YY} g \big\rangle - 2\big\langle g \big| \Sigma_{YX} f \big\rangle \\
&= \big\| \Sigma_{XX}^{1/2} f \big\|^2 - 2\big\langle V_{XY} \Sigma_{YY}^{1/2} g \big| \Sigma_{XX}^{1/2} f \big\rangle + \big\| \Sigma_{YY}^{1/2} g \big\|^2
\end{aligned}$$

Rearranging the first two terms we obtain

$$\begin{aligned}
\mathcal{E}_g(f) &= \big\| \Sigma_{XX}^{1/2} f - V_{XY} \Sigma_{YY}^{1/2} g \big\|^2 + \big\langle g \big| \Sigma_{YY} g \big\rangle - \big\langle V_{XY} \Sigma_{YY}^{1/2} g \big| V_{XY} \Sigma_{YY}^{1/2} g \big\rangle \\
&= \big\| \Sigma_{XX}^{1/2} f - V_{XY} \Sigma_{YY}^{1/2} g \big\|^2 + \big\langle g \big| \Sigma_{YY} g \big\rangle - \big\langle g \big| \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2} g \big\rangle \\
&= \big\| \Sigma_{XX}^{1/2} f - V_{XY} \Sigma_{YY}^{1/2} g \big\|^2 + \big\langle g \big| (\Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}) g \big\rangle \\
&= \big\| \Sigma_{XX}^{1/2} f - V_{XY} \Sigma_{YY}^{1/2} g \big\|^2 + \big\langle g \big| \Sigma_{YY|X} g \big\rangle
\end{aligned}$$

Then clearly $\big\langle g \big| \Sigma_{YY|X} g \big\rangle \leq \mathcal{E}_g(f)$ for any $f \in \mathcal{H}_x$. Furthermore, let $\varepsilon > 0$, since $V_{XY} \Sigma_{YY}^{1/2}$ and $\Sigma_{XX}^{1/2}$ have the same range, there is an f_0 such that $\big\| \Sigma_{XX}^{1/2} f_0 - V_{XY} \Sigma_{YY}^{1/2} g \big\|^2 \leq \varepsilon^2$. Thus, for any ε^2, there is a number $\mathcal{E}_g(f_0) \leq \big\langle g \big| \Sigma_{YY|X} g \big\rangle + \varepsilon^2$ and then $\big\langle g \big| \Sigma_{YY|X} g \big\rangle$ is the greatest lower bound of the set of these numbers. This proves the first claim of the theorem.

For the second assertion, use the following well-known relation

$$\mathrm{Var}[X] = \mathrm{Var}_Y[E[X|Y]] + E_Y[\mathrm{Var}[X|Y]]$$

which states that the variance is the sum of the variance of the conditional mean and the mean of the conditional variance. Applying it to $g(Y) - f(X)$ leads to

$$\begin{aligned}
\big\langle g \big| \Sigma_{YY|X} g \big\rangle &= \inf_{f \in \mathcal{H}_x} E\Big[\big((g(Y) - E[g(Y)]) - (f(X) - E[f(X)]\big)^2\Big] \\
&= \inf_{f \in \mathcal{H}_x} \mathrm{Var}\big[g(Y) - f(X)\big] \\
&= \inf_{f \in \mathcal{H}_x} \Big(\mathrm{Var}_X\big[E[g(Y) - f(X) \big| X]\big] + E_X\big[\mathrm{Var}[g(Y) - f(X) \big| X]\big] \Big) \\
&= \inf_{f \in \mathcal{H}_x} \Big(\mathrm{Var}_X\big[E[g(Y) \big| X] - f(X)\big] \Big) + E_X\big[\mathrm{Var}[g(Y) \big| X]\big]
\end{aligned}$$

since the variance is invariant under translation.

Clearly, if $E[g(.) \big| X] \in \mathcal{H}_x$, then the result holds, but, this is not guaranteed. However, if $E[g(.) \big| X]$ is P_x square integrable, then from the assumption of density, the infimum in the precious equation can be made equal to zero.

We know that $\mathrm{Var}[E[g(Y) \big| X]] = \mathrm{Var}[g(Y)] - E[\mathrm{Var}[g(Y)|X]]$ and therefore $\mathrm{Var}[E[g(Y) \big| X]] \leq \mathrm{Var}[g(Y)]$. We have supposed that the kernels involved are such that $E[k(Y,Y)] < +\infty$ which imply that $\mathcal{H}_y \in L^2(P_Y)$. Indeed, for $g \in \mathcal{H}_y$, $E[g(Y)^2] = E[< g(.) | k(.,Y) >^2] \leq \|g\|_{\mathcal{H}_y} E[k(Y,Y)]$. Thus, $\mathrm{Var}[E[g(Y) \big| X]]$ belongs to $L^2(P_X)$. Since $\mathcal{H}_x + \mathbb{R}$ is dense into $L^2(P_X)$, for any $\varepsilon > 0$ there are a $f \in \mathcal{H}_x$ and a $c \in \mathbb{R}$ such that $E[(E[g(Y)\big|X]-f(X)-c)^2] < \varepsilon^2$. But $\mathrm{Var}_X\big[E[g(Y)\big|X]-f(X)\big] = \mathrm{Var}_X\big[g[E[g(Y)\big|X]-f(X)-c\big] \le E[(E[g(Y)\big|X]-f(X)-c)^2] \le \varepsilon^2$. Since ε is arbitrary, we thus have $\inf_{f\in\mathcal{H}_x}\Big(\mathrm{Var}_X\big[E[g(Y)\big|X]-f(X)\big]\Big) = 0$. $\blacksquare$

We have seen in the second part of the proof that the theorem does not require $E[g(.)\big|X] \in \mathcal{H}_x$. However if we suppose so, the statement and results are more direct. In that case, we have for any $g \in \mathcal{H}_y$

$$\Sigma_{XX}E[g(.)\big|X] = \Sigma_{XY}g(.)$$

and this provides a means of calculating the conditional mean in a RKHS if the covariance is invertible. We now turn to the conditional cross-covariance, assuming that the conditional mean of function of RKHS belongs to the proper RKHS.

The we have

$$\begin{aligned}
\big\langle f\big|\Sigma_{XY|Z}g\big\rangle &= \mathrm{Cov}\ [f(X),g(Y)] - \big\langle f\big|\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}g\big\rangle \\
&= \mathrm{Cov}\ [f(X),g(Y)] - \big\langle \Sigma_{ZX}f\big|\Sigma_{ZZ}^{-1}\Sigma_{ZY}g\big\rangle \\
&= \mathrm{Cov}\ [f(X),g(Y)] - \big\langle \Sigma_{ZX}f\big|E[g(.)|Z]\big\rangle \\
&= \mathrm{Cov}\ [f(X),g(Y)] - \big\langle \Sigma_{ZZ}E[f(.)|Z]\big|E[g(.)|Z]\big\rangle \\
&= \mathrm{Cov}\ [f(X),g(Y)] - \mathrm{Cov}\ _Z\big[E[f(.)|Z],E[g(.)|Z]\big] \\
&= E[f(X)g(Y)] - E_Z\big[E[f(.)|Z]E[g(.)|Z]\big] \\
&= E_Z\Big[E[f(X)g(Y)|Z] - E[f(.)|Z]E[g(.)|Z]\Big] \\
&= E_Z\Big[\mathrm{Cov}\ [f(X),g(Y)|Z]\Big]
\end{aligned}$$

The question is now to know whether conditional independence can be measured using the conditional covariance operator. The first result shows that a zero conditional covariance operator is not equivalent to conditional independence, but equivalent to a weaker form. The second result shows how to slightly modify the covariance operator to obtain the equivalence.

We suppose in the following that the all the kernels in the game are characteristics, and that the conditional mean involved belongs to the proper RKHS.

Theorem 7.3 *Let X, Y, Z be three random vectors embedded in corresponding RKHS. Then we have*

1. $\Sigma_{XY|Z} = 0 \Longleftrightarrow P_{XY} = E_Z[P_{X|Z} \otimes P_{Y|Z}]$
2. $\Sigma_{(XZ)Y|Z} = 0 \Longleftrightarrow X \perp Y|Z.$

Proof **First assertion.** We have seen that

$$\big\langle f\big|\Sigma_{XY|Z}g\big\rangle = E[f(X)g(Y)] - E_Z\big[E[f(.)|Z]E[g(.)|Z]\big]$$

which can be written as

$$\begin{aligned}\langle f\big|\Sigma_{XY|Z}g\rangle &= \int f(x)g(y)P_{XY}(dx,dy) - \int P_Z(dz)\int f(x)g(y)P_{X|Z}(dx,z)P_{Y|Z}(dy,z)\\ &= \int f(x)g(y)P_{XY}(dx,dy) - \int f(x)g(y)\int P_Z(dz)P_{X|Z}(dx,z)P_{Y|Z}(dy,z)\end{aligned}$$

Thus obviously, if for all A and B in the adequate sigma algebra

$$P_{XY}(A,B) = \int P_Z(dz)P_{X|Z}(A,z)P_{Y|Z}(B,z)$$

we have $\langle f\big|\Sigma_{XY|Z}g\rangle = 0$ for all f and g leading necessarily to $\Sigma_{XY|Z} = 0$. Now if the covariance operator is zero then we have for all f and g $E_{P_{XY}}[f(X)g(Y)] = E_Q\big[f(X)g(Y)\big]$ where $Q = E_Z[P_{X|Z}\otimes P_{Y|Z}]$. Working in the tensorial product $\mathcal{H}_x\otimes\mathcal{H}_y$ where we have assumed $k_x\otimes k_y$ as a characteristic kernel allows to conclude that $Q = P_{XY}$.

Second assertion. Let A, B, C be elements of the sigma algebra related to X, Y and Z respectively. Let $\mathbf{1}_A$ the characteristic function of set A. Then we have

$$\begin{aligned}&P_{XZY}(A,C,B) - E_Z[P_{XZ|Z}(A,C)P_{Y|Z}(B)]\\ &= E[\mathbf{1}_{A\times C}(X,Z)\mathbf{1}_B(Y)] - E_Z\big[E[\mathbf{1}_{A\times C}(X,Z)|Z]E[\mathbf{1}_B(Y)|Z]\big]\\ &= E_Z\big[E[\mathbf{1}_{A\times C}(X,Z)\mathbf{1}_B(Y)]\big] - E_Z\big[\mathbf{1}_C(X)E[\mathbf{1}_A(X)|Z]E[\mathbf{1}_B(Y)|Z]\big]\\ &= E_Z\big[\mathbf{1}_C(Z)E[\mathbf{1}_A(X)\mathbf{1}_B(Y)|Z]\big] - E_Z\big[\mathbf{1}_C(X)E[\mathbf{1}_A(X)|Z]E[\mathbf{1}_B(Y)|Z]\big]\\ &= E_Z\big[\mathbf{1}_C(Z)\big(E[\mathbf{1}_A(X)\mathbf{1}_B(Y)|Z] - E[\mathbf{1}_A(X)|Z]E[\mathbf{1}_B(Y)|Z]\big)\big]\\ &= \int_C P_Z(dz)\Big(P_{X,Y|Z}(A,B,z) - P_{X|Z}(A,z)P_{X|Z}(B,z)\Big)\end{aligned}$$

If $\Sigma_{(XZ)Y|Z} = 0$ then the first assertion implies $P_{XZY} = E_Z[P_{XZ|Z}\otimes P_{Y|Z}]$ and the previous integral is equal to zero for any C, which in turn implies that $P_{X,Y|Z}(A,B,z) - P_{X|Z}(A,z)P_{X|Z}(B,z)$ almost everywhere (P_Z) for any A, B. But this is precisely the defintion of conditional independence. The converse is evident. ■

Estimation We now develop the estimators of the conditional measures, and give their representation in terms of Gram matrices. In the following, we suppress the indication of the RKHS in which we work and this for the sake of readability. We recall that we have for N identically distributed observations (X_i, Y_i, Z_i)

$$\widehat{\Sigma}_{XY}f = \frac{1}{N}\sum_k \tilde{k}_x(.,X_k)\langle\tilde{k}_y(.,Y_k)\big|f\rangle$$

where the tilde means the kernel are centered. In the sequel, we also need to know how the inverse (assuming it exists) acts on a function. For a covariance operator this can be studied using orthonormal bases of eigen functions. For the empirical operator, we have the following heuristic. Since it is not of full rank, the operator cannot be inverted. However, regularizing it allows to invert it. Regularization can be done by adding a small diagonal operator. Let $\widehat{\Sigma}_{r,XX}$ this regularized opertor. Now we now that applying its inverse to the direct operator, we should obtain the identity, thus

$$\widehat{\Sigma}_{r,XX}^{-1}\widehat{\Sigma}_{r,XX}f = f$$

Since we work in the subspace generated by the N $k_x(.,X_i)$ we should have the two following relations

$$\begin{aligned}\widehat{\Sigma}^{-1}_{r,XX}\widehat{\Sigma}_{r,XX}k_x(.,X_n) &= k_x(.,X_n)\\ \widehat{\Sigma}^{-1}_{r,XX}k_x(.,X_n) &= \sum_k(\boldsymbol{H}_x)_{nk}k_x(.,X_k)\end{aligned}$$

Therefore,

$$\begin{aligned}\tilde{k}_x(.,X_n) &= \widehat{\Sigma}^{-1}_{r,XX}\frac{1}{N}\sum_k\tilde{k}_x(.,X_k)\big\langle\tilde{k}_x(.,X_k)\big|k_x(.,X_n)\big\rangle\\ &= \frac{1}{N}\sum_{k,l}(\widetilde{\boldsymbol{K}}_{r,x})_{kn}(\boldsymbol{H}_x)_{lk}k_x(.,X_l)\end{aligned}$$

This is obtain if $\sum_k(\widetilde{\boldsymbol{K}}_{r,x})_{kn}(\boldsymbol{H}_x)_{lk} = N\delta_{nl}$. Therefore, $\boldsymbol{H}_x = N\widetilde{\boldsymbol{K}}^{-1}_{r,x}$, and we have

$$\widehat{\Sigma}^{-1}_{r,XX}k_x(.,X_n) = N\sum_k(\widetilde{\boldsymbol{K}}^{-1}_{r,x})_{nk}k_x(.,X_k)$$

We can now obtain the representation of the empirical measures in terms of Gram matrices.

Matrix representation. We evaluate $\big\langle f\big|\Sigma_{XY|Z}g\big\rangle$ for $f(.)=\sum_i\alpha_i\tilde{k}(.,X_i)$ and $g(.)=\sum_i\alpha_i\tilde{k}(.,Y_i)$ when we observe N triple X_i,Y_i,Z_i identically distributed. We already saw that

$$\big\langle f\big|\widehat{\Sigma}_{XY}g\big\rangle = \frac{1}{N}\boldsymbol{\beta}^\top\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}_x\boldsymbol{\alpha}$$

Then we have

$$\begin{aligned}\big\langle f\big|\widehat{\Sigma}_{XZ}\widehat{\Sigma}^{-1}_{ZZ}\widehat{\Sigma}_{ZY}g\big\rangle &= \sum_{i,j}\alpha_i\beta_j\big\langle\tilde{k}_x(.,X_i)\big|\widehat{\Sigma}_{XZ}\widehat{\Sigma}^{-1}_{ZZ}\widehat{\Sigma}_{ZY}\tilde{k}_y(.,Y_j)\big\rangle\\ &= \frac{1}{N^2}\sum_{i,j,k}\alpha_i\beta_j\big\langle\tilde{k}_x(.,X_l)\big|\tilde{k}_x(.,X_i)\big\rangle\big\langle\tilde{k}_z(.,Z_l)\big|\widehat{\Sigma}^{-1}_{ZZ}\tilde{k}_z(.,Z_k)\big\rangle\big\langle\tilde{k}_y(.,Y_k)\big|\tilde{k}_y(.,Y_j)\big\rangle\\ &= \frac{1}{N}\sum_{i,j,k,l,m}\alpha_i\beta_j(\widetilde{\boldsymbol{K}}_y)_{kj}(\widetilde{\boldsymbol{K}}_x)_{li}(\widetilde{\boldsymbol{K}}^{-1}_{r,z})_{km}(\widetilde{\boldsymbol{K}}_z)_{ml}\\ &= \frac{1}{N}\boldsymbol{\beta}^\top\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}^{-1}_{r,z}\widetilde{\boldsymbol{K}}_z\widetilde{\boldsymbol{K}}_x\boldsymbol{\alpha}\end{aligned}$$

Thus we get

$$\big\langle f\big|\widehat{\Sigma}_{XY|Z}g\big\rangle = \frac{1}{N}\boldsymbol{\beta}^\top\Big(\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}_x - \widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}^{-1}_{r,z}\widetilde{\boldsymbol{K}}_z\widetilde{\boldsymbol{K}}_x\Big)\boldsymbol{\alpha}$$

Hilbert-Schmidt norms. We have

$$\begin{aligned}\Big\|\widehat{\Sigma}_{XY|Z}\Big\|^2_{HS} &= \sum_i\big\langle\widehat{\Sigma}_{XY|Z}\varphi_i\big|\widehat{\Sigma}_{XY|Z}\varphi_i\big\rangle\\ &= \Big\|\widehat{\Sigma}_{XY}\Big\|^2_{HS} + \Big\|\widehat{\Sigma_{XZ}}\widehat{\Sigma^{-1}_{ZZ}}\widehat{\Sigma_{ZY}}\Big\|^2_{HS} - 2\sum_i\big\langle\widehat{\Sigma}_{XY}\varphi_i\big|\widehat{\Sigma_{XZ}}\widehat{\Sigma^{-1}_{ZZ}}\widehat{\Sigma_{ZY}}\varphi_i\big\rangle\end{aligned}$$

We develop the calculation for the double product term we call P. We have

$$\begin{array}{rcl}
P & := & \sum_i \big\langle \widehat{\Sigma}_{XY}\varphi_i \big| \widehat{\Sigma_{XZ}}\widehat{\Sigma^{-1}_{ZZ}}\widehat{\Sigma_{ZY}}\varphi_i \big\rangle \\
& = & \frac{1}{N}\sum_{i,k} \big\langle \tilde{k}_y(.,Y_k)\big|\varphi_i\big\rangle \big\langle \tilde{k}_x(.,X_k)\big| \widehat{\Sigma_{XZ}}\widehat{\Sigma^{-1}_{ZZ}}\widehat{\Sigma_{ZY}}\varphi_i \big\rangle \\
& = & \frac{1}{N^2}\sum_{i,k,l} \big\langle \tilde{k}_y(.,Y_k)\big|\varphi_i\big\rangle \big\langle \tilde{k}_y(.,Y_l)\big|\varphi_i\big\rangle \big\langle \tilde{k}_x(.,X_k)\big| \widehat{\Sigma_{XZ}}\widehat{\Sigma^{-1}_{ZZ}}\tilde{z}_y(.,Z_l)\big\rangle \\
& = & \frac{1}{N}\sum_{k,l,m} (\widetilde{\boldsymbol{K}}_y)_{kl}(\widetilde{\boldsymbol{K}}^{-1}_{r,z})_{lm} \big\langle \tilde{k}_x(.,X_k)\big| \widehat{\Sigma_{XZ}} k_z(.,Z_m)\big\rangle \\
& = & \frac{1}{N^2}\sum_{k,l,m,n} (\widetilde{\boldsymbol{K}}_y)_{kl}(\widetilde{\boldsymbol{K}}^{-1}_{r,z})_{lm} \big\langle \tilde{k}_x(.,X_k)\big| \tilde{k}_x(.,X_n)\big\rangle \big\langle \tilde{k}_z(.,Z_n)\big| \tilde{k}_z(.,Z_m)\big\rangle \\
& = & \frac{1}{N^2}\sum_{k,l,m,n} (\widetilde{\boldsymbol{K}}_y)_{kl}(\widetilde{\boldsymbol{K}}^{-1}_{r,z})_{lm}(\widetilde{\boldsymbol{K}}_x)_{kn}(\widetilde{\boldsymbol{K}}_z)_{mn} \\
& = & \frac{1}{N^2}\sum_{k,m} (\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}^{-1}_{r,z})_{km}(\widetilde{\boldsymbol{K}}_z\widetilde{\boldsymbol{K}}_x)_{mk} \\
& = & \frac{1}{N^2}\mathrm{Tr}\Big(\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}^{-1}_{r,z}\widetilde{\boldsymbol{K}}_z\widetilde{\boldsymbol{K}}_x\Big)
\end{array}$$

Carrying the same calculation on the last term allows to obtain

$$\Big\|\widehat{\Sigma}_{XY|Z}\Big\|^2_{HS} \;=\; \frac{1}{N^2}\mathrm{Tr}\Big(\widetilde{\boldsymbol{K}}_x\widetilde{\boldsymbol{K}}_y - 2\widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}^{-1}_{r,z}\widetilde{\boldsymbol{K}}_z\widetilde{\boldsymbol{K}}_x + \widetilde{\boldsymbol{K}}_y\widetilde{\boldsymbol{K}}_z\widetilde{\boldsymbol{K}}^{-1}_{r,z}\widetilde{\boldsymbol{K}}_x\widetilde{\boldsymbol{K}}^{-1}_{r,z}\widetilde{\boldsymbol{K}}_z\Big)$$

8 Embedding for inferring

Here we rederive the recent approach developed by Song, Fulumizu and Gretton of the embedding of conditional distribution, something that allow to do bayesian inference in RKHS.

First, we have seen that the conditional expectation of $g(Y)$ given X satisfies

$$E[g(Y)|X=.] = (C^{-1}_{XX}C_{XY}g)(.)$$

provided $E[g(Y)|X=.] \in \mathcal{H}_x$, an assumption we adopt. Let μ_X and μ_Y be the embeddings of X and Y respectively in $\mathcal{H}_x$ and $\mathcal{H}_x$. Then we have $\mu_Y = C_{YX}C^{-1}_{XX}\mu_X$ since

$$\begin{array}{rcl}
\big\langle C_{YX}C^{-1}_{XX}\mu_X\big|g\big\rangle_{\mathcal{H}_y} & = & \big\langle C^{-1}_{XX}\mu_X\big|C_{XY}g\big\rangle_{\mathcal{H}_x} \\
& = & \big\langle \mu_X\big|C^{-1}_{XX}C_{XY}g\big\rangle_{\mathcal{H}_x} \\
& = & \big\langle \mu_X\big|E[g(Y)|X]\big\rangle_{\mathcal{H}_x} \\
& = & E_X[E[g(Y)|X]] \\
& = & E[g(Y)] \\
& = & \big\langle \mu_Y\big|g\big\rangle_{\mathcal{H}_y}
\end{array}$$

In particular, if we set $P(dX) = \delta_x(dX)$ we end up with $\mu_{Y|x} = E[k_y(.,Y)|x] = C_{YX}C_{XX}^{-1}k_x(.,x)$.

Estimate. When confronted to data (X_i, Y_i) $\mu_{Y|x}$ is estimated by

$$\widehat{\mu}_{Y|x} \;=\; \widehat{C}_{YX}\widehat{C}_{r,XX}^{-1}k_x(.,x)$$

where $_r$ stands for regularized. The estimators are defined on and send to the finite dimensional spaces spanned by the $k_x(.,X_i)$ and $k_y(.,Y_i)$.

To study $\widehat{C}_{r,XX}^{-1}k_x(.,x)$, let us write it as $\sum_i \alpha_i k_x(.,X_i)+\perp$, where $\perp$ stands for a function orthogonal to the space spanned by the $k_x(.,X_i)$. Then we have

$$\begin{aligned} k_x(.,x) &= (\widehat{C}_{XX} + \lambda I)(\sum_i \alpha_i k_x(.,X_i)+\perp) \\ &= \sum_{ij} \alpha_i \boldsymbol{K}_{x,ij} k_x(.,X_j) + \lambda \sum_i \alpha_i k_x(.,X_i) + \lambda \perp \end{aligned}$$

Taking the scalar product with $k_x(.,X_k)$ leads to

$$k_x(X_k,x) \;=\; \sum_{ij} \alpha_i \boldsymbol{K}_{x,ij}\boldsymbol{K}_{x,kj} + \lambda \sum_i \alpha_i \boldsymbol{K}_{x,ki}$$

or

$$\boldsymbol{k}_x(x) = \boldsymbol{K}_x\boldsymbol{K}_x\boldsymbol{\alpha} + \lambda\boldsymbol{K}_x\boldsymbol{\alpha} = (\boldsymbol{K}_x \;\; + \lambda I)\boldsymbol{K}_x\boldsymbol{\alpha}$$

Now

$$\begin{aligned} \widehat{C}_{YX}\widehat{C}_{r,XX}^{-1}k_x(.,x) &= \widehat{C}_{YX}(\sum_i \alpha_i k_x(.,X_i)+\perp) \\ &= \sum_{ij} \alpha_i \boldsymbol{K}_{x,ij} k_y(.,Y_j) \\ &= \sum_j (\boldsymbol{K}_x\boldsymbol{\alpha})_j k_y(.,Y_j) \\ &= \sum_j [(\boldsymbol{K}_x + \lambda I)^{-1}\boldsymbol{k}_x(x)]_j k_y(.,Y_j) \end{aligned}$$

Then if we want to find an estimate conditional expectation $E[g(Y)|x]$ for $g = \sum_i \gamma_i k_y(.,Y_i)$ we have

$$\begin{aligned} E[g(Y)|x] &= \langle g|\widehat{\mu}_{Y|x}\rangle \\ &= \sum_{ij} \gamma_i [(\boldsymbol{K}_x + \lambda I)^{-1}\boldsymbol{k}_x(x)]_j \boldsymbol{K}_{y,ij} \\ &= \boldsymbol{\gamma}^\top \boldsymbol{K}_y(\boldsymbol{K}_x + \lambda I)^{-1}\boldsymbol{k}_x(x) \end{aligned}$$

8.1 Bayes laws using embeddings

We saw that the embedding for the conditional mean is given by

$$\mu_{Y|x} = E[k_y(.,Y)|x] = C_{YX}C_{XX}^{-1}k_x(.,x)$$

Here, the couple (X,Y) is distributed according to the joint $P(X,Y)$. Therefore C_{YX} which is the covariance operator may interpretated as the embedding of the the joint into the tensor product $\mathcal{H}_y \otimes \mathcal{H}_x$ whereas C_{XX} can be interpreted as the embedding of $P(X,X)$!! into the tensor product $\mathcal{H}_x \otimes \mathcal{H}_x$.

In a Bayesian approach, we need the embedding of the *a posteriori*probability measure, when a *prior* π is chosen. Therefore we can use the previous result but when the joint is $Q(X,Y) = P(X|Y)\pi(Y)$. Thus we use the superscript π to mention this fact and have the embedding of the *a posteriori*given by

$$\mu_{Y|x}^{\pi} = E^{\pi}[k_y(.,Y)|x] = C_{YX}^{\pi}C_{XX}^{\pi-1}k_x(.,x)$$

Now, it is possible to relate C_{YX}^{π} and $C_{XX}^{\pi-1}$ to embeddings evaluated on the joint P. We have

$$\begin{aligned}
C_{XY}^{\pi} &= E_Q[k_x(.,X) \otimes k_y(.,Y)] \\
&= E_{\pi}\left[E[k_x(.,X) \otimes k_y(.,Y)\big|Y]\right] \\
&= E_{\pi}\left[E[k_x(.,X)\big|Y] \otimes k_y(.,Y)\right] \\
&= E_{\pi}\left[\mu_{X|Y} \otimes k_y(.,Y)\right] \\
&= E_{\pi}\left[C_{X|Y}k_y(.,Y) \otimes k_y(.,Y)\right] \\
&= C_{X|Y}C_{YY}^{\pi}
\end{aligned}$$

The second line can also be interpretated as the average of the embedding of $P(Y,X|X)$ and therefore we have an alternative expression as

$$\begin{aligned}
C_{XY}^{\pi} &= E_{\pi}\left[E[k_x(.,Y) \otimes k_y(.,Y)\big|Y]\right] \\
&= E_{\pi}\left[\mu_{XY|Y}\right] \\
&= E_{\pi}\left[C_{XY|Y}k_y(.,Y)\right] \\
&= C_{XY|Y}\mu_Y^{\pi}
\end{aligned}$$

Likewise we have

$$\begin{aligned}
C_{XX}^{\pi} &= E\left[E[k_x(.,X) \otimes k_x(.,X)]\right] \\
&= E_{\pi}E\left[E[k_x(.,X) \otimes k_x(.,X)\big|Y]\right] \\
&= E_{\pi}E\left[\mu_{XX|Y}\right] \\
&= E_{\pi}E\left[C_{XX|Y}k(.,Y)\right] \\
&= C_{XX|Y}\mu_Y^{\pi}
\end{aligned}$$

Finally we get

$$\begin{aligned}
\mu_{Y|x} &= C_{YX}^{\pi}C_{XX}^{\pi-1}k_x(.,x) \\
&= (C_{XY}^{\pi})^{\top}C_{XX}^{\pi-1}k_x(.,x)
\end{aligned}$$

where

$$\begin{aligned}
C^{\pi}_{XY} &= C_{X|Y}C^{\pi}_{YY} \\
&= C_{XY}C^{-1}_{YY}C^{\pi}_{YY} \\
\text{or} &= C_{XY|Y}\mu^{\pi}_{Y} \\
&= C_{(XY)Y}C^{-1}_{YY}\mu^{\pi}_{Y} \\
C^{\pi}_{XX} &= C_{XX|Y}\mu^{\pi}_{Y} \\
&= C_{(XX)Y}C^{-1}_{YY}\mu^{\pi}_{Y}
\end{aligned}$$

Interpretation The operators C^{π}_{XY} and C^{π}_{XX} have simple interpretation when considered as embeddings.

C^{π}_{XX} corresponds to the embedding of the law $Q(X) = \int P(X|Y)\pi(dY)$ into the tensorial product $\mathcal{H}_x \otimes \mathcal{H}_x$. C^{π}_{XY} is the embedding into $\mathcal{H}_x \otimes \mathcal{H}_y$ of $Q(X,Y) = P(X|Y)\pi(Y)$. Thus C^{π}_{XX} can be seen as the embedding of the sum rule, and is thus called **kernel sum rule**, whereas C^{π}_{XY} is the embedding of the chain rule, and is thus called **kernel chain rule**. Obviously, Bayesian manipulation are a succession of applications of these rules.

It remains now to get estimators for all these operators!

8.2 Estimators

We will use

$$\begin{aligned}
\mu_{Y|x} &= C^{\pi}_{YX}C^{\pi-1}_{XX}k_x(.,x) \\
C^{\pi}_{XY} &= C_{(XY)Y}C^{-1}_{YY}\mu^{\pi}_{Y} \\
C^{\pi}_{XX} &= C_{(XX)Y}C^{-1}_{YY}\mu^{\pi}_{Y}
\end{aligned}$$

The last two operators are seen as linear operator into respectively $\mathcal{H}_x \otimes \mathcal{H}_y$ and $\mathcal{H}_x \otimes \mathcal{H}_x$.

Let us find an estimator for C^{π}_{XY}. The other will be immediate.

We suppose to have an estimator for the function μ^{π}_{Y} which is written

$$\mu^{\pi}_{Y}(.) = \sum_{i=1}^{N_\pi} \gamma_i k_y(., Y^{\pi}_i)$$

where the upperscript is placed to remember that those Y are distributed according to the prior.

We suppose to have a training sample $\{X_i, Y_i\}, i = 1, \dots, N$ distributed according to the joint P. We denote by $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ the Gram matrices evaluated on this sample.

To evaluate an estimate of C^{π}_{XY}, first we find an estimate of $C^{-1}_{YY}\mu^{\pi}_{Y}$. It is a function in $\mathcal{H}_y$ and we look for it in the form $\sum_i \beta_i k_y(., Y_i) + \perp$. Further we use the regularized version of the inverse.

Then we have

$$\begin{aligned}\mu^\pi_Y(.) &= (C_{YY}+\lambda I)(\sum_i \beta_i k_y(.,Y_i) + \perp)\\ &= \sum_{ij}\beta_i \boldsymbol{K}_{y,ij} k_y(.,Y_j) + \lambda\sum_i \beta_i k_y(.,Y_i) + \lambda \perp\\ \mu^\pi_Y(Y_k) &= \sum_{ij}\beta_i \boldsymbol{K}_{y,ij}\boldsymbol{K}_y(Y_k,Y_j) + \lambda\sum_i \beta_i k_y(Y_k,Y_i)\end{aligned}$$

Let $\boldsymbol{\mu}^\pi_Y$ the vector containing the $\mu^\pi_Y(Y_k)$. We then have $\boldsymbol{\mu}^\pi_Y = (\boldsymbol{K}_y + \lambda I)\boldsymbol{K}\boldsymbol{\beta}$. Then applying $C_{(XY)Y}$ to $C_{YY}^{-1}\mu^\pi_Y$ we have

$$\begin{aligned}C^\pi_{XY} &= \sum_{ij}\beta_i \boldsymbol{K}_{y,ij} k(.,X_j)\otimes k(.,Y_j)\\ &= \sum_j \mu_j k(.,X_j)\otimes k(.,Y_j) \text{ where}\\ \boldsymbol{\mu} &= (\boldsymbol{K}_y+\lambda I)^{-1}\boldsymbol{\mu}^\pi_Y\end{aligned}$$

Likewise

$$\begin{aligned}C^\pi_{XX} &= \sum_j \mu_j k(.,X_j)\otimes k(.,X_j) \text{ where}\\ \boldsymbol{\mu} &= (\boldsymbol{K}_y+\lambda I)^{-1}\boldsymbol{\mu}^\pi_Y\end{aligned}$$

To get an estimate for $\mu_{Y|x}$ note that $C^\pi_{YX} = C^{\pi,\top}_{XY} = \sum_j \mu_j k(.,Y_j)\otimes k(.,X_j)$.

Since C^π_{XX} is note insured to be positive definite, use the regularization $(C^2+\varepsilon I)^{-1}C$ as the inverse. Doing as above, searching for $\mu_{Y|x}(.) = \sum_j \zeta_j(x) k_y(.,Y_j)$, we end up we the vector

$$\begin{aligned}\boldsymbol{\zeta}(x) &= \Lambda\left((\boldsymbol{K}_x\Lambda)^2+\varepsilon I\right)^{-1}\boldsymbol{K}_x\Lambda\boldsymbol{k}_X(x)\\ &= \Lambda\boldsymbol{K}_x\left((\boldsymbol{K}_x\Lambda)^2+\varepsilon I\right)^{-1}\Lambda\boldsymbol{k}_X(x)\end{aligned}$$

where $\boldsymbol{k}_X(x) = (k_x(X_1,x),\dots,k_x(X_N,x))^\top$ and $\Lambda = \mathrm{Diag}(\boldsymbol{\mu})$.

8.3 Another stuff on estimators

More generally, the function

$$g(.) = C_{XY}C_{YY}^{-1}f(.)$$

is estimated by

$$\begin{aligned}\hat g(.) &= \sum_i \boldsymbol{\mu}_i k_x(.,X_i) \quad \text{where}\\ \boldsymbol{\mu} &= (\boldsymbol{K}_y+\lambda \boldsymbol{I})^{-1}\boldsymbol{f} \quad \text{with } \boldsymbol{f} = (f(Y_1),\dots,f(Y_N))^\top\end{aligned}$$

or

8.4 Application in filtering

Suppose we want to estimate an hidden state x_k from past observation $y_{1:k}$. Assuming the state is Markovian and the observation conditionally white, the solution of the problem is given by the well-known recursion

$$p(x_k|y_{1:k}) = \frac{p(y_k|x_k)p(x_k|y_{1:k-1})}{\int p(y_k|x_k)p(x_k|y_{1:k-1})dx_k}$$

which is nothing but the Bayes law where the prior is $p(x_k|y_{1:k-1})$. Thus we can use the kernel Bayes rules to realize this in RKHS.

Let $m_{z,k|l}$ be the embedding of $p(z_k|y_{1:l})$. Since $p(x_k|y_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|y_{1:k-1})dx_{k-1}$, $m_{x,k|k-1}$ can be obtained by applying the kernel sum rule to $p(x_k|x_{k-1})$ using the prior $p(x_{k-1}|y_{1:k-1})$ whose embedding is $m_{x,k-1|k-1}$ Thus we have

$$m_{x,k|k-1} = C_{x_k x_{k-1}} C^{-1}_{x_{k-1}x_{k-1}} m_{x,k-1|k-1}$$

Then we need to apply the sum rule for $p(y_k|y_{1:k-1}) = \int p(y_k|x_k)p(x_k|y_{1:k-1})dx_k$ and the kernel chain rule for $p(y_k|x_k)p(x_k|y_{1:k-1})$. The application of both will result in the embedding of $m_{x,k|k}$

The embedding $m_{y,k|k-1}$ of $p(y_k|y_{1:k-1})$ into $\mathcal{H}$ satisfies, according to the sum rule

$$m_{y,k|k-1} = C_{y_k x_k} C^{-1}_{x_k x_k} m_{x,k|k-1}$$

whereas its embedding into $\mathcal{H} \otimes \mathcal{H}$is given by

$$c_{yy,k|k-1} = C_{y_k y_k x_k} C^{-1}_{x_k x_k} m_{x,k|k-1}$$

The chain rule gives

$$c_{x,k|k} = C_{y_k x_k} C^{-1}_{x_k x_k} m_{x,k|k-1}$$

and we finally get

$$m_{x,k|k} = c_{x,k|k} c^{-1}_{yy,k|k-1} k_y(.,y_k)$$

We need matrix representation for all these rules. To this aim, we suppose to have access to $N+1$ samples of the couple (X_k, Y_k). At time $k-1$ we suppose that the kernel conditional mean is given by

$$m_{x,k-1|k-1}(.) = \sum_{i=1}^{N} \alpha_i^{k-1} k_x(.,X_i) = \boldsymbol{k}_X(.)\boldsymbol{\alpha}^{k-1}$$

Thus, we have

$$m_{x,k|k-1}(.) \quad = \quad \boldsymbol{k}_{X+}(.)(\boldsymbol{K}_x + \lambda I)^{-1}\boldsymbol{K}_x\boldsymbol{\alpha}^{k-1}$$

where $\boldsymbol{k}_{X+}(.) = (k_x(.,X_2),\dots,k_x(.,X_{N+1}))$ and $\boldsymbol{K}_x$ is the Gram matrix built on $X_1,\dots,X_N$.

We then have

$$\begin{aligned}
c_{x,k|k} &= C_{y_k x_k} C_{x_k x_k}^{-1} m_{x,k|k-1} \\
&= \boldsymbol{k}_Y(.)(\boldsymbol{K}_x + \lambda I)^{-1} \big(m_{x,k|k-1}(X_1), \dots m_{x,k|k-1}(X_N)\big)^\top \\
&= \boldsymbol{k}_Y(.)(\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_{XX+} (\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_x \boldsymbol{\alpha}^{k-1} \\
&= \sum_{i=1}^{N} \boldsymbol{\mu}_i^k k_y(., Yi) \quad \text{where} \\
\boldsymbol{\mu}^k &= (\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_{XX+} (\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_x \boldsymbol{\alpha}^{k-1}
\end{aligned}$$

We also get

$$c_{yy,k|k-1} = \sum_i \boldsymbol{\mu}_i^k k_y(., Yi) \otimes k_y(., Yi)$$

and therefore we finally get

$$\boldsymbol{\alpha}^k \quad = \quad \Lambda^k \boldsymbol{K}_y \Big((\boldsymbol{K}_y \Lambda^k)^2 + \varepsilon I \Big)^{-1} \Lambda^k \boldsymbol{k}_Y(y_k)$$

where $\Lambda^k = \mathrm{Diag}(\boldsymbol{\mu}^k)$.

Let us synthesize: the kernel conditional mean is given by

$$m_{x,k|k}(.) = \sum_{i=1}^{N} \alpha_i^k k_x(., X_i) = \boldsymbol{k}_X(.) \boldsymbol{\alpha}^k$$

where the vector $\boldsymbol{\alpha}^k$ satisfies the recursion

$$\begin{aligned}
\boldsymbol{\mu}^k &= (\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_{XX+} (\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_x \boldsymbol{\alpha}^{k-1} \\
\Lambda^k &= \mathrm{Diag}(\boldsymbol{\mu}^k) \\
\boldsymbol{\alpha}^k &= \Lambda^k \boldsymbol{K}_y \Big((\boldsymbol{K}_y \Lambda^k)^2 + \varepsilon I \Big)^{-1} \Lambda^k \boldsymbol{k}_Y(y_k)
\end{aligned}$$

Note that the matrix $(\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_{XX+} (\boldsymbol{K}_x + \lambda I)^{-1} \boldsymbol{K}_x$ can obviously be pre-computed. To initialize, we can use $\widehat{\pi}(x_1) = E[k_x(., x_1)] = C_{xy} C_{yy}^{-1} k(., y_1)$ and thus $\boldsymbol{\alpha}^1 = (\boldsymbol{K}_y + \lambda I)^{-1} \boldsymbol{k}_Y(y_1)$.

The outcome of the algorithm is an estimation of the embedding of the *a posteriori*measure. If we want an estimate of $E[f(x_k)|y_1, \dots, y_k]$ where $f \in \mathcal{H}_x$ we simply use the definition $E[f(x_k)|y_1, \dots, y_k] = \big\langle f|x, m_{k|k} \big\rangle$. However, if f does not belong to the RKHS we are in trouble. Another possibility is to find the pre-image x^k whose image $k(., x^k)$ is the closest to the embedding of the posterior. For radial kernel $k(., x) = f(\|. - x\|^2)$ this can be solved efficiently if closeness is measured using the RKHS norm. Indeed, searching for $\min_x \|k_x(., x) - \sum_i k_x(., X_i) \alpha_i^t$ lead to the fixed point condition $x = \sum_i X_i f'(\|x - X_i\|^2) \alpha_i^t / \sum_i f'(\|x - X_i\|^2) \alpha_i^t$. A solution can be obtained sequentially as

$$x_n^t = \frac{\sum_i X_i f'(\|x_{n-1}^t - X_i\|^2) \alpha_i^t}{\sum_i f'(\|x_{n-1}^t - X_i\|^2) \alpha_i^t}$$

9 The Bayesian point of view