

# Hilbert Embeddings

(Inmersiones en espacios de Hilbert)

Objetivo: mapeo de distribuciones a espacios de características utilizando kernels, comparando y manipulando distribuciones mediante operaciones en el espacio de características.

Operaciones básicas en probabilidades:

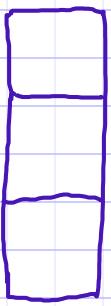
$$Q(x) = \sum_y P(x,y) = \sum_y P(x|y)P(y) \rightarrow \text{SUMA}$$

$$Q(x,y) = P(x|y)P(y) \rightarrow \text{PRODUCTO}$$

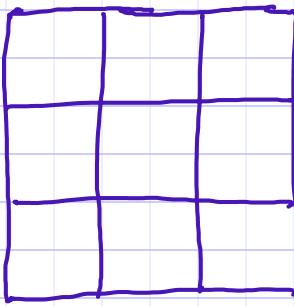
$$Q(y|x) = \frac{P(x|y)P(y)}{P(x)} \rightarrow \text{REGLA DE BAYES}$$

Caso discreto:

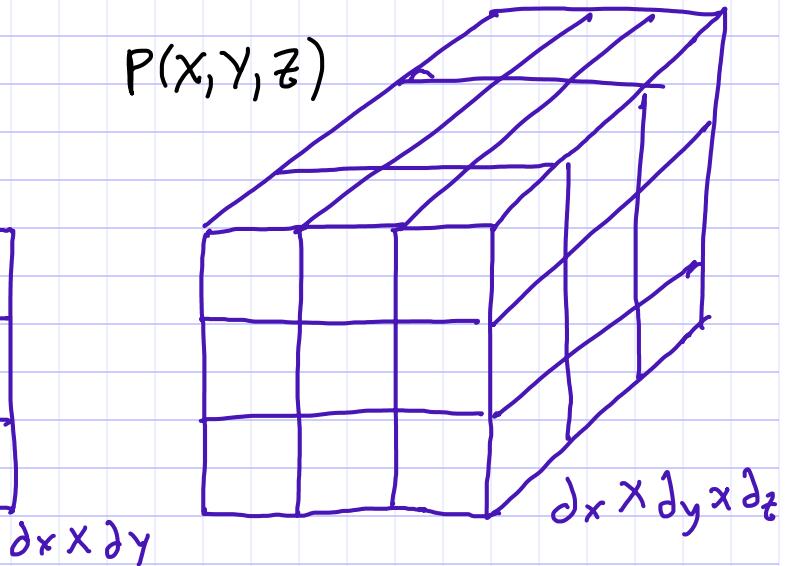
$$P(x)$$



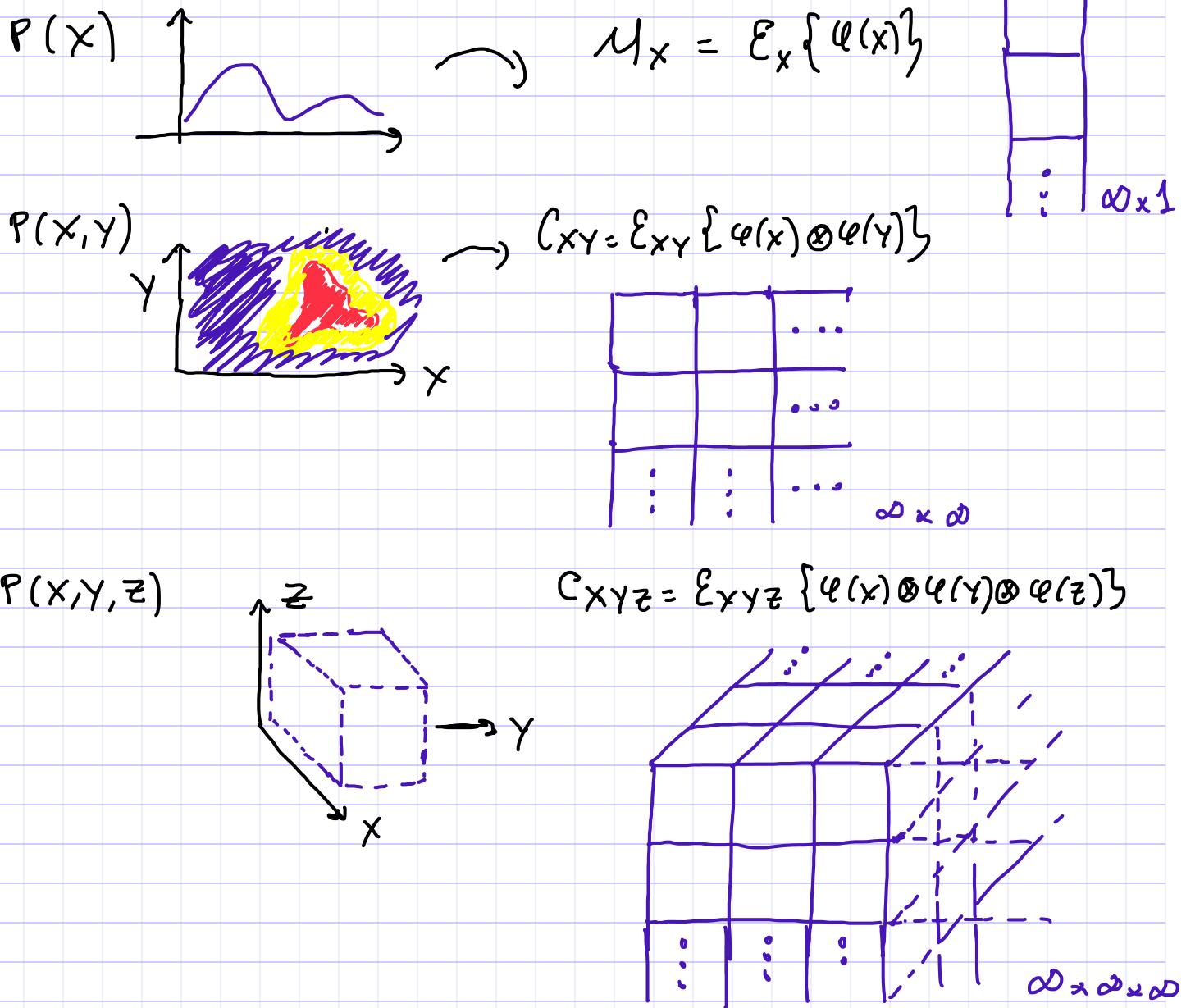
$$P(x,y)$$



$$P(x,y,z)$$



# Kernel Embedding

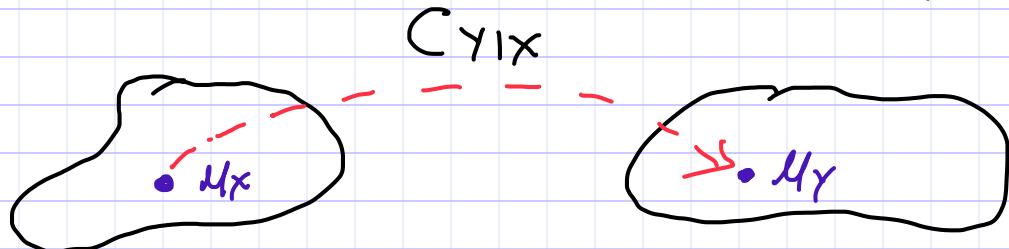


$$M_x = C_{X|Y} M_y \rightarrow \text{SUMA}$$

$$C_{XY} = C_{Y|X} C_{XX} \rightarrow \text{PRODUCTO}$$

$$M_{Y|X} = C_{Y|X} \varphi(x) \rightarrow \text{BAYES}.$$

$|$   
 $|$   $C_{Y|X}$ : punto en  
 $|$  RKHS / OPERADOR  
 $|$  ENTRE RKHSs.



## Esperanza desde productos internos.

→ En espacios de Hilbert de dimensión finita, el operador esperanza se puede definir como:

$$\text{Ej: } \varphi(x) = k(\cdot, x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}; \quad f(\cdot) = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}} = [a \ b] \begin{bmatrix} x \\ x^2 \end{bmatrix} = ax + bx^2$$

$$x \in \mathcal{X}; \quad \varphi(x) \in \mathcal{F}; \quad f: \mathcal{X} \rightarrow \mathcal{F}.$$

→ Sea  $x \sim P$  una variable aleatoria con función de distribución  $P$ .

$$\mathbb{E}_P \{ f(x) \} = \mathbb{E}_P \{ \langle f, \varphi(x) \rangle_{\mathcal{F}} \} = \langle f, \mathbb{E}_P \{ \varphi(x) \} \rangle_{\mathcal{F}}$$

$$\mathbb{E}_P \{ \varphi(x) \} = \int \varphi(x) dP(x)$$

$$\mathbb{E}_P \{ f(x) \} = \langle f, \mu_P \rangle_{\mathcal{F}}$$

Para el ejemplo anterior:

$$\mathbb{E}_P \{ f(x) \} = [a \ b] \begin{bmatrix} \mathbb{E}_P \{ x \} \\ \mathbb{E}_P \{ x^2 \} \end{bmatrix}$$

Qué pasa en espacios de infinita dimensión?

→ Un operador lineal  $A: \mathcal{F} \rightarrow \mathbb{R}$  es acotado cuando:

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}}; \quad \forall f \in \mathcal{F}$$

**Teorema de Riesz**: En un espacio de Hilbert  $\mathcal{F}$ , todos los operadores lineales acotados  $A$  se pueden expresar como  $\langle \cdot, g_A \rangle_{\mathcal{F}}$ ;  $g_A \in \mathcal{F}$ :

$$Af = \langle f, g_A \rangle_{\mathcal{F}}.$$

**Existencia de la inmersión media**: Si

$$\mathbb{E}_P \{ (K(x,x))^{\frac{1}{2}} \} < \infty, \text{ entonces } M_P \in \mathcal{F}.$$

El operador lineal  $T_P f = \mathbb{E}_P \{ f(x) \}$ ;  $\forall f \in \mathcal{F}$ , es acotado, dado que:

$$|T_P f| = |\mathbb{E}_P \{ f(x) \}| \leq \mathbb{E}_P \{ |f(x)| \} = \mathbb{E}_P \{ |\langle f, \ell(x) \rangle_{\mathcal{F}}| \}$$

$\downarrow$

Desigualdad de Jensen.

**TAREA**: Probar desigualdad de Jensen.

$$\mathbb{E}_P \{ |\langle f, \ell(x) \rangle_{\mathcal{F}}| \} \leq \mathbb{E}_P \{ \|f\|_{\mathcal{F}} \|\ell(x)\|_{\mathcal{F}} \} = \mathbb{E}_P \{ \sqrt{K(x,x)} \|f\|_{\mathcal{F}} \}$$

$$|T_P f| \leq \mathbb{E}_P \{ \sqrt{K(x,x)} \|f\|_{\mathcal{F}} \}$$

Por ende existe un  $M_P \in \mathcal{F}$  tal que  $T_P f = \langle f, M_P \rangle_{\mathcal{F}}$

Si:  $f(x) = \ell(x) = K(x, \cdot)$ :

$$M_P(x) = \langle M_P, K(x, \cdot) \rangle_{\mathcal{F}} = \mathbb{E}_P \{ K(x, x') \}.$$

## Mean Embedding (Inmersión media)

Una inmersión núcleo (Kernel embedding) representa una distribución de probabilidad como un elemento en RKHS :

$$M_x = \mathbb{E}_P \{ \varphi(x) \} = \int_{\mathcal{X}} \varphi(x) dP(x) ; \quad x \in \mathcal{X}.$$

$\varphi : \mathcal{X} \rightarrow \mathcal{F}$  ;  $\mathcal{F}$  puede tener dimensión infinita.

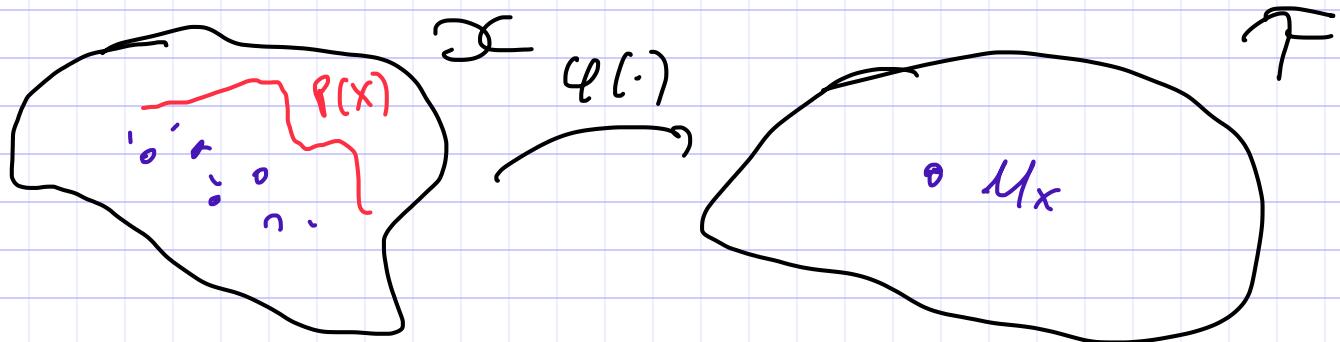
**NOTA:** El mean embedding es diferente a la estimación de densidad por kernels; por ejemplo

Parten:  $\hat{p}(x) = \frac{1}{N} \sum_{i=1}^N \hat{k}_{\sigma}(x, x_i)$ .

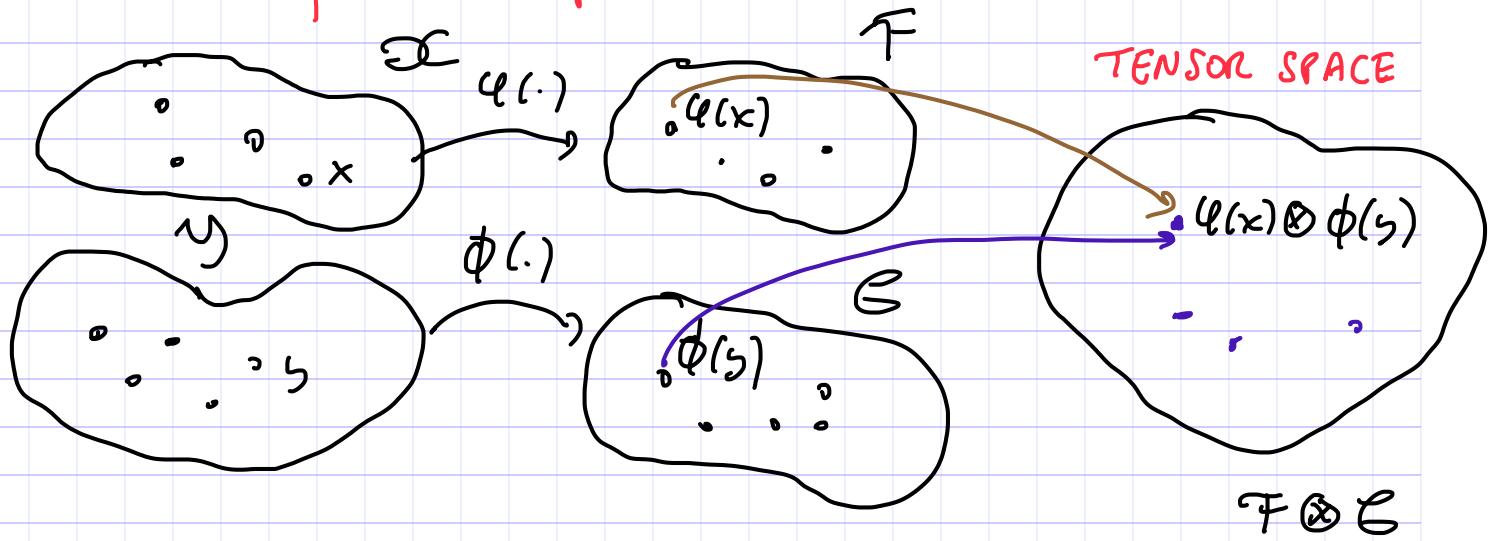
En general  $\hat{k}(x, x_i) \neq \langle \varphi(x), \varphi(x_i) \rangle$ .

En caso de ser kernel; su ancho de banda cambia dependiendo del número de puntos  $N$ .

\* No es una inmersión de distribución a un espacio fijo de características.



# Inmersiones a distribuciones conjuntas. (Tensor product spaces).



$\text{Si } X, Y \in \mathcal{X};$  la inmersión conjunta al espacio tensor  $F \otimes F$  se define como :

$$C_{XY} = \sum_{x,y} \{\varphi(x) \otimes \varphi(y)\}^y = \int_{\mathcal{X} \times \mathcal{X}} \varphi(x) \otimes \varphi(y) dP(x,y)$$

Para el caso anterior  $F = \mathcal{G}$ ;  $\varphi = \phi$ .

**NOTA:** El espacio tensor desde RKHS<sub>S</sub> da origen al operador de covarianza  $\Rightarrow$  análogo a la matriz de covarianza en espacios de infinita dimensión.

## Operadores de Hilbert-Schmidt

Sea  $\mathcal{F}$  y  $\mathcal{G}$  dos espacios de Hilbert separables; sea  $\{e_i \in \mathcal{F}\}_{i=1}^I$  y  $\{f_j \in \mathcal{G}\}_{j=1}^J$  dos conjuntos de bases ortonormales.  $I, J$  pueden ser  $\infty$ .

Sean  $L: \mathcal{G} \rightarrow \mathcal{F}$  y  $M: \mathcal{G} \rightarrow \mathcal{F}$  dos operadores lineales compactos; la norma Hilbert-Schmidt de  $L, M$  se define como:

$$\|L\|_{HS}^2 = \sum_{j \in J} \|Lf_j\|_{HS}^2 = \sum_{i \in I} \sum_{j \in J} |\langle Lf_j, e_i \rangle_{\mathcal{F}}|^2$$

$L$  es Hilbert-Schmidt si  $\|L\|_{HS}^2 < \infty$

→ El producto interno entre los operadores  $L, M$ :

$$\langle L, M \rangle_{HS} = \sum_{j \in J} \langle Lf_j, Mf_j \rangle_{\mathcal{F}} ; f_j \in \mathcal{G}.$$

Desde  $e_i \in \mathcal{F}$ :

$$\langle L, M \rangle_{HS} = \sum_{i \in I} \sum_{j \in J} \langle Lf_j, e_i \rangle_{\mathcal{F}} \langle Mf_j, e_i \rangle_{\mathcal{F}}$$

**Prueba:** Representando las bases ortonormales:

$$Lf_j = \sum_{i \in I} \alpha_i^j e_i ; Mf_j = \sum_{i' \in I} \beta_{i'}^j e_{i'}$$

$$\begin{aligned} \langle L, M \rangle_{HS} &= \sum_{j \in J} \langle Lf_j, Mf_j \rangle_{\mathcal{F}} = \sum_j \left( \sum_i \alpha_i^j e_i \sum_{i'} \beta_{i'}^j e_{i'} \right) \\ &= \sum_j \sum_i \alpha_i^j \beta_i^j ; e_i e_{i'} = \begin{cases} 1 & i = i' \\ 0 & i \neq i' \end{cases} \end{aligned}$$

## Operadores de rango 1 → Tensor product

Sea  $b \in \mathcal{G}$  y  $a \in \mathcal{F}$ ; el producto tensor  $a \otimes b$  se define como el operador de rango 1:

$$(b \otimes a)f \mapsto \langle f, a \rangle_{\mathcal{F}} b$$

Para comprobar si es HS;  $\|a \otimes b\|_{HS}^2 < \infty$ :

$$\begin{aligned} \|a \otimes b\|_{HS}^2 &= \sum_{j \in J} \|(a \otimes b)f_j\|_{HS}^2 = \sum_{j \in J} \|a \langle b, f_j \rangle_{\mathcal{G}}\|_{HS}^2 \\ &= \|a \sum_{j \in J} \langle b, f_j \rangle_{\mathcal{G}}\|_{HS}^2 = \|a\|_{\mathcal{F}}^2 \sum_j |\langle b, f_j \rangle_{\mathcal{G}}|^2 \end{aligned}$$

$$\boxed{\|a \otimes b\|_{HS}^2 = \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2}; \quad f_j \in \mathcal{G}$$

$$\begin{aligned} \text{NOTA: } \|b\|_{\mathcal{G}}^2 &= \langle b, b \rangle_{\mathcal{G}} = \left\langle \sum_j \alpha_j f_j, \sum_j \alpha_j f_j \right\rangle_{\mathcal{G}} \\ &= \sum_j \langle \alpha_j, \alpha_j \rangle = \sum_j |\alpha_j|^2 \end{aligned}$$

$$\alpha_j^* = \arg \min_{\alpha_j} \|b - \sum_j \alpha_j f_j\|_{\mathcal{G}}^2$$

$$\text{s.t. } \langle f_j, f_j \rangle_{\mathcal{G}} = \begin{cases} 1 & j=j' \\ 0 & j \neq j' \end{cases}$$

$$\alpha_j^* = \frac{\langle b, f_j \rangle_{\mathcal{G}}}{\|f_j\|_{\mathcal{G}}^2} = \langle b, f_j \rangle_{\mathcal{G}}.$$

Ahora: para  $L: \mathcal{G} \rightarrow \mathcal{F}$ ;  $a \in \mathcal{F}$ ;  $b \in \mathcal{G}$ .

$$\langle L, a \otimes b \rangle_{HS} = \langle a, Lb \rangle_{\mathcal{F}}$$

En particular:

$$\langle u \otimes v, a \otimes b \rangle_{HS} = \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}}; \quad u, a \in \mathcal{F} \\ b, v \in \mathcal{G}.$$

**Prueba:** Expandiendo  $b$  en sus bases ortonormales

$$b = \sum_j \alpha_j f_j = \sum_j \langle b, f_j \rangle_{\mathcal{G}} f_j$$

$$\begin{aligned} \langle L, a \otimes b \rangle_{HS} &= \langle a, Lb \rangle_{\mathcal{F}} \\ &= \langle a, L \left( \sum_j \langle b, f_j \rangle_{\mathcal{G}} f_j \right) \rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j \rangle_{\mathcal{G}} \langle a, Lf_j \rangle_{\mathcal{F}} \end{aligned}$$

$$\langle L, a \otimes b \rangle_{HS} = \langle a \otimes b, L \rangle_{HS} = \sum_j \langle b, f_j \rangle_{\mathcal{G}} \langle a, Lf_j \rangle_{\mathcal{F}}$$

$$\text{Si } u \otimes v = L \quad \hookrightarrow \quad (b \otimes a) f \mapsto \langle f, a \rangle_{\mathcal{F}} b$$

$$\begin{aligned} \langle u \otimes v, a \otimes b \rangle_{HS} &= \langle a, (u \otimes v) b \rangle_{\mathcal{F}}; \quad a, u \in \mathcal{F} \\ &= \langle a, u \langle b, v \rangle_{\mathcal{G}} \rangle_{\mathcal{F}} \\ &= \langle b, v \rangle_{\mathcal{G}} \langle a, u \rangle_{\mathcal{F}} \end{aligned}$$

$$\boxed{\langle u \otimes v, a \otimes b \rangle_{HS} = \langle b, v \rangle_{\mathcal{G}} \langle a, u \rangle_{\mathcal{F}}}$$

## Operador de covarianza cruzada

Sean  $F, G \in \mathcal{RHS}_S$ ,  $\varphi : \mathcal{X} \rightarrow F$   
 $\psi : \mathcal{X} \rightarrow G$

En  $\mathcal{X}$ ; para  $x, y \in \mathcal{X}$ :

$$C_{xy} = \mathbb{E}_{xy}\{xy^T\}; \quad F^T C_{xy} g = \mathbb{E}_{xy}\{(F^T x)(g^T x)\}$$

Para el caso centrado:

$$\tilde{C}_{xy} = C_{xy} - \mu_x \mu_y^T; \quad \mathbb{E}_x\{x\} = \mu_x$$

$\uparrow$

$$\mathbb{E}_y\{y\} = \mu_y$$

TAREA: Demostrar

Ahora, sea el producto cruzado  $\varphi(x) \otimes \psi(y) \in HS$ ; es decir  $\varphi(x) \otimes \psi(y) \in F \otimes G$

→ (a forma lineal  $\langle \varphi(x) \otimes \psi(y), A \rangle_{HS}$  es  
 medible; dado que  $\|A\|_{HS} < \infty$ ; incluso  
 para dimensiones infinitas..)

→ Se requiere que  $\sum_{x,y} \{\|\varphi(x) \otimes \psi(y)\|_{HS}\} < \infty$

→ Por ende se requiere que:

$$\langle C_{xy}, A \rangle_{HS} = \mathbb{E}_{xy} \left\{ \langle \varphi(x) \otimes \psi(y), A \rangle_{HS} \right\}$$

Prueba:  $T_{xy} : \mathcal{H}\mathcal{S}(G, F) \rightarrow \mathbb{R}$

$$A \mapsto \sum_{x,y} \{ \langle \varphi(x) \otimes \phi(y), A \rangle_{\mathcal{H}\mathcal{S}} \}$$

$T_{xy}$  es acotado si  $\sum_{x,y} \{ \| \varphi(x) \otimes \phi(y) \| \}_{\mathcal{H}\mathcal{S}} < \infty$

Aplicando Jensen:  $(|\sum \{ x \}| \leq \sum_{x,y} \{ |x| \})$   
 $|\sum_{x,y} \{ \langle \varphi(x) \otimes \phi(y), A \rangle_{\mathcal{H}\mathcal{S}} \}| \leq \sum_{x,y} \{ |\langle \varphi(x) \otimes \phi(y), A \rangle_{\mathcal{H}\mathcal{S}}| \}$

Aplicando Cauchy:  $(\langle x, y \rangle \leq \|x\| \|y\|)$

$$|\sum_{x,y} \{ \langle \varphi(x) \otimes \phi(y), A \rangle_{\mathcal{H}\mathcal{S}} \}| \leq \|A\|_{\mathcal{H}\mathcal{S}} \sum_{x,y} \{ \| \varphi(x) \otimes \phi(y) \| \}$$

$$\|A\|_{\mathcal{H}\mathcal{S}} < \rho; \quad y :$$

$$\begin{aligned} \sum_{x,y} \{ \| \varphi(x) \otimes \phi(y) \| \} &= \sum_{x,y} \{ \| \varphi(x) \|_F \| \phi(y) \|_E \} \\ &= \sum_{x,y} \{ \sqrt{k_x(x,x) k_y(y,y)} \} < \infty. \end{aligned}$$

Por ende  $C_{xy}$  existe.

Para un elemento  $f \otimes g \in \mathcal{H}\mathcal{S}$ :

$$\begin{aligned} \langle f, C_{xy} g \rangle_F &= \langle C_{xy}, f \otimes g \rangle_{\mathcal{H}\mathcal{S}} \\ &= \sum_{x,y} \{ \langle \varphi(x) \otimes \phi(y), f \otimes g \rangle_{\mathcal{H}\mathcal{S}} \} \end{aligned}$$

$$\langle f, C_{xy} g \rangle = \langle C_{xy}, f \otimes g \rangle$$

$$= \sum_{x,y} \{ \langle f, \varphi(x) \rangle_F \langle g, \phi(y) \rangle_E \}$$

$$= \sum_{x,y} \{ f(x) g(y) \} = \text{cov}(f, g).$$

Estimador Básico de  $C_{xy}$ :

Dado el conjunto i.i.d  $\{x_i, y_i \in \mathcal{X}\}_{i=1}^N$ :

$$\hat{C}_{xy} = \frac{1}{N} \sum_{i=1}^N \varphi(x_i) \otimes \phi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y$$

$$\hat{\mu}_x = \frac{1}{N} \sum_{i=1}^N \varphi(x_i); \quad \hat{\mu}_y = \frac{1}{N} \sum_{i=1}^N \phi(y_i)$$

$$\hat{C}_{xy} = \frac{1}{N} \Phi_x H \Phi_y^T$$

TAREA: Demostrar.

$$H = I - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$$

$$\Phi_x = [\varphi(x_i)]_{i=1}^N$$

$$\Phi_y = [\phi(y_i)]_{i=1}^N; \quad K_{ij} = [\Phi_x^T \Phi_y]_{ij}$$

$$L_{ij} = [\Phi_y^T \Phi_y]_{ij}$$

$$\tilde{K} = H K H; \quad \tilde{L} = H L H$$

## Aplicaciones:

### 1. Medida de dependencia.

$$\max_{f, g} \langle g, \hat{C}_{xy} f \rangle_G$$

$$\text{s.t. } \|f\|_F = 1$$

$$\|g\|_G = 1$$

TAREA: Resolver por lagrangiano y comparar con KCCA.

### 2. Criterio de independencia.

HSIC: Hilbert-Schmidt independence criterion.

$$HSIC(F, G, P_{xy}) = \|C_{xy} - M_x \otimes M_y\|_{HS}^2$$

$$= \|C_{xy}\|_{HS}^2 - 2 \langle C_{xy}, M_x \otimes M_y \rangle_{HS} - \|M_x \otimes M_y\|_{HS}^2$$

$$\begin{aligned} \|C_{xy}\|_{HS}^2 &= \langle C_{xy}, C_{xy} \rangle = \sum_{x,y} \{ \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle \} \\ &= \sum_{x,y} \sum_{x'y'} \{ \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle \} \\ &= \sum_{x,y} \sum_{x'y'} \{ \langle \varphi(x), \varphi(x') \rangle_F \langle \phi(y), \phi(y') \rangle_G \} \\ &= \sum_x \sum_{x'y'} \{ K_x(x, x') K_y(y, y') \} = A \end{aligned}$$

$$\begin{aligned}
\langle M_x \otimes M_y, M_x \otimes M_y \rangle_{H^2} &= \|M_x \otimes M_y\|_{H^2}^2 \\
&= \langle M_x, M_x \rangle_F \langle M_y, M_y \rangle_F \\
&= \sum_{x \in X} \left\{ K_x(x, x) \right\} \sum_{y \in Y} \left\{ K_y(y, y) \right\} \\
&= P
\end{aligned}$$

$$\begin{aligned}
\langle C_{xy}, M_x \otimes M_y \rangle_{H^2} &= \sum_{x \in X} \left\{ \left\langle \ell(x) \otimes \phi(y), M_x \otimes M_y \right\rangle \right\} \\
&= \sum_{x \in X} \left\{ \langle \ell(x), M_x \rangle_F \langle \phi(y), M_y \rangle_F \right\} \\
&= \sum_{x \in X} \left\{ \sum_{x' \in X} \left\{ K_x(x, x') \right\} \sum_{y' \in Y} \left\{ K_y(y, y') \right\} \right\} \\
&= B
\end{aligned}$$

$$HSIC^2(F, G, X, Y) = A - 2B + D.$$

**TAREA:** Comparar HSIC con CKA. Y escribir HSIC en forma matricial.

**NOTA:** En estimadores :  $mse(\hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$

$$\begin{aligned}
&= b^2(\hat{\theta}) + \text{var}(\hat{\theta})
\end{aligned}$$

Siendo  $\hat{\theta}$  estimador de  $\theta$ , y :

$$b(\hat{\theta}) = \mathbb{E}\{\hat{\theta}\} - \theta \rightarrow \text{sesgo}$$

$$\text{var}(\hat{\theta}) = \mathbb{E}\{(\hat{\theta} - \mathbb{E}\{\hat{\theta}\})^2\} \rightarrow \text{varianza.}$$

**TAREA:** Consulte el estimador sin sesgo de HSIC



