Kernel Methods for Comparing Probability Distributions

Problem: testing for differences and/or dependences b/w r.v.'s in high dimensional spaces ("curse of dimensionality") with limited samples is very difficult

By working in ∞-dimensional spaces, Kernel methods allow us to do this relatively well for arbitrary data structures (e.g. discrete/continuous)

Key: Mean embedding:

"Kernel trick": $f(x) = \langle f, \phi_x \rangle_{\mathcal{F}}$

$\Rightarrow$ "mean trick": $\underbrace{\mathbb{E}_{x \sim p}[f(x)]}_{} = \langle \mu_p, f \rangle_{\mathcal{F}}$

you can prove μ_p exists via the Riesz theorem

> More generally,
$\mu_p(x) = \langle \mu_p, \phi_x \rangle_{\mathcal{F}}$
$= \langle \mu_p, k(x, \cdot) \rangle_{\mathcal{F}}$
$\xrightarrow[\text{("mean trick")}]{\text{by definition}} = \mathbb{E}_{x \sim p} k(x, X)$
$(= k(x, \mathbb{E}_{x \sim p} X))$

Empirically, you can estimate μ_p via:

$$\hat{\mu}_p = \frac{1}{N} \sum_{i=1}^{N} \phi_{x_i} \quad , \quad x_i \overset{\text{i.i.d.}}{\sim} P$$

Note that $\hat{\mu}_p$ is an infinitely dimensional vector in the RKHS $\mathcal{F}$, i.e. it is a function! So, we can estimates of the mean in different regions of $\mathcal{F}$ by indexing via the function argument, Analagously to how we do this in finite space:

$$\text{For } \{x^{(i)}\} \subseteq \mathbb{R}^D, \quad \underset{\substack{\uparrow \\ \text{find mean on} \\ d\text{th dimension}}}{\bar{x}_d} = \frac{1}{N} \sum_i x_d^{(i)} = \frac{1}{N} \sum_i \underset{\substack{\uparrow \\ \text{projection onto} \\ d\text{th basis vector}}}{x^{(i)T} e_d} \qquad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} 1 \\ \\ \\ d \\ \\ D \end{matrix}$$

probability feature map

$$\text{For } \{x^{(i)}\} \subseteq \mathcal{X}, \quad \mu_p(d) = \frac{1}{N} \sum_i \langle \phi_{x^{(i)}}, \phi_d \rangle = \frac{1}{N} \underbrace{\sum_i k(x^{(i)}, d)}_{\text{expectation of a kernel!}}$$

$d \in \mathcal{X}$

We can formally compare means in feature space via the Maximum mean discrepancy:

$$\mathrm{MMD}(\overset{\text{probability distributions}}{P, Q}; \mathcal{F}) = \sup_{f \in \mathcal{F}} \left[\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right], \quad X \sim P,\ Y \sim Q$$

$$= \sup_{f \in \mathcal{F}} \left[\langle f, \mu_P \rangle_{\mathcal{F}} - \langle f, \mu_Q \rangle_{\mathcal{F}} \right]$$

→ MMD = 0 iff P = Q when $\mathcal{F}$ is the unit ball in a characteristic RKHS

$$= \sup_{f \in \mathcal{F}} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$

$$= \| \mu_P - \mu_Q \|_{\mathcal{F}}$$

But, of course, we never have access to P and Q, only samples. So, we estimate MMD as:

$$\mathrm{MMD}^2 = \| \mu_P - \mu_Q \|^2 = \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle$$

$$= \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2 \langle \mu_P, \mu_Q \rangle$$

$$\simeq \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} K(x_i, x_j) + \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i} K(y_i, y_j) - \frac{2}{NM} \sum_{i,j} K(x_i, y_j)$$

where $x_i \overset{\text{i.i.d.}}{\sim} P$, $y_i \overset{\text{i.i.d.}}{\sim} Q$

We now want to make sure that MMD does what we want: i.e. MMD = 0 iff P = Q. As mentioned above, this holds whenever $\mathcal{F}$ is a characteristic RKHS.

For a translation invariant and periodic kernel K, prove below that K is a characteristic kernel (with associated characteristic RKHS $\mathcal{F}$) iff $\forall \ell \ \hat{K}_\ell \neq 0$:

Supposing $p(x)$ is defined only over $x \in [-\pi, \pi]$,

$$\mu_P(z) = \mathbb{E}_{x \sim P} K(z, x)$$

$$= \mathbb{E}_{x \sim P} K(z - x) \qquad K \text{ is translation invariant}$$

$$= \int_{-\pi}^{\pi} K(z - x)\, dP(x) \qquad K \text{ is periodic}$$

The Fourier transform of μ_P is then given by:

$$\hat{\mu}_{P,l} = \iint_{-\pi}^{\pi} K(z-x) e^{-ilz} \, dP(x) \, dz$$

Let $v = z - x \iff dv = dz$

$$= \iint_{-\pi}^{\pi} K(v) e^{-il(x+v)} \, dP(x) \, dv$$

$$= \int K(v) e^{-ilv} \, dv \int_{-\pi}^{\pi} e^{-ilx} \, dP(x)$$

$$= \hat{K}_l \varphi_{P,l}$$

Recall now that for an RKHS $\mathcal{F}$ associated with a periodic and translation invariant kernel K, $\|f\|^2_{\mathcal{F}} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{K}_l}$

So, we now have

$$MMD^2 = \|\mu_P - \mu_Q\|^2_{\mathcal{F}} = \sum_{l=-\infty}^{\infty} \frac{|\hat{K}_l \varphi_{P,l} - \hat{K}_l \varphi_{Q,l}|^2}{\hat{K}_l}$$

$$= \sum_{l=-\infty}^{\infty} |\varphi_{P,l} - \varphi_{Q,l}|^2 \hat{K}_l$$

Two points:

1. Note that as long as $\hat{K}_l \neq 0$ for all l, $MMD^2 = 0$ only when $P = Q$, making K a characteristic kernel
2. Remember that $\hat{K}_l$ decays with l, so this definition of MMD can be interpreted as penalizing differences in lower frequencies more than in higher frequencies

~~For arbitrary kernel $K(x,y)$ and probability distribution $P(x)$ on $\mathbb{R}^D$, we use a similar proof by invoking~~

For an arbitrary translation invariant kernel $K(x-y)$ on $\mathbb{R}^D$, we invoke Bochner's theorem: $K(x,y) = K(x-y) = K(z) = \int_{\mathbb{R}^D} e^{-iz^T\omega} \, d\Lambda(\omega)$

$\Rightarrow K$ characteristic for prob. measure on $\mathbb{R}^D$ iff $\text{supp}(\Lambda) = \mathbb{R}^D$

Fourier transform of K

Having established MMD as a good metric of similarity b/w probability distributions, lets now think about how we can use it to make inferences from data via <u>hypothesis testing</u>:

$$H_0 : P = Q \qquad \text{given } \{x_i\}_{i=1}^N,\ x_i \overset{iid}{\sim} P$$
$$H_1 : P \neq Q \qquad \{y_j\}_{j=1}^M,\ y_j \overset{iid}{\sim} Q$$

We want to reject H_0 if MMD is far from zero:

We can then define this by picking a threshold criterion based on the asymptotic distribution $(N\to\infty)$ of $\widehat{MMD}^2$

$$\widehat{MMD}^2 = \frac{1}{N(N-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i),(x_j, y_j))}$$

for $N=M$, $N \to \infty$

* note that $\widehat{MMD}$ is missing some terms from our previous empirical estimate of MMD: $\widehat{MMD}$ is still an unbiased estimator, but no longer minimum variance

When $P \neq Q$,

$$\sqrt{N}\left(\widehat{MMD}^2 - MMD^2\right) \sim \mathcal{N}(0, \sigma_u^2)$$

$$\text{where } \sigma_u^2 = 4\left(\mathbb{E}_z\left[(\mathbb{E}_{z'} h(z,z'))^2\right] - \left[\mathbb{E}_{z,z'} h(z,z')\right]^2\right)$$

$$z := (x_i, y_i)$$

When $P = Q$,

$$N \cdot MMD \sim \sum_{l=1}^{\infty} \lambda_l (z_l^2 - 2)$$

this is an infinite sum of χ^2 distributions

$$\text{where } z_l \overset{iid}{\sim} \mathcal{N}(0,2)$$

$$\int_{\mathcal{X}} \underbrace{\tilde{k}(x,x')}_{\text{(centered)}} \phi_i(x)\, dP(x) = \lambda_i \phi_i(x')$$

$\Rightarrow$ this is a <u>degenerate</u> U-statistic, can't compute the <u>null</u> distribution!

Pearson moment matching — permutation methods

Can we use a similar kernel-based metric to test for <u>dependence</u> b/w two random variables?

If X, Y are independent, then

$$P(X,Y) = P(X)\,P(Y) \qquad (*)$$

Lets call these three distributions $P_{x,y}$, P_x, P_y, which are probability measures on $\mathcal{X}\times\mathcal{Y}$, $\mathcal{X}$, $\mathcal{Y}$ respectively.

We now move into arbitrary feature space by assuming two RKHS's $\mathcal{F}$ and $\mathcal{G}$ w/ associated kernels k and l such that $\mu_{P_x} \in \mathcal{F}$, $\mu_{P_y} \in \mathcal{G}$.

If $X \sim P_x$ and $Y \sim P_y$ are independent, then, for a function $f(x, y)$ with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ eqn. $(*)$ tells us that

$$\mathbb{E}_{P_{xy}} f = \mathbb{E}_{P_x P_y} f,$$ where f is a matrix (i.e. a Hilbert-Schmidt operator) in the Hilbert ~~[illegible]~~ space $\mathcal{F}\times\mathcal{G}$

$$\Leftrightarrow \mathbb{E}_{P_{xy}} f - \mathbb{E}_{P_x P_y} f = 0 \quad (**)$$

We can then use an analog of MMD ~~[illegible]~~ called the <u>Hilbert-Schmidt Independence Criterion</u> (HSIC):

$$\mathrm{HSIC}(P_{xy}, P_x P_y) = \mathrm{MMD}(P_{xy}, P_x P_y; \mathcal{F}\times\mathcal{G})^2$$

$$= \left(\sup_{\|f\|=1} \mathbb{E}_{P_{xy}} f - \mathbb{E}_{P_x P_y} f \right)^2$$

$$= \| \mu_{P_{xy}} - \mu_{P_x P_y} \|^2_{\mathcal{F}\times\mathcal{G}}$$

which should equal 0 if X, Y are independent (eqn. $(**)$).

Noting that

$$\langle \mu_{P_{xy}}, \phi_x \otimes \psi_y \rangle = \mu_{P_{xy}}(x,y) = \ldots \qquad \left(\begin{array}{l} x \in \mathcal{X},\ y \in \mathcal{Y}, \\ \phi: \mathcal{X} \to \mathcal{F}, \\ \psi: \mathcal{Y} \to \mathcal{G} \end{array} \right)$$

$$\langle \mu_{P_x P_y}, \phi_x \otimes \psi_y \rangle = \mu_{P_x P_y}(x,y) = \ldots$$

We can expand the HSIC into:

$$HSIC(P_{xy}, P_x P_y) = \| \mu_{P_{xy}} - \mu_{P_x P_y} \|^2$$

$$= \langle \mu_{P_{xy}}, \mu_{P_{xy}} \rangle + \langle \mu_{P_x P_y}, \mu_{P_x P_y} \rangle - 2 \langle \mu_{P_{xy}}, \mu_{P_x P_y} \rangle$$

$$= \mathbb{E}_{xy \sim P_{xy}} \mathbb{E}_{x'y' \sim P_{xy}} k(x,x') l(y,y') + \mathbb{E}_{x \sim P_x} \mathbb{E}_{x' \sim P_x} k(x,x') \mathbb{E}_{y \sim P_y} \mathbb{E}_{y' \sim P_y} l(y,y')$$

$$- 2 \mathbb{E}_{xy \sim P_{xy}} \left[\mathbb{E}_{x' \sim P_x} k(x,x') \right] \left[\mathbb{E}_{y' \sim P_y} l(y,y') \right]$$

For scratched out part on other side:

$$\mu_{P_{xy}} := \mathbb{E}_{x,y \sim P_{xy}} \phi_x \otimes \psi_y \quad ($$

$$\mu_{P_{xy}}(x,y) = \langle \mu_{P_{xy}}, \phi_x \otimes \psi_y \rangle_{\mathcal{F} \times G}$$

you can show that $\mu_{P_{xy}}$ exists via Riesz theorem: exists in a Hilbert space $\mathcal{H}(\mathcal{F}, G)$ w/ Hilbert-Schmidt norm, dot product ↳ from which you prove

$$= \langle \mathbb{E}_{x,y \sim P_{xy}} \phi_X \otimes \psi_Y, \phi_x \otimes \psi_y \rangle_{\mathcal{F} \times G}$$

$$= \mathbb{E}_{x,y \sim P_{xy}} \langle \phi_X \otimes \psi_Y, \phi_x \otimes \psi_y \rangle_{\mathcal{F} \times G}$$

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

$$= \mathbb{E}_{x,y \sim P_{xy}} \langle \phi_X, \phi_x \rangle_{\mathcal{F}} \langle \psi_Y, \psi_y \rangle_G$$

$$= \mathbb{E}_{x,y \sim P_{xy}} k(X,x) l(Y,y)$$

$$\mu_{P_x P_y} := \mathbb{E}_{X \sim P_x} \mathbb{E}_{Y \sim P_y} \phi_X \otimes \psi_Y$$

As above,

$$\mu_{P_x P_y}(x,y) = \mathbb{E}_{X \sim P_x} \mathbb{E}_{Y \sim P_y} \langle \phi_X \otimes \psi_Y, \phi_x \otimes \psi_y \rangle$$

$$= \mathbb{E}_{X \sim P_x} k(X,x) \mathbb{E}_{Y \sim P_y} k(Y,y)$$

From this, you can show the above expansion of the HSIC.

Given a set of data points $\{(x_i, y_i)\}_{i=1}$, we can estimate the HSIC with

$$\hat{\mu}_{P_x} = \frac{1}{N}\sum_{i=1}^{N} \phi_{x_i}, \quad \hat{\mu}_{P_y} = \frac{1}{N}\sum_{i=1}^{N} \psi_{y_i}, \quad \hat{\mu}_{P_{xy}} = \frac{1}{N}\sum_{i=1}^{N} \phi_{x_i} \otimes \psi_{y_i}$$

$$\hat{\mu}_{P_x P_y} = \hat{\mu}_{P_x} \otimes \hat{\mu}_{P_y}, \quad K_{ij} = k(x_i, x_j), \quad L_{ij} = l(y_i, y_j),$$

yielding the empirical $\text{HSIC} = \frac{1}{N^2} \underbrace{\text{Tr}(KHLH)}$ (Frobenius product of K, HLH)

where $H = I - \frac{1}{N} \underbrace{\mathbb{1}_N}_{N\times 1 \text{ vector of 1's}} \mathbb{1}_N^T$

such that HKH is centered: has its column and row means subtracted

* In fact, this estimate is biased because it ~~takes~~ products b/w repeated data points. To get an unbiased estimate we include only products b/w kernels of different data points.

A different approach to measuring dependence is to find the pair of mappings $f: \mathcal{X} \to \mathcal{F}$, $g: \mathcal{Y} \to G$ that maximizes the covariance b/w $f(x)$ and $g(y)$, $(x \sim P_x, y \sim P_y)$ under a smoothness constraint on f and g. This is formalized as the constrained covariance (COCO):

$$\text{COCO}(P_{xy}, P_x, P_y; \mathcal{F}, G) = \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_G=1} \left(\mathbb{E}_{x,y\sim P_{xy}} f(x) g(y) - \mathbb{E}_{x\sim P_x} f(x)\, \mathbb{E}_{y\sim P_y} g(y) \right)$$

In feature space,

$$= \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_G=1} \langle \mu_{P_{xy}}, f \otimes g \rangle_{\mathcal{F}\times G} - \langle \mu_{P_x}, f \rangle_{\mathcal{F}} \langle \mu_{P_y}, g \rangle_G$$

$$= \quad '' \quad '' \quad - \langle \mu_{P_x} \otimes \mu_{P_y}, f \otimes g \rangle_{\mathcal{F}\times G}$$

$$= \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_G=1} \langle \underbrace{\mu_{P_{xy}} - \mu_{P_x} \otimes \mu_{P_y}}_{C_{xy} \text{ (covariance)}}, f \otimes g \rangle_{\mathcal{F}\times G}$$

$F \times G \quad F \quad G$

$\langle L, (a \otimes b) \rangle_{F \times G}$

$= \langle a, Lb \rangle_F$

$$= \sup_{\|f\|_F = 1, \|g\|_G = 1} \langle f, C_{XY}\, g \rangle_F$$

$\Rightarrow$ Empirical estimate of C_{XY}:

$$\hat{C}_{XY} = \frac{1}{N} \sum_{i=1}^{N} \phi_{x_i} \otimes \psi_{y_i} - (\hat{\mu}_{P_x} \otimes \hat{\mu}_{P_y}),$$

$$\hat{\mu}_{P_x} = \frac{1}{N} \sum_{j=1}^{N} \phi_{x_j}, \quad \hat{\mu}_{P_y} \text{ analogous}$$

For $X = \left[\phi_{x_1} \cdots \phi_{x_N} \right]$, $Y = \left[\psi_{y_1} \cdots \psi_{y_N} \right]$

$$K = X^T X \text{ s.t. } K_{ij} = k(x_i, x_j), \quad L = Y^T Y \text{ s.t. } L_{ij} = l(y_i, y_j)$$

$$\hat{C}_{XY} = \frac{1}{N} X H Y^T, \quad H = I - \frac{1}{N} \mathbb{1}_N$$

Pf. $\hat{C}_{XY} = \left[\frac{1}{N} \sum_{i=1}^{N} \phi_{x_i} \otimes \psi_{y_i} \right] - \left(\frac{1}{N} \sum_{i=1}^{N} \phi_{x_i} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \psi_{y_i} \right)$

$= \frac{1}{N} X Y^T - \left(\frac{1}{N} X \mathbb{1}_{N \times 1} \otimes \frac{1}{N} Y \mathbb{1}_{N \times 1} \right)$ where $\mathbb{1}_{N \times 1}$ is an $N \times 1$ vector of 1's, such that $A \mathbb{1}_{N \times 1}$ = sum of the columns of A

$= \frac{1}{N} X Y^T - \frac{1}{N^2} (X \mathbb{1}_{N \times 1})(Y \mathbb{1}_{N \times 1})^T$

$= \frac{1}{N} X Y^T - \frac{1}{N^2} X \mathbb{1}_{N \times N} Y^T$

$= \frac{1}{N} X \underbrace{\left(I - \frac{1}{N} \mathbb{1}_{N \times N} \right)}_{H} Y^T$

$= \frac{1}{N} X H Y^T$

We now solve the optimization problem above by assuming f, g are linear combinations of the mean-subtracted data in feature space $\{\tilde{\phi}_{x_i}\}_{i=1}^N$, $\{\tilde{\psi}_{y_i}\}_{i=1}^N$ respectively, where $\tilde{\phi}_{x_i} = \phi_{x_i} - \frac{1}{N}\sum_{i=1}^N \phi_{x_i}$

$$f = XH\alpha, \quad g = YH\beta$$

where again $H = I - \frac{1}{N}\mathbb{1}_{N \times N}$ such that $XH = [\tilde{\phi}_{x_1} \cdots \tilde{\phi}_{x_N}]$.

We then solve the following Lagrangian:

$$\mathcal{L}(f, g, \lambda_1, \lambda_2)$$

$$= f^T \hat{C}_{xy} g - \frac{\lambda_1}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\lambda_2}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)$$

$[H = H^T]$
$$= \alpha^T H X^T (XHY^T) YH\beta - \frac{\lambda_1}{2}\left(\alpha^T H X^T X H\alpha - 1\right) - \frac{\lambda_2}{2}\left(\beta^T H Y^T Y H\beta - 1\right)$$

$$= \frac{1}{N}\alpha^T HKHLH\beta - \frac{\lambda_1}{2}\left(\alpha^T HKH\alpha - 1\right) - \frac{\lambda_2}{2}\left(\beta^T HLH\beta - 1\right)$$

$[H = HH]$

$$= \frac{1}{N}\alpha^T \tilde{K}\tilde{L}\beta - \frac{\lambda_1}{2}\left(\alpha^T \tilde{K}\alpha - 1\right) - \frac{\lambda_2}{2}\left(\beta^T \tilde{L}\beta - 1\right)$$

$[\tilde{K} = (XH)^T(XH) = HKH]$

Taking the derivative w.r.t. α, β and setting to 0:

$$\frac{\partial}{\partial\alpha}\mathcal{L} = \frac{1}{N}\tilde{K}\tilde{L}\beta - \lambda_1\tilde{K}\alpha = 0$$

$$\frac{\partial}{\partial\beta}\mathcal{L} = \frac{1}{N}\tilde{L}\tilde{K}\alpha - \lambda_2\tilde{L}\beta = 0$$

If we multiply both sides by α^T for eqn 1 and by β^T for eqn 2,

$$\frac{1}{N}\alpha^T\tilde{K}\tilde{L}\beta = \lambda_1\alpha^T\tilde{K}\alpha$$
$$\frac{1}{N}\beta^T\tilde{L}\tilde{K}\alpha = \lambda_2\beta^T\tilde{L}\beta$$

Noting that $\alpha^T \tilde{K} \alpha = \|f\|_{\mathcal{F}}^2 = 1$ and $\beta^T \tilde{L} \beta = \|g\|_{\mathcal{G}}^2 = 1$ and that $\underset{(\text{scalar})}{\alpha^T \tilde{K} \tilde{L} \beta} = \mathrm{Tr}(\alpha^T \tilde{K} \tilde{L} \beta) = \mathrm{Tr}(\beta^T \tilde{L}^T \tilde{K}^T \alpha) = \mathrm{Tr}(\beta^T \tilde{L} \tilde{K} \alpha) = \beta^T \tilde{L} \tilde{K} \alpha$, we have $\lambda_1 = \lambda_2$. Thus, we can solve for α, β by the following eigenvalue eqn.:

$$\begin{bmatrix} 0 & \frac{1}{N}\tilde{K}\tilde{L} \\ \frac{1}{N}\tilde{L}\tilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

where $\lambda = \lambda_1 = \lambda_2$

Effectively, the constraints $\|f\|_{\mathcal{F}} = 1$ and $\|g\|_{\mathcal{G}} = 1$ enforce smoothness on f and g (recall that $\|f\|_{\mathcal{F}}^2 = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell^2}{\hat{k}_\ell}$), allowing only lower frequency dependencies to show up in the COCO. Higher frequency covariance can only be detected via COCO with large sample sizes.

~~Turns out~~? $HSIC = \sum_{i=1}^{N} \gamma_i^2$, where γ_i is the i^{th} largest eigenvalue of $\begin{bmatrix} 0 & \frac{1}{N}\tilde{K} \\ \frac{1}{N}\tilde{L} & 0 \end{bmatrix}$ (from eigenvalue equation above), in the limit of infinite samples

(missing something here about how γ_i's relate to f and g)

$\tilde{C}_{xy}$ is defined as the matrix such that

$$\langle A, \tilde{C}_{xy} \rangle_{HS} = \mathbb{E}_{xy} \langle A, \phi(x) \otimes \psi(y) \rangle$$

μ_x is defined as the vector such that

$$\langle f, \mu_x \rangle = \mathbb{E}_x \langle f, x \rangle$$

---

we have $\|\hat{C}_{xy}\|^2 = \langle \hat{C}_{xy}, \hat{C}_{xy} \rangle$

$$= \frac{1}{N^2} \sum_{i,j} k(x_i, x_j)\, l(y_i, y_j)$$

$$= \frac{1}{N^2} \mathrm{Tr}(KL)$$

obviously, this is a biased estimate, since we are including terms like $k(x_i, x_i)\, l(y_i, y_i)$

How biased is it?

unbiased estimate! $\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} k(x_i, x_j)\, l(y_i, y_j)$

Difference:

$$\frac{1}{N^2} \sum_{i}^{N} k(\;)l(\;) + \left(\frac{1}{N^2} - \frac{1}{N(N-1)}\right) \sum_{i} \sum_{j \neq i} k(\;)\, l(\;)$$

$$= \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^{N} k(x_i, x_i)\, l(y_i, y_i) - \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} k(\;)\, l(\;) \right]$$

Taking expectations:

$$= \left[\mathbb{E}_{xy} k(x,x)\, \ell(y,y) - \mathbb{E}_{xy} \mathbb{E}_{x'y'} k(x,x')\, \ell(y,y') \right] \frac{1}{N}$$

$\Rightarrow$ i.e. the expected difference (BIAS) drops with $\frac{1}{N}$

(*But, be careful about kernel such that $k(x,x)$, $\ell(y,y)$ not too large such that they dominate)

Statistical Testing w/ HSIC

The (biased) empirical estimate of HSIC is a V-statistic when $P_{xy} \neq P_x P_y$

Under the Null, however, ($P_{xy} = P_x P_y$), again it is a ~~[illegible]~~ degenerate U-statistic (infinite weighted sum of χ^2 distributions)

But we can find the 1st and 2nd moments and approximate the infinite sum of χ^2's w/ a Gamma distribution with matched first two moments.

Or, we can use a permutation test.

Hilbert-Schmidt Operators

$L, M : G \to \mathcal{F}$, i.e. $L, M \in \mathcal{F} \times G$

Let $\{f_i\}_{i \in I}$ be basis of $\mathcal{F}$

$\{g_j\}_{j \in J}$ " " " G

Dot Product:

$$\langle L, M \rangle_{HS} = \sum_{j \in J} \langle L g_j, M g_j \rangle_{\mathcal{F}}$$

$$= \sum_{i \in I} \sum_{j \in J} \langle L g_j, f_i \rangle_{\mathcal{F}} \langle M g_j, f_i \rangle_{\mathcal{F}}$$

Norm:

$$\| L \|_{HS}^2 = \sum_{j \in J} \| L g_j \|_{\mathcal{F}}^2 = \sum_{i \in I} \sum_{j \in J} \left| \langle L g_j, f_i \rangle_{\mathcal{F}} \right|^2$$

Rank 1 Operators: $(b \otimes a) f \longrightarrow \langle f, a \rangle_{\mathcal{F}}\, b$

for $a \in \mathcal{F}$, $b \in G$ forming rank 1 operator $a \otimes b : G \to \mathcal{F}$

Property #1: $$\| a \otimes b \|_{HS}^2 = \sum_{j \in J} \| (a \otimes b) g_j \|_{\mathcal{F}}^2$$

$$= \sum_{j \in J} \| \langle g_j, b \rangle_G\, a \|_{\mathcal{F}}^2$$

$$= \| a \|_{\mathcal{F}}^2 \sum_{j \in J} \left| \langle g_j, b \rangle_G \right|^2$$

$$= \| a \|_{\mathcal{F}}^2 \| b \|_G^2$$

Property #2: $\langle L, a\otimes b\rangle_{HS} = \langle a, Lb\rangle_{\mathcal{F}}$

Pf. $$\langle L, a\otimes b\rangle_{HS} = \sum_{j\in J}\langle Lg_j, (a\otimes b)g_j\rangle_{\mathcal{F}}$$

$$= \sum_{j\in J}\langle Lg_j, a\langle b, g_j\rangle_G\rangle_{\mathcal{F}}$$

$$= \sum_{j\in J}\langle Lg_j, a\rangle_{\mathcal{F}}\langle b, g_j\rangle_G$$

$$= \left\langle \sum_{j\in J} L\langle b, g_j\rangle_G g_j, a\right\rangle_{\mathcal{F}}$$

$$= \left\langle L\sum_{j\in J}\langle b, g_j\rangle g_j, a\right\rangle_{\mathcal{F}}$$

$$= \langle Lb, a\rangle_{\mathcal{F}}$$

Property #3: $\langle u\otimes v, a\otimes b\rangle_{HS} = \langle u, a\rangle_{\mathcal{F}}\langle v, b\rangle_G$

Pf. From above, $\langle u\otimes v, a\otimes b\rangle_{HS} = \langle a, (u\otimes v)b\rangle_{\mathcal{F}}$

$$= \langle a, \langle v, b\rangle_G u\rangle_{\mathcal{F}}$$

$$= \langle a, u\rangle_{\mathcal{F}}\langle v, b\rangle_G$$