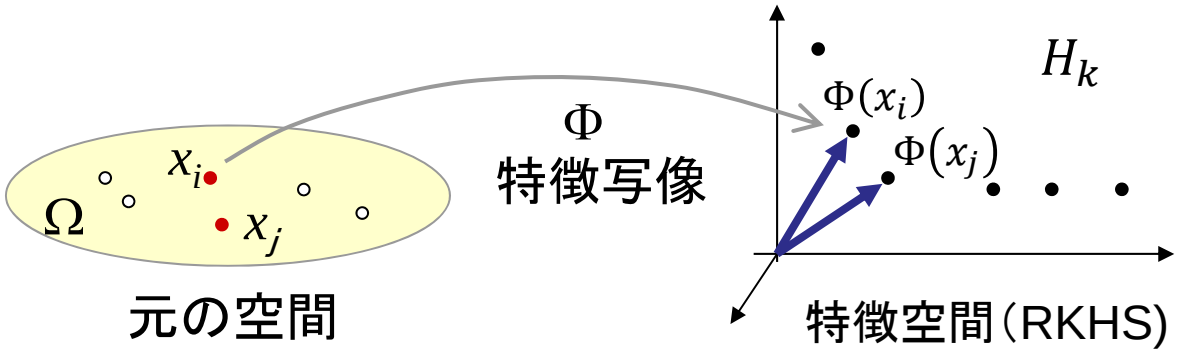
カーネル法入門

6. カーネル平均を用いたノンパラメトリック推論

福水健次
統計数理研究所／総合研究大学院大学

大阪大学大学院基礎工学研究科・集中講義
2014 September

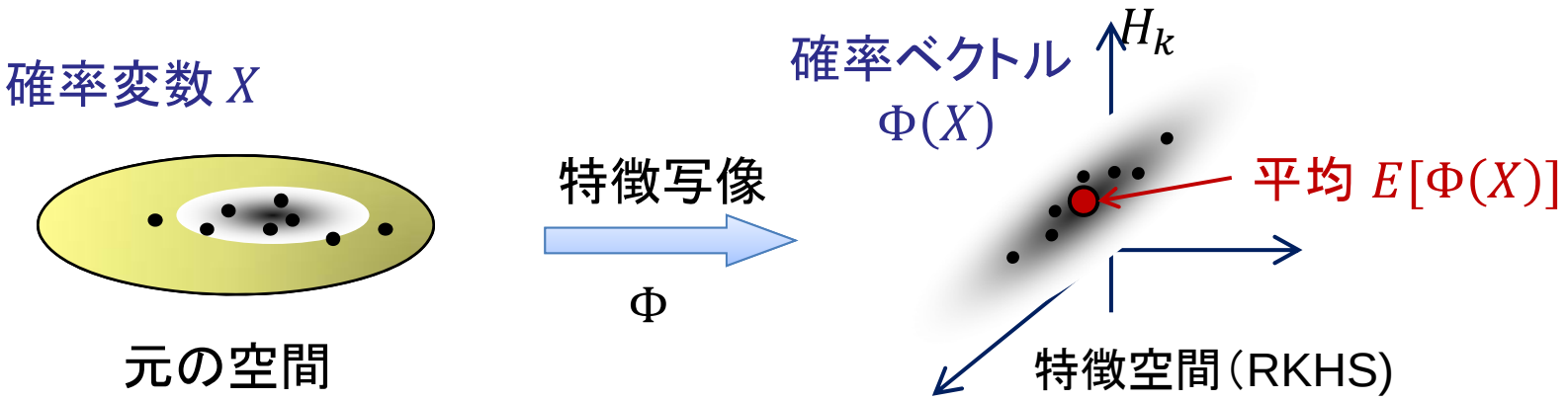
カーネル法概観：復習



$$\Phi: \quad \Omega \quad \rightarrow \quad H, \qquad x \mapsto \Phi(x) = k(\cdot, x)$$

- 「データ点」を特徴ベクトルに写像し，データ解析手法を適用する

カーネル平均による方法：概観



ランダムな特徴ベクトル$\Phi(X)$の平均 $E[\Phi(X)]$ だけで表現する，

カーネル平均

X: 可測空間 (Ω, B) に値を取る確率変数，分布 P.
k: Ω上の可測な正定値カーネル，　　H: kの定めるRKHS

Def.　Xの（Hにおける）カーネル平均

$$m_X := E[\Phi(X)] = E[k(\cdot, X)] = \int k(\cdot, x) dP(x)$$

- m_X もRKHSの元，すなわち関数.
- 確率 P によって m_P とも書くことにする.
- 例：　ガウスカーネルのとき

$$m_P(y) = \int \exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right) p(x) dx \qquad \text{[convolution]}$$

- 厳密には，$E\|\Phi(X)\| = E[\sqrt{k(X,X)}] < \infty$ のときに平均が存在（Bochner積分）.

■ カーネル平均の性質

– 再生性

$$\langle f, m_X \rangle = E[f(X)] \qquad \forall f \in H_k.$$

言い換えると

$$\langle f, E[\Phi(X)] \rangle = E[\langle f, \Phi(X) \rangle]$$

平均操作と内積は交換可能

– 高次モーメントの表現

例）

Taylor展開 $k(u,x) = c_0 + c_1 ux + c_2 (ux)^2 + \cdots \qquad (c_i > 0),$

e.g. $k(u,x) = e^{ux}$

$$m_X(u) = c_0 + c_1 E[X]u + c_2 E[X^2]u^2 + \cdots$$

モーメント母関数の役割を果たす

特性的な正定値カーネル

(Fukumizu et al. JMLR 2004, AoS 2009; Sriperumbudur et al. JMLR2010)

定義. 可測空間 (Ω, B) 上の可測かつ有界な正定値カーネル k が特性的（characteristic）であるとは

$$\{ (\Omega, B) \text{ 上の確率} \} \to H_k, \quad P \mapsto m_P$$

が単射であること, すなわち,

$$E_{X\sim P}[k(\cdot, X)] = E_{Y\sim Q}[k(\cdot, Y)] \Leftrightarrow P = Q$$

であることをいう.

特性的なカーネルによって, カーネル平均 m_P は確率Pを一意に定める

– 「特性関数」$E[e^{\sqrt{-1}X^T\omega}]$ の類似.
  - 特性関数はユークリッド空間上のBorel確率測度を一意に定める.
  - 特性的なカーネルはユークリッド空間に限らない (Fukumizu et al. 2009).
  - カーネルトリックにより推定量の計算が容易.

– 例: Gaussian, Laplace カーネルは特性的

  多項式カーネルは特性的でない（d次のモーメントまでしか表せない）

– 特性的なRKHSは十分広い空間である.

  定理6．1　(Fukumizu et al 2009)

  可測空間(Ω, B)上の有界で可測な正定値カーネル k が特性的であるための必要十分条件は，(Ω, B)の任意の確率Pに対して$H + \mathbf{R}$（和空間）が$L^2(P)$で稠密なことである.

  （証明は福水．補題8．6）

カーネル平均を用いた統計的推論

原理: 特性的なカーネルを用いると,
確率 P に関する推論問題 ⇒ ベクトル m_P に関する推論問題

- 2標本問題 → $m_P = m_Q$? (Gretton et al. JMLR 2012)
- 独立性検定 → $m_{XY} = m_X \otimes m_Y$?
- ベイズ推論 → 事後確率のカーネル平均表示.

カーネル平均の推定

$X_1, \ldots, X_n$: i.i.d.標本

定義. 標本カーネル平均

$$\widehat{m}_X^{(n)} := \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)$$

一致性

定理　6. 2.

$E[k(X,X)] < \infty$ のとき,

$$\|\widehat{m}_X^{(n)} - m_X\|_H = O_p\left(n^{-\frac{1}{2}}\right) \quad (n \to \infty).$$

補題6．３　$\|m_X\|^2 = E\left[k\left(X, \tilde{X}\right)\right]$　　（$\tilde{X}$: independent copy of X）

証明）

$$
\begin{aligned}
& E\left\|\hat{m}_X^{(n)} - m_X\right\|_H^2 \\
&= E\left[\frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j) - \frac{2}{n}\sum_{i=1}^{n} E[k(X_i, X)] + E\left[k\left(X, \tilde{X}\right)\right]\right] \\
&= E\left[\frac{1}{n^2}\sum_{i\neq j} k(X_i, X_j) + \frac{1}{n^2}\sum_{i=1}^{n} k(X_i, X_i) - \frac{2}{n}\sum_{i=1}^{n} E[k(X_i, X)]\right] + E\left[k\left(X, \tilde{X}\right)\right] \\
&= \frac{n^2 - n}{n^2} E[k(X, \tilde{X})] + \frac{n}{n^2} E[k(X, X)] - 2E\left[k\left(X, \tilde{X}\right)\right] + E\left[k\left(X, \tilde{X}\right)\right] \\
&= \frac{1}{n}\left\{E[k(X, X)] - E\left[k\left(X, \tilde{X}\right)\right]\right\}
\end{aligned}
$$

Chebyshev の不等式により

$$
\Pr\left(\left\|\hat{m}_X^{(n)} - m_X\right\|_H \geq \frac{a}{\sqrt{n}}\right) \leq \frac{nE\left\|\hat{m}_X^{(n)} - m_X\right\|_H^2}{a^2} = \frac{E[k(X, X)] - E\left[k\left(X, \tilde{X}\right)\right]}{a^2}.
$$

共分散作用素

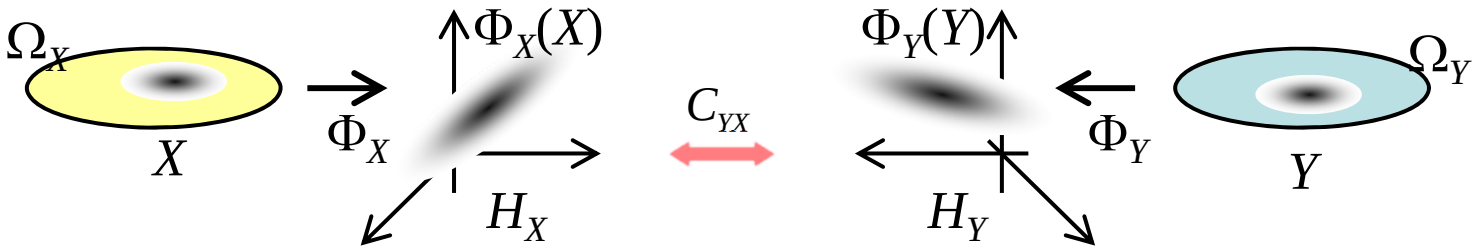
$(X, Y) : \Omega_X \times \Omega_Y$ に値を取る確率変数.

$(H_X, k_X), (H_Y, k_Y)$: Ω_X, Ω_Y 上のRKHS.

定義. (中心化しない) 共分散作用素 $C_{YX}: H_X \to H_Y$, $C_{XX}: H_X \to H_X$

$$C_{YX} := E\left[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle_{H_X}\right], \quad C_{XX} = E\left[\Phi_X(X)\langle \Phi_X(X), \cdot \rangle_{H_X}\right]$$

$$C_{YX} f = \int k_Y(\cdot, y) f(x) dP(x, y), \quad C_{XX} f = \int k_X(\cdot, x) f(x) dP_X(x)$$



– ユークリッド空間に値を取る通常の確率ベクトル X, Y の共分散行列 $V_{YX} = E[YX^T]$ の自然な拡張

– 再生性

$$\langle g, C_{YX} f \rangle_{H_Y} = E[f(X)g(Y)] \quad \forall f \in H_X, g \in H_Y$$

– 共分散作用素は，X と Y の関係を表現している

– 共分散作用素＝積空間でのカーネル平均

「線形写像とテンソル積は同一視できる」（次ページ）

$$C_{YX} \quad \cong \quad m_{XY} := E[\Phi_X(X) \otimes \Phi_Y(Y)]$$

共分散作用素 $H_X \to H_Y$

積空間での(X, Y)のカーネル平均 $H_X \otimes H_Y$

■ テンソル積と線形写像

– 内積空間 V, W のテンソル積 $V \otimes W$ は，線形写像の空間 $L(V,W) := \{T : V \to W \mid T \text{ は線形写像}\}$ と同一視される.

$$V \otimes W \longrightarrow L(V,W)$$

$$v \otimes w \mapsto T_{v\otimes w} : \tilde{v} \mapsto (v,\tilde{v})_V\, w$$

一般には

$$\xi = \sum_{i=1}^{m} v_i \otimes w_i \mapsto T_\xi : \tilde{v} \mapsto \sum_{i=1}^{m} (v_i,\tilde{v})_V\, w_i$$

– Hilbert空間の場合は，$H_X \to H_Y$のHilbert-Schmidt 作用素全体と，積空間$H_X \otimes H_Y$が同一視される.

共分散作用素の推定量

$(X_1, Y_1,), \ldots, (X_n, Y_n) \sim P$, i.i.d.

標本共分散作用素:

$$\hat{C}_{YX} f = \frac{1}{n} \sum_{i=1}^{n} k_Y(\cdot, Y_i) \langle k_X(\cdot, X_i), f \rangle$$

$$= \frac{1}{n} \sum_{i=1}^{n} k_Y(\cdot, Y_i) f(X_i)$$

- $\hat{C}_{YX} : H_X \to H_Y$ RKHS間の作用素，ランクは高々n.
- $\sqrt{n}$ 一致性を持つ（カーネル平均とみなせばよい）

2標本問題への応用

2標本問題

$\mathbf{X_n} = X_1, \ldots, X_n \sim P$, i.i.d.

$\mathbf{Y_m} = Y_1, \ldots, Y_m \sim Q$, i.i.d.　　$\mathbf{X_n}$ と $\mathbf{Y_m}$ は互いに独立

2標本問題：　P = Q？

- 例
  - 服薬を行ったグループと行っていないグループで，血糖値に違いがあるか？
  - 中学校Aと中学校Bでテストの点数に違いがあるか？

- 古典的なアプローチ：　X, Y の平均値の違いを見る.
- カーネル法によるアプローチ:　$m_P = m_Q$?

  $\left\| \hat{m}_P - \hat{m}_Q \right\|_H^2$ により検定を行う.

2標本問題の従来法

- パラメトリックモデルを用いる（正規分布など）
  - 2標本 t 検定
    - 1次元分布，等分散
    - 1次元正規分布，不等分散（Welch’s test）

- ノンパラメトリック検定
  - 1次元の場合：よい方法が知られている
    - Mann-Whitney U-test （等分散，順序統計量に基づく）
    - Kolmogorov-Smirnov test （不等分散でもOK，c.d.f.に基づく）
    - Wald–Wolfowitz run test （順序統計量の連に基づく）

  - 多次元の場合：
    - Friedman Rafsky (1979) による KS, WW test の多次元拡張.
      Minimum spanning tree を用いる.

MMD: maximum mean discrepancy

定義.（Gretton et al NIPS19）

X, Y: 可測空間 (Ω, B) 上に値を取る確率変数.

(H, k): Ω上の特性的な正定値カーネル

$$\mathrm{MMD}^2(X, Y) \coloneqq \|m_X - m_Y\|_H^2$$

サンプルによる推定量

$$\mathrm{MMD}^2_{\mathrm{emp}}(\mathbf{X_n}, \mathbf{Y_m}) \coloneqq \|\hat{m}_X - \hat{m}_Y\|_H^2$$

– 名前の由来

$$\|m_X - m_Y\|_H = \sup_{\|f\|_H=1} |E[f(X)] - E[f(Y)]|$$

平均の差が最大になる値を見る.

MMDによる2標本検定

- ノンパラメトリック検定

  帰無仮説　H_0: $P = Q$

  対立仮説　H_1: $P \neq Q$

- 検定統計量

$$MMD^2_{emp}(\mathbf{X_n}, \mathbf{Y_m}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(Y_i, Y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j)$$

あるいは，不偏化した U-統計量

$$T_{n,m} = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(X_i, X_j) + \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k(Y_i, Y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i, Y_j)$$

＊ 2標本 U-統計量

カーネル $h(x_1, x_2; y_1, y_2)$ を持つ2標本U-統計量

$$T_{n,m} = \frac{1}{\binom{n}{r}\binom{m}{s}}\sum_{\alpha}\sum_{\beta} h(X_{\alpha_1}, X_{\alpha_2}; Y_{\beta_1}, Y_{\beta_2})$$

α: $\{1, \ldots, n\}$の要素数rの部分集合
β: $\{1, \ldots, m\}$の要素数sの部分集合

$$h(x_1, x_2; y_1, y_2) = k(x_1, x_2) + k(y_1, y_2) - \frac{1}{2}\{k(x_1, y_1) + k(x_1, y_2) + k(x_2, y_1) + k(x_2, y_2)\}$$

とおけばよい [Exercise: これを確かめよ]

– 帰無分布（退化したU-統計量）

- 漸近分布

$\frac{n}{n+m} \to \gamma,\ \frac{m}{n+m} \to 1-\gamma$ （$n+m \to \infty$）と仮定

$$(n+m)T_{n,m} \implies \sum_{i=1}^{\infty} \lambda_i \left(Z_i^2 - \frac{1}{\gamma(1-\gamma)} \right) \qquad (n.m \to \infty)$$

$Z_i\ (i = 1,2, \dots,) \sim N\left(0, \frac{1}{\gamma(1-\gamma)}\right), i.i.d.$

$\lambda_i\ (i = 1,2, \dots)$: 次の積分作用素$B$の非負固有値

$$B: L^2(P) \to L^2(P), \qquad Bf = \int \tilde{k}(y,x) f(x) dP(x)$$

$$\tilde{k}(x,y) = k(x,y) - E[k(X,y)] - E[k(x,X)] + E[k\left(X, \tilde{X}\right)]$$

- 中心化グラム行列による固有値の推定が可能（Gretton et al NIPS 22）

Experiments

Comparison of two databases.

	Data size /	Dim
Neural I:	4000 /	63
Neural II:	1000 /	100
Health:	25 /	12600
Subtype:	25 /	2118

Data set	Attr.	MMD-B	WW	KS
Neural I	Same	96.5	97.0	95.0
	Different	0.0	0.0	10.0
Neural II	Same	94.6	95.0	94.5
	Different	3.3	0.8	31.8
Health	Same	95.5	94.7	96.1
	Different	1.0	2.8	44.0
Subtype	Same	99.1	94.6	97.3
	Different	0.0	0.0	28.4

Gaussian kernels are used.

WW: Wald-Walfovitz test
KS: Kolmogorov-Smirnov test

-- Classical methods (see Appendix)

Percentage of accepting $P = Q$.
Significance level $\alpha = 0.05$.

(Gretton et al. JMLR 2012)

Kolmogorov-Smirnov (K-S) test for two samples

One-dimensional variables

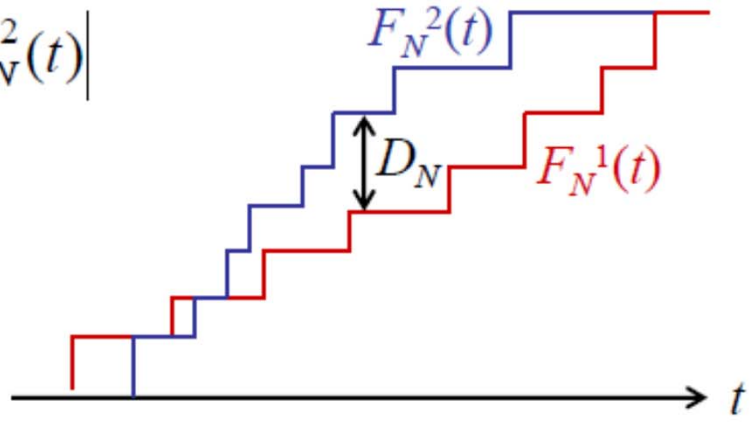
- Empirical distribution function

$$F_N(t) = \frac{1}{N}\sum_{i=1}^{N} I(X_i \le t)$$

- KS test statistics

$$D_N = \sup_{t \in \mathbf{R}} \left| F_N^1(t) - F_N^2(t) \right|$$

- Asymptotic null distribution is known (not shown here).



Wald-Wolfowitz run test

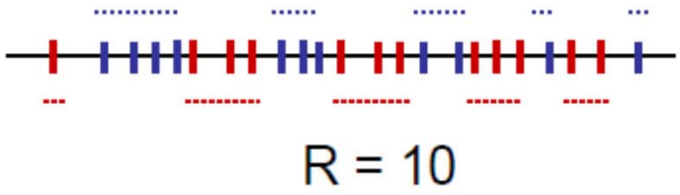
One-dimensional samples

- Combine the samples and plot the points in ascending order.
- Label the points based on the original two groups.
- Count the number of "runs", i.e. consecutive sequences of the same label.

R = Number of runs

- Test statistics

$$T_N = \frac{R - E[R]}{\sqrt{Var[R]}} \quad \Rightarrow \quad N(0,1)$$



R = 10

- In one-dimensional case, less powerful than KS test

依存性尺度と独立性検定

復習：独立性

■ 定義

– m次元確率ベクトル X と n次元確率ベクトル Y が独立であるとは，
任意の可測集合 $A \in B(\mathbf{R}^m), B \in B(\mathbf{R}^n)$に対し，

$$\Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$$

が成り立つことをいう． $X \perp Y$ と書く．

■ Fact.

1. $X \perp Y$ ならば
$$E[f(X)g(Y)] = E[f(X)]E[g(Y)].$$

2. (X, Y) の分布が密度関数 $p_{XY}(x, y)$ を持つとき，Xと Yの周辺分布の密度関数をそれぞれ $p_X(x), p_Y(y)$ とすると，
$$X \perp Y \iff p_{XY}(x, y) = p_X(x)p_Y(y)$$

復習:ガウス変数の独立性

– 多変量ガウス確率変数

$$X = (X_1, \ldots, X_m) \sim \ N(\mu, V)$$

– 独立性

(X, Y): $(m + \ell)$次元ガウス確率変数

$$\text{分散共分散行列 } V = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix} \begin{matrix} m \\ \ell \end{matrix}$$

<u>命題</u>

$$X \perp Y \quad \Leftrightarrow \quad V_{XY} = O$$

$$\Leftrightarrow \quad E[XY^T] = E[X]E[Y]^T$$

If $V_{XY} = O$

$$p_{XY}(x, y) = \frac{1}{(2\pi)^{\frac{m+n}{2}}|V_{XX}||V_{YY}|} \exp\left(-\frac{1}{2}(x^T, y^T)\begin{pmatrix} V_{XX}^{-1} & O \\ O & V_{YY}^{-1} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right)$$

$$= \frac{1}{(2\pi)^{\frac{m}{2}}|V_{XX}|} \exp\left(-\frac{1}{2}x^T V_{XX}^{-1} x\right) \frac{1}{(2\pi)^{\frac{n}{2}}|V_{YY}|} \exp\left(-\frac{1}{2}y^T V_{YY}^{-1} y\right) = p_X(x)p_Y(y)$$

独立性の特徴づけ

– 復習:

$$X \perp Y \quad \Leftrightarrow \quad \mathrm{Cov}[f(X), g(Y)] = 0 \qquad (\forall f, g \text{ 可測関数})$$

– 依存性尺度？

$$\sup_{f,g:measurable} |\mathrm{Cov}[f(X), g(Y)]|$$

- すべての可測関数を評価することはできない.
- 有限サンプルによってどう推定するか？

HSIC

Hilbert-Schmidt Independence Criteria (Gretton et al 2005)

(X, Y): $\Omega_X \times \Omega_Y$に値を取る任意の確率変数（Euclid空間とは限らない）

$(H_x, k_x), (H_y, k_y)$: RKHSと正定値カーネル

$\Sigma_{YX} := C_{YX} - m_Y \otimes m_X,$　（中心化した）共分散作用素

$\hat{\Sigma}_{YX} := \hat{C}_{YX} - \hat{m}_Y \otimes \hat{m}_X$

<u>定義</u>

$$\mathrm{HSIC}(X, Y) = \|\Sigma_{YX}\|_{HS}^2$$

$$\mathrm{HSIC}_{emp}(\mathbf{X_n}, \mathbf{Y_n}) = \left\|\hat{C}_{YX} - \hat{m}_Y \otimes \hat{m}_X\right\|_{HS}^2$$

– 共分散の2乗和

$$\|\Sigma_{YX}\|_{HS}^2 = \sum_{i,j} \mathrm{Cov}[\phi_i(X), \psi_j(Y)]^2$$

Note:

$$\langle \Sigma_{YX}\phi_i, \psi_j \rangle = \mathrm{Cov}[\phi_i(X), \psi_j(Y)]$$

HSICはすべての基底関数の組に関して共分散が0か調べている.

– MMDの特別なケース

$$\|\Sigma_{YX}\|_{HS}^2 = \|C_{YX} - m_Y \otimes m_X\|_{HS}^2 = MMD^2(P_{XY}, P_X P_Y)$$

同時分布と，周辺分布の積をMMDで比較

– 積分表示を持ち，推定が容易

$$\mathrm{HSIC}(X,Y) = \|C_{YX} - m_Y \otimes m_X\|_{HS}^2$$
$$= E\left[k_x(X,\tilde{X})k_y(Y,\tilde{Y})\right] - 2E\left[E\left[k_x(X,\tilde{X})k_y(Y,\tilde{Y})\middle|X,Y\right]\right]$$
$$+ E\left[k_x(X,\tilde{X})\right]E\left[k_y(Y,\tilde{Y})\right].$$

[Exercise] 上式を示せ．　　$(\tilde{X},\tilde{Y})$: independent copy of (X,Y)

<u>推定量</u>

$$\mathrm{HSIC}_{emp}(\mathbf{X_n},\mathbf{Y_n}) = \frac{1}{n^4}\sum_{i,j,s,t} k_x(X_i,X_j)k_y(Y_s,Y_t)$$
$$-\frac{2}{n^3}\sum_{i,j,s} k_x(X_i,X_j)k_y(Y_s,Y_j) + \frac{1}{n^2}\sum_{i,j} k_x(X_i,X_j)\sum_{i,j} k_y(Y_i,Y_j)$$

$$= \frac{1}{n^2}\mathrm{Tr}[\tilde{K}_X\tilde{K}_Y]$$

$\tilde{K}_X = Q_n K_X Q_n$: 中心化グラム行列

K_X: Gram行列，$Q_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$

Hilbert-Schmidt ノルム

定義.　ヒルベルト空間の間の作用素 $T: H_1 \to H_2$ がHilbert-Schmidtであるとは，H_1, H_2の任意の正規直交基底 $\{\phi_i\}_i, \{\psi_j\}_j$ に対して，

$$\sum_{i,j} |\langle \psi_j, T\phi_i \rangle|_{H_2}^2 < \infty$$

であることをいう.

このとき，左辺の値は正規直交基底の取り方に依らず，

$$\|T\|_{HS} := \sqrt{\sum_{i,j} |\langle \psi_j, T\phi_i \rangle|_{H_2}^2}$$

はHilbert-Schmidt作用素全体のなすベクトル空間にノルムを定める.

共分散作用素と独立性

定理

$(\Omega_x, B_x), (\Omega_y, B_y)$: 可測空間

(X, Y): $\Omega_X \times \Omega_Y$に値を取る任意の確率変数

$(H_x, k_x), (H_y, k_y)$: Ω_x, Ω_y上のRKHSと正定値カーネル

仮定: 積カーネル $k_x k_y$ は $\Omega_x \times \Omega_y$ 上特性的

$$X \perp Y \quad \Leftrightarrow \quad \Sigma_{YX} = O.$$

系

上の定理の仮定の下,

$$X \perp Y \quad \Leftrightarrow \quad \mathrm{HSIC}(\mathrm{X}, \mathrm{Y}) = 0.$$

HSICによる独立性検定

- 独立性検定

  $(X_1 Y_1), \ldots, (X_n, Y_n) \sim P_{XY}, i.i.d.$

  H0: X と Y は独立

  H1: 独立でない

- 検定統計量

$$T_n := n\left\|\hat{\Sigma}_{YX}\right\|_{HS}^2$$

  基本はMMDと同じ （$\|\Sigma_{YX}\|_{HS}^2 = MMD^2(P_{XY}, P_X P_Y)$ ）

- 棄却域の決定
  - 漸近的（$n \to \infty$）な帰無分布を用いる:

$$T_n \implies \sum_{i=1}^{\infty} \lambda_i \left(Z_i^2 - 2\right), \qquad Z_i \sim N(0,2), i.i.d.$$

  - 並べ替え検定／リサンプリング

独立性検定の従来法

- カテゴリカル／分割表
  - χ^2-test

$$T_n = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{\left(\hat{P}_{ij} - \hat{P}_{i\cdot}\hat{P}_{\cdot j}\right)^2}{\hat{P}_{i\cdot}\hat{P}_{\cdot j}}$$

	A	B
a	12	23
b	5	32

- 連続値＆ノンパラメトリック
  - Spearman's rank test

$$\rho_n = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n}, \qquad D_i = X_{n:i} - Y_{n:i}.$$

極限帰無分布は t-分布

  - Power divergence

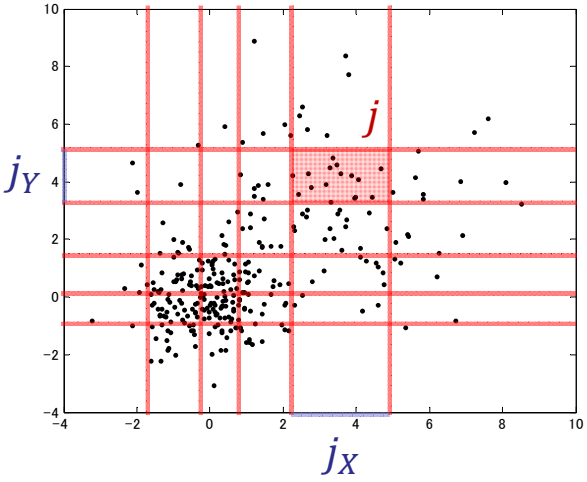
■ Jensen-Tsallis / Power Divergence (Read&Cressie, Ku&Fine05)

– 各次元を q 個に分割(等頻度など). $\{A_j\}_{j\in J}$ $(|J| = q^{d_x+d_y})$

– $\hat{p}_j$: 各小領域の頻度. $\hat{p}_{j_X}, \hat{p}_{j_Y}$: j を構成する各変数の領域の周辺頻度

$$I^{\lambda}(X,Y) := \frac{1}{\lambda(\lambda+2)} \sum_{j\in J} \hat{p}_j \left\{ \left(\frac{\hat{p}_j}{\hat{p}_{j_X}\hat{p}_{j_Y}} \right)^{\lambda} - 1 \right\}$$

- $I^0(X,Y) =$ 相互情報量
- $I^2(X,Y) =$ χ^2-ダイバージェンス
  / mean square contingency



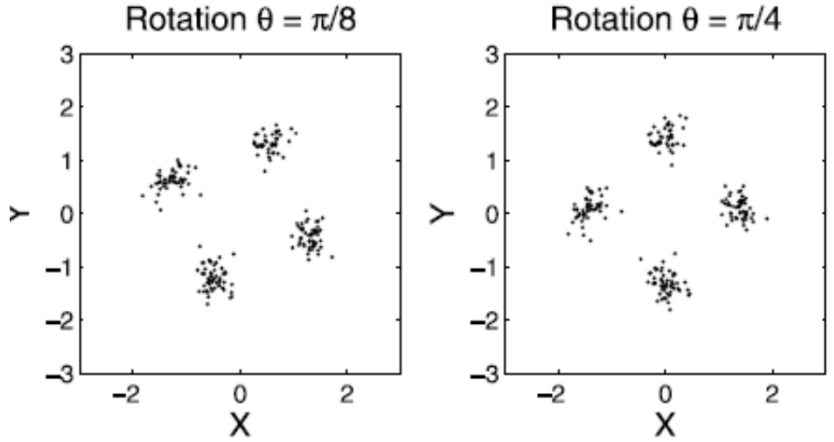
– $n \to \infty$における帰無分布

$$2nI^{\lambda}(X,Y) \Rightarrow \chi^2_{\left(q^{d_x}-1\right)\left(q^{d_y}-1\right)}$$

– 分割による方法は，次元が高いと困難

数値実験

– データ: X, Y: 1 dim プラス ノイズの次元



回転角$0 \leq \theta \leq \frac{\pi}{4}$. $\theta = 0$ なら独立, $\theta = \frac{\pi}{4}$で最も依存性高い

– 方法

- HSIC モーメントマッチングによるGamma曲線近似
- HSIC リサンプリング
- Power divergence（$\lambda = 3/2$）
- functional correlation (Dauxois and G. M. Nkiet AoS 1998)

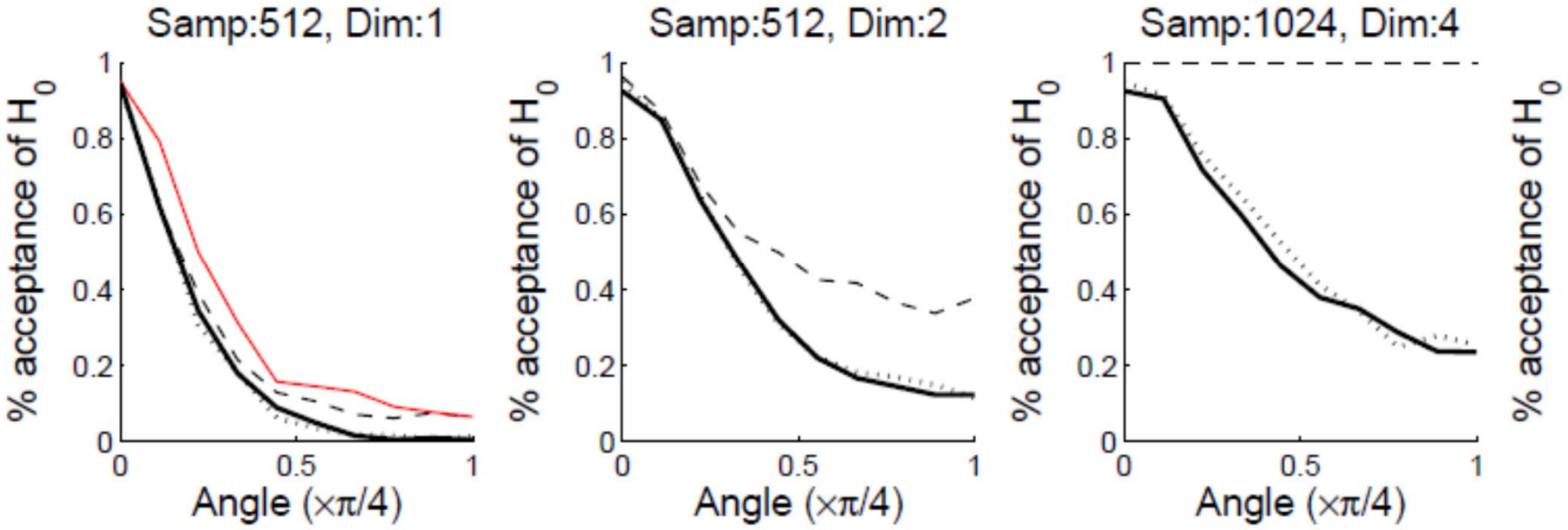
—— HSIC (Gamma近似)

····· HSIC (resampling)

- - - Power divergence ($\lambda = 3/2$, 等確率による分割)

—— functional Correlation

Type II errors



文書の独立性検定

- Data: Official records of Canadian Parliament in English and French.
  - Dependent data: 5 line-long parts from English texts and their French translations.
  - Independent data: 5 line-long parts from English texts and random 5 line-parts from French texts.
- Kernel: Bag-of-words and spectral kernel

Results of permutations test with HS measure

Topic	Match	BOW(N=10)	Spec(N=10)	BOW(N=50)	Spec(N=50)
Agri-culture	Random	0.94	0.95	0.93	0.95
	Same	0.18	0.00	0.00	0.00
Fishery	Random	0.94	0.94	0.93	0.95
	Same	0.20	0.00	0.00	0.00
Immig-ration	Random	0.96	0.91	0.94	0.95
	Same	0.09	0.00	0.00	0.00

Acceptance rate (α = 5%)

(Gretton et al. 07)

dCovとの比較

■ Distance covariance

Def.

X, Y: 確率変数(Euclidean空間に値を取る)

$$\mathrm{dCov}^2(X,Y) := E[\|X-X'\|\,\|Y-Y'\|] - 2E[\|X-X'\|\,\|Y-Y''\|] + E[\|X-X'\|]\,E[\|Y-Y'\|].$$

(X', Y'), (X'', Y''): independent copies of (X, Y).

c.f. $\mathrm{HSIC}(X,Y) = E\left[k_x(X,X')k_y(Y,Y')\right] - 2E\left[E[k_x(X,X')k_y(Y,Y'')]\right] + E[k_x(X,X')]E\left[k_y(Y,Y')\right]$

Note: $\|X-X'\|$ は正定値ではない.

– Distance covariance (distance correlation) は最近注目されている依存性尺度 (Székely, Rizzo, Bakirov, AoS 2007).

■ dCovの拡張

- Ω 上のsemi-metric ρ
  - $\rho(z, z') = \rho(z', z)$　　[対称性]
  - $\rho(z, z') \geq 0$, 等号成立は $z = z'$, [正値性]

  （三角不等式は仮定しない）

- semi-metric ρ_x, ρ_y に対し，generalized distance covariance を以下で定義.

$$\mathrm{dCov}^2_{\rho_X,\rho_Y}(X, Y) := E[\rho_X(X, X')\rho_Y(Y, Y')] - 2E[\rho_X(X, X')\rho_Y(Y, Y'')] + E[\rho_X(X, X')]\, E[\rho_Y(Y, Y')].$$

定理 (Sejdinovic et al. AoS 2013). [dCovはHSICの特別な例]

ρ_x, ρ_y が負定値なsemi-metric であるとき,

$k(z,z') := \frac{1}{2}\{\rho(z,z_0) + \rho(z',z_0) - \rho(z,z')\}$ により定まる正定値カーネル（補題4.8参照）を k_x, k_y とおくと,

$$\mathrm{HSIC}_{k_x,k_y}(X,Y) = \mathrm{dCov}_{\rho_X,\rho_Y}(X,Y).$$

証明は略. 直接計算.

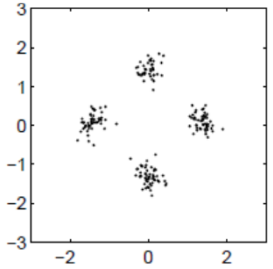
例）ユークリッド空間上 $\|x-y\|^q$ $(0 < q \leq 2)$ は負定値なsemi-metric (定理4.10 Remark 1)

$$\rho(z,z') = \|z-z'\|^q, \qquad k_\rho(z,z') = \frac{1}{2}\{\|z\|^q + \|z'\|^q - \|z-z'\|^q\}$$

$$\mathrm{HSIC}_{k_\rho}(X,Y) = \mathrm{dCov}_\rho^2(X,Y) = E[\|X-X'\|^q\|Y-Y'\|^q] - 2E[\|X-X'\|^q\|Y-Y''\|^q] + E[\|X-X'\|^q]\,E[\|Y-Y'\|^q].$$

■ 数値実験

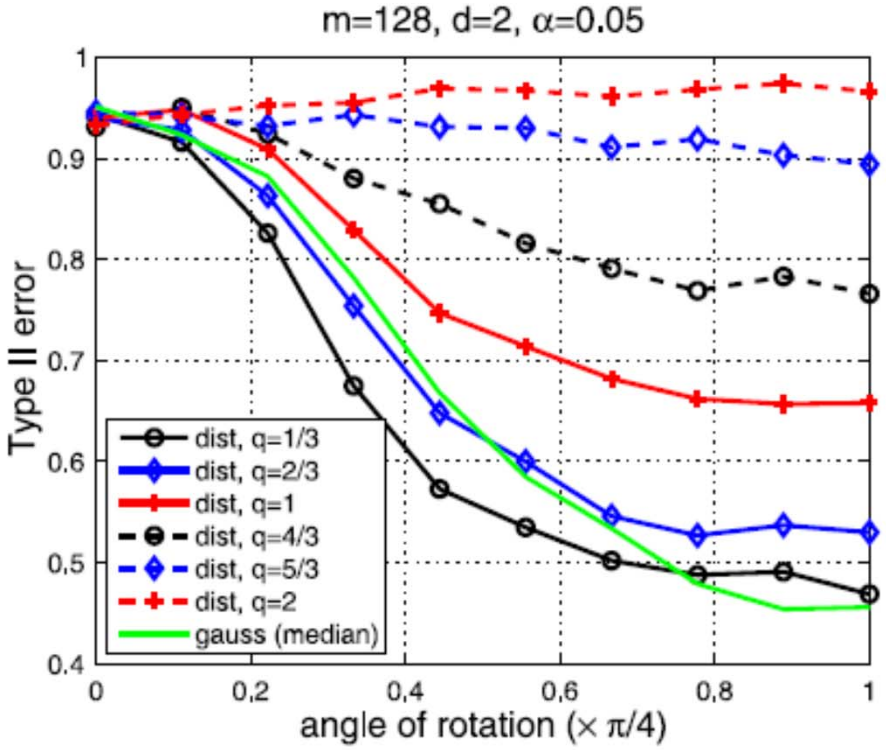
$\rho(z, z') = \|z - z'\|^q$

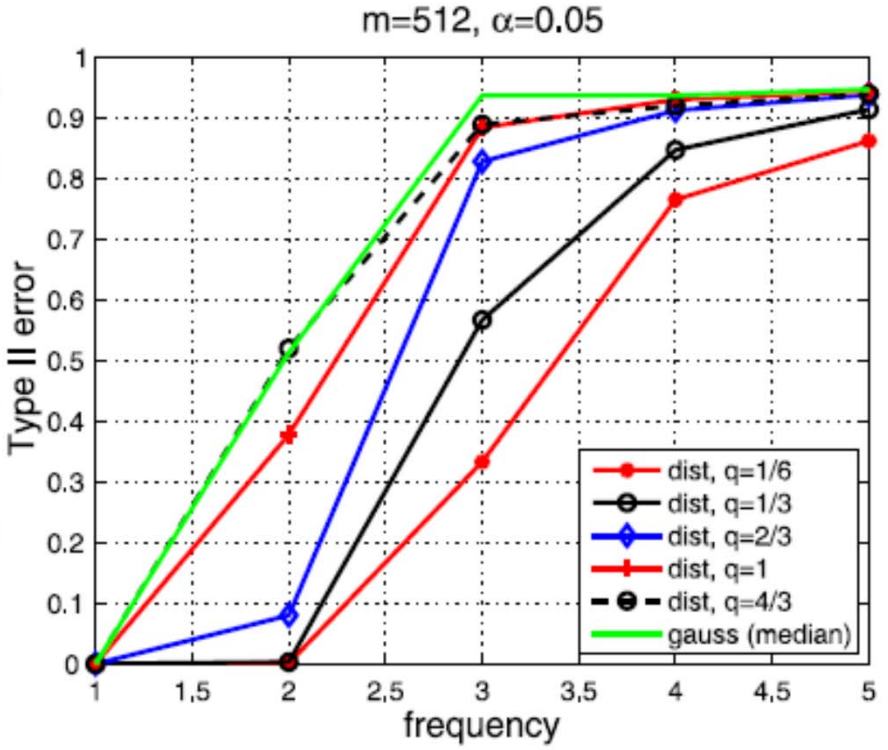


(A)

(B)

$p(x, y) \propto 1 + \sin(\ell x) \sin(\ell y)$





independent ⟷ dependent

easier ⟷ harder

カーネル選択

■ MMD/HSICによるノンパラメトリック検定の場合

– カーネルは特性的なものがよい.

– カーネルパラメータの選択

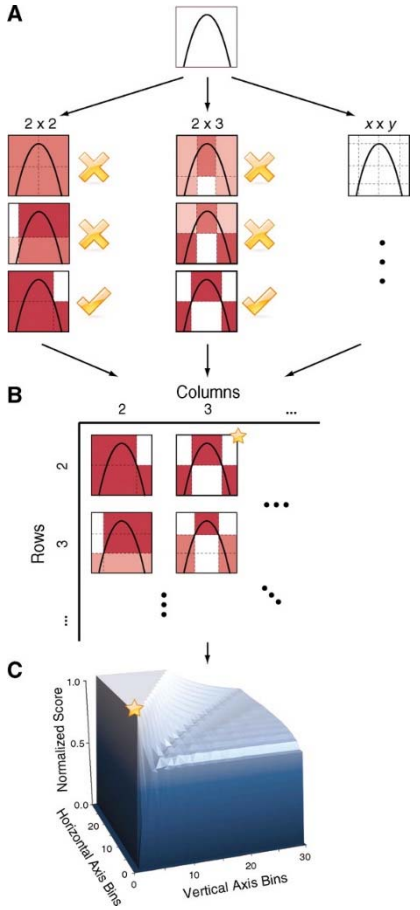
- 本来は，powerやefficiencyを用いるのがよい.
  しかし，必ずしも理論解析できない（参考：Gretton et al NIPS 2012)

- Heuristics
  - $\sigma = \mathrm{median}\left\{ \left\| X_i - X_j \right\| \mid i, j = 1, \ldots n \right\}$
  - supMMD

    $\sup_{\sigma} MMD_{emp}(X, Y)$ を検定統計量に用いる.

今後の研究が必要.

参考： MIC



– Maximal Information coefficient, MIC

  Reshef, et al. "Detecting Novel Associations in Large Data Sets". *Science 6062,* 2011.

  さまざまな分割の仕方を組み合わせて使う.

  “A Correlation for the 21st Century”

  by Terry Speed, *Science 6062,* 2011

– HSICとの比較

  http://www.slideshare.net/motivic/tokyo-r-lt-25759212

HSICの応用： Kernelized Sorting
(Quadrianto, Smola, Song, Tuytelaars IEEE PAMI 2010)

– マッチング問題： オブジェクトの対応を取る

- しかし，対応の情報（辞書）は全く使えないとする.

– 「各ドメイン内で」オブジェクト間の類似性の情報を使う.

Domain A: $X_1, \ldots, X_n$.　　Domain B: $Y_1, \ldots, Y_n$

Permutation π で, X_i と $Y_{\pi(i)}$ $(i = 1, \ldots, n)$ が対応するものを探す.

カーネル k_X, k_Y によるグラム行列 K_X, K_Y は計算可能と仮定.

– HSIC = グラム行列の類似尺度

$$HSIC_{emp}(\mathbf{X_n}, \mathbf{Y_n}) = \frac{1}{n^2}\mathrm{Tr}[\widetilde{K}_X \widetilde{K}_Y]$$

– Kernelized Sorting

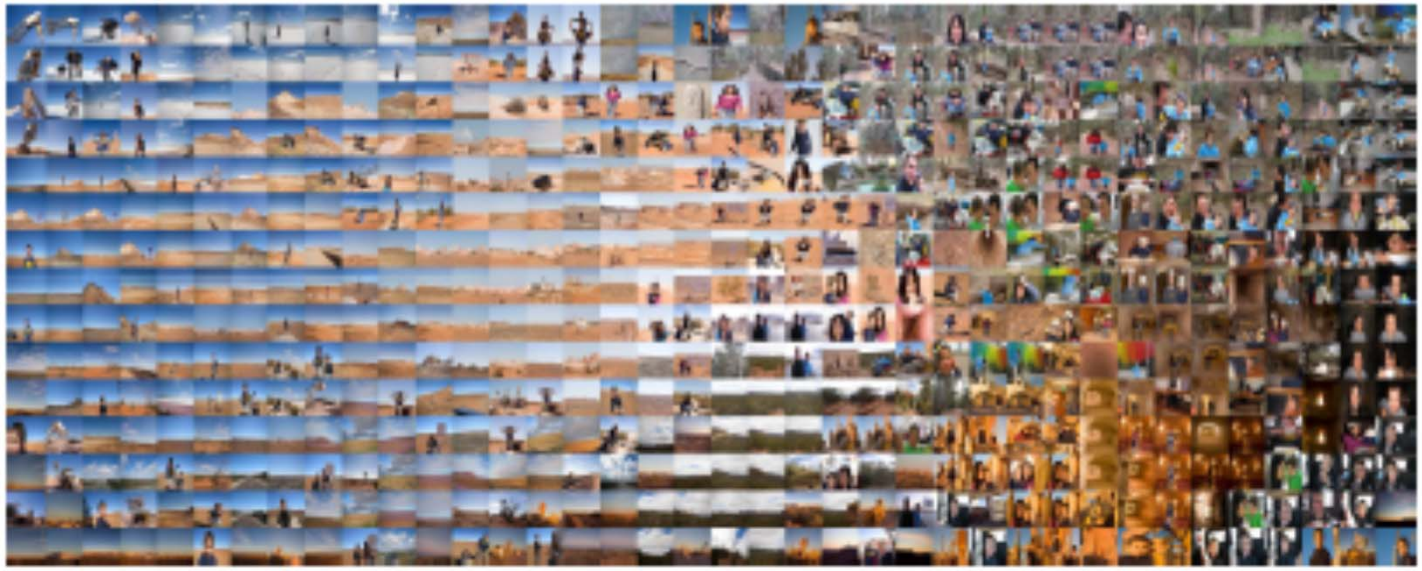
$$\max_{\pi \in \Pi_n} HSIC_{emp}(\mathbf{X_n}, \mathbf{Y_n^\pi}) = \max_{\pi} \mathrm{Tr}[\widetilde{K}_X A_\pi^T \widetilde{K}_Y A_\pi] \qquad A_\pi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

or

$$\max_{\pi} \|\widetilde{K}_X A_\pi^T - \widetilde{K}_Y A_\pi\|_F^2$$

A_πを [0, 1]値の確率行列に緩和して解く – Convex Kernelized Sorting (Djuric et al AAAI2012)

Fig. 4: Layout of 570 images into a 2D grid of size 15 by 38 using bag-of-visual-words based Kernelized Sorting. Several object categories, like books, cars, planes, and people are grouped into proximal locations.



(a) Photos summarization by color based Kernelized Sorting.

(Quadrianto et al. IEEE PAMI 2010)

条件付独立性

条件付共分散作用素

(X, Y, Z): $\Omega_x \times \Omega_y \times \Omega_z$に値を取る確率変数.

k_x, k_y, k_z: $\Omega_x, \Omega_y, \Omega_z$上の正定値カーネル

条件付相互共分散作用素

$$\Sigma_{YX|Z} := \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$$

– 分解: $\Sigma_{YZ} = \Sigma_{YY}^{1/2} W_{YZ} \Sigma_{ZZ}^{1/2}$, $\|W_{YZ}\| \leq 1$ (Baker 1973).

W_{YZ} は相関を表す. $W_{YZ} = \Sigma_{YY}^{-1/2}\Sigma_{YZ}\Sigma_{ZZ}^{-1/2}$.

条件付独立性

- すべてのカーネルは特性的とする.

  $\Sigma_{YX|Z} = O$ は条件付独立 $X \perp Y \mid Z$ より弱い（次ページ参照）.

  $\Sigma_{Y(X,Z)|Z} = O$ if and only if $X \perp Y \mid Z$.

  (X, Z): 変数の組み. 積カーネル $k_X k_Z$を用いる.

- 条件付独立性尺度:

$$\mathrm{HSCONIC}(X, Y|Z) \coloneqq \left\|\Sigma_{(Y,Z)(X,Z)|Z}\right\|_{\mathrm{HS}}^2$$

- 推定量:

$$\mathrm{HSCONIC}_{\mathrm{emp}}(X, Y|Z) \coloneqq \mathrm{Tr}[G_{\tilde{X}} G_{\tilde{Y}} - 2G_{\tilde{X}} R_Z G_{\tilde{Y}} + G_{\tilde{X}} R_Z G_{\tilde{Y}} R_Z]$$

$$R_Z \coloneqq G_Z (G_Z + n\epsilon_n I_n)^{-1}.$$

ϵ_n: regularization coefficient

定理（福水 定理9．６）

k_z が特性的なカーネルであるとき，

$$\langle g, \Sigma_{YX|Z} f \rangle = E\left[\mathrm{Cov}[f(X), g(Y)|Z]\right] \qquad (\forall f \in H_x, g \in H_y)$$

c.f.

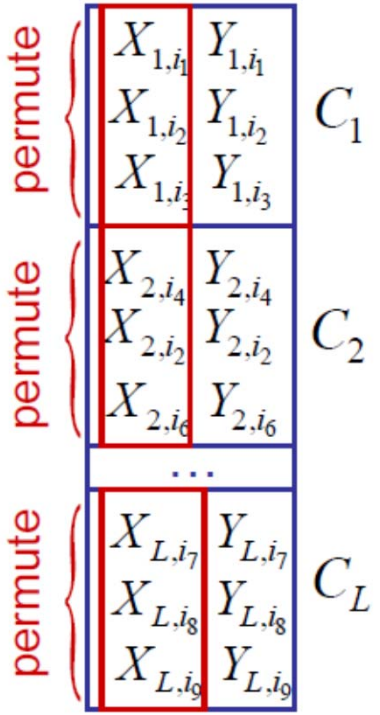
$$X \perp Y|Z \quad \Leftrightarrow \quad \mathrm{Cov}[f(X), g(Y)|Z = z] = 0$$

for any square integrable f, g, and a.e. z.

条件付独立性検定

– HSICとは違い，漸近的な帰無分布の形は知られていない．
（正則化の扱いが難しい）

– 並べ替え検定／リサンプリングは，Z が連続変数の場合，簡単ではない．
  - Zの値の離散化が必要
    → 厳密な条件付独立性とシミュレートしていない．



正規化共分散作用素による依存性尺度

共分散作用素の正規化

– 命題 （Baker 1973)

相互共分散作用素 $\Sigma_{YX}: H_x \to H_y$ に対し，作用素 $W_{YX}: H_x \to H_y$ で $\|W_{YX}\| \leq 1$,

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2}$$

かつ，$R(W_{YX}) \subset \overline{R\,(\Sigma_{YY})}$, $N(W_{YX})^{\perp} \subset \overline{R(\Sigma_{XX})}$ となるものが一意に存在する.

– 正規化相互共分散作用素（Normalized Cross-Covariance Operator）

NOCCO $\quad W_{YX} = \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2}$

「相関」を表す作用素

NOCCOによる依存性尺度

Characterization of independence

With characteristic kernels,

$$W_{YX} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

Dependence measure

Assume W_{XY} etc. are Hilbert-Schmidt.

$$HSNIC = \|W_{YX}\|_{HS}^2$$

Kernel-free Integral Expression

Theorem (Fukumizu et al. NIPS 21, 2008)

Assume

P_{XY} have density $p_{XY}(x, y)$

$H_X \otimes H_Y$ is characteristic.

W_{YX} is Hilbert-Schmidt.

Then,

$$\| W_{YX} \|_{HS}^2 = \iint \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} - 1 \right)^2 p_X(x) p_Y(y) dx dy$$

- Kernel-free expression, though the definitions are given by kernels!
- χ^2-divergence (aka mean square contingency) is expressed by kernels!
- A conditional version exists.

Empirical Estimator

- Empirical estimation is straightforward with the empirical cross-covariance operator $\hat{\Sigma}_{YX}^{(N)}$.
- Inversion → regularization: $\Sigma_{XX}^{-1} \rightarrow \left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon I\right)^{-1}$
- Replace the covariances in $W_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ by the empirical ones given by the data $\Phi_X(X_1),\ldots, \Phi_X(X_N)$ and $\Phi_Y(Y_1),\ldots, \Phi_Y(Y_N)$

$$HSNIC_{emp} = \mathrm{Tr}[R_X R_Y] \qquad \text{(dependence measure)}$$

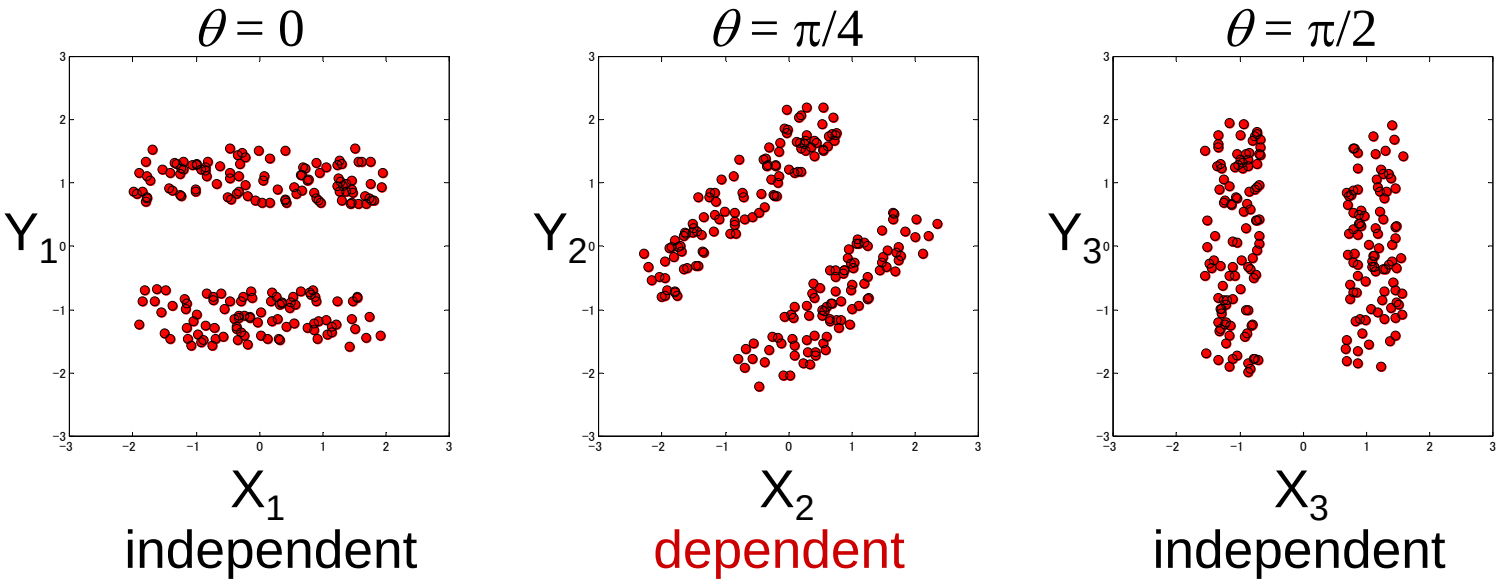
where $R_X \equiv G_X \left(G_X + N\varepsilon_N I_N\right)^{-1}$

$$G_X = \left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)K_X\left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right) \qquad K_X = \left(k(X_i, X_j)\right)_{i,j=1}^{N}$$

- $\mathrm{HSNIC}_{\mathrm{emp}}$ gives a new kernel estimator for the χ^2-divergence. Consistency is known.

Application to Independence Test

- Toy example



They are all uncorrelated, but dependent for $0 < \theta < \pi/2$

N = 200.
Permutation test is used for independence test except contingency table.

	indep. → more dependent					
Angle	0.0	4.5	9.0	13.5	18.0	22.5
HSIC (Median)	93	92	63	5	0	0
HSIC (Asymp. Var.)	93	44	1	0	0	0
HSNIC ($\varepsilon = 10^4$, Median)	94	**23**	**0**	**0**	**0**	**0**
HSNIC ($\varepsilon = 10^6$, Median)	92	**20**	**1**	**0**	**0**	**0**
HSNIC ($\varepsilon = 10^8$, Median)	93	**15**	**0**	**0**	**0**	**0**
HSNIC (Asymp. Var.)	94	**11**	**0**	**0**	**0**	**0**
MI (#NN = 1)	93	62	11	0	0	0
MI (#NN = 3)	96	43	0	0	0	0
MI (#NN = 5)	97	49	0	0	0	0
Power Diverg. (#Bins=3)	96	92	43	9	1	0
Power Diverg. (#Bins=4)	98	29	0	0	0	0
Power Diverg. (#Bins=5)	94	60	2	0	0	0

acceptance of independence out of 100 tests (significance level = 5%)

まとめ

■ カーネル平均による分布の表現

- カーネル平均埋め込み： 特徴ベクトルの平均により分布を一意に表現することが可能である – 特性的なカーネル.
- 密度関数の推定などの従来の方法と異なる，ノンパラメトリック推論の方法が構築できる.
- 再生性により，推定量の計算が容易である.

■ 共分散作用素

- 2つの確率変数の関係は，RKHS上の共分散作用素により表現できる.
- 特性的なカーネルを用いると，共分散作用素＝0により独立性を特徴づけられる.
- 独立性尺度： HSIC＝共分散作用素のHSノルム^2
- 独立性検定の新しい統計量としてHSICが有効である.

■ 条件付独立性

- 条件付共分散作用素により，条件付き独立性を特徴づけることが可能である.
- 多次元連続変数の場合の条件付独立性検定は，困難を伴う.

■ 正規化された相互共分散作用素

- HSノルムがχ^2-divergence（カーネルに依存しない量）を表す.

References

- Gretton, A., K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola; A Kernel Two-Sample Test. *Journal of Machine Learning Research* 13(Mar):723–773, 2012.
- Gretton, A., Bousquet, O., Smola, A., and Schoelkopf, B., Measuring Statistical Dependence with Hilbert-Schmidt Norms, *Proc. 16th International Conference on Algorithmic Learning Theory*: Springer-Verlag, 2005.
- Gretton, A., K. Fukumizu, C.-H. Teo, L. Song, B. Scholkopf, A. Smola. A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems 20*, 585-592, MIT Press (2008).
- Fukumizu, K., L. Song, A. Gretton (2014) Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *Journal of Machine Learning Research. 14:3753–3783.*
- Gretton, A., K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, A. Smola. A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems 20*, 585-592, MIT Press (2008).
- Gretton, A., Z. Harchaoui, K. Fukumizu, B. Sriperumbudur. A Fast, Consistent Kernel Two-Sample Test. *Advances in Neural Information Processing Systems 22*, 673-681, MIT Press (2010)

- A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu (2012) Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pp.1214-1222
- Székely, G. J. Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing independence by correlation of distances, *Annals of Statistics,* 35/6, 2769–2794
- Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics* 41, 5, 2263-2702
- Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C. (2011). "Detecting Novel Associations in Large Data Sets". *Science* **334** (6062): 1518–1524.
- Fukumizu, K., A. Gretton, X. Sun, and B. Scholkopf: Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 20*, 489-496, MIT Press (2008).