Measuring Dependence and Conditional Dependence with Kernels

Kenji Fukumizu

The Institute of Statistical Mathematics, Japan

June 25, 2014.

ICML 2014, Causality Workshop

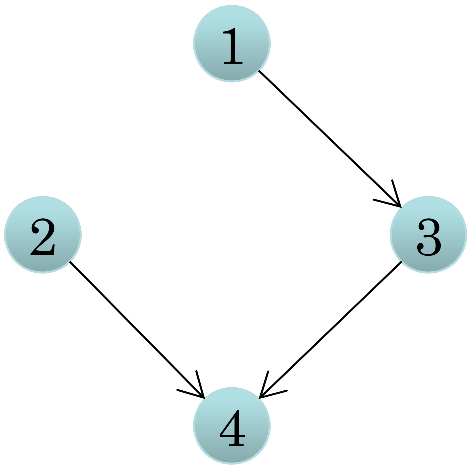
Introduction

Dependence Measures

- Dependence measures and causality
  - Constraint methods for causal structure learning are based on measuring or testing (conditional) dependence.

e.g. PC Algorithm (Spirtes et al. 1991, 2001)

(Conditional) independence tests with χ^2-tests.



$$X_1 \perp X_2$$

$$X_1 \perp X_4 \mid X_3$$

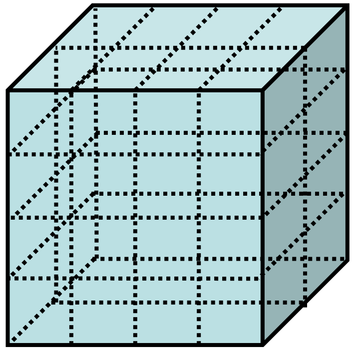
$$X_2 \not\perp X_3 \mid X_4$$

etc.

■ Problems

– Tests for structure learning may involve many variables.

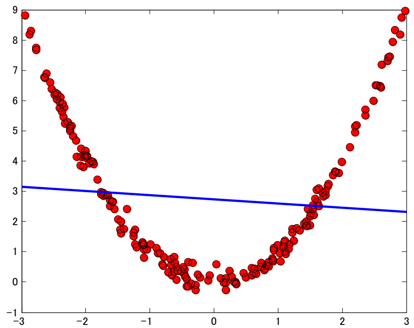
– (Conditional) independence test for continuous, high-dimensional domains are not easy.



- Discretization causes many bins, requiring a large data size.
- Nonparametric methods are often weak for high-dimensionality.

  KDE, smoothing kernel, ...

– Linear correlations may not be sufficient for complex relations.



- This talk
  - As building blocks of causal learning, kernel methods for measuring (in)dependence and conditional (in)dependence are discussed.

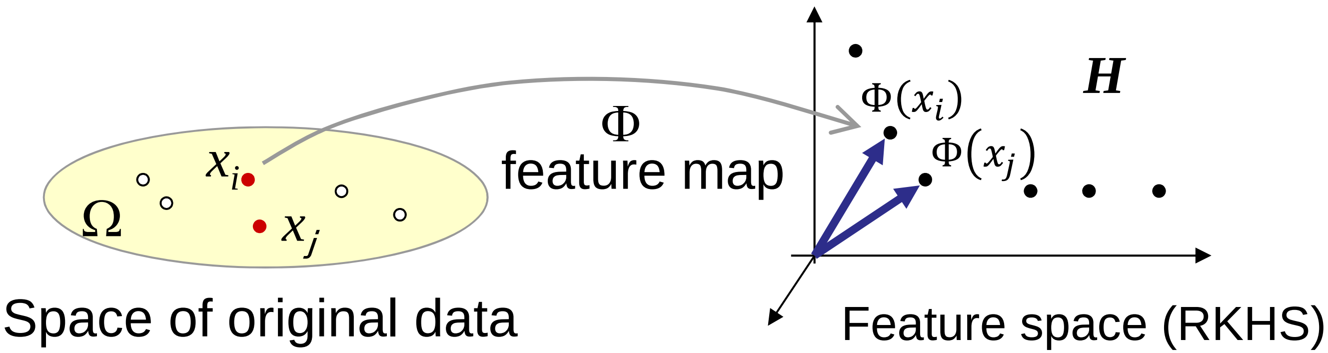
Outline

1. Introduction
2. Kernel measures for independence
3. Relations with distance covariance
4. How to choose a kernel
5. Conditional independence
6. Conclusions

Kernel measures for independence

Kernel methods

Feature map and kernel methods



Do linear analysis in the feature space.

– Feature map

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

Feature vectors

$$X_1 \dots, X_n \mapsto \Phi(X_1), \dots, \Phi(X_n)$$

Do kernel methods work well for high dimensional data?

– Empirical comparison: pos. def. kernel and smoothing kernel

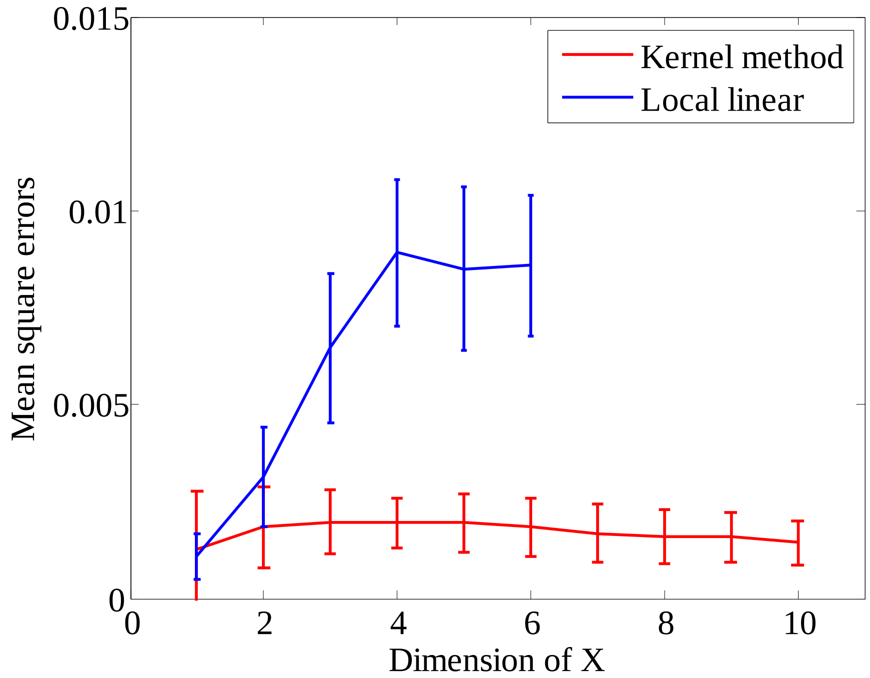
Nonparametric regression

$$Y = 1/(1.5 + ||X||^2) + Z, \qquad X \sim N(0, I_d),\ Z \sim N(0, 0.1^2)$$

- Kernel ridge regression
  (Gaussian kernel)
- Local linear regression
  (Epanechnikov kernel,
  'locfit' in R is used)

n = 100, 500 runs

Bandwidth parameters are chosen by CV.



– Theory?

Representing probabilities

X: random variable taking values on Ω. k: pos. def. kernel on Ω.

Feature map defines a RKHS-valued random variable $\Phi(X)$.

The kernel mean $E[\Phi(X)]$ represents the probability distribution of X.

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

– Kernel mean can express higher-order moments of X.

Suppose $k(u,x) = c_0 + c_1 ux + c_2(ux)^2 + \cdots \qquad (c_i \geq 0)$,   e.g., e^{ux}

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

c.f. moment generating function

Comparing two probabilities

- MMD (Maximum Mean Discrepancy, Gretton et al 2005)

  $X \sim P$, $Y \sim Q$ (two probabilities on Ω). k: pos. def. kernel on Ω.

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

$$= \sup_{\|f\|=1, f\in H} |\langle m_X - m_Y, f\rangle_H|^2$$

$$= \sup_{\|f\|=1, f\in H} |E[f(X)] - E[f(Y)]|^2$$

Comparing the moments through various functions

  – Characteristic kernels are defined so that

$$\mathrm{MMD}(P,Q) = 0 \quad \text{if and only if} \quad P = Q.$$

  e.g. Gaussian and Laplace kernels

  Kernel mean m_X determines the distribution of X uniquely.

  MMD is a metric on the probabilities.

HSIC: Independence measure

- Hilbert-Schmidt Independence Criterion (HSIC)

(X, Y) : random vector taking values on $\Omega_X \times \Omega_Y$.

(H_X, k_X), (H_Y, k_Y): RKHS on Ω_X and Ω_Y, resp.

Compare the joint probability P_{YX} and the product of the marginal $P_Y P_X$

Def.
$$\mathrm{HSIC}(X,Y) \coloneqq \mathrm{MMD}^2(P_{YX}, P_Y P_X)$$
$$= \|m_{YX} - m_Y \otimes m_X\|^2_{H_X \otimes H_Y}$$

Theorem

Assume: product kernel $k_X k_Y$ is characteristic on $\Omega_X \times \Omega_Y$.

$\mathrm{HSIC}(X,Y) = 0$ if and only if $X \perp Y$

Covariance operator

Operator expression:

$$\langle m_{YX} - m_Y \otimes m_X, g \otimes f \rangle_{H_Y \otimes H_X} = E[f(X)g(Y)] - E[f(X)]E[g(Y)]$$

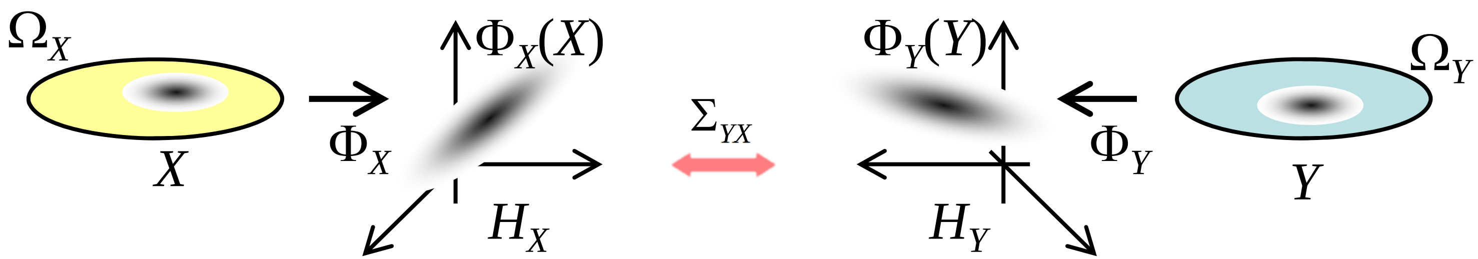
<u>Def.</u> covariance operators $\Sigma_{YX}: H_X \to H_Y$, $\Sigma_{XX}: H_X \to H_X$

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

$$\langle h, \Sigma_{XX} f \rangle_{H_X} = E[f(X)h(X)] - E[f(X)]E[h(X)] \qquad (\forall f, h \in H_X)$$

Simply, extension of covariance matrix (linear map)

$$V_{YX} = E[YX^T] - E[Y]E[X^T], \quad b^T V_{YX} a = E[b^T Y \cdot a^T X] - E[b^T Y]E[a^T X]$$



Expressions of HSIC

- $\mathrm{HSIC}(X,Y) = \|\Sigma_{YX}\|_{HS}^2$ Hilbert-Schmidt norm (same as Frobenius norm)

  $$\|A\|_{HS}^2 := \sum_i \sum_j \langle \psi_j, A\phi_i \rangle^2$$

  $A: H \to G$. $\{\phi_i\}_i, \{\psi_j\}_j$: ONB of H and G, (resp).

- $$\mathrm{HSIC}(X,Y) = E[k_X(X,X')k_Y(Y,Y')] - 2E[k_X(X,X')k_Y(Y,Y'')] + E[k_X(X,X')]E[k_Y(Y,Y')]$$

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

- Empirical estimator (Gram matrix expression)

  $$\mathrm{HSIC}_{emp}(X,Y) = \frac{1}{n^2} Tr[Q_n G_X Q_n G_Y]$$ → Test statistic

  $G_{X,ij} = k_X(X_i, X_j)$, $G_{Y,ij} = k_Y(Y_i, Y_j)$, $Q_n := I_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T$ (centering)

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

Independence test with HSIC

Theorem: null distribution (Gretton, Fukumizu, etc. NIPS2007)

If X and Y are independent, then

$$n\,\mathrm{HSIC}_{emp}(X,Y) \overset{\text{law}}{\Longrightarrow} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \qquad (n \to \infty).$$

where Z_i : i.i.d. ~ $N(0,1)$,

$\{\lambda_i\}_{i=1}^{\infty}$ is the eigenvalues of an integral operator.

Theorem: consistency of test (Gretton, Fukumizu, etc. NIPS2007)

If $HSIC(X,Y) \neq 0$, then

$$\sqrt{n}\left(\mathrm{HSIC}_{\mathrm{emp}}(X,Y) - \mathrm{HSIC}(X,Y)\right) \overset{\text{law}}{\Rightarrow} N(0,\sigma^2) \qquad (n \to \infty).$$

where

$$\sigma^2 = 16\left(E_a\left[E_{b,c,d}[h(U_a,U_b,U_c,U_d]^2\right] - M_{YX}\right)$$

Independent test with HSIC:

– How to compute the critical region given significance level.

- Simulation of the null distribution (Gretton, Fukumizu et al NIPS2009).
  The eigenvalues can be estimated with the Gram matrices.

- Approximation with two-parameter Gamma by moment matching (Gretton, Fukumizu et al NIPS2007).

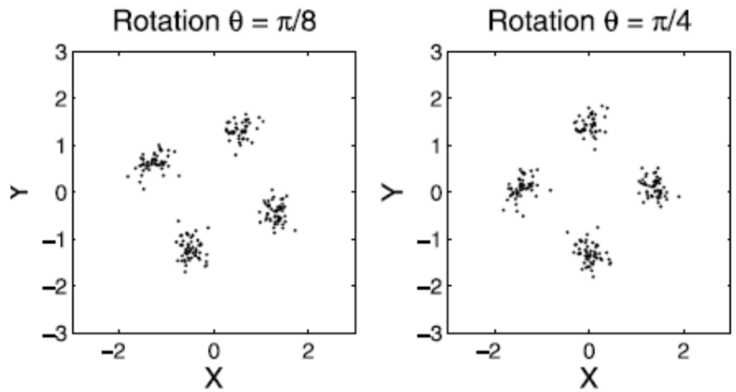
- Permutation test / Bootstrap
  Always possible, but time consuming.

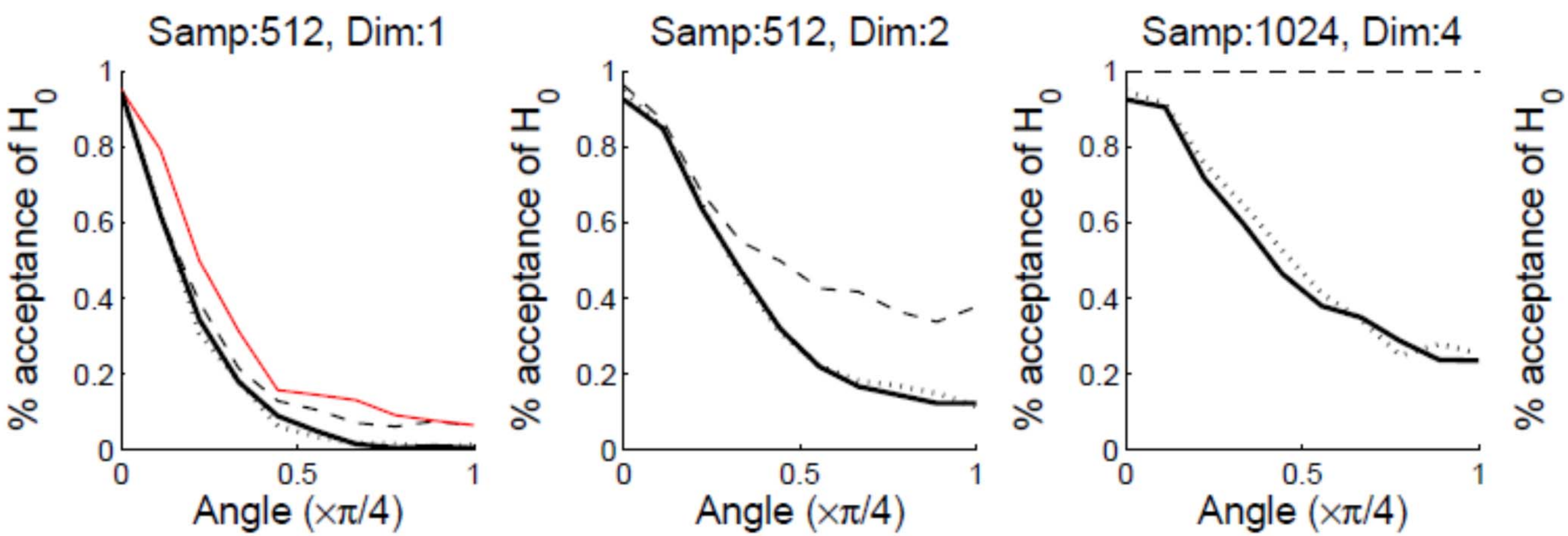
Experiments: independence test

X, Y: 1 dim + noise components

—— HSIC (Gamma approx.)

- - - Power divergence ($\lambda = 3/2$) with discretization (equi-probable)





Type II errors

Power divergence

Each dimension is partitioned into q parts.

Partitions $\{A_j\}_{j\in J}$. $(|J| = q^d)$

$$T_n^{(\lambda)} := \frac{2n}{\lambda(\lambda+2)} \sum_{j\in J} \hat{p}_j \left\{ \prod_{k=1}^{d} \left(\frac{\hat{p}_j}{\hat{p}_{j_k}^{(k)}} \right)^{\lambda} - 1 \right\}$$

$\hat{p}_j$: frequency in A_j

$\hat{p}_{j_k}^{(k)}$: marginal frequency in k-th dimension

$$T_n^{(\lambda)} \Rightarrow \chi^2_{q^d - qd + d - 1}$$

$\lambda = 0$: Mutual information

$\lambda = 2$: χ^2-divergence (mean square contingency)

Relation to distance covariance

Distance covariance

- Distance covariance (distance correlation) is a recent measure of independence for continuous variables (Székely, Rizzo, Bakirov, AoS 2007). It is very popular among statistical community.
- HSIC is closely related to (more general than, in fact) dCov.

Distance covariance

<u>Def.</u>

X, Y: random vectors (on Euclidean spaces)

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

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

Note: $\|X-X'\|$ is NOT positive definite.

For be a semi-metric ρ on Ω, ($\rho(z,z') = \rho(z',z)$, and $\rho(z,z') \geq 0$ with equality $z = z'$), define generalized distance covariance by

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

Theorem (Sejdinovic et al. AoS 2013). Assume ρ is of negative type, i.e.,

$$\textstyle\sum_{i=1}^n c_i \rho\left(z_i, z_j\right) \leq 0 \quad \text{for any } (c_i) \text{ with } \sum_{i=1}^n c_i = 0.$$

Then, $k(z,z') \coloneqq \frac{1}{2}\{\rho(z,z_0) + \rho(z',z_0) - \rho(z,z')\}$ is positive definite, and with k_X and k_Y induced by ρ_X and ρ_Y, resp.,

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

Example:

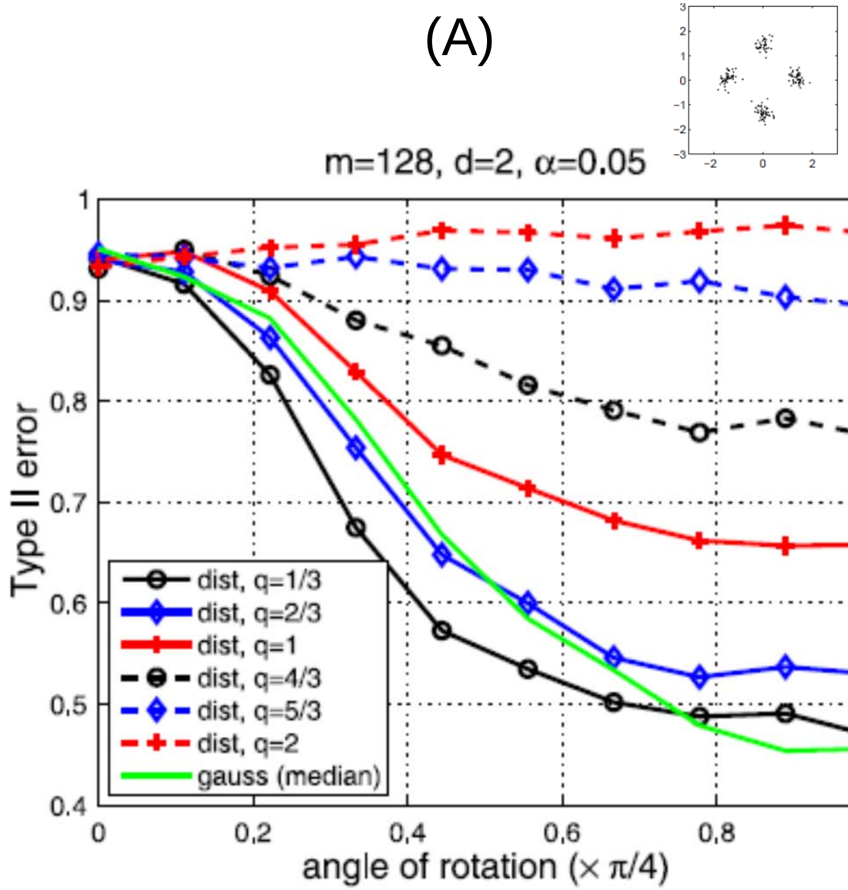
$$\rho(z,z') = \|z - z'\|^q \quad (0 < q \leq 2), \quad k_\rho(z,z') = \frac{1}{2}\{\|z\|^q + \|z'\|^q - \|z - z'\|^q\}$$

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

Experiments

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

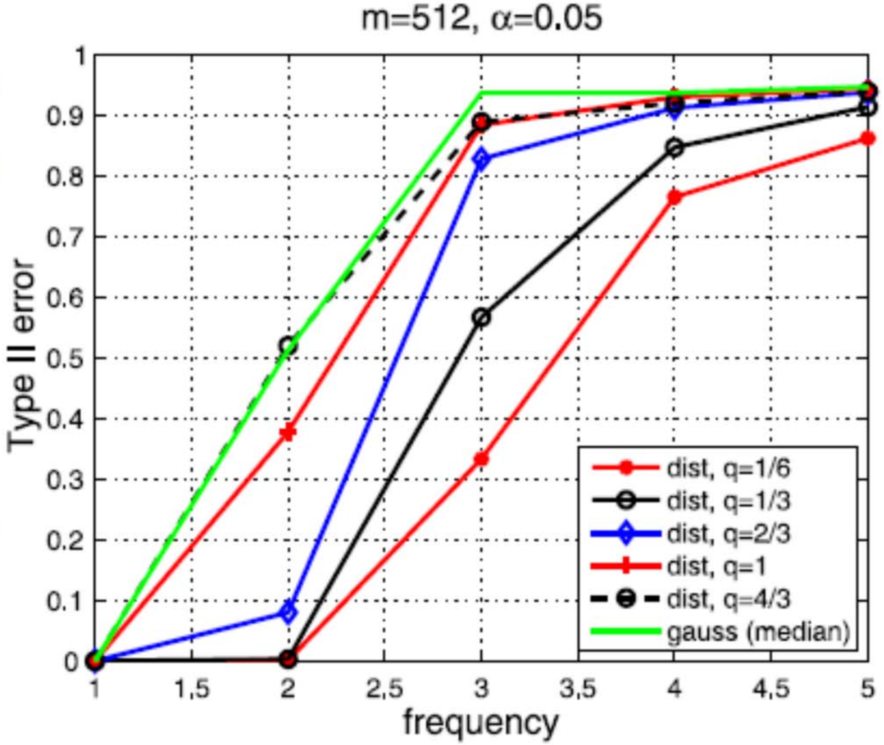
(A)



independent ⟷ dependent

(B)

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



easier ⟷ harder

How to choose a kernel

Kernel Choice

– The power of a test depends on the choice of kernels.

e.g. bandwidth σ in Gaussian kernel $\exp(-\frac{1}{2\sigma^2}\|x-y\|^2)$.

- Heuristics for σ: median of $\{\|X_i - X_j\|\}_{ij}$ (Gretton et al NIPS2006)
- Maximization of HSIC value (Sriperumbudur, Fukumizu et al. NIPS2009)

$$\sup_k HSIC^k_{emp}(X, Y)$$

  – No theoretical optimality, but empirically good.

- Power of the test (Gretton, Fukumizu, et al. NIPS 2010)
  – Developed for a simple version of MMD.
  – May be extended to HSIC.

Power of linear-time MMD test

Linear-time MMD

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

$\mathrm{MMD}_{emp}(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} \{ k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(Y_i, X_j) \}$

$\text{L-MMD}_{emp}(X, Y) := \frac{2}{n} \sum_{i=1}^{n/2} \{ k(X_{2i-1}, X_{2i}) + k(Y_{2i-1}, Y_{2i}) - k(X_{2i-1}, Y_{2i}) - k(Y_{2i-1}, X_{2i}) \}$

- Consistent estimator of MMD(X, Y).
  - Less accurate, but less computational cost
  - Easier asymptotics

$$\sqrt{n}(\text{L-MMD}_{emp}(X, Y) - MMD(X, Y)) \implies N(0, 2\sigma^2)$$

$$\sigma^2 = Var[h], \;\; h = k(X, X') + k(Y, Y') - k(X, Y') - k(Y, X')$$

Power of test

- $t_{k,\alpha}$ threshold for level α: $\Pr\left(LMMD_{emp} > t_{k,\alpha} \middle| H_0\right) = \alpha$.
- Under alternative ($MMD_k(X,Y) > 0$), the type II error is

$$\Pr\left(\mathrm{LMMD_{emp}} < t_{k,\alpha}\right) \to \Phi^{-1}\left(\Phi(1-\alpha) - \frac{\sqrt{n}\,\mathrm{MMD_k}(X,Y)}{\sqrt{2}\sigma_k^2}\right)$$

Φ: c.d.f. of $N(0,1)$.

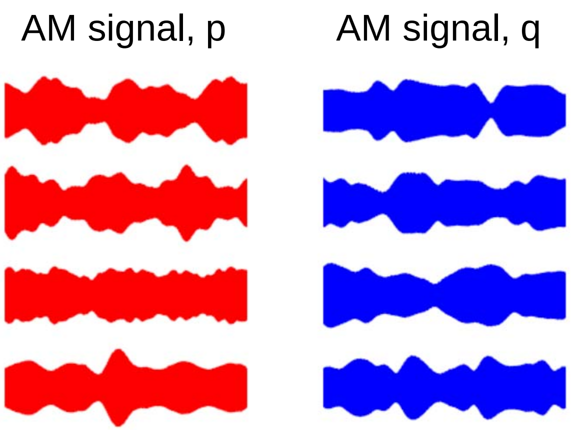
- To minimize the type II error, choose a kernel such that

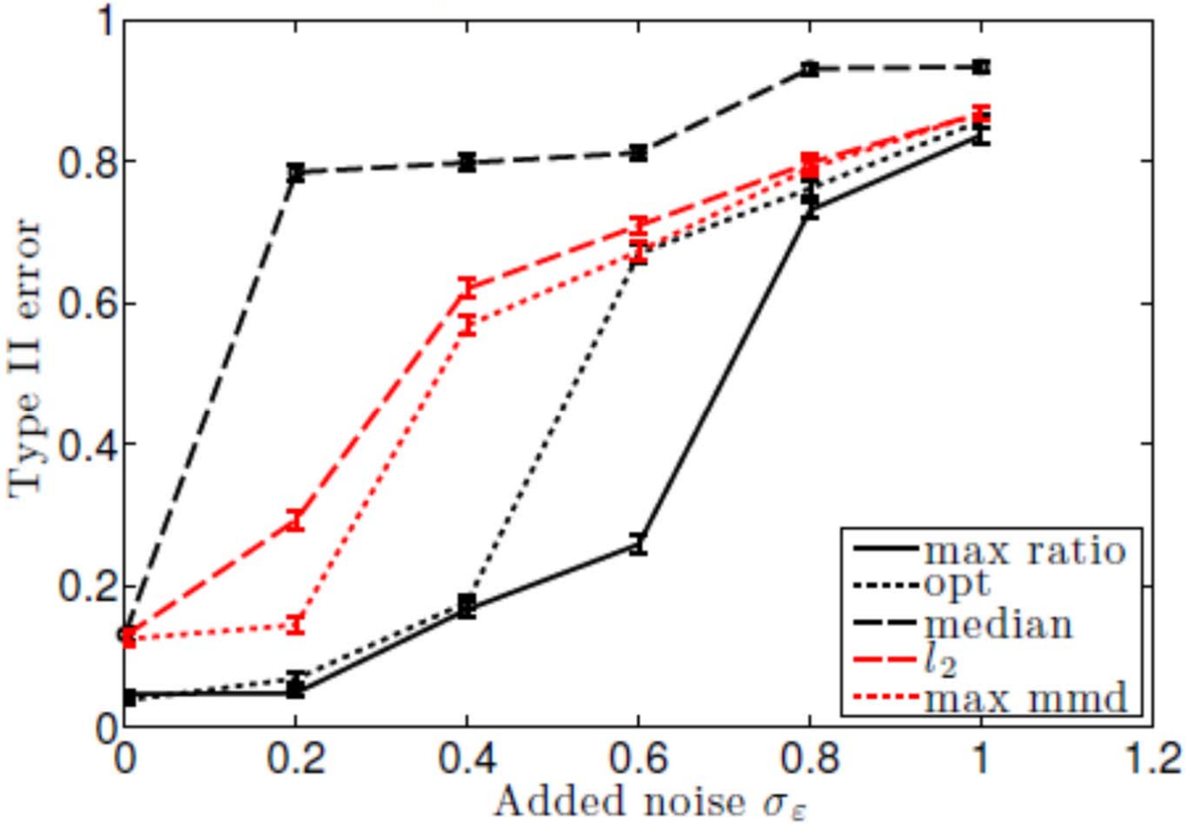
$$\max_{k\in F} \frac{\mathrm{LMMD}_{emp,k}(X,Y)}{\hat{\sigma}_k^2}$$

Experiment

– Two AM signals (songs with different instruments)

$$y(t) = \left(As(t) + o_{offset}\right)\cos(\omega_{carrier} t) + noise(t)$$





Gaussian kernels
with different bandwidth

Conditional independence

Conditional covariance

Conditional covariance operator

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

- Decomposition $\Sigma_{YZ} = \Sigma_{YY}^{1/2} W_{YZ} \Sigma_{ZZ}^{1/2}$ is possible with $\|W_{YZ}\| \le 1$ (Baker 1973).

  V_{YZ} is a "correlation" operator. c.f. $V_{YY}^{-1/2} V_{YZ} V_{ZZ}^{-1/2}$.

Conditional independence

– Assume kernels are characteristic.

$\Sigma_{YX|Z} = O$ is weaker than the cond. independence $X \perp Y \mid Z$.

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

paired variable: product kernel is used.

– Conditional independence measure:

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

– Empirical estimator:

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

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

ϵ_n: regularization coefficient

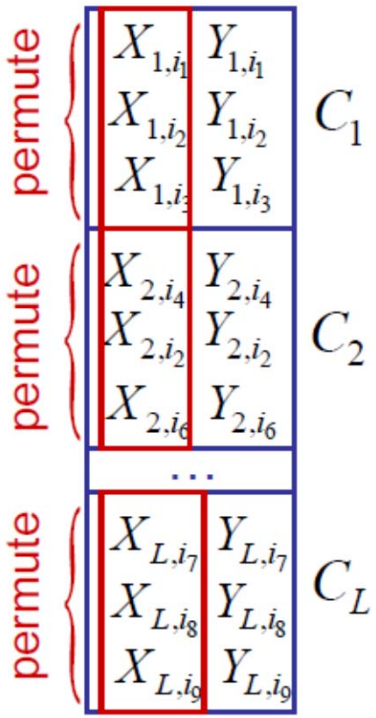
– The estimator is consistent, but the asymptotic distribution is NOT known.

  - Regularized inversion makes it difficult.

– Permutation test for continuous variables is not straightforward.

  - Discretization / neighbor data are needed to simulate the conditionally independent data.

    → Not rigorous conditional independence.



Conclusions

Dependence measures with kernels

- HSIC and HSCONIC are defined by the kernel mean embedding of probabilities.

- Show better performance than classical methods for high dimensional cases.
  - Theoretical backup is needed, but still open.

- As a special case, HSIC includes the distance covariance, which is a recent popular independence measure in statistics.

- For linear time MMD, a kernel can be chosen so that the power is maximized asymptotically.
  - Extension to other cases is needed.