Lecture 3: Advanced Applications of Probability Mappings to RKHS

Lille, 2014

Arthur Gretton

Gatsby Unit, CSML, UCL

Advanced topics

- Energy distance and the MMD
- Two-sample testing for big data
- Testing three-way interactions
- Bayesian inference without models

Energy Distance and the MMD

Energy distance is a special case of MMD

Two-sample testing:

- Energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, 2005]

$$D_E(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_{\mathbf{P}}\|x - x'\| + \mathbf{E}_{\mathbf{Q}}\|y - y'\| - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}}\|x - y\|$$

- Compare with MMD:

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F) = \mathbf{E}_{\mathbf{P}}k(x, x') + \mathbf{E}_{\mathbf{Q}}k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}}k(x, y)$$

Independence testing:

- $HSIC(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y)$ looks like distance covariance [Székely and Rizzo, 2009, Székely et al., 2007]

Energy distance is a special case of MMD

Two-sample testing:

- Energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, 2005]

$$D_E(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_{\mathbf{P}}\|x - x'\| + \mathbf{E}_{\mathbf{Q}}\|y - y'\| - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}}\|x - y\|$$

- Compare with MMD:

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F) = \mathbf{E}_{\mathbf{P}}k(x, x') + \mathbf{E}_{\mathbf{Q}}k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}}k(x, y)$$

Independence testing:

- $HSIC(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y)$ looks like distance covariance [Székely and Rizzo, 2009, Székely et al., 2007]

Energy distance is a special case of MMD

[Lyons, 2011][ICML12,AOS13]

Semimetrics of negative type

Semimetric: Let $\mathcal{Z}$ be a non-empty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$,

1. $\rho(z, z') = 0$ if and only if $z = z'$, and
2. $\rho(z, z') = \rho(z', z)$.

Not enforced: triangle inequality

Then $(\mathcal{Z}, \rho)$ is a semimetric space and ρ is a semimetric on $\mathcal{Z}$.

Semimetrics of negative type

Semimetric: Let $\mathcal{Z}$ be a non-empty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$,

1. $\rho(z, z') = 0$ if and only if $z = z'$, and
2. $\rho(z, z') = \rho(z', z)$.

Not enforced: triangle inequality

Then $(\mathcal{Z}, \rho)$ is a semimetric space and ρ is a semimetric on $\mathcal{Z}$.

Negative type: The semimetric space $(\mathcal{Z}, \rho)$ is said to have negative type if $\forall n \geq 2$, $z_1, \ldots, z_n \in \mathcal{Z}$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0. \tag{2}$$

Negative type, extension: If ρ satisfies (1), then so does ρ^q, for $0 < q < 1$.

Semimetrics and Hilbert spaces

Theorem: [Berg et al., 1984] ρ is a semimetric of negative type iff if there exists a Hilbert space $\mathcal{H}$ and an injective map $\varphi : \mathcal{Z} \to \mathcal{H}$, such that

$$\rho(z, z') = \left\| \varphi(z) - \varphi(z') \right\|_{\mathcal{H}}^2 .$$

Semimetrics and Hilbert spaces

Theorem: [Berg et al., 1984] ρ is a semimetric of negative type iff if there exists a Hilbert space $\mathcal{H}$ and an injective map $\varphi : \mathcal{Z} \to \mathcal{H}$, such that

$$\rho(z, z') = \left\| \varphi(z) - \varphi(z') \right\|_{\mathcal{H}}^2 .$$

Distance induced kernels: (positive definite)

$$k(z, z') = \frac{1}{2} \left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z') \right]$$

centred at $z_0 \in \mathcal{Z}$.

Semimetrics and Hilbert spaces

Theorem: [Berg et al., 1984] ρ is a semimetric of negative type iff if there exists a Hilbert space $\mathcal{H}$ and an injective map $\varphi : \mathcal{Z} \to \mathcal{H}$, such that

$$\rho(z, z') = \left\|\varphi(z) - \varphi(z')\right\|_{\mathcal{H}}^2 .$$

Distance induced kernels: (positive definite)

$$k(z, z') = \frac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right]$$

centred at $z_0 \in \mathcal{Z}$.

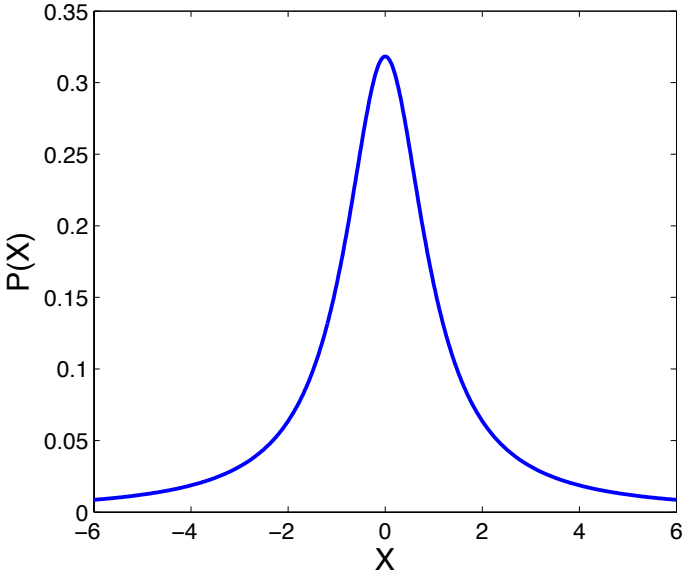
Special case: $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\rho_q(z, z') = \|z - z'\|^q$. Then ρ_q is a valid semimetric of negative type for $0 < q \leq 2$.

Energy distance when $q = 1$: just MMD with a particular kernel

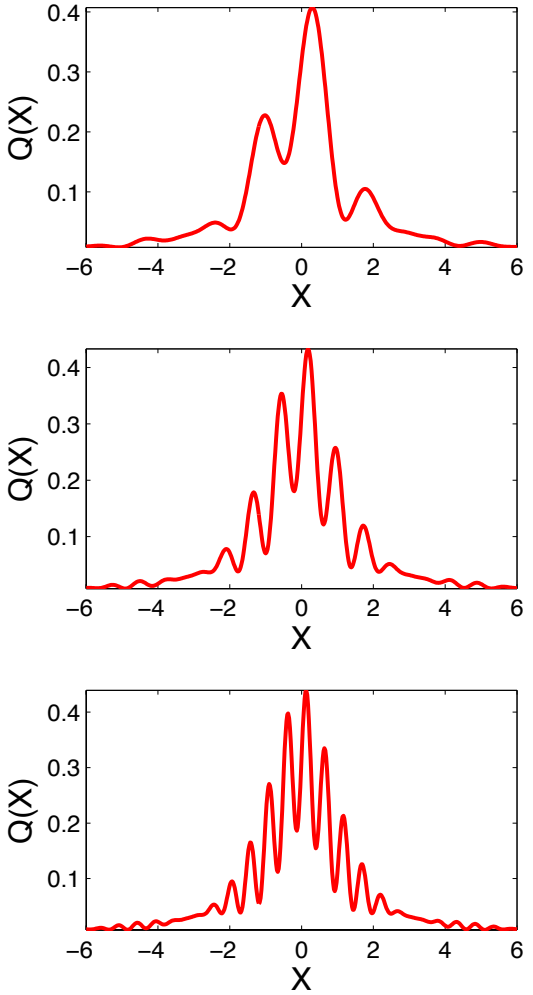
(nothing special about $q = 1$)

Two-sample testing benchmark

Two-sample testing example in 1-D:

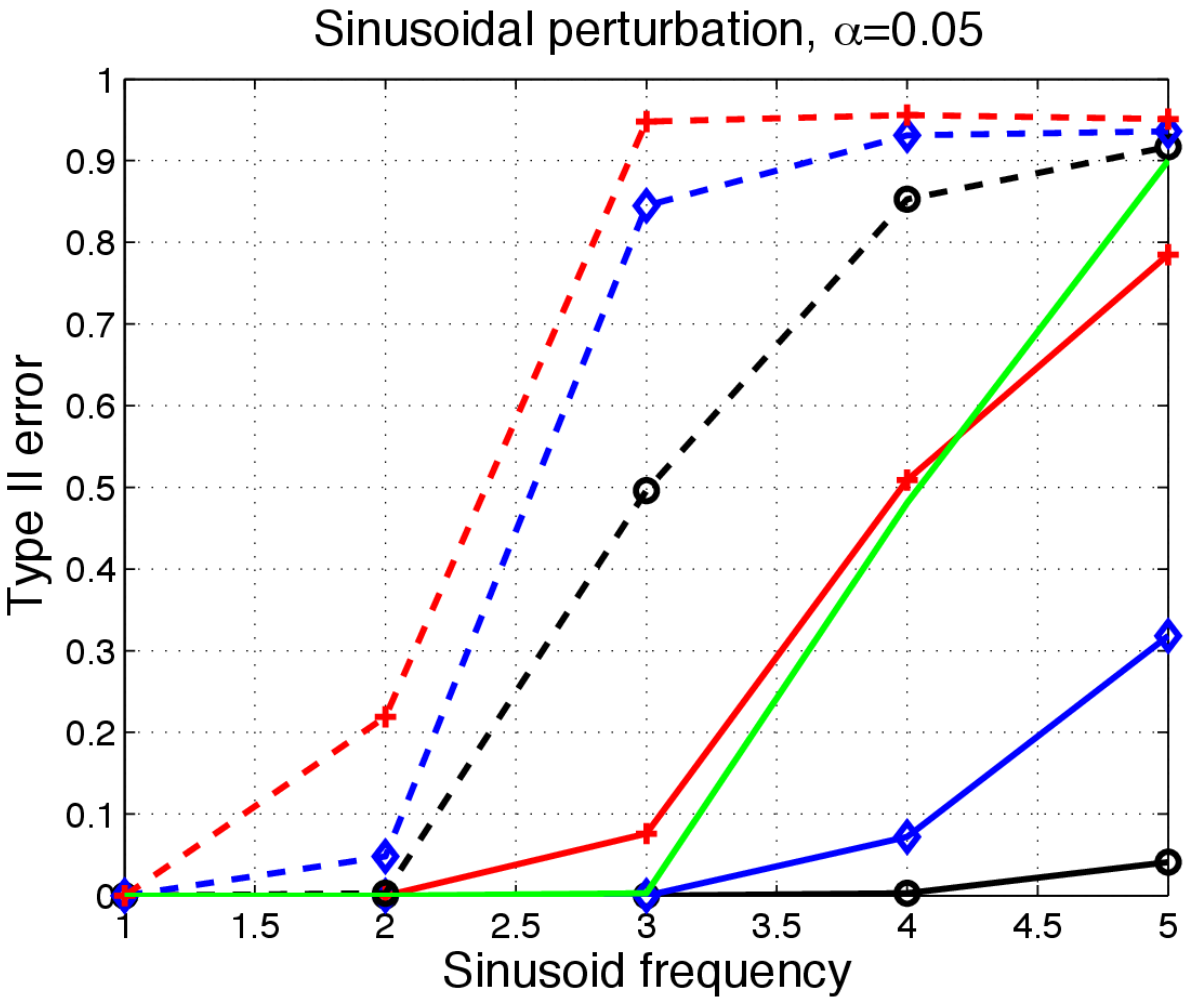


VS



Two-sample test, MMD with distance kernel

Obtain more powerful tests on this problem when $q \neq 1$ (exponent of distance)

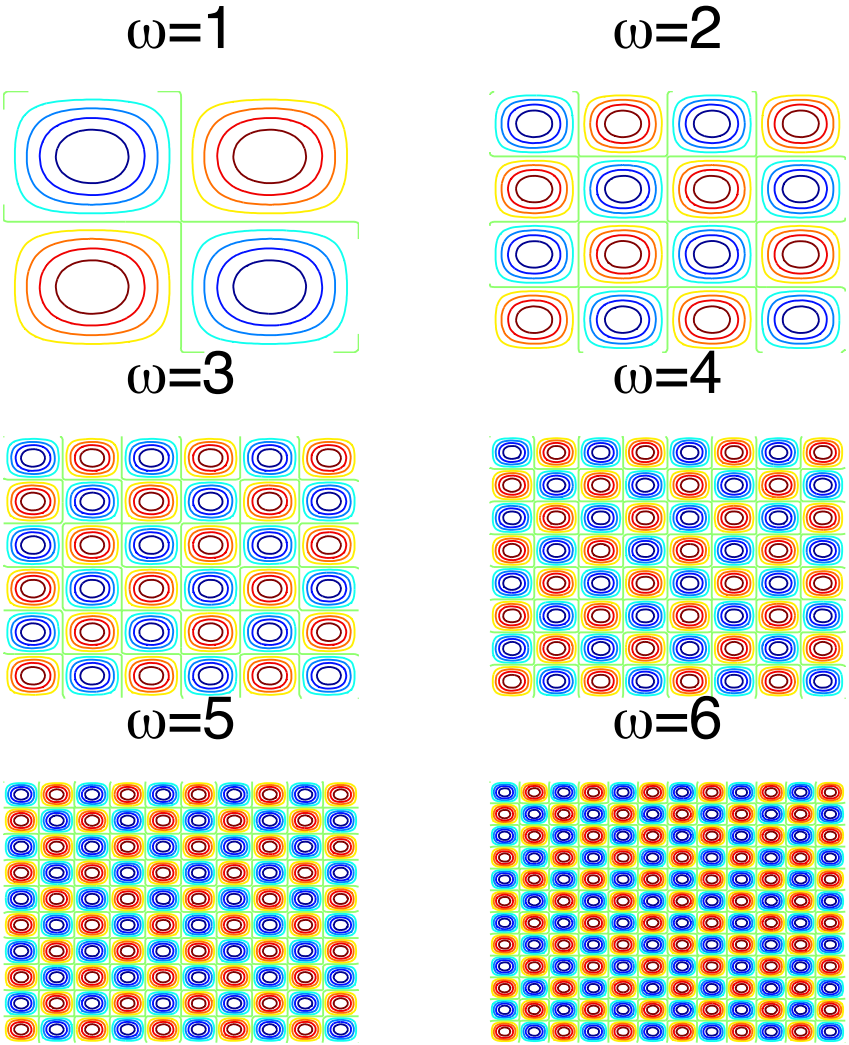


Key:

- Gaussian kernel
- $q = 1$
- Best: $q = 1/3$
- Worst: $q = 2$

Independence testing benchmark

Independence testing dataset

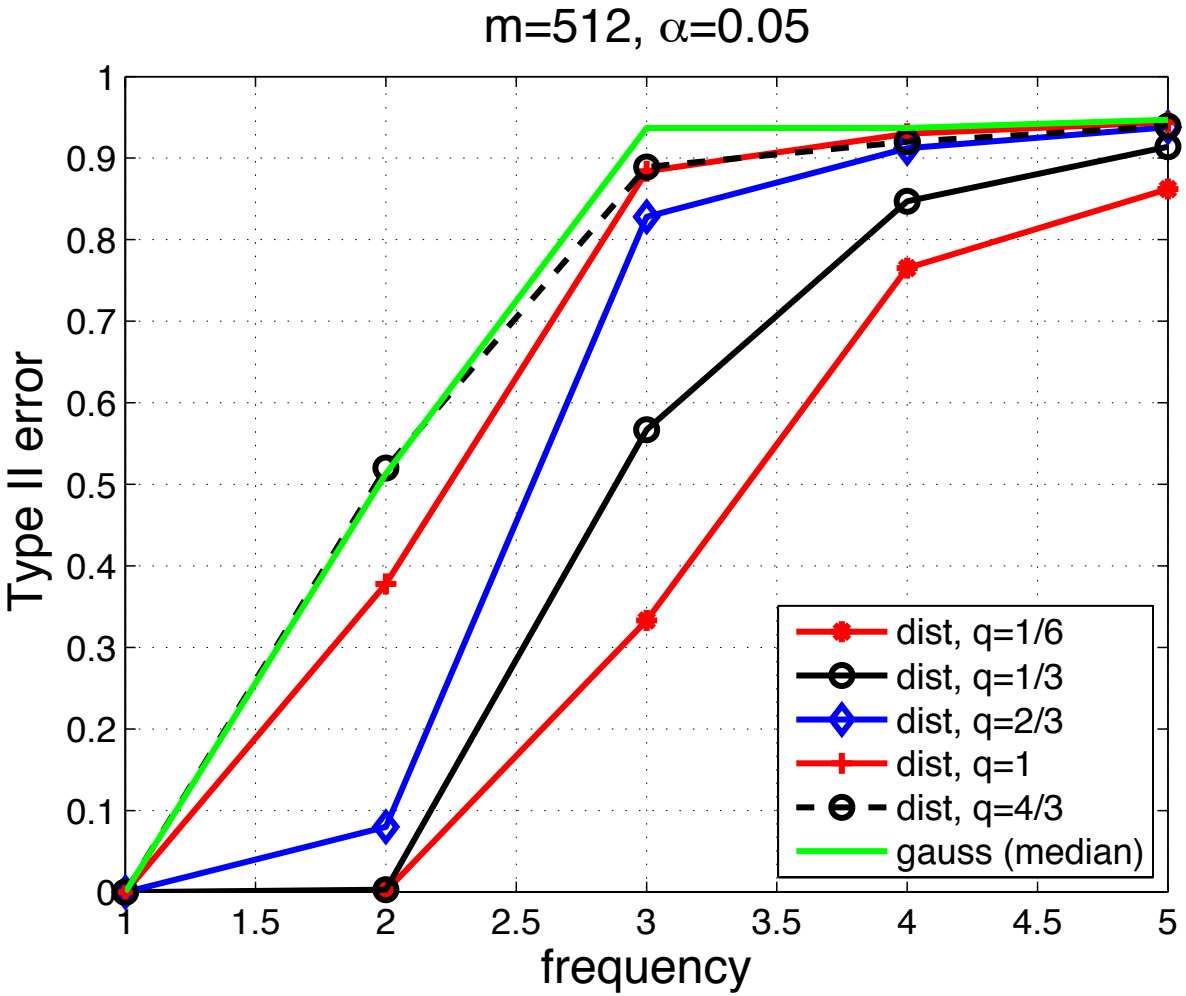


Density takes the form:

$$\mathbf{P}_{x,y} \propto 1 + \sin(\omega x)\sin(\omega y)$$

Independence test, HSIC with distance kernel

Obtain more powerful tests on this problem when $q \neq 1$ (exponent of distance)



Key:

- Gaussian kernel
- $q = 1$
- Best: $q = 1/6$
- Worst: $q = 4/3$

Kernel two-sample tests for big data, optimal kernel choice

Quadratic time estimate of MMD

$$\mathrm{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)$$

Quadratic time estimate of MMD

$$\text{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)$$

Given i.i.d. $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_m\}$ from $\mathbf{P}, \mathbf{Q}$, respectively:

The earlier estimate: (quadratic time)

$$\widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j)$$

Quadratic time estimate of MMD

$$\text{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)$$

Given i.i.d. $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_m\}$ from $\mathbf{P}, \mathbf{Q}$, respectively:

The earlier estimate: (quadratic time)

$$\widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j)$$

New, linear time estimate:

$$\begin{aligned} \widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') &= \frac{2}{m} \left[k(x_1, x_2) + k(x_3, x_4) + \ldots \right] \\ &= \frac{2}{m} \sum_{i=1}^{m/2} k(x_{2i-1}, x_{2i}) \end{aligned}$$

Linear time MMD

Shorter expression with explicit k dependence:

$$\text{MMD}^2 =: \eta_k(p, q) = \mathbb{E}_{xx'yy'} h_k(x, x', y, y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') - k(x', y),$$

and $v := [x,\ x',\ y,\ y']$.

Linear time MMD

Shorter expression with explicit k dependence:

$$\mathrm{MMD}^2 =: \eta_k(p, q) = \mathbb{E}_{xx'yy'} h_k(x, x', y, y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') - k(x', y),$$

and $v := [x,\ x',\ y,\ y']$.

The linear time estimate again:

$$\check{\eta}_k = \frac{2}{m} \sum_{i=1}^{m/2} h_k(v_i),$$

where $v_i := [x_{2i-1},\ x_{2i},\ y_{2i-1},\ y_{2i}]$ and
$h_k(v_i) := k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(x_{2i}, y_{2i-1})$

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence...
- ...a much less powerful test for a given m

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence...
- ...a much less powerful test for a given m

Advantages of the linear time MMD vs quadratic time MMD

- Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of χ^2)
- Both test statistic and threshold computable in $O(m)$, with storage $O(1)$.
- Given unlimited data, a given Type II error can be attained with less computation

Asymptotics of linear time MMD

By central limit theorem,

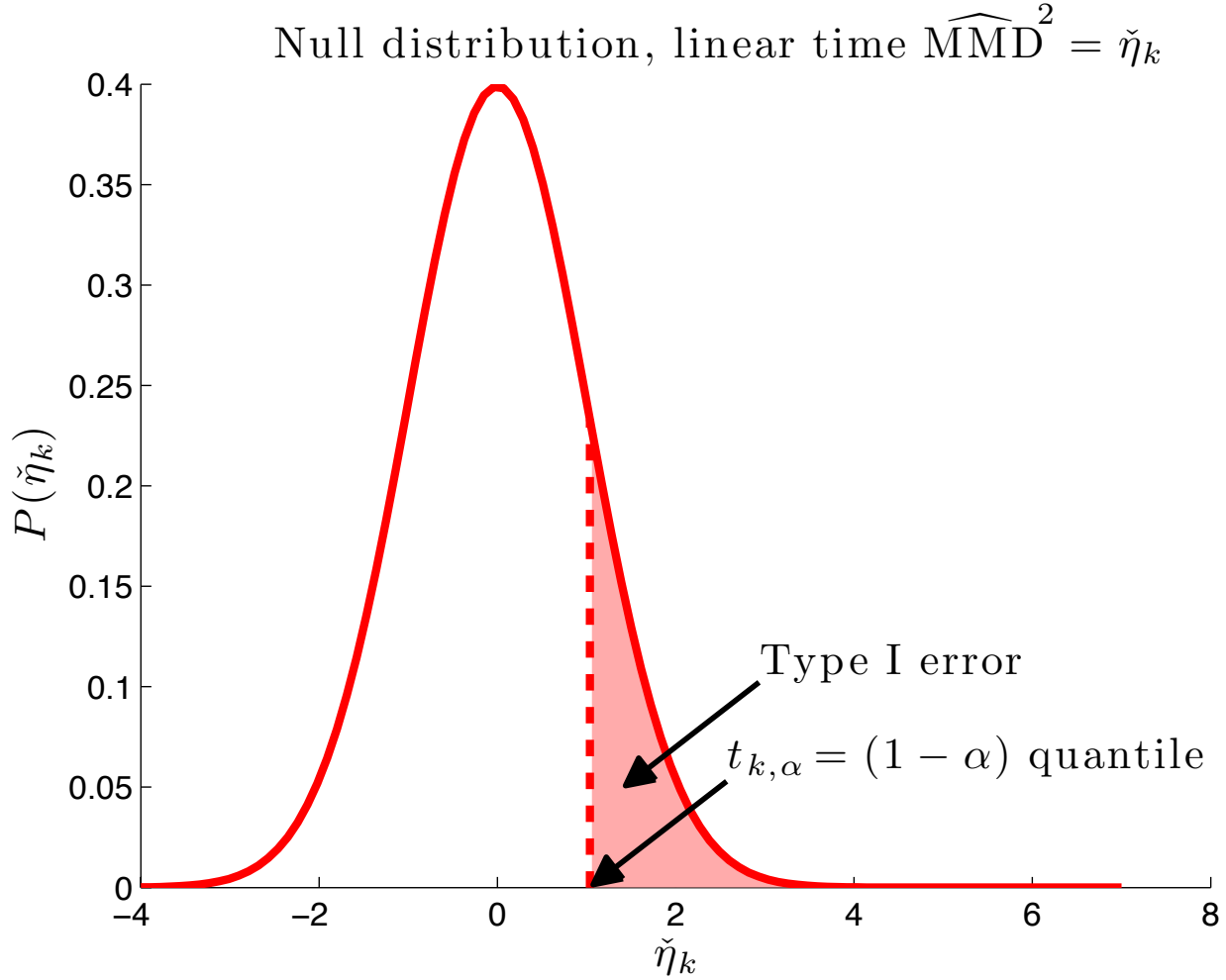
$$m^{1/2}\left(\check{\eta}_k - \eta_k(p,q)\right) \xrightarrow{D} \mathcal{N}(0, 2\sigma_k^2)$$

- assuming $0 < \mathbb{E}(h_k^2) < \infty$ (true for bounded k)
- $\sigma_k^2 = \mathbb{E}_v h_k^2(v) - \left[\mathbb{E}_v(h_k(v))\right]^2$.

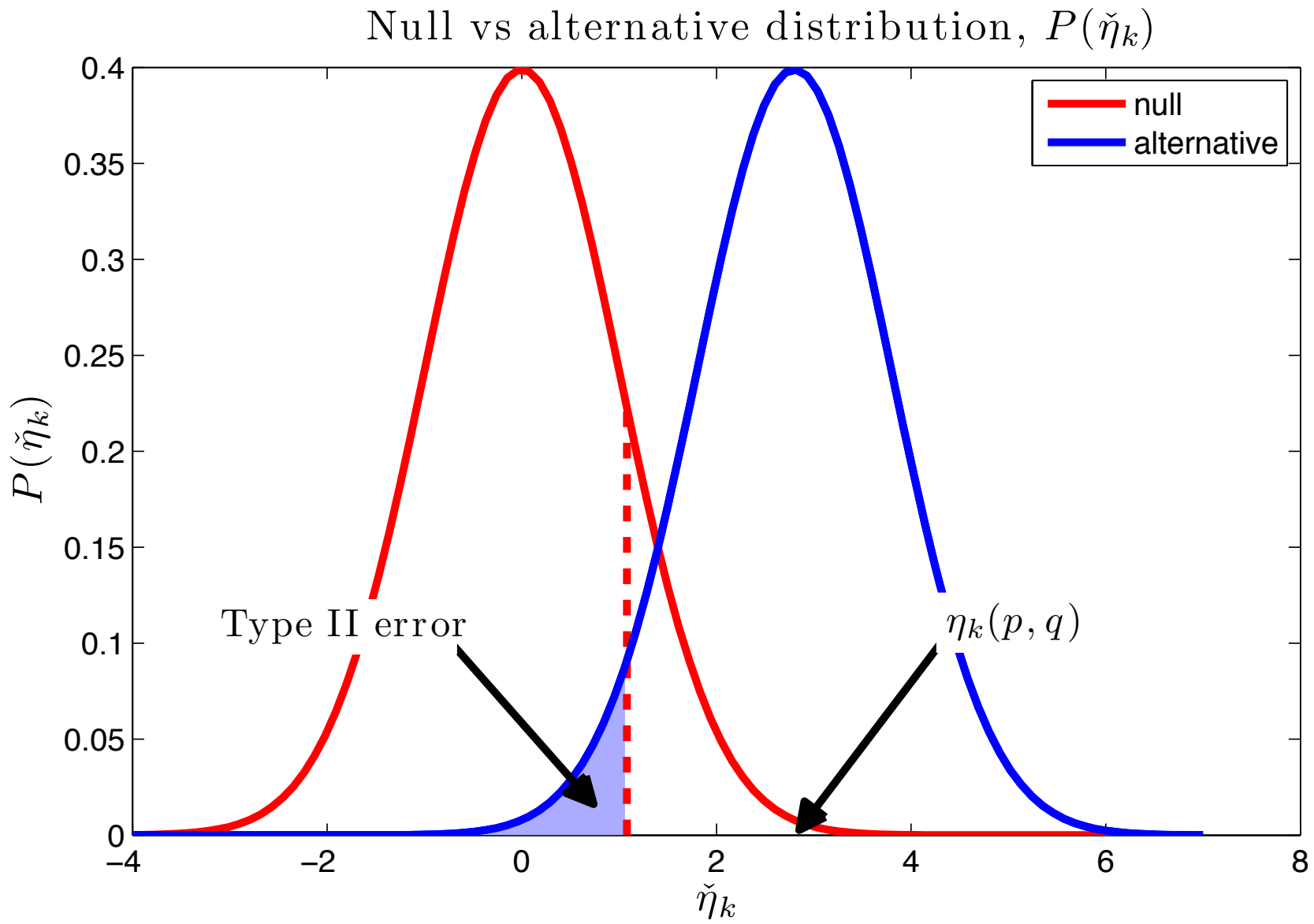
Hypothesis test

Hypothesis test of asymptotic level α:

$t_{k,\alpha} = m^{-1/2}\sigma_k\sqrt{2}\Phi^{-1}(1-\alpha)$ where Φ^{-1} is inverse CDF of $\mathcal{N}(0,1)$.



Type II error



The best kernel: minimizes Type II error

Type II error: $\check{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p,q) > 0$.

Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\eta_k(p,q)\sqrt{m}}{\sigma_k\sqrt{2}}\right)$$

where Φ is a Normal CDF.

The best kernel: minimizes Type II error

Type II error: $\check{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p,q) > 0$.

Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\eta_k(p,q)\sqrt{m}}{\sigma_k\sqrt{2}}\right)$$

where Φ is a Normal CDF.

Since Φ monotonic, best kernel choice to minimize Type II error prob. is:

$$k_* = \arg\max_{k\in\mathcal{K}} \eta_k(p,q)\sigma_k^{-1},$$

where $\mathcal{K}$ is the family of kernels under consideration.

Learning the best kernel in a family

Define the family of kernels as follows:

$$\mathcal{K} := \left\{ k \ : \ k = \sum_{u=1}^{d} \beta_u k_u, \ \|\beta\|_1 = D, \ \beta_u \geq 0, \ \forall u \in \{1, \ldots, d\} \right\}.$$

Properties: if at least one $\beta_u > 0$

- all $k \in \mathcal{K}$ are valid kernels,
- If all k_u charateristic then k characteristic

Test statistic

The squared MMD becomes

$$\eta_k(p,q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p,q),$$

where $\eta_u(p,q) := \mathbb{E}_v h_u(v)$.

Test statistic

The squared MMD becomes

$$\eta_k(p,q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p,q),$$

where $\eta_u(p,q) := \mathbb{E}_v h_u(v)$.

Denote:

- $\beta = (\beta_1, \beta_2, \ldots, \beta_d)^\top \in \mathbb{R}^d$,
- $h = (h_1, h_2, \ldots, h_d)^\top \in \mathbb{R}^d$,
  - $h_u(x, x', y, y') = k_u(x, x') + k_u(y, y') - k_u(x, y') - k_u(x', y)$
- $\eta = \mathbb{E}_v(h) = (\eta_1, \eta_2, \ldots, \eta_d)^\top \in \mathbb{R}^d$.

Quantities for test:

$$\eta_k(p,q) = \mathbb{E}(\beta^\top h) = \beta^\top \eta \qquad \sigma_k^2 := \beta^\top \mathrm{cov}(h)\beta.$$

Optimization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad\qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta},$$

$\hat{Q}$ is empirical estimate of $\mathrm{cov}(h)$.

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ on data to be tested (why?)

Optimization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad\qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta},$$

$\hat{Q}$ is empirical estimate of $\mathrm{cov}(h)$.

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ on data to be tested (why?)

Objective:

$$\begin{aligned}
\hat{\beta}^* &= \arg\max_{\beta \succeq 0} \ \hat{\eta}_k(p,q)\hat{\sigma}_{k,\lambda}^{-1} \\
&= \arg\max_{\beta \succeq 0} \ \left(\beta^\top \hat{\eta}\right) \left(\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta\right)^{-1/2} \\
&=: \alpha(\beta; \hat{\eta}, \hat{Q})
\end{aligned}$$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\beta \succeq 0$ s.t. $\alpha(\beta; \hat{\eta}, \hat{Q}) > 0$.

Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\beta \succeq 0$ s.t. $\alpha(\beta; \hat{\eta}, \hat{Q}) > 0$.

Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\hat{\beta}^* = \arg\max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q})$.

Quadratic program:

$$\min\{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta : \beta^\top \hat{\eta} = 1,\ \beta \succeq 0\}$$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\beta \succeq 0$ s.t. $\alpha(\beta; \hat{\eta}, \hat{Q}) > 0$.

Thus: $\alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\hat{\beta}^* = \arg\max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q})$.

Quadratic program:

$$\min\{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta : \beta^\top \hat{\eta} = 1,\ \beta \succeq 0\}$$

What if $\hat{\eta}$ has no positive entries?

Test procedure

1. Split the data into testing and training.
2. On the training data:
   (a) Compute $\hat{\eta}_u$ for all $k_u \in \mathcal{K}$
   (b) If at least one $\hat{\eta}_u > 0$, solve the QP to get β^*, else choose random kernel from $\mathcal{K}$
3. On the test data:
   (a) Compute $\check{\eta}_{k^*}$ using $k^* = \sum_{u=1}^{d} \beta^* k_u$
   (b) Compute test threshold $\check{t}_{\alpha,k^*}$ using $\check{\sigma}_{k^*}$
4. Reject null if $\check{\eta}_{k^*} > \check{t}_{\alpha,k^*}$

Convergence bounds

Assume bounded kernel, σ_k, bounded away from 0.

If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P\left(m^{-1/3}\right).$$

Convergence bounds

Assume bounded kernel, σ_k, bounded away from 0.

If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P\left(m^{-1/3}\right).$$

Idea:

$$\begin{aligned}
&\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| \\
&\leq \sup_{k \in \mathcal{K}} \left| \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \eta_k \sigma_{k,\lambda}^{-1} \right| + \sup_{k \in \mathcal{K}} \left| \eta_k \sigma_{k,\lambda}^{-1} - \eta_k \sigma_k^{-1} \right| \\
&\leq \frac{\sqrt{d}}{D\sqrt{\lambda_m}} \left(C_1 \sup_{k \in \mathcal{K}} |\hat{\eta}_k - \eta_k| + C_2 \sup_{k \in \mathcal{K}} |\hat{\sigma}_{k,\lambda} - \sigma_{k,\lambda}| \right) + C_3 D^2 \lambda_m,
\end{aligned}$$

Experiments

Competing approaches

- Median heuristic
- Max. MMD: choose $k_u \in \mathcal{K}$ with the largest $\hat{\eta}_u$
  - same as maximizing $\beta^\top \hat{\eta}$ subject to $\|\beta\|_1 \leq 1$
- ℓ_2 statistic: maximize $\beta^\top \hat{\eta}$ subject to $\|\beta\|_2 \leq 1$
- Cross validation on training set

Also compare with:

- Single kernel that maximizes ratio $\eta_k(p, q)\sigma_k^{-1}$

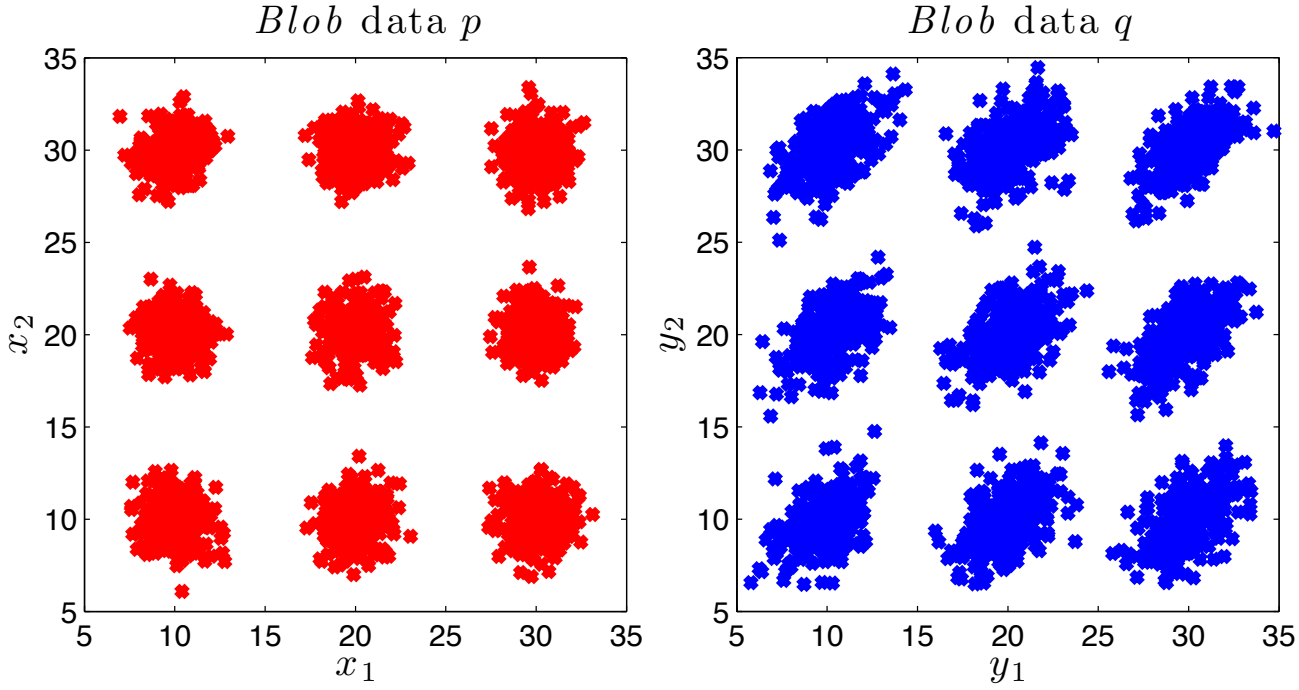
Blobs: data

Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

Blobs: data

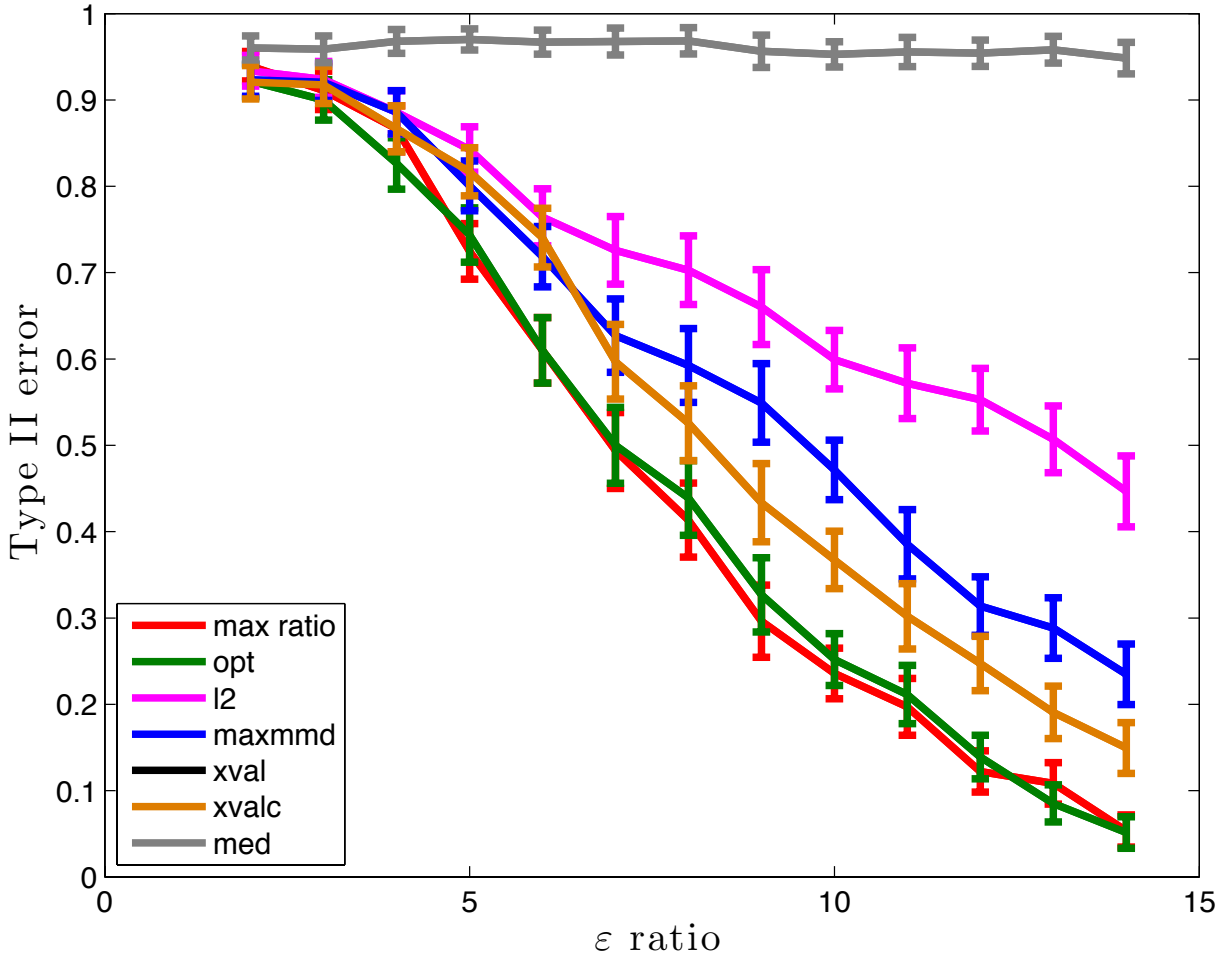
Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

We distinguish a field of Gaussian blobs with different covariances.



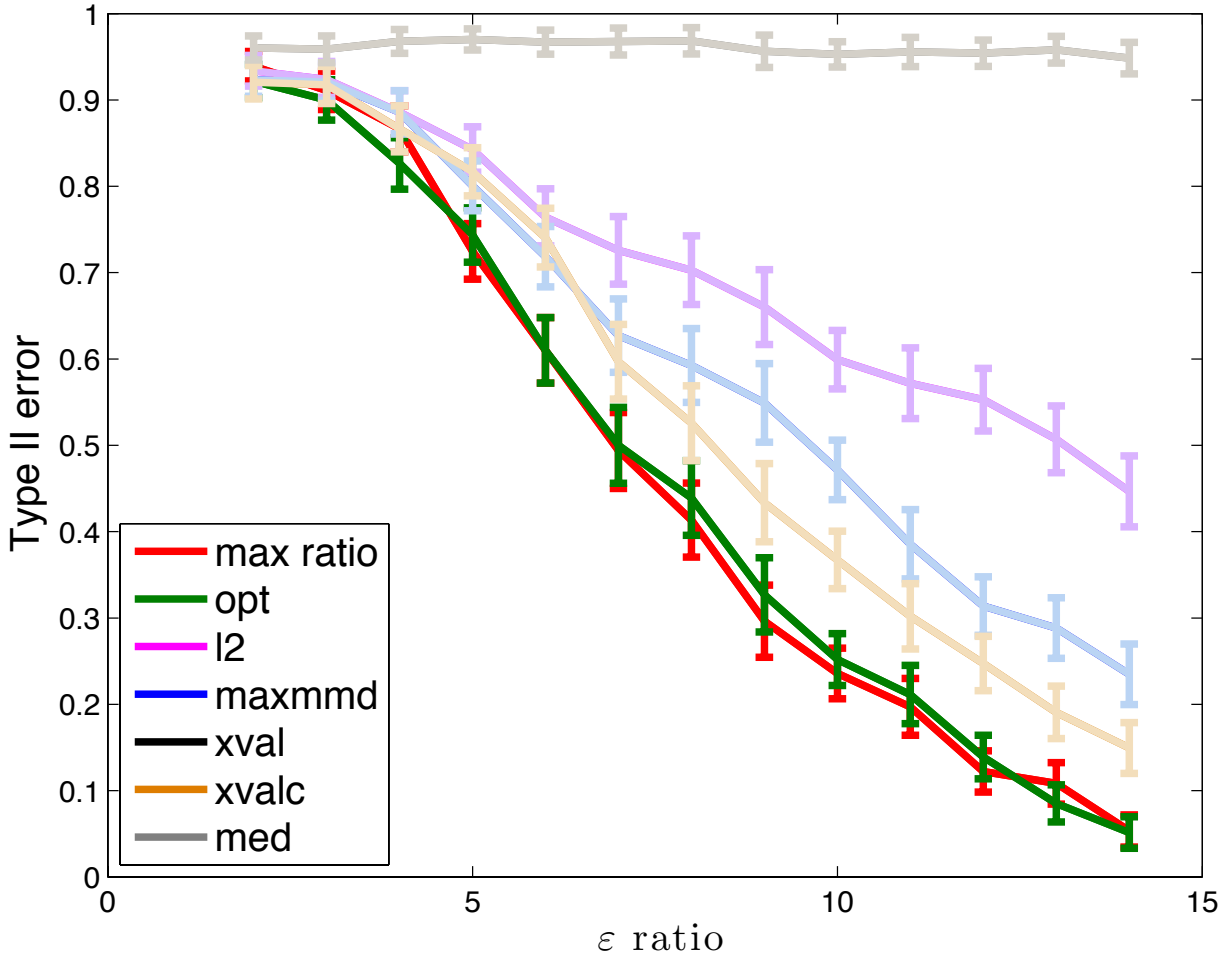
Ratio $\varepsilon = 3.2$ of largest to smallest eigenvalues of blobs in q.

Blobs: results



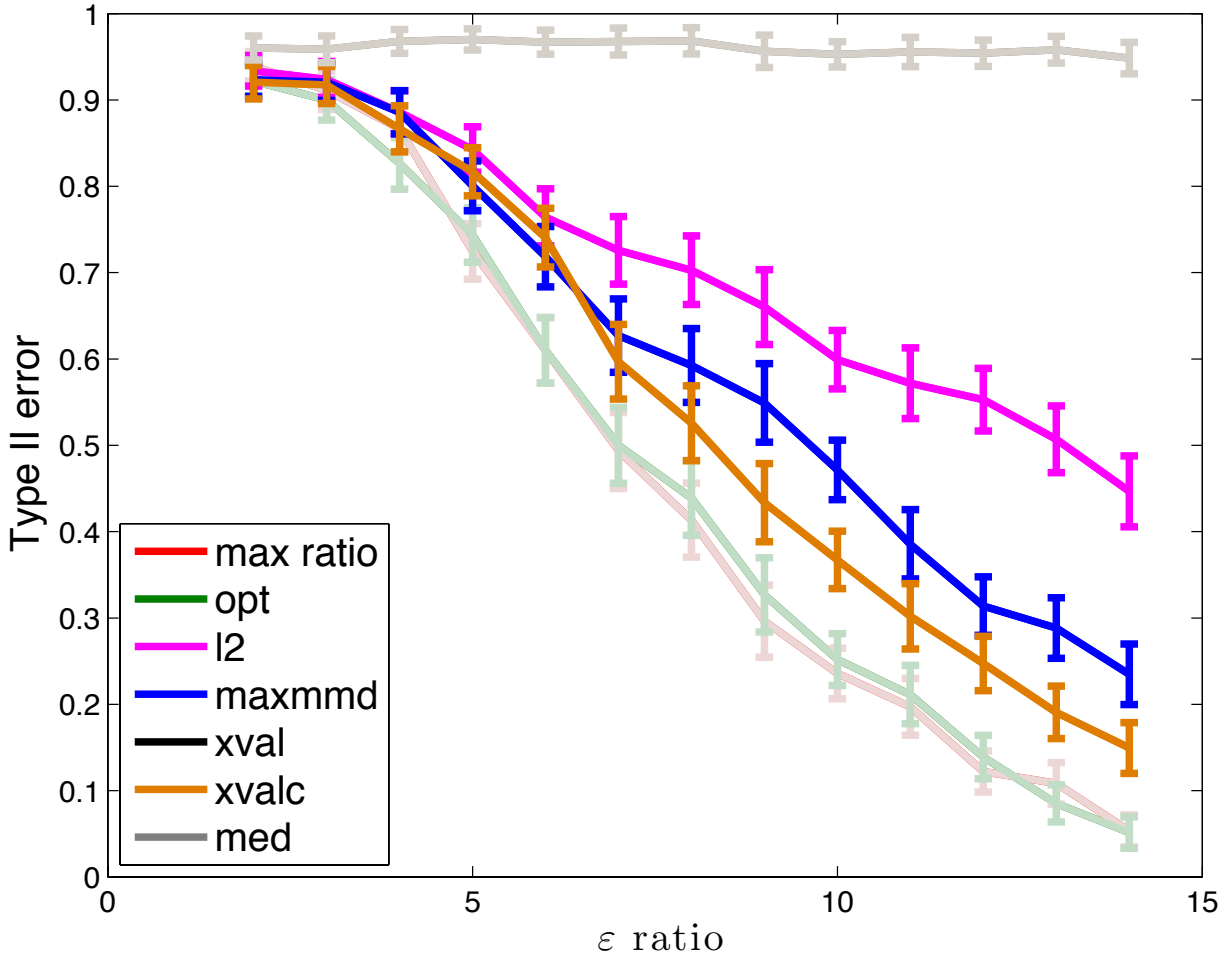
Parameters: $m = 10,000$ (for training and test). Ratio ε of largest to smallest eigenvalues of blobs in q. Results are average over 617 trials.

Blobs: results



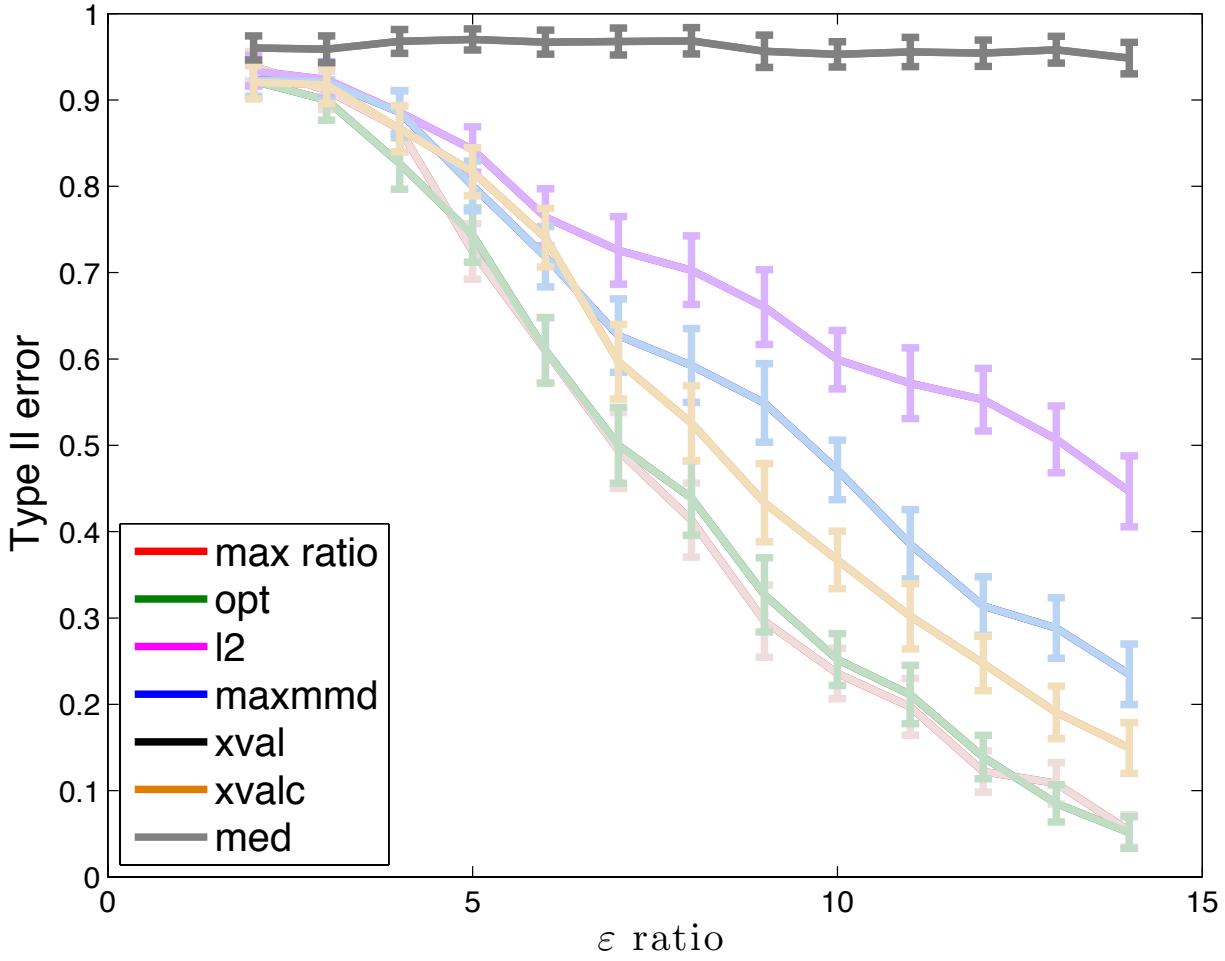
Optimize ratio $\eta_k(p,q)\sigma_k^{-1}$

Blobs: results



Maximize $\eta_k(p, q)$ with β constraint

Blobs: results



Median heuristic

Feature selection: data

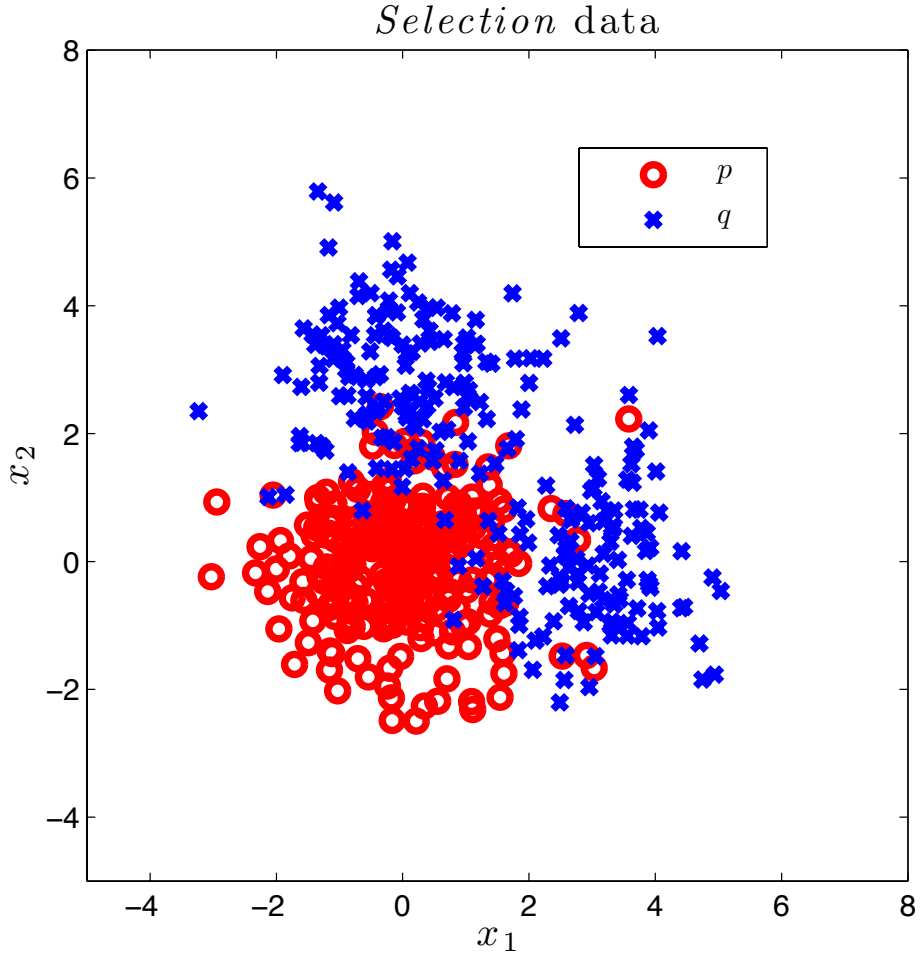
Idea: no single best kernel.

Each of the k_u are univariate (along a single coordinate)

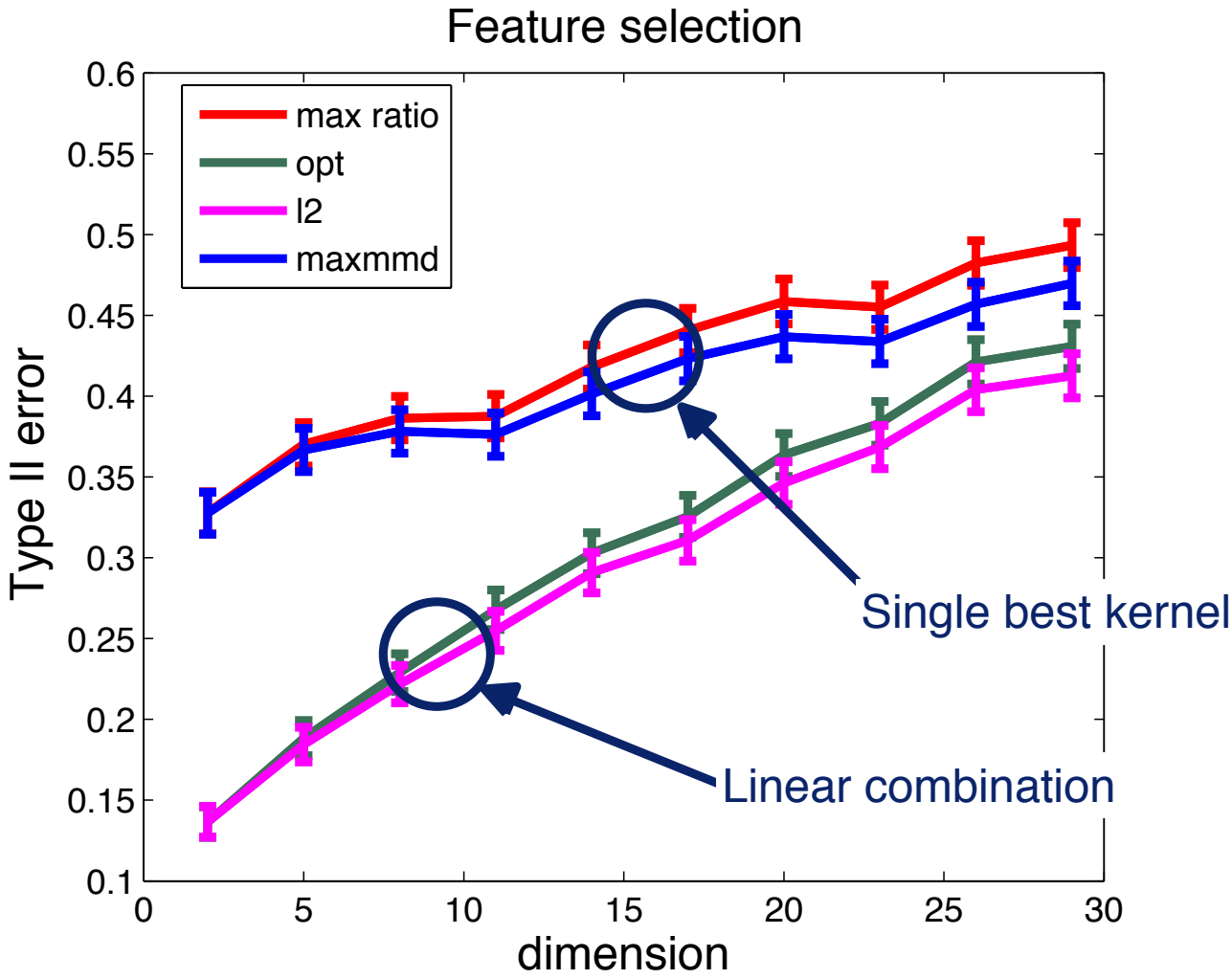
Feature selection: data

Idea: no single best kernel.

Each of the k_u are univariate (along a single coordinate)



Feature selection: results



$m = 10,000$, average over 5000 trials

Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

$$u(t) = \sin(\omega_c t)\,[a\,s(t) + l]$$

- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

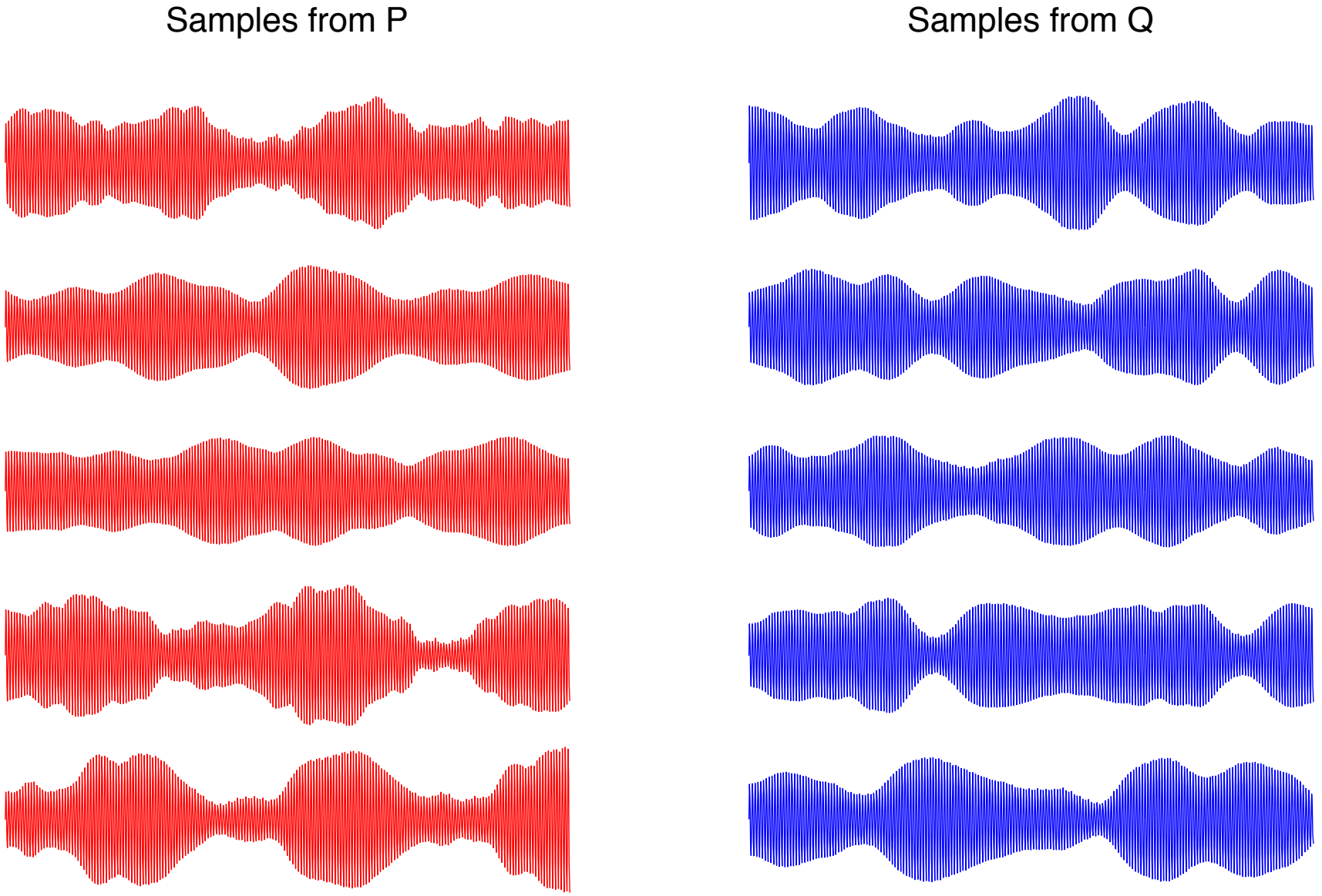
$$u(t) = \sin(\omega_c t)\,[a\,s(t) + l]$$

- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

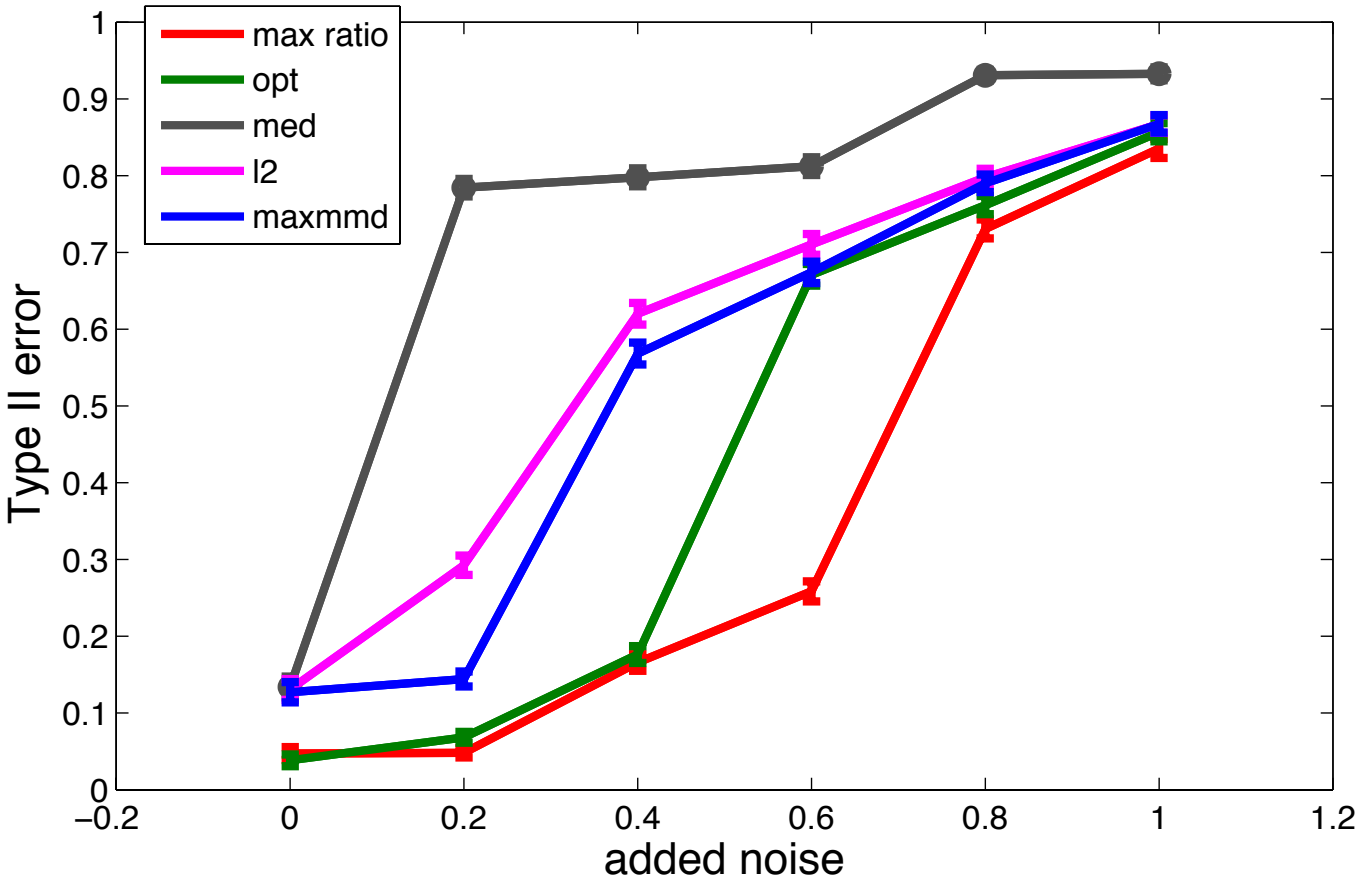
Two amplitude modulated signals from same artist (in this case, Magnetic Fields).

- Music sampled at 8KHz (very low)
- Carrier frequency is 24kHz
- AM signal observed at 120kHz
- Samples are extracts of length $N = 1000$, approx. 0.01 sec (very short).
- Total dataset size is 30,000 samples from each of p, q.

Amplitude modulated signals



Results: AM signals



$m = 10,000$ (for training and test) and scaling $a = 0.5$. Average over 4124 trials. Gaussian noise added.

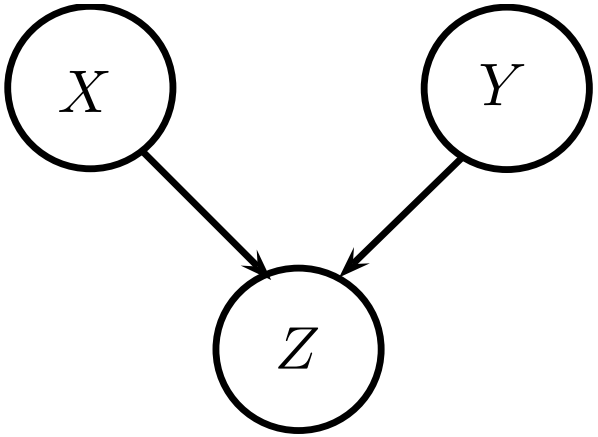
Observations on kernel choice

- It is possible to choose the best kernel for a kernel two-sample test
- Kernel choice matters for "difficult" problems, where the distributions differ on a lengthscale different to that of the data.
- Ongoing work:
  - quadratic time statistic
  - avoid training/test split

Lancaster (3-way) Interactions

Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?

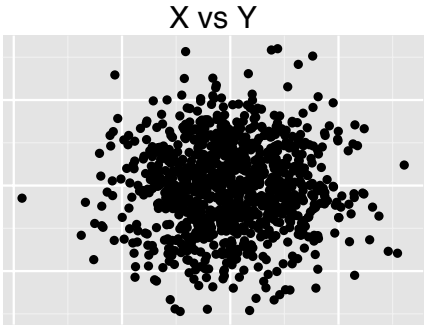


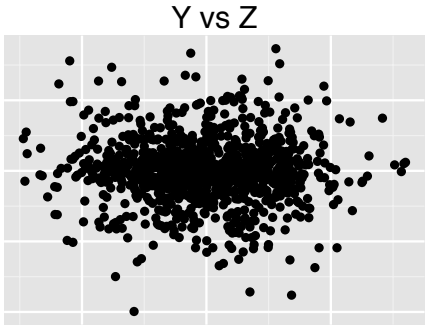
Detecting a higher order interaction

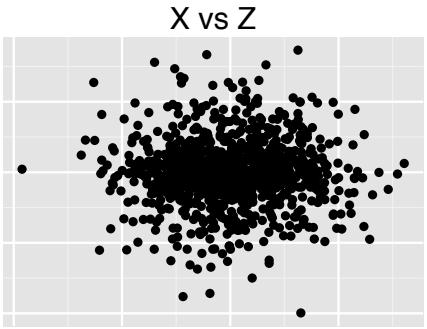
- How to detect V-structures with pairwise weak (or nonexistent) dependence?

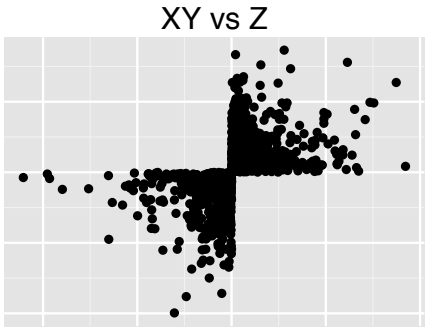
Detecting a higher order interaction

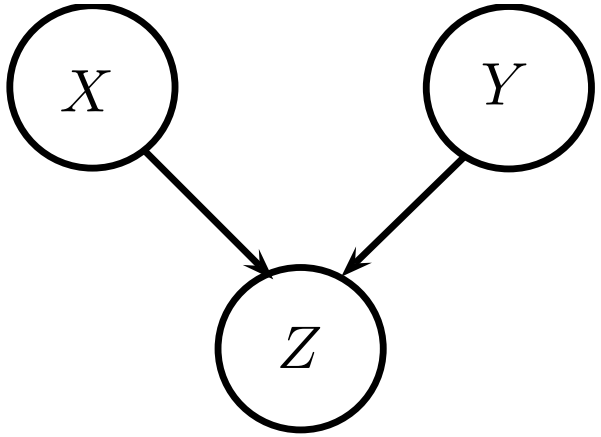
- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$







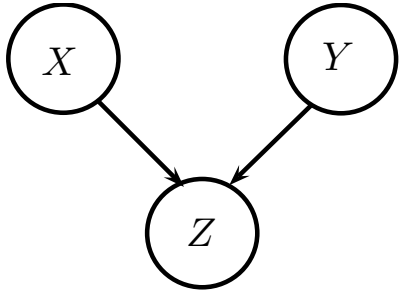




- $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z|\ X, Y \sim \text{sign}(XY)Exp(\frac{1}{\sqrt{2}})$

Faithfulness violated here

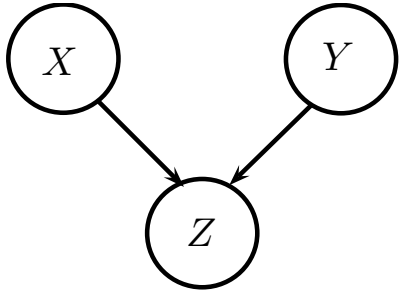
V-structure Discovery



Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ (Zhang et al 2011) or

V-structure Discovery

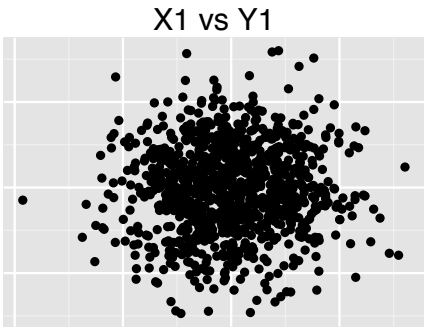


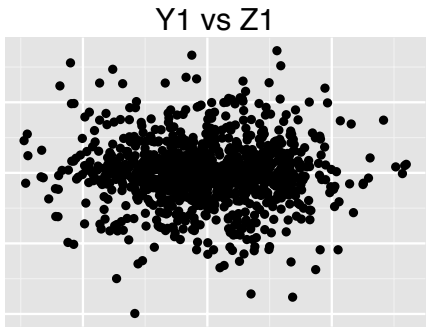
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

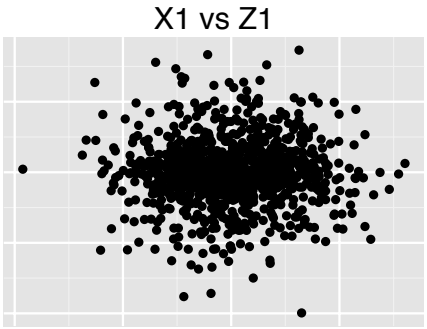
- CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ (Zhang et al 2011) or
- Factorisation test: $\mathbf{H_0} : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)
  - compute p-values for each of the marginal tests for $(Y, Z) \perp\!\!\!\perp X$, $(X, Z) \perp\!\!\!\perp Y$, or $(X, Y) \perp\!\!\!\perp Z$
  - apply Holm-Bonferroni (**HB**) sequentially rejective correction (Holm 1979)

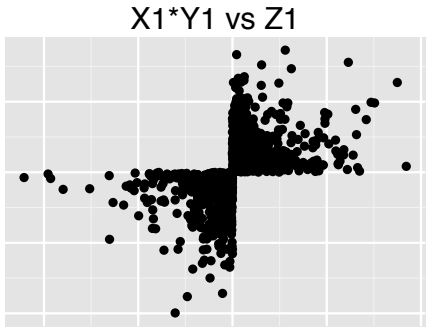
V-structure Discovery (2)

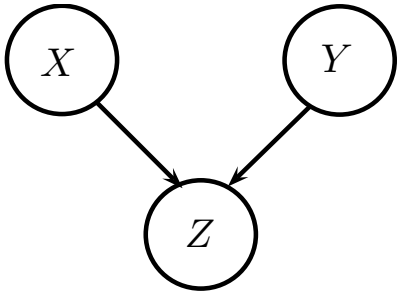
- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$











- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z_1 |\, X_1, Y_1 \sim \text{sign}(X_1 Y_1) Exp(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$ Faithfulness violated here

V-structure Discovery (3)

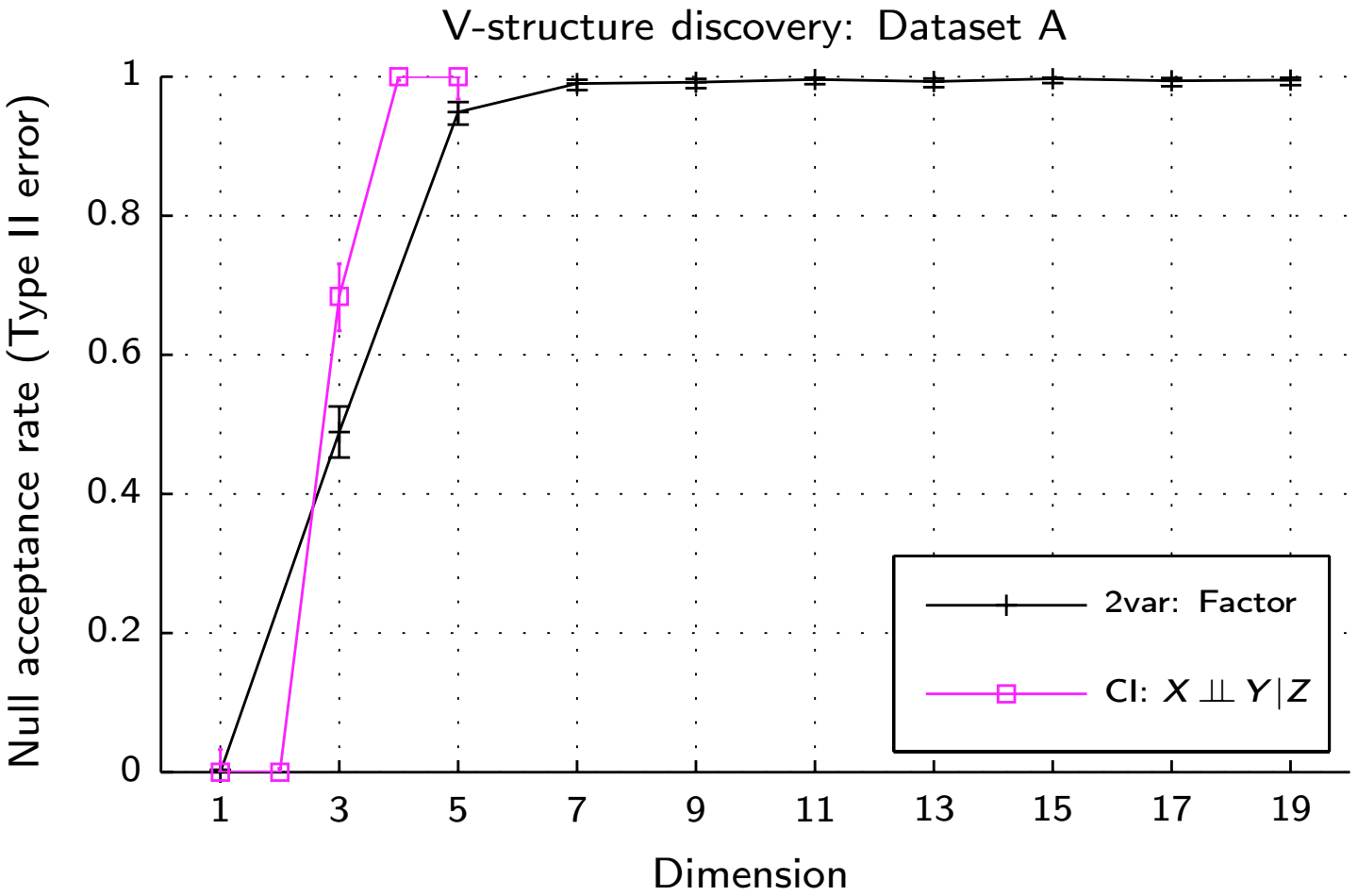


Figure 1: CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), and a factorisation test with a **HB** correction, $n = 500$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\qquad \Delta_L P = P_{XY} - P_X P_Y$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

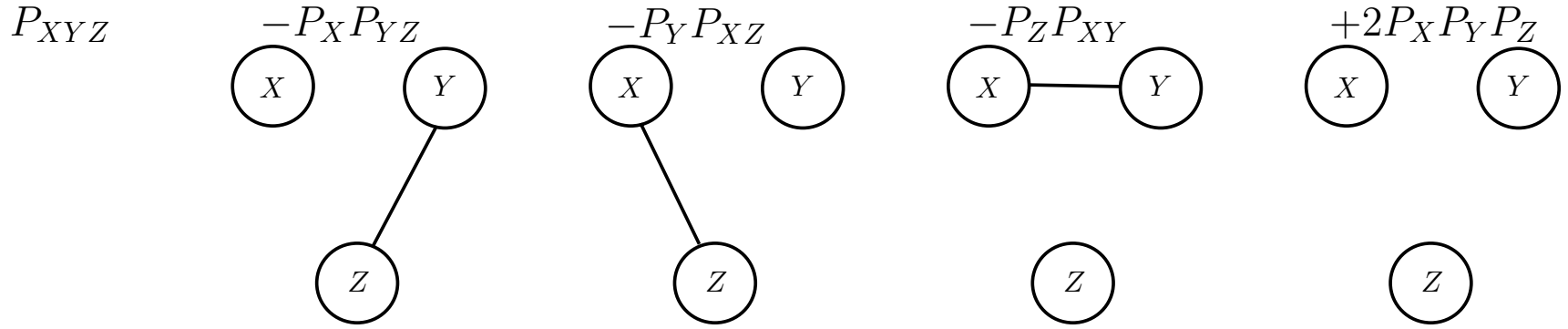
- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

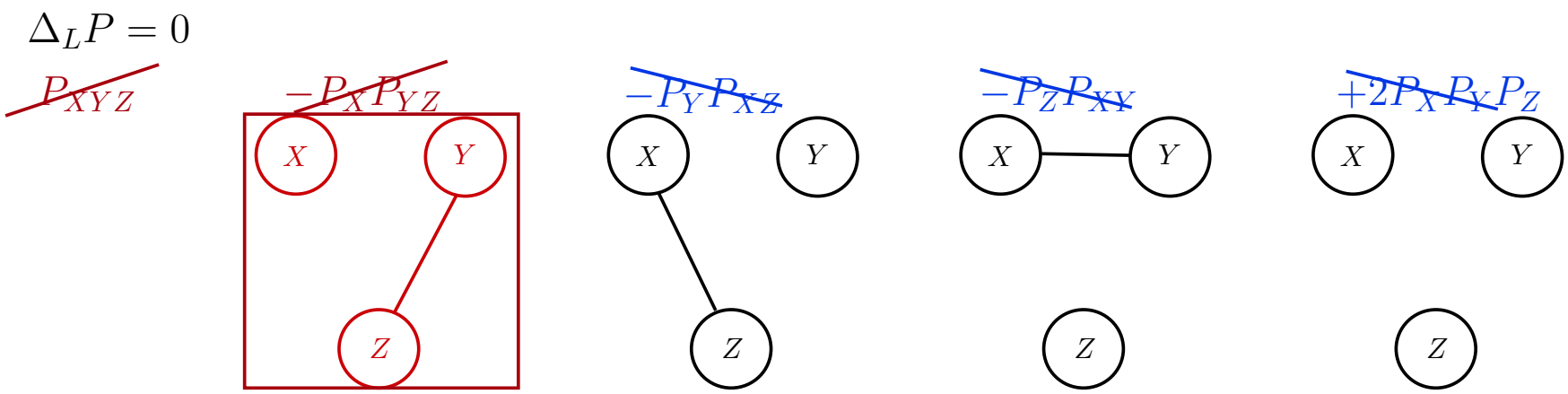
$\Delta_L P =$



Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$



Case of $P_X \perp\!\!\!\perp P_{YZ}$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

$$(X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X \Rightarrow \Delta_L P = 0.$$

...so what might be missed?

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

$$\Delta_L P = 0 \not\Rightarrow (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$$

Example:

$P(0,0,0) = 0.2$	$P(0,0,1) = 0.1$	$P(1,0,0) = 0.1$	$P(1,0,1) = 0.1$
$P(0,1,0) = 0.1$	$P(0,1,1) = 0.1$	$P(1,1,0) = 0.1$	$P(1,1,1) = 0.2$

A Test using Lancaster Measure

- Test statistic is empirical estimate of $\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa}$, where $\kappa = k \otimes l \otimes m$:

$$\|\mu_\kappa(P_{XYZ} - P_{XY}P_Z - \cdots)\|^2_{\mathcal{H}_\kappa} =$$
$$\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{\mathcal{H}_\kappa} - 2\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z \rangle_{\mathcal{H}_\kappa} \cdots$$

Inner Product Estimators

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(\mathbf{K} \circ \mathbf{L} \circ \mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M} \circ \mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+ \circ \mathbf{L}_+ \circ \mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K} \circ \mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K} \circ \mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L} \circ \mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_XP_YP_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

Table 1: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$

Inner Product Estimators

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(\mathbf{K} \circ \mathbf{L} \circ \mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M} \circ \mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+ \circ \mathbf{L}_+ \circ \mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K} \circ \mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K} \circ \mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L} \circ \mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_XP_YP_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

Table 2: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$

$$\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa} = \frac{1}{n^2}\left(H\mathbf{K}H \circ H\mathbf{L}H \circ H\mathbf{M}H\right)_{++}.$$

Empirical joint central moment in the feature space

Example A: factorisation tests

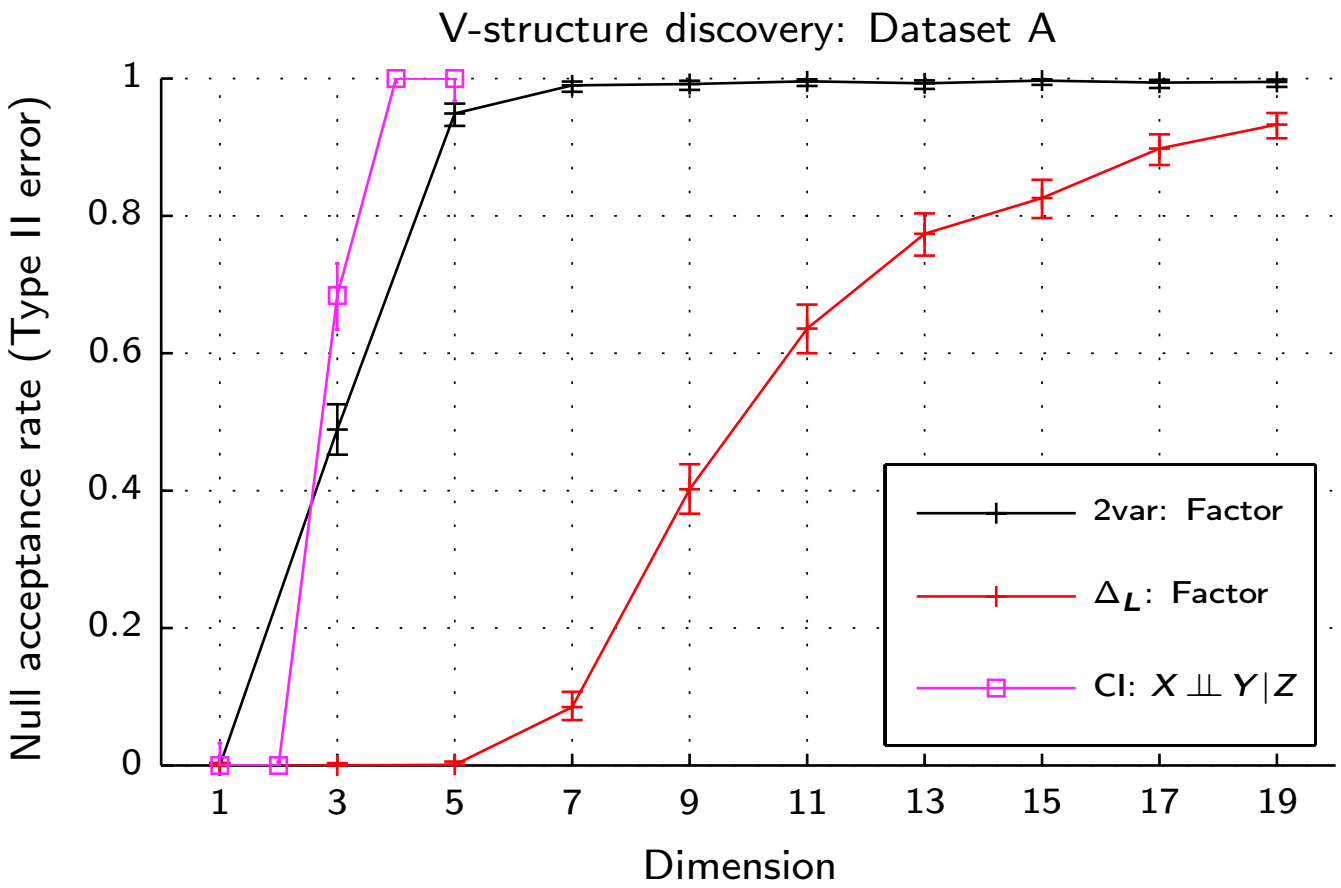


Figure 2: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Example B: Joint dependence can be easier to detect

- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $Z_1 = \begin{cases} X_1^2 + \epsilon, & w.p.\ 1/3, \\ Y_1^2 + \epsilon, & w.p.\ 1/3, \\ X_1 Y_1 + \epsilon, & w.p.\ 1/3, \end{cases}$ where $\epsilon \sim \mathcal{N}(0, 0.1^2)$.
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$
- dependence of Z on pair (X, Y) is stronger than on X and Y individually
- Satisfies faithfulness

Example B: factorisation tests

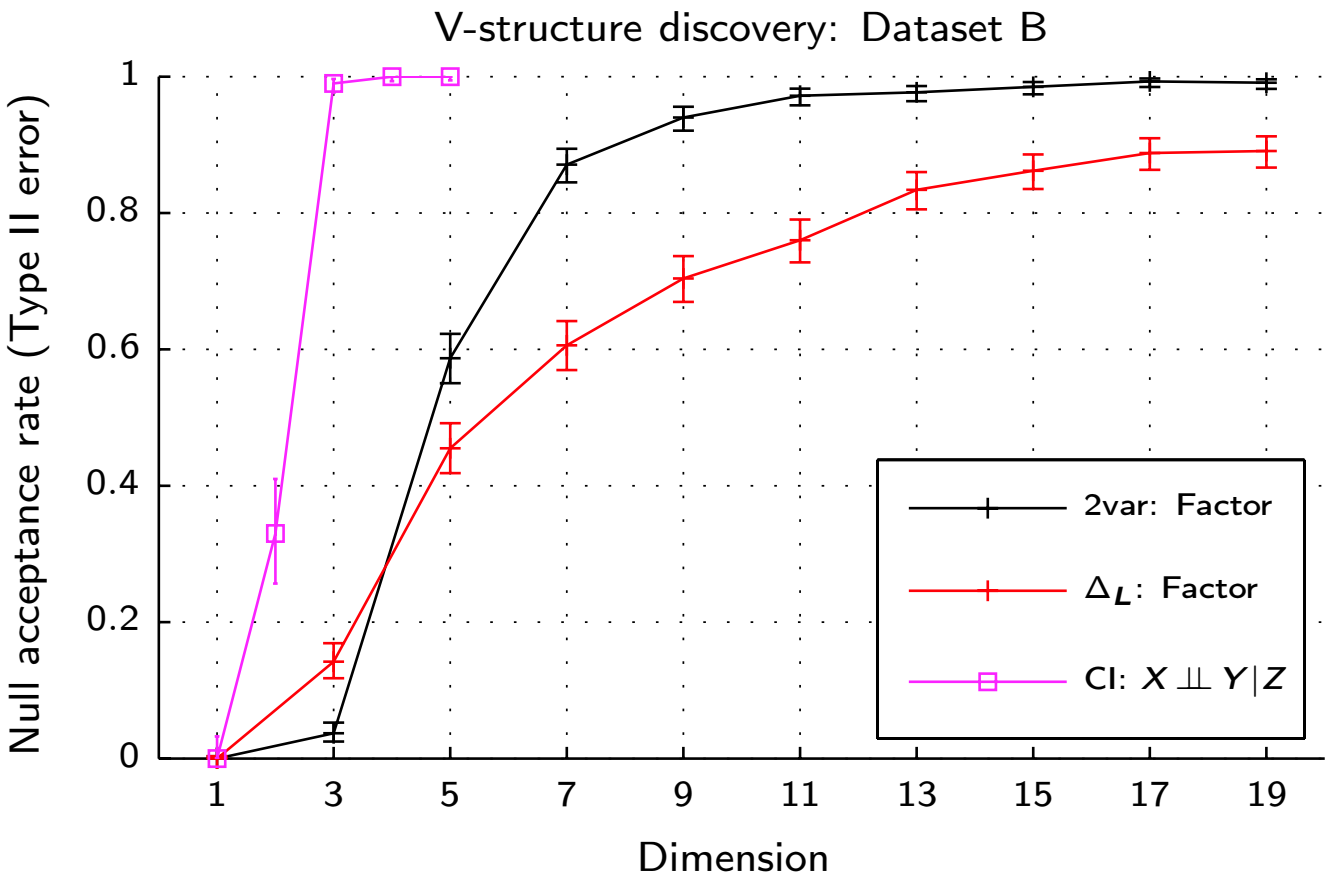


Figure 3: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} \, (|\pi| - 1)! J_\pi P$$

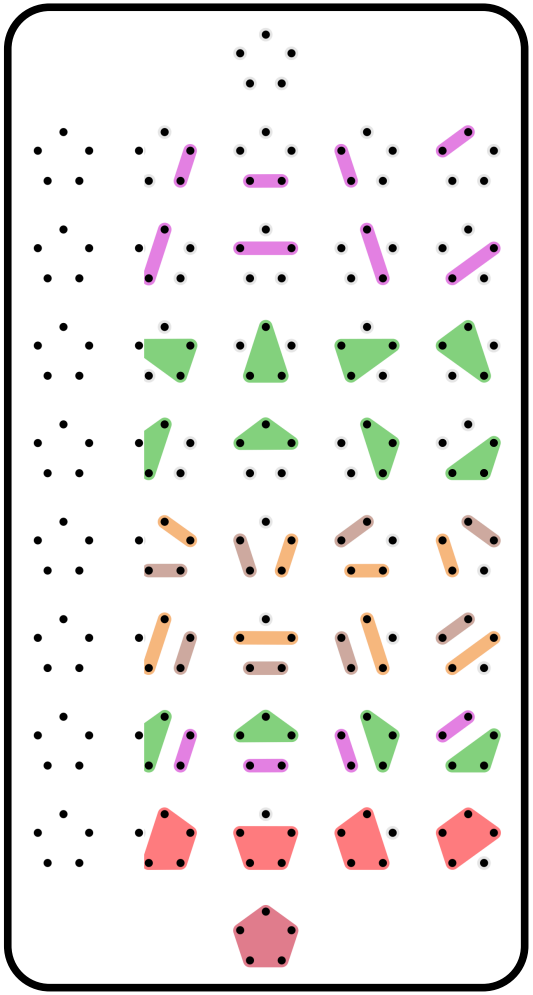
  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_\pi P$$

  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

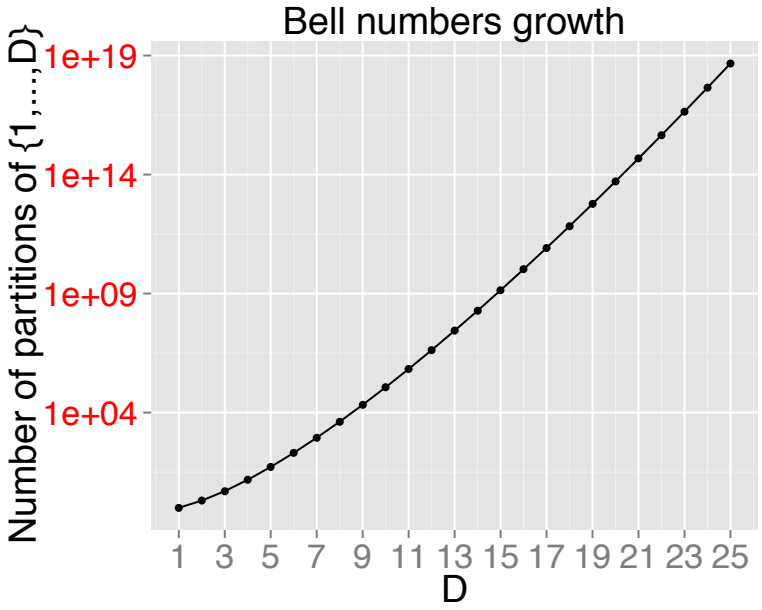


Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} \, (|\pi| - 1)! J_\pi P$$

  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.



joint central moments (Lancaster interaction)

vs.

joint cumulants (Streitberg interaction)

Total independence test

- Total independence test:

  $\mathbf{H_0} : P_{XYZ} = P_X P_Y P_Z$ vs. $\mathbf{H}_1 : P_{XYZ} \neq P_X P_Y P_Z$

Total independence test

- Total independence test:
  $\mathbf{H_0} : P_{XYZ} = P_X P_Y P_Z$ vs. $\mathbf{H}_1 : P_{XYZ} \neq P_X P_Y P_Z$
- For $(X_1, \ldots, X_D) \sim P_{\mathbf{X}}$, and $\kappa = \bigotimes_{i=1}^{D} k^{(i)}$:

$$\left\| \mu_\kappa \underbrace{\left(\hat{P}_{\mathbf{X}} - \prod_{i=1}^{D} \hat{P}_{X_i} \right)}_{\Delta_{tot}\hat{P}} \right\|^2_{\mathcal{H}_\kappa} = \frac{1}{n^2} \sum_{a=1}^{n} \sum_{b=1}^{n} \prod_{i=1}^{D} K_{ab}^{(i)} - \frac{2}{n^{D+1}} \sum_{a=1}^{n} \prod_{i=1}^{D} \sum_{b=1}^{n} K_{ab}^{(i)} + \frac{1}{n^{2D}} \prod_{i=1}^{D} \sum_{a=1}^{n} \sum_{b=1}^{n} K_{ab}^{(i)}.$$

- Coincides with the test proposed by Kankainen (1995) using empirical characteristic functions: similar relationship to that between dCov and HSIC (DS et al, 2013)

Example B: total independence tests

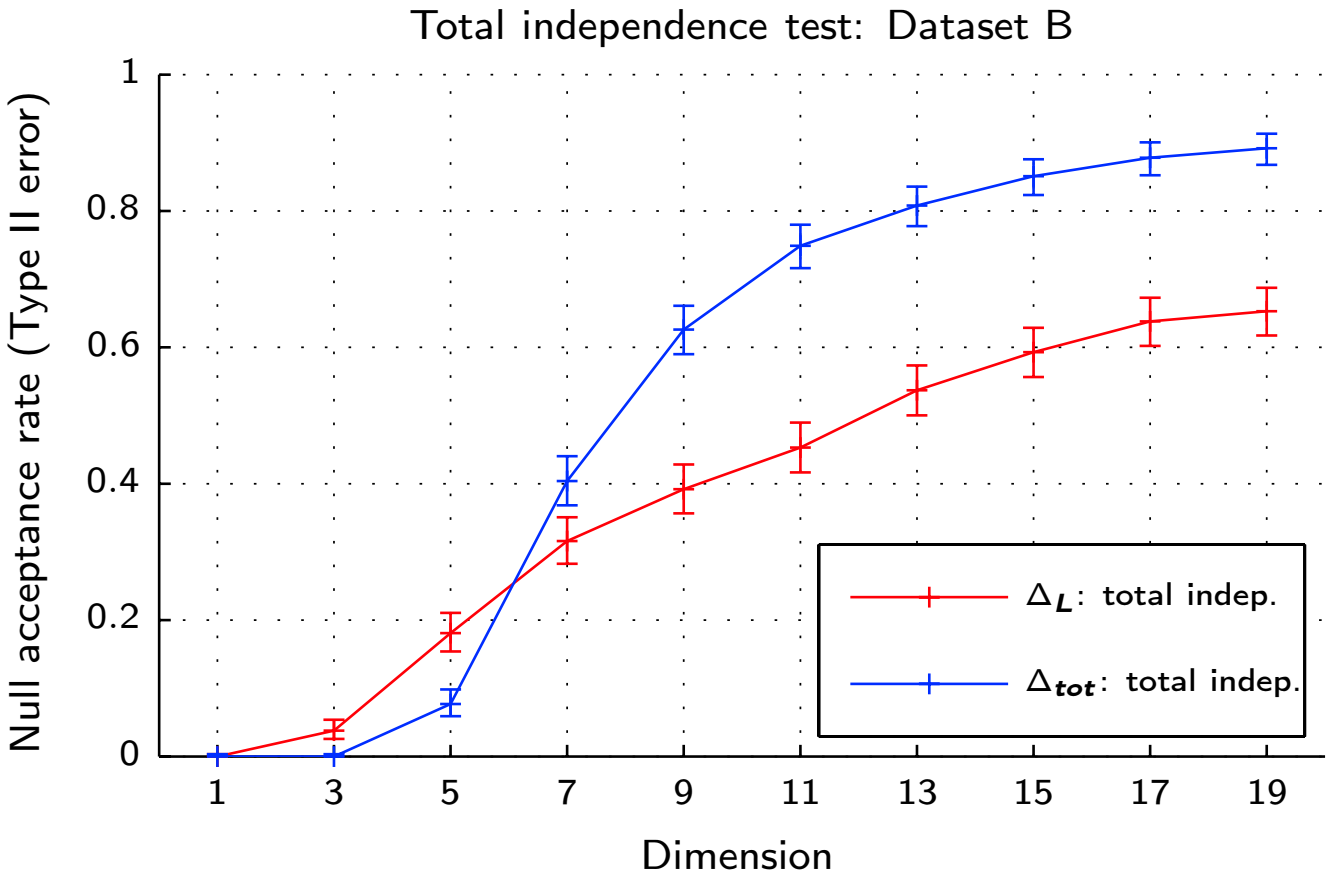
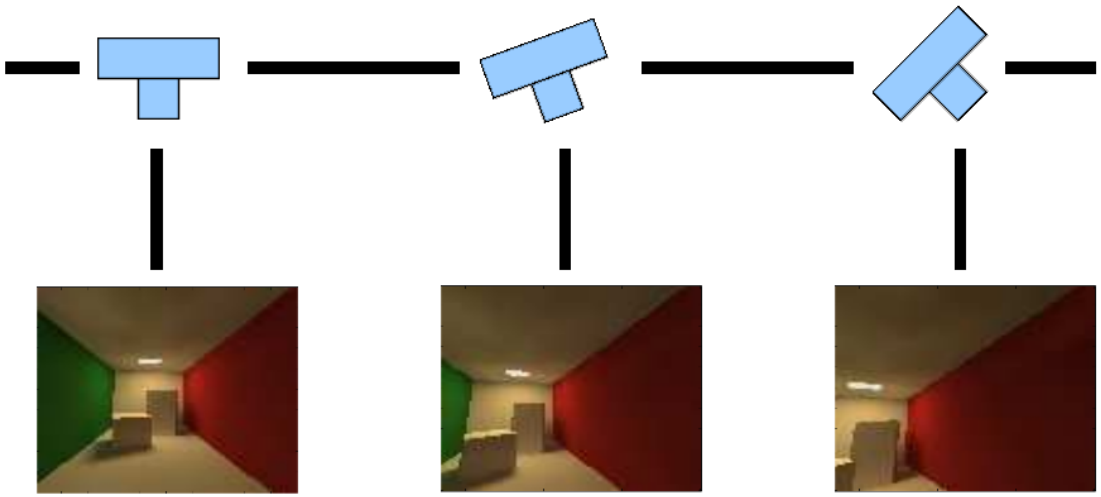


Figure 4: Total independence: $\Delta_{tot}\hat{P}$ vs. $\Delta_L\hat{P}$, $n = 500$

Nonparametric Bayesian inference using distribution embeddings

Motivating Example: Bayesian inference without a model



- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.

Challenges:

- No parametric model of camera dynamics (only samples)
- No parametric model of map from camera angle to image (only samples)
- Want to do filtering: Bayesian inference

ABC: an approach to Bayesian inference without a model

Bayes rule:

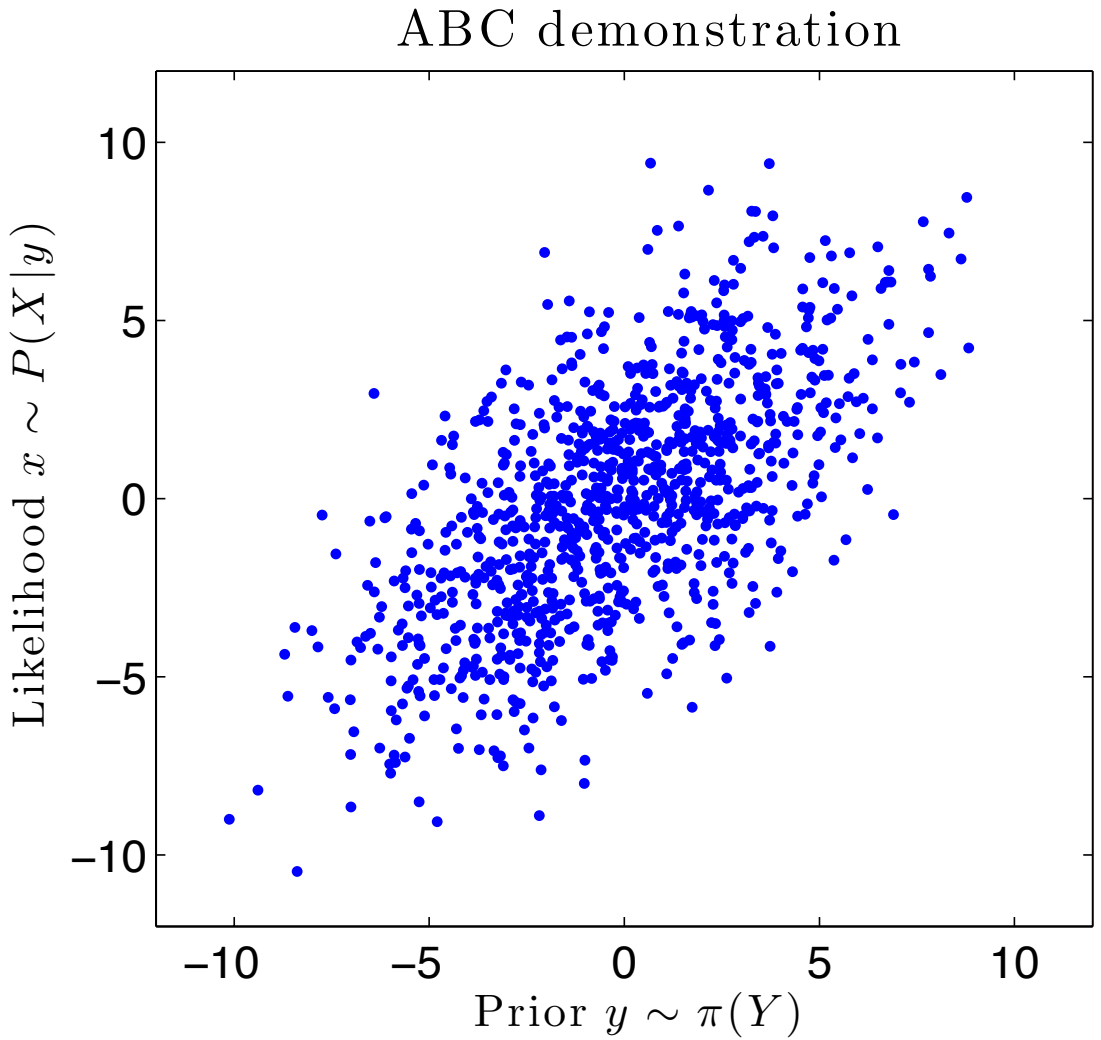
$$\mathbf{P}(y|x) = \frac{\mathbf{Q}(x|y)\pi(y)}{\int \mathbf{Q}(x|y)\pi(y)dy}$$

- $\mathbf{Q}(x|y)$ is likelihood
- $\pi(y)$ is prior

One approach: Approximate Bayesian Computation (ABC)

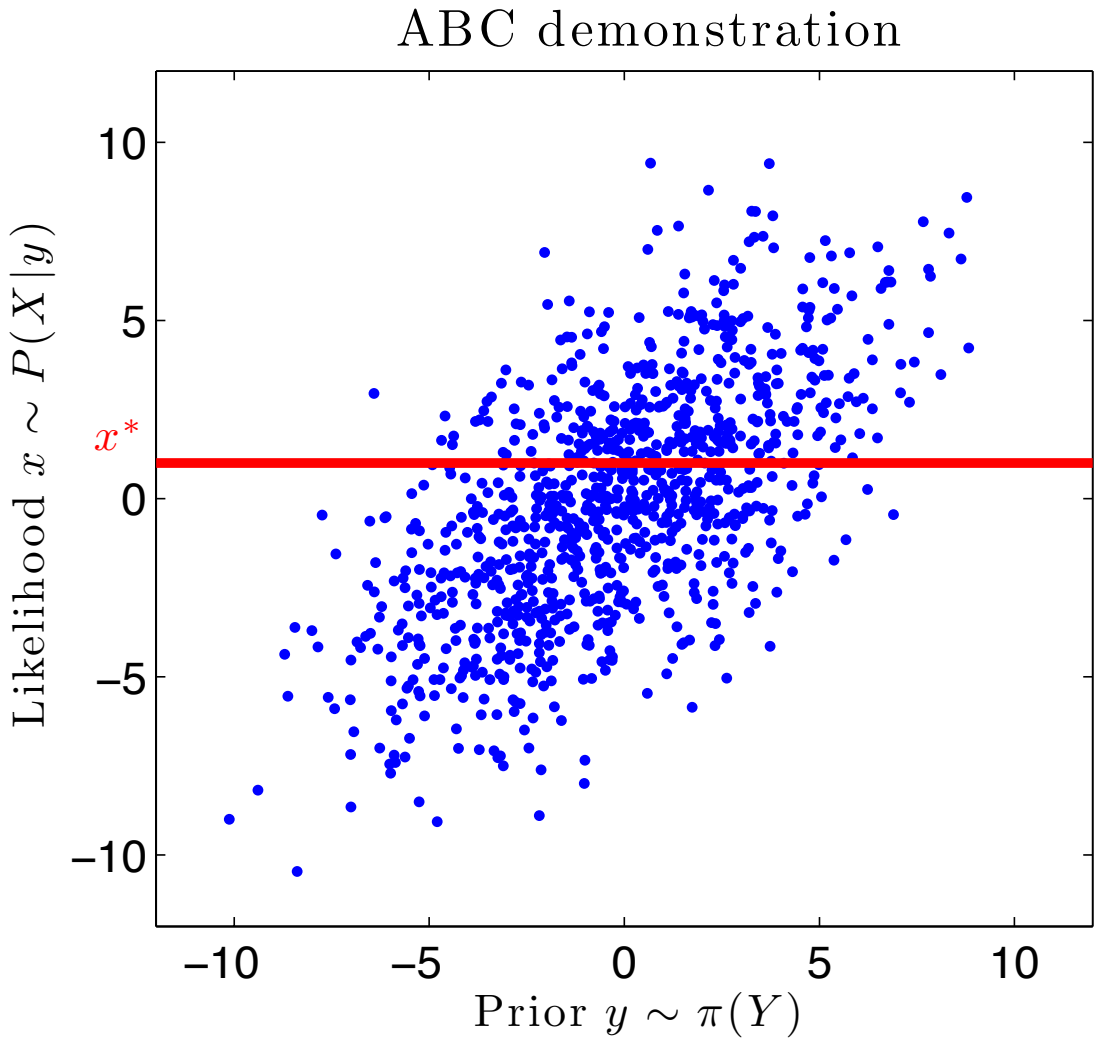
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



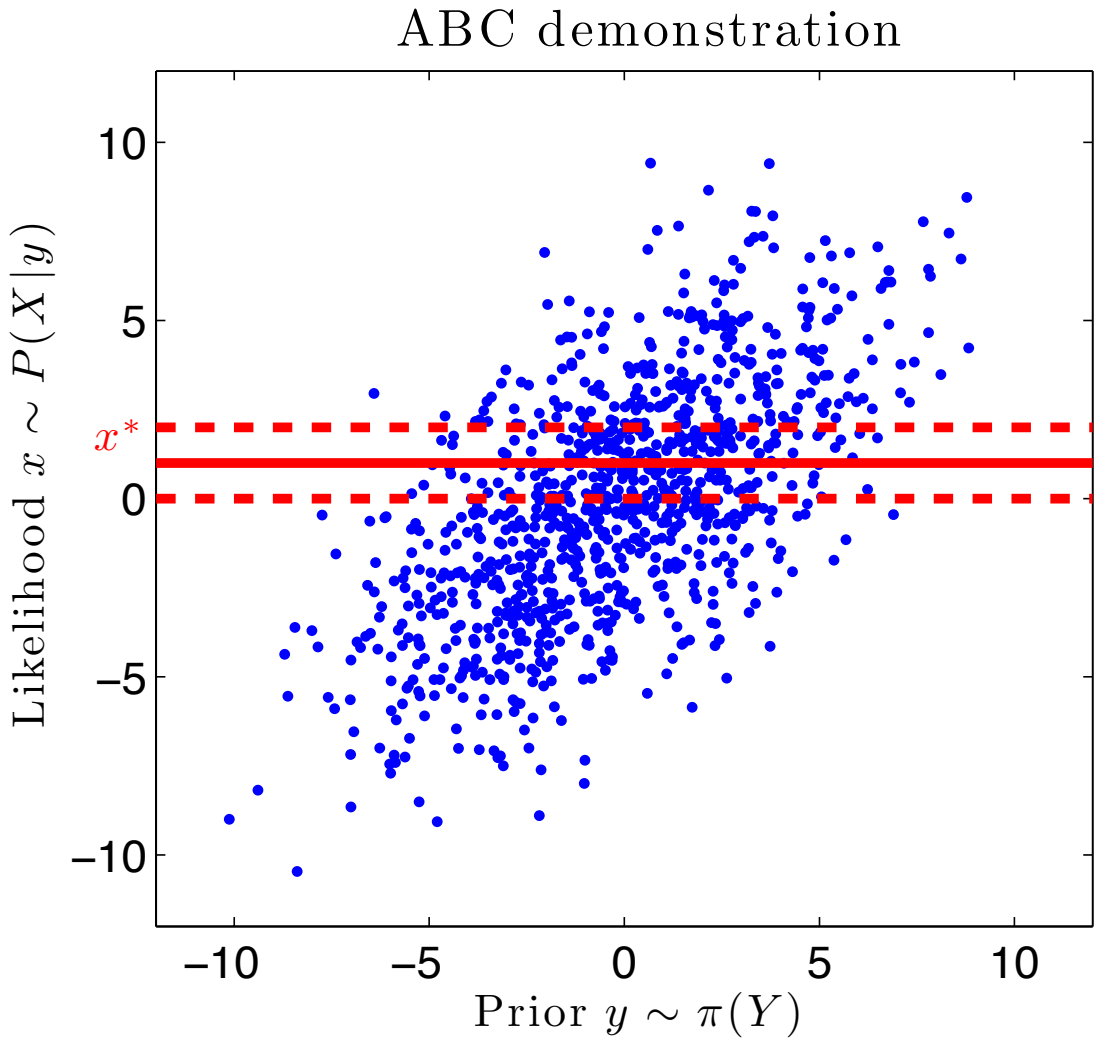
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



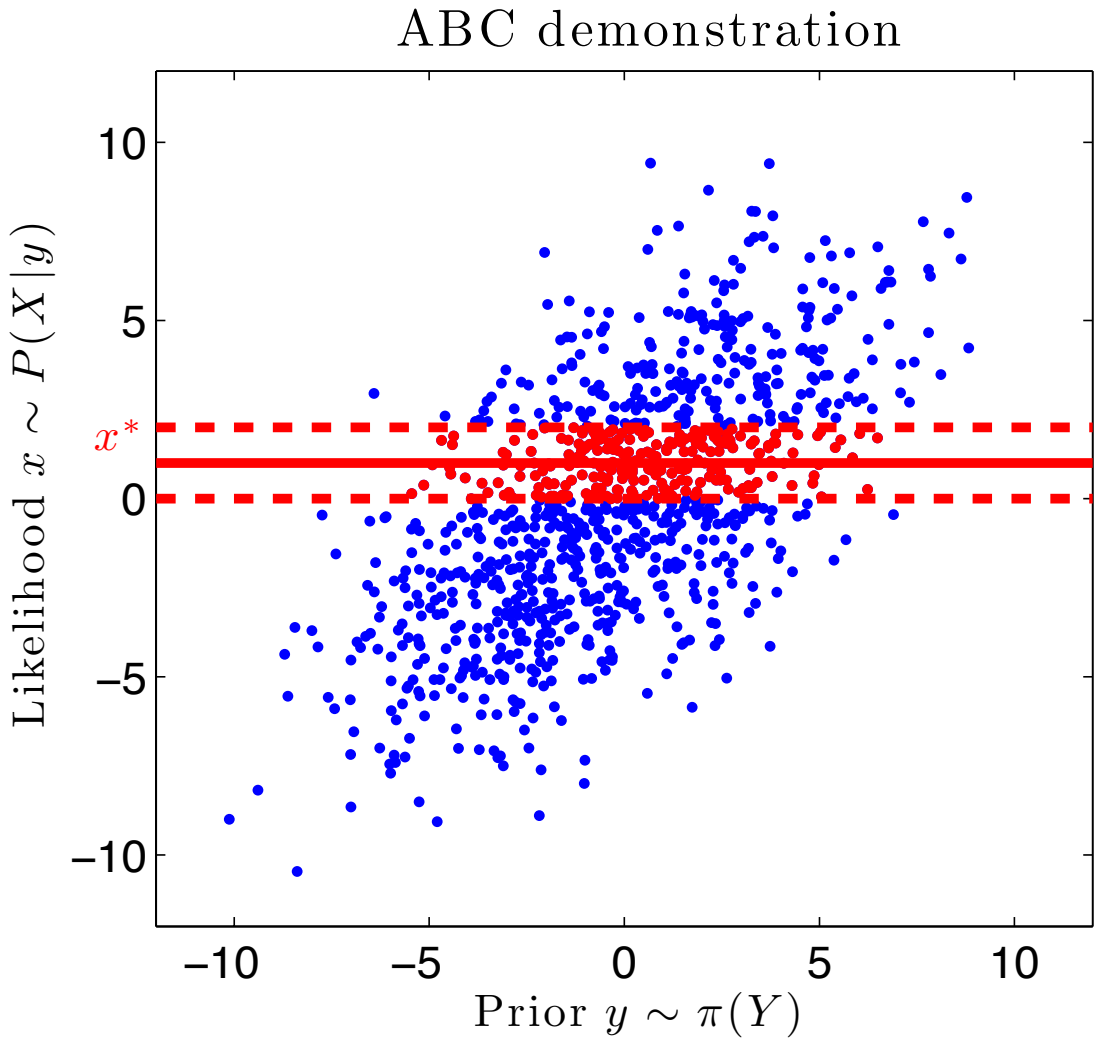
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



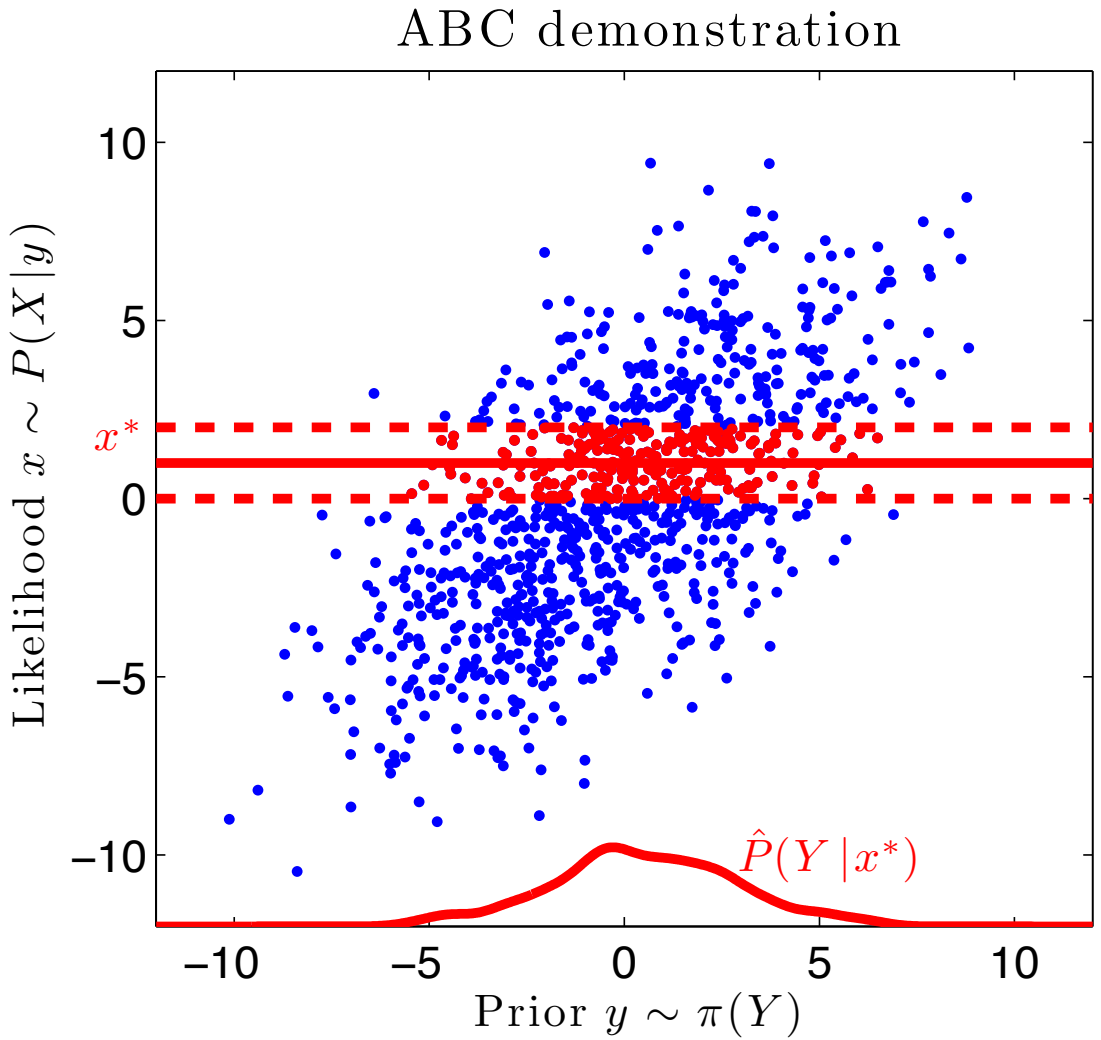
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



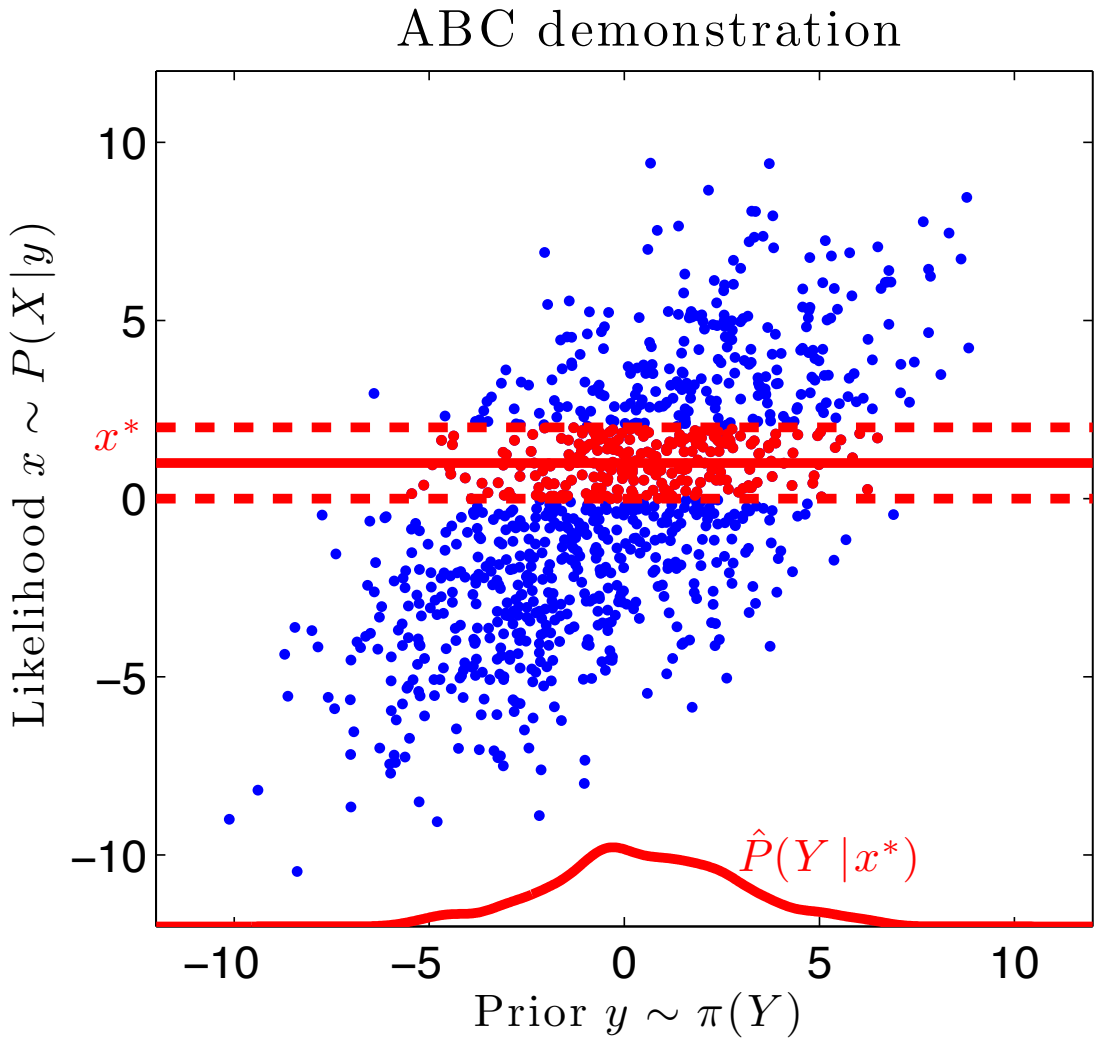
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):

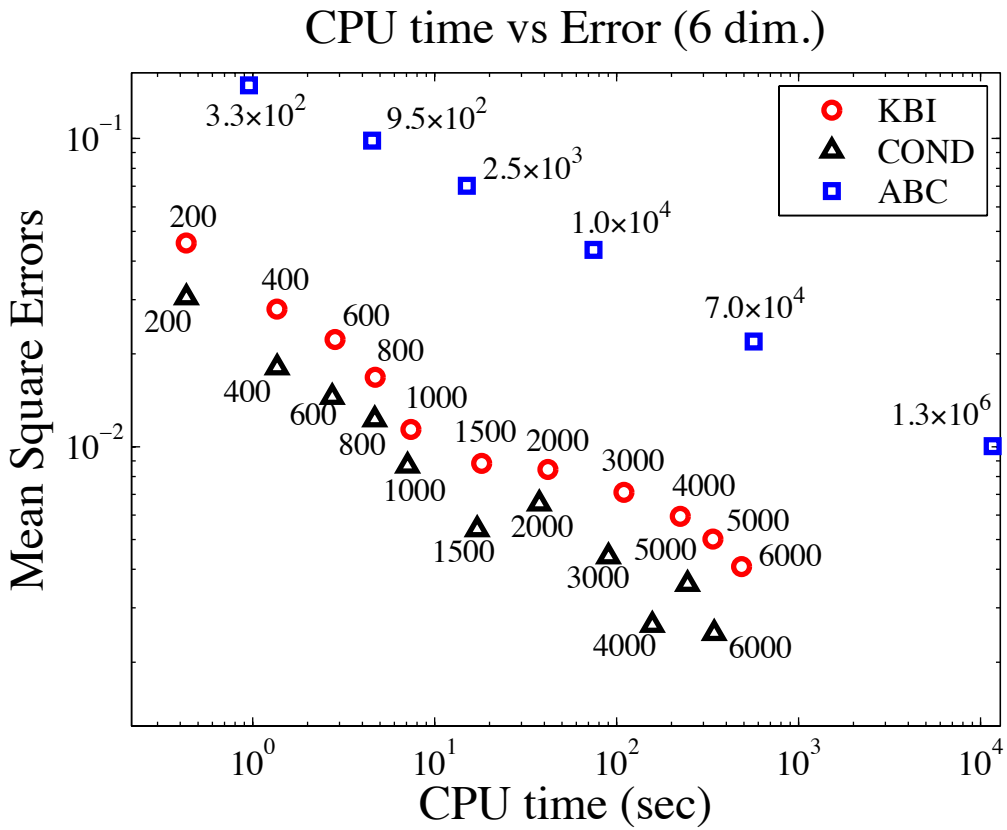


Needed: distance measure D, tolerance parameter τ.

Motivating example 2: simple Gaussian case

- $p(x, y)$ is $\mathcal{N}((0, \mathbf{1}_d^T)^T, V)$ with V a randomly generated covariance

Posterior mean on x: ABC vs kernel approach



Reminder: feature embeddings of probabilities

For all $f \in \mathcal{F}$,

The kernel trick:

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{F}}$$

The mean trick:

$$\mathbf{E}_{\mathbf{P}}(f(X)) = \langle \mu_{\mathbf{P}}, f \rangle_{\mathcal{F}}$$

$\mu_{\mathbf{P}}$ gives you expectations of all RKHS functions

When k characteristic, then $\mu_{\mathbf{P}}$ unique, e.g. Gauss, Laplace, ...

Bayes again

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{Q}(x|y)\pi(y)}{\int \mathbf{Q}(x|y)\pi(y)dy}$$

- $\mathbf{Q}(x|y)$ is likelihood
- π is prior

How would this look with kernel embeddings?

Bayes again

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{Q}(x|y)\pi(y)}{\int \mathbf{Q}(x|y)\pi(y)dy}$$

- $\mathbf{Q}(x|y)$ is likelihood
- π is prior

How would this look with kernel embeddings?

We need a conditional mean embedding: for all $f \in \mathcal{F}$,

$$\mathbf{E}_{Y|x^*} f(Y) = \langle f, \mu_{\mathbf{P}(y|x^*)} \rangle_{\mathcal{F}}$$

This will be obtained by RKHS-valued ridge regression

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \dots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \dots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Solve

$$\breve{A} = \arg\min_{A \in \mathbb{R}^{d' \times d}} \left(\|Y - AX\|^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right),$$

where

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^\top A) = \sum_{i=1}^{\min\{d,d'\}} \gamma_{A,i}^2$$

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \dots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Solve

$$\breve{A} = \quad \arg\min_{A \in \mathbb{R}^{d' \times d}} \left(\|Y - AX\|^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right),$$

where

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^\top A) = \sum_{i=1}^{\min\{d,d'\}} \gamma_{A,i}^2$$

Solution: $\breve{A} = C_{YX} \left(C_{XX} + m\lambda I \right)^{-1}$

Ridge regression and the conditional feature mean

Prediction at new point x:

$$\begin{aligned} y^* &= \breve{A}x \\ &= C_{YX}\left(C_{XX} + m\lambda I\right)^{-1} x \\ &= \sum_{i=1}^{m} \beta_i(x) y_i \end{aligned}$$

where

$$\beta_i(x) = (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \dots & k(x_m, x) \end{bmatrix}^\top$$

and

$$K := X^\top X \qquad\qquad k(x_1, x) = x_1^\top x$$

Ridge regression and the conditional feature mean

Prediction at new point x:

$$\begin{aligned} y^* &= \breve{A}x \\ &= C_{YX}\left(C_{XX} + m\lambda I\right)^{-1} x \\ &= \sum_{i=1}^{m} \beta_i(x) y_i \end{aligned}$$

where

$$\beta_i(x) = (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \dots & k(x_m, x) \end{bmatrix}^\top$$

and

$$K := X^\top X \qquad\qquad k(x_1, x) = x_1^\top x$$

What if we do everything in kernel space?

Ridge regression and the conditional feature mean

Recall our setup:

- Given training *pairs:*

$$(x_i, y_i) \sim \mathbf{P}_{XY}$$

- $\mathcal{F}$ on $\mathcal{X}$ with feature map φ_x and kernel $k(x, \cdot)$
- $\mathcal{G}$ on $\mathcal{Y}$ with feature map ψ_y and kernel $l(y, \cdot)$

We define the covariance between feature maps:

$$C_{XX} = \mathbf{E}_X \ (\varphi_X \otimes \varphi_X) \qquad C_{XY} = \mathbf{E}_{XY} \ (\varphi_X \otimes \psi_Y)$$

and matrices of feature mapped training data

$$X = \begin{bmatrix} \varphi_{x_1} & \dots & \varphi_{x_m} \end{bmatrix} \quad Y := \begin{bmatrix} \psi_{y_1} & \dots & \psi_{y_m} \end{bmatrix}$$

Ridge regression and the conditional feature mean

Objective: [Weston et al. (2003), Micchelli and Pontil (2005), Caponnetto and De Vito (2007), Grunewalder et al. (2012, 2013)]

$$\breve{A} = \arg \min_{A \in \mathrm{HS}(\mathcal{F},\mathcal{G})} \left(\|Y - AX\|_{\mathcal{G}}^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right), \qquad \|A\|_{\mathrm{HS}}^2 = \sum_{i=1}^{\infty} \gamma_{A,i}^2$$

Solution same as vector case:

$$\breve{A} = C_{YX} \left(C_{XX} + m\lambda I \right)^{-1},$$

Prediction at new x using kernels:

$$\begin{aligned} \breve{A}\varphi_x &= \begin{bmatrix} \psi(y_1) & \dots & \psi(y_m) \end{bmatrix} (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \dots & k(x_m, x) \end{bmatrix} \\ &= \sum_{i=1}^{m} \beta_i(x) \psi_{y_i} \end{aligned}$$

where $K_{ij} = k(x_i, x_j)$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := \breve{A}\varphi_x$$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := \breve{A}\varphi_x$$

We need $\breve{A}$ to have the property

$$\begin{aligned}
\mathbf{E}_{Y|x} g(Y) &\approx \langle g, \mu_{Y|x} \rangle_{\mathcal{G}} \\
&= \langle g, \breve{A}\varphi_x \rangle_{\mathcal{G}} \\
&= \langle \breve{A}^* g, \varphi_x \rangle_{\mathcal{F}} = (\breve{A}^* g)(x)
\end{aligned}$$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := \breve{A}\varphi_x$$

We need $\breve{A}$ to have the property

$$\begin{aligned}\mathbf{E}_{Y|x} g(Y) &\approx \langle g, \mu_{Y|x}\rangle_{\mathcal{G}} \\ &= \langle g, \breve{A}\varphi_x\rangle_{\mathcal{G}} \\ &= \langle \breve{A}^* g, \varphi_x\rangle_{\mathcal{F}} = (\breve{A}^* g)(x)\end{aligned}$$

Natural risk function for conditional mean

$$\mathcal{L}(A, \mathbf{P}_{XY}) := \sup_{\|g\|\leq 1} \mathbf{E}_X \left[\underbrace{\left(\mathbf{E}_{Y|X} g(Y)\right)}_{\text{Target}}(X) - \underbrace{(A^* g)}_{\text{Estimator}}(X) \right]^2,$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[\langle g, \psi_Y \rangle_{\mathcal{G}} - \langle A^* g, \varphi_X \rangle_{\mathcal{F}} \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - \left(A^* g\right)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - \left(A^* g\right)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[\langle g, \psi_Y \rangle_{\mathcal{G}} - \langle g, A\varphi_X \rangle_{\mathcal{G}} \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - \left(A^* g\right)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - \left(A^* g\right)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2 \\
&\leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

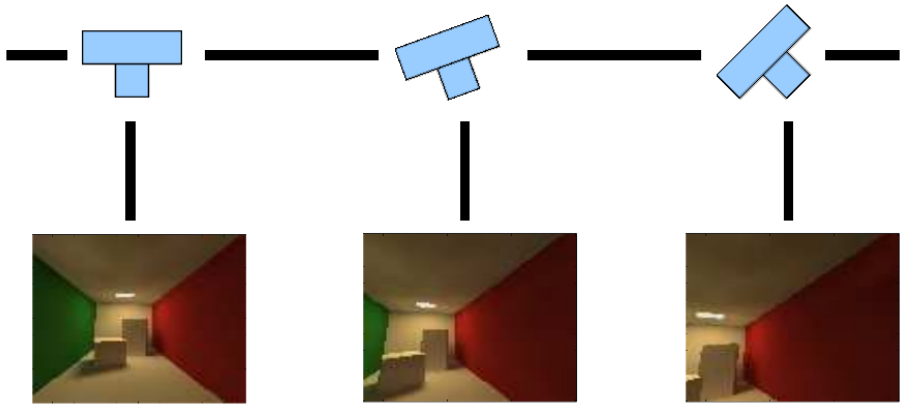
Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2 \\
&\leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2
\end{aligned}$$

If we assume $\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$ then upper bound tight.

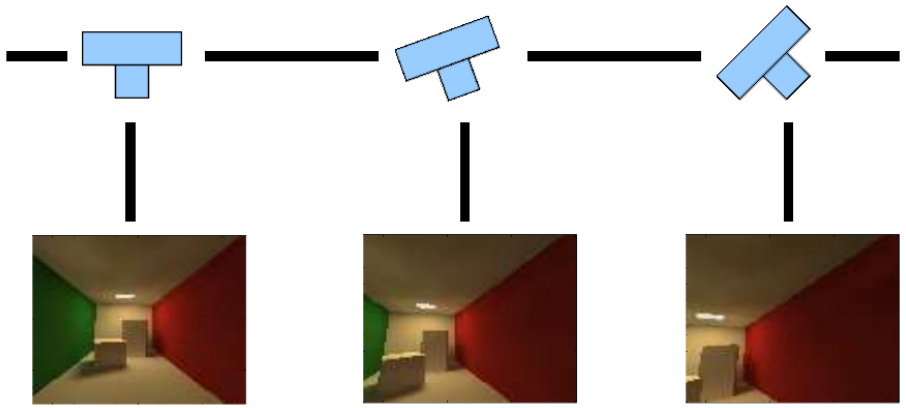
Experiment: Kernel Bayes' law vs EKF

- Compare with extended Kalman filter (EKF) on camera orientation task
- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0, 1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.



Experiment: Kernel Bayes' law vs EKF

- Compare with extended Kalman filter (EKF) on camera orientation task
- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.



Average MSE and standard errors (10 runs)

	KBR (Gauss)	KBR (Tr)	Kalman (9 dim.)	Kalman (Quat.)
$\sigma^2 = 10^{-4}$	0.210 ± 0.015	0.146 ± 0.003	1.980 ± 0.083	0.557 ± 0.023
$\sigma^2 = 10^{-3}$	0.222 ± 0.009	0.210 ± 0.008	1.935 ± 0.064	0.541 ± 0.022

Summary

- Bayesian inference without models [NIPS11, JMLR13]
- Three-way interactions [NIPS13]
- Energy distance/distance covariance is special case [ICML12, AOS13]
- Linear test on big data, kernel selection strategy [NIPS12]
- ...

Co-authors

- **From UCL:**
  - Luca Baldasssarre
  - Steffen Grunewalder
  - Guy Lever
  - Sam Patterson
  - Massimiliano Pontil
  - Dino Sejdinovic
- **External:**
  - Karsten Borgwardt, MPI
  - Wicher Bergsma, LSE
  - Kenji Fukumizu, ISM
  - Zaid Harchaoui, INRIA
  - Bernhard Schoelkopf, MPI
  - Alex Smola, CMU/Google
  - Le Song, Georgia Tech
  - Bharath Sriperumbudur, Cambridge

Selected references

Characteristic kernels and mean embeddings:

- Smola, A., Gretton, A., Song, L., Schoelkopf, B. (2007). A hilbert space embedding for distributions. ALT.
- Sriperumbudur, B., Gretton, A., Fukumizu, K., Schoelkopf, B., Lanckriet, G. (2010). Hilbert space embeddings and metrics on probability measures. JMLR.
- Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., Smola, A. (2012). A kernel two- sample test. JMLR.

Two-sample, independence, conditional independence tests:

- Gretton, A., Fukumizu, K., Teo, C., Song, L., Schoelkopf, B., Smola, A. (2008). A kernel statistical test of independence. NIPS
- Fukumizu, K., Gretton, A., Sun, X., Schoelkopf, B. (2008). Kernel measures of conditional dependence.
- Gretton, A., Fukumizu, K., Harchaoui, Z., Sriperumbudur, B. (2009). A fast, consistent kernel two-sample test. NIPS.
- Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., Smola, A. (2012). A kernel two- sample test. JMLR

Energy distance, relation to kernel distances

- Sejdinovic, D., Sriperumbudur, B., Gretton, A.,, Fukumizu, K., (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. Annals of Statistics.

Three way interaction

- Sejdinovic, D., Gretton, A., and Bergsma, W. (2013). A Kernel Test for Three-Variable Interactions. NIPS.

Selected references (continued)

Conditional mean embedding, RKHS-valued regression:

- Weston, J., Chapelle, O., Elisseeff, A., Schölkopf, B., and Vapnik, V., (2003). Kernel Dependency Estimation, NIPS.
- Micchelli, C., and Pontil, M., (2005). On Learning Vector-Valued Functions. Neural Computation.
- Caponnetto, A., and De Vito, E. (2007). Optimal Rates for the Regularized Least-Squares Algorithm. Foundations of Computational Mathematics.
- Song, L., and Huang, J., and Smola, A., Fukumizu, K., (2009). Hilbert Space Embeddings of Conditional Distributions. ICML.
- Grunewalder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., Pontil, M. (2012). Conditional mean embeddings as regressors. ICML.
- Grunewalder, S., Gretton, A., Shawe-Taylor, J. (2013). Smooth operators. ICML.

Kernel Bayes rule:

- Song, L., Fukumizu, K., Gretton, A. (2013). Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine.
- Fukumizu, K., Song, L., Gretton, A. (2013). Kernel Bayes rule: Bayesian inference with positive definite kernels, JMLR

References

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.

M. Arcones and E. Giné. On the bootstrap of u and v statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *J. Multivariate Anal.*, 88:190–206, 2004.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups.* Springer, New York, 1984.

R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, UK, 2002.

Andrey Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 1.* John Wiley and Sons, 2nd edition, 1994.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function.* PhD thesis, University of Jyväskylä, 1995.

R. Lyons. Distance covariance in metric spaces. arXiv:1106.5758, to appear in Ann. Probab., June 2011.

C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

T. Read and N. Cressie. *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis.* Springer-Verlag, New York, 1988.

R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

G. Székely and M. Rizzo. A new test for multivariate normality. *J. Multivariate Anal.*, 93:58–80, 2005.

G. Székely and M. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 4(3):1233–1303, 2009.

G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Stat.*, 35(6):2769–2794, 2007.