ICA and Kernel Distribution Testing

(Lecture notes, MLSS 06, Canberra)

Arthur Gretton
Department of Biological Cybernetics
Max-Planck Institute, Tübingen, Germany.
arthur@tuebingen.mpg.de

Overview

- Independent component analysis: recover the linear mixing that combines independent sources
- Kernel independence testing: given a sample of m pairs $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, are the random variables x and y independent?
- The two sample problem: are samples $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$ generated from the same distribution?

Some notation and conventions

- Random variables are written *sans serif*, eg x, $\boldsymbol{\mathsf{x}}$
- Vector spaces are written in caligraphic font, eg $x \in \mathcal{X}$
- Probability distributions and densities are $\mathbf{\mathsf{P}}_{\mathsf{x}}(A)$, expectations are $\mathbf{\mathsf{E}}_{\mathsf{x}}(\mathsf{x})$
- Covariance matrices are written

$$\mathbf{C}_{xy} := \mathbf{\mathsf{E}}_{\boldsymbol{\mathsf{x}},\boldsymbol{\mathsf{y}}}(\boldsymbol{\mathsf{x}}\boldsymbol{\mathsf{y}}^\top) - \mathbf{\mathsf{E}}_{\boldsymbol{\mathsf{x}}}(\boldsymbol{\mathsf{x}})\mathbf{\mathsf{E}}_{\boldsymbol{\mathsf{y}}}(\boldsymbol{\mathsf{y}}^\top)$$

ICA

...where to be careful when doing it

ICA (Population version)

- Indepdendent component analysis: we assume

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

  - $\mathbf{x}$ vector of observations, $\mathbf{A}$ (**unknown**) mixing matrix,
  - $\mathbf{s}$ a vector of l **unknown, independent inputs:** $\mathbf{P}_{\mathbf{s}} = \prod_{i=1}^{l} \mathbf{P}_{\mathbf{s}_i}$
  - **$\mathbf{B}$ is our estimate of $\mathbf{A}^{-1}$**

- We want to find
  - An estimate $\mathbf{y}$ of $\mathbf{s}$, using...
  - ...an estimate $\mathbf{B}$ of $\mathbf{A}^{-1}$:

$$\hat{\mathbf{s}} := \mathbf{y} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{s}$$

ICA (empirical version)

- Indepdendent component analysis: we assume

$$\mathbf{X} = \mathbf{A}\mathbf{S},$$

- Data matrices are $l \times m$, where

$$\mathbf{X} := \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_l \end{bmatrix} \qquad \text{and} \qquad \mathbf{S} := \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_l \end{bmatrix}$$

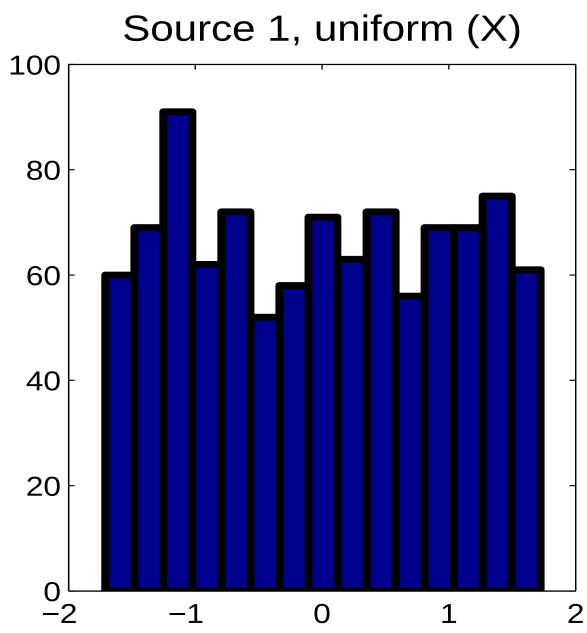
- Vectors $\mathbf{x}_i$ and $\mathbf{s}_i$ contain m i.i.d. samples

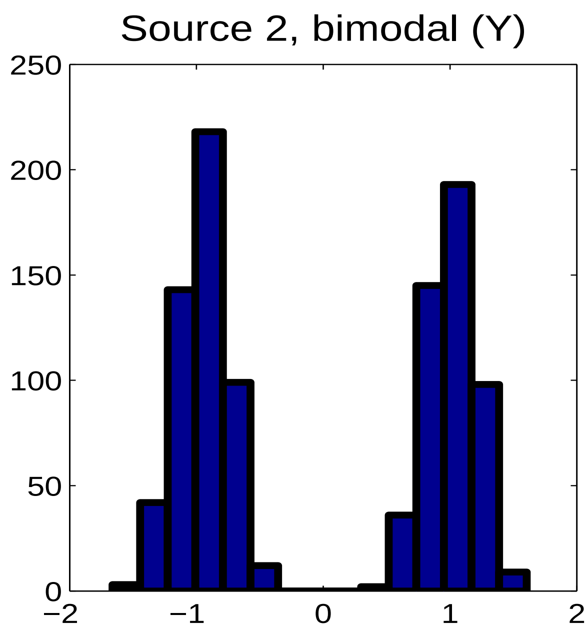
ICA examples

- Sounds mixed together (“cocktail party” problem)
- EEG recordings (brain, fetal heartbeat)
- Economics
- Image processing

A toy example (1)

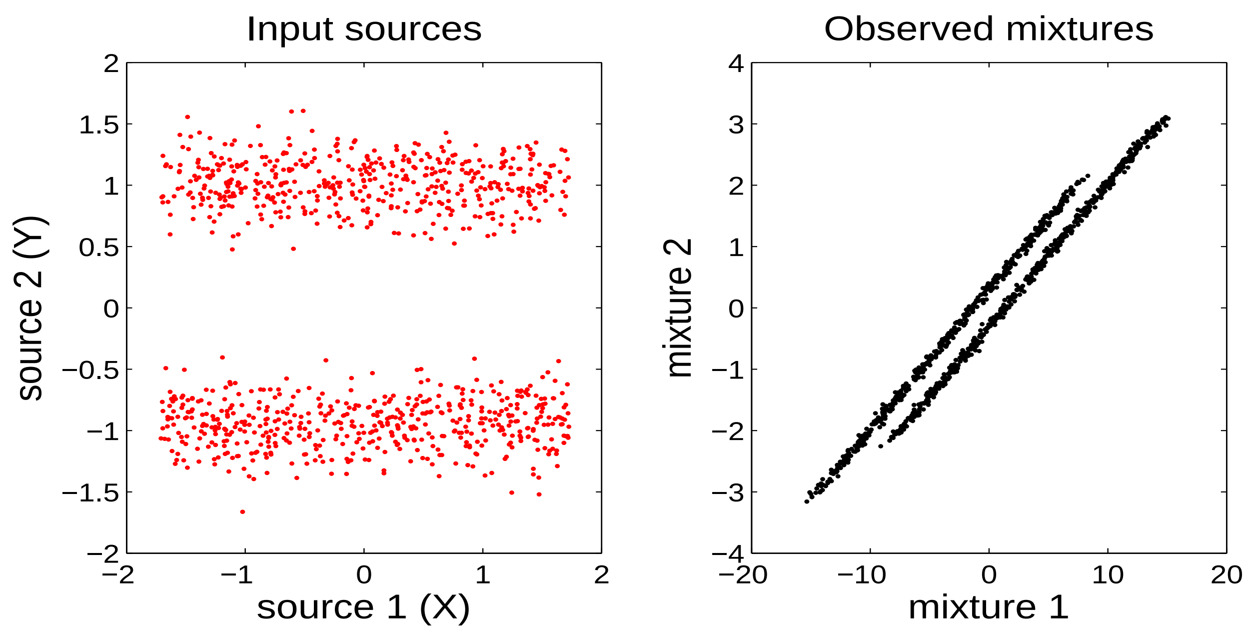
- We have two distributions: $\mathbf{P_x}$ is uniform, $\mathbf{P_y}$ is bimodal





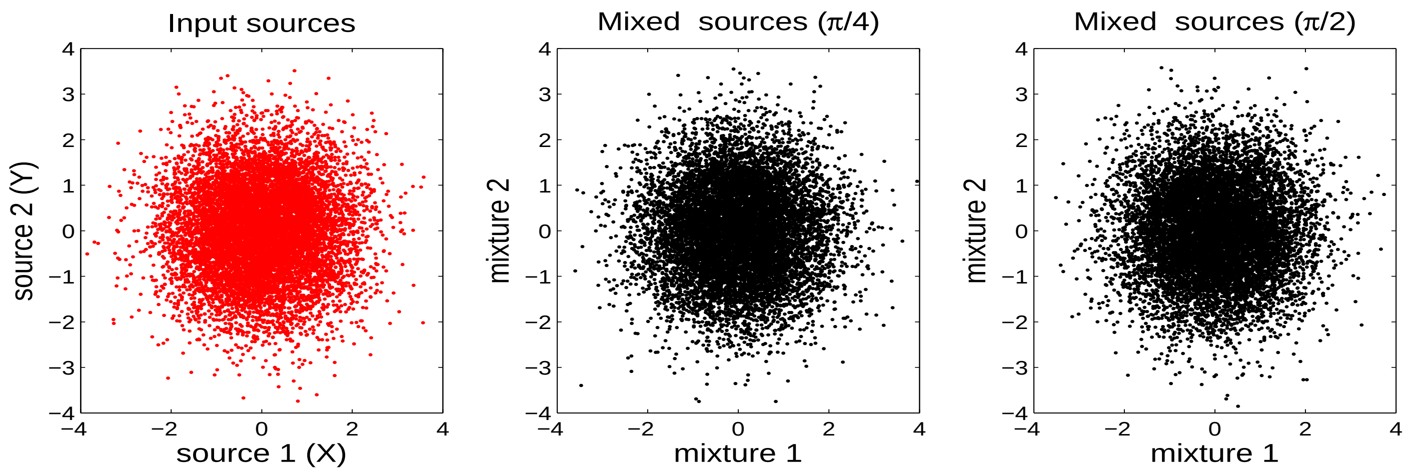
A toy example (2)

- Initial unmixed RVs in red, mixed RVs in black



Things that are impossible for ICA (1)

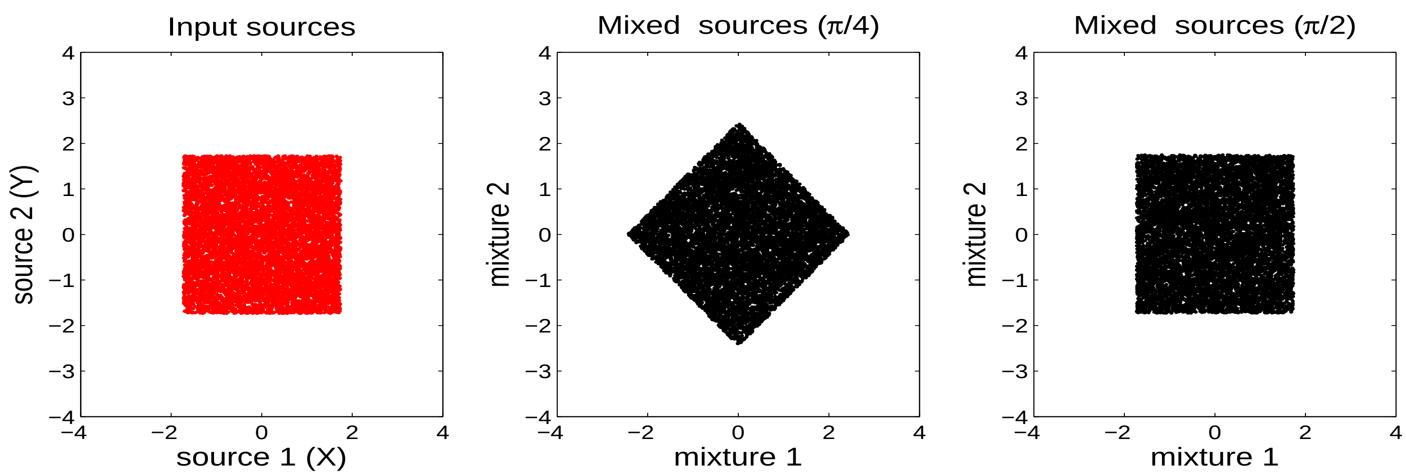
- Assuming we **know** what the original signals look like, can we determine **how observations were mixed**?
  - Reminder: ICA doesn't care about the sources: it only tries to recover the mixing matrix
- First example:
  - Both PDFs Gaussian
  - Observe mixtures at different rotation angles
  - Can we ever recover the mixing?



Input sources
source 2 (Y)
source 1 (X)
Mixed sources (π/4)
mixture 2
mixture 1
Mixed sources (π/2)
mixture 2
mixture 1

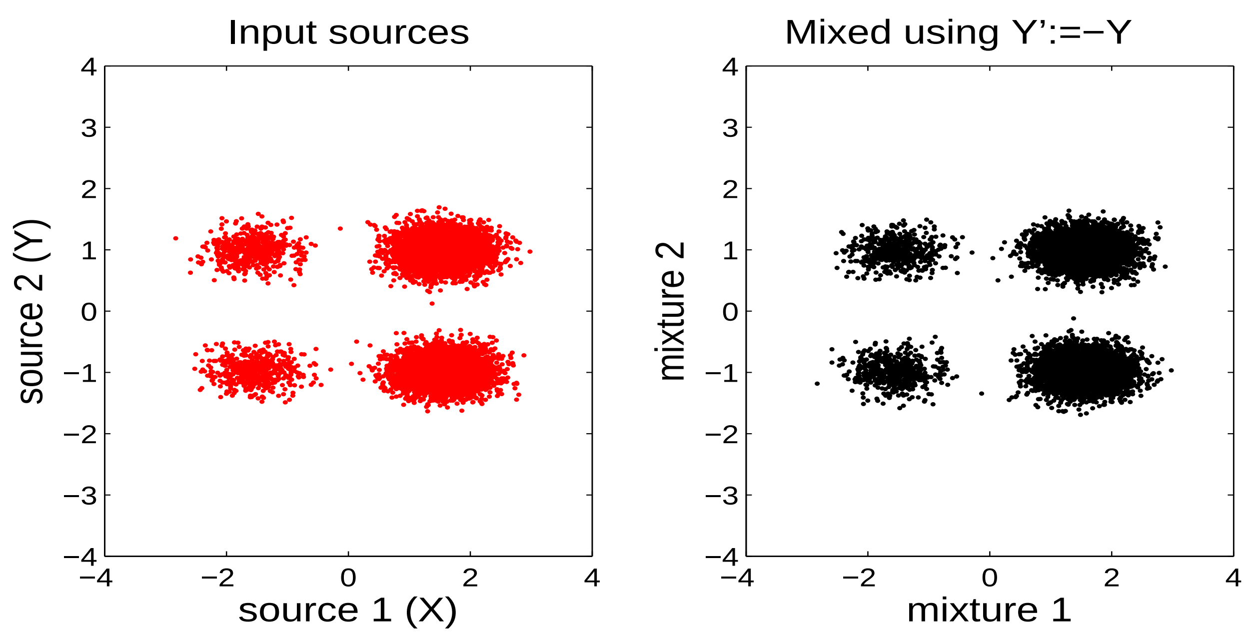
Things that are impossible for ICA (2)

- Second example:
  - Both PDFs uniform, symmetric about origin
  - Observe mixtures at different rotation angles
  - What happens when rotation angle is maximum $(\pi/2)$?



Things that are impossible for ICA (3)

- Third example:
  - RV on x-axis has asymmetric PDF, that on y-axis has symmetric pdf
  - What happens if the mixing matrix negates the Y variable?



Things that are impossible for ICA (4)

- Separating RVs that are everywhere constant
- Separating multiple Gaussians
- Recovering signal order
- Recovering signal amplitude

ICA Step 1

Decorrelation

First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between observations $\mathbf{x}$
- Call whitened signals $\mathbf{t}$: we haven't reached unmixed signals $\mathbf{y}$
- Whiten the observations:

$$\mathbf{t} = \mathbf{B}_w \mathbf{x} \quad \text{where} \quad \mathbf{C}_{tt} := \mathbf{E}_\mathbf{t}(\mathbf{t}\mathbf{t}^\top) - \mathbf{E}_\mathbf{t}(\mathbf{t})\mathbf{E}_\mathbf{t}(\mathbf{t}^\top) = \mathbf{I}$$

- We thus break up $\mathbf{B}$ as follows:

$$\mathbf{B} = \mathbf{B}_r \mathbf{B}_w$$

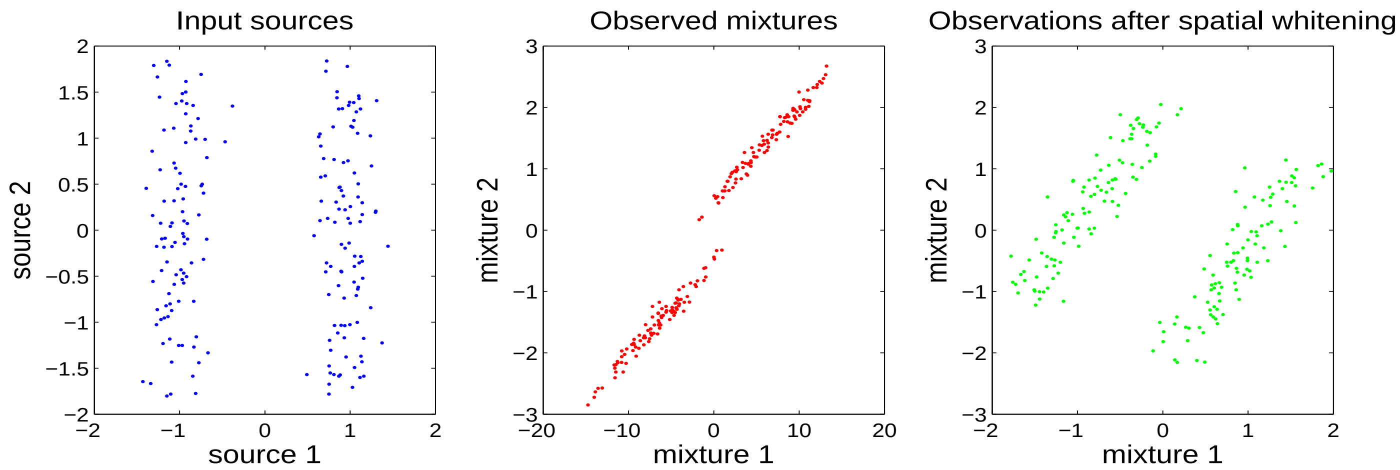
  - $\mathbf{B}_w$ is a whitening matrix
  - $\mathbf{B}_r$ is remaining demixing operation (more soon!)

- Reminder: this is done by using the SVD of $\mathbf{C}_{tt} = \mathbf{S}\Lambda\mathbf{S}^\top$:

$$\mathbf{B}_w = \Lambda^{-1/2}\mathbf{S}^\top$$

Example: what does decorrelation achieve?

- A uniform distribution on the interval $[-2, 2]$
- A mixture of two Gaussians with equal probability, means $+1$ and -1



Decorrelation: a drawback

A small warning: in theory, it is better not to break up the unmixing matrix in this way, since there is a loss in accuracy (statistically less efficient).

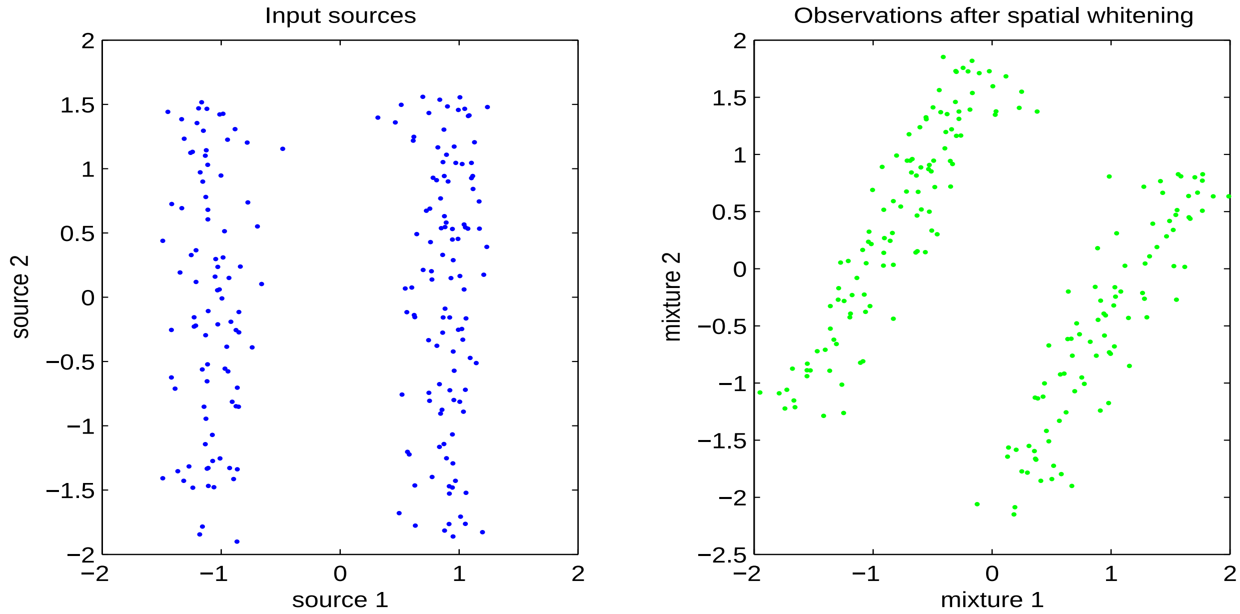
In practice, most ICA methods do decorrelation first, and the effect is not really noticeable.

ICA Step 2(a)

Rotation: maximum likelihood

What is left: *rotation*

- To recover original signal, need to rotate (see figure)
- We assume from now on that only the rotation remains to be done



Rotation (continued)

- For two signals, the rotation is expressed

$$\mathbf{B}_r = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

- This generalises to higher dimensions, eg for $l = 3$,

$$\mathbf{B}_r := \begin{bmatrix} \cos(\theta_z) & -\sin(\theta_z) & 0 \\ \sin(\theta_z) & \cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos(\theta_y) & 0 & -\sin(\theta_y) \\ 0 & 1 & 0 \\ \sin(\theta_y) & 0 & \cos(\theta_y) \end{bmatrix}$$
$$\times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) \\ 0 & \sin(\theta_x) & \cos(\theta_x) \end{bmatrix}$$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}^{-1}, \hat{\mathsf{P}}_{\mathsf{s}})$
  - Reminder: we use $\mathbf{B}^{-1}$ here since $\mathbf{B}$ the *unmixing* matrix
  - Another reminder: model must have $\hat{\mathsf{P}}_{\mathsf{s}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{\mathsf{s}_i}$
- With this model, our **estimated** density of observations is

$$\hat{\mathsf{P}}_{\mathsf{x}} = |\det(\mathbf{B}^{-1})|^{-1} \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{Bx}) = |\det(\mathbf{B})| \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{Bx})$$

- Maximise the *expected log likelihood*,

$$L := \mathsf{E}_{\mathsf{x}} \left[\log \hat{\mathsf{P}}_{\mathsf{x}} \right] = \mathsf{E}_{\mathsf{x}} \left[\log |\det(\mathbf{B})| + \log \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{Bx}) \right]$$

- Empirical expression:

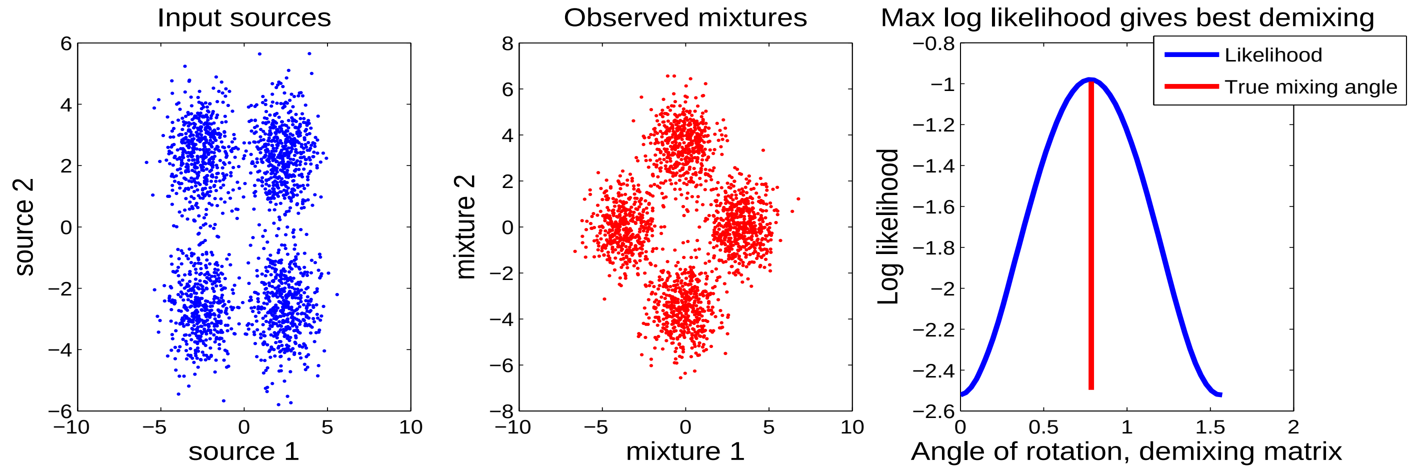
$$\widehat{L} := \log |\det(\mathbf{B})| + \frac{1}{m} \sum_{j=1}^{m} \log \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{Bx}_j)$$

Maximum likelihood: example

- The probability distribution of both source densities is

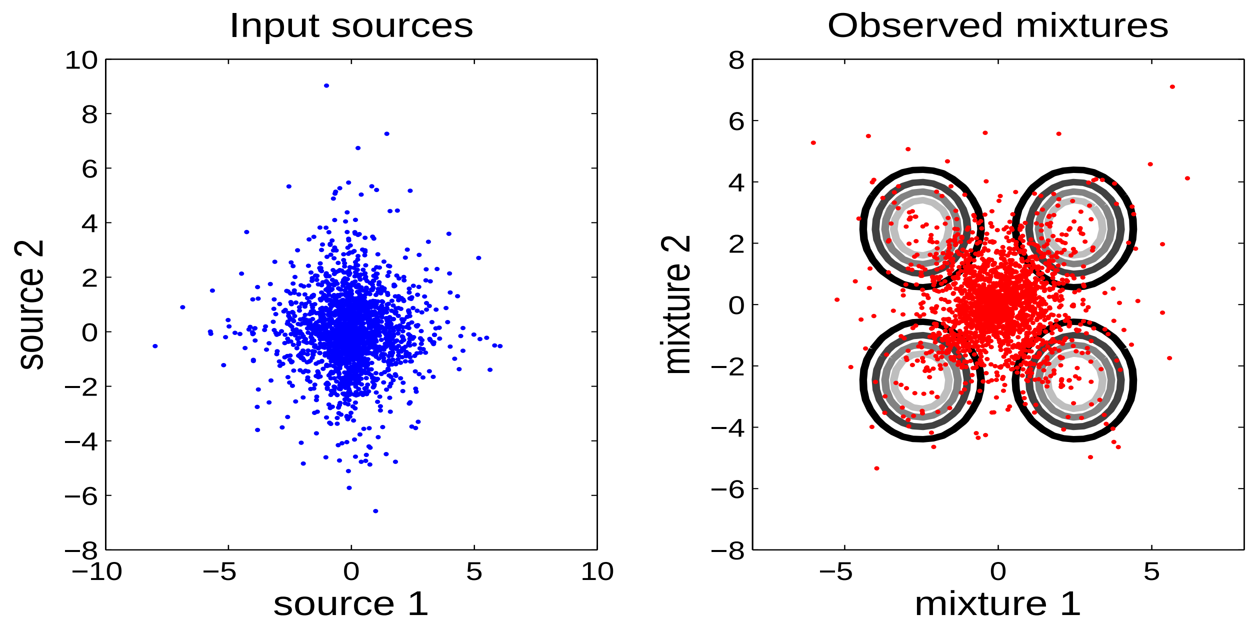
$$\frac{1}{2}\left(\mathcal{N}(-2.5,1)+\mathcal{N}(2.5,1)\right),$$

where $\mathcal{N}(\mu,\sigma^2)$ is a Gaussian with mean μ and variance σ^2



Maximum likelihood: where it fails

- Model as before, but true source densities are Laplace.
- Why is this so wrong?



ICA Step 2(b)

Rotation: contrast functions

What is a copy?

- The random vector $\mathbf{s}$ is a copy of $\mathbf{x}$ if and only if $\mathbf{x} = \mathbf{C}\mathbf{s}$, where $\mathbf{C}$ does only:
  - Permutations, e.g. $\mathbf{C} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
  - Sign swaps, e.g. $\mathbf{C} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$
  - Rescalings, e.g. $\mathbf{C} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$
  - Some combination of several of the above
- The most we can hope for in ICA is to recover a **copy** of the signals

Contrast functions

- Ideally: contrast $\phi(\mathbf{y}) = 0$ if and only if all components of $\mathbf{y}$ mutually independent:

$$\mathsf{P}_{\mathbf{y}} = \prod_{i=1}^{l} \mathsf{P}_{\mathsf{y}_i}.$$

  - Under our mixing assumptions: contrast $\phi(\mathbf{Cs}) = 0$ if and only if $\mathbf{Cs}$ a **copy** of $\mathbf{s}$
  - How people *really* use it: contrast should be "smallest" when random variables are "most independent"

- There exist contrast functions that have nothing to do with max likelihood...

- ...but max likelihood induces the "best" contrast (when correct!)

Contrast functions and maximum likelihood

How does the maximum likelihood relate to contrast functions?

- The max likelihood solution induces a contrast function:

$$L := \mathbf{E}_{\mathbf{x}} \left[\log \hat{\mathbf{P}}_{\mathbf{x}} \right] = -D_{\mathrm{KL}}(\mathbf{P}_{\mathbf{Bx}} || \hat{\mathbf{P}}_{\mathbf{s}}) + \text{const}$$

- What is KL divergence? Given two densities $\mathbf{P}_{\mathbf{x}}$, $\mathbf{Q}_{\mathbf{x}}$ defined on $\mathcal{X} \subset \mathbb{R}^n$, then

$$D_{\mathrm{KL}}(\mathbf{P}_{\mathbf{x}} || \mathbf{Q}_{\mathbf{x}}) = \int_{\mathcal{X}} \mathbf{P}_{\mathbf{x}}(\mathbf{x}) \log \left(\frac{\mathbf{P}_{\mathbf{x}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{x}}(\mathbf{x})} \right) d\mathbf{x}.$$

- $D_{\mathrm{KL}}(\mathbf{P}_{\mathbf{x}} || \mathbf{Q}_{\mathbf{x}}) \geq 0$ with equality if and only if $\mathbf{P}_{\mathbf{x}} = \mathbf{Q}_{\mathbf{x}}$ almost everywhere.

- ...thus $\phi_{ML}(\mathbf{y}) = D_{\mathrm{KL}}(\mathbf{P}_{\mathbf{Bx}} || \hat{\mathbf{P}}_{\mathbf{s}})$ is a contrast *as long as* $\hat{\mathbf{P}}_{\mathbf{s}} = \mathbf{P}_{\mathbf{s}}$

Contrast functions and mutual information (1)

- The mutual information is just the KL divergence between the joint distribution and the product of the marginals:

$$I(\mathsf{y}_i, \mathsf{y}_j) = \int_{\mathcal{Y}} \mathbf{P}_{\mathsf{y}_i,\mathsf{y}_j}(y_i, y_j) \log\left(\frac{\mathbf{P}_{\mathsf{y}_i,\mathsf{y}_j}(y_i, y_j)}{\mathbf{P}_{\mathsf{y}_i}(y_i)\mathbf{P}_{\mathsf{y}_j}(y_j)}\right) dy_i\, dy_j$$

- This is also a contrast function:

$$I(\mathsf{y}_i, \mathsf{y}_j) = 0 \quad \text{iff} \quad \mathbf{P}_{\mathsf{y}_i,\mathsf{y}_j} = \mathbf{P}_{\mathsf{y}_i}\mathbf{P}_{\mathsf{y}_j}$$

- Little used in ICA:
  - Hard to find good empirical estimates
  - Hard to optimise

Contrast functions and mutual information (2)

- Simplification: when rotation only is considered, need only 1-D entropies (see [8] in references)
- Reason:

$$D_{\mathrm{KL}}\left(\mathbf{P}_{\mathsf{y}} \middle\| \prod_{i=1}^{l} \mathbf{P}_{\mathsf{y}_i}\right) = \sum_{i=1}^{l} h\left(\mathsf{y}_i\right) - h\left(\mathbf{x}\right) - \log\left|\det \mathbf{B}\right|.$$

  where $h(\mathsf{y}) = -\mathbf{E}_{\mathsf{y}} \log(\mathbf{P}_{\mathsf{y}}(y))$
- $h\left(\mathbf{x}\right)$ constant wrt $\mathbf{B}$: only function of observations $\mathbf{x}$
- $\log\left|\det \mathbf{B}\right| = 1$ when $\mathbf{B}$ are rotations
- Entropies are also hard to compute: IDEA: use

$$\phi(\mathbf{y}) = \sum_{j=1}^{l} \mathbf{E}_{\mathsf{y}_j}\left(f\left(\mathsf{y}_j\right)\right)$$

  for some other nonlinear $f\left(y\right)$

Contrast functions (3): Some famous cases

This slide represents a gross simplification of what really goes on. Read the papers!

- What kind of nonlinear $f(y)$ can we use to make our contrasts?
- Infomax-type contrast:

  $$f(y) = a - exp(-y^2/2)\text{sech}^2(y)$$

  for some $a \geq 1$
- Fast ICA-type contrast:

  $$f(y) = \frac{1}{a}\log\cosh(ay),$$

  where $a \geq 1$.
- Jade-type contrast:

  $$f(y) = y^4$$

Kurtosis: an important concept

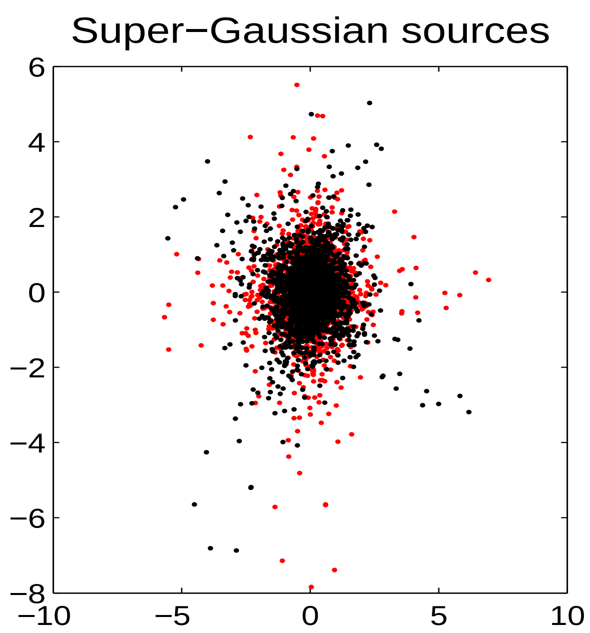
- Kurtosis definition: when mean is zero,

$$\kappa_4 = \mathbf{E}_{\mathsf{x}}\left(\mathsf{x}^4\right) - 3\left(\mathbf{E}_{\mathsf{x}}\left(\mathsf{x}^2\right)\right)^2 .$$

- Source densities can be super-Gaussian (positive kurtosis) or sub-Gaussian (negative kurtosis)
- Zero kurtosis does not mean Gaussian!
- Certain popular contrast functions depend explicitly on kurtosis of unmixed signals
- Other contrast functions only work when kurtosis is positive or negative

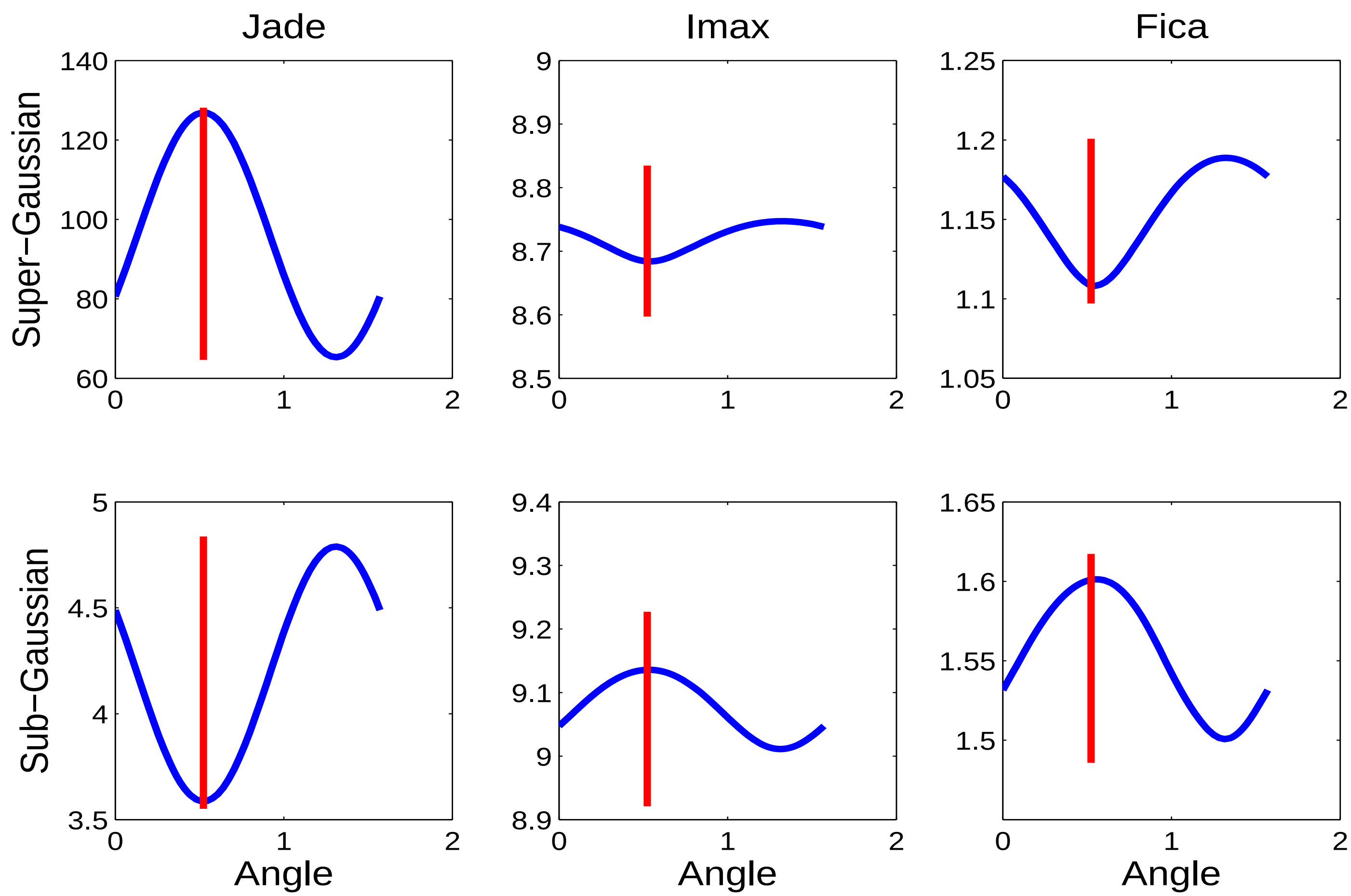
Contrast functions: Example (1)

- Samples drawn from Super- and Sub-Gaussian distributions below:



Contrast functions: Example (2)

- Results for Jade, Infomax, and Fast ICA contrasts



Disclaimer!

- The implementations of Jade, Fast ICA, and Infomax on the internet work for positive and negative kurtoses! I.e. real life algorithms are more complicated.
- That said, the foregoing demonstrates the danger of blindly using random ICA software on the internet without knowing what it does.

ICA for non-i.i.d. processes

ICA for non-i.i.d. signals (1)

- We can get extra information from sources not being i.i.d.
- Assume zero mean.
- Assume that our observation vector $\mathbf{x}(t)$ now depends on *time shifted values* $\mathbf{x}(t+\tau)$, where $\tau \geq 1$, and that the process is *stationary*
- Define the covariance

$$\mathbf{C}_{xx}(\tau) = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t+\tau)),$$

where the above is indpendent of τ due to stationarity

- Hint: the ideas we're about to use were described for decorrelation in i.i.d. case

ICA for non-i.i.d. signals (2)

- Our assumption that the *inputs* are uncorrelated causes the following to hold:

$$\begin{aligned} \Lambda &= \mathbf{E}\left(\mathbf{s}(t)\mathbf{s}^\top(t)\right) = \mathbf{E}\left(\left(\mathbf{A}^{-1}\mathbf{x}(t)\right)\left(\mathbf{A}^{-1}\mathbf{x}(t)\right)^\top\right) \\ &= \mathbf{A}^{-1}\mathbf{C}_{xx}(0)\left(\mathbf{A}^{-1}\right)^\top \end{aligned}$$

  where $\boldsymbol{\Lambda}$ is a diagonal matrix

- But the following can also be assumed: for any $\tau \geq 1$,

$$\widetilde{\Lambda} = \mathbf{E}\left(\mathbf{s}(t)\mathbf{s}^\top(t+\tau)\right) = \mathbf{A}^{-1}\mathbf{C}_{xx}(\tau)\left(\mathbf{A}^{-1}\right)^\top$$

- Combining both criteria: get

$$\mathbf{C}_{xx}(0)\mathbf{C}_{xx}^{-1}(\tau)\mathbf{A} = \mathbf{A}\left(\Lambda\widetilde{\Lambda}^{-1}\right)$$

- Methods exist to solve for a greater number of delays (see references): procedure is called *joint diagonalisation*

Advanced (kernel!) independence measures

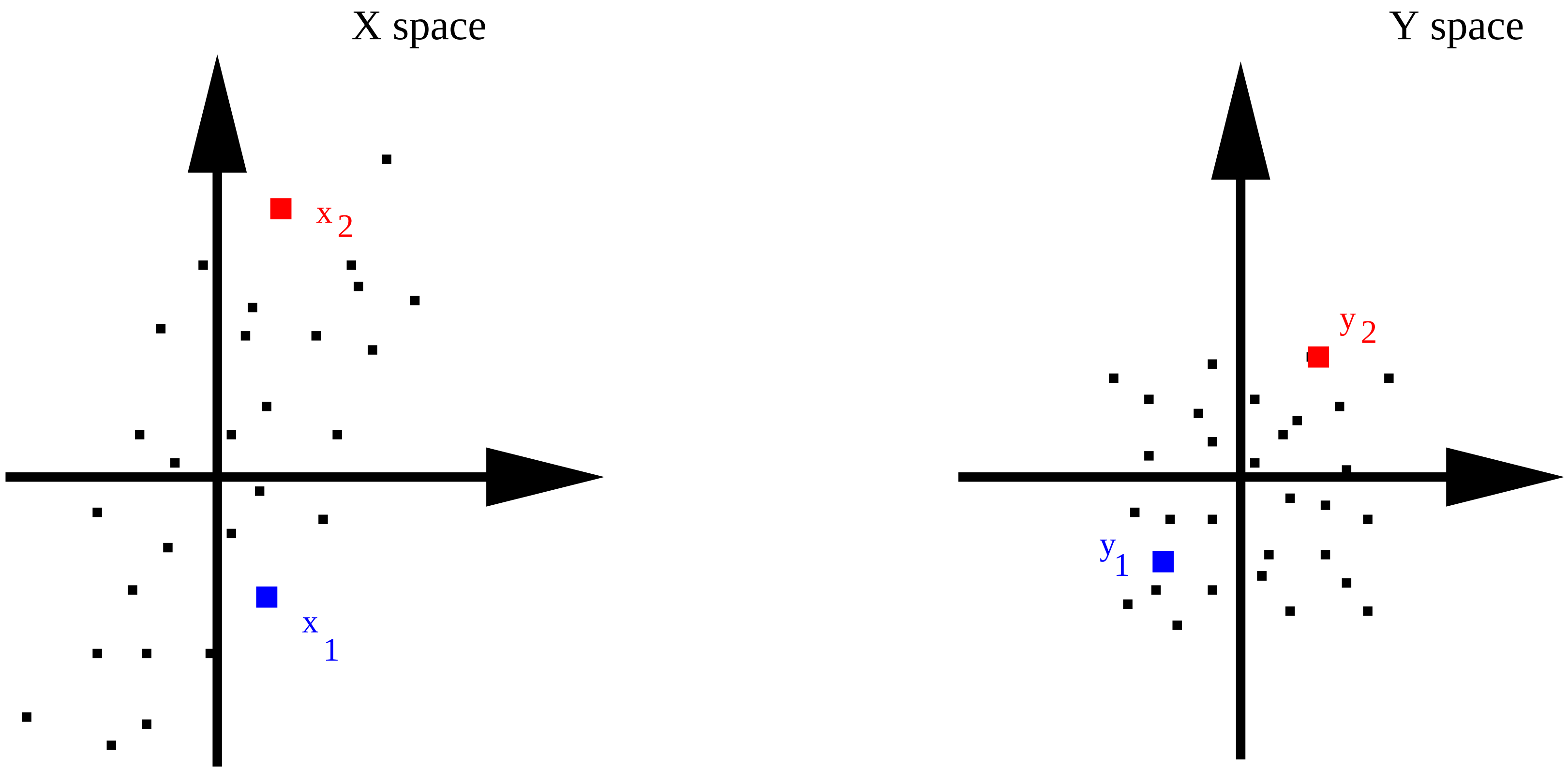
Kernel dependence measures

- Kernel dependence measures
  - Zero only at independence
  - Take into account high order moments
  - Make "sensible" assumptions about smoothness
- Applications
  - Independent component analysis (ICA)
  - Feature selection (Fukumizu *et al.*)
  - Dependence detection between voxel activity in Macaque visual cortex (V1)

Outline

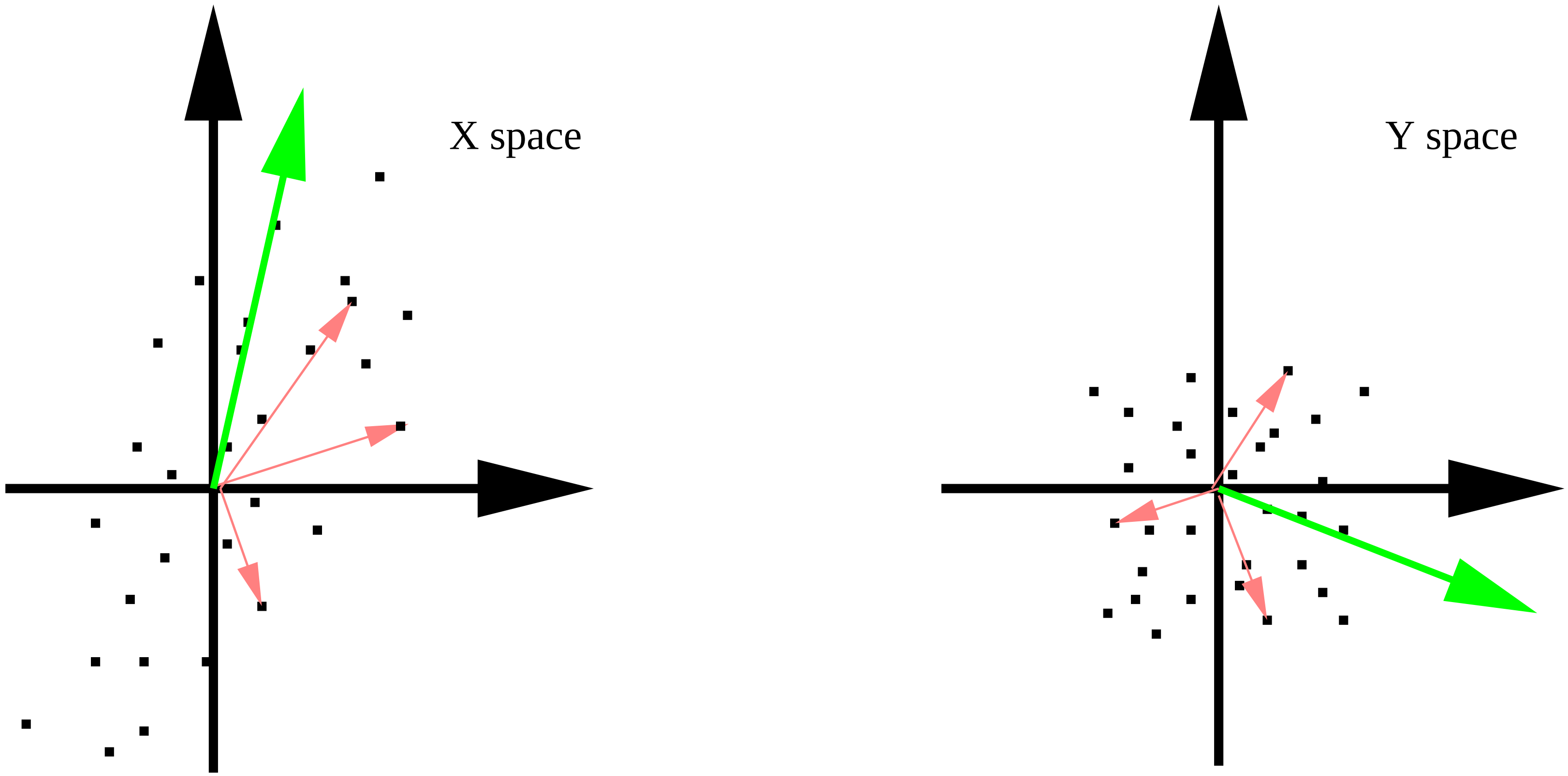
- Constrained covariance (COCO)
  - Covariance in RKHSs
  - Three useful properties of COCO
    * Independence measure when kernels universal
    * How to derive independence test from independence measure
      - Cases where dependence hard to detect
      - How to choose kernel?
    * Error prob. of test drops quickly as sample size increases
- Use of COCO (and other kernel dependence measures) in ICA

Dependence detection



- Get m pairs of points in different spaces
- Are the RVs x and y dependent?

A second order method



- Choose directions, get dot product with all points.
- Directions chosen such that the vectors of projections have biggest covariance. **Is covariance 0?**

Take *nonlinear* features

- Points in each space mapped to vectors of nonlinear features:
  - $x \to \mathbf{x} := \begin{bmatrix} \sqrt{\lambda_1}\varphi_1(x) & \sqrt{\lambda_2}\varphi_2(x) & \dots & \sqrt{\lambda_n}\varphi_n(x) & \dots \end{bmatrix}$
  - $y \to \mathbf{y} := \begin{bmatrix} \sqrt{\lambda_1}\varphi_1(y) & \sqrt{\lambda_2}\varphi_2(y) & \dots & \sqrt{\lambda_n}\varphi_n(y) & \dots \end{bmatrix}$
  - $\mathbf{x} \in \mathcal{H}_\mathcal{X}$ and $\mathbf{y} \in \mathcal{H}_\mathcal{Y}$, can be infinite dimensional
  - As n increases, λ_n smaller and φ_n less smooth
- Define projection vectors in each space: $\mathbf{f} \in \mathcal{H}_\mathcal{X}$, $\mathbf{g} \in \mathcal{H}_\mathcal{Y}$.
- Formal definition of COCO:

$$\mathrm{COCO}(\mathbf{P}_{x,y}; \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) := \sup_{f \in \mathcal{H}_\mathcal{X},\, g \in \mathcal{H}_\mathcal{Y}} \frac{\mathrm{cov}\left(\mathbf{f}^\top \mathbf{x}, \mathbf{g}^\top \mathbf{y}\right)}{\|\mathbf{f}\|_{\mathcal{H}_\mathcal{X}} \|\mathbf{g}\|_{\mathcal{H}_\mathcal{Y}}}$$

The kernel trick (1)

- Must we really consider infinite dimensional vectors?
- Differentiating COCO wrt $\mathbf{f}$ and $\mathbf{g}$, want biggest eigenvalue

$$\begin{bmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{f}_i \\ \mathbf{g}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \mathbf{f}_i \\ \mathbf{g}_i \end{bmatrix}$$

- When we rely on a finite sample,

$$\widehat{\mathbf{C}}_{xy} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_m \end{bmatrix} \mathbf{H} \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_m^\top \end{bmatrix}$$

The kernel trick (2)

- This means:

$$\mathbf{f} = \sum_{l=1}^{m} c_l \mathbf{x}_l,$$
$$\mathbf{g} = \sum_{l=1}^{m} d_l \mathbf{y}_l.$$

- Inner product in reproducing kernel Hilbert spaces given by kernel

$$\mathbf{x}_1^\top \mathbf{x}_2 = k\left(x_1 - x_2\right)$$
$$\mathbf{y}_1^\top \mathbf{y}_2 = k\left(y_1 - y_2\right)$$

An empirical estimate

- Kernel covariance then **largest eigenvalue** γ_i **of**

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}^{(x)}_{mm}\widetilde{\mathbf{K}}^{(y)}_{mm} \\ \widetilde{\mathbf{K}}^{(y)}_{mm}\widetilde{\mathbf{K}}^{(x)}_{mm} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \widetilde{\mathbf{K}}^{(x)}_{mm} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{K}}^{(y)}_{mm} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}.$$

- $\widetilde{\mathbf{K}}^{(x)}_{mm}$ is matrix of inner products between centred observations in feature space:

$$\widetilde{\mathbf{K}}^{(x)}_{mm} = \mathbf{H}\mathbf{K}^{(x)}_{mm}\mathbf{H}$$

  where

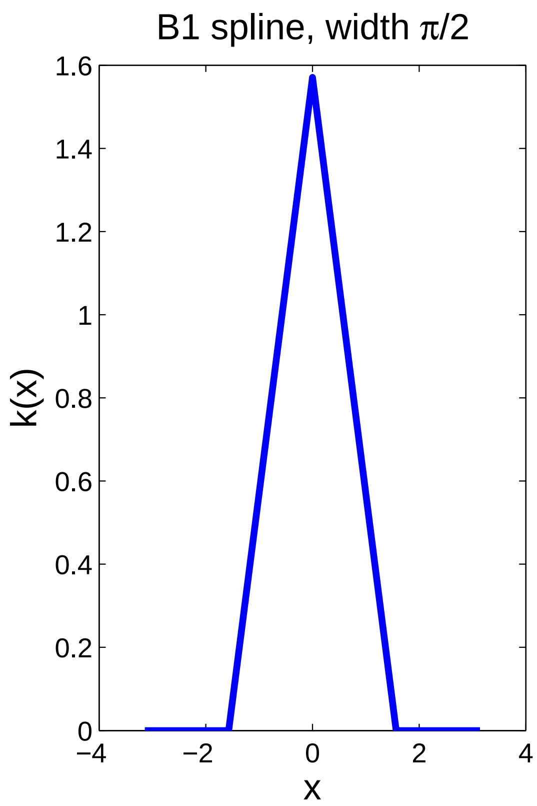
$$\mathbf{H} = \mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$$

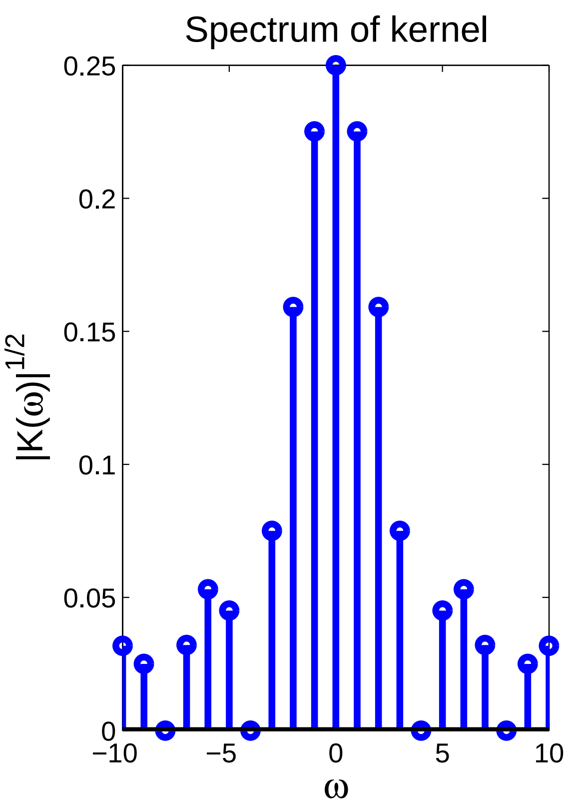
COCO measures independence

- $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}) = 0$ iff x, y independent, when $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are RKHSs induced by universal kernels (eg. Gaussian kernels, Laplace kernels, ...)
- Also true of
  - Kernel canonical correlation: as above, but normalising by the variance in the RKHS [1]
  - Kernel mutual information: an upper bound on the MI near independence [6]
  - Kernel generalised variance: a looser upper bound on the MI near independence [1]

Why universal?

- What happens when kernel is not universal?
- Example: spline kernel





Background: statistical tests (1)

- Probability measure $\mathbf{P}_z$ in $\mathcal{P}_0$ or $\overline{\mathcal{P}_0}$
- Two hypotheses:
  - H_0: null hypothesis ($\mathbf{P}_z \in \mathcal{P}_0$)
  - H_1: alternative hypothesis
- Observe a sample $\boldsymbol{z}$
- If sample is in
  - Rejection/critical region R: reject H_0
  - Acceptance region: accept H_0
- Region defined using test statistic $\Delta(\boldsymbol{z})$
  - Example: sample mean (is mean greater than some threshold?)

Background: statistical tests (2)

- How good is a test?
  - Type I error: We reject H_0 although it is true
  - Type II error: We accept H_0 although it is false
- Power of test:

$$\beta(\mathbf{P_z}) := \mathbf{P_z}(\boldsymbol{z} \in R)$$

  - Should be ~ 0 for $\mathbf{P_z} \in \mathcal{P}_0$, ~ 1 for $\mathbf{P_z} \in \overline{\mathcal{P}_0}$
- Level of test: for $0 \leq \alpha \leq 1$

$$\alpha \geq \sup_{\mathbf{P_z} \in \mathcal{P}_0} \beta(\mathbf{P_z})$$

  - Upper bound on worst possible type I error
  - Note: size of test is true worst type I error

When is dependence hard to detect?

- NO test can detect all dependence for finite samples.
- Example: Set $\mathcal{P}$ of prob. distrib. $\mathbf{P_x}$ over n variables
  - $\mathcal{P}_i$ generates independent random variables,
  - $\mathcal{P}_d$ gives dependent RVs
- Test: $\Delta(\boldsymbol{x})$ takes m i.i.d. samples, returns

$$\Delta(\boldsymbol{x}) = 1 \;:\; \boldsymbol{x} \sim \mathbf{P}^{(d)}_{\mathbf{x}^m}, \qquad \Delta(\boldsymbol{x}) = 0 \;:\; \boldsymbol{x} \sim \mathbf{P}^{(i)}_{\mathbf{x}^m}$$

- Uncertainty due to empirical estimate: α-test

$$\sup_{\mathbf{P}^{(i)}_{\mathbf{x}} \in \mathcal{P}_i} \mathbf{E}_{\boldsymbol{x} \sim \mathbf{P}^{(i)}_{\mathbf{x}^m}} (\Delta(\boldsymbol{x}) = 1) \leq \alpha$$

- There exists $\mathbf{P_x} \notin \mathcal{P}_i$ such that for small ϵ,

$$\mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathbf{x}^m}} (\Delta(\boldsymbol{x}) = 0) \geq 1 - \alpha - \epsilon$$

Hard-to-detect dependence (2)

- COCO can be ≈ 0 for dependent RVs with highly non-smooth densities:

$$\mathsf{P}_{\mathsf{x},\mathsf{y}} = \alpha + \beta \varphi_l(\mathsf{x}) \varphi_l(\mathsf{y}),$$

  - l large
  - β non-trivial

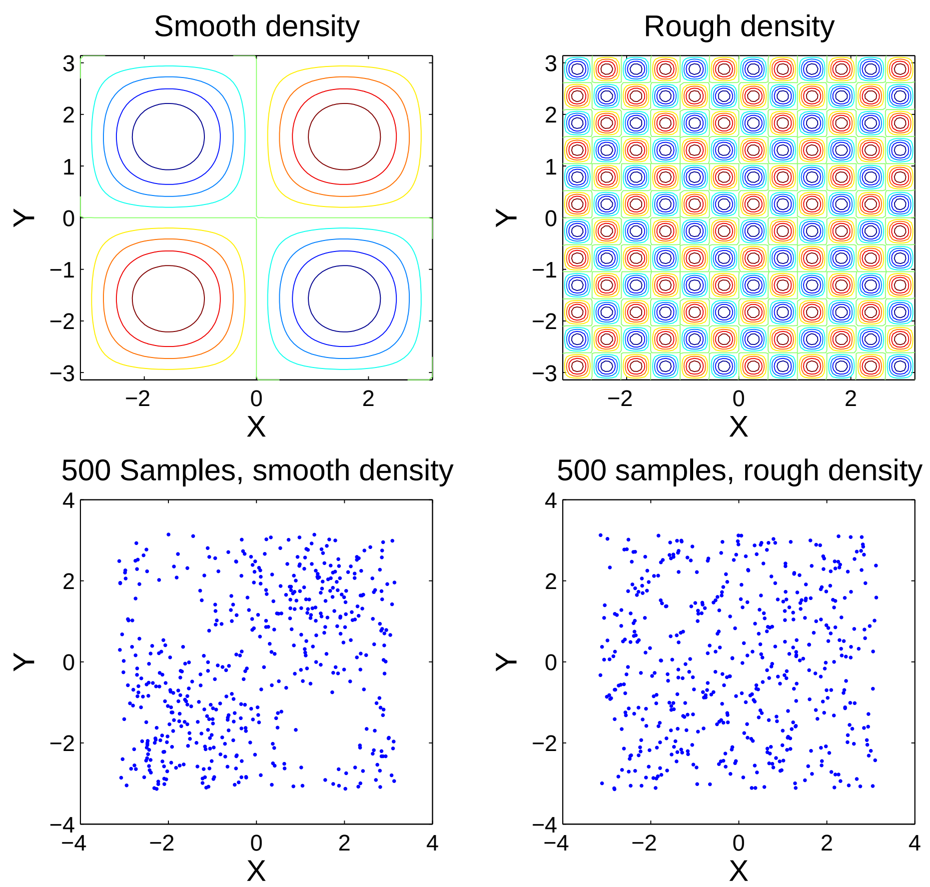
- COCO "as small as you want" (depends on l)

- Reason: norms in the denominator

$$\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}) := \sup_{f \in \mathcal{H}_{\mathcal{X}},\, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\mathrm{cov}\left(\mathbf{f}^\top \mathbf{x}, \mathbf{g}^\top \mathbf{y}\right)}{\|\mathbf{f}\|_{\mathcal{H}_{\mathcal{X}}} \|\mathbf{g}\|_{\mathcal{H}_{\mathcal{Y}}}}$$

- RESULT: not detectable with finite sample size

Hard-to-detect dependence (3)

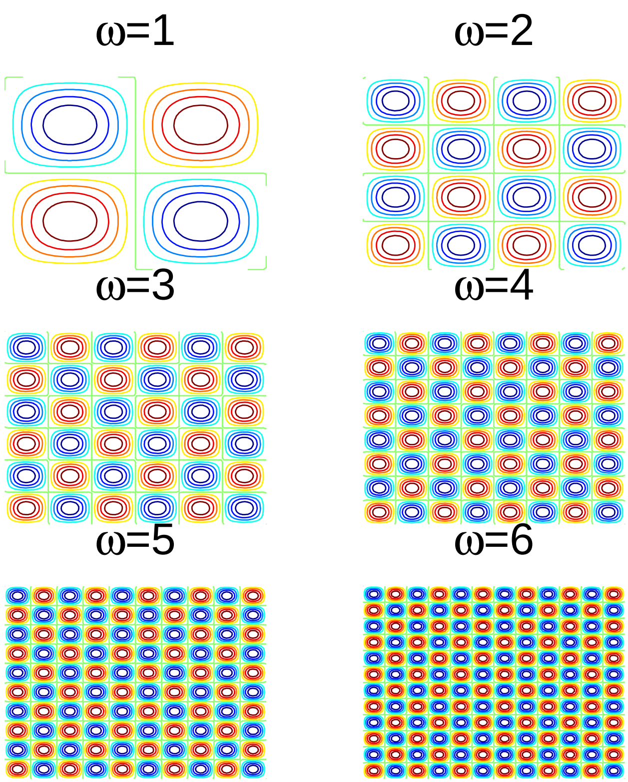


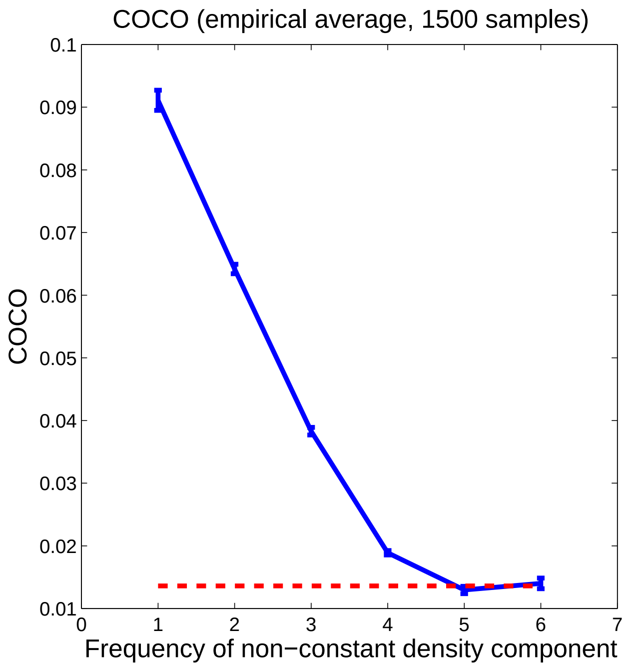
Density takes the form:

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

Hard-to-detect dependence (4)

- Example: sinusoids of increasing frequency



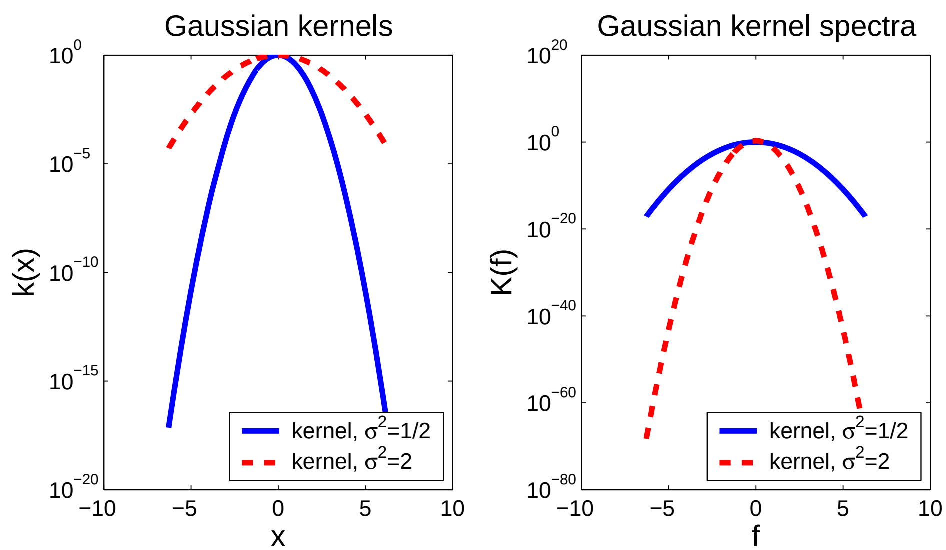


A test of independence

- Empirical COCO converges to the population COCO at speed $1/\sqrt{n}$.
- A dependence test: $\Delta(\boldsymbol{z})$ is the indicator that COCO larger than $C\sqrt{\log(1/\alpha)/n}$
- $\Delta(\boldsymbol{z})$ is an α-test
  - Reminder: α upper bounds prob. that test returns dependence when random variables independent
- Type II approaches zero as $1/\sqrt{n}$.
  - Reminder: Type II error is prob. that test returns independence when random variables dependent
- No slow learning rates for dependence tests!
- Finite sample results!

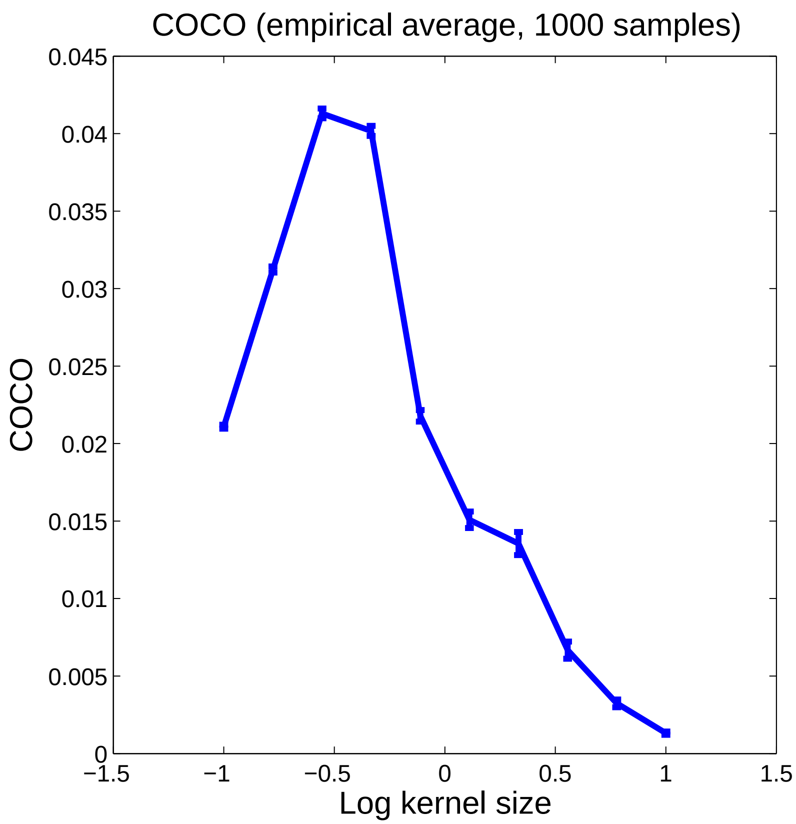
Choosing kernel size (1)

- Reminder: the RKHS norm of a function is $\|f\|^2_{\mathcal{H}_\mathcal{X}} := \sum_{i=1}^{\infty} \tilde{f}_i^2 \left(\tilde{k}_i\right)^{-1}$.
- If kernel decays quickly, its spectrum decays slowly:
  - then non-smooth functions have smaller RKHS norm
- Example: spectrum of two Gaussian kernels



Choosing kernel size (2)

- Could we just decrease kernel size?
- Yes, but only up to a point

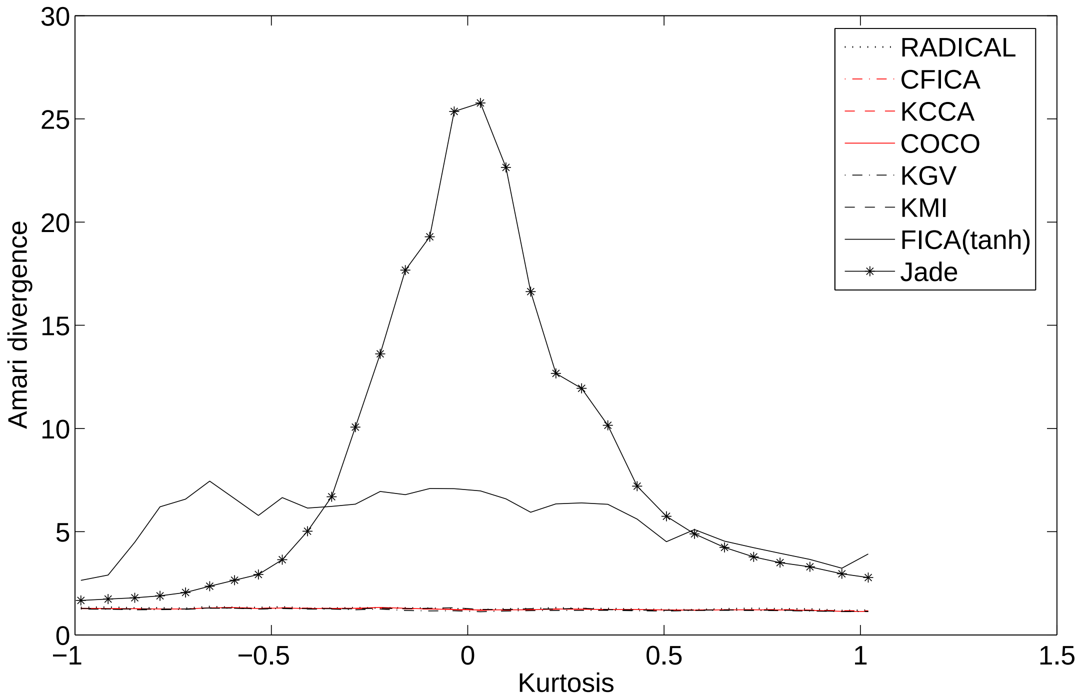


Application to ICA

- ICA can be done by optimising over kernel dependence measures (contrast function)
- State-of the art performance for small to medium scale problems
- Still too slow for large-scale ($\gtrapprox$ 16 sources) problems
- Better outlier resistance than alternatives
- Source kurtosis does not affect performance

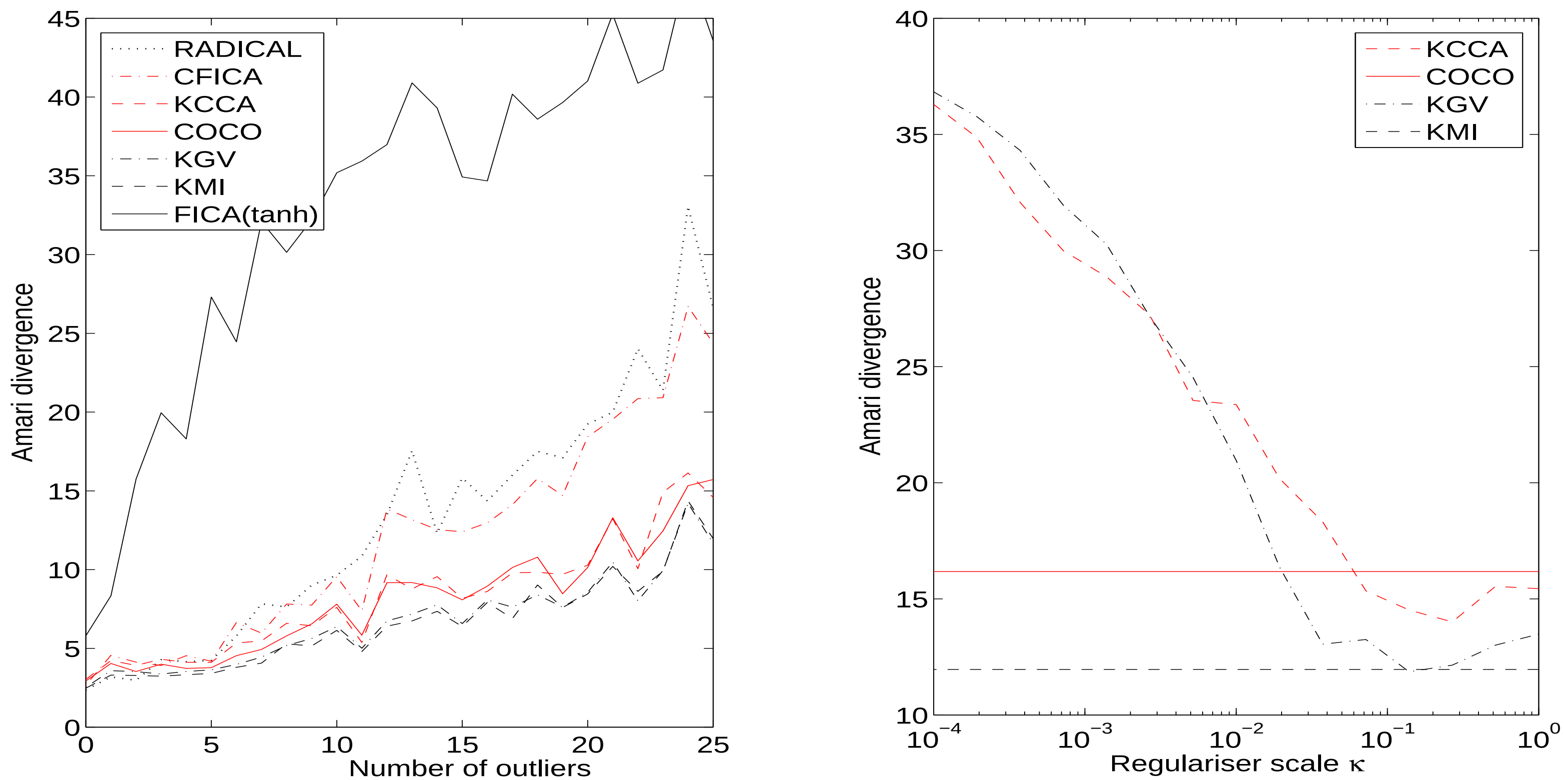
Positive, Negative, and Zero kurtosis

- Amari divergence mesaures distance between estimated and true mixing matrix
- Invariant to source order swapping and source scaling
- Bigger $\longrightarrow$ worse performance



Outlier resistance

- Outlier noise added to the mixed sources



The Two-Sample Problem

The two-sample problem

- Test if same distribution generated two samples
- Our criterion: the maximum mean discrepancy
  - Given a type I error, type II error converges fast $(1/\sqrt{n})$
  - No assumptions about generating distributions
- Applications
  - Neuroscience: test whether spikes on different days are from the same neuron
  - Speaker identification
  - Comparison of paintings using hyperspectral photography
  - Merging databases

The MMD (1)

- $\mathcal{F}$ a universal RKHS, $F := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq 1\}$ the unit ball in $\mathcal{F}$.
- The population MMD is defined as

$$MMD(\mathbf{P}_{\mathsf{x}}, \mathbf{P}_{\mathsf{y}}; F) := \left(\sup_{f \in F} \left[\mathbf{E}_{\mathsf{x}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{y}} f(\mathsf{y}) \right] \right)^2 .$$

- $MMD(\mathbf{P}_{\mathsf{x}}, \mathbf{P}_{\mathsf{y}}; F) = 0$ if and only if $\mathbf{P}_{\mathsf{x}} = \mathbf{P}_{\mathsf{y}}$, for universal kernels

The MMD (2)

- How to get it wrt kernels
  - Mean elements corresponding to $\phi(\mathsf{x})$ and $\phi(\mathsf{y})$:

$$\begin{aligned}\langle \mu_x, f\rangle_{\mathcal{F}} &:= \mathbf{E}_{\mathsf{x}}\left[\langle \phi(\mathsf{x}), f\rangle_{\mathcal{F}}\right] &= \mathbf{E}_{\mathsf{x}}(f(\mathsf{x})),\\ \langle \mu_y, f\rangle_{\mathcal{F}} &:= \mathbf{E}_{\mathsf{y}}\left[\langle \phi(\mathsf{y}), f\rangle_{\mathcal{F}}\right] &= \mathbf{E}_{\mathsf{y}}(f(\mathsf{y})).\end{aligned}$$

  - The norm is also written as

$$\|\mu\|_{\mathcal{F}} := \sup_{f\in F}\langle f, \mu\rangle_{\mathcal{F}}$$

- The MMD in terms of kernels:

$$\begin{aligned} MMD(\mathbf{P}_{\mathsf{x}}, \mathbf{P}_{\mathsf{y}}; F) &= \left(\sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2 \\ &= \| \mu_x - \mu_y \|_{\mathcal{F}}^2 \\ &= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathsf{y},\mathsf{y}'} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathsf{x},\mathsf{y}} k(\mathsf{x}, \mathsf{y}), \end{aligned}$$

- x' is a R.V. independent of x with distribution $\mathbf{P}_{\mathsf{x}}$
- y' is a R.V. independent of y with distribution $\mathbf{P}_{\mathsf{y}}$.

Empirical estimate

- Given data $\boldsymbol{x}$ of size m drawn from P_x and $\boldsymbol{y}$ of size n drawn from P_y
- An unbiased empirical estimate (quadratic cost):

$$
\begin{aligned}
KMD(\boldsymbol{x}, \boldsymbol{y}; \mathcal{F}) \quad := \quad & \underbrace{\frac{1}{m(m-1)} \sum_{i \neq j} k(x_{i_1}, x_{i_2})}_{\text{(a)}} \\
& + \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j} k(y_{j_1}, y_{j_2})}_{\text{(b)}} \\
& - \underbrace{\frac{2}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j)}_{\text{(c)}}.
\end{aligned}
$$

How fast does empirical converge to population?

- For testing purposes, need only positive deviation
- Use 1- and 2-sample U-statistic bounds from Hoeffding
- Assume $0 \leq k(x, y) \leq R$ almost everywhere, $m \leq n$.
- For all $n > 2$ and all $0 < \delta < 1$, with probability at least $1 - \delta$, for all $\mathbf{P}_x$ and $\mathbf{P}_y$,

$$KMD(\boldsymbol{x}, \boldsymbol{y}; \mathcal{F}) - KMD(\mathbf{P}_x, \mathbf{P}_y; \mathcal{F}) \leq \frac{R}{\beta}\sqrt{\frac{\log(3/\delta)}{n}},$$

  - Here $\beta = \frac{1+(1-\sqrt{2})r}{1+r(2-r)}$
  - $r = \sqrt{n/m}$.

A 2-sample test based on MMD

- Test statistic is $KMD(\boldsymbol{x}, \boldsymbol{y}; F)$
- Null hypothesis H_0 is $\mathsf{P}_\mathsf{x} = \mathsf{P}_\mathsf{y}$
- The test: accept H_0 if

$$KMD(\boldsymbol{x}, \boldsymbol{y}; F) \leq \frac{R}{\beta}\sqrt{\frac{\log(3/\alpha)}{n}}$$

- gives a test of level α
- Type 2 error asymptotically drops as $1/\sqrt{n}$
- What is p-value? We get an upper bound using

$$p \leq 3\exp\left(\frac{-KMD^2(\boldsymbol{x}, \boldsymbol{y}; F)\beta^2 n}{R^2}\right).$$

Further reading

Some references on ICA and independence measurement

- Start with Cardoso's excellent introduction [3], and the tutorial by Hyvärninen [7]
- For kernel methods, look at [6] (this talk), [1], and [5] (final paper deals with *conditional* independence)
- Some alternative recent methods with "adaptive" contrast functions: [10, 8]
- Classic algorithms for time series separation with second order methods: [9, 2]
- An important paper for optimising over rotation matrices: [4]

References

[1] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[3] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90(8):2009–2026, 1998.

[4] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[5] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[6] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schoelkopf. Kernel methods for measuring independence. *JMLR*, 6:2075–2129, 2005.

[7] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[8] E. Learned-Miller and J. Fisher III. Independent components analysis by direct entropy minimization. Technical Report UCB/CSD-3-1221, Computer Science Division, University of California Berkeley, 2003.

[9] L. Molgedey and H. Schuster. Separation of a mixture of independent signals using time delayed correlation. *Physical Review Letters*, 72(23):3634–3637, 1994.

[10] D.-T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 2002. Submitted.