Independent Component Analysis

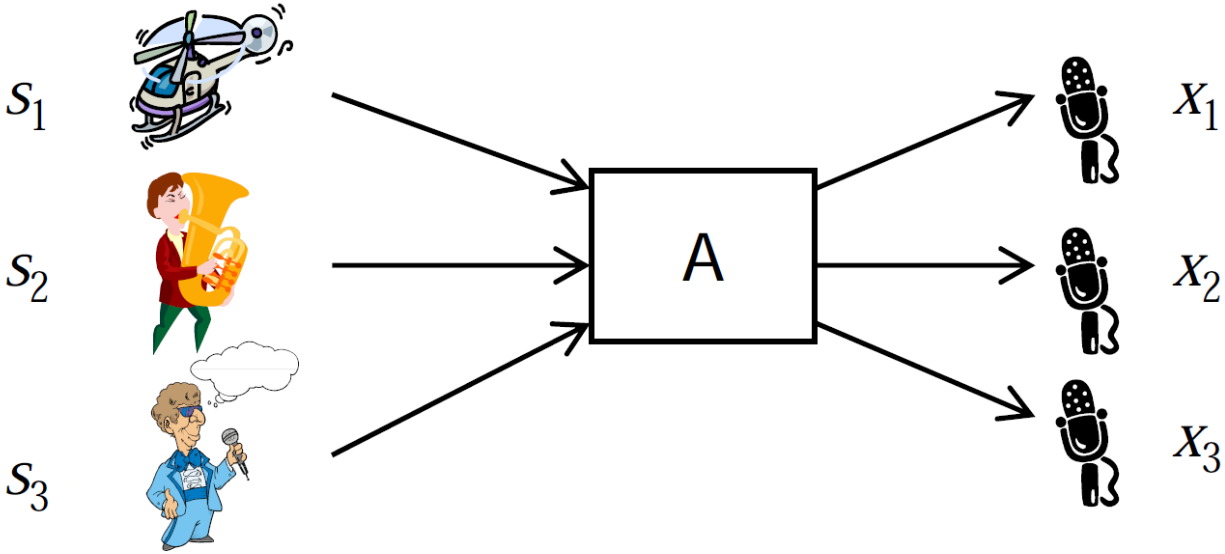
Arthur Gretton

Carnegie Mellon University

2009

ICA: setting

Independent component analysis:

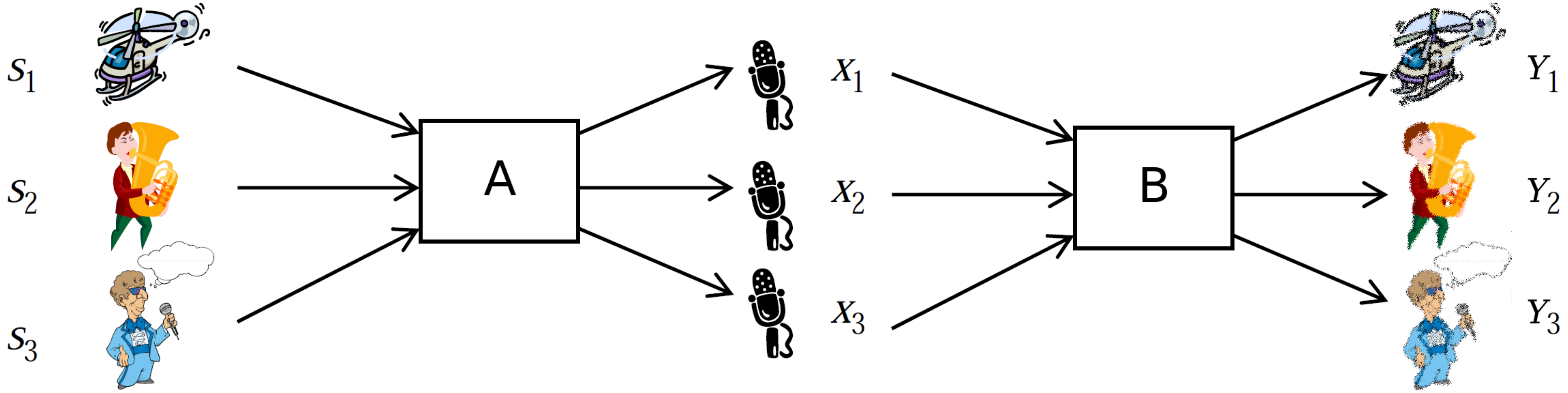


$$S \quad \times A \quad = X$$

- $\boldsymbol{S}$ a vector of l unknown, independent sources: $\mathsf{P}_{\boldsymbol{S}} = \prod_{i=1}^{l} \mathsf{P}_{S_i}$
- $\boldsymbol{X}$ vector of mixtures
- $\mathbf{A}$ is $l \times l$ mixing matrix (full rank)

ICA: setting

Independent component analysis:

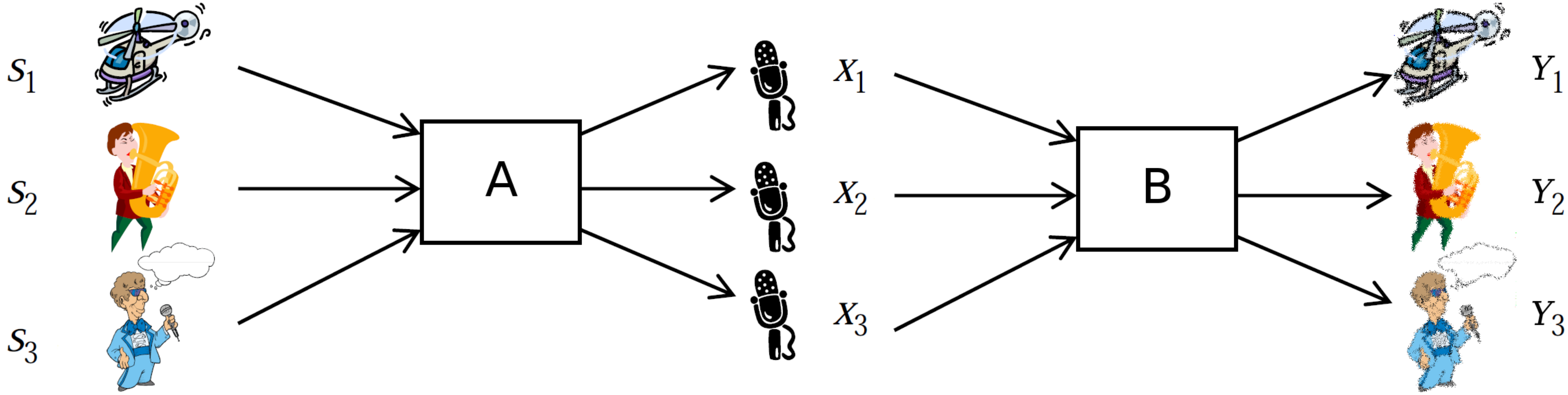


$$X \quad \times B \quad = Y$$

- $\mathbf{B}$ is estimated $\mathbf{A}^{-1}$, we solve for this
- $\boldsymbol{Y}$ vector of estimated sources

ICA: setting

Independent component analysis:



$$X \quad \times B \quad = Y$$

- $\mathbf{B}$ is estimated $\mathbf{A}^{-1}$, we solve for this
- $\boldsymbol{Y}$ vector of estimated sources

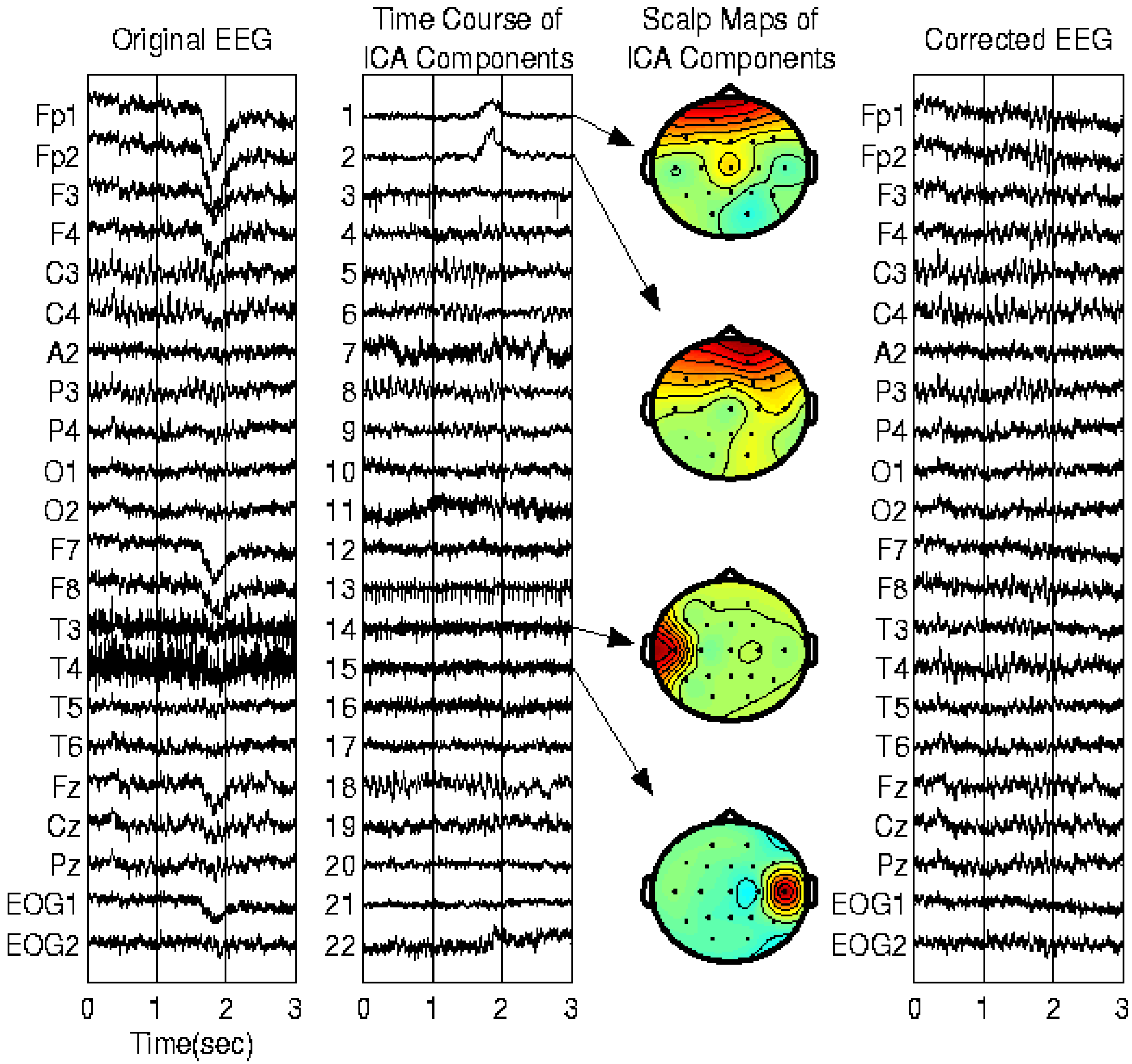
Neglect time dependence: m i.i.d. mixture observations

ICA: another example

- Mixtures X are original EEG

[Jung et al., 2000]

- Estimated sources Y are ICA components
- Scalp map from B

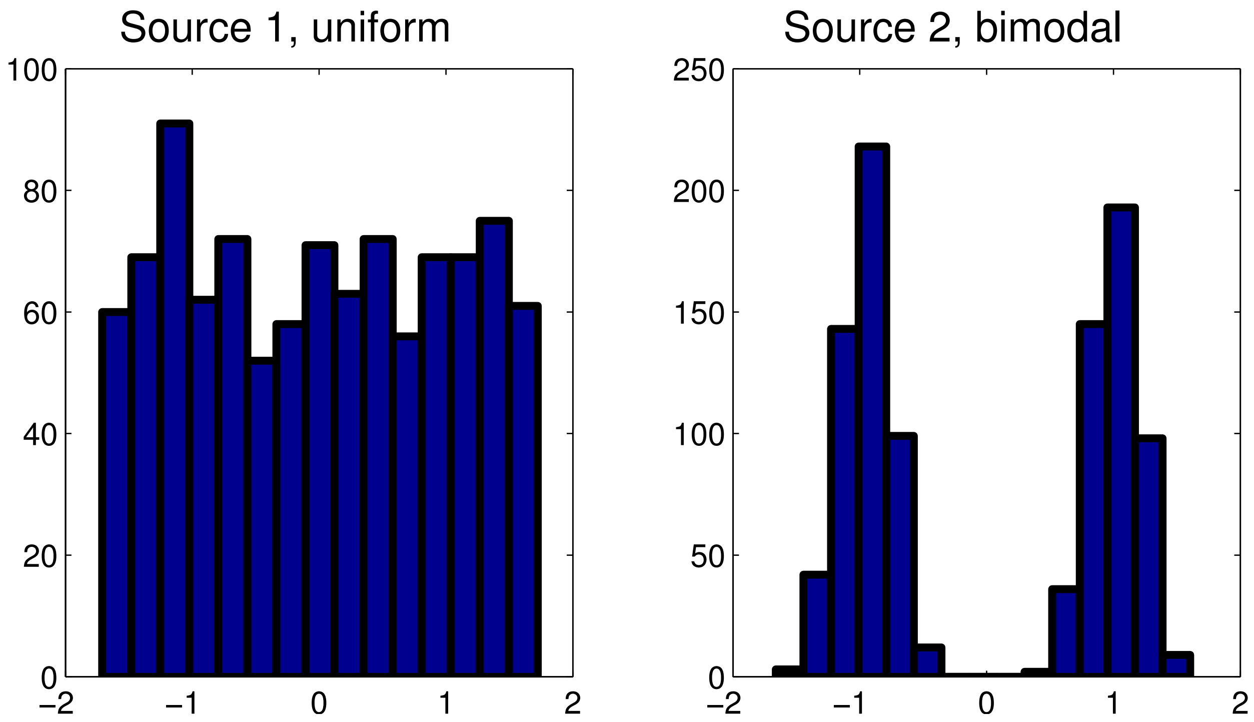


ICA examples

- We've seen:
  - Sounds mixed together ("cocktail party" problem) [Hyvärinen et al., 2001]
  - EEG recordings (brain, fetal heartbeat) [Jung et al., 2000, Stögbauer et al., 2004]
- Some further examples:
  - Extracting independent activity from fMRI [Calhoun et al., 2003]
  - Financial data [Kiviluoto and Oja, 1998]
  - Linear edge filters for image patch coding? (Possibly not: [Bethge, 2006])

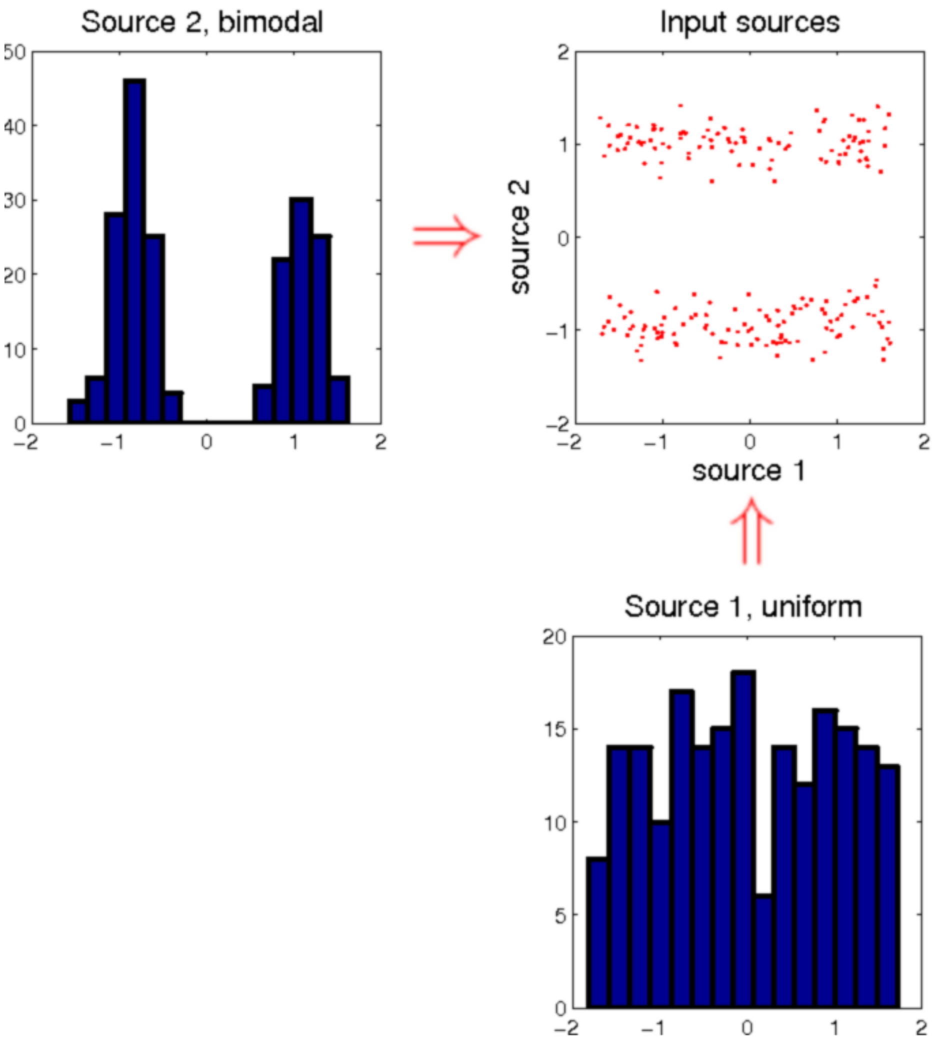
A toy example

- Two distributions: $\mathbf{P}_{\boldsymbol{S}_1}$ is uniform, $\mathbf{P}_{\boldsymbol{S}_2}$ is bimodal



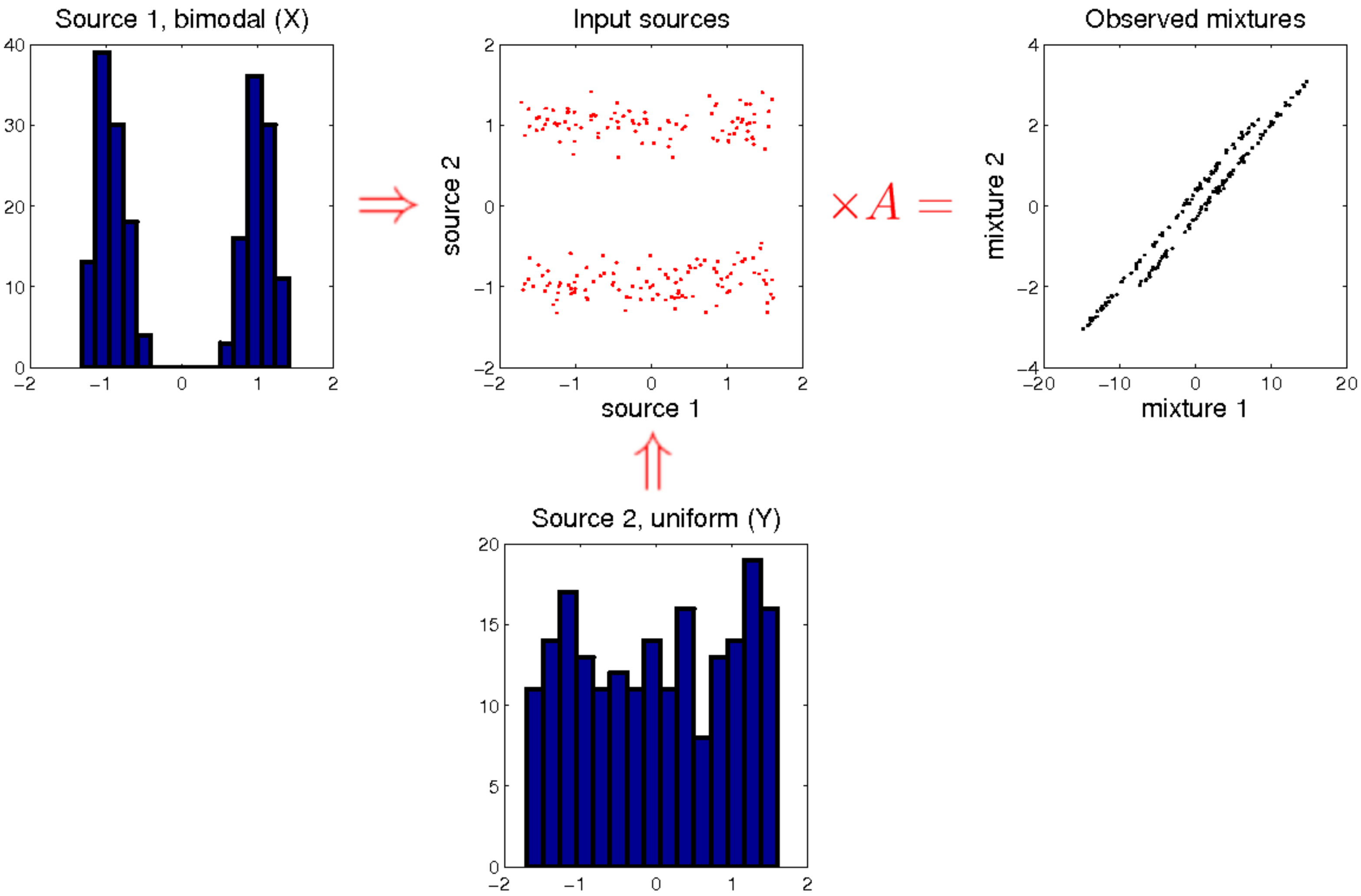
A toy example

- Two distributions: $\mathbf{P}_{\boldsymbol{S}_1}$ is uniform, $\mathbf{P}_{\boldsymbol{S}_2}$ is bimodal



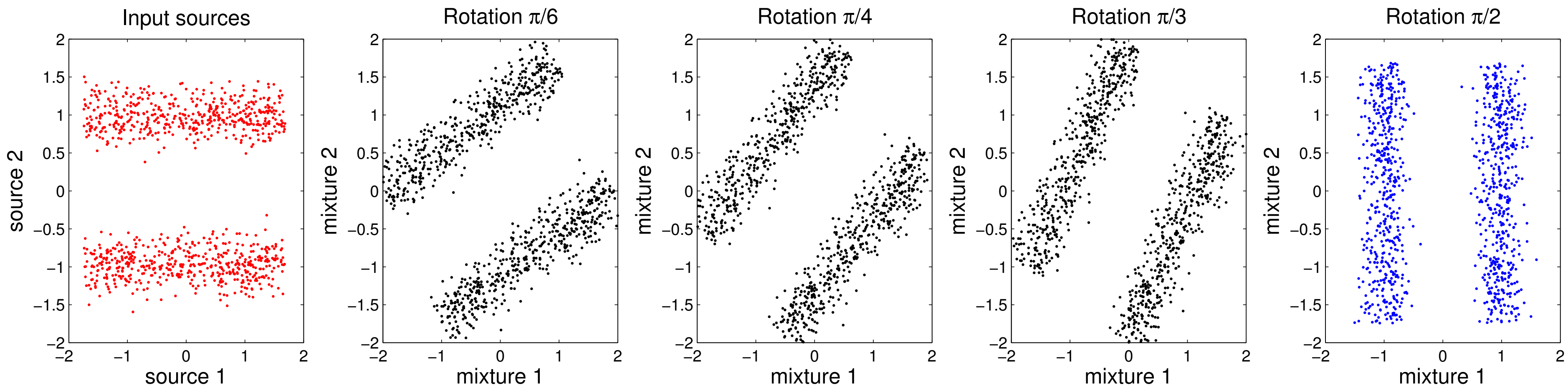
A toy example

- Two distributions: $\mathbf{P}_{S_1}$ is uniform, $\mathbf{P}_{S_2}$ is bimodal



First indeterminacy: ordering

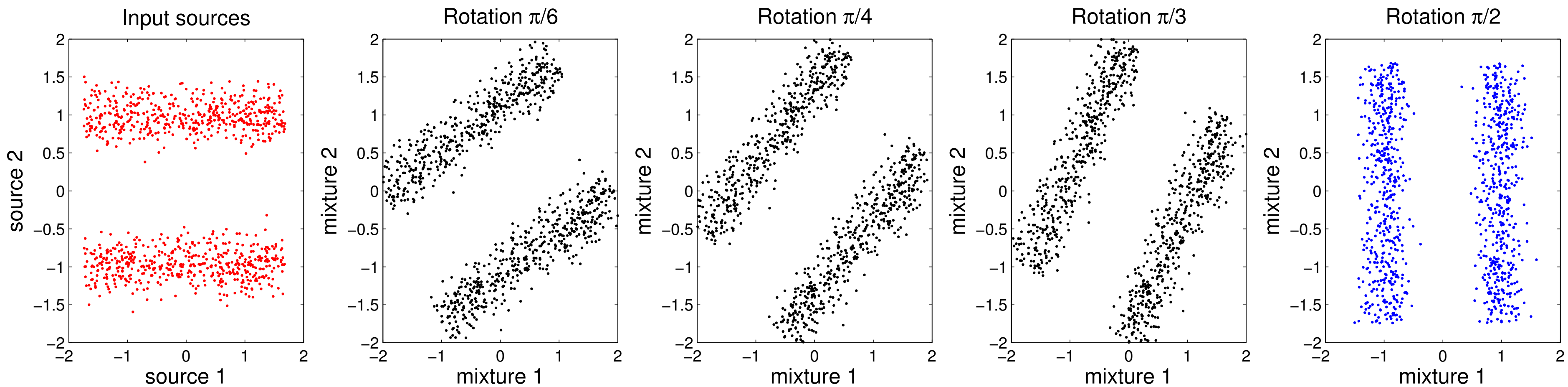
- Initial unmixed RVs in red



- Independent at rotation $\pi/2$

First indeterminacy: ordering

- Initial unmixed RVs in red

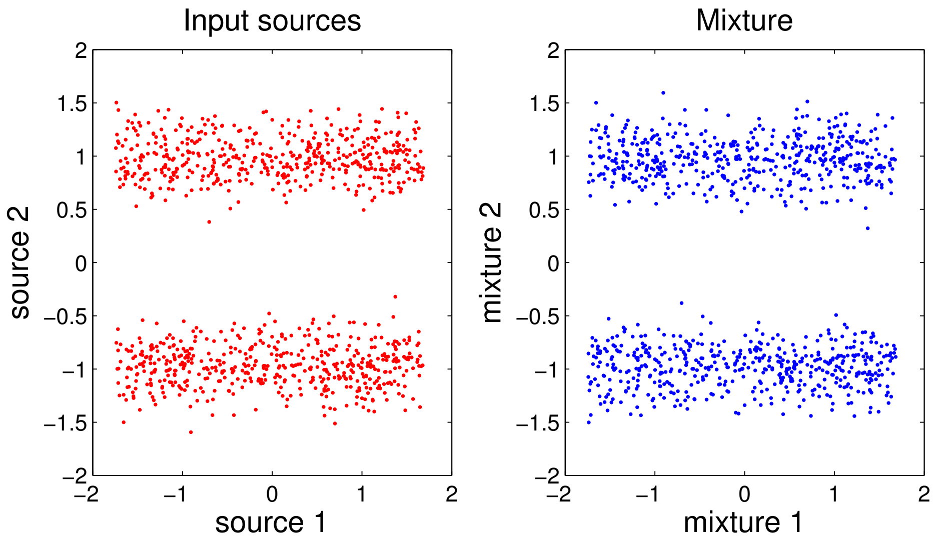


- Independent at rotation $\pi/2$

Ignore source order

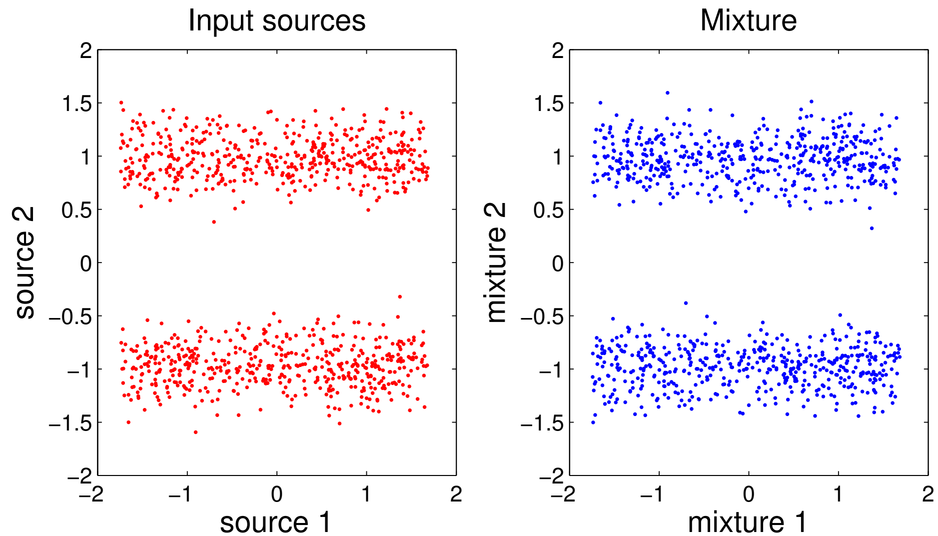
Second indeterminacy: sign

- Initial unmixed RVs in red
- Source 2 sign reversed in blue



Second indeterminacy: sign

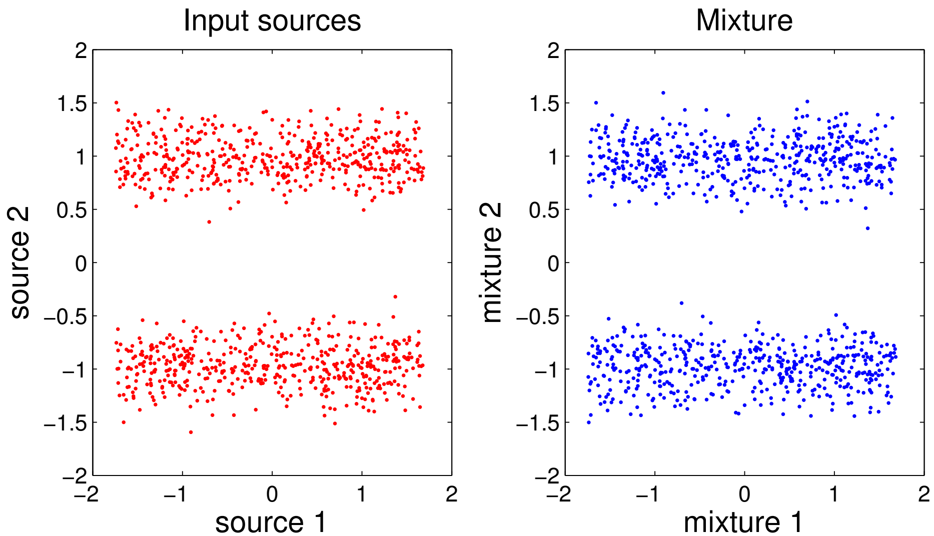
- Initial unmixed RVs in red
- Source 2 sign reversed in blue



Ignore source sign

Second indeterminacy: sign

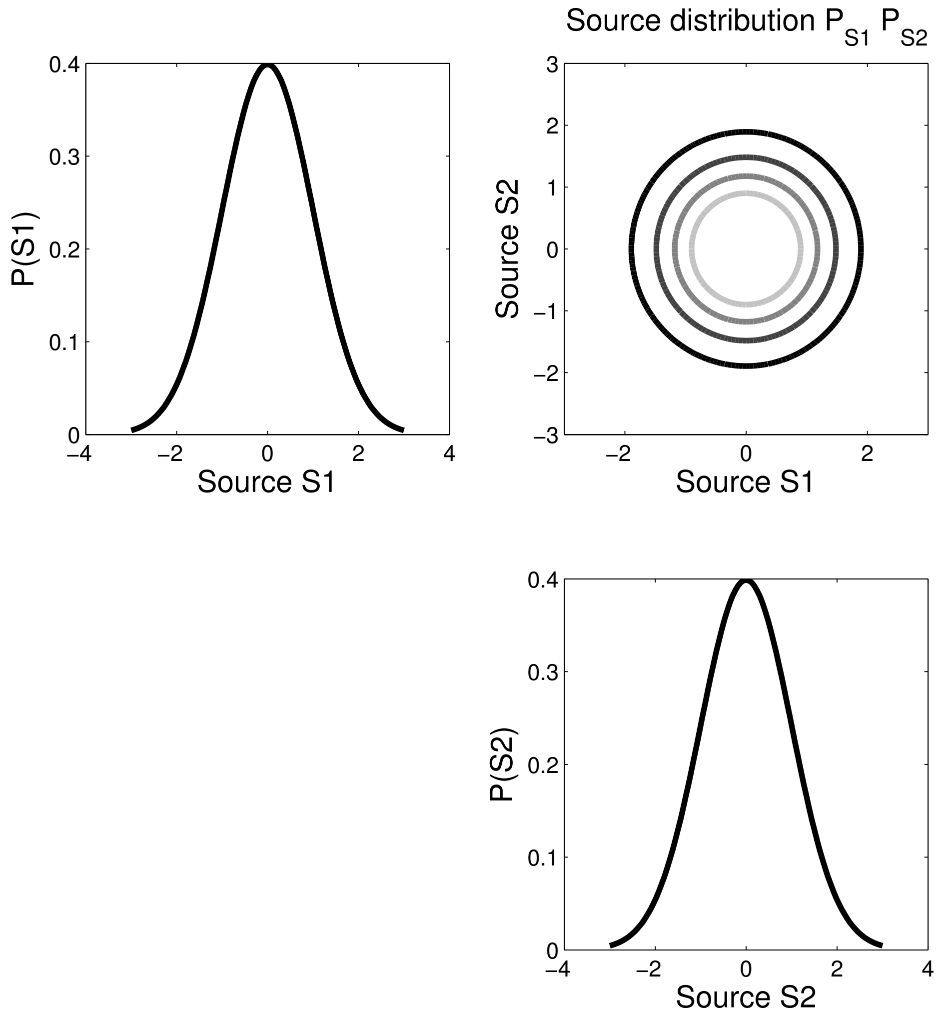
- Initial unmixed RVs in red
- Source 2 sign reversed in blue



- More generally: S_1 and S_2 independent iff aS_1 and S_2 independent for $a \neq 0$
  – Assume sources have unit variance

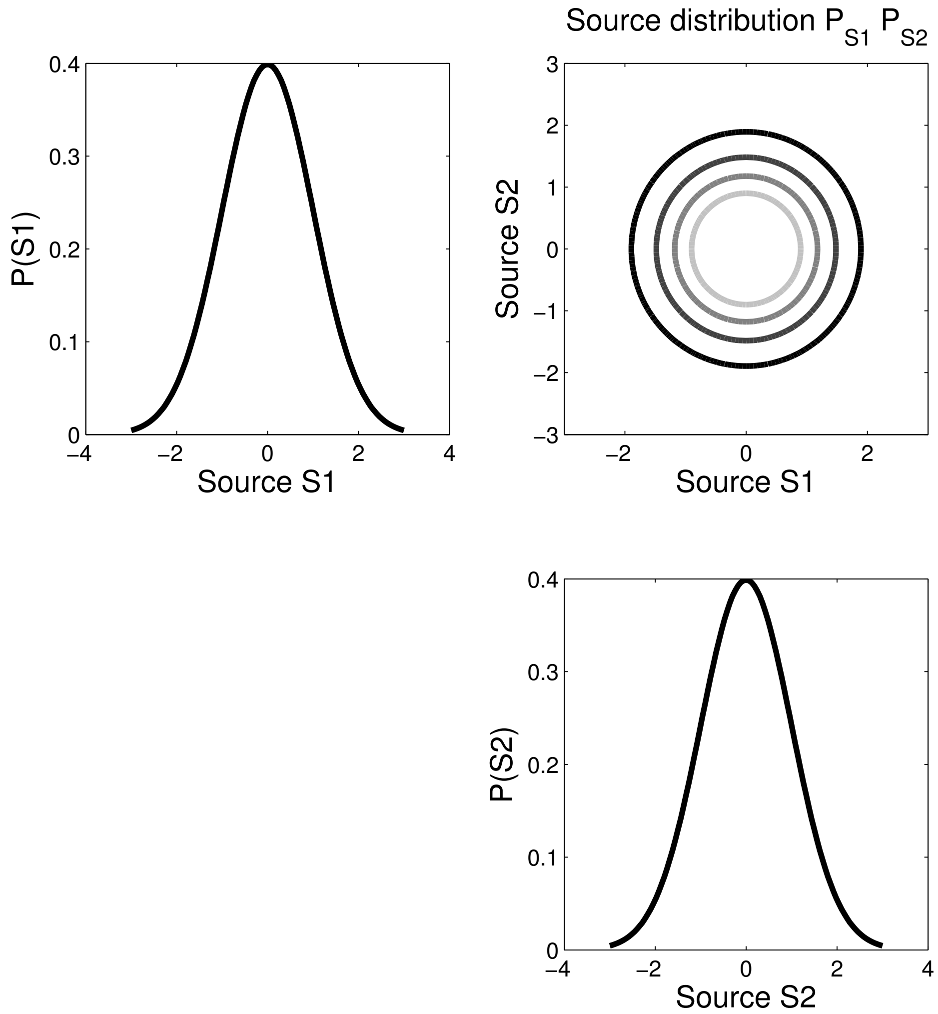
Third indeterminacy: Gaussians

Both sources Gaussian



Third indeterminacy: Gaussians

Both sources Gaussian



Meaningless to “unmix” Gaussians

Things that are impossible for ICA

Using independence alone, we cannot . . .

- recover signal order,
- recover signal sign (or amplitude) ,
- separate multiple Gaussians.

Things that are impossible for ICA

Using independence alone, we cannot . . .

- recover signal order,
- recover signal sign (or amplitude) ,
- separate multiple Gaussians.

We can recover

$$B^* = PDA^{-1}$$

- P is a permutation matrix
- D diagonal, $d_{ii} \in \{-1, 1\}$

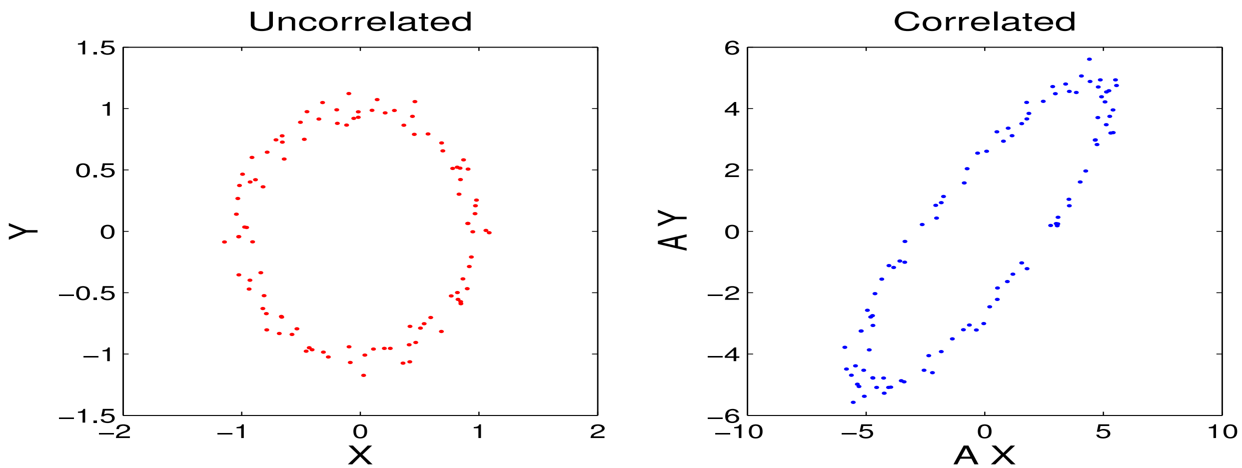
(as long as no more than one Gaussian source)

First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between mixtures $\boldsymbol{X}$

First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between mixtures $\boldsymbol{X}$



First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between mixtures $\boldsymbol{X}$
- New signals have unit covariance:

$$\boldsymbol{T} = \mathbf{B}_w \boldsymbol{X} \qquad \mathbf{C}_t = \mathbf{I}$$

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

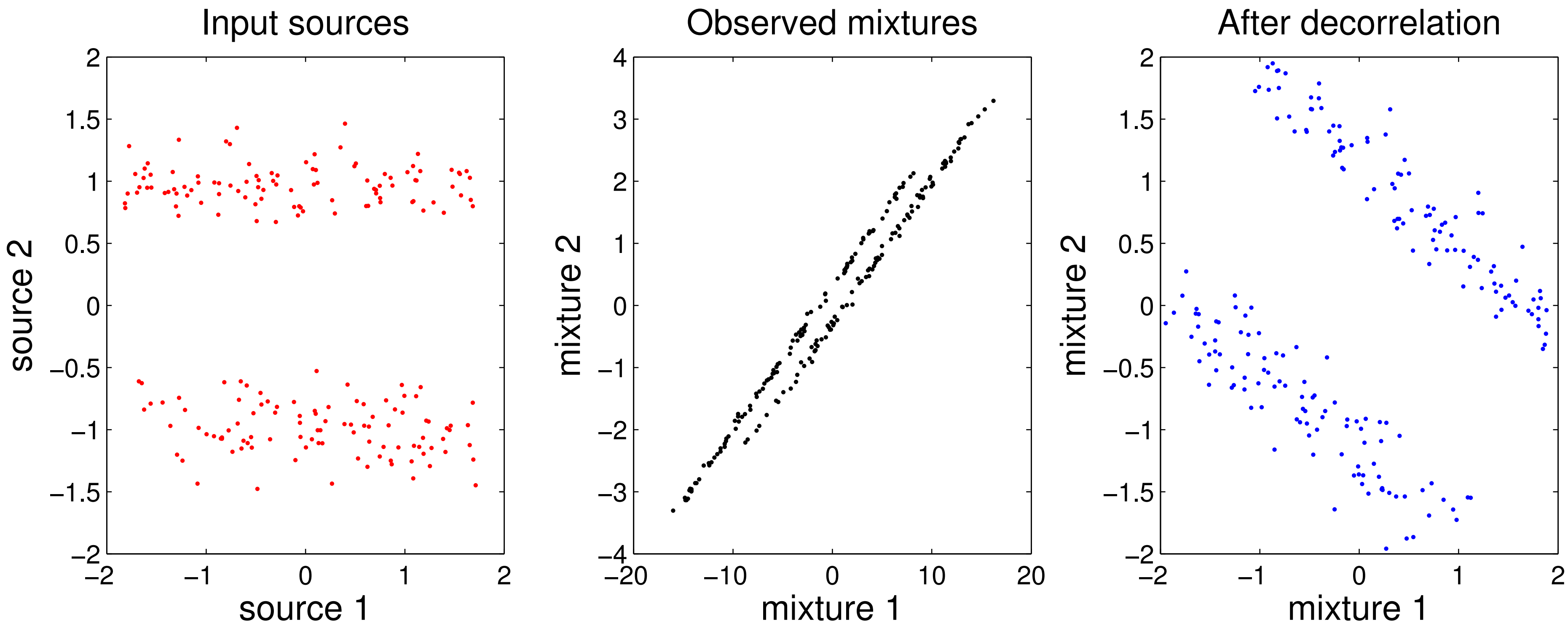
$$\mathbf{B} = \mathbf{B}_r \mathbf{B}_w$$

  - $\mathbf{B}_w$ is a whitening matrix
  - $\mathbf{B}_r$ is remaining demixing operation
- Use the SVD of mixture covariance $\mathbf{C}_x = \mathbf{U}\Lambda\mathbf{U}^\top$:

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

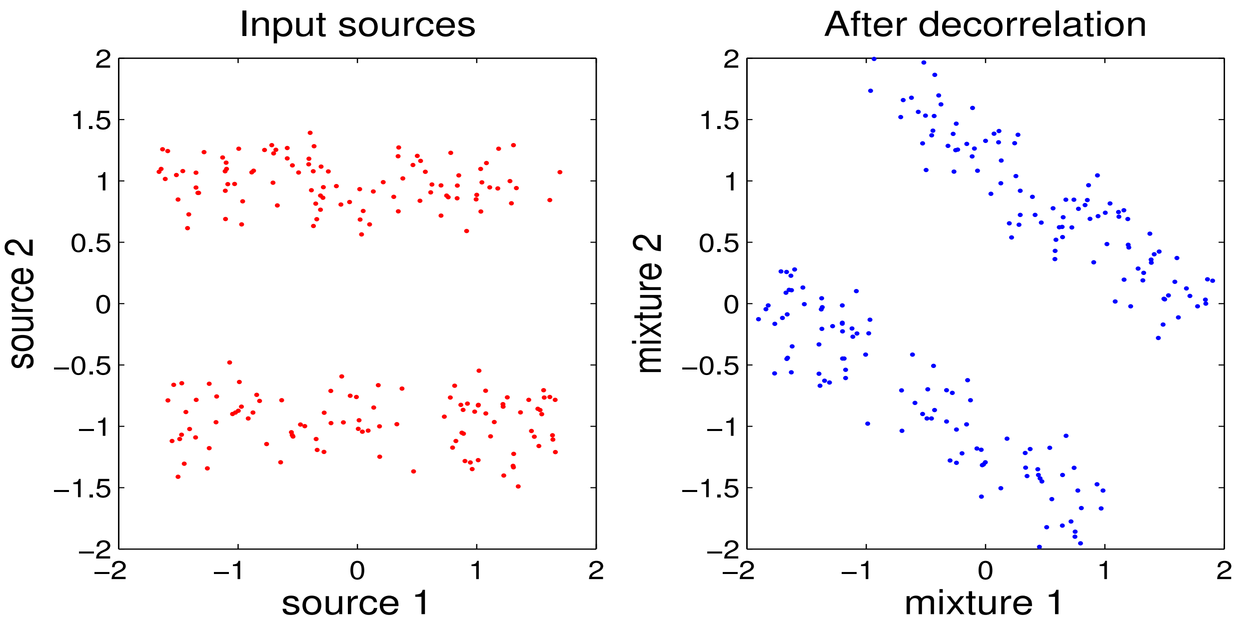
What does decorrelation achieve?

- Two distributions: $\mathbf{P}_{\boldsymbol{S}_1}$ is uniform, $\mathbf{P}_{\boldsymbol{S}_2}$ is bimodal



Problem remaining: *rotation*

- Assume correlation has already been removed
- To recover original signal, need to rotate

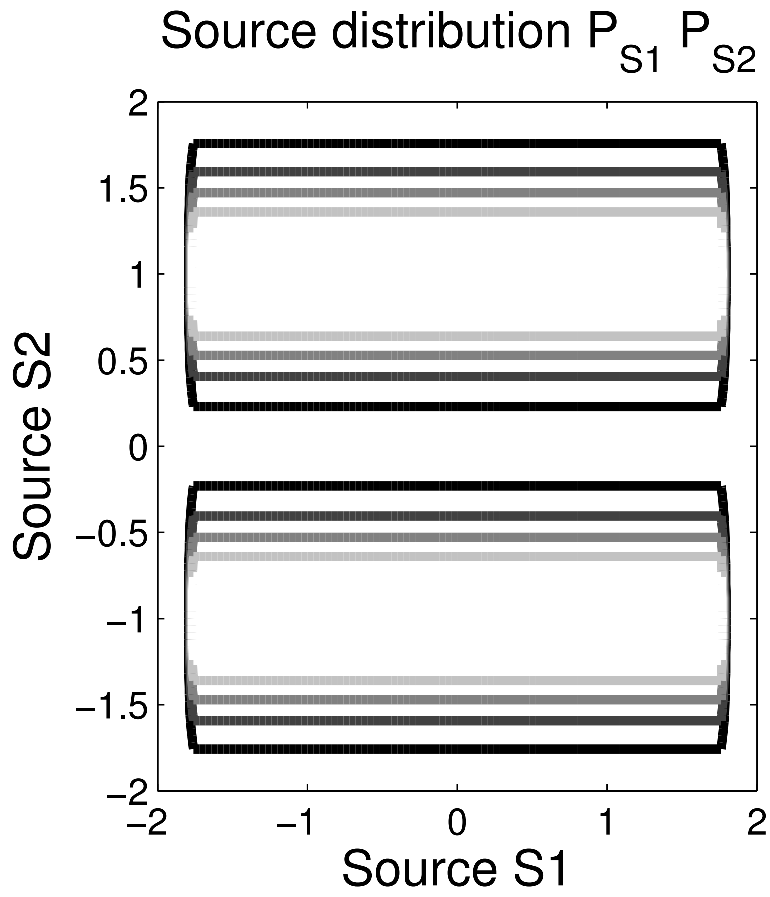


- In remainder: unmixing matrix $\mathbf{B}$ is rotation,

$$\mathbf{B}^\top \mathbf{B} = \mathbf{I}$$

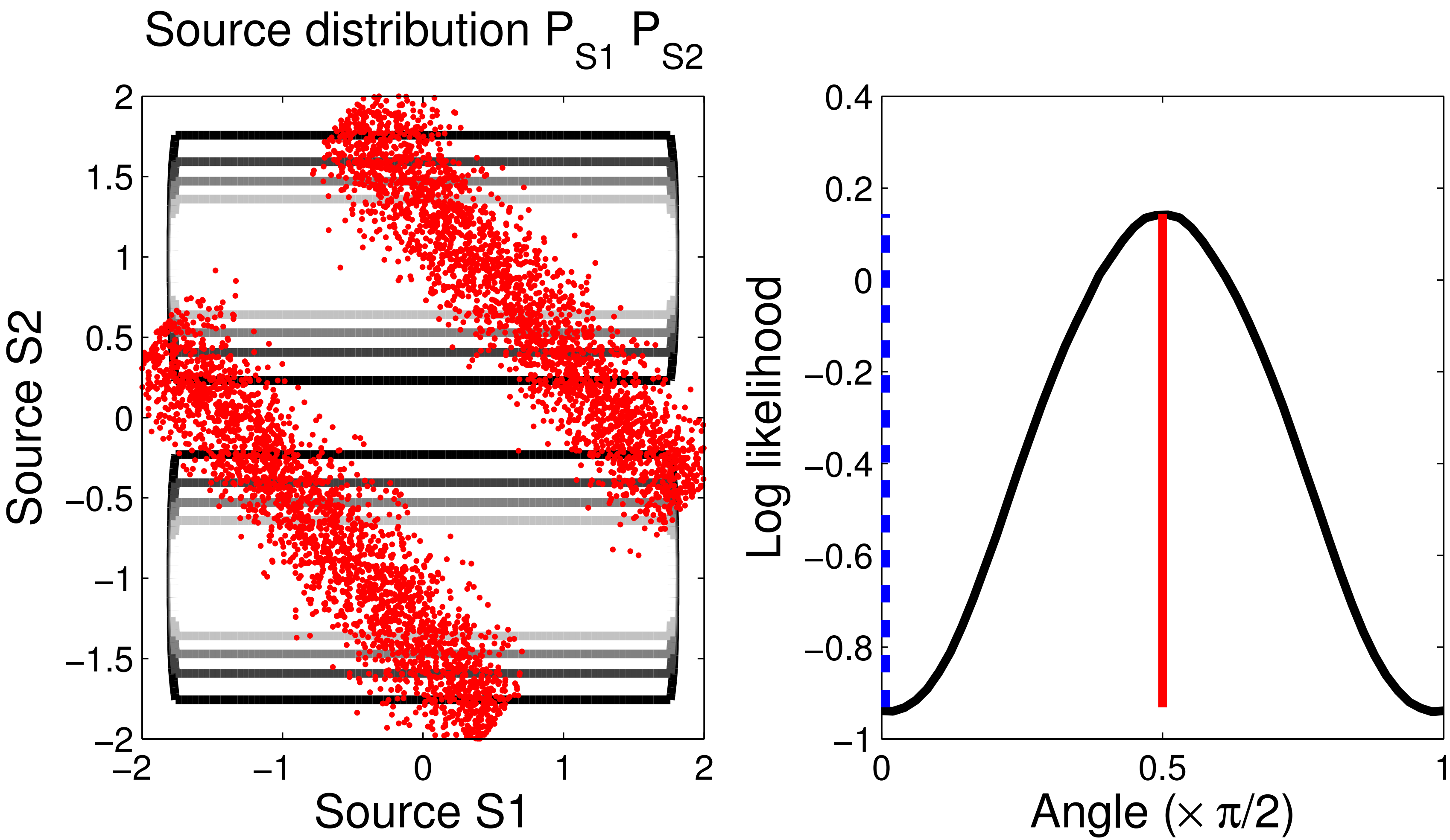
ICA: maximum likelihood

- Model for mixtures parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\boldsymbol{S}})$



ICA: maximum likelihood

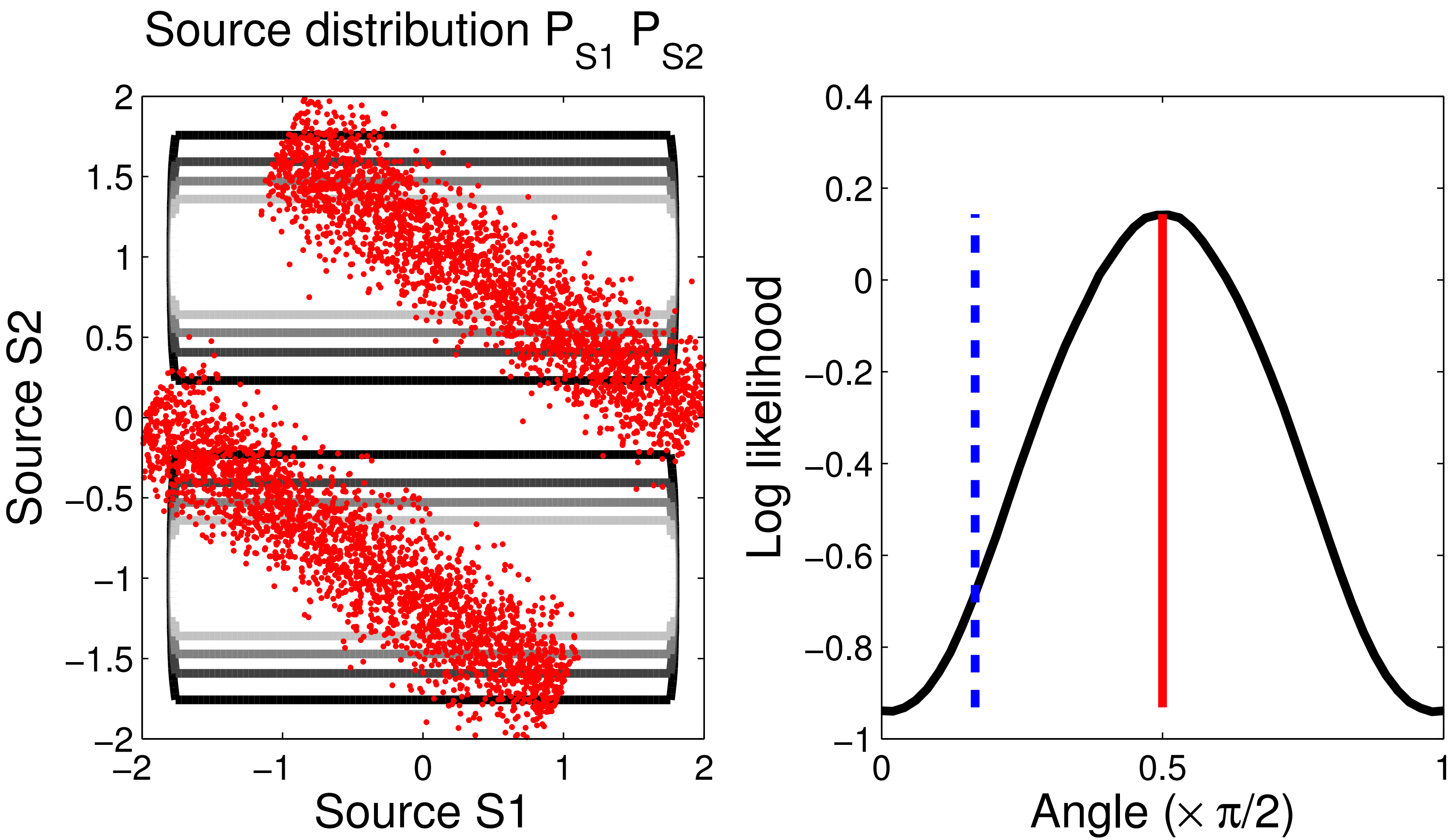
- Model for mixtures parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_S)$



Unmixing angle for B: 0

ICA: maximum likelihood

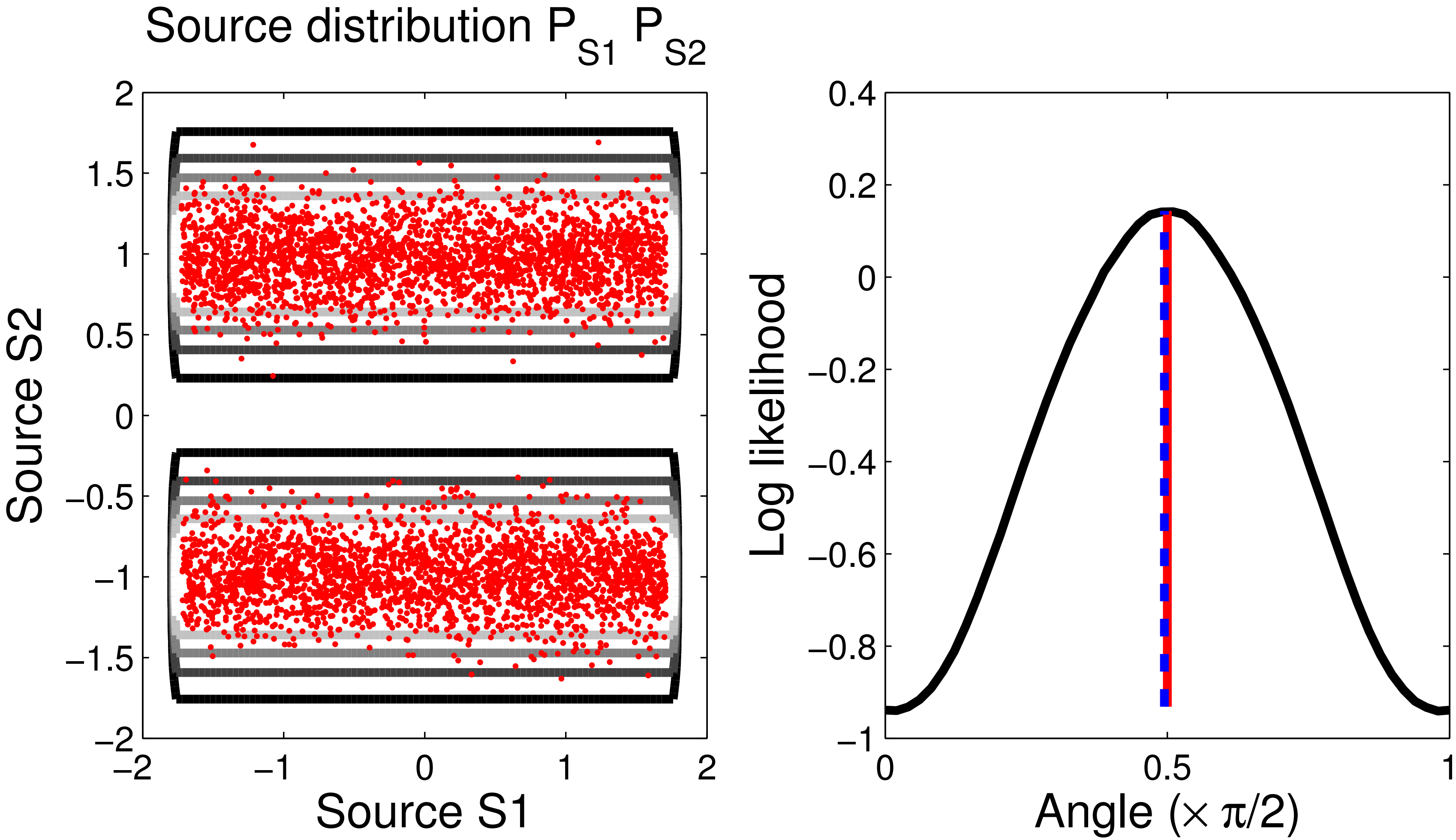
- Model for mixtures parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_S)$



Unmixing angle for B: $\pi/12$

ICA: maximum likelihood

- Model for mixtures parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_S)$



Unmixing angle for B: $\pi/4$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\boldsymbol{S}})$
  - Model must have $\hat{\mathsf{P}}_{\boldsymbol{S}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{S_i}$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\boldsymbol{S}})$
  - Model must have $\hat{\mathsf{P}}_{\boldsymbol{S}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{S_i}$
- With this model, our **estimated** density of observations is

$$\hat{\mathsf{P}}_{\boldsymbol{X}} = |\det(\mathbf{B})|\, \hat{\mathsf{P}}_{\boldsymbol{S}}(\mathbf{B}\boldsymbol{X})$$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\boldsymbol{S}})$
  - Model must have $\hat{\mathsf{P}}_{\boldsymbol{S}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{S_i}$
- With this model, our **estimated** density of observations is

$$\hat{\mathsf{P}}_{\boldsymbol{X}} = \cancel{|\det(\mathbf{B})|}\, \hat{\mathsf{P}}_{\boldsymbol{S}}(\mathbf{B}\boldsymbol{X})$$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\boldsymbol{S}})$
  - Model must have $\hat{\mathsf{P}}_{\boldsymbol{S}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{S_i}$
- With this model, our **estimated** density of observations is

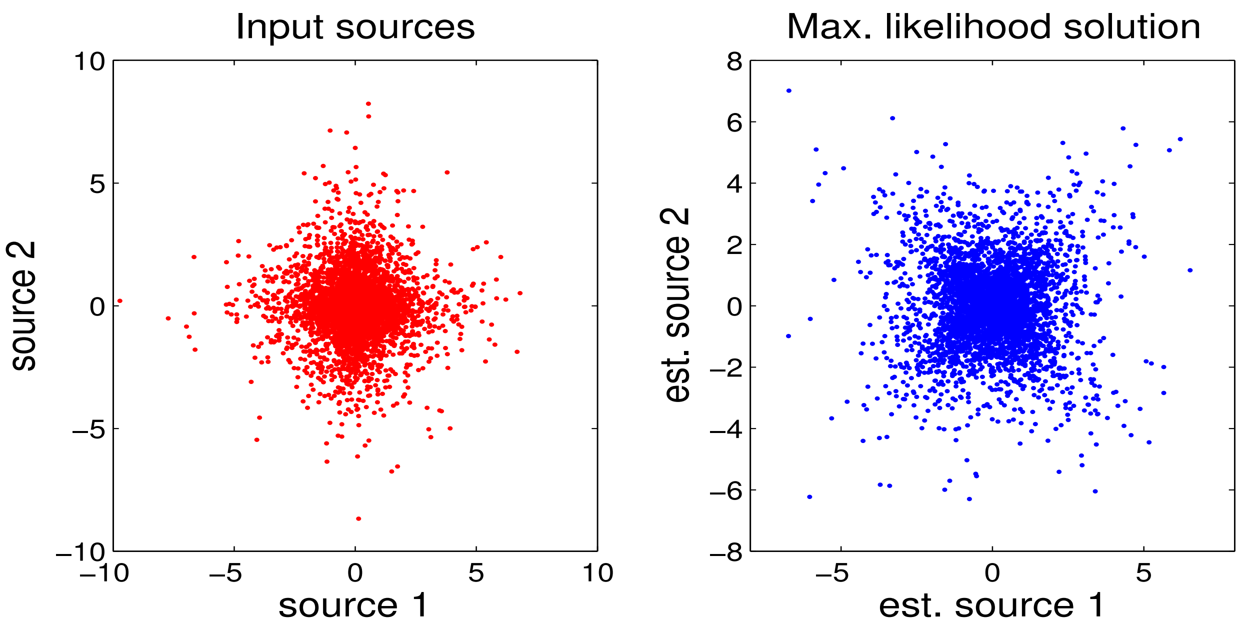
$$\hat{\mathsf{P}}_{\boldsymbol{X}} = \hat{\mathsf{P}}_{\boldsymbol{S}}(\mathbf{B}\boldsymbol{X})$$

- Maximise the expected log likelihood,

$$L := \mathbf{E}_{\boldsymbol{X}} \left[\log \hat{\mathsf{P}}_{\boldsymbol{X}} \right]$$

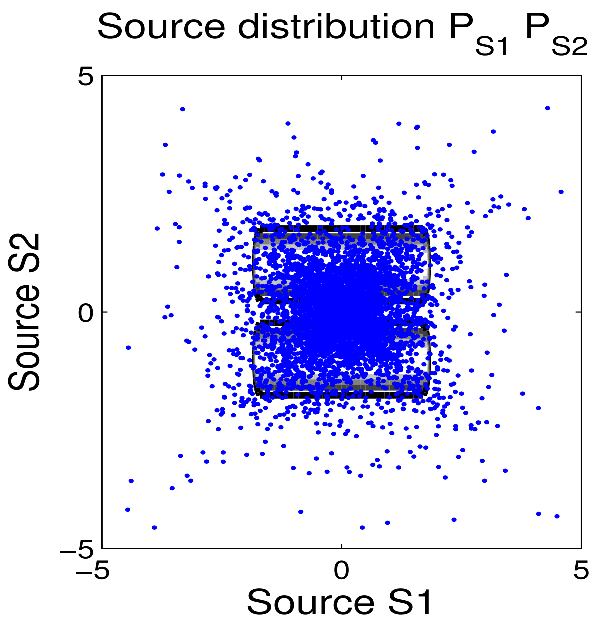
Maximum likelihood: where it fails

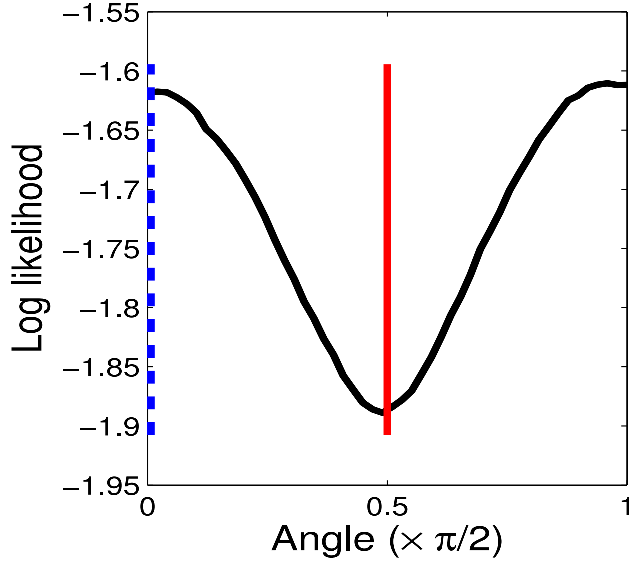
- Model as before, but true source densities are Laplace.
- Why is this so wrong?



Maximum likelihood: where it fails

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





Back to original setting: independence

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

$$\mathsf{P}_{\boldsymbol{Y}} = \prod_{i=1}^{l} \mathsf{P}_{Y_i}.$$

  – Under our mixing assumptions: $\boldsymbol{Y}$ are original sources $\boldsymbol{S}$ besides permutations, sign swaps

Back to original setting: independence

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

$$\mathsf{P}_{\boldsymbol{Y}} = \prod_{i=1}^{l} \mathsf{P}_{Y_i}.$$

  - Under our mixing assumptions: $\boldsymbol{Y}$ are original sources $\boldsymbol{S}$ besides permutations, sign swaps

- How it's *really* used: contrast should be "smallest" when random variables are "most independent"

Mutual information

- The mutual information:

$$I(\boldsymbol{Y}) = D_{\mathrm{KL}}\left(\mathsf{P}_{\boldsymbol{Y}} \,\middle\|\, \prod_{i=1}^{l} \mathsf{P}_{Y_i}\right)$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathsf{P}_{\boldsymbol{Y}} = \prod_{i=1}^{l} \mathsf{P}_{Y_i}$

Mutual information

- The mutual information:

$$I(\boldsymbol{Y}) = D_{\text{KL}}\left(\mathbf{P}_{\boldsymbol{Y}} \middle\| \prod_{i=1}^{l} \mathbf{P}_{Y_i}\right)$$

- $D_{\text{KL}} \geq 0$ with equality iff $\mathbf{P}_{\boldsymbol{Y}} = \prod_{i=1}^{l} \mathbf{P}_{Y_i}$
- Simplification: when $\mathbf{B}$ is a rotation,

$$D_{\text{KL}}\left(\mathbf{P}_{\boldsymbol{Y}} \middle\| \prod_{i=1}^{l} \mathbf{P}_{Y_i}\right) = \sum_{i=1}^{l} h\left(Y_i\right) - h\left(\boldsymbol{X}\right) - \log\left|\det \mathbf{B}\right|.$$

where $h(Y) = -\mathbf{E}_Y \log(\mathbf{P}_Y(y))$

Mutual information

- The mutual information:

$$I(\boldsymbol{Y}) = D_{\mathrm{KL}}\left(\mathsf{P}_{\boldsymbol{Y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{Y_i}\right)$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathsf{P}_{\boldsymbol{Y}} = \prod_{i=1}^{l} \mathsf{P}_{Y_i}$
- Simplification: when $\mathbf{B}$ is a rotation,

$$D_{\mathrm{KL}}\left(\mathsf{P}_{\boldsymbol{Y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{Y_i}\right) = \sum_{i=1}^{l} h\left(Y_i\right) - \underbrace{h\left(\boldsymbol{X}\right) - \log\left|\det \mathbf{B}\right|}_{\text{constant}}.$$

where $h(Y) = -\mathbf{E}_Y \log(\mathsf{P}_Y(y))$

Mutual information

- The mutual information:

$$I(\boldsymbol{Y}) = D_{\mathrm{KL}}\left(\mathsf{P}_{\boldsymbol{Y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{Y_i}\right)$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathsf{P}_{\boldsymbol{Y}} = \prod_{i=1}^{l} \mathsf{P}_{Y_i}$
- Simplification: when $\mathbf{B}$ is a rotation,

$$D_{\mathrm{KL}}\left(\mathsf{P}_{\boldsymbol{Y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{Y_i}\right) = \sum_{i=1}^{l} h\left(Y_i\right) - \underbrace{h\left(\boldsymbol{X}\right) - \log\left|\det \mathbf{B}\right|}_{\text{constant}}.$$

where $h(Y) = -\mathbf{E}_Y \log(\mathsf{P}_Y(y))$

Contrast: $\phi_{KL}(\boldsymbol{Y}) := \sum_{i=1}^{l} h\left(Y_i\right)$

Maximum likelihood revisited

- Mutual information contrast: minimize

$$\phi_{KL}(\boldsymbol{Y}) := \sum_{i=1}^{l} -\mathbf{E}_Y \log(\mathsf{P}_Y(y))$$

- Maximum likelihood: maximize

$$\begin{aligned} L &:= \mathbf{E}_{\boldsymbol{X}} \left[\log \hat{\mathsf{P}}_{\boldsymbol{S}}(\mathbf{B}\boldsymbol{X})\right] \\ &= \sum_{i=1}^{l} \mathbf{E}_Y \log(\mathsf{P}_Y(y)) \end{aligned}$$

- Same thing!

Maximum likelihood revisited

- Mutual information contrast: minimize

$$\phi_{KL}(\boldsymbol{Y}) := \sum_{i=1}^{l} -\mathbf{E}_Y \log(\mathsf{P}_Y(y))$$

- Maximum likelihood: maximize

$$\begin{aligned} L &:= \mathbf{E}_{\boldsymbol{X}} \left[\log \hat{\mathsf{P}}_{\boldsymbol{S}}(\mathbf{B}\boldsymbol{X})\right] \\ &= \sum_{i=1}^{l} \mathbf{E}_Y \log(\mathsf{P}_Y(y)) \end{aligned}$$

- Same thing!
- The difference is in approach:
  - For max. likelihood we assumed a model $\hat{\mathsf{P}}_{\boldsymbol{S}}$
  - Now we assume no model for $\mathsf{P}_{\boldsymbol{Y}}$ (though we still make assumptions)

Contrast functions with fixed nonlinearities

- Entropies hard to compute/optimize: replace with

$$\phi_f(\boldsymbol{Y}) = \sum_{j=1}^{l} \mathbf{E}(f(Y_j))$$

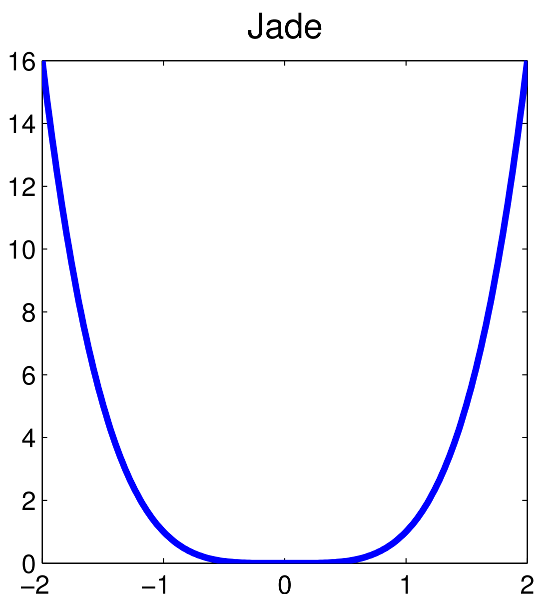
for some other nonlinear $f(y)$

Contrast functions with fixed nonlinearities

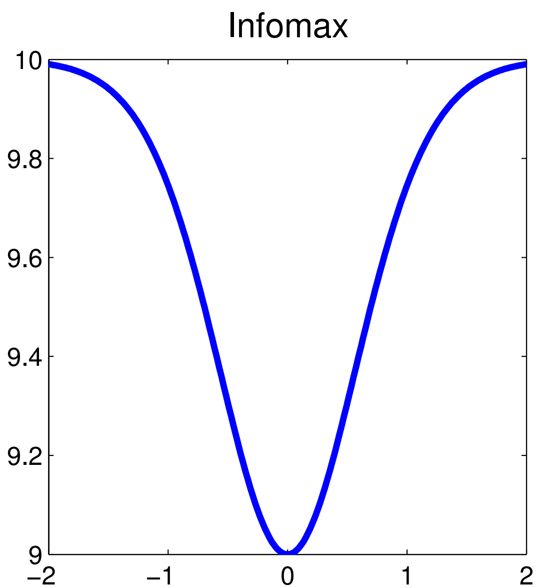
- Entropies hard to compute/optimize: replace with

$$\phi_f(\boldsymbol{Y}) = \sum_{j=1}^{l} \mathbf{E}(f(Y_j))$$

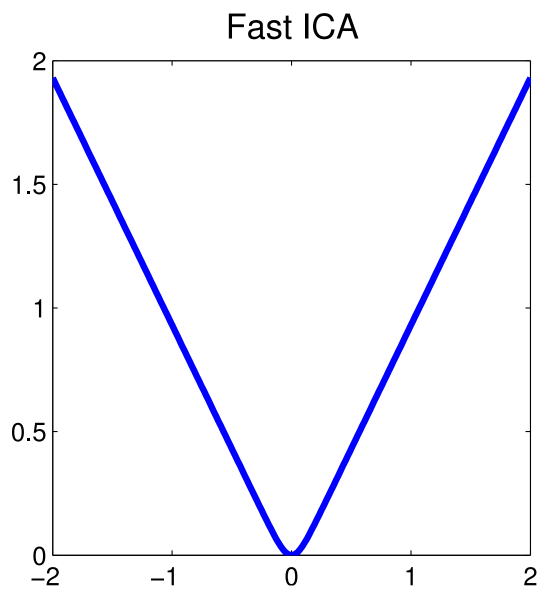
for some other nonlinear $f(y)$



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

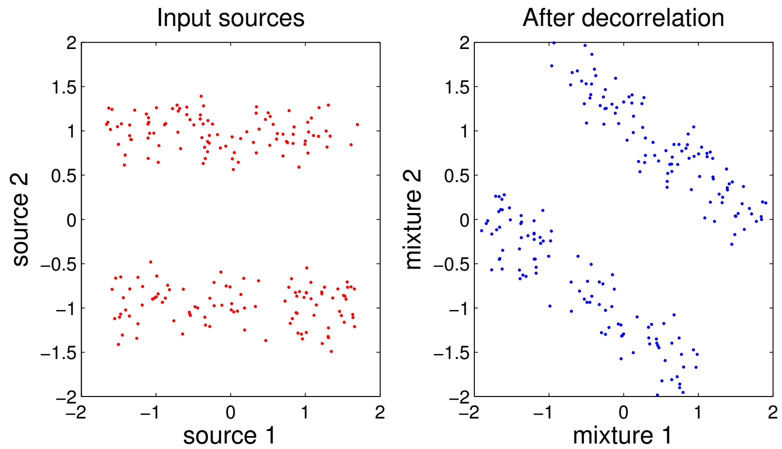


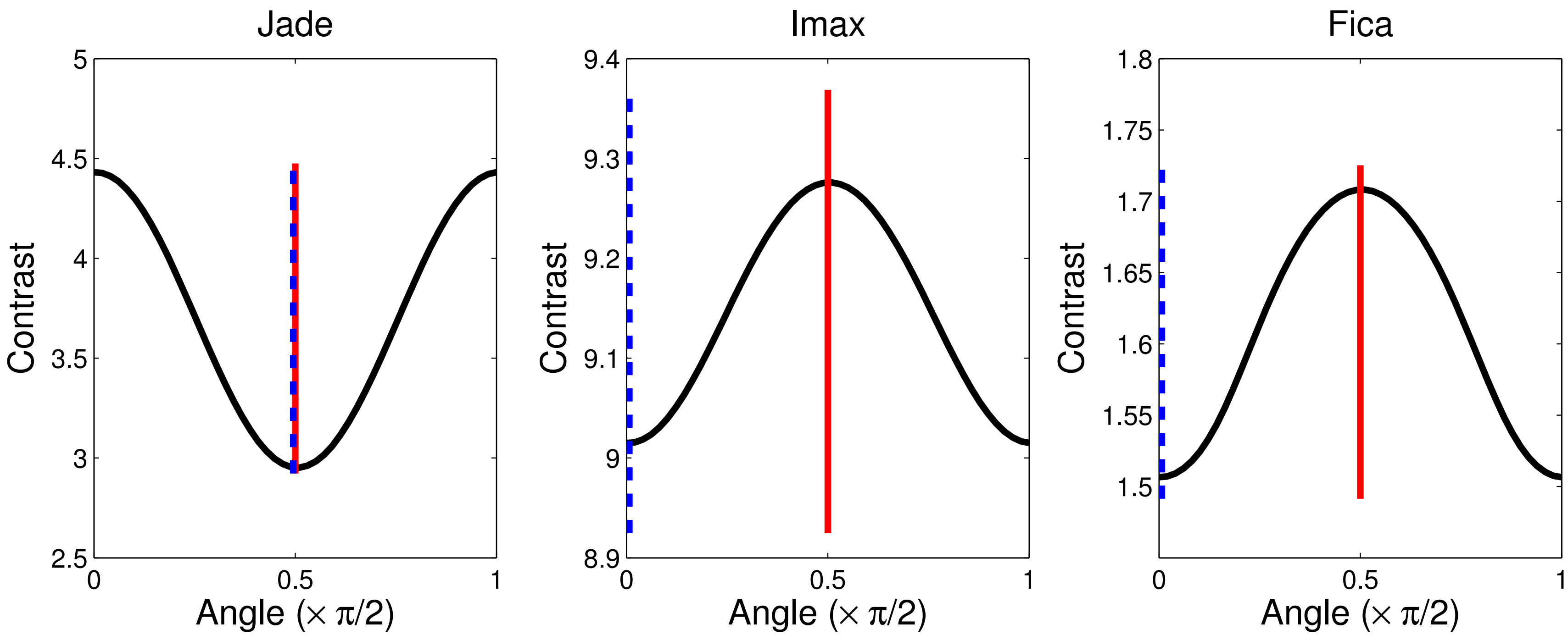
$$f(y) = a - exp(-y^2/2)\text{sech}^2(y)$$



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

Our example again



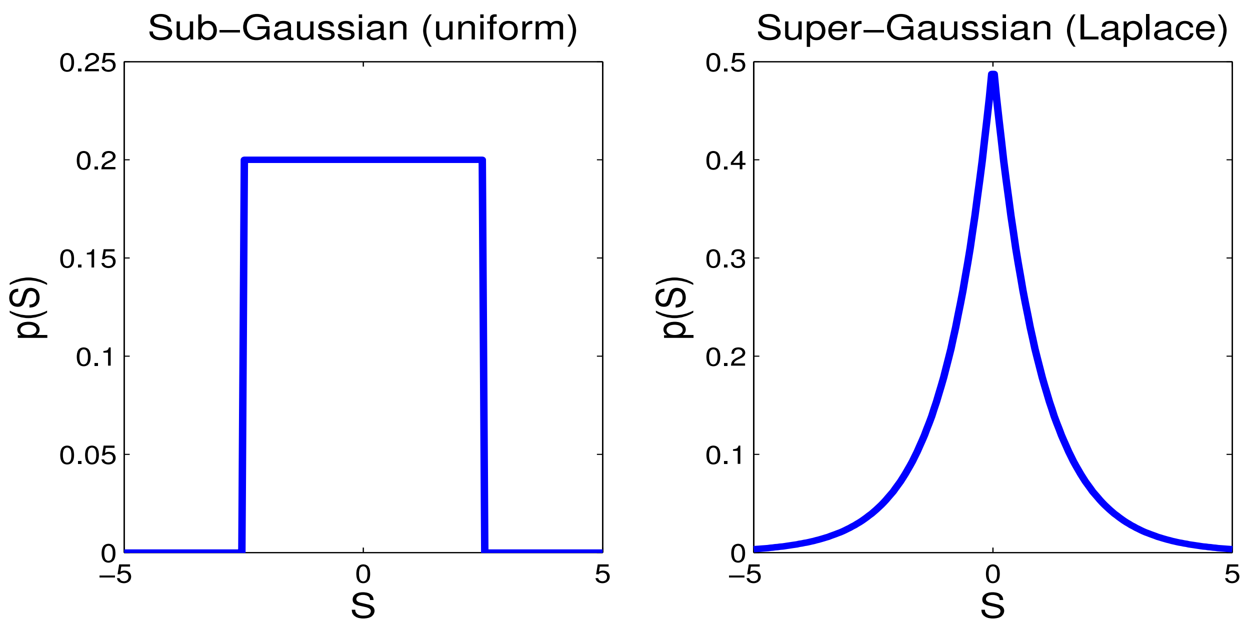


Kurtosis: an important concept

- Kurtosis definition: when mean is zero,

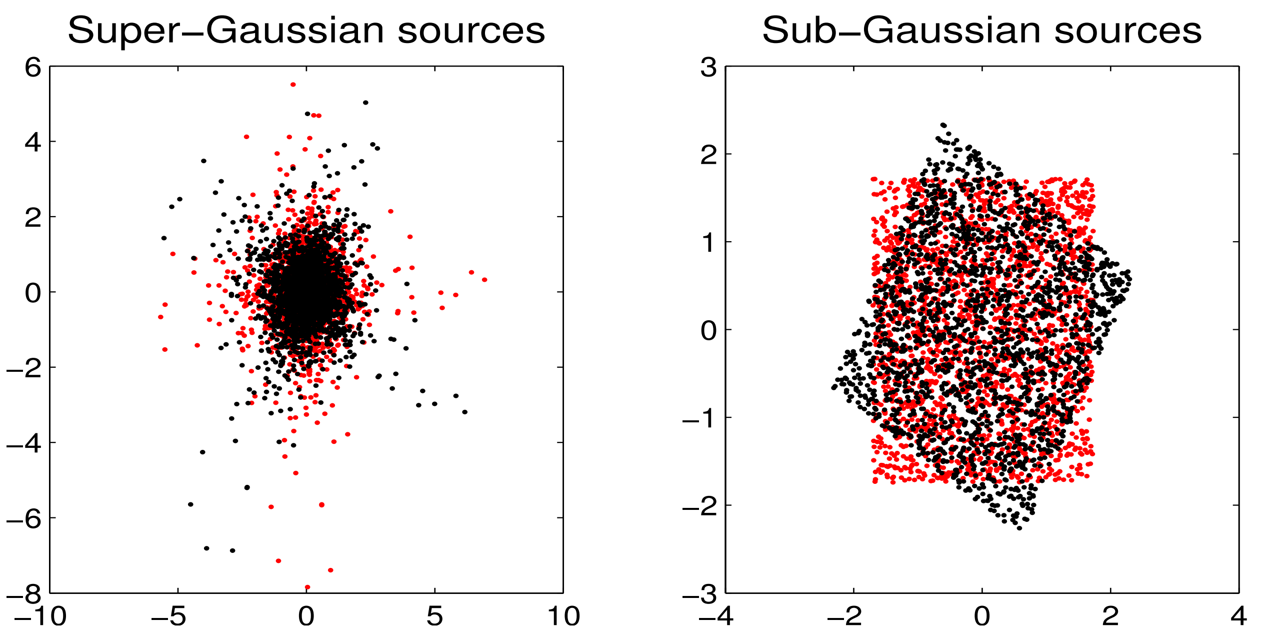
$$\kappa_4 = \mathbf{E}\left(\mathsf{x}^4\right) - 3\left(\mathbf{E}\left(\mathsf{x}^2\right)\right)^2.$$

- Source densities can be super-Gaussian (positive kurtosis) or sub-Gaussian (negative kurtosis)
- Zero kurtosis does not mean Gaussian!



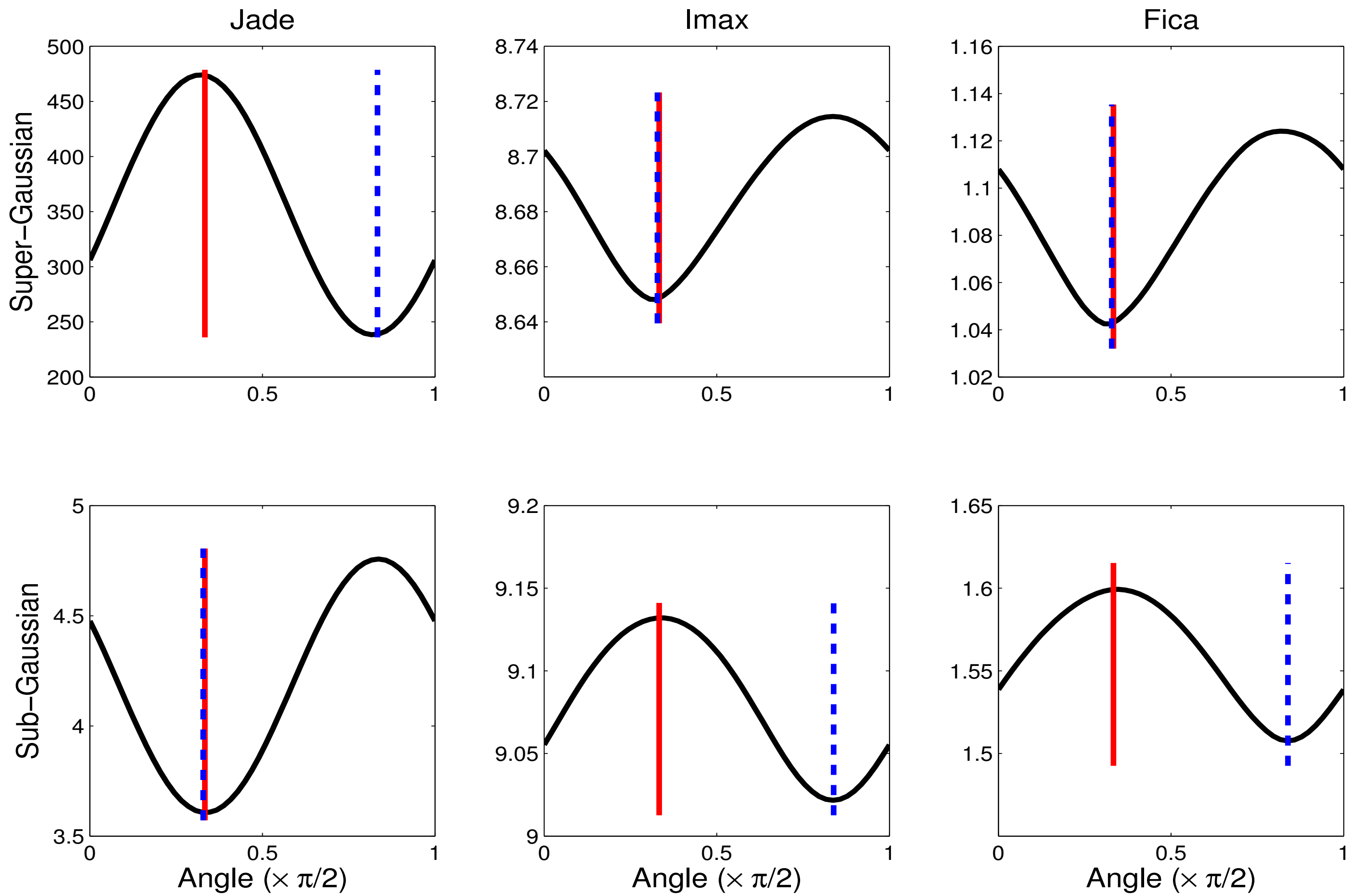
Demo: contrasts with fixed nonlinearities

- Super-Gaussian (Laplace) and sub-Gaussian (Uniform) sources
- Unmixed sources in red
- Mixture (angle $\pi/6$) in black



Demo: contrasts with fixed nonlinearities

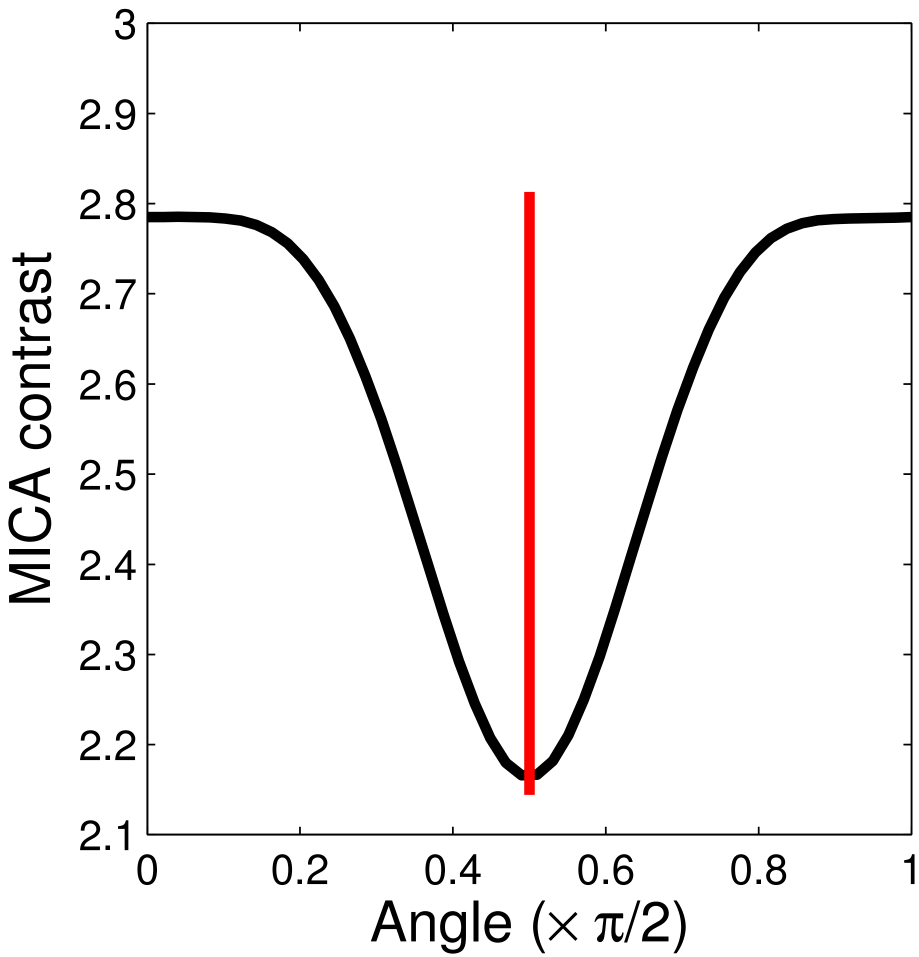
- Results for Jade, Infomax, and Fast ICA contrasts

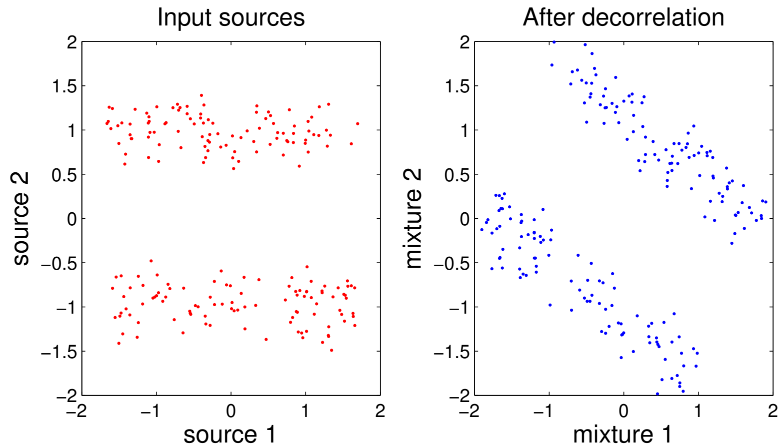


Care needed when using fixed contrasts!

Contrast functions using entropy estimates

- Simplest option: convolve with spline kernel, then compute discrete entropy via space partition [Pham, 2004]

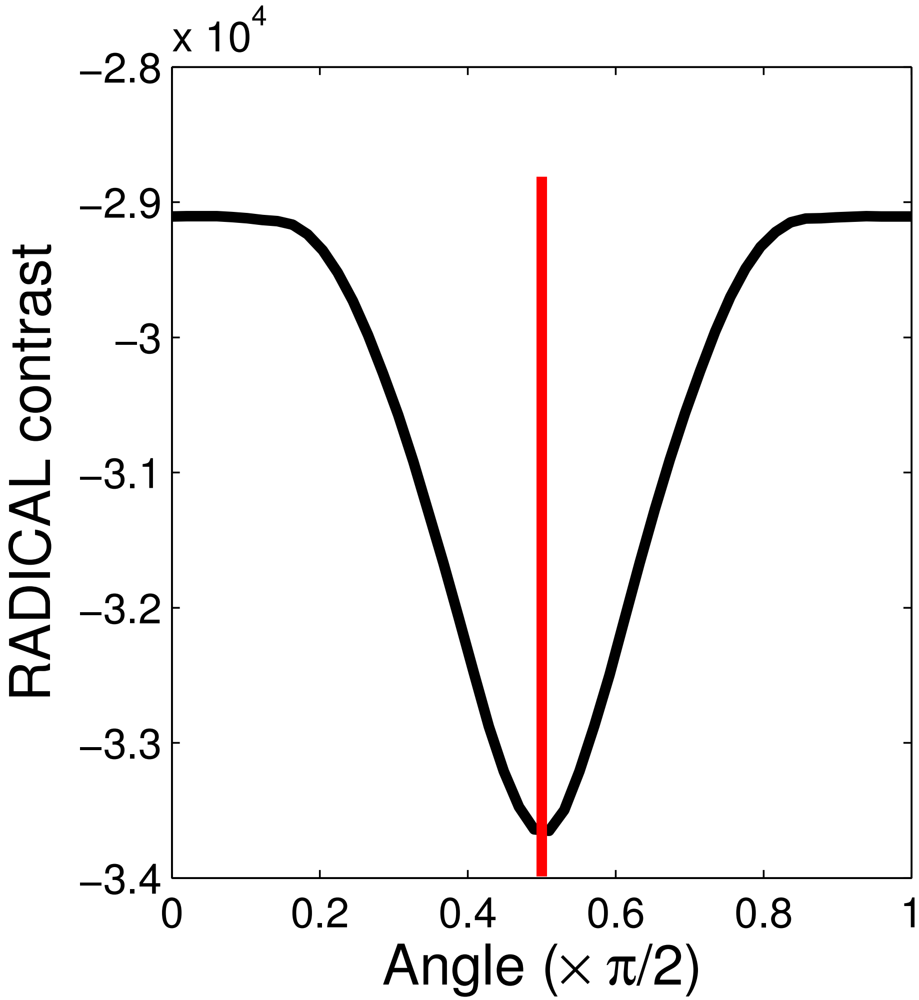


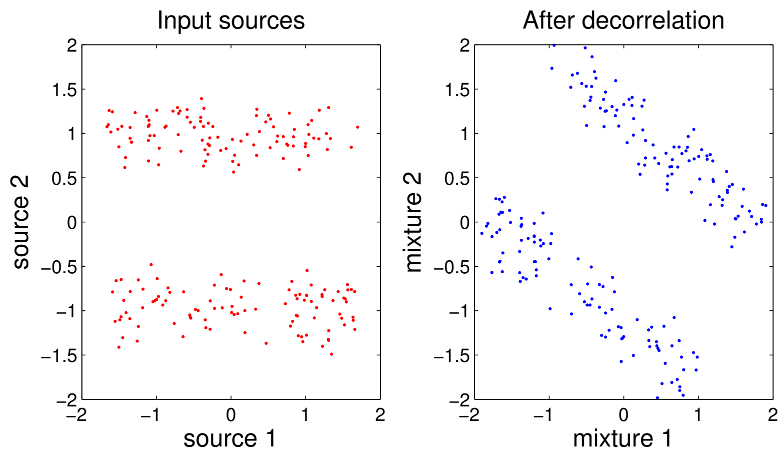


Contrast functions using spacings entropy estimate

- More sophisticated option: spacings estimate of entropy

[Learned-Miller and Fisher III, 2003]





Contrast functions using spacings entropy estimate

- More sophisticated option: spacings estimate of entropy

  [Learned-Miller and Fisher III, 2003]

- Sort sample $Y_1, \ldots, Y_m$ in increasing order: $Y_{(i)} \leq Y_{(i+1)}$
- Prob. density estimate based on spacings

$$\hat{\mathbf{P}}(y; Y_1, \ldots, Y_m) = \frac{1}{(m+1)(Y_{(i+1)} - Y_{(i)})}, \qquad Y_{(i)} \leq y < Y_{(i+1)}$$

- Entropy estimate based on spacings

$$\hat{h}(Y) = \frac{1}{m-1} \sum_{i=1}^{m-1} \log(m+1)(Y_{(i+1)} - Y_{(i)})$$

- Smoothing: add "extra" mixture points (noisy copies of original mixtures)
- Hard to optimize

Other independence measures as contrasts

- Why mutual information?
  - Same as maximum likelihood (good if model is correct)
  - Contrast function is sum of entropies: fast
- Other independence measures?

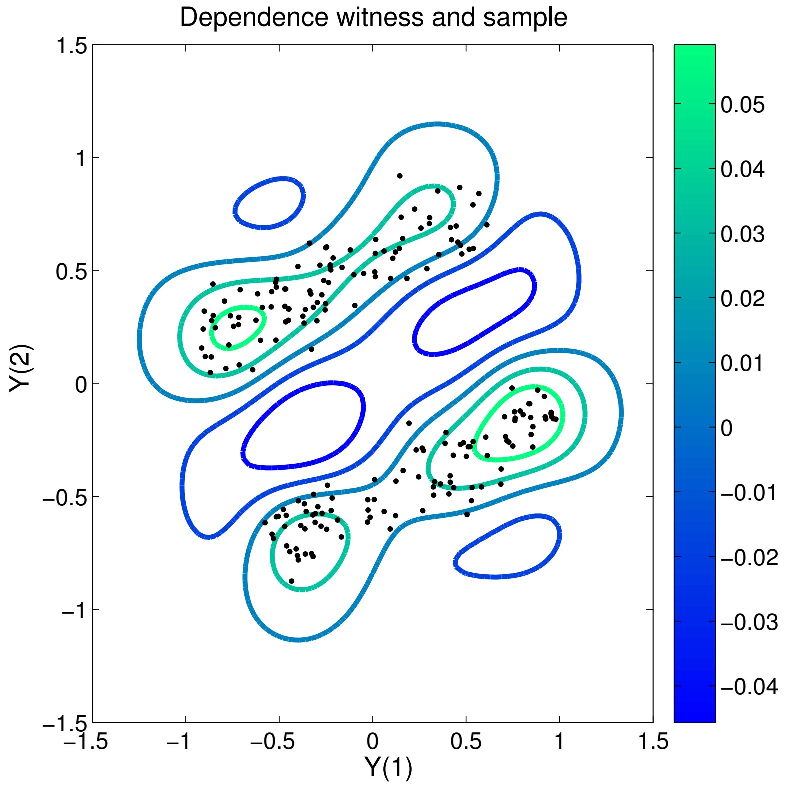
Other independence measures as contrasts

- Why mutual information?
  - Same as maximum likelihood (good if model is correct)
  - Contrast function is sum of entropies: fast
- Other independence measures?
- Most common: kernel/characteristic function-based
  - Characteristic function-based ICA [Eriksson and Koivunen, 2003, Chen and Bickel, 2005]
  - Kernel ICA (covariance): COCO, KMI, HSIC [Gretton et al., 2005, Shen et al., 2007, 2009]
  - Kernel ICA (correlation): KCCA, KGV [Bach and Jordan, 2002]
- HSIC same as characteristic function-based (for the purposes of ICA) [Shen et al., 2009]

Kernel contrast function: HSIC

- Dependence measure:

$$\mathrm{HSIC}(\mathbf{P}_{UV}, F) := \left(\sup_{f \in F} \left[\mathbf{E}_{UV} f - \mathbf{E}_U \mathbf{E}_V f \right] \right)^2$$



HSIC: empirical expression

- Empirical HSIC:

$$\text{HSIC} := \frac{1}{m^2}\text{tr}(KHLH)$$

  - K Gram matrix for $(u_1, \ldots, u_m)$
  - L Gram matrix for $(v_1, \ldots, v_m)$
  - Centering $H = I - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top$

Contrast functions: a small selection

Contrast functions

- Sum of expectations of a fixed nonlinearity
  - Fast ICA, Infomax, Jade
- Sum of entropies/mutual information...
  - ... using fast, smoothed entropy estimates
  - ... using spacings/k-nn entropy estimates
- Kernel/characteristic function dependence measures

Contrast functions: a small selection

Contrast functions

- Sum of expectations of a fixed nonlinearity
  - Fast ICA, Infomax, Jade
- Sum of entropies/mutual information…
  - … using fast, smoothed entropy estimates
  - … using spacings/k-nn entropy estimates
- Kernel/characteristic function dependence measures

How do we optimize?

Optimization (Jacobi)

- For two signals, the rotation is expressed

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

- Higher dimensions, eg for $l = 3$,

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

- Coordinate descent, exhaustive search, etc...

Optimization (Newton)

- Unmixing matrix B satisfies $B^\top B = I$
- Local parameterisation Ω about B: at iteration k,

$$B_{k+1} = B_k \exp(\Omega) \qquad \Omega = -\Omega^\top$$

- How to choose direction and size of Ω?

Optimization (Newton)

- Unmixing matrix B satisfies $B^\top B = I$
- Local parameterisation Ω about B: at iteration k,

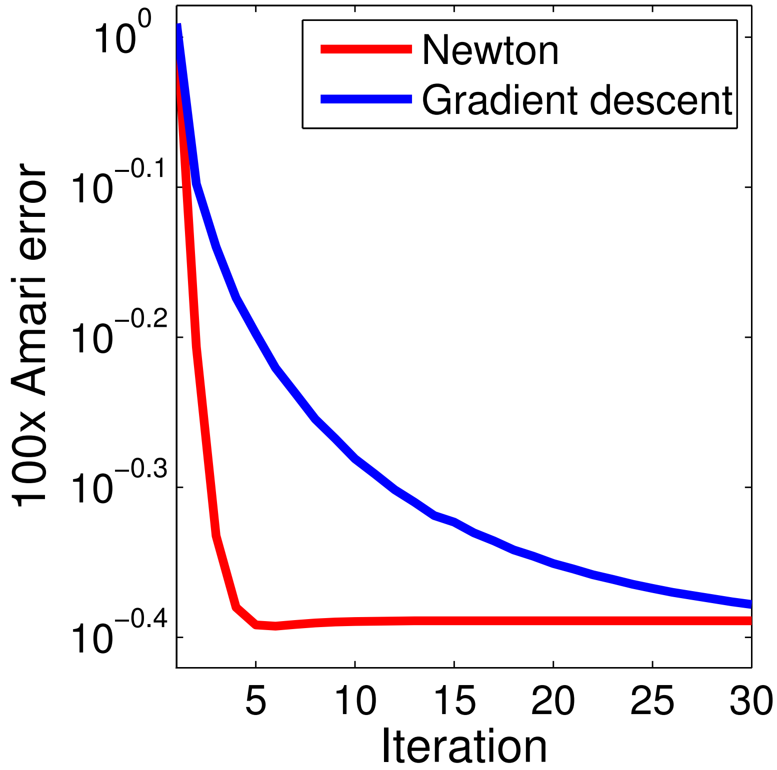
$$B_{k+1} = B_k \exp(\Omega) \qquad \Omega = -\Omega^\top$$

- How to choose direction and size of Ω?
- Newton-like method: solve the linear system for $\Omega \in \mathbb{R}^{m(m-1)/2}$

$$\mathcal{H}_{B_k}(\phi)\Omega = -\nabla_{B_k}(\phi)$$

- Approximate Hessian as diagonal: FastICA [Shen and Hüper, 2006]

Gradient descent vs Newton



What if we have time dependence?

- We can get extra information from sources not being i.i.d.
- Mixture $\mathbf{x}(t)$ now stationary random process, depends on $\mathbf{x}(t-\tau)$
- Define mixture covariances

$$\mathbf{C}_0 = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t)), \qquad \mathbf{C}_\tau = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t-\tau)),$$

  – $\mathbf{C}_\tau$ indpendent of t (stationarity)

What if we have time dependence?

- We can get extra information from sources not being i.i.d.
- Mixture $\mathbf{x}(t)$ now stationary random process, depends on $\mathbf{x}(t-\tau)$
- Define mixture covariances

$$\mathbf{C}_0 = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t)), \qquad \mathbf{C}_\tau = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t-\tau)),$$

  - $\mathbf{C}_\tau$ indpendent of t (stationarity)
- Decorrelate:

$$\mathbf{B}\mathbf{C}_0\mathbf{B}^\top = \Lambda \qquad \mathbf{B}\mathbf{C}_\tau\mathbf{B}^\top = \widetilde{\Lambda}$$

  - Λ and $\widetilde{\Lambda}$ diagonal
- Combining both requirements:

$$\mathbf{B}\mathbf{C}_0\mathbf{C}_\tau^{-1} = \left(\Lambda\widetilde{\Lambda}^{-1}\right)\mathbf{B}$$

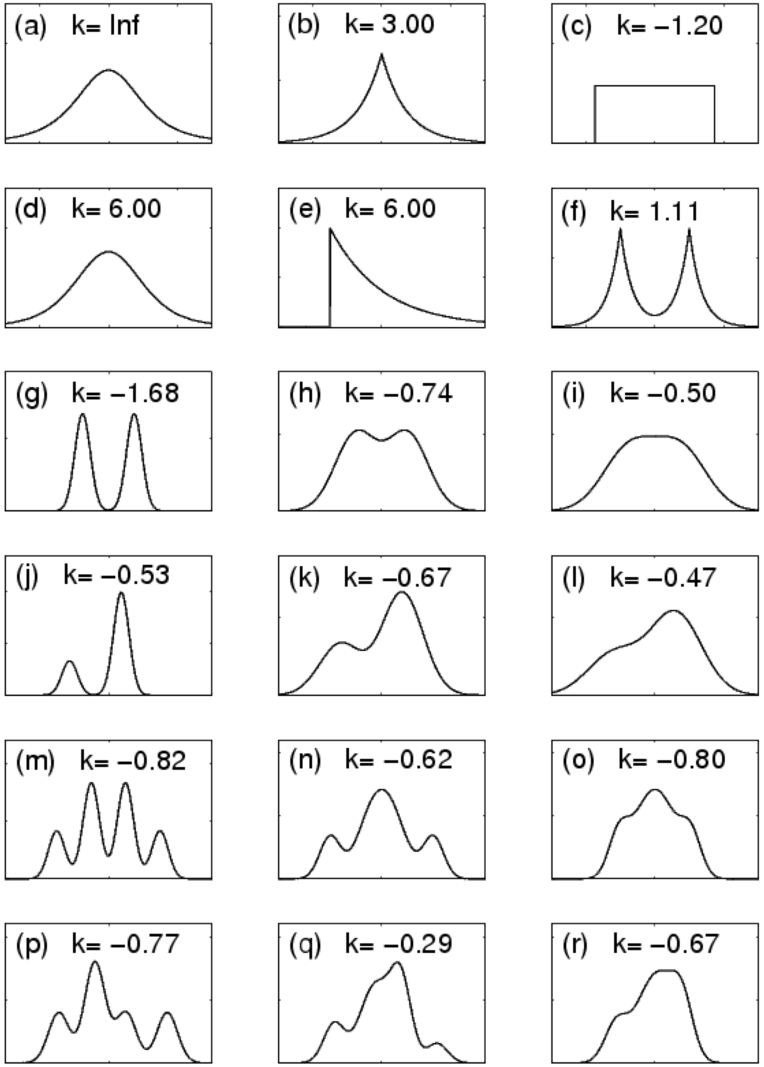
- Greater number of delays: joint diagonalisation

What's the best method?

A basic benchmark

- $l = 8$ sources
- $m = 40,000$ samples
- Benchmark data from

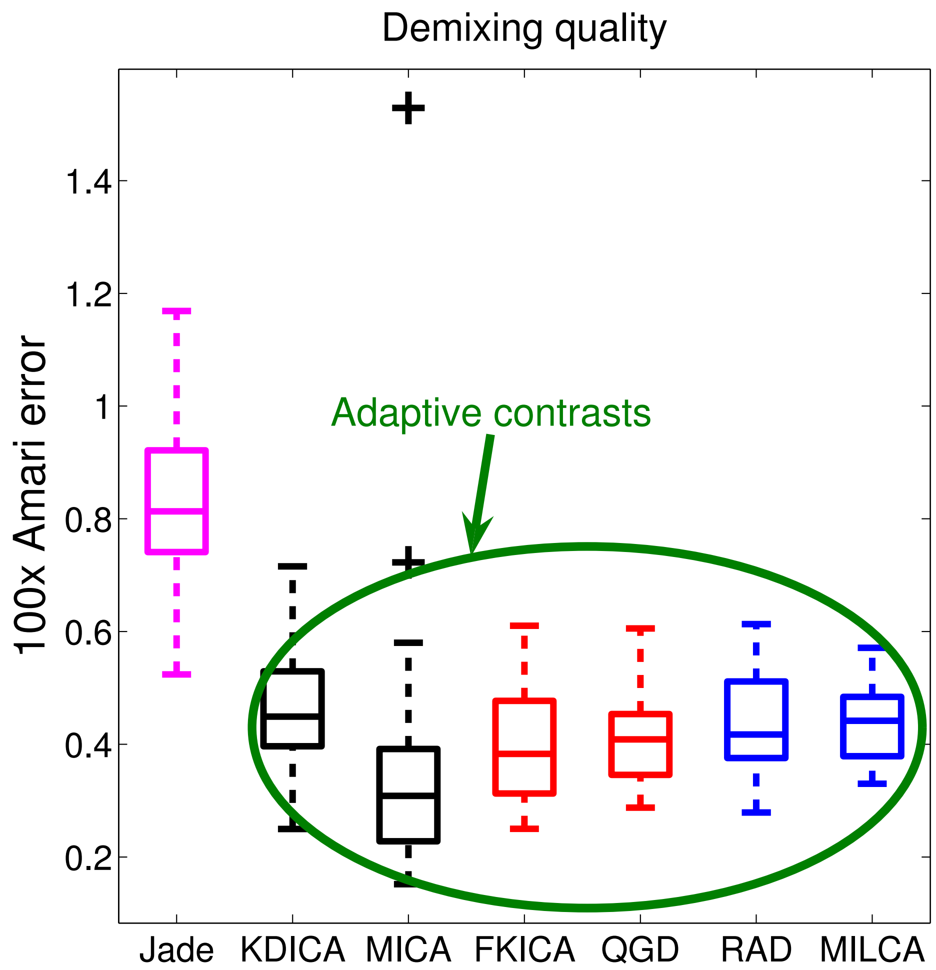
  [Bach and Jordan, 2002]
- Average over 24 repetitions



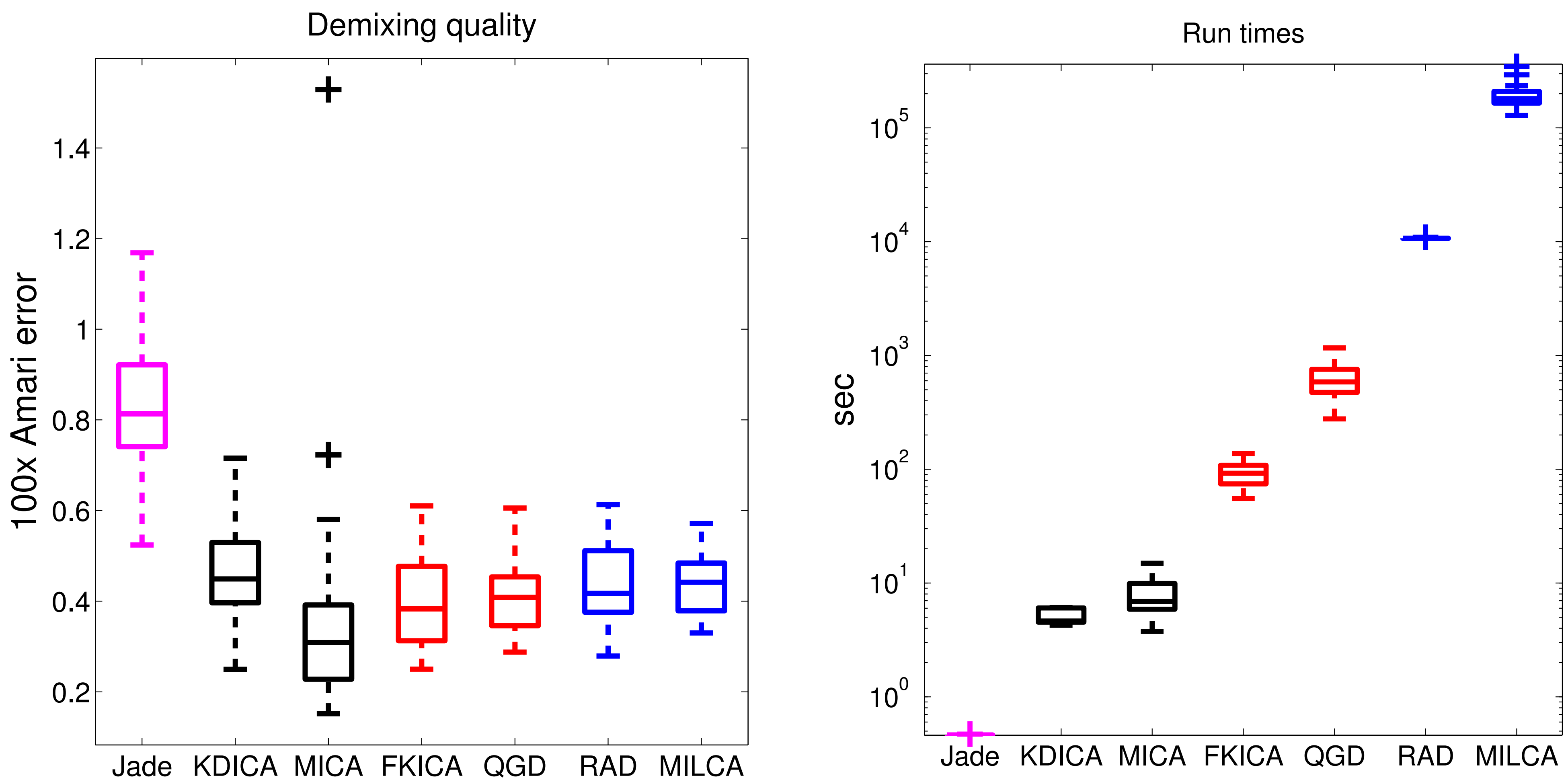
A basic benchmark: results

A basic benchmark: results

Adaptive contrasts outperform fixed nonlinearities

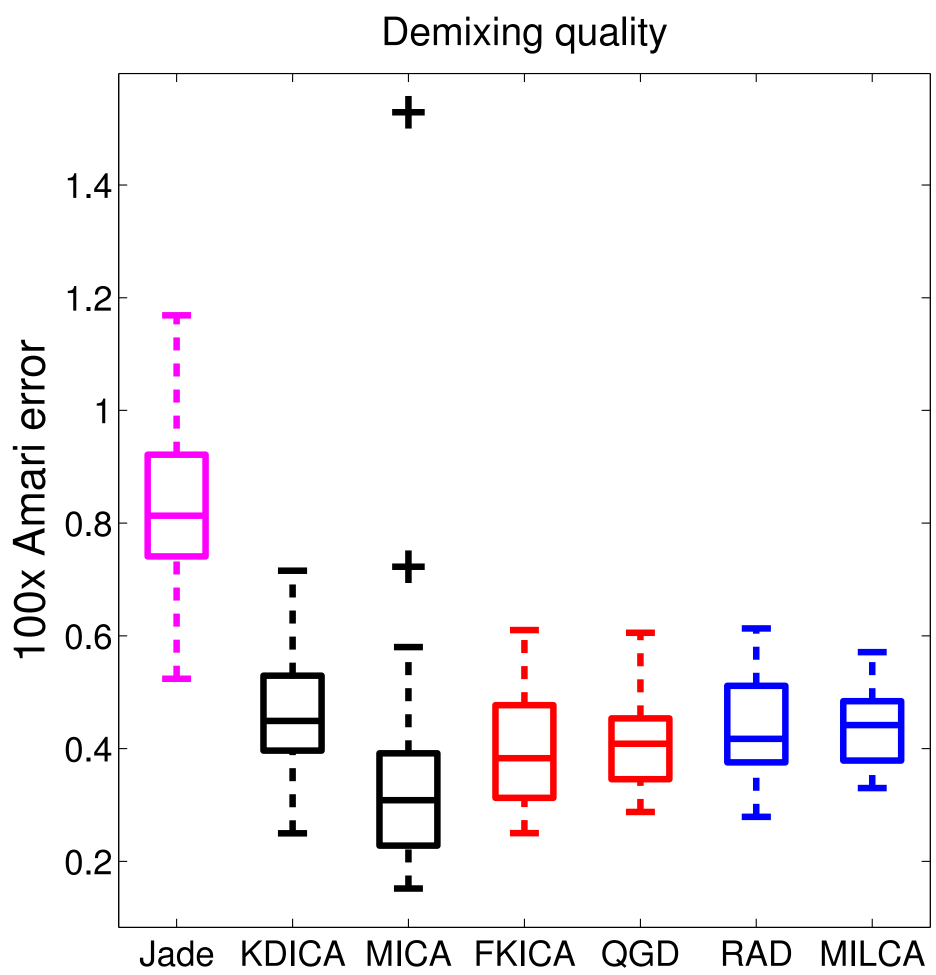


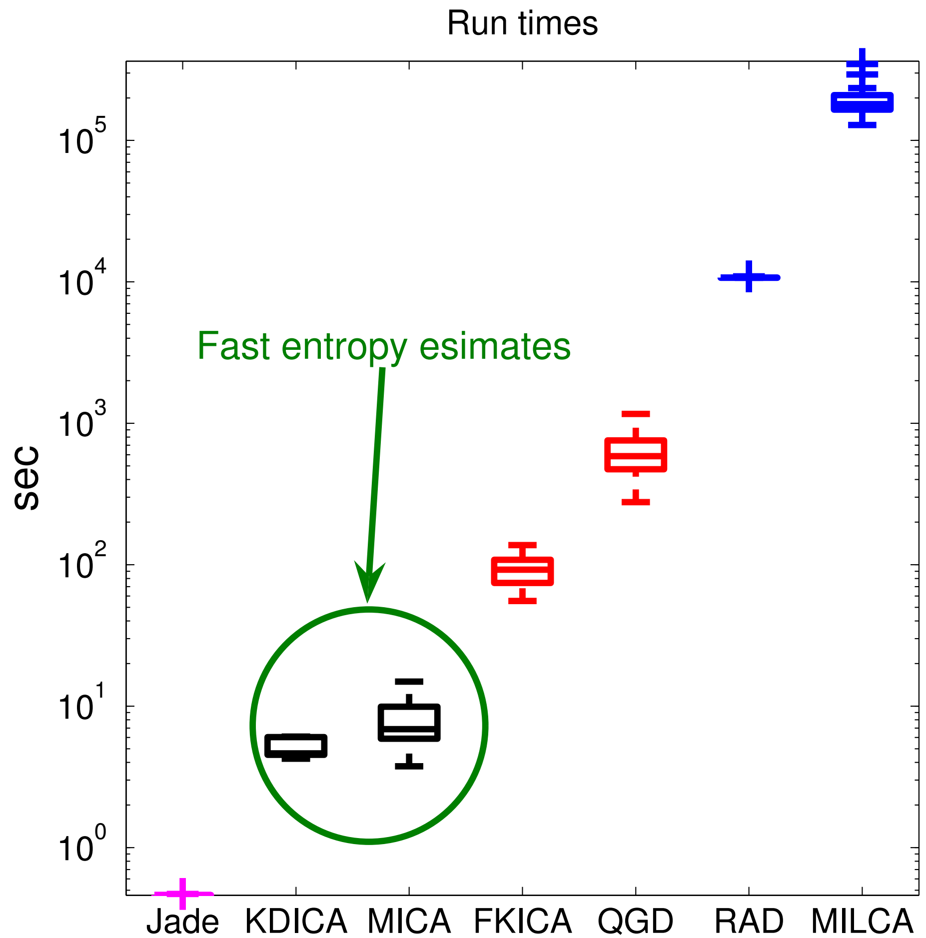
A basic benchmark: computational cost



A basic benchmark: computational cost

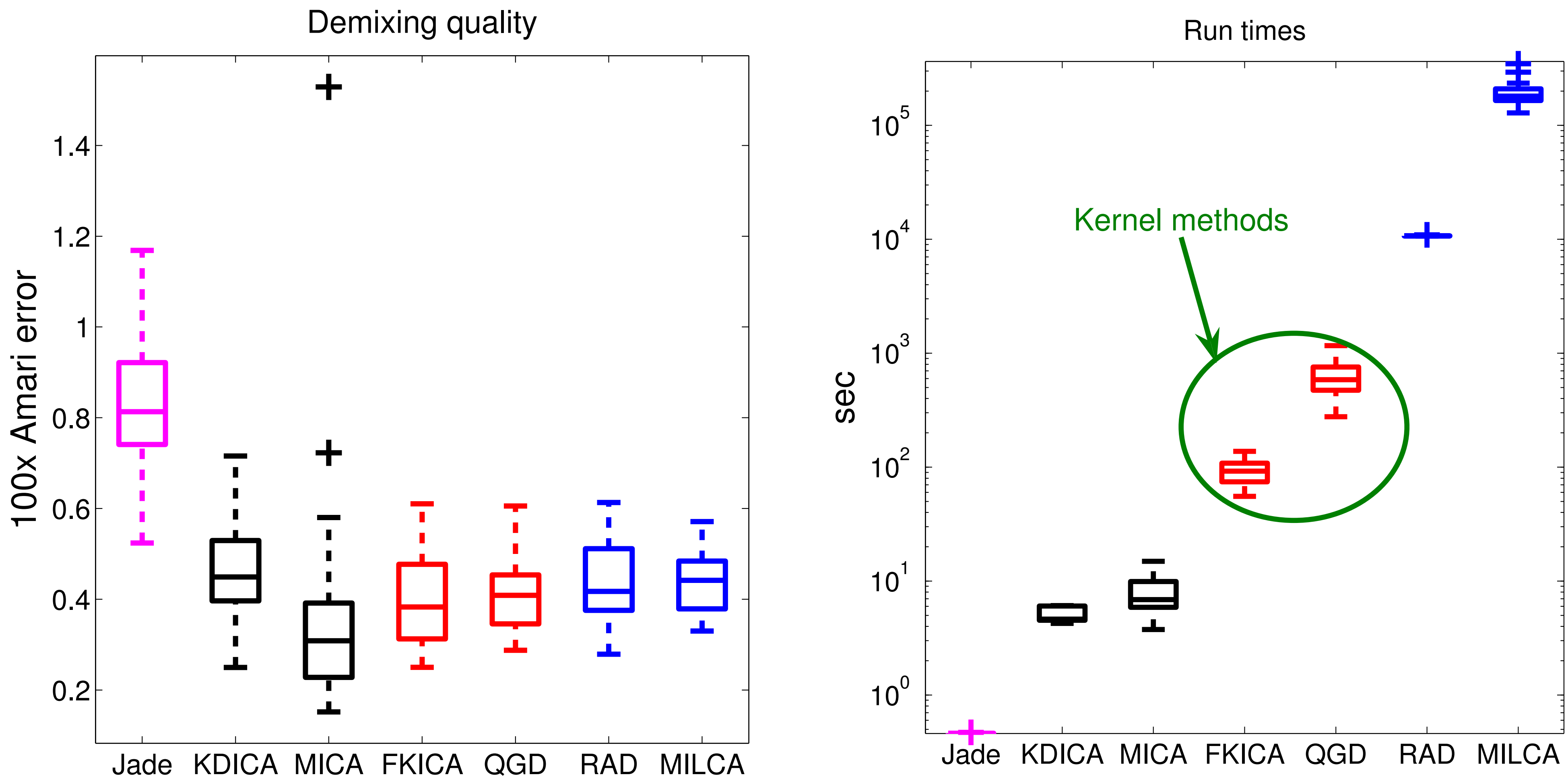
Best runtime (adaptive): fast entropy estimates





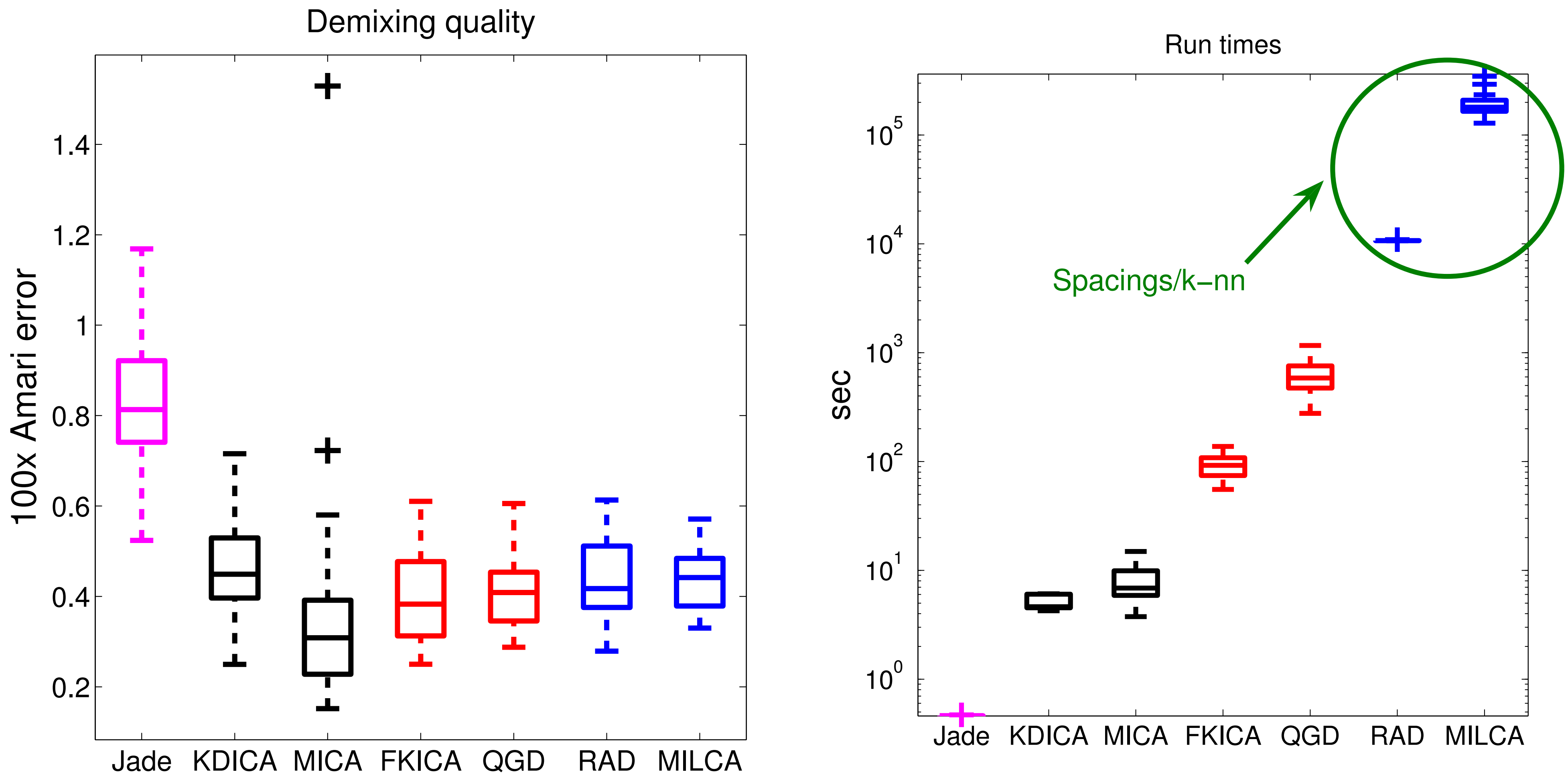
A basic benchmark: computational cost

Kernel methods: Newton outperforms Gradient Descent



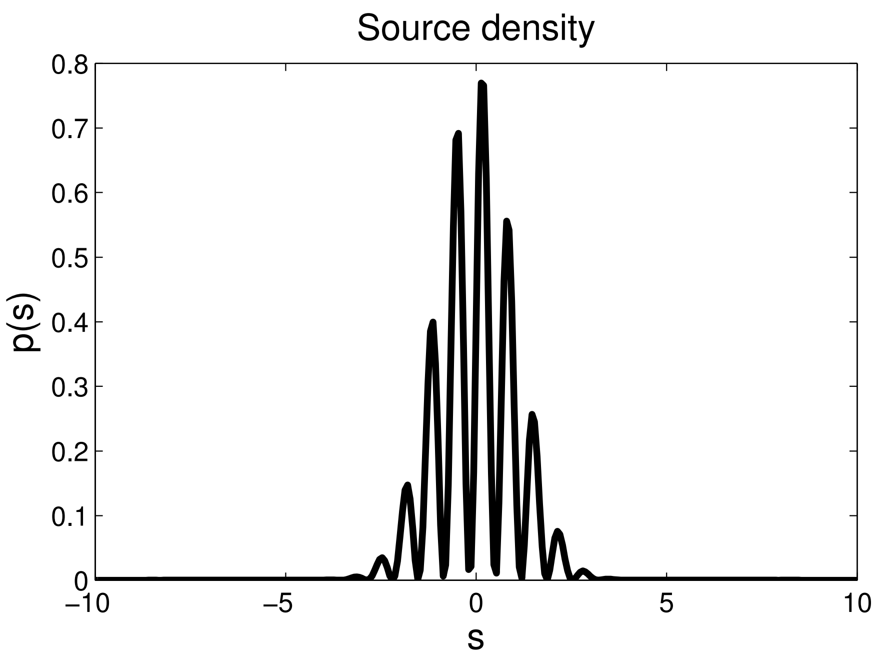
A basic benchmark: computational cost

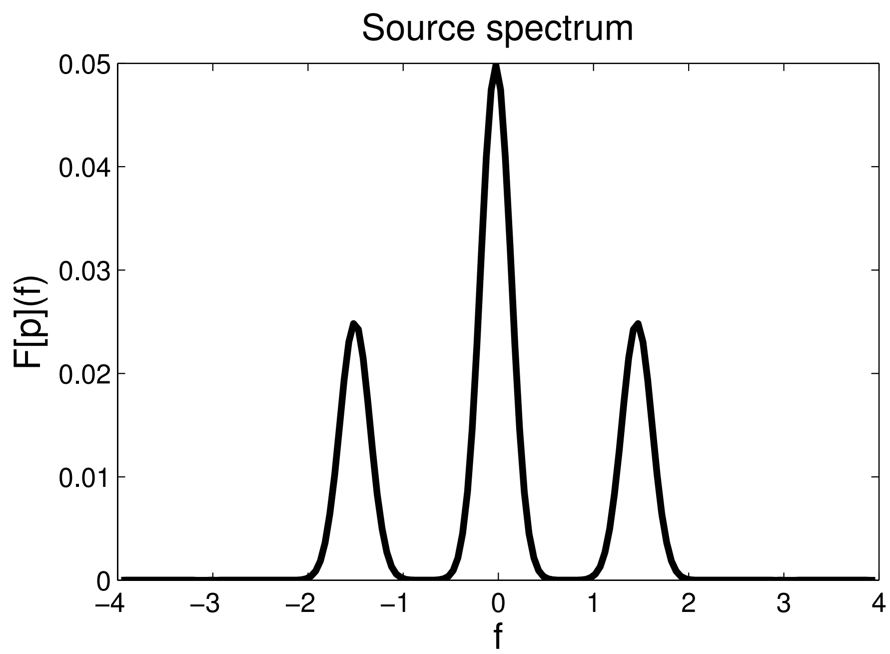
Spacings/k-nn entropy contrasts slowest



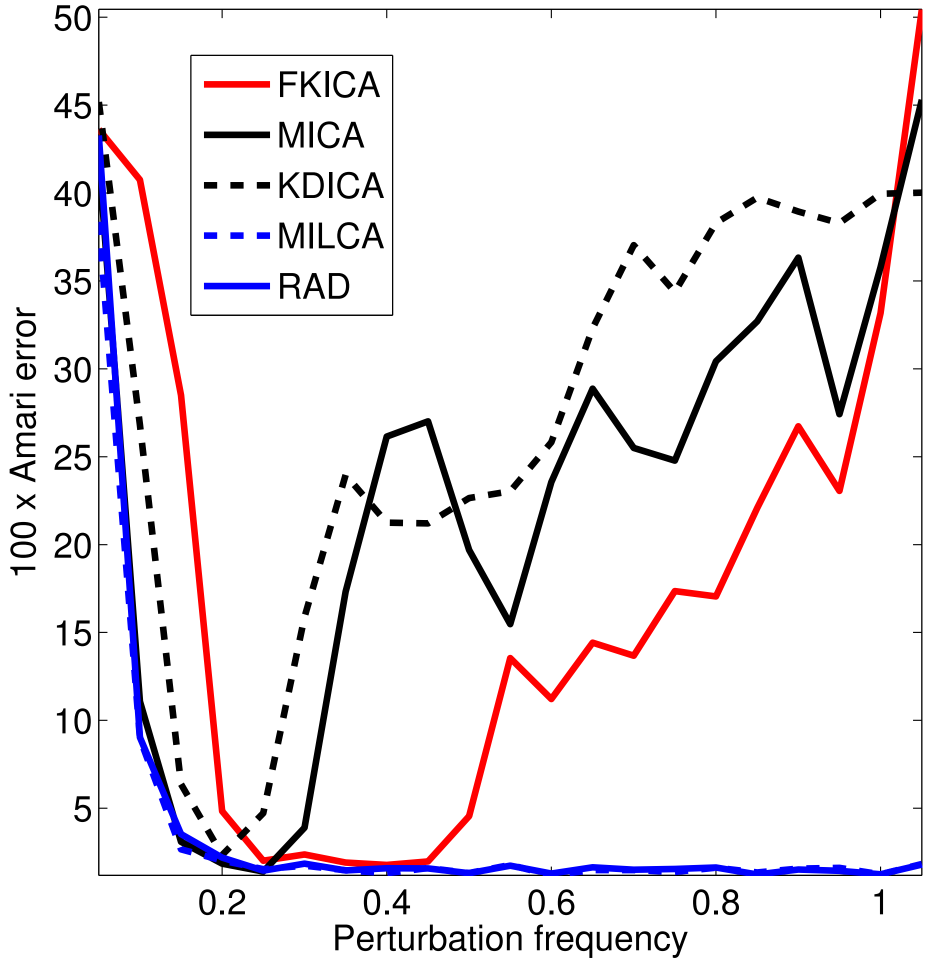
High frequency perturbations

- Two sources, sinusoidal perturbations to Gaussian
- Random mixing angle.
- Results averaged over 25 datasets, $m = 1000$





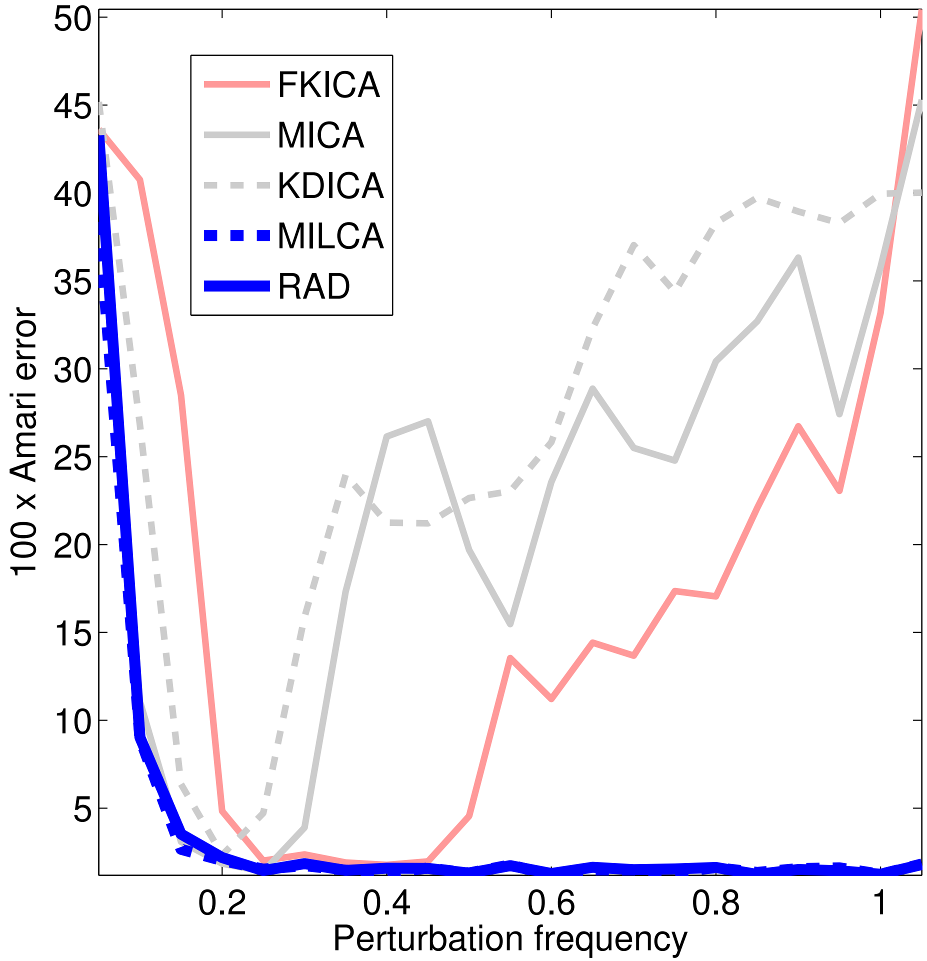
High frequency perturbations



High frequency perturbations

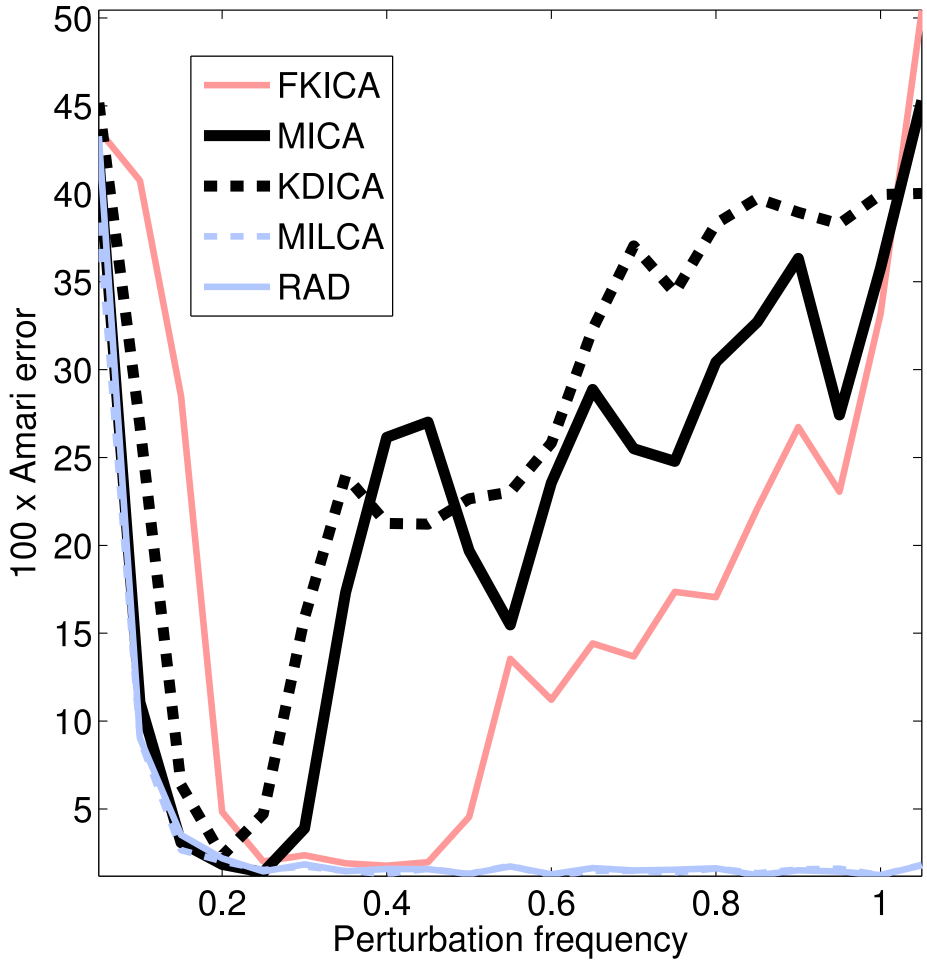
Spacings/k-nn methods perform best

(but slow)



High frequency perturbations

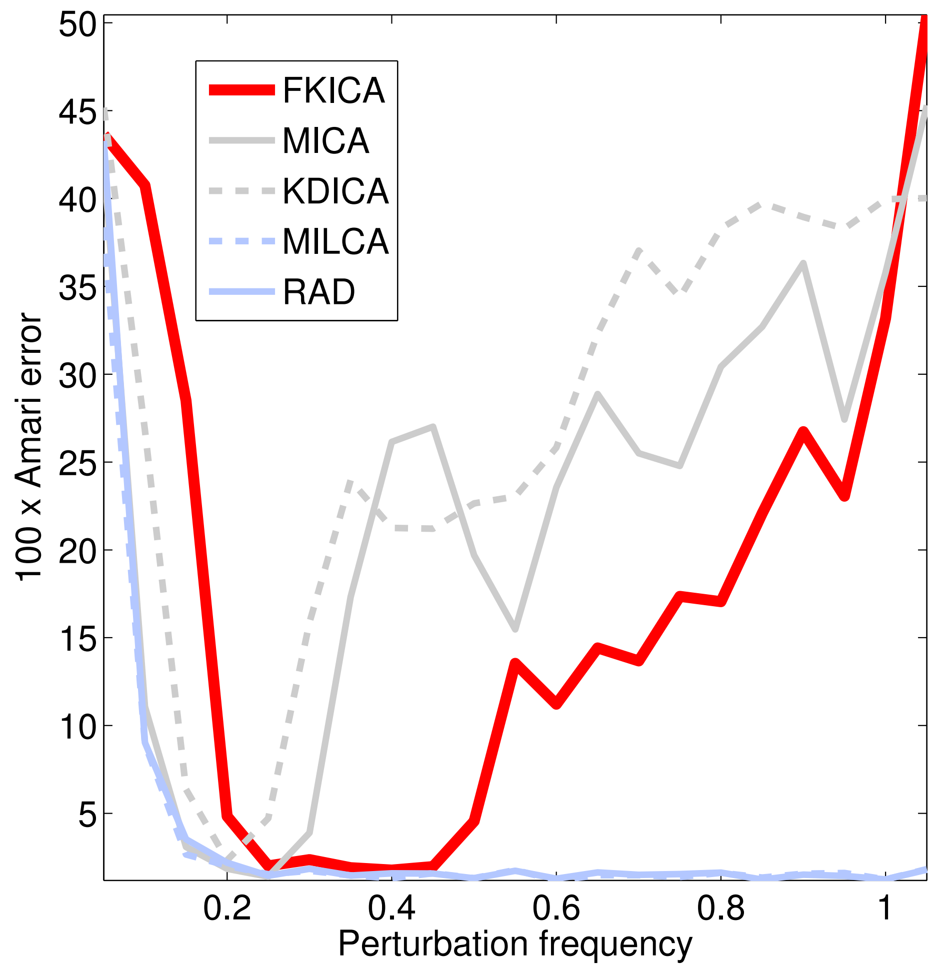
Fast entropy estimates: narrowest range



High frequency perturbations

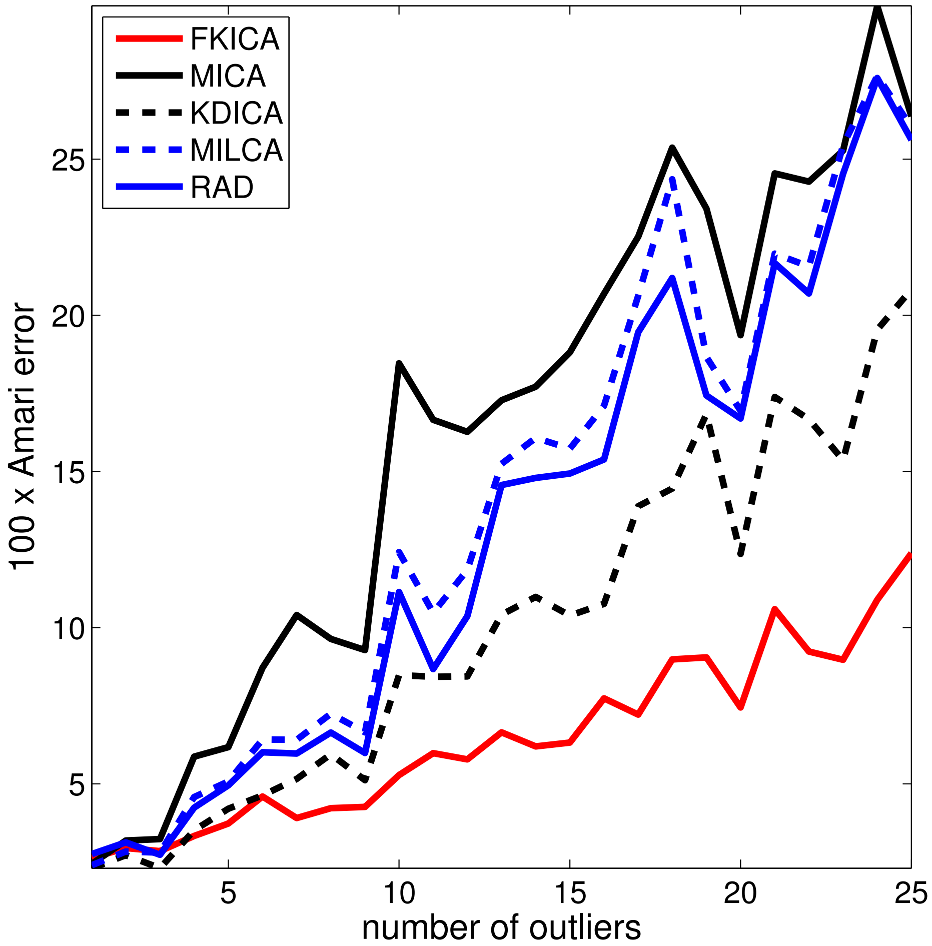
Fast Kernel ICA: peforms in between

(good performance/runtime tradeoff)



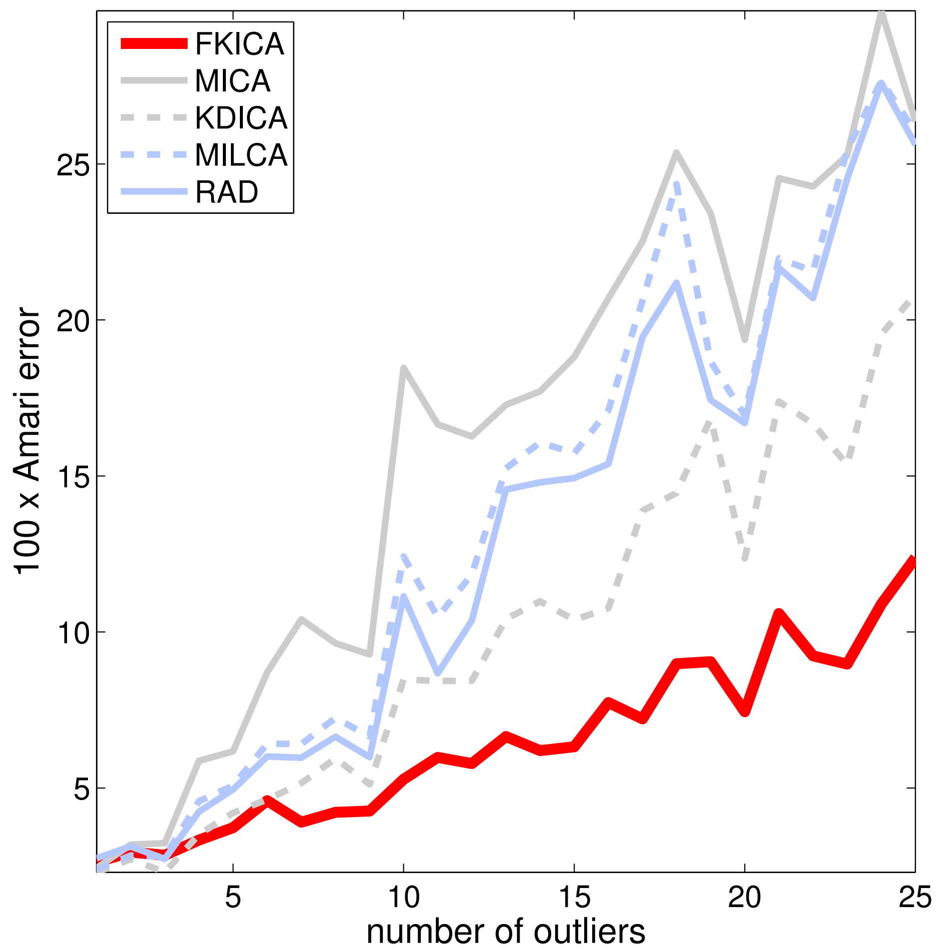
Outlier resistance

Two sources, outliers added to both *mixtures*



Outlier resistance

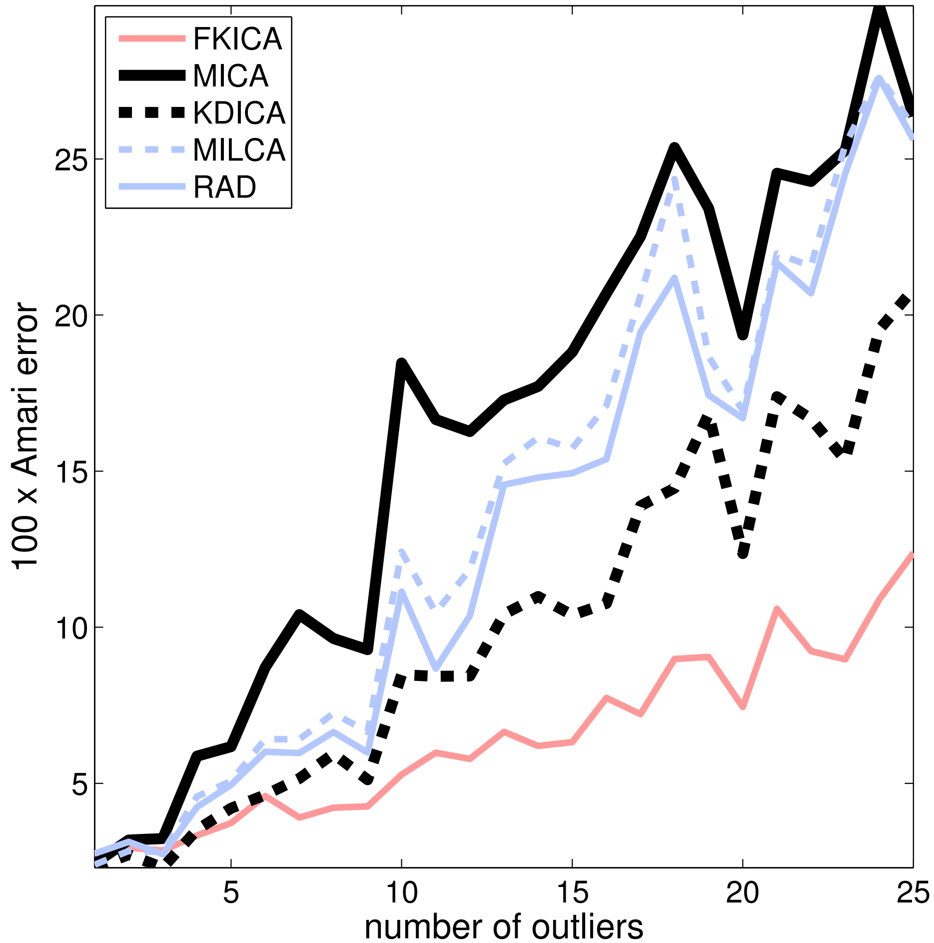
Kernel ICA performs best



Outlier resistance

Fast entropy estimates: less good

KDICA initialized with kernel ICA solution!



ICA algorithm choice

- Choosing kernel ICA approach
  - Fastest (by far): Fast ICA [Hyvärinen et al., 2001], Jade [Cardoso, 1998]
  - Good tradeoff between speed and performance: MICA [Pham, 2004]
  - Tricky cases (outliers, non-smooth sources): Fast KICA [Shen et al., 2007, 2009]
  - Small sample size: KGV very good [Bach and Jordan, 2002]

ICA algorithm choice

- Choosing kernel ICA approach
  - Fastest (by far): Fast ICA [Hyvärinen et al., 2001], Jade [Cardoso, 1998]
  - Good tradeoff between speed and performance: MICA [Pham, 2004]
  - Tricky cases (outliers, non-smooth sources): Fast KICA [Shen et al., 2007, 2009]
  - Small sample size: KGV very good [Bach and Jordan, 2002]
- Some further hints:
  - Use multiple restarts (non-convex)
  - Independence test to check answer

ICA algorithm choice

- Choosing kernel ICA approach
  - Fastest (by far): Fast ICA [Hyvärinen et al., 2001], Jade [Cardoso, 1998]
  - Good tradeoff between speed and performance: MICA [Pham, 2004]
  - Tricky cases (outliers, non-smooth sources): Fast KICA [Shen et al., 2007, 2009]
  - Small sample size: KGV very good [Bach and Jordan, 2002]
- Some further hints:
  - Use multiple restarts (non-convex)
  - Independence test to check answer
- Comparing (usually fixed contrast) algorithms:
  - One approach "better" than another?
  - Example: sources l very large, samples m small (wrt l), e.g. microarray data [Lee and Batzoglou, 2003]

Selected ICA references

- Start with Cardoso's excellent introduction [Cardoso, 1998], and the book by Hyvärninen *et al.* [Hyvärinen et al., 2001]
- Fast kernel ICA is described in [Shen et al., 2007, 2009]. Characteristic function-based ICA is described in [Eriksson and Koivunen, 2003, Chen and Bickel, 2005]. For earlier kernel ICA methods, see [Bach and Jordan, 2002, Gretton et al., 2005]
- Mutual information/entropy based: [Pham, 2004, Learned-Miller and Fisher III, 2003, Stögbauer et al., 2004, Chen, 2006]
- Classic algorithms for *time series* separation with second order methods (not covered much in this talk): [Molgedey and Schuster, 1994, Belouchrani et al., 1997]
- An important paper for optimising over orthogonal matrices: [Edelman et al., 1998]. The Newton-like method: [Hüper and Trumpf, 2004].

References

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

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.

M. Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A*, 23(6):1253–1268, 2006.

V. Calhoun, T. Adali, L. Hansen, J. Larsen, and J. Pekar. Ica of functional mri data: An overview. In *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 281–288, 2003.

J.-F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*, 86(10):2009–2025, 1998.

A. Chen. Fast kernel density independent component analysis. In *Sixth International Conference on ICA and BSS*, volume 3889, pages 24–31, Berlin/Heidelberg, 2006. Springer-Verlag.

A. Chen and P. J. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

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

J. Eriksson and K. Koivunen. Characteristic-function based independent component analysis. *Signal Processing*, 83(10):2195–2208, 2003.

A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

K. Hüper and J. Trumpf. Newton-like methods for numerical optimisation on manifolds. In *Proceedings of Thirty-eighth Asilomar Conference on Signals, Systems and Computers*, pages 136–139, 2004.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. McKeown, V. Iragui, and T. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–178, 2000.

K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *In Proc. Int. Conf. on Neural Information Processing (ICONIP)*, volume 2, pages 895–898, 1998.

E. G. Learned-Miller and J. W. Fisher III. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

S.I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biology*, 4(11):R76, 2003.

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

D.-T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 52(10):2690–2700, 2004.

H. Shen and K. Hüper. Newton-like methods for parallel independent component analysis. In *MLSP 16*, pages 283–288, Maynooth, Ireland, 2006.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel ICA using an approximate Newton method. In *AISTATS 11*, pages 476–483. Microtome, 2007.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 2009. In press.

H. Stögbauer, A. Kraskov, S. Astakhov, and P. Grassberger. Least dependent component analysis based on mutual information. *Phys. Rev. E*, 70(6): 066123, 2004.