

Mercer Kernels, Independence and Mutual Information

Francis R. Bach

fbach@cs.berkeley.edu

Michael I. Jordan

jordan@cs.berkeley.edu

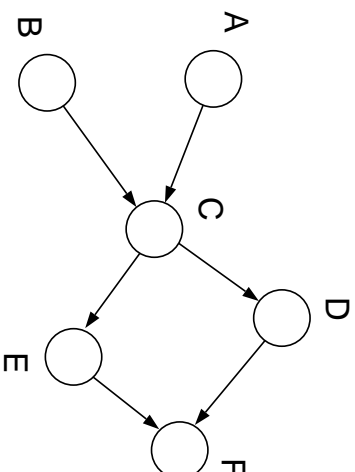
Computer Science and Statistics
University of California, Berkeley

Outline

- Semiparametric methods for graphical models
- The ICA problem
- Semiparametric approach to ICA based on canonical correlations in a reproducing kernel Hilbert space
- Tree-dependent component analysis

Semiparametric graphical models

- Graphical models require a distribution for each node in the graph:

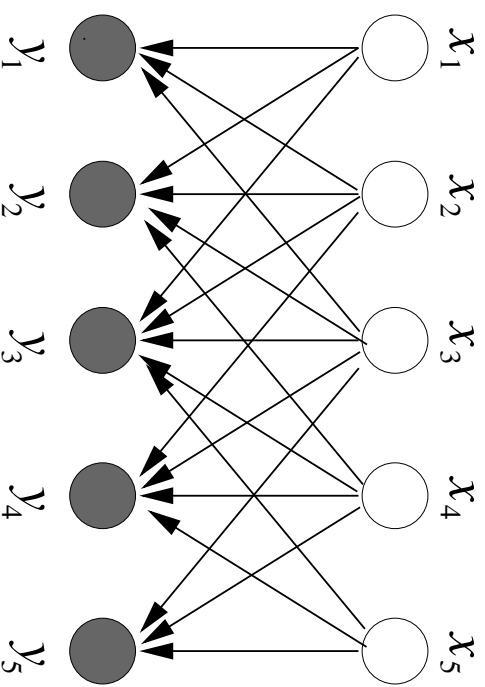


$$p(A, B, C, D, E, F) = p(A)p(B)p(C | A, B)p(D | C)p(E | C)p(F | D, E)$$

- In practice, however, we may not know what distribution to use for many of the variables in the problem
- The *semiparametric approach* leaves these distributions unspecified, and attempts to estimate the parameters of the specified distributions under all possible choices of unspecified distributions

Independent Component Analysis (ICA)

- Recover source vector $x = (x_1, \dots, x_m)^T$ from m unknown linear combinations $y = (y_1, \dots, y_m)^T$:

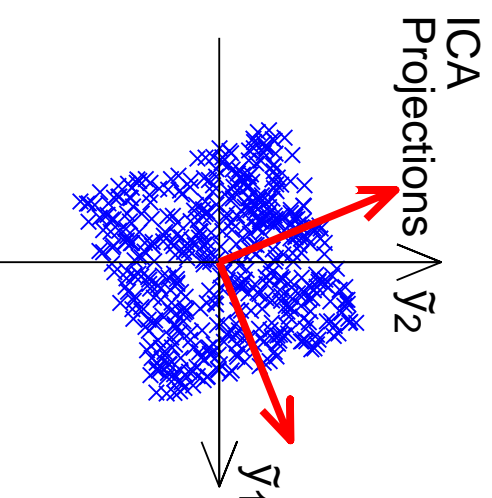
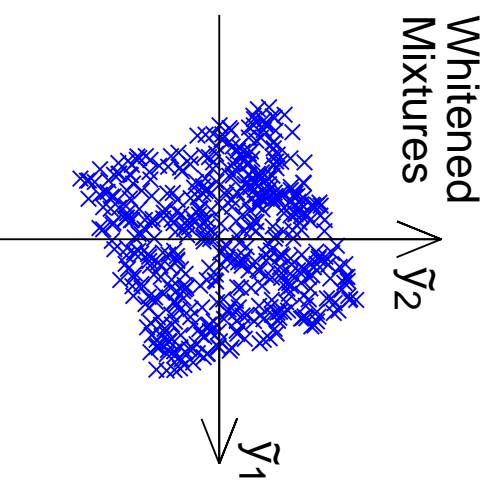
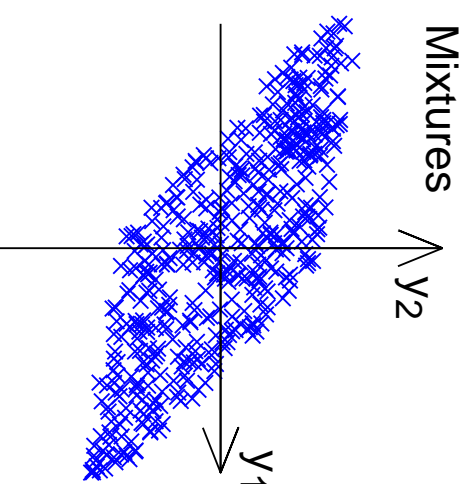
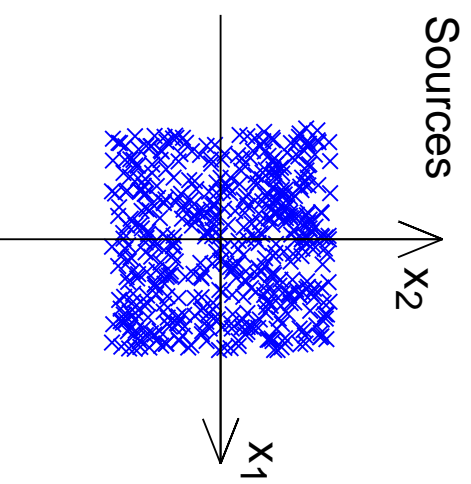


- We do not assume that we know the distributional form of the sources x_i
- Current graphical model technology doesn't know how to handle this

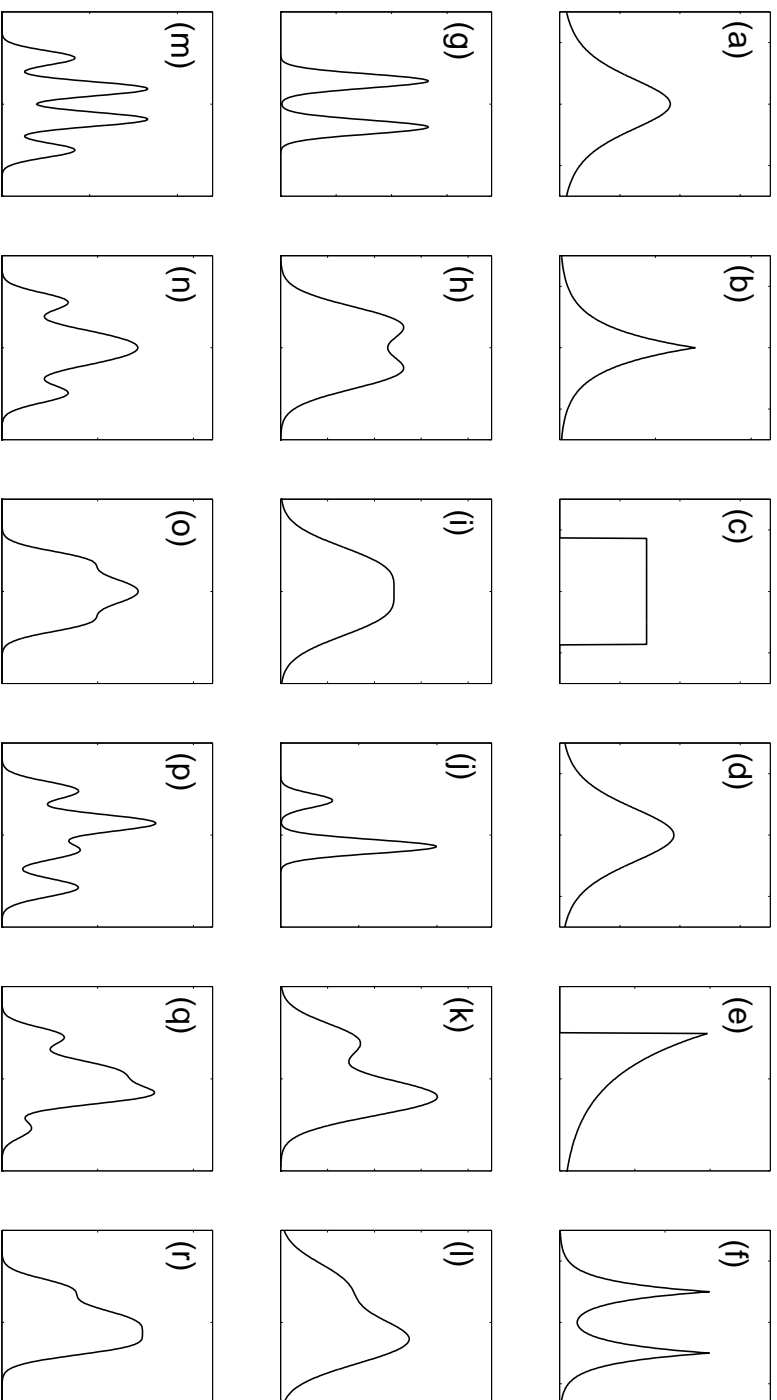
The ICA model

- Recover latent random vector $x = (x_1, \dots, x_m)^T$ from m unknown linear combinations $y = (y_1, \dots, y_m)^T$.
- Model : $y = Ax$, where both x and A are unknown
- Goal : estimate A from samples $\{y^1, \dots, y^N\}$ in \mathbb{R}^m
- This is a *semiparametric* statistical model
 - A is the parametric part, which we wish to estimate
 - the distribution of x is the nonparametric part—an infinite-dimensional nuisance parameter
- Identifiable for non-Gaussian source distributions
- Numerous applications (cf. ICA annual conference)

An example of ICA



Source distributions



Approaches to estimation of the ICA model

- Treat as a parametric model
 - make an assumption about the density of x
 - form the parametric likelihood
 - show that the approach is robust to the assumption
- Treat as a semiparametric model
 - the semiparametric likelihood is equivalent to the mutual information
 - approximate the mutual information
 - leads to cumulants and other expectations
 - compute and minimize empirical versions of these expectations
- General concept of a *contrast function*
 - provides a characterization of independence (expectation is equal to zero if and only if the estimated components are independent)
 - a general focus on choice of a single nonlinear function, or a small parameterized set of nonlinear functions

KernelICA

- Two new contrast functions based on *reproducing kernel Hilbert space* ideas
 - rather than using a single fixed nonlinear function (such as a cumulant), use an entire function space of nonlinear functions
 - requires the solution of a *generalized eigenvector problem* as an inner loop
- One contrast function uses first eigenvalue
 - reproducing kernel Hilbert spaces and *independence*
- The other contrast function uses the full spectrum
 - reproducing kernel Hilbert spaces and *mutual information*
- Goal: a flexible, robust, extensible ICA methodology

Mercer kernels

- Map each data point x into a “feature space” \mathcal{F} via a nonlinear function Φ :

$$x \xrightarrow{\Phi} \Phi(x)$$

(in our case, there will be m such feature spaces—one for each component of the recovered source vector)

- where Φ is chosen such that inner products in \mathcal{F} can be evaluated compactly:

$$\langle \Phi(x), \Phi(y) \rangle = K(x, y)$$

for some *kernel function* $K(x, y)$ —this is the so-called “kernel trick”

- This can be done for any function $K(x, y)$ that is positive semidefinite
 - e.g., $K(x, y) = e^{-\frac{1}{2}\|x-y\|^2}$, monomials, string kernels, tree kernels, etc.

Kernelization

- Any algorithm based only on inner products can be “kernelized”
 - Conceptually, map all data points x^i to the feature space via $x^i \rightarrow \Phi(x^i)$, and apply the standard algorithm
 - Computationally, use the standard algorithm, but replace any inner product $\langle \Phi(x^i), \Phi(x^j) \rangle$ with a kernel evaluation $K(x^i, x^j)$
- Yields nonlinear versions of:
 - PCA
 - Fisher discriminant
 - perceptron
 - maximal margin classifier (yields the “support vector machine”)
- Relevance to ICA?

The reproducing kernel Hilbert space (RKHS) perspective

- The “feature space” is a *function space*:

$$x \xrightarrow{\Phi} K(\cdot, x)$$

- Take the span and complete to a Hilbert space:

$$\mathcal{F} = \overline{\text{span}}(\{K(\cdot, x) : x \in \mathcal{X}\})$$

- The *reproducing property* of the kernel:

$$\langle K(\cdot, x), f(\cdot) \rangle = f(x) \quad \forall f \in \mathcal{F}$$

- A suggestive fact from L_2 (which is a Hilbert space but not an RKHS):

$$\int \delta(y, x) f(y) dy = f(x) \quad \forall f \in L_2$$

The reproducing kernel Hilbert space (RKHS) perspective (cont.)

- Use the reproducing property evaluated on the kernel itself:

$$\langle K(\cdot, x), K(\cdot, y) \rangle = K(x, y)$$

- Recalling that $\Phi(x) = K(\cdot, x)$:

$$\langle \Phi(x), \Phi(y) \rangle = K(x, y)$$

and thus the RKHS perspective yields a particular (elegant) instantiation of the kernel trick (it is a coordinate-free instantiation)

The \mathcal{F} -correlation

- Measure the dependence between random variables x_1 and x_2 using the correlation of functions of the variables:

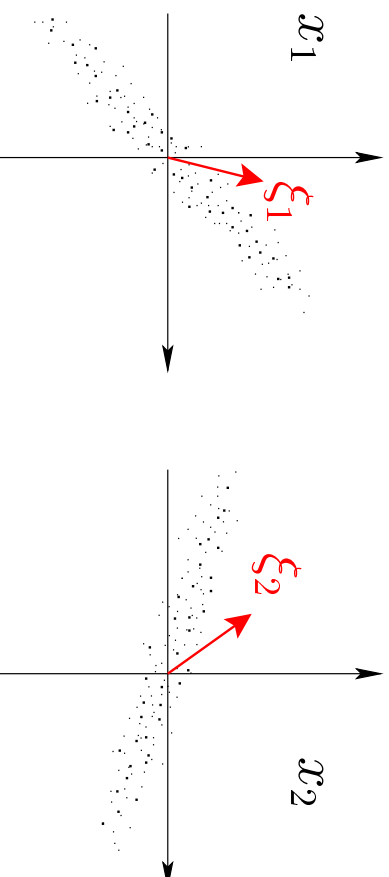
$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} \text{corr}(f_1(x_1), f_2(x_2)) = \max_{f_1, f_2 \in \mathcal{F}} \frac{\text{cov}(f_1(x_1), f_2(x_2))}{(\text{var } f_1(x_1))^{1/2} (\text{var } f_2(x_2))^{1/2}}.$$

- If \mathcal{F} is “big enough”, then $\rho_{\mathcal{F}} = 0$ if and only if x_1 and x_2 are independent.
- When \mathcal{F} is a reproducing kernel Hilbert space, i.e., $f(x) = \langle \Phi(x), f \rangle$, then

$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} \text{corr}(\langle \Phi(x_1), f_1 \rangle, \langle \Phi(x_2), f_2 \rangle)$$

$\Rightarrow \rho_{\mathcal{F}}$ is the first **canonical correlation** between $\Phi(x_1)$ and $\Phi(x_2)$

Canonical correlation analysis



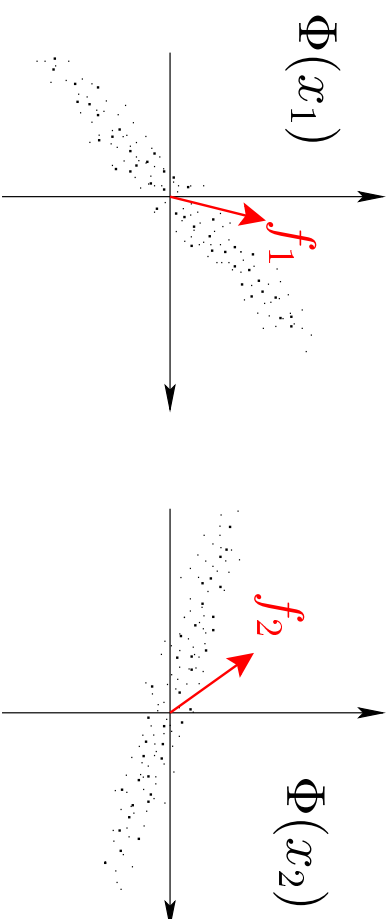
- Given two multivariate random variables x_1 and x_2 , find the pair of directions ξ_1, ξ_2 with maximal correlation of the projections:

$$\rho(x_1, x_2) = \max_{\xi_1, \xi_2} \text{corr}(\xi_1^T x_1, \xi_2^T x_2) = \max_{\xi_1, \xi_2} \frac{\xi_1^T C_{12} \xi_2}{(\xi_1^T C_{11} \xi_1)^{1/2} (\xi_2^T C_{22} \xi_2)^{1/2}}$$

- Generalized eigenvalue problem:

$$\begin{pmatrix} 0 & C_{12} \\ C_{21} & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \rho \begin{pmatrix} C_{11} & 0 \\ 0 & C_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}.$$

Canonical correlation analysis in function space



- Given two random variables x_1 and x_2 and their images in feature space, $\Phi(x_1)$ and $\Phi(x_2)$, find the pair of functions f_1, f_2 yielding maximal correlation of the projections:

$$\rho_{\mathcal{F}} = \max_{f_1, f_2 \in \mathcal{F}} \text{corr}(\langle \Phi(x_1), f_1 \rangle, \langle \Phi(x_2), f_2 \rangle)$$

- This turns out to be a valid contrast function in the case of Gaussian kernels (it is equal to zero if and only if x_1 and x_2 are independent)

Kernel canonical correlation analysis

- Given data $\{x_1^i\}$ and $\{x_2^i\}$, form the **Gram matrices** K_1 , K_2 :

$$K_1 = \left\{ K(x_1^i, x_1^j) \right\}_{ij} \quad K_2 = \left\{ K(x_2^i, x_2^j) \right\}_{ij}$$

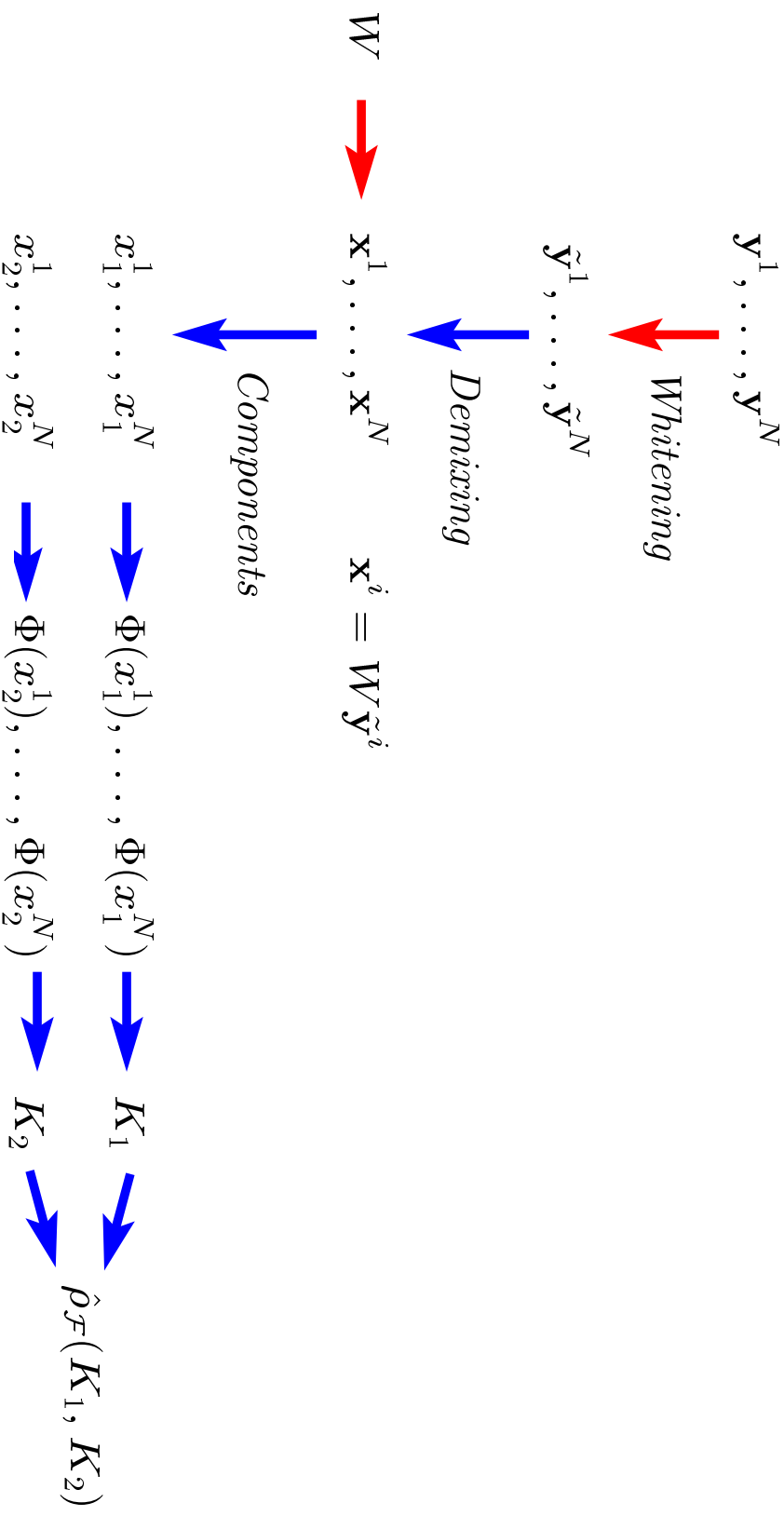
- Canonical correlation analysis in \mathcal{F} reduces to:

$$\hat{\rho}_{\mathcal{F}}(K_1, K_2) = \max_{\alpha_1, \alpha_2 \in \mathbb{R}^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\alpha_1^T K_1^2 \alpha_1)^{1/2} (\alpha_2^T K_2^2 \alpha_2)^{1/2}}$$

- which is solved by finding the maximal generalized eigenvalue of:

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

KERNELICA algorithm (for $m = 2$ sources)



- Minimize $-\frac{1}{2} \log \hat{\rho}_{\mathcal{F}}$ with respect to W

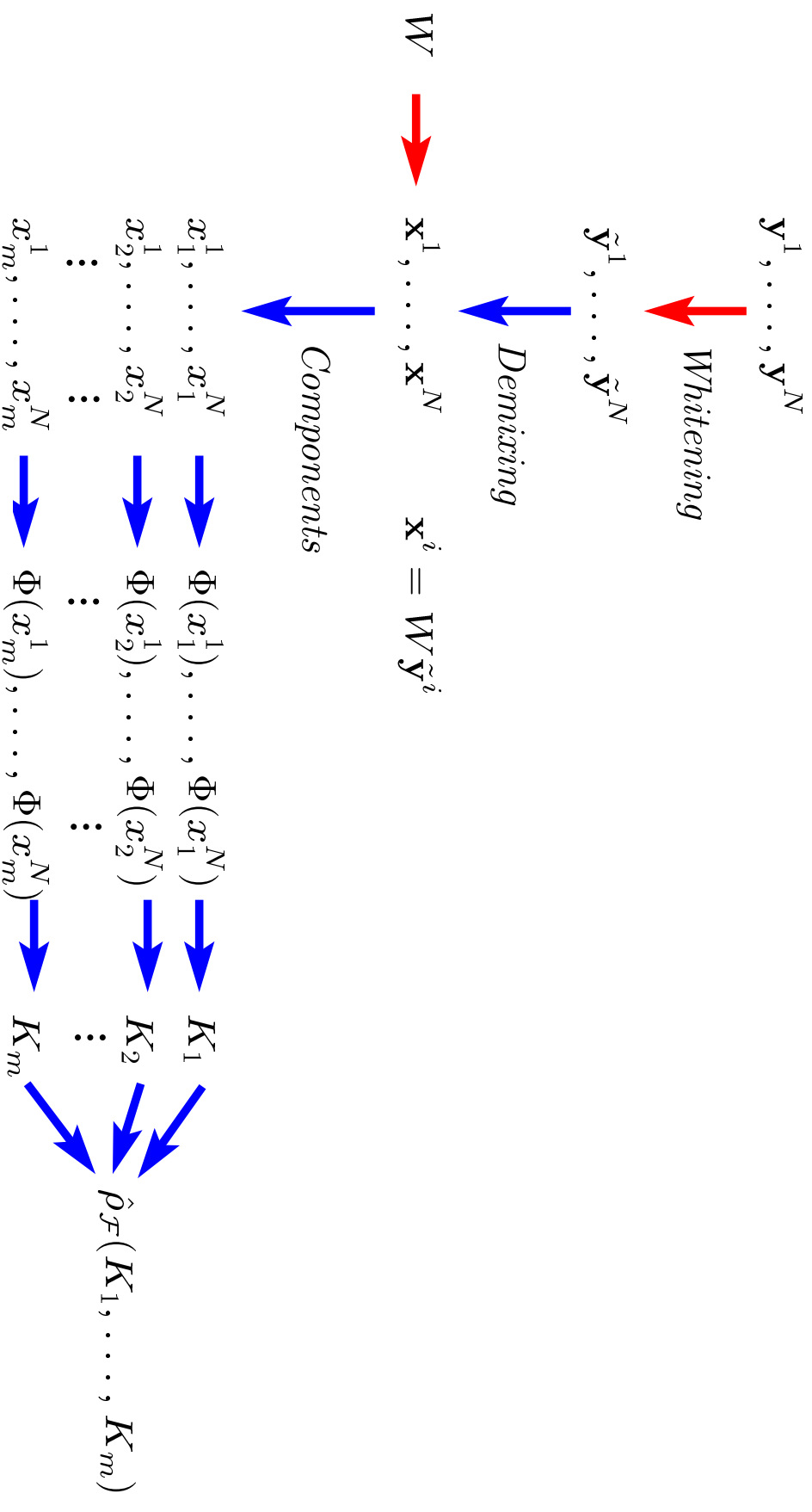
Generalization to $m > 2$ variables

- Using a generalization of CCA to $m > 2$ variables, find the smallest generalized eigenvalue of:

$$\begin{pmatrix} K_1^2 & K_1 K_2 & \cdots & K_1 K_m \\ K_2 K_1 & K_2^2 & \cdots & K_2 K_m \\ \vdots & \vdots & \ddots & \vdots \\ K_m K_1 & K_m K_2 & \cdots & K_m^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \lambda \begin{pmatrix} K_1^2 & 0 & \cdots & 0 \\ 0 & K_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_m^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

- The corresponding population functional is again a valid contrast function

KERNELICA algorithm



Generalized variance

- For Gaussian random variables, the full canonical correlation spectrum gives the mutual information:

$$\begin{aligned} M(x_1, x_2) &= -\frac{1}{2} \log \left(\frac{\det C}{\det C_{11} \det C_{22}} \right) \\ &= -\frac{1}{2} \log \det(I - AA^T) \text{ where } A = C_{11}^{-1/2} C_{12} C_{22}^{-1/2} \\ &= -\frac{1}{2} \sum_{i=1}^p \log(1 - \rho_i^2) \end{aligned}$$

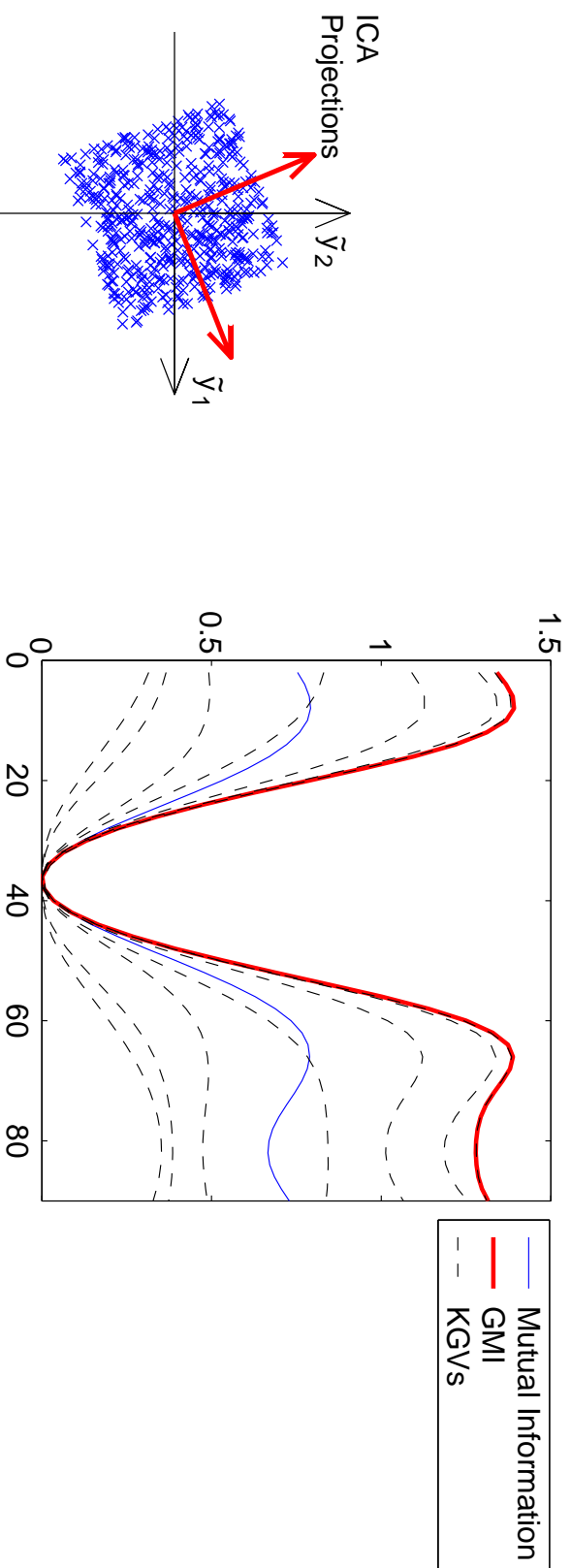
- Generalized variance $\doteq \frac{\det C}{\det C_{11} \det C_{22}}$

Kernel generalized variance

- Kernel generalized variance: $D(K_1, K_2) = \frac{\det \begin{pmatrix} K_1^2 & K_1 K_2 \\ K_2 K_1 & K_2^2 \end{pmatrix}}{\det K_1^2 \det K_2^2}$
- New contrast function: $M_D = -\frac{1}{2} \log D$

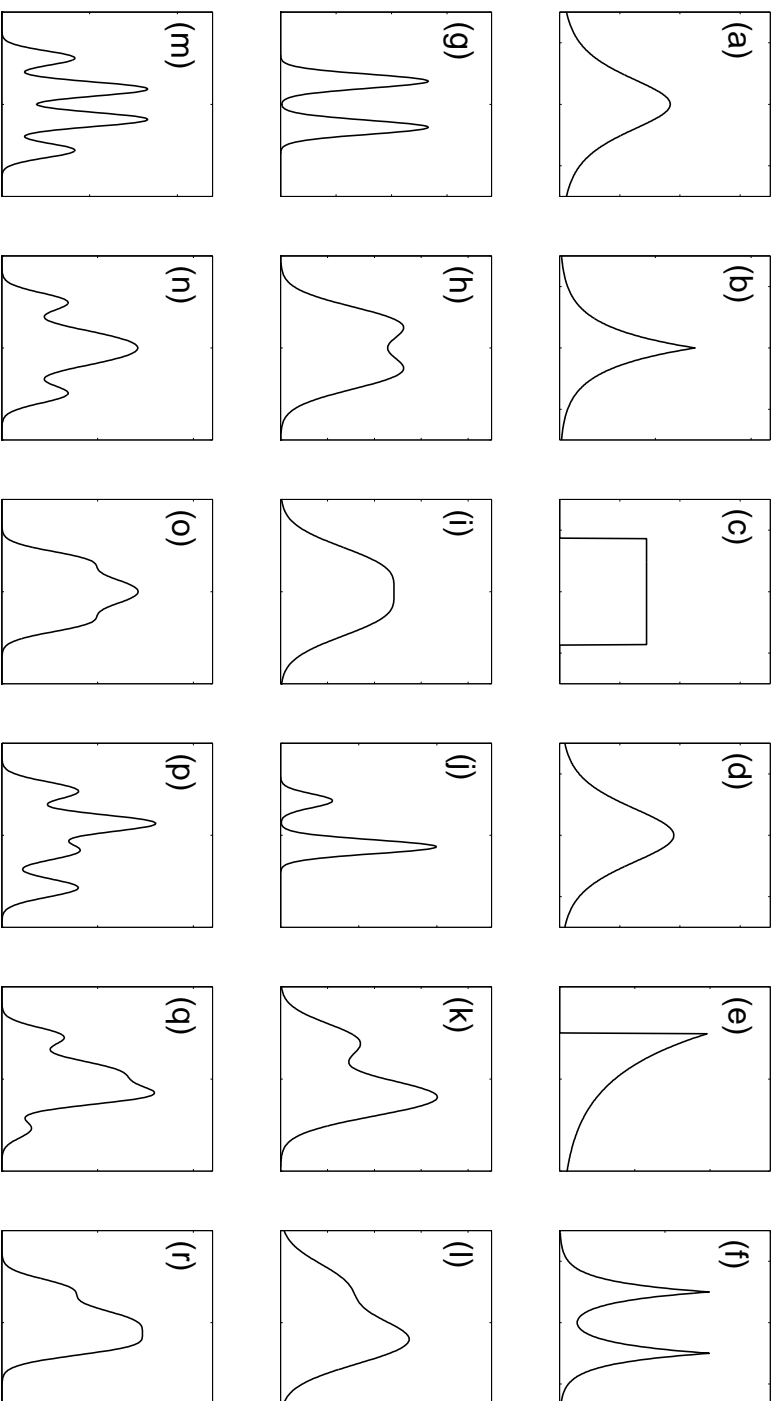
Kernel generalized variance and mutual information

- Translation-invariant kernels: $K(x, y) = k\left(\frac{x-y}{\sigma}\right)$
- When σ tends to zero, $M_D(\sigma)$ has a limit $\mathcal{I}(x_1, x_2)$
- $\mathcal{I}(x_1, x_2)$ is equal to the mutual information up to second order “around independence”



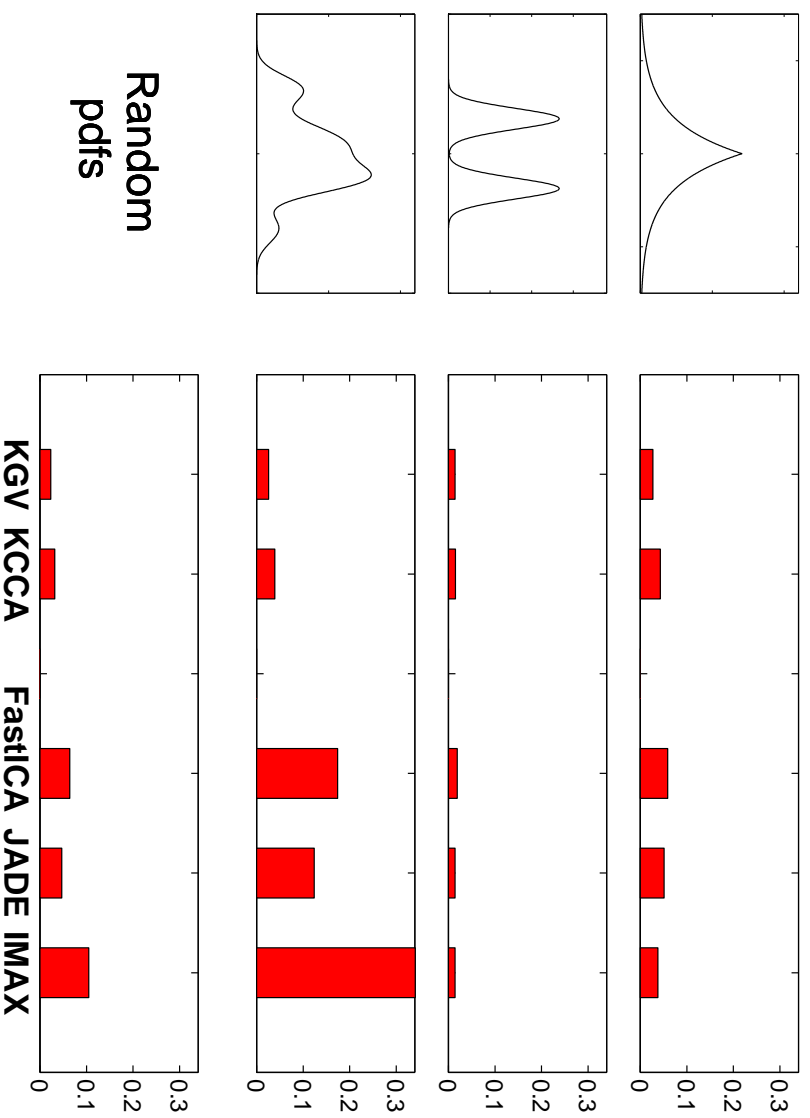
Empirical results

- Source distributions



Robustness to source distributions

- Comparison with other algorithms: FastICA, Jade, Extended Infomax
- Amari error: standard ICA distance from true sources



pdfs	FastICA	JADE	IMAX	KCCA	KGV
a	4.6	4.4	3.1	4.7	3.3
b	5.9	5.1	3.8	4.3	2.7
c	2.4	1.6	2.0	2.7	1.7
d	1.9	1.4	1.4	1.5	1.4
e	5.2	4.0	3.3	4.5	3.4
f	10.7	7.1	6.8	10.2	9.6
g	7.8	6.0	54.9	1.5	1.4
h	6.2	4.2	3.8	3.0	2.6
i	10.7	8.1	11.4	5.0	4.4
j	5.9	5.0	7.1	7.7	6.0
k	5.4	4.2	4.5	1.7	1.4
l	3.3	2.6	1.5	1.5	1.3
m	4.0	2.7	4.4	2.3	1.3
n	5.5	4.0	28.9	2.9	1.8
o	4.1	2.9	3.9	5.0	3.3
p	3.7	2.8	10.3	2.3	1.8
q	17.4	12.4	41.1	3.9	2.6
r	6.2	4.6	5.0	4.2	3.1
mean	6.2	4.6	11.0	3.8	2.9
rand	6.4	4.7	10.5	3.2	2.3

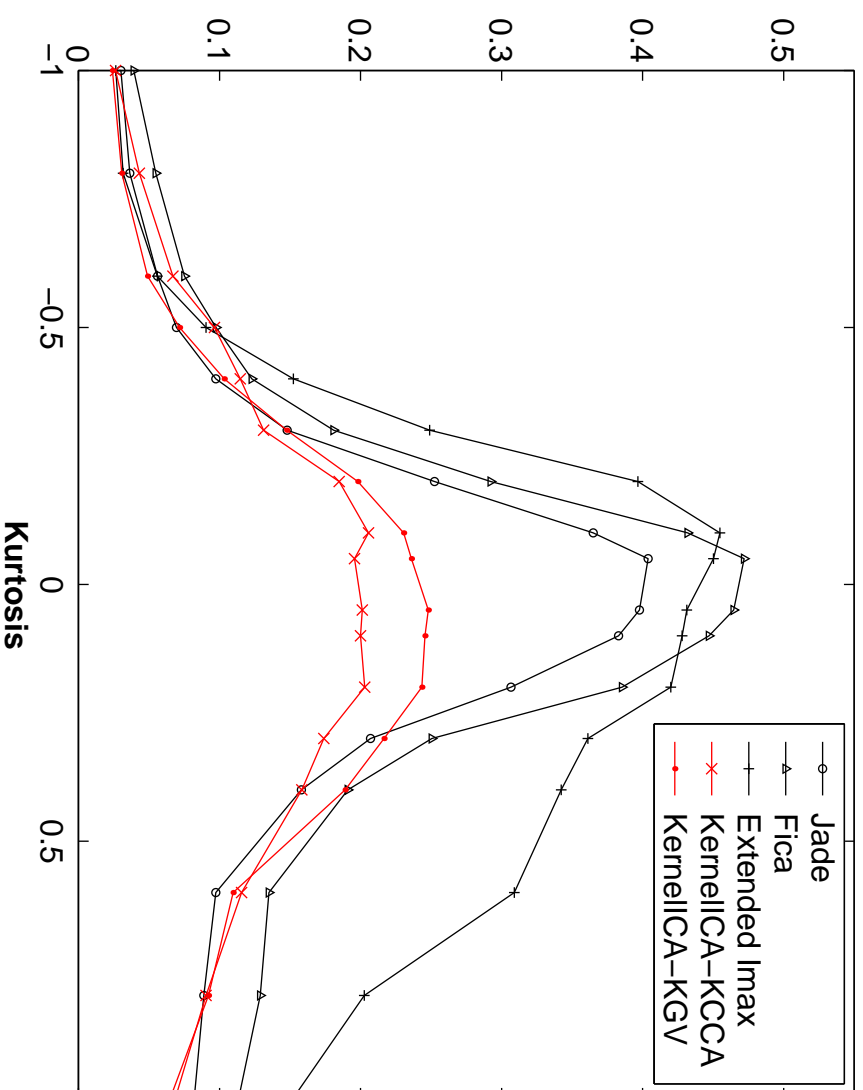
More sources

m	N	# repl	FastICA	JADE	IMAX	KCCA	KGV
2	250	1000	11	9	28	7	6
	1000	1000	6	5	11	3	2
4	1000	100	19	15	35	13	9
	4000	20	9	6	22	5	3
8	1000	20	34	39	87	31	32
	4000	20	19	16	51	12	7
16	4000	10	40	36	99	22	13

The Amari errors (multiplied by 100) for m components with N samples: m (generally different) pdfs were chosen uniformly at random among the 18 possible pdfs. The results are averaged over the stated number of replications.

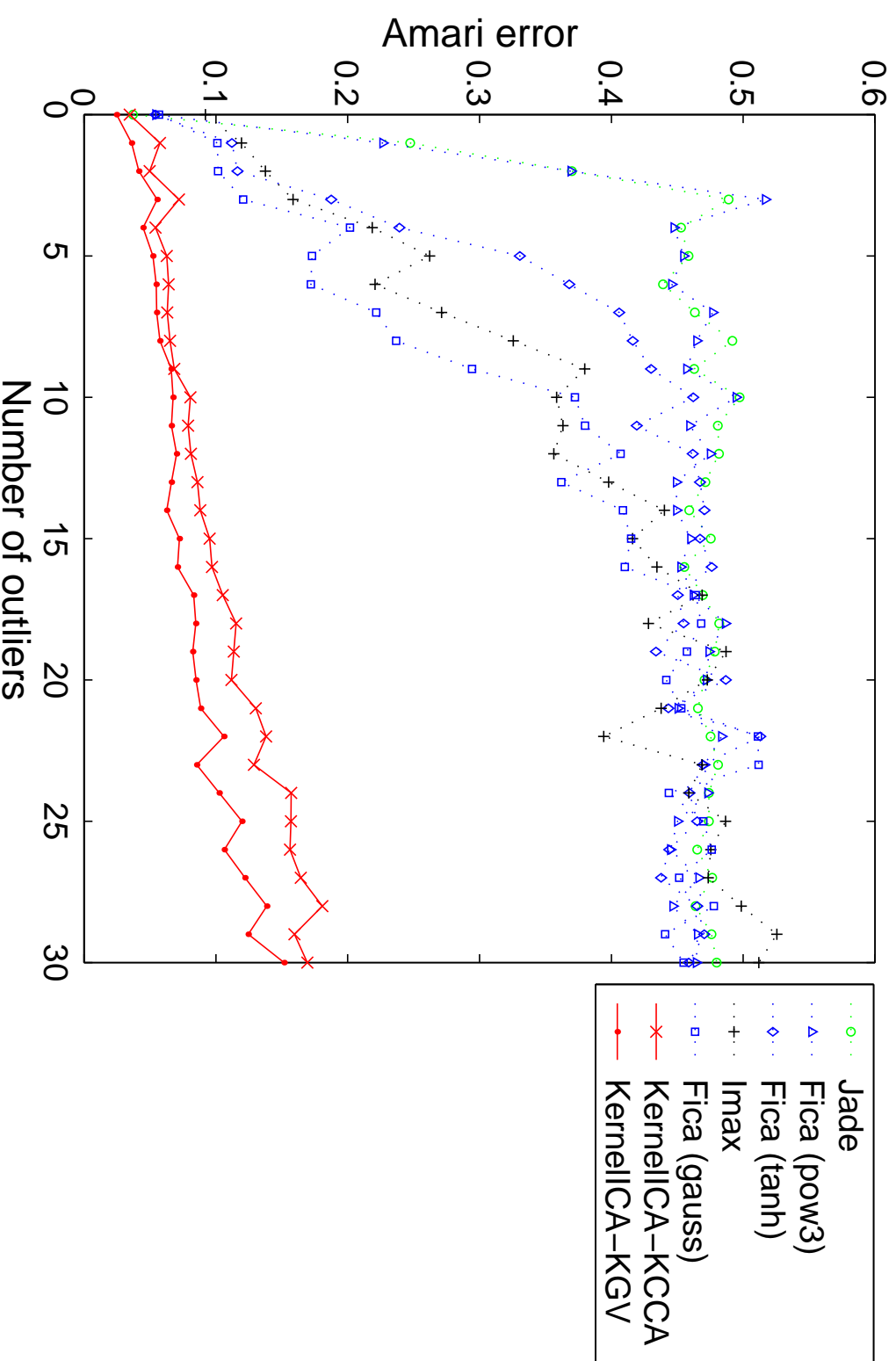
Robustness to Gaussianity

- When sources are nearly Gaussian, ICA is more challenging



Robustness to outliers

- Large values added to randomly selected data points



Tree-dependent component analysis

- Tree-dependent component analysis
 - Find a linear transform W so that Wx factors according to a tree T
- T -mutual information:

$$\begin{aligned} I^T(x) &\triangleq \min_{q \in \mathcal{D}^T} D(p||q) = D(p||p_T) \\ &= I(x_1, \dots, x_m) - \sum_{(u,v) \in T} I(x_u, x_v) \end{aligned}$$

- Use the KGV to approximate and optimize
 - Chou-Liu inner loop
- Yields a general multivariate density estimator that requires only bivariate density estimation

Summary and open questions

- New contrast functions for ICA
 - more computationally-demanding
 - more flexible and robust
- Link between kernel methods and mutual information
 - expansion of mutual information around *independence*, not around *Gaussianity*
- Extensions from tree-dependent components to more general graph-dependent components?
- Kernels on strings, trees, etc. \implies ICA on strings, trees, etc.?

Relationship between likelihood and mutual information

- We wish to minimize the KL divergence between the true distribution p^* and the model p : $D(p^*(y) \parallel p(y))$.
- Parameterize the model using $W = A^{-1}$, so that $x = Wy$. Invariance of the KL divergence implies our problem is equivalent to that of minimizing $D(p^*(x) \parallel p(x))$.
- Let $\tilde{p}(x)$ denote the joint probability distribution obtained by taking the product of the marginals of $p^*(x)$. We have:

$$D(p^*(x) \parallel p(x)) = D(p^*(x) \parallel \tilde{p}(x)) + D(\tilde{p}(x) \parallel p(x)),$$

for any distribution $p(x)$ with independent components.

- Consequently, for a given A , the minimum over all possible $p(x)$ is attained precisely at $p(x) = \tilde{p}(x)$, and the minimal value is $D(p^*(x) \parallel \tilde{p}(x))$, the mutual information between the components of $x = Wy$.

Cumulant-based approach

- Recall the Gram-Charlier expansion, appropriate for pdf's near Gaussianity:

$$p(x) \approx \phi(x) \left(1 + \kappa_3(X) \frac{H_3(x)}{3!} + \kappa_4(X) \frac{H_4(x)}{4!} \right)$$

- Plug into the mutual information formula, manipulate, and obtain an approximation of mutual information in terms of the third and fourth cumulants
- Use this approximation as a contrast function for ICA (i.e., estimate the cumulants from data and minimize the approximation with respect to W)
- Not a very well-motivated heuristic

Regularization

- Numerical and statistical issues

$$\begin{pmatrix} (K_1 + \kappa I)^2 & K_1 K_2 & \dots & K_1 K_m & \alpha_1 \\ K_2 K_1 & (K_2 + \kappa I)^2 & \dots & K_2 K_m & \alpha_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ K_m K_1 & K_m K_2 & \dots & (K_m + \kappa I)^2 & \alpha_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

$$= \lambda \begin{pmatrix} (K_1 + \kappa I)^2 & 0 & \dots & 0 & \alpha_1 \\ 0 & (K_2 + \kappa I)^2 & \dots & 0 & \alpha_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (K_m + \kappa I)^2 & \alpha_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

Running time complexity

- Naive implementation: $O(m^3 N^3)$
- We manage to reduce to linear time complexity in N :
 - Very low rank approximations: $K = GG^T$ where G is $N \times M$ and $M \ll N$.
 - Possible because Gram matrices have geometrically decaying spectrum
 - Symmetric positive semidefiniteness:
incomplete Cholesky Decomposition can be used
 - complexity of decomposition: $O(M^2 N)$
- Final complexity: $O(m^2 N)$

Derivatives

- Needed if contrast functions are to be used in gradient-based optimization algorithms.
- Major challenge: $\partial_t K$ is symmetric but not semi-definite positive or negative
 \Rightarrow incomplete Cholesky cannot be used.

- In the case of Gaussian kernels, $\partial_t K$ is the “Gram” matrix obtained from the function $f(x) = x^2 e^{-x^2/2\sigma^2}$. $\hat{f}(\omega)$ is not positive but:

$$\begin{aligned} f(x) &= \sigma^2 e^{-x^2/2\sigma^2} & - & e^{-x^2/2\sigma^2} (\sigma^2 - x^2) \\ \hat{f}(\omega) &= \sigma^2 e^{-\omega^2\sigma^2/2} & - & \sigma^2 \omega^2 \sqrt{2\pi\sigma} e^{-\omega^2\sigma^2/2} \\ &> 0 & & > 0 \end{aligned}$$

$\Rightarrow \partial_t K$ is a difference between two low-rank semi-definite matrices.

Optimization

- Stiefel manifolds (Edelman, et al., 1999)
- Conjugate gradient, with line search along geodesics