Kernels, Independence and Dimension Reduction

Michael I. Jordan

Computer Science Division and Department of Statistics
University of California, Berkeley

http://www.cs.berkeley.edu/∼jordan

Joint work with: Francis Bach and Kenji Fukumizu

Mercer kernels

- Positive definite kernel $k(x, x')$:

$$\begin{aligned} \text{polynomial} \quad k(x, x') &= (\langle x, x' \rangle + c)^d \\ \text{RBF} \quad k(x, x') &= \exp\left(-\|x - x'\|^2 / (2\,\sigma^2)\right) \\ \text{string kernels} \quad k(x, x') &= \text{number of overlapping substrings in } x \text{ and } x' \end{aligned}$$

- Expansion in eigenfunctions: $k(x, x') = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')$

- "Kernelization":
  - transform x to a "feature space" according to $x \xrightarrow{\Phi} (\phi_1(x), \phi_2(x), \ldots)$
  - inner products in the "feature space" are kernel evaluations $k(x, x')$

- Cottage industry of "kernelizations" (maximum margin, Fisher discriminant, PCA, canonical correlations, logistic regression, Kalman filter, etc)

Reproducing kernel Hilbert space

- Given a kernel, consider the mapping

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

- Take the span and complete to a Hilbert space:

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

- The *reproducing property* of the kernel:

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

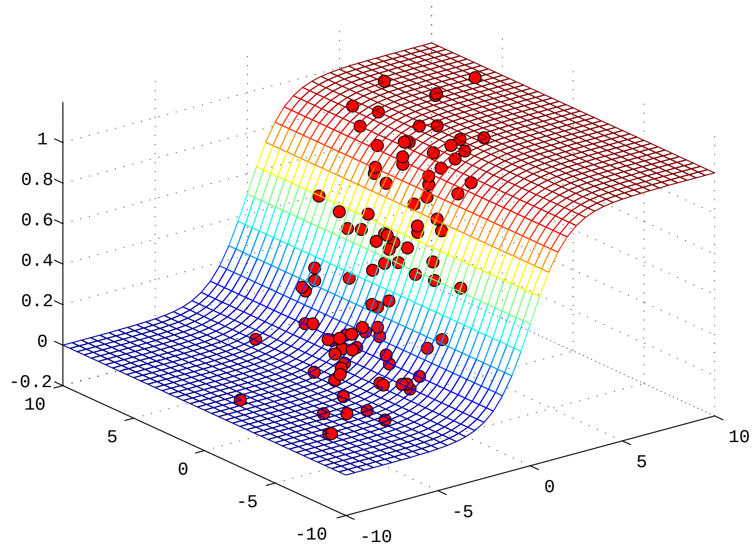
- Yields a coordinate-free interpretation of Mercer's theorem:

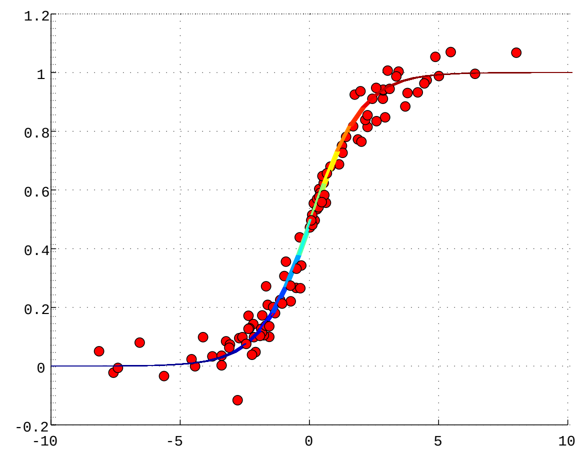
$$k(x, x') = \langle k(\cdot, x), k(\cdot, x')\rangle$$

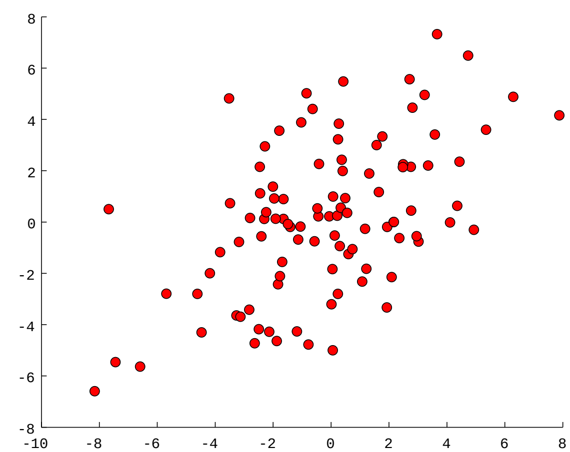
This talk

- A different usage of reproducing kernel Hilbert spaces (RKHS)
- Use RKHS to characterize mutual independence and conditional independence
  - a nonparametric, variational characterization
  - relationship to information theory
- Use this characterization to derive contrast functions for various semiparametric estimation problems:
  - dimension reduction for regression
  - independent component analysis
  - tree-dependent component analysis

Dimension reduction for regression





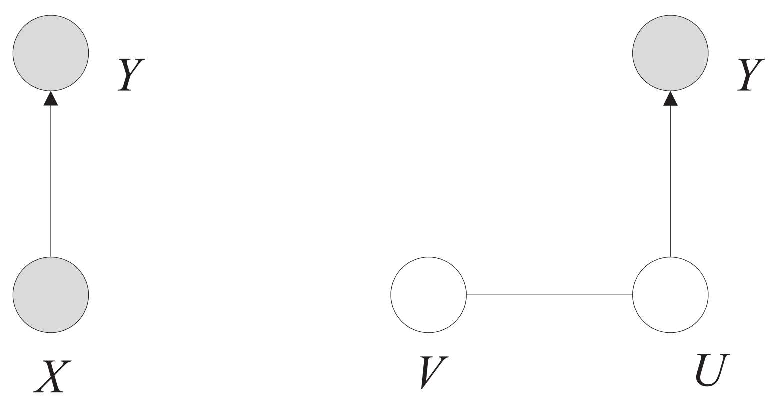


$$Y = \frac{1}{1+\exp(-X_1)} + N(0, 0.1^2)$$

Effective subspace = direction of X_1

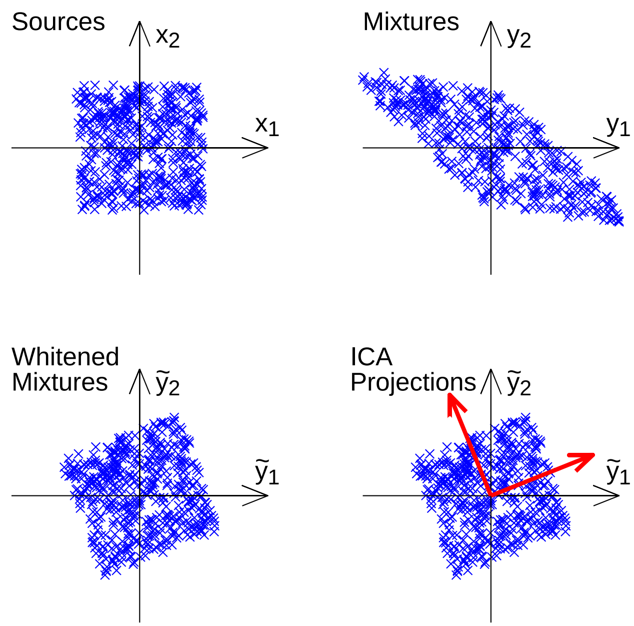
Dimension reduction for regression (cont.)

- Nonparametric regression problem: $p(y|x)$
- Effective subspace: determined by a matrix B such that $p(y|x) = p(y|B^T x)$
  - now it's a semiparametric problem
- Conditional independence interpretation:
  - let $(U, V) = (B^T x, C^T x)$, for $R = (B, C)$ an orthogonal matrix
  - $p(y|x) = p(y|B^T x)$ if and only if $Y \perp\!\!\!\perp V | U$



Independent component analysis (ICA)

- Model : $y = Ax$, where x is a latent *source* vector
- Goal : estimate A from samples $\{y^1, \dots, y^N\}$

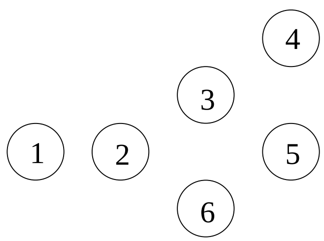
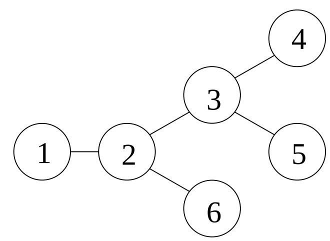


- The components of x are assumed independent, but the distribution of x is otherwise unknown

From ICA to TCA

- Model : $y = Ax$

- Instead of assuming mutual independence, assume that the underlying source distribution factors according to a tree

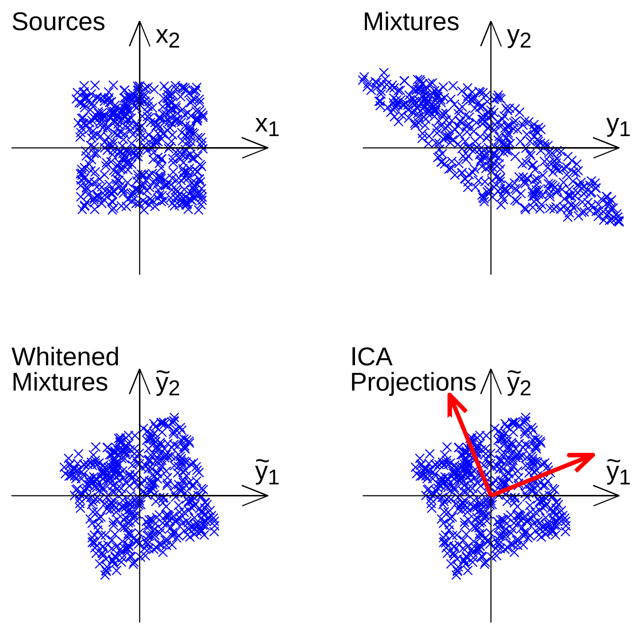
Mutual independence	Tree-dependence
	
$p_{ML}(x) = \prod_{i=1}^{m} p(x_i)$	$p_{ML}(x) = \prod_{(u,v)\in T} \frac{p(x_u, x_v)}{p(x_u)p(x_v)} \prod_{u=1}^{m} p(x_u)$

- Can be interpreted as a set of conditional independence statements

- This is again a semiparametric estimation problem

Independent component analysis (ICA)

- Model : $y = Ax$, where x is a latent *source* vector
- Goal : estimate A from samples $\{y^1, \ldots, y^N\}$



- The components of x are assumed independent, but the distribution of x is otherwise unknown

Estimation of the ICA model

$\mathbf{y}^1, \ldots, \mathbf{y}^N$

↓ *Whitening*

$\tilde{\mathbf{y}}^1, \ldots, \tilde{\mathbf{y}}^N$

↓ *Demixing*

W → $\mathbf{x}^1, \ldots, \mathbf{x}^N$ $\quad \mathbf{x}^i = W\mathbf{y}^i$

↓ *Components*

$x_1^1, \ldots, x_1^N$

$x_2^1, \ldots, x_2^N$

$\vdots \quad \vdots$

$x_m^1, \ldots, x_m^N$

→ *Contrast function* $\mathcal{I}(\{x_1^i\}, \ldots, \{x_m^i\})$

Contrast functions

- Standard ICA contrast functions
  - Mutual information
  - Edgeworth expansions
  - Ad-hoc nonlinearities f_1 and f_2: $\mathcal{J}(x_1, x_2) = E(f_1(x_1)f_2(x_2))$.
- Our approach: a contrast function based on an RKHS characterization of independence

The $\mathcal{F}$-correlation

- Measures dependence between x_1 and x_1 using correlation of functions of the variables, $f_1(x_1)$ and $f_2(x_2)$, for f_1, f_2 belonging to a function space $\mathcal{F}$:

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

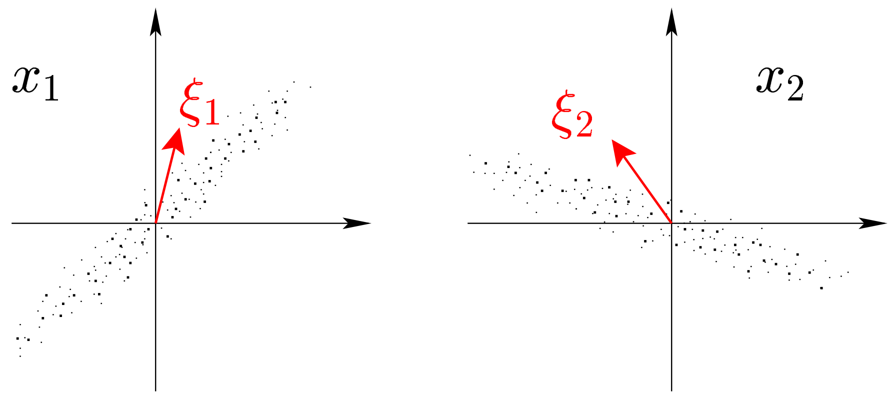
- If $\mathcal{F}$ is "big enough," then $\rho_{\mathcal{F}} = 0$ if and only if x_1 and x_2 are independent.

- When $\mathcal{F}$ is an RKHS, i.e. $f(x) = \langle \Phi(x), f \rangle$, then

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

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

Canonical correlation analysis



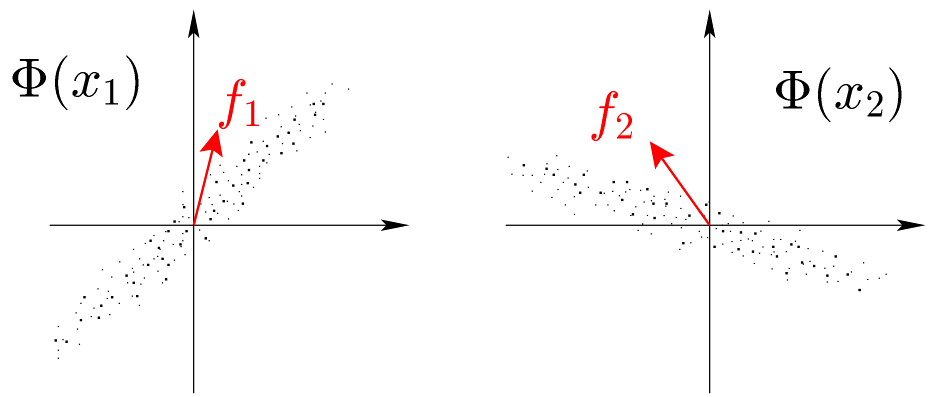
- Given two multivariate random variables x_1 and x_2, find the pair of directions ξ_1, ξ_2 with maximum correlation:

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

- 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 feature 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 with maximum correlation:

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

Kernel Canonical Correlation Analysis

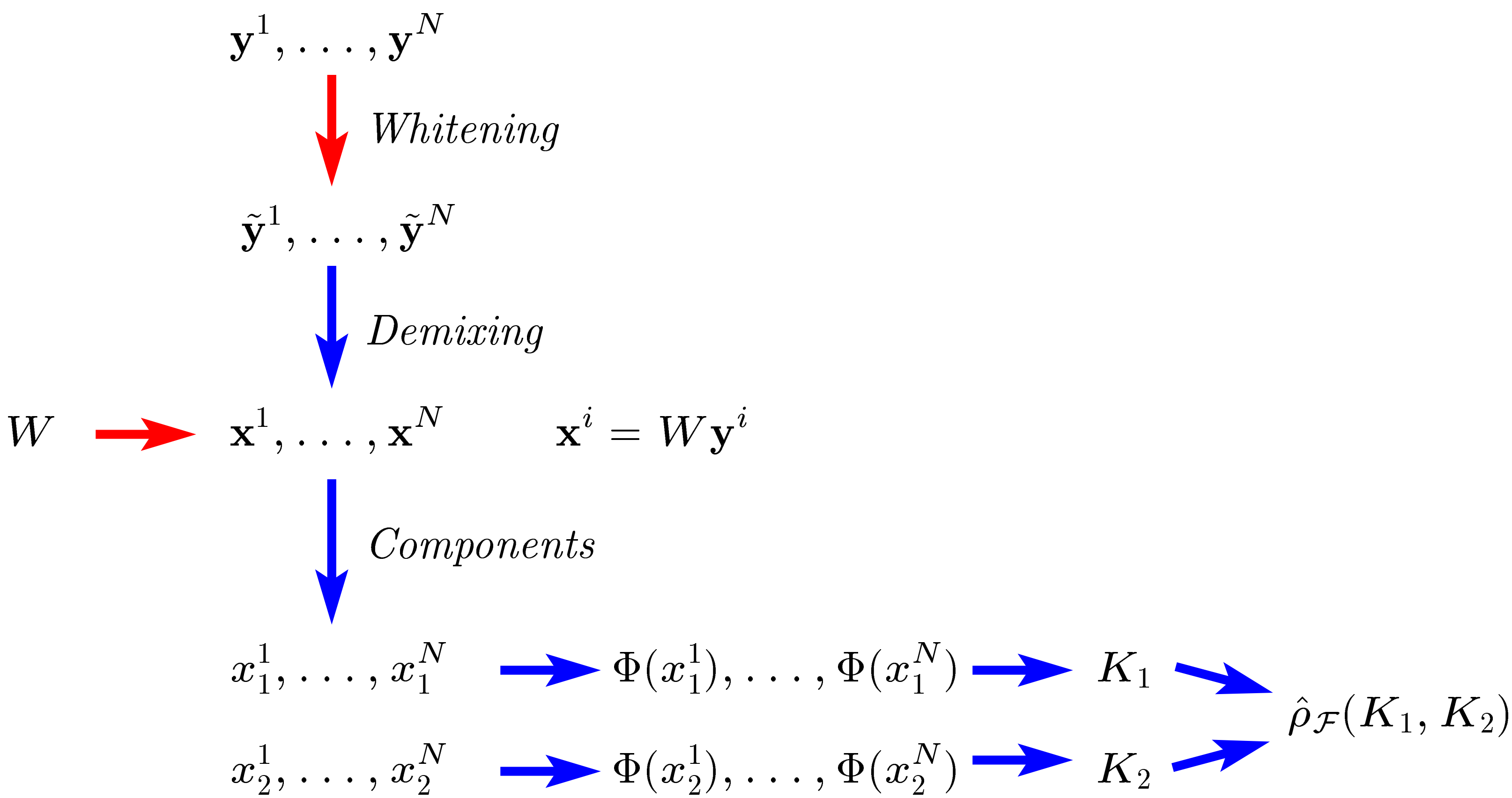
- Consider spans $\sum_i \alpha_1^i k(\cdot, x_i^1)$ and $\sum_i \alpha_2^i k(\cdot, x_i^2)$
- Let K_1, K_2 be the Gram matrices for $\{x_1^i\}$ and $\{x_2^i\}$

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

- 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



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

Generalization to $m > 2$ variables

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

$$\begin{pmatrix} K_1^2 & K_1K_2 & \cdots & K_1K_m \\ K_2K_1 & K_2^2 & \cdots & K_2K_m \\ \vdots & \vdots & & \vdots \\ K_mK_1 & K_mK_2 & \cdots & K_m^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

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

Regularization

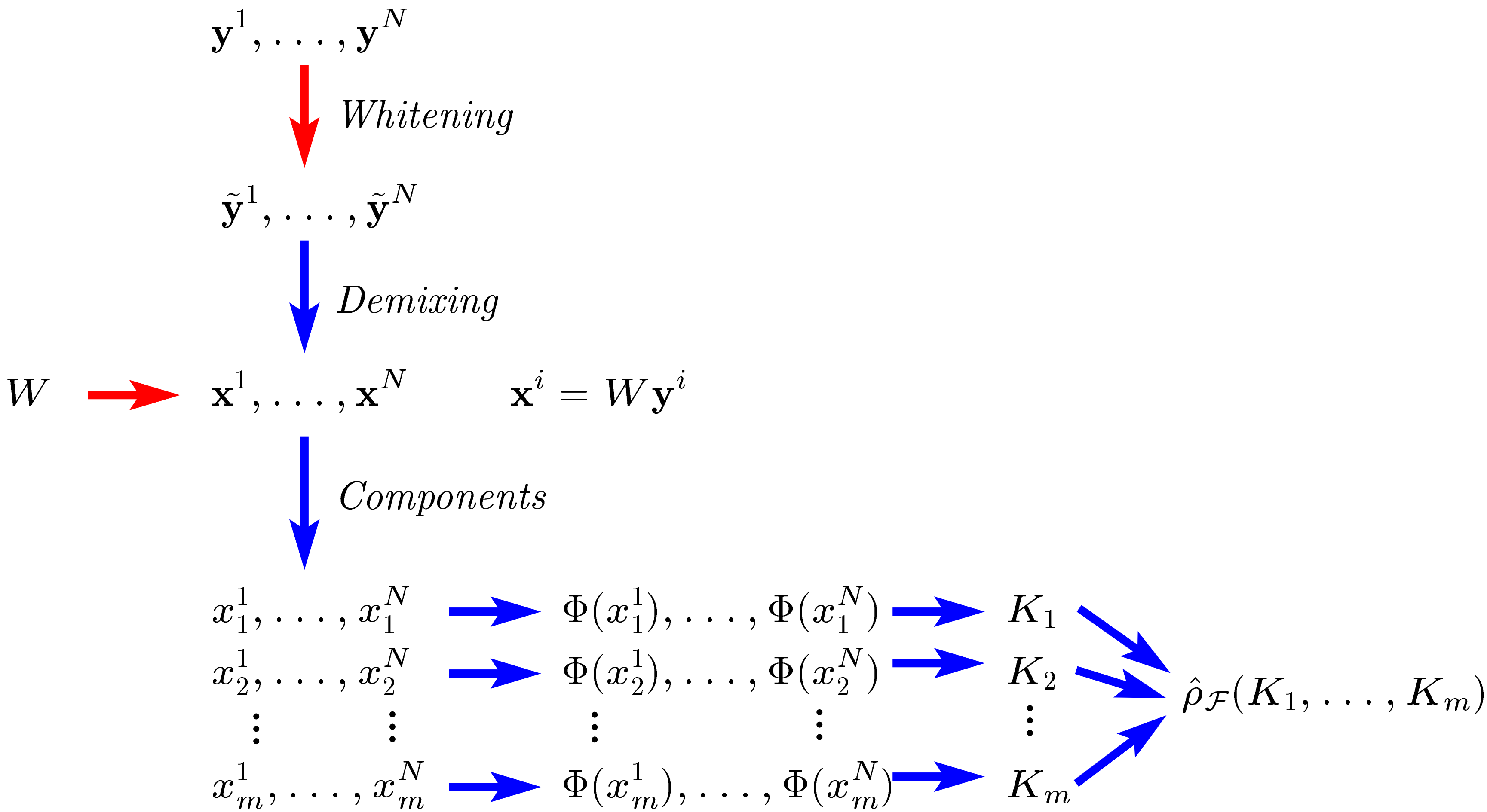
$$\begin{pmatrix} (K_1+\kappa I)^2 & K_1K_2 & \cdots & K_1K_m \\ K_2K_1 & (K_2+\kappa I)^2 & \cdots & K_2K_m \\ \vdots & \vdots & & \vdots \\ K_mK_1 & K_mK_2 & \cdots & (K_m+\kappa I)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}$$

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

Running time complexity

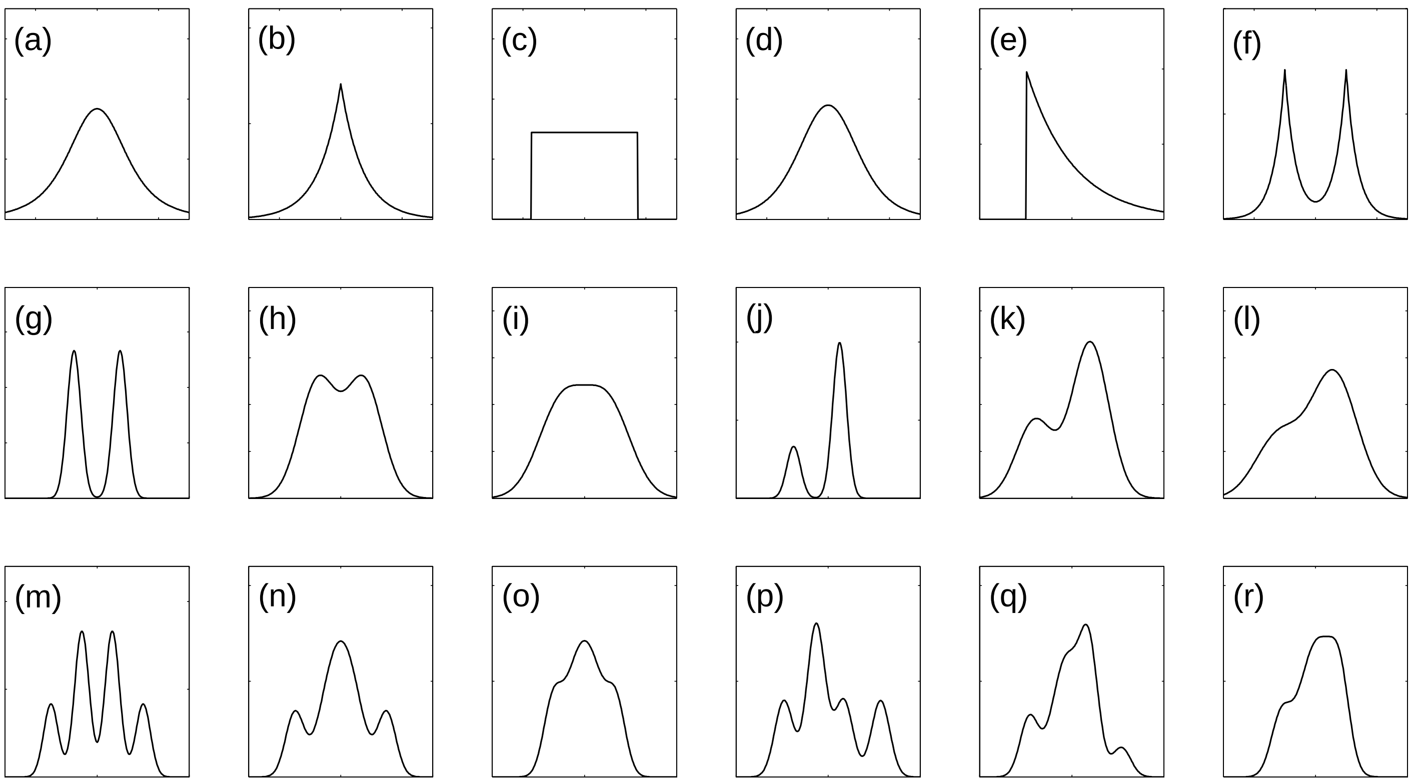
- Naive implementation: $O(m^3N^3)$
- Manage to reduce to linear time complexity in N:
  - Very low rank approximations: $K = GG^T$ where G is $N \times M$ and $M \ll N$.
  - Possible because Gram matrices have geometrically decaying spectrum
  - Symmetric positive semidefiniteness:
    incomplete Cholesky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

KernelICA algorithm



Empirical results

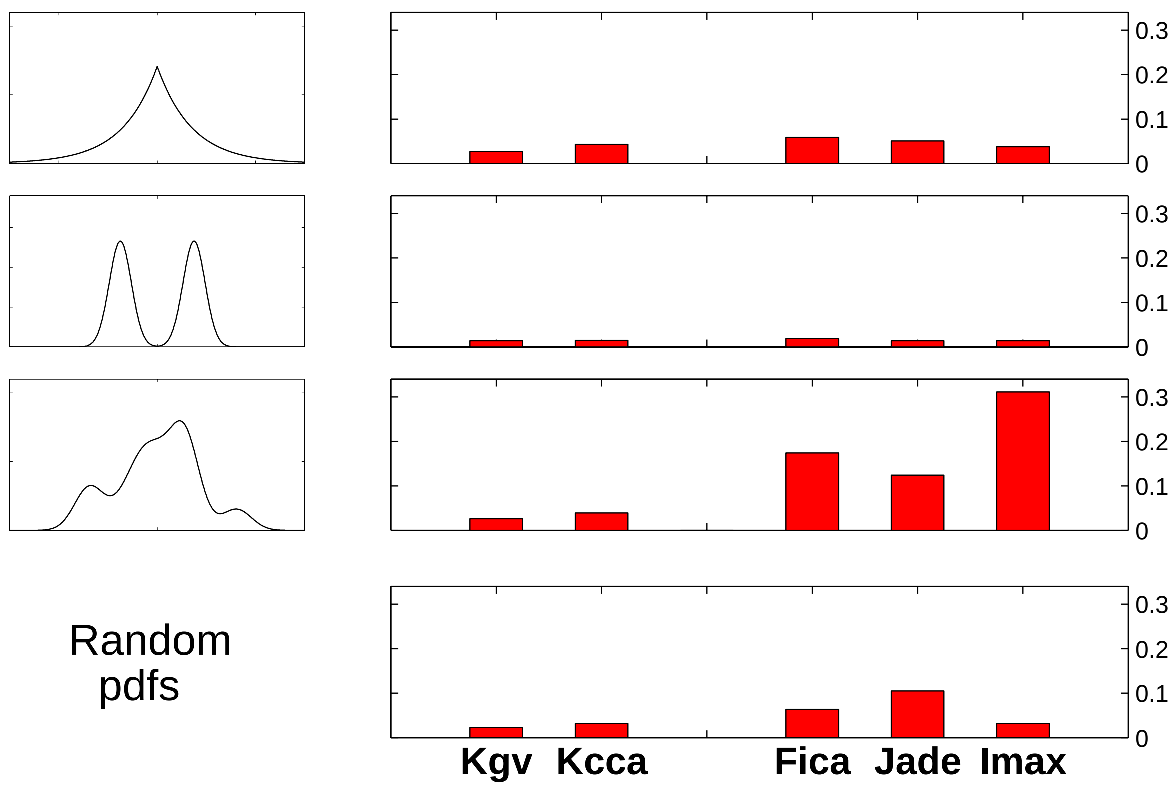
- Source distributions



Robustness to source distributions

- Comparison with other algorithms: FastICA (Hyvarinen,1999), Jade (Cardoso, 1998), Extended Infomax (Lee, 1999)

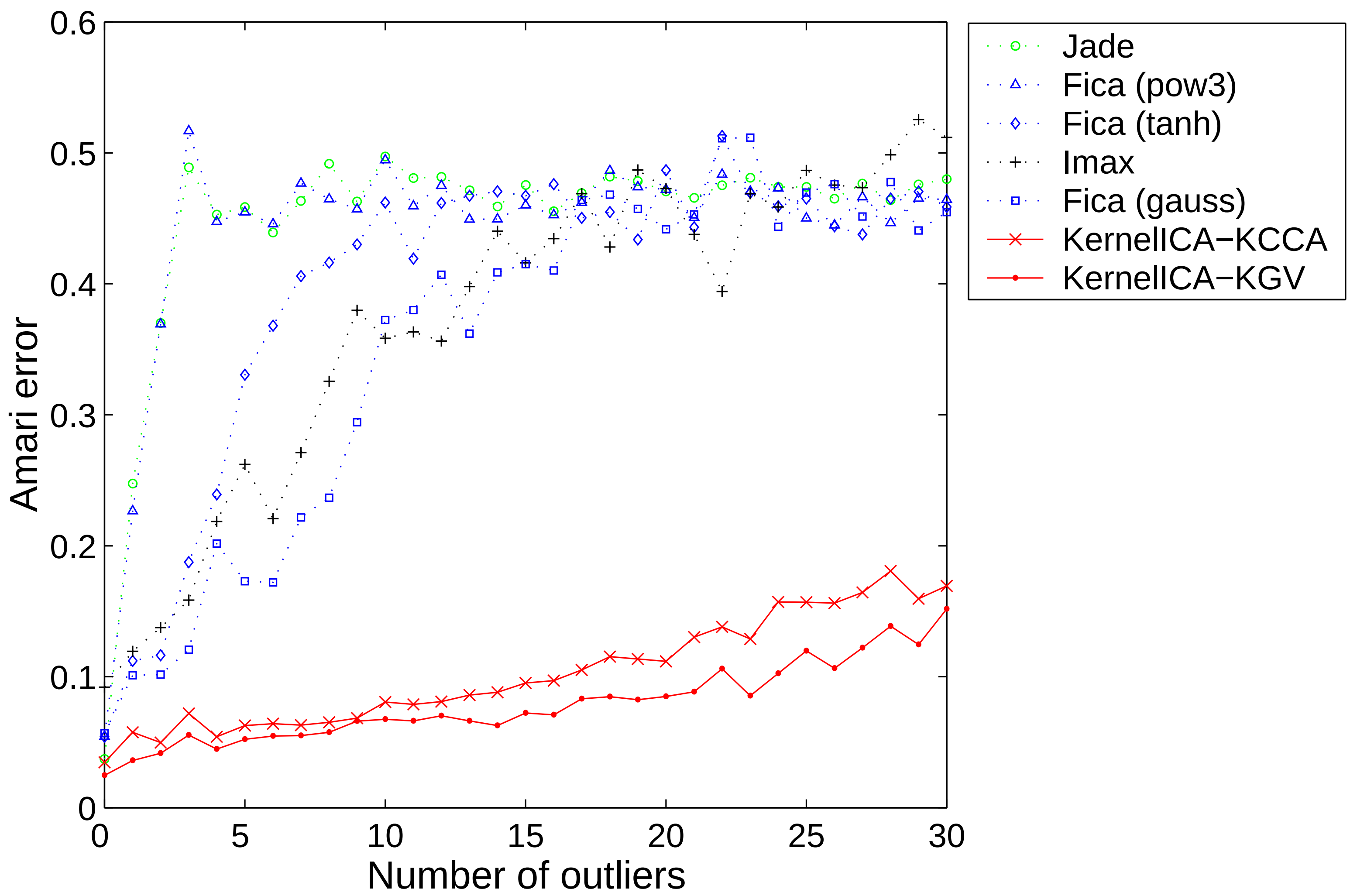
- Amari error : standard ICA distance from true sources



pdfs	Fica	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**

Robustness to outliers

- Large values added to randomly selected data points



Kernel Generalized Variance

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

$$M(x_1, x_2) = -\frac{1}{2} \log \left(\frac{\det C}{\det C_{11} \det C_{22}} \right) = -\frac{1}{2} \sum_{i=1}^{p} \log(1 - \rho_i^2)$$

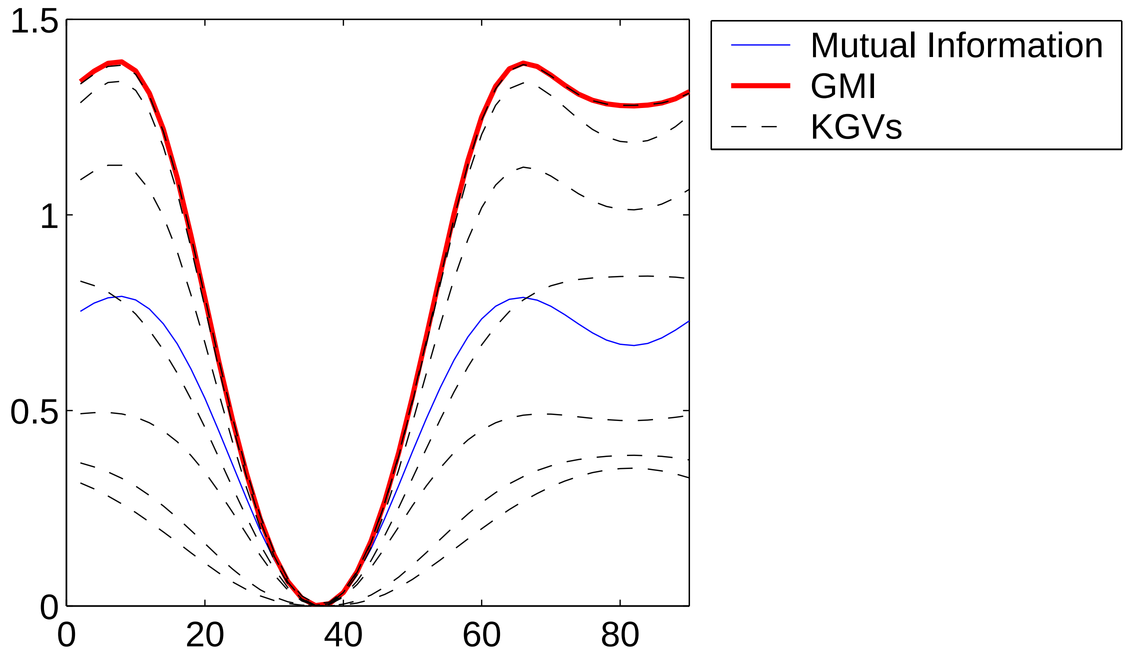
- Generalized variance $= \dfrac{\det C}{\det C_{11} \det C_{22}}$

- Kernel generalized variance $D(K_1, K_2) = \dfrac{\det \begin{pmatrix} K_1^2 & K_1 K_2 \\ K_2 K_1 & K_2^2 \end{pmatrix}}{\det K_1^2 \det K_2^2}$

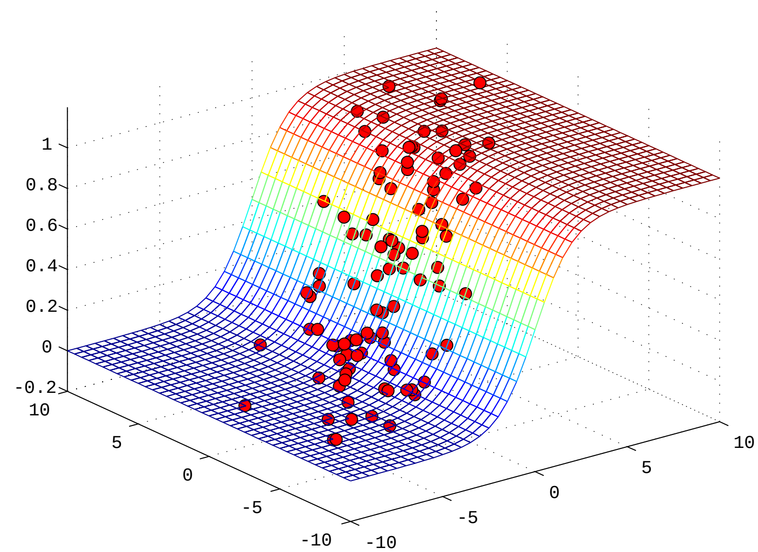
- New contrast function $M_D = -\frac{1}{2} \log D$

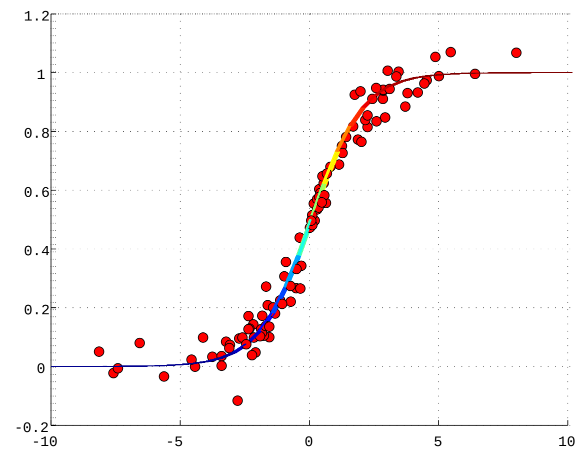
Kernel generalized variance and mutual information

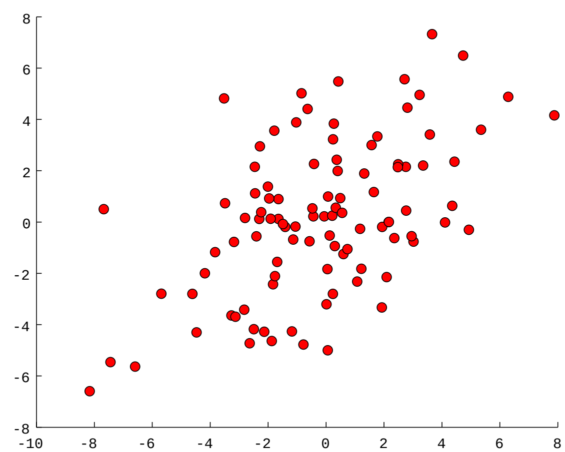
- Translation invariant kernels $K(x, y) = k\left(\frac{x-y}{\sigma}\right)$
- When σ tends to zero, $M_D(\sigma)$ has a limit $\mathcal{I}(x_1, x_2)$
- $\mathcal{I}(x_1, x_2)$ is equal to the mutual information up to second order "around independence"



Dimension reduction for regression





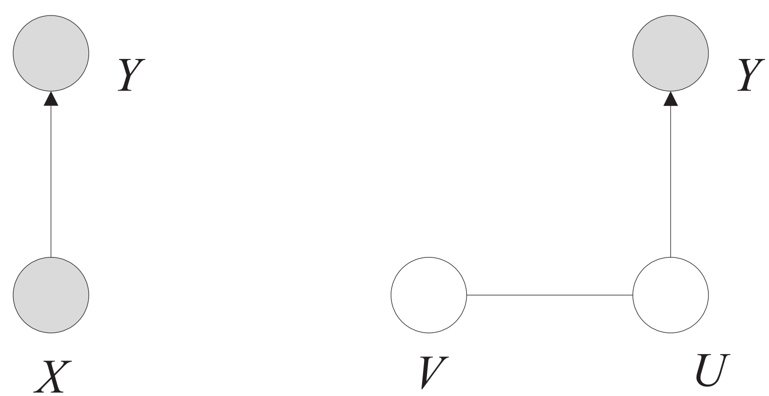


$$Y = \frac{1}{1+\exp(-X_1)} + N(0, 0.1^2)$$

Effective subspace = direction of X_1

Dimension reduction for regression (cont.)

- Nonparametric regression problem: $p(y|x)$
- Effective subspace: determined by a matrix B such that $p(y|x) = p(y|B^T x)$
  - now it's a semiparametric problem
- Conditional independence interpretation:
  - let $(U, V) = (B^T x, C^T x)$, for $R = (B, C)$ an orthogonal matrix
  - $p(y|x) = p(y|B^T x)$ if and only if $Y \perp\!\!\!\perp V | U$



Cross-covariance operator

- Consider reproducing kernel Hilbert spaces H_X and H_Y, for random variables X and Y

- Define a bounded operator $\Sigma_{YX} : H_X \to H_Y$ by:

$$\langle g, \Sigma_{YX} f \rangle_{H_Y} = E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)]$$

- Σ_{YX} is called a *cross-covariance operator*

Theorem *Given reproducing kernel Hilbert spaces H_X and H_Y, based on Gaussian kernels for random variables X and Y, respectively, X and Y are independent if and only if $\Sigma_{YX} = 0$.*

RKHS and conditional independence

- Conditional covariance operators:

$$\left\langle f, (\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})g \right\rangle_{H_Y} = E_X\left[\mathrm{Cov}_{Y|X}[f(Y)g(Y)|X]\right]$$

  where we assume Σ_{XX}^{-1} exists and assume $E_{Y|X}[g(Y)|X] \in H_X$ for all $g \in H_Y$

- We refer to $\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ as a *conditional covariance operator*

- Monotonicity of conditional covariance operator:

$$\Sigma_{YY|U} \geq \Sigma_{YY|X}$$

  where $Y, X = (U, V)$ random vectors, and the inequality is the sense of self-adjoint operators

RKHS and conditional independence (cont.)

Theorem

- $X = (U, V)$ *and* Y *are random vectors*
- H_X, H_Y, H_U*: RKHS with Gaussian kernels* k_X, k_Y, k_U*, respectively*
- $E_{Y|X}[g(Y)|X] \in H_X$ *and* $E_{Y|U}[g(Y)|U] \in H_U$ *for all* $g \in H_Y$
- *Then*

$$Y \perp\!\!\!\perp V | U \quad \textit{if and only if} \quad \Sigma_{YY|U} = \Sigma_{YY|X}$$

RKHS and conditional independence (cont.)

- Suggests the following formulation of the semiparametric estimation problem:

$$\min_{B:U=B^T X} \Sigma_{YY|U}$$

- What norm?
  - trace norm
  - operator norm
  - determinant norm

Kernel dimensionality reduction

- Estimation of the conditional covariance operator via plug-in:

$$\hat{\Sigma}_{YY|U} = \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY}$$

where

$$\hat{\Sigma}_{UU} = (K_U + \epsilon I)^2, \quad \hat{\Sigma}_{YY} = (K_Y + \epsilon I)^2, \quad \hat{\Sigma}_{UY} = \hat{\Sigma}_{UU}\hat{\Sigma}_{YY}$$

for centered Gram matrices K_U and K_Y

Kernel dimensionality reduction (cont.)

$$
\begin{aligned}
& \min_B && \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY} \\
\iff & \min_B && \det\left(I - \hat{\Sigma}_{YY}^{-1/2}\hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY}\hat{\Sigma}_{YY}^{-1/2}\right) \\
\iff & \min_B && \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}
\end{aligned}
$$

where

$$
\hat{\Sigma}_{[YU][YU]} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YU} \\ \hat{\Sigma}_{UY} & \hat{\Sigma}_{UU} \end{pmatrix}
$$

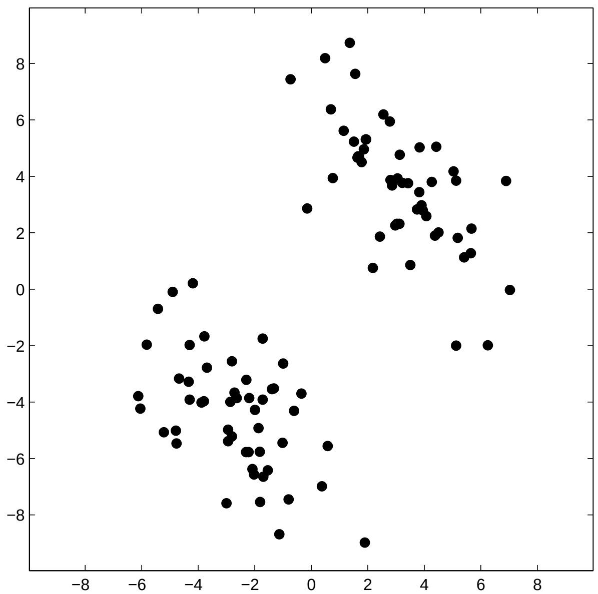
- But this is the kernel generalized variance, which we know how to optimize numerically (incomplete Cholesky)

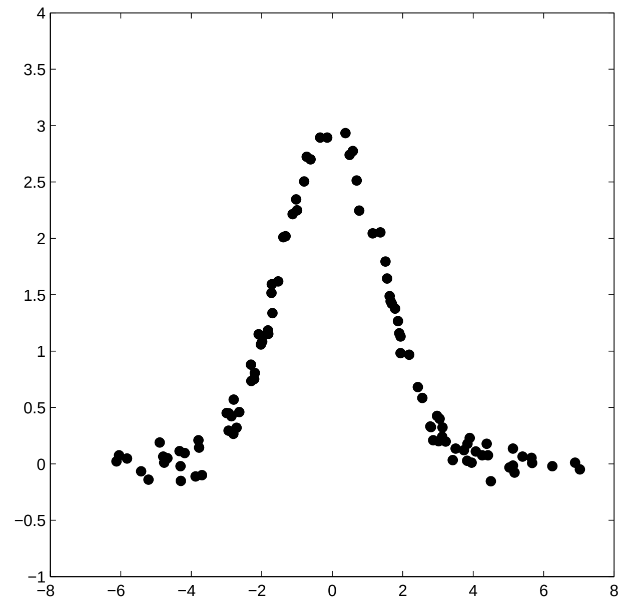
Existing methods

- Sliced inverse regression (SIR; Li, 1991)
  - PCA on $E[X|Y]$ for slices of Y
  - no assumptions on $p(y|x)$; elliptic assumption on $p(x)$
- Principal Hessian direction (pHd; Li, 1992)
  - based on eigenvectors of average Hessian
  - Gaussian assumption on $p(x)$; univariate Y
- Projection pursuit (Friedman, et al., 1981)
  - additive model assumed for $E[X|Y]$
- Canonical correlation (CCA) / Partial least squares (PLS)
  - linear model assumed for $E[X|Y]$

Synthetic data

- $Y \sim 2\exp(-X_1^2) + N(0, 0.1^2)$; 100 data points

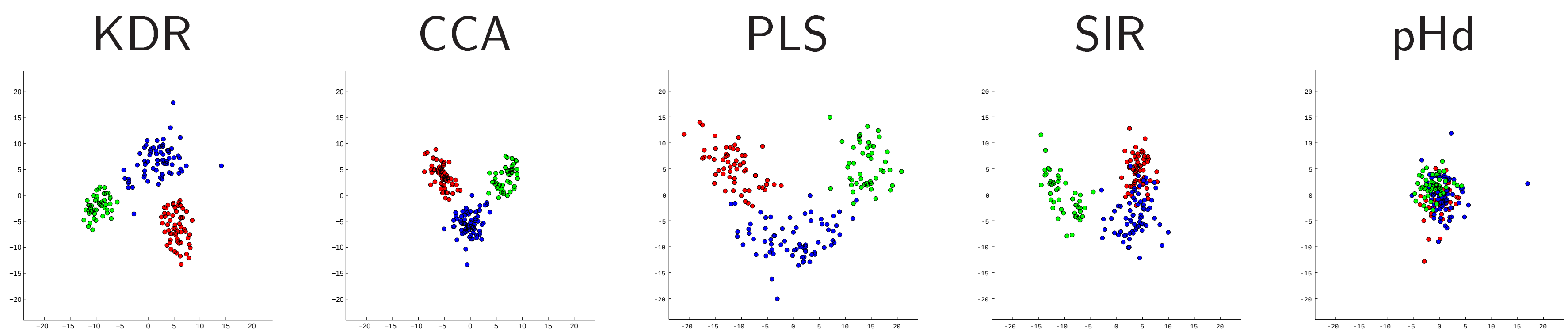




Method	SIR	pHd	CCA	PLS	KDR
Angle	-86.5	57.0	-10.4	-26.1	0.3

Wine data

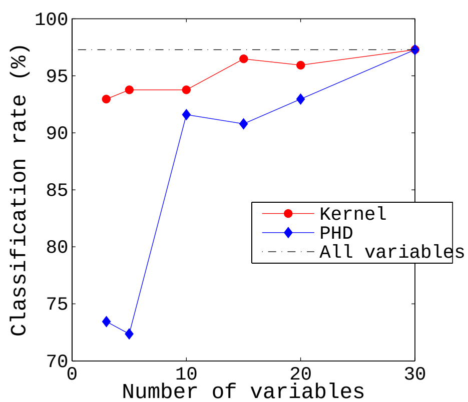
- 178 points in 13 dimensions



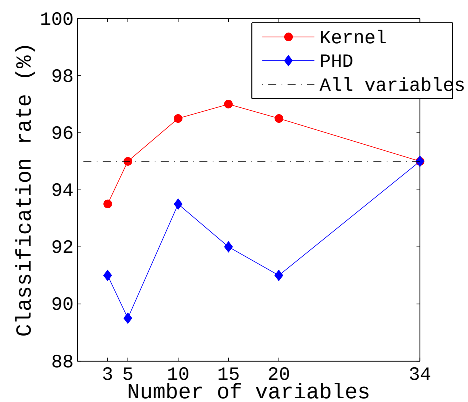
Binary classification

- Data sets from UCI repository for binary classification
- Only pHd and KDR are applicable to binary classification
- Support vector machine used for $p(y|x)$

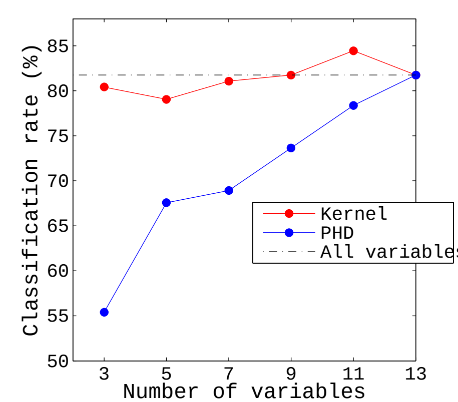
Breast cancer



Ionosphere



Heart disease



Beyond independent components: trees and clusters

- Model:

$$x = As, \; s \in \mathbb{R}^m, \; x \in \mathbb{R}^m, \; A \in \mathbb{R}^{m \times m}$$

- Relax the assumption of independence of the sources and estimate the pattern of dependence

- Using tree-structured distributions $\Rightarrow$ **Tree-dependent component analysis** (TCA)

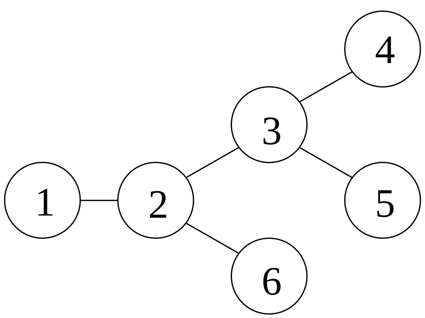
- Motivation:
  - multidimensional ICA $\Leftrightarrow$ clusters of sources
    (fetal ECG, music and instruments)
  - density estimation

Tree-structured distributions

- Equivalent definitions of tree-structured distributions:
  - **Factorization**

$$p(x) = \prod_{(u,v)\in T} \frac{p(x_u, x_v)}{p(x_u)p(x_v)} \prod_{u=1}^{m} p(x_u)$$

  - **Conditional independence and graph separation**: given the neighbors of x_i, x_i is independent from the rest of the graph

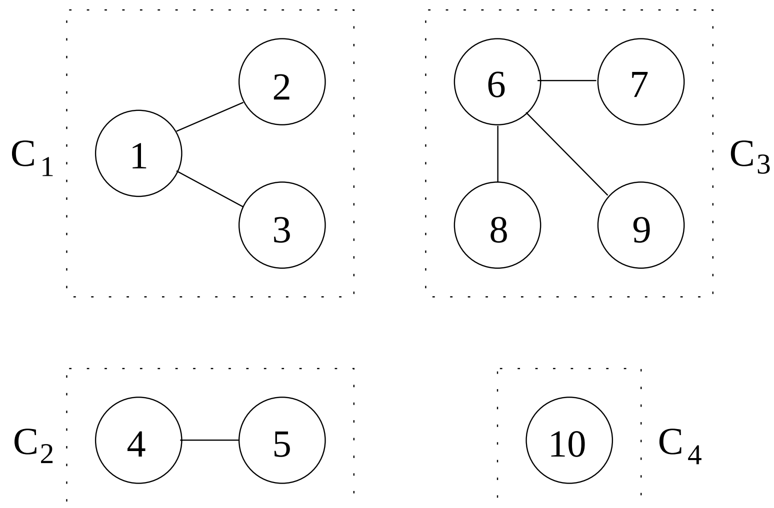


$$p(x) = \frac{p(x_2, x_1)p(x_3, x_2)p(x_4, x_3)p(x_5, x_3)p(x_6, x_2)}{p(x_2)^2 p(x_3)^2}$$

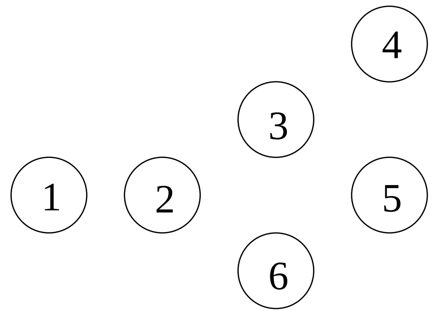
$$x_3 \perp x_1 | x_2, \quad x_3 \perp x_6 | x_2, \quad x_2 \perp x_5 | x_3$$

Trees, forests and clusters

- **Clusters of components**: components are dependent within clusters and independent between clusters
- Modeling clusters using **non-spanning trees** (or forests):
  - Each connected component represents a cluster
  - Each cluster marginal distribution is modeled by a tree
- Estimating the forest ⇒ estimating the number and sizes of the clusters



Empty graph (no edges)



- $p(x) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5) \Rightarrow$ independent components
- ICA is a submodel of TCA

From ICA to TCA: semiparametric contrast function

	ICA	TCA
Parametric part	demixing matrix W	demixing matrix W tree T
Nuisance parameters	source marginal densities	source marginal and pairwise densities
Contrast function	Mutual information of the estimated sources $s = Wx$ $I(s_1, \ldots, s_m)$	T-mutual information of the estimated sources $I(s_1, \ldots, s_m) - \sum_{(u,v) \in T} I(s_u, s_v)$

Estimation of the TCA contrast function

- Contrast function :

$$
\begin{aligned}
J(x, W, T) &= I(s_1, \ldots, s_m) - \sum_{(u,v)\in T} I(s_u, s_v) \\
&= -H(s) + H(s_1) + \cdots + H(s_m) - \sum_{(u,v)\in T} \left(H(s_u) + H(s_v) - H(s_u, s_v)\right)
\end{aligned}
$$

- Solution 1: Kernel generalized variance (KGV)
  - Compute m-fold and 2-fold approximations of the mutual information via the KGV
  - Overall computation is $O(mN)$

- Solution 2: Kernel density estimation (KDE)
  - Only need 2D entropies because $H(s) = H(Wx) = H(x) + \log\det W$
  - Overall computation of $J(x, W, T)$ is $O(mN)$

Optimization

- Contrast:

$$J(W,T) = I(s_1,\ldots,s_m) - \sum_{(u,v)\in T} I(s_u,s_v), \text{ where } s = Wx$$

- Alternating minimization
  - with respect to W: gradient descent
  - with respect to T: maximum weight spanning tree
    solvable in polynomial time by greedy algorithms (Chow-Liu algorithm, 1968).

Alternative optimization

- "Marginalize" the tree T, i.e. minimize with respect to W:

$$\widetilde{J}(W) = \min_T \left\{ I(s_1, \ldots, s_m) - \sum_{(u,v)\in T} I(s_u, s_v) \right\}, \text{ where } s = Wx$$

- $\widetilde{J}(W)$ is continuous piecewise differentiable function

- Minimize using coordinate descent (i.e. Jacobi rotations if imposing whiteness constraint)

Extension to time series

- Modeling assumption: **stationary Gaussian time series** $s(t) = (s_1(t), \ldots, s_m(t)), t \in \mathbb{Z}$

- Enough for demixing if sources have linearly independent spectral density functions

- Algorithms for ICA can be defined either
  - in the **time domain**, using the autocovariance function:

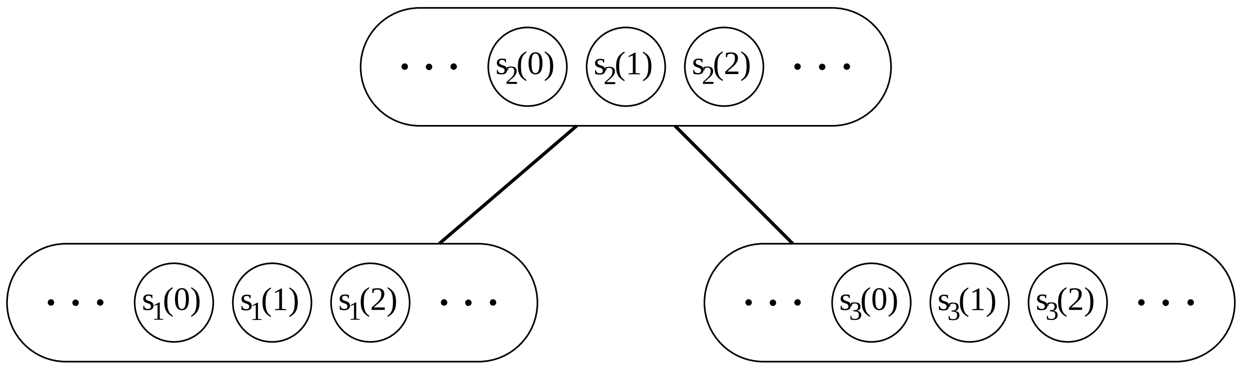
    $$\Gamma(h) = E[s(t)s(t+h)^\top]$$

    SOBI (Belouchrani et. al, 1997), TDSEP (Ziehe et. al, 1998)
  - in the **frequency domain**, using the spectral density (Pham, 2000):

    $$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \Gamma(h) e^{-ih\omega}$$

Graphical model for Gaussian stationary time series

- Conditional/marginal independence between **entire time series** (Brillinger, 1996, Dahlhaus, 2000)



- **Natural expression in the frequency domain** ⇒ most of the results from Gaussian graphical models can be extended by replacing entropies by entropy rates

$$h(x) = \frac{1}{4\pi}\int_0^{2\pi} \log\det(4\pi^2 e f(\omega))d\omega$$

Estimation for Gaussian stationary time series

- Entropy rate of projection

$$h(s) = h(Wx) = \frac{1}{4\pi} \int_0^{2\pi} \log \det(4\pi^2 e W f(\omega) W^\top) d\omega$$

- Estimated using **smoothed periodogram**

Conclusions

- Characterizations of independence via reproducing kernel Hilbert spaces
  - RKHS are rich enough for the statistical problem, but small enough to yield efficient algorithms
- Applications
  - independent component analysis
  - tree-dependent component analysis
  - dimension reduction for regression
- For details: **http://www.cs.berkeley.edu/~jordan**