Independent Component Analysis

Background paper:

http://www-stat.stanford.edu/∼hastie/Papers/ica.pdf

ICA Problem

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

where

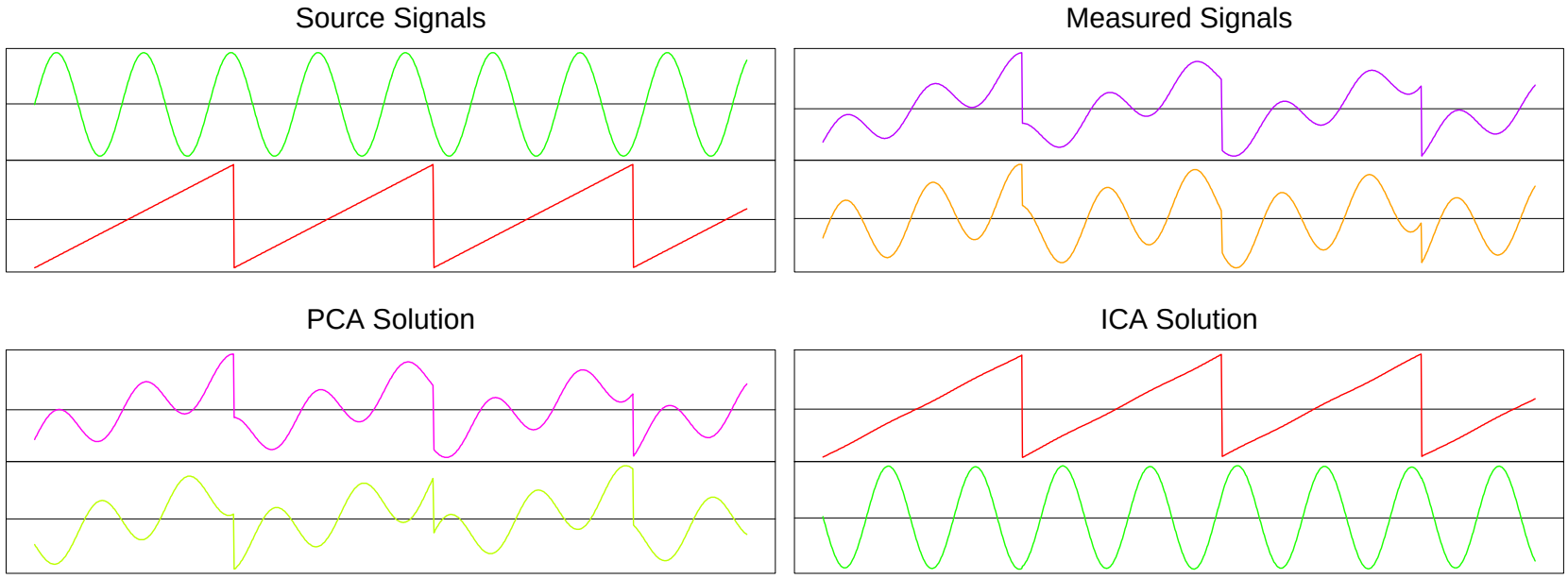
- X is a random p-vector representing multivariate input measurements.
- S is a latent source p-vector whose components are independently distributed random variables.
- $\mathbf{A}$ is $p \times p$ mixing matrix.

Given realizations $x_1, x_2, \ldots, x_N$ of X, the goals of ICA are to

- Estimate $\mathbf{A}$
- Estimate the source distributions $S_j \sim f_{S_j},\ j = 1, \ldots, p$.

Cocktail Party Problem

In a room there are p independent sources of sound, and p microphones placed around the room hear different mixtures.



Here each of the $x_{ij} = x_j(t_i)$ and recovered sources are a time-series sampled uniformly at times t_i.

Independent vs Uncorrelated

WoLOG can assume that $\mathrm{E}(S) = 0$ and $\mathrm{Cov}(S) = \mathbf{I}$, and hence $\mathrm{Cov}(X) = \mathrm{Cov}(\mathbf{A}S) = \mathbf{A}\mathbf{A}^T$.

Suppose $X = \mathbf{A}S$ with S unit variance, uncorrelated

Let $\mathbf{R}$ be any orthogonal $p \times p$ matrix. Then

$$X = \mathbf{A}S = \mathbf{A}\mathbf{R}^T\mathbf{R}S = \mathbf{A}^*S^*$$

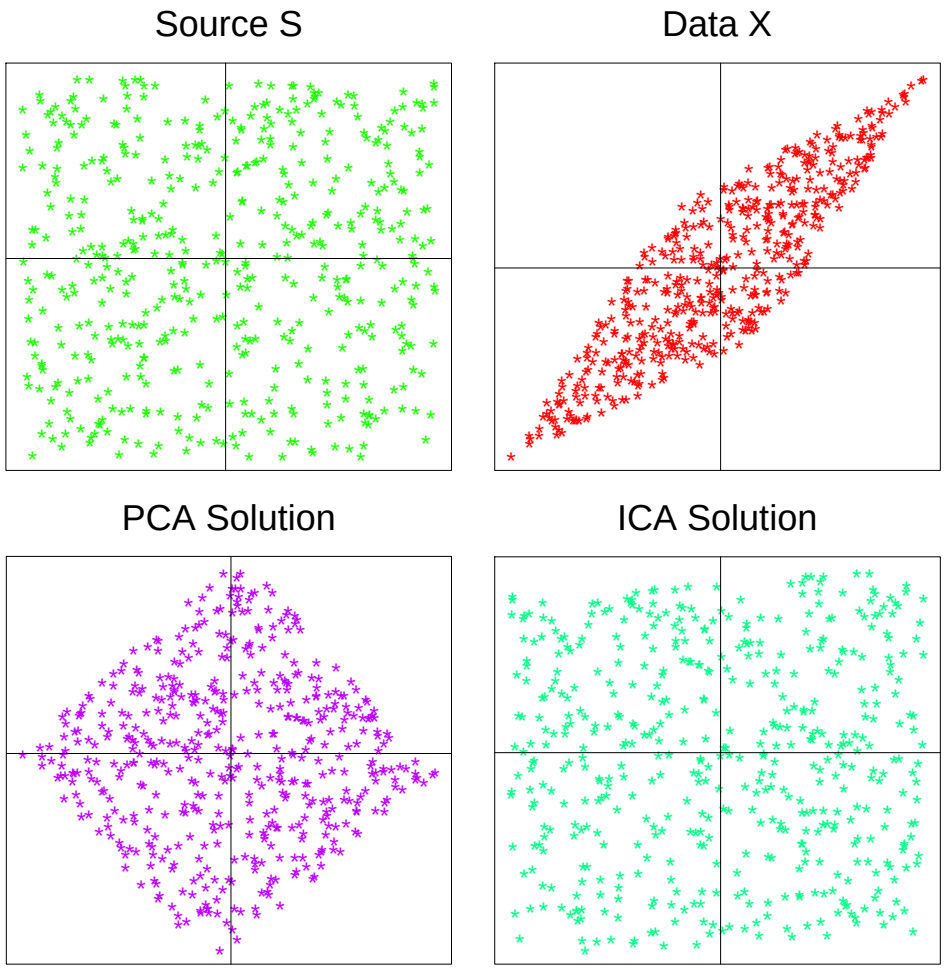
and $\mathrm{Cov}(S^*) = I$

It is not enough to find uncorrelated variables, as they are not unique under rotations.

Hence methods based on second order moments, like principal components and Gaussian factor analysis, cannot recover $\mathbf{A}$.

ICA uses independence, and non-Gaussianity of S, to recover $\mathbf{A}$ — e.g. higher order moments.

Independent vs Uncorrelated Demo

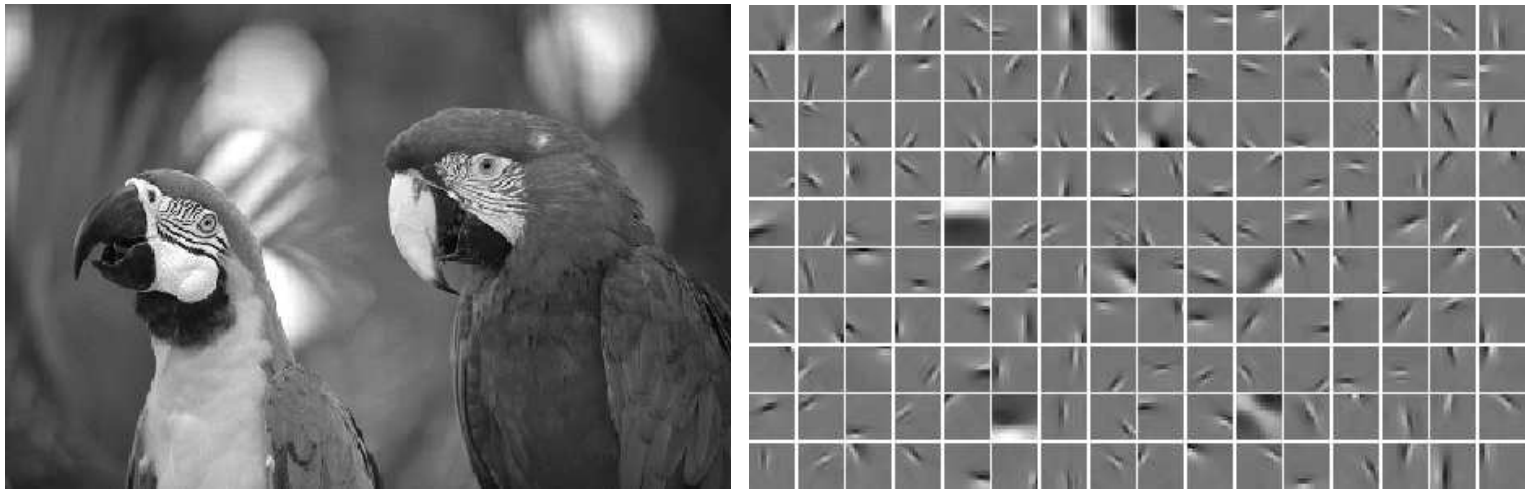


Principal components are uncorrelated linear combinations of X, chosen to successively maximize variance.

Independent components are also uncorrelated linear combinations of X, chosen to be as independent as possible.

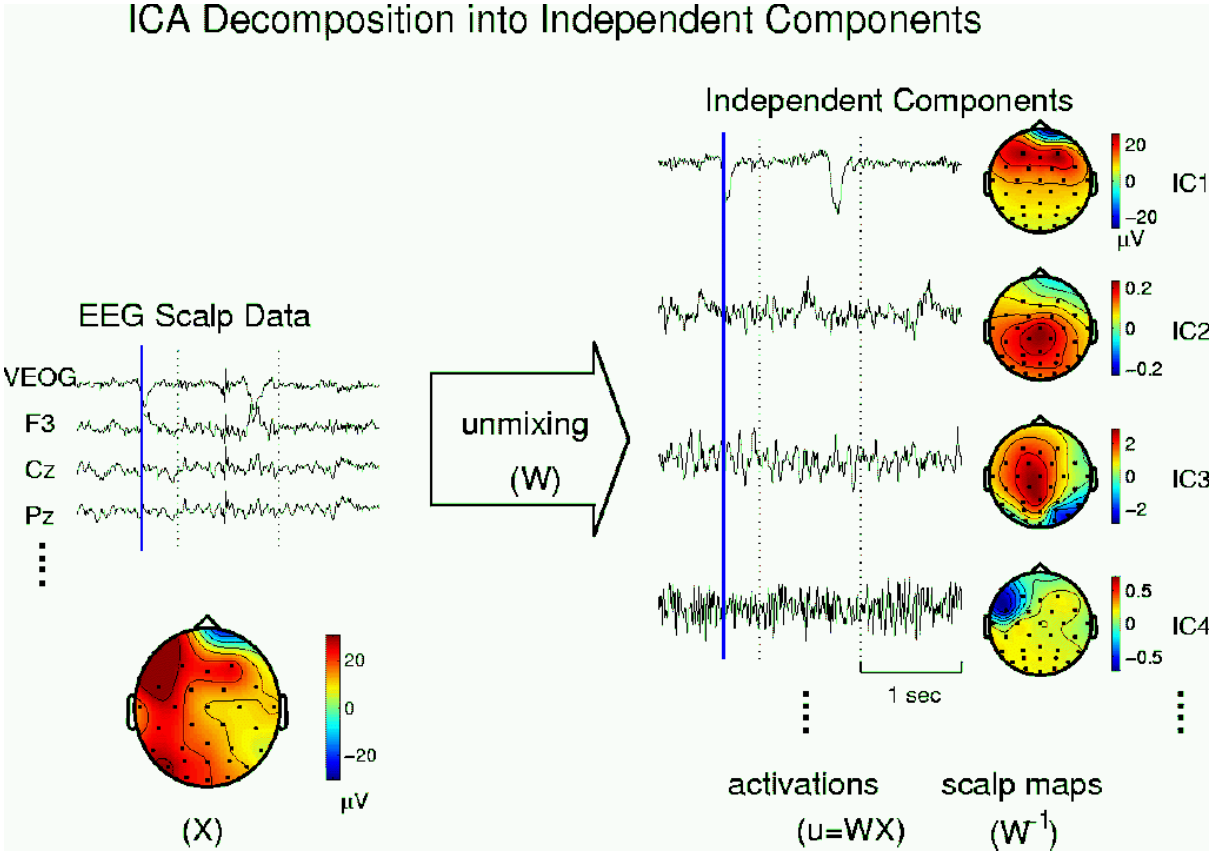
Example: ICA representation of Natural Images

Pixel blocks are treated as vectors, and then the collection of such vectors for an image forms an image database. ICA can lead to a sparse coding for the image, using a natural basis.



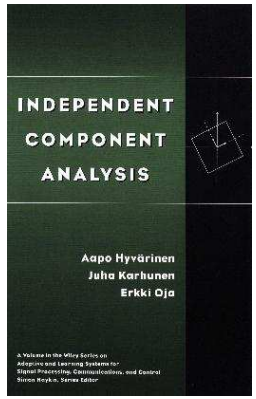
see http://www.cis.hut.fi/projects/ica/imageica/(Patrik Hoyer and Aapo Hyvärinen, Helsinki University of Technology)

Example: ICA and EEG data



See http://www.cnl.salk.edu/~tewon/ica_cnl.html (Scott Makeig, Salk Institute)

Approaches to ICA



ICA literature is HUGE. Recent book by Hyvärinen, Karhunen & Oja (Wiley, 2001) is a great source for learning about ICA, and some good computational tricks.

- Mutual Information and Entropy, maximizing non-Gaussianity — FastICA (HKO 2001), Infomax (Bell and Sejnowski, 1995)
- Likelihood methods — ProdDenICA (Hastie +Tibshirani)-later
- Nonlinear decorrelation — Y_1 independent Y_2 iff $\max_{g,f} \mathrm{Corr}[f(Y_1), g(Y_2)] = 0$ (Hérault-Jutten, 1984), KernelICA (Bach and Jordan, 2001)
- Tensorial moment methods

Simplification

Let $\mathbf{\Sigma} = \mathrm{Cov}(X)$ and suppose $X = \mathbf{A}S$, and $\mathbf{B} = \mathbf{A}^{-1}$. Then we can write $\mathbf{B} = \mathbf{W}\mathbf{\Sigma}^{-\frac{1}{2}}$ for some nonsingular $\mathbf{W}$. Then $S = \mathbf{B}X = \mathbf{W}\mathbf{\Sigma}^{-\frac{1}{2}}X$ with $\mathrm{Cov}(S) = \mathbf{I}$ and $\mathbf{W}$ orthonormal.

So operationally we sphere the data $\tilde{X} = \mathbf{\Sigma}^{-\frac{1}{2}}X$, and then seek an orthogonal matrix $\mathbf{W}$ so that the components $S = \mathbf{W}\tilde{X}$ are independent.

Entropy and Mutual Information

Entropy: $H(Y) = -\int f(y) \log f(y) dy$ — maximized by $f(y) = \phi(y)$, the Gaussian (for fixed variance).

Mutual Information: $I(Y) = \sum_{j=1}^{p} H(Y_j) - H(Y)$

- Y is a random vector with joint density $f(y)$ and entropy $H(Y)$
- $H(Y_j)$ is the (marginal) entropy of component Y_j, with marginal density $f_j(y_j)$.
- $I(Y)$ is the Kullback-Leibler divergence between $f(Y)$ and it's independence version $\prod_1^p f_j(y_j)$ (which is the KL closest of all independence densities to $f(y)$)
- Hence $I(Y)$ is a measure of dependence between the components of a random vector Y.

Entropy, Mutual Information, and ICA

If $\mathrm{Cov}(X) = \mathbf{I}$ and $\mathbf{W}$ is orthogonal then simple calculations show

$$I(\mathbf{W}X) \quad = \quad \sum_{j=1}^{p} H(w_j^T X) - H(X)$$

Hence

$$\begin{aligned}
\min_{\mathbf{W}} I(\mathbf{W}X) \quad &\Longleftrightarrow \quad \min_{\mathbf{W}} \{\text{dependence between } w_j^T X\} \\
&\Longleftrightarrow \quad \min_{\mathbf{W}} \{\text{the sum of the entropies of the } w_j^T X\} \\
&\Longleftrightarrow \quad \max_{\mathbf{W}} \{\text{departures from Gaussianity of the } w_j^T X\}
\end{aligned}$$

- Many methods for ICA look for low-entropy or non-gaussian projections (Hyvärinen, Karhunen & Oja, 2001)
- Strong similarities with projection pursuit (Friedman and Tukey, 1974)

Negentropy and *FastICA*

Negentropy: $J(Y_j) = H(Y_j) - H(Z_j)$, where Z_j is a Gaussian RV with same variance as Y_j. Measures the departure from Gaussianity.

FastICA uses simple approximations to negentropy

$$J(w_j^T X) \approx [\mathrm{E}G(w_j^T X) - \mathrm{E}G(Z_j)]^2,$$

and with data replaced expectations by a sample averages. They use

$$G(y) = \frac{1}{a} \log \cosh y;, \quad 1 \leq a \leq 2$$

ICA Density Model

Under the ICA model, the density of S:

$$f_S(s) = \prod_{j=1}^{p} f_j(s_j)$$

with each f_j a univariate density, with mean 0 and variance 1.

Next we develop two direct ways of fitting this density model.

Discrimination from Gaussian

- "Unsupervised vs supervised" idea
- Consider case of $p = 2$ variables. Idea: observed data is assigned to class $G = 1$, and a background sample is generated from spherical Gaussian $g_0(x)$ and assigned to class $G = 0$.
- We fit to this data a projection pursuit model of the form

$$\log \frac{Pr(G = 1)}{1 - Pr(G = 1)} = f_1(a_1^T X) + f_2(a_2^T X) \qquad (1)$$

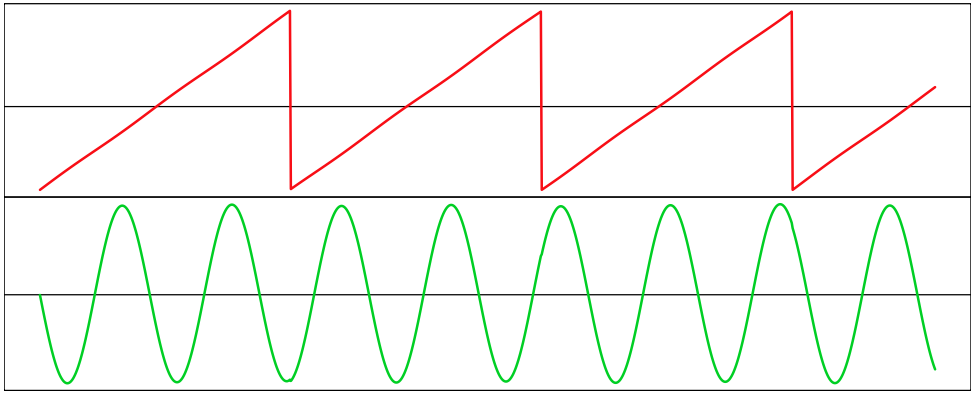
  where $[a_1, a_2]$ is an orthonormal basis.
- This implies the data density model

$$g(X) = g_0(X) \cdot \exp f_1(a_1^T X) \cdot \exp f_2(a_2^T X)$$

  Can also show that $g_0(X)$ factors into a product $h_1(a_1^T X) \cdot h_2(a_2^T X)$, since $[a_1, a_2]$ is an orthonormal basis.

- Hence model (1) is a product density for the data. If also looks like a neural network
- Can fit it by a) sphering ("whitening") the data and then b) applying `ppr` function in R, combined with IRLS (iteratively reweighted least squares).

Projection Pursuit Solutions



ICA Density Model -direct fitting

Density of $X = \mathbf{A}S$:

$$f_X(x) = |\mathbf{B}| \prod_{j=1}^{p} f_j(b_j^T x)$$

with $\mathbf{B} = \mathbf{A}^{-1}$.

Log-likelihood, given $x_1, x_2, \ldots, x_N$:

$$\ell(\mathbf{B}, \{f_j\}_1^p) = \log |\mathbf{B}| + \sum_{i=1}^{N} \sum_{j=1}^{p} \log f_j(b_j^T x_i)$$

ICA Density Model -fitting ctd

- As before: let $\mathbf{\Sigma} = \mathrm{Cov}(X)$; then for any $\mathbf{B}$ we can write $\mathbf{B} = \mathbf{W}\mathbf{\Sigma}^{-\frac{1}{2}}$ for some nonsingular $\mathbf{W}$. Then $S = \mathbf{B}X$ and $\mathrm{Cov}(S) = \mathbf{I} \implies \mathbf{W}$ is orthonormal. So we sphere the data $\tilde{X} = \mathbf{\Sigma}^{-\frac{1}{2}} X$ and seek an orthonormal $\mathbf{W}$.
- Let

$$f_j(s_j) = \phi(s_j) e^{g_j(s_j)},$$

  a tilted Gaussian density. Here ϕ is the standard Gaussian density, and g_j satisfies the normalization conditions.

Then

$$\ell(\mathbf{W}, \mathbf{\Sigma}, \{g_j\}_1^p) = -\frac{1}{2} \log |\mathbf{\Sigma}| + \sum_{i=1}^{N} \sum_{j=1}^{p} \Big[\log \phi_j(w_j^T \tilde{x}_i) + g_j\big(w_j^T \tilde{x}_i\big) \Big]$$

- We estimate $\hat{\mathbf{\Sigma}}$ by the sample covariance of the x_i, and ignore it thereafter — pre-whitening in the ICA literature.

ProdDenICA algorithm

Maximizes log-likelihood on previous slide, over $\mathbf{W}$ and $g_j(\cdot)$, $j = 1, 2, \ldots p$, with smoothness constraints on $g_j(\cdot)$.

Details on next 3 slides.

Restrictions on g_j

Our model is still over-parameterized. We maximize instead a penalized log-likelihood:

$$\sum_{j=1}^{p}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log\phi(w_j^T\tilde{x}_i)+g_j(w_j^T\tilde{x}_i)\right]-\lambda_j\int\{g_j'''(t)\}^2(t)dt\right]$$

w.r.t. $\mathbf{W}^T=(w_1,w_2,\ldots,w_p)$ and $g_j,\ j=1,\ldots,p$

subject to

- $\mathbf{W}^T\mathbf{W}=\mathbf{I}$
- each $f_j(s)=\phi(s)e^{g_j(s)}$ is a density, with mean 0 and variance 1.
- solutions $\hat{g}_j$ are piecewise quartic splines

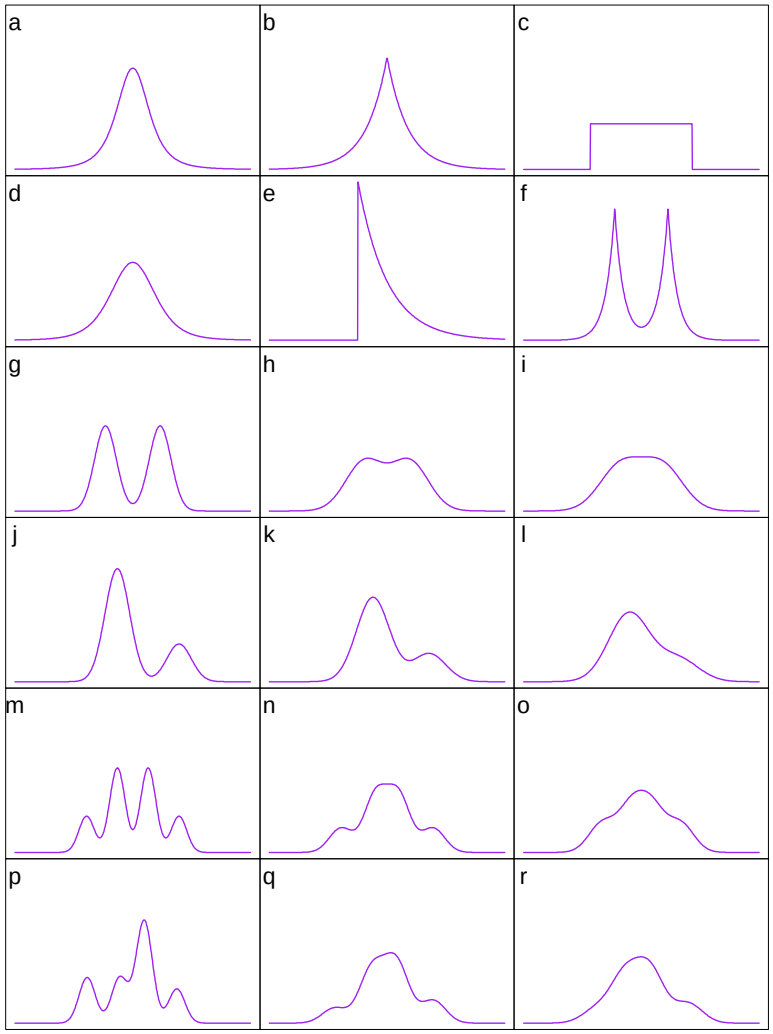
ProDenICA: Product Density ICA algorithm

1. Initialize $\mathbf{W}$ (random Gaussian matrix followed by orthogonalization).
2. Alternate until convergence of $\mathbf{W}$, using the Amari metric.
   (a) Given $\mathbf{W}$, optimize the penalized log-likelihood w.r.t. g_j (separately for each j), using the penalized density estimation algorithm.
   (b) Given $g_j,\ j = 1, \ldots, p$, perform one step of a fixed point algorithm towards finding the optimal $\mathbf{W}$.

Penalized density estimation for step (a)

- Discretize the data
- Fit a generalized additive model with Poisson family of distributions (see text chapter 9)

Simulations



Taken from Bach and Jordan (2001)

- Each distribution used to generate 2-dim S and X (with random $\mathbf{A}$) — 30 reps each
- 300 reps with 4-dim S — distributions of S_j picked at random.

Compared FastICA (homegrown Splus version), KernelICA (Francis Bach's matlab version), and ProDenICA (Splus).

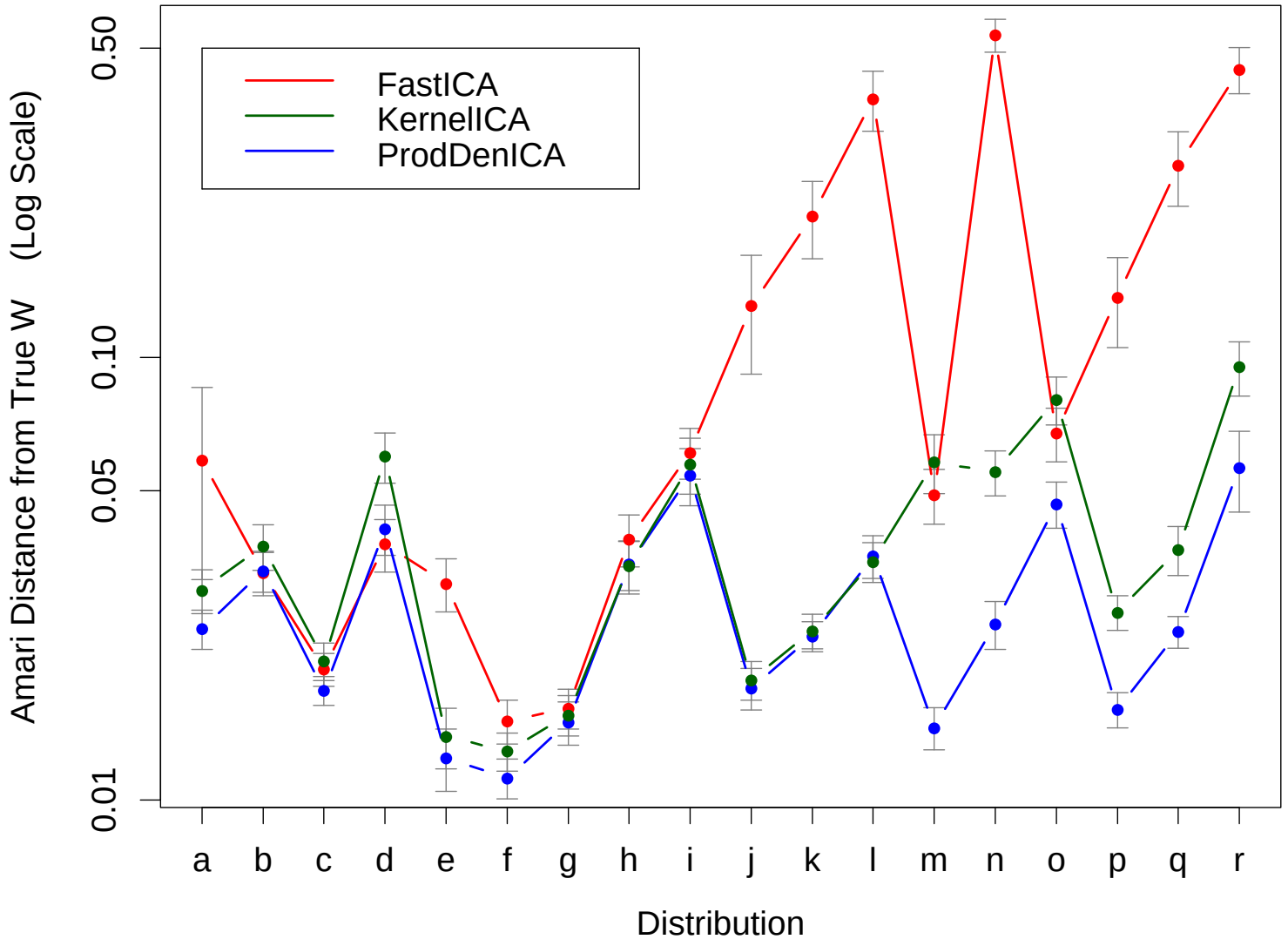
Amari Metric

We evaluate solutions by comparing their estimated $\mathbf{W}$ with the true $\mathbf{W}_0$, using the Amari metric (HKO 2001, Bach and Jordan 2001):

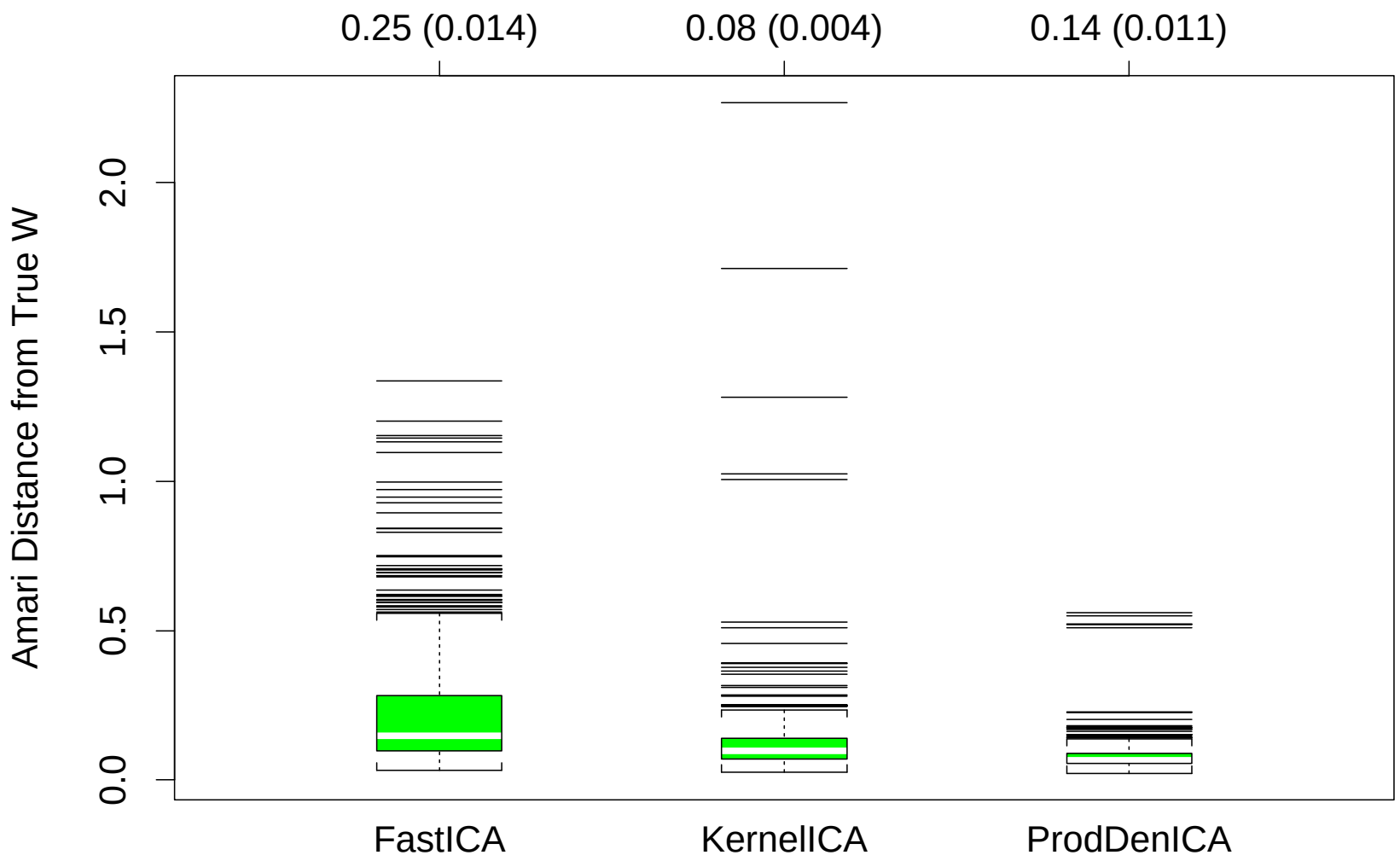
$$d(\mathbf{W}_0, \mathbf{W}) = \frac{1}{2p} \sum_{i=1}^{p} \left(\frac{\sum_{j=1}^{p} |r_{ij}|}{\max_j |r_{ij}|} - 1 \right) + \frac{1}{2p} \sum_{j=1}^{p} \left(\frac{\sum_{i=1}^{p} |r_{ij}|}{\max_i |r_{ij}|} - 1 \right)$$

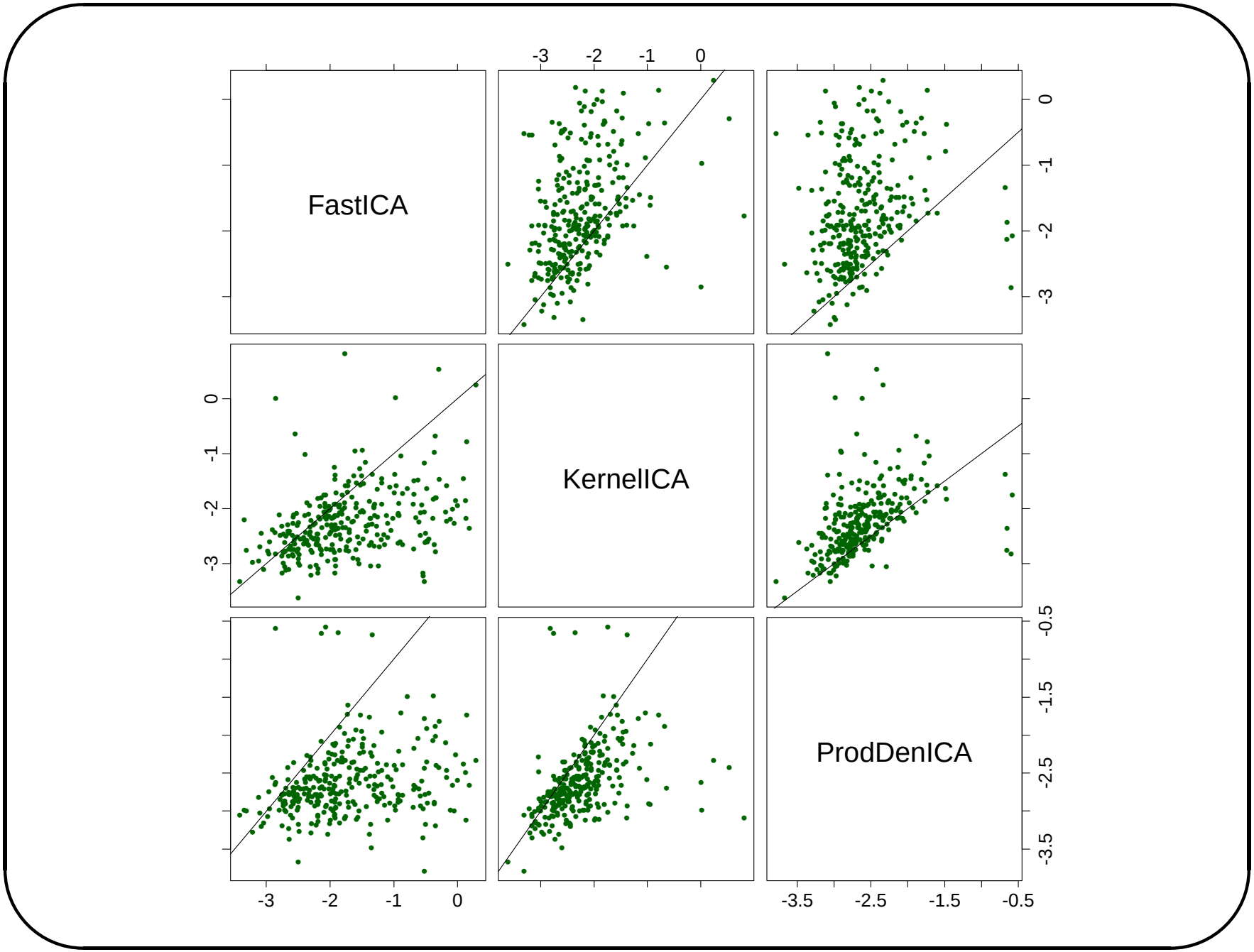
where $r_{ij} = (\mathbf{W}_0 \mathbf{W}^{-1})_{ij}$.

2-Dim examples



4-dim examples




FastICA
KernelICA
ProdDenICA
-3
-2
-1
0
-3.5
-2.5
-1.5
-0.5