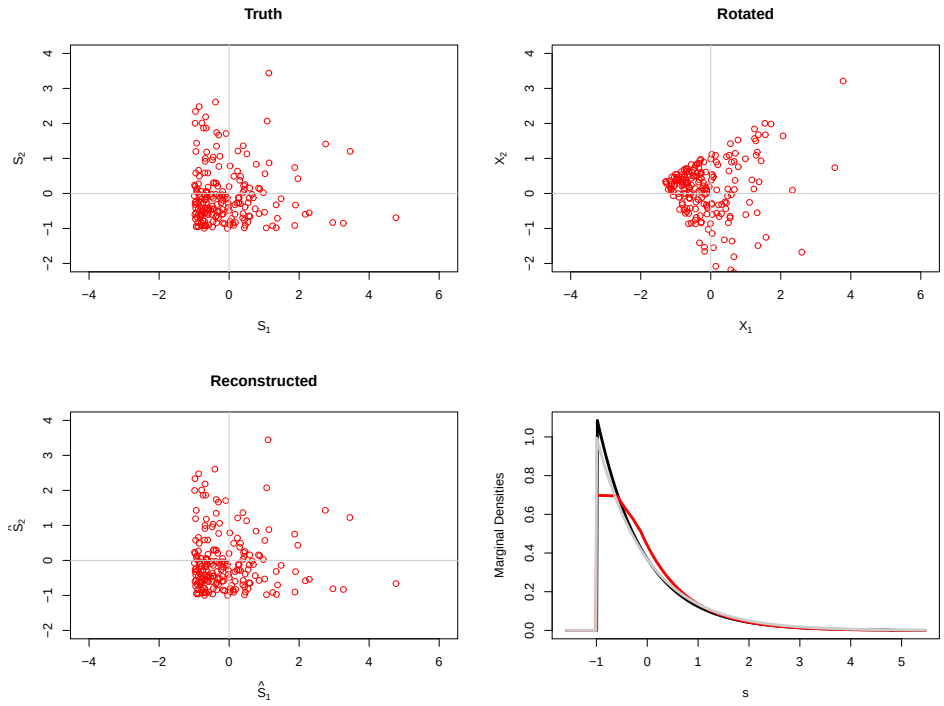
Independent Component Analysis via nonparametric maximum likelihood estimation



Richard Samworth, University of Cambridge

Joint work with Ming Yuan

What are ICA models?

ICA is a special case of a *blind source separation* problem, where from a set of mixed signals, we aim to infer both the source signals and mixing process; e.g. cocktail party problem.

It was pioneered by Comon (1994), and has become enormously popular in signal processing, machine learning, medical imaging...

Mathematical definition

In the simplest, noiseless case, we observe replicates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of

$$\underset{d\times 1}{X} = \underset{d\times d}{A}\ \underset{d\times 1}{S},$$

where the *mixing* matrix A is invertible and S has independent components. Our main aim is to estimate the *unmixing* matrix $W = A^{-1}$; estimation of marginals $P_1, \ldots, P_d$ of $S = (S_1, \ldots, S_d)$ is a secondary goal.

This semiparametric model is therefore related to PCA.

Different previous approaches

- **Postulate parametric family for marginals $P_1, \ldots, P_d$; optimise contrast function involving $(W, P_1, \ldots, P_d)$. Contrast usually represents mutual information or maximum entropy; or non-Gaussianity (Eriksson et al., 2000, Karvanen et al., 2000).**
- **Postulate smooth (log) densities for marginals (Bach and Jordan, 2002; Hastie and Tibshirani, 2003; Samarov and Tsybakov, 2004, Chen and Bickel, 2006).**

Our approach

S. and Yuan (2012)

To avoid assumptions of existence of densities, and choice of tuning parameters, we propose to maximise the log-likelihood

$$\log|\det W| + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\log f_j(w_j^{\mathsf{T}}\mathbf{x}_i)$$

over all $d \times d$ **non-singular matrices** $W = (w_1, \ldots, w_d)^{\mathsf{T}}$**, and univariate log-concave densities** $f_1, \ldots, f_d$**.**

To understand how this works, we need to understand log-concave ICA projections.

Notation

Let $\mathcal{P}_k$ be the set of probability distributions P on $\mathbb{R}^k$ with $\int_{\mathbb{R}^k} \|x\| \, dP(x) < \infty$ and $P(H) < 1$ for all hyperplanes H.

Let $\mathcal{F}_k$ be the set of upper semi-continuous log-concave densities on $\mathbb{R}^k$. The condition $P \in \mathcal{P}_d$ is necessary and sufficient for the existence of a unique log-concave projection $\psi^* : \mathcal{P}_d \to \mathcal{F}_d$ given by

$$\psi^*(P) = \operatorname*{argmax}_{f \in \mathcal{F}_d} \int_{\mathbb{R}^d} \log f \, dP.$$

(Cule, S. and Stewart, 2010; Cule and S., 2010; Dümbgen, S., Schuhmacher, 2011).

ICA notation

Let $\mathcal{W}$ be the set of $d \times d$ invertible matrices. The ICA model $\mathcal{P}_d^{\mathrm{ICA}}$ consists of those $P \in \mathcal{P}_d$ with

$$P(B) = \prod_{j=1}^{d} P_j(w_j^{\mathsf{T}} B), \quad \forall \text{ Borel } B,$$

for some $W \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$.

The log-concave ICA model $\mathcal{F}_d^{\mathrm{ICA}}$ consists of $f \in \mathcal{F}_d$ with

$$f(x) = |\det W| \prod_{j=1}^{d} f_j(w_j^{\mathsf{T}} x) \text{ with } W \in \mathcal{W}, f_1, \ldots, f_d \in \mathcal{F}_1.$$

If X has density $f \in \mathcal{F}_d^{\mathrm{ICA}}$, then $w_j^{\mathsf{T}} X$ has density f_j.

Log-concave ICA projections

Let

$$\psi^{**}(P) = \operatorname*{argmax}_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP.$$

We also write $L^{**}(P) = \sup_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP$**.**

The condition $P \in \mathcal{P}_d$ **is necessary and sufficient for** $L^{**}(P) \in \mathbb{R}$ **and then** $\psi^{**}(P)$ **defines a non-empty, proper subset of** $\mathcal{F}_d^{\mathrm{ICA}}$**.**

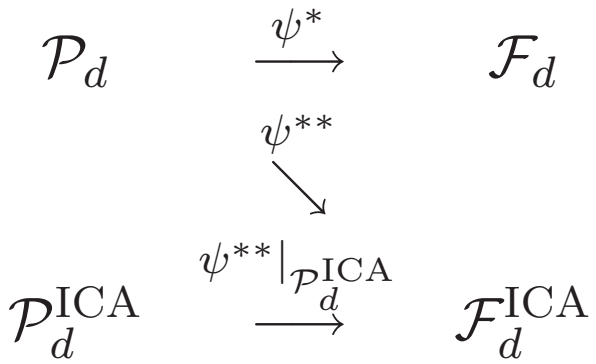
An example

Suppose P is the uniform distribution on the unit Euclidean disk in $\mathbb{R}^2$.

Then $\psi^{**}(P)$ consists of those $f \in \mathcal{F}_d^{\mathrm{ICA}}$ that can be represented by an arbitrary $W \in \mathcal{W}$ and

$$f_1(x) = f_2(x) = \frac{2}{\pi}(1 - x^2)^{1/2} \mathbb{1}_{\{x \in [-1,1]\}}.$$

Schematic picture of maps



Log-concave ICA projection on $\mathcal{P}_d^{\mathrm{ICA}}$

If $P \in \mathcal{P}_d^{\mathrm{ICA}}$, then $\psi^{**}(P)$ defines a unique element of $\mathcal{F}_d^{\mathrm{ICA}}$. The map $\psi^{**}|_{\mathcal{P}_d^{\mathrm{ICA}}}$ coincides with $\psi^{*}|_{\mathcal{P}_d^{\mathrm{ICA}}}$. Moreover, suppose that $P \in \mathcal{P}_d^{\mathrm{ICA}}$, so that

$$P(B) = \prod_{j=1}^{d} P_j(w_j^{\mathsf{T}} B), \quad \forall \text{ Borel } B,$$

for some $W \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. Then

$$f^{**}(x) := \psi^{**}(P)(x) = |\det W| \prod_{j=1}^{d} f_j^{*}(w_j^{\mathsf{T}} x),$$

where $f_j^{*} = \psi^{*}(P_j)$.

Identifiability

Comon (1994), Eriksson and Koivunen (2004)

Suppose a probability measure P on $\mathbb{R}^d$ satisfies

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\mathsf{T} B) = \prod_{j=1}^{d} \tilde{P}_j(\tilde{w}_j^\mathsf{T} B) \quad \forall \text{ Borel } B,$$

where $W, \tilde{W} \in \mathcal{W}$ and $P_1, \ldots, P_d, \tilde{P}_1, \ldots, \tilde{P}_d$ are probability measures on $\mathbb{R}$. Then there exists a permutation π and scaling vector $\epsilon \in (\mathbb{R} \setminus \{0\})^d$ such that $\tilde{P}_j(B_j) = P_{\pi(j)}(\epsilon_j B_j)$ and $\tilde{w}_j = \epsilon_j^{-1} w_{\pi(j)}$ iff none of $P_1, \ldots, P_d$ is a Dirac mass and not more than one of them is Gaussian.

Consequence: If $P \in \mathcal{P}_d^{ICA}$, then $\psi^{**}(P)$ is identifiable iff P is identifiable.

Convergence

Suppose that $P, P^1, P^2, \ldots \in \mathcal{P}_d$ satisfy $d(P^n, P) \to 0$, where d denotes Wasserstein distance. Then

$$\sup_{f^n \in \psi^{**}(P^n)} \inf_{f \in \psi^{**}(P)} \int_{\mathbb{R}^d} |f^n - f| \to 0.$$

If $P \in \mathcal{P}_d^{\mathrm{ICA}}$ is identifiable and $(W, P_1, \ldots, P_d) \overset{\mathrm{ICA}}{\sim} P$, then

$$\sup_{f^n \in \psi^{**}(P^n)} \sup_{(W^n, f_1^n, \ldots, f_d^n) \overset{\mathrm{ICA}}{\sim} f^n} \inf_{\pi^n \in \Pi_d} \inf_{\epsilon_1^n, \ldots, \epsilon_d^n \in \mathbb{R} \setminus \{0\}}$$

$$\left\{ \|(\epsilon_j^n)^{-1} w^n_{\pi^n(j)} - w_j\| + \int_{-\infty}^{\infty} \left| |\epsilon_j^n| f^n_{\pi^n(j)}(\epsilon_j^n x) - f_j^*(x) \right| dx \right\} \to 0,$$

for each $j = 1, \ldots, d$, where $f_j^* = \psi^*(P_j)$. Consequently, for large n, every $f^n \in \psi^{**}(P^n)$ is identifiable.

Estimation procedure

Now suppose $(W^0, P_1^0, \ldots, P_d^0) \overset{\mathrm{ICA}}{\sim} P^0 \in \mathcal{P}_d^{\mathrm{ICA}}$**, and we have data** $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} P^0$ **with** $n \geq d+1$**.**

We propose to estimate P^0 **by** $\psi^{**}(\hat{P}^n)$**, where** $\hat{P}^n$ **is the empirical distribution of the data. That is, we maximise**

$$\ell^n(W, f_1, \ldots, f_d) = \log|\det W| + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\log f_j(w_j^{\mathsf{T}}\mathbf{x}_i)$$

over $W \in \mathcal{W}$ **and** $f_1, \ldots, f_d \in \mathcal{F}_1$**.**

Consistency

Suppose P^0 is identifiable. For any maximiser $(\hat{W}^n, \hat{f}_1^n, \ldots, \hat{f}_d^n)$ of $\ell^n(W, f_1, \ldots, f_d)$, there exist $\hat{\pi}^n \in \Pi_d$ and $\hat{\epsilon}_1^n, \ldots, \hat{\epsilon}_d^n \in \mathbb{R} \setminus \{0\}$ such that

$$(\hat{\epsilon}_j^n)^{-1} \hat{w}^n_{\hat{\pi}^n(j)} \xrightarrow{a.s.} w_j^0 \text{ and } \int_{-\infty}^{\infty} \big| |\hat{\epsilon}_j^n| \hat{f}^n_{\hat{\pi}^n(j)}(\hat{\epsilon}_j^n x) - f_j^*(x) \big| \, dx \xrightarrow{a.s.} 0,$$

for $j = 1, \ldots, d$, where $f_j^* = \psi^*(P_j^0)$.

Pre-whitening

Pre-whitening is a standard pre-processing step in ICA algorithms to improve stability. We replace the data with $\mathbf{z}_1 = \hat{\Sigma}^{-1/2}\mathbf{x}_1, \ldots, \mathbf{z}_n = \hat{\Sigma}^{-1/2}\mathbf{x}_n$**, and maximise the log-likelihood over** $O \in O(d)$ **and** $g_1, \ldots, g_d \in \mathcal{F}_1$**.**

If $(\hat{O}^n, \hat{g}_1^n, \ldots, \hat{g}_d^n)$ **is a maximiser, we then set** $\hat{\hat{W}}^n = \hat{O}^n\hat{\Sigma}^{-1/2}$ **and** $\hat{\hat{f}}_j^n = \hat{g}_j^n$**.**

Thus to estimate the d^2 **parameters of** W^0**, we first estimate the** $d(d+1)/2$ **free parameters of** Σ**, then maximise over the** $d(d-1)/2$ **free parameters of** O**.**

Equivalence of pre-whitened algorithm

Suppose P^0 is identifiable and $\int_{\mathbb{R}^d} \|x\|^2 \, dP^0(x) < \infty$. With probability 1 for large n, a maximiser $(\hat{\hat{W}}^n, \hat{\hat{f}}_1^n, \ldots, \hat{\hat{f}}_d^n)$ of $\ell^n(W, f_1, \ldots, f_d)$ over $W \in O(d)\hat{\Sigma}^{-1/2}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$ exists. For any such maximiser, there exist $\hat{\hat{\pi}}^n \in \Pi_d$ and $\hat{\hat{\epsilon}}_1^n, \ldots, \hat{\hat{\epsilon}}_d^n \in \mathbb{R} \setminus \{0\}$ such that

$$(\hat{\hat{\epsilon}}_j^n)^{-1} \hat{\hat{w}}^n_{\hat{\hat{\pi}}^n(j)} \xrightarrow{a.s.} w_j^0 \quad \text{and} \quad \int_{-\infty}^{\infty} \big| |\hat{\hat{\epsilon}}_j^n| \hat{\hat{f}}^n_{\hat{\hat{\pi}}^n(j)}(\hat{\hat{\epsilon}}_j^n x) - f_j^*(x) \big| \, dx \xrightarrow{a.s.} 0,$$

where $f_j^* = \psi^*(P_j^0)$.

Computational algorithm

With (pre-whitened) data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, consider maximising

$$\ell^n(W, f_1, \ldots, f_d)$$

over $W \in O(d)$ and $f_1, \ldots, f_d \in \mathcal{F}_1$.

(1) Initialise W according to Haar measure on $O(d)$

(2) For $j = 1, \ldots, d$, update f_j with the log-concave MLE of $w_j^\mathsf{T}\mathbf{x}_1, \ldots, w_j^\mathsf{T}\mathbf{x}_n$ (Dümbgen and Rufibach, 2011)

(3) Update W using projected gradient step

(4) Repeat (2) and (3) until negligible relative change in log-likelihood.

Projected gradient step

The set $SO(d)$ is a $d(d-1)/2$-dimensional Riemannian submanifold of $\mathbb{R}^{d^2}$. The tangent space at $W \in SO(d)$ is $T_W SO(d) := \{WY : Y = -Y^\mathsf{T}\}$.

The unique geodesic passing through $W \in SO(d)$ with tangent vector WY (where $Y = -Y^\mathsf{T}$) is the map $\alpha : [0,1] \to SO(d)$ given by $\alpha(t) = W \exp(tY)$, where $\exp$ is the usual matrix exponential.

Projected gradient step 2

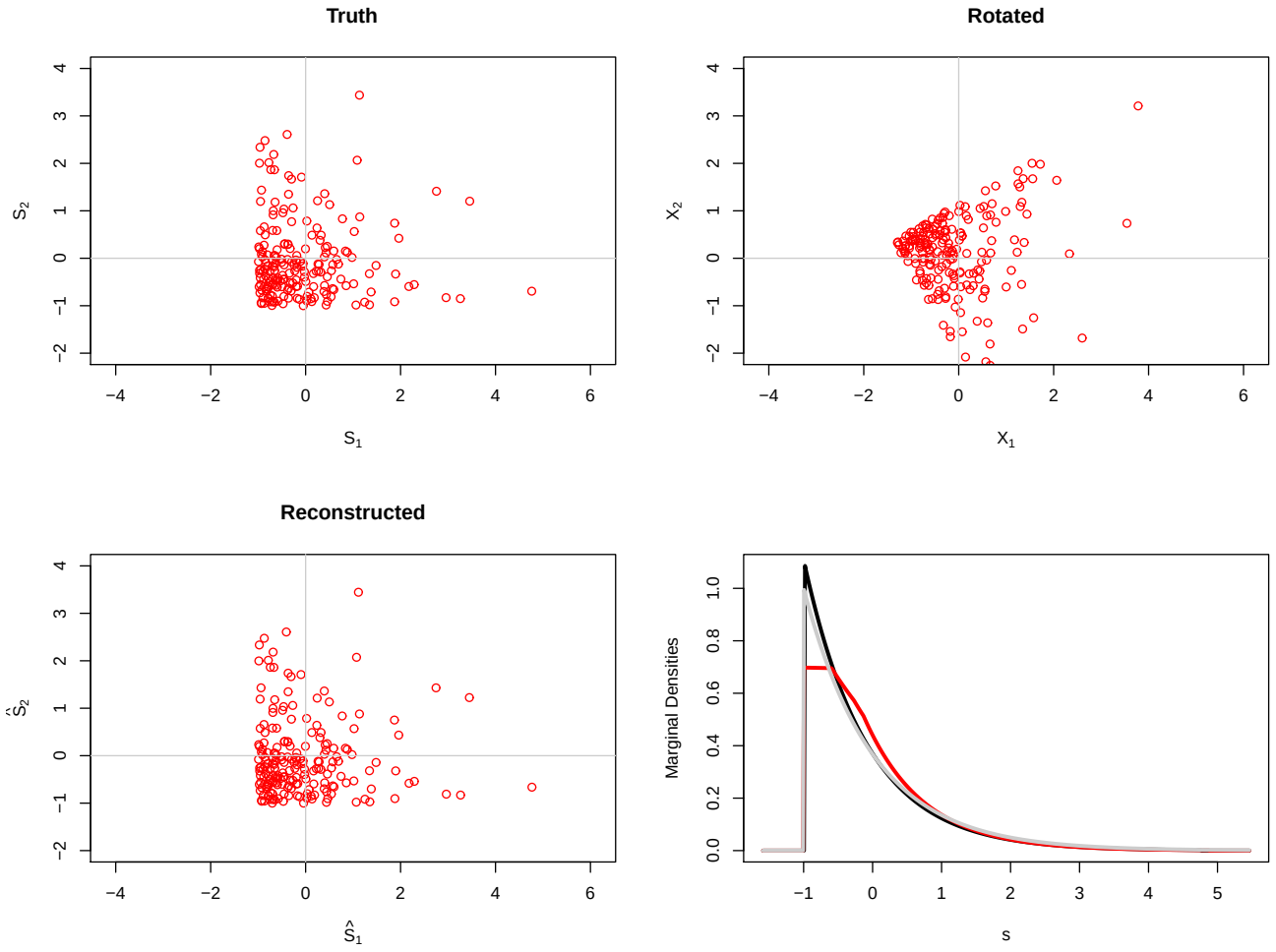
On $[\min(w_j^\mathsf{T}\mathbf{x}_1, \ldots, w_j^\mathsf{T}\mathbf{x}_n), \max(w_j^\mathsf{T}\mathbf{x}_1, \ldots, w_j^\mathsf{T}\mathbf{x}_n)]$**, we have**

$$\log f_j(x) = \min_{k=1,\ldots,m_j} (b_{jk}x - \beta_{jk}).$$

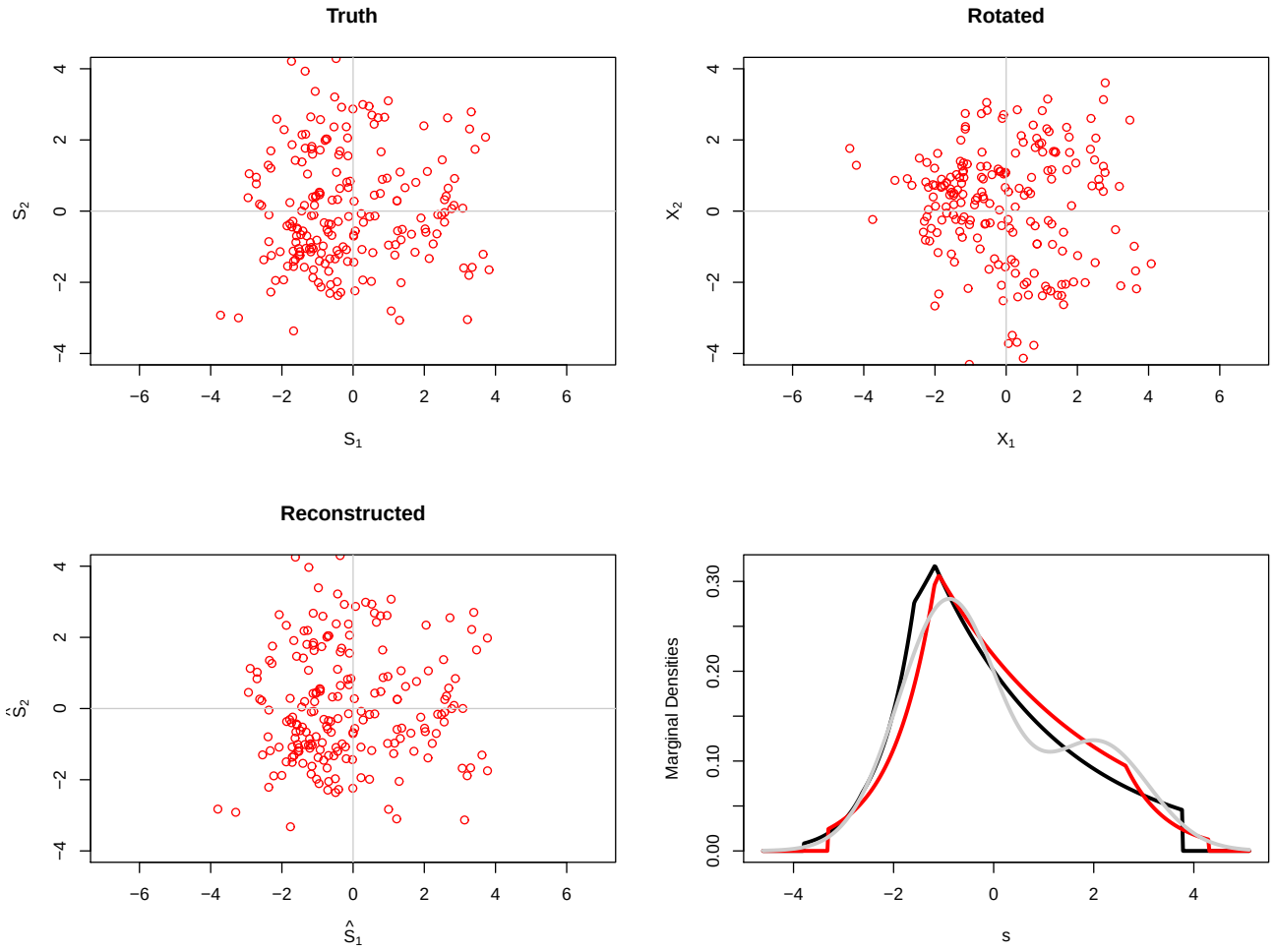
For $1 < s < r < d$**, let** $Y_{r,s}$ **denote the** $d \times d$ **matrix with** $Y_{r,s}(r,s) = 1/\sqrt{2}$**,** $Y_{r,s}(s,r) = -1/\sqrt{2}$ **and zero otherwise. Then** $\mathcal{Y}^+ = \{Y_{r,s} : 1 < s < r < d\}$ **forms an o.n.b. for the skew-symmetric matrices. Let** $\mathcal{Y}^- = \{-Y : Y \in \mathcal{Y}^+\}$**. Choose** $Y^{\max} \in \mathcal{Y}^+ \cup \mathcal{Y}^-$ **to maximise the one-sided directional derivative** $\nabla_{WY} g(W)$**, where**

$$g(W) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \min_{k=1,\ldots,m_j} (b_{jk}w_j^\mathsf{T}\mathbf{x}_i - \beta_{jk}).$$

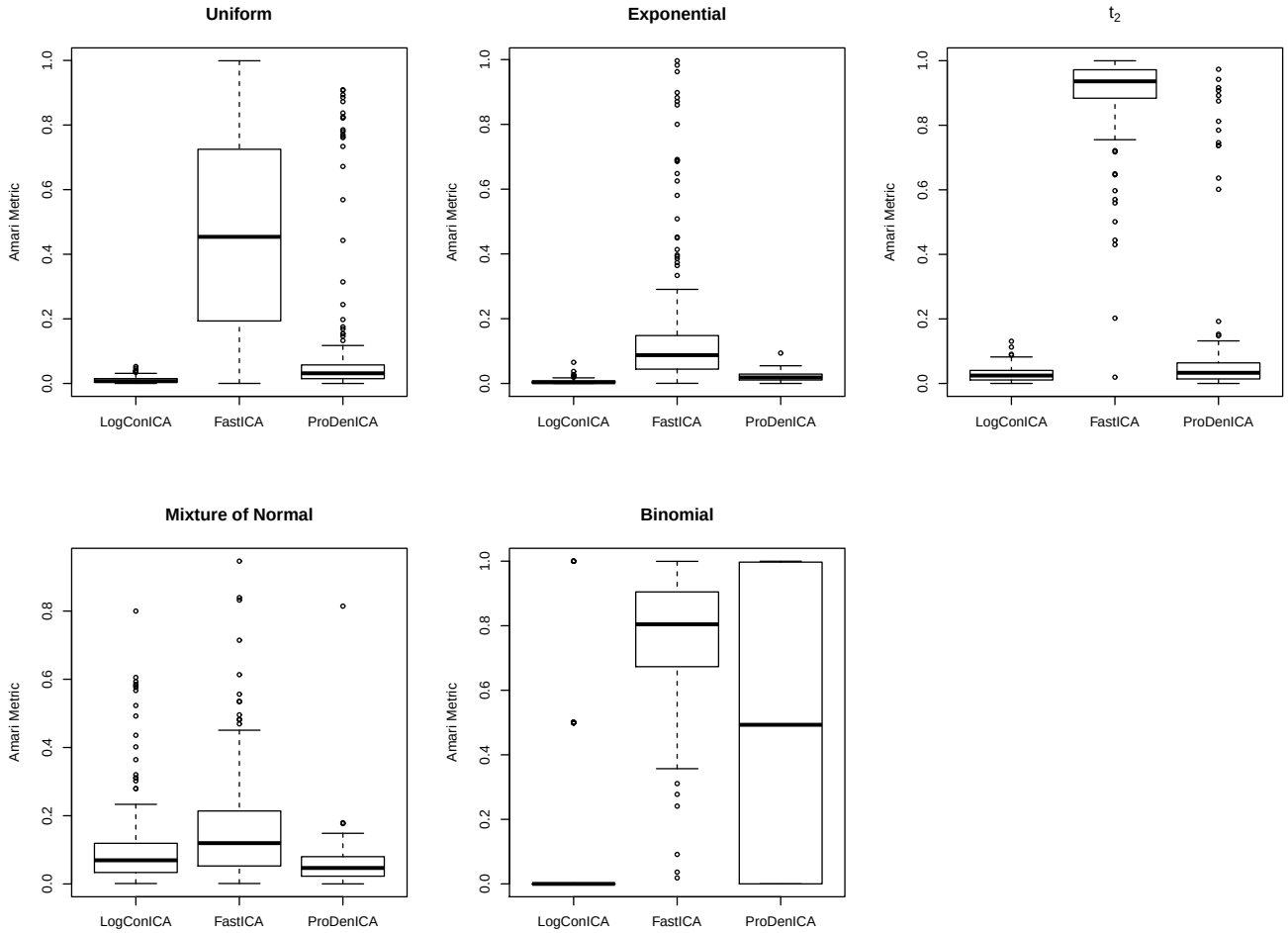
Exp(1)-1


Truth
Rotated
Reconstructed
S_1
S_2
X_1
X_2
$\hat{S}_1$
$\hat{S}_2$
Marginal Densities
s

$$0.7N(-0.9, 1) + 0.3N(2.1, 1)$$



Performance comparison



References

- **Bach, F., Jordan, M. I. (2002) Kernel independent component analysis. *Journal of Machine Learning Research*, 3, 1–48.**
- **Chen, A. and Bickel, P. J. (2006) Efficient independent component analysis, *The Annals of Statistics*, 34, 2825–2855.**
- **Comon, P. (1994) Independent component analysis, A new concept? *Signal Proc.*, 36, 287–314.**
- **Cule, M., Samworth, R. (2010) Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. Electron. J. Stat., 4, 254-270.**
- **Cule, M., Samworth, R. and Stewart, M. (2010), Maximum likelihood estimation of a multi-dimensional log-concave density, J. Roy. Statist. Soc., Ser. B. (with discussion), 72, 545-607.**

- **Dümbgen, L. and Rufibach, K. (2011)** `logcondens`**: Computations Related to Univariate Log-Concave Density Estimation.** ***J. Statist. Software*****, 39, 1–28.**

- **Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011) Approximation by log-concave distributions, with applications to regression.** ***Ann. Statist.*****, 39, 702–730.**

- **Eriksson, J. and Koivunen, V. (2004) Identifiability, separability and uniqueness of linear ICA models.** ***IEEE Signal Processing Letters*****, 11, 601–604.**

- **Hastie, T. and Tibshirani, R. (2003) Independent component analysis through product density estimation. In** ***Advances in Neural Information Processing Systems 15 (Becker, S. and Obermayer, K., eds)*****, MIT Press, Cambridge, MA. pp 649–656.**

- **Hastie, T. and Tibshirani, R. (2003)** `ProDenICA`**: Product Density Estimation for ICA using tilted Gaussian density estimates. R package version 1.0.** **`http://cran.r-project.org/web/packages/ProDenICA/`**

- **Samarov, A. and Tsybakov, A. (2004), Nonparametric independent component analysis.** ***Bernoulli*****,**

10, 565–582.

- **Samworth, R. J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. `http://arxiv.org/abs/1206.0457`**