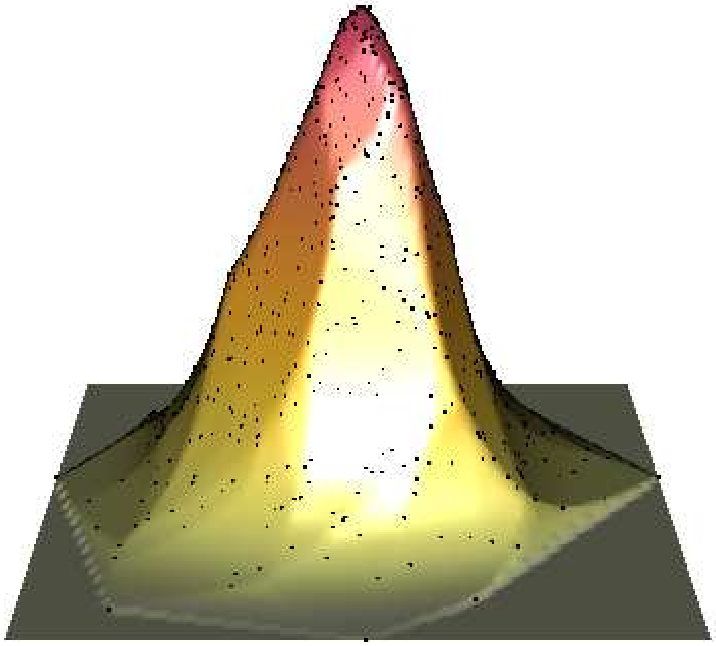
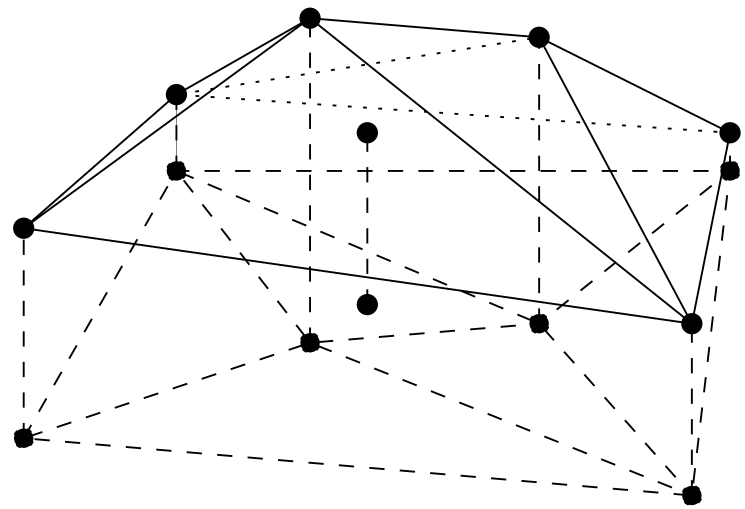
Log-concavity: new theory and methodology





Co-authors: Y. Chen, M. Cule, L. Dümbgen, R. Gramacy

A. Kim, D. Schuhmacher, M. Stewart, M. Yuan

The original problem

Let $X_1, \ldots, X_n$ be a random sample from a density f_0 in $\mathbb{R}^d$.

$$\boxed{\textbf{How should we estimate } f_0\textbf{?}}$$

Two main alternatives:

- **Parametric models: use e.g. MLE. Assumptions often too restrictive.**
- **Nonparametric models: use e.g. kernel density estimate. Choice of bandwidth difficult, particularly for $d > 1$.**

Shape-constrained estimation

Nonparametric shape constraints are becoming increasingly popular (Groeneboom et al. 2001, Walther 2002, Pal et al. 2007, Dümbgen and Rufibach 2009, Schuhmacher et al. 2011, Seregin and Wellner 2010, Koenker and Mizera 2010 . . .)**.** **E.g. log-concavity, r-concavity, k-monotonicity, convexity.**

A density f is log-concave if $\log f$ is concave.

- **Univariate examples: normal, logistic, Gumbel densities, as well as Weibull, Gamma, Beta densities for certain parameter values.**

Characterising log-concave densities

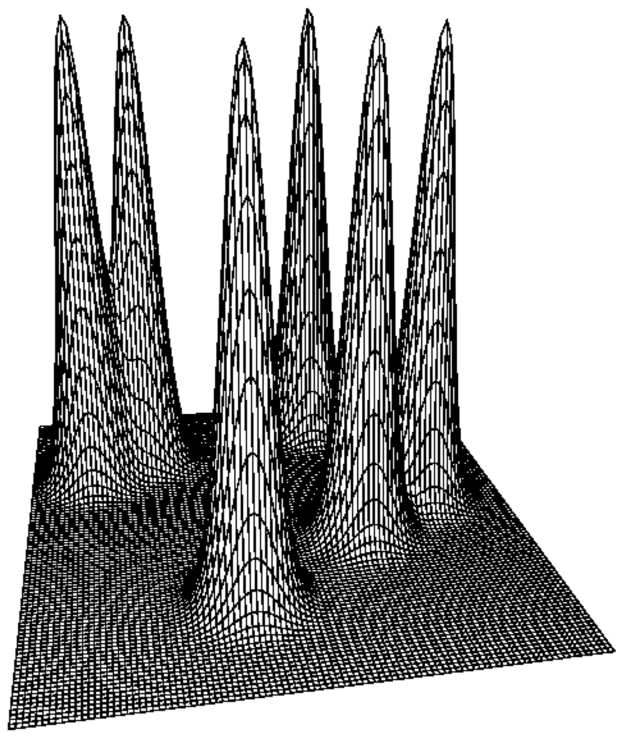
Cule, S. and Stewart (2010)

Let X have density f in $\mathbb{R}^d$. For a subspace V of $\mathbb{R}^d$, let $P_V(x)$ denote the orthogonal projection of x onto V. Then in order that f be log-concave, it is:

1. necessary that for any subspace V, the marginal density of $P_V(X)$ is log-concave (Prékopa 1973), and the conditional density $f_{X|P_V(X)}(\cdot|t)$ of X given $P_V(X) = t$ is log-concave for each t
2. sufficient that, for every $(d-1)$-dimensional subspace V, the conditional density $f_{X|P_V(X)}(\cdot|t)$ of X given $P_V(X) = t$ is log-concave for each t.

Unbounded likelihood!

Consider maximizing the likelihood $L(f) = \prod_{i=1}^{n} f(X_i)$ **over all densities** f**.**



Existence and uniqueness

Walther (2002), Cule, S. and Stewart (2010)

Let $X_1, \ldots, X_n$ be independent with density f_0 in $\mathbb{R}^d$, and suppose that $n \geq d+1$. Then, with probability one, a log-concave maximum likelihood estimator $\hat{f}_n$ exists and is unique.

Sketch of proof

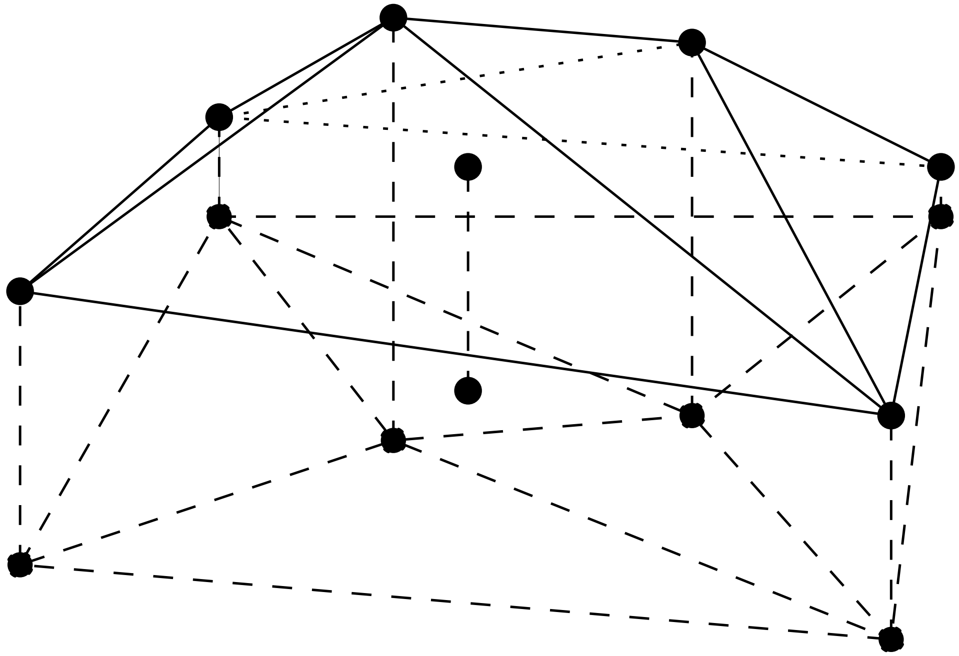
Consider maximising over all log-concave *functions*

$$\psi_n(f) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i) - \int_{\mathbb{R}^d} f(x)\,dx.$$

Any maximiser $\hat{f}_n$ must satisfy:

1. $\hat{f}_n(x) > 0$ **iff** $x \in C_n \equiv \mathrm{conv}(X_1, \ldots, X_n)$
2. **Fix** $y = (y_1, \ldots, y_n)$ **and let** $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ **be the smallest concave function with** $\bar{h}_y(X_i) \geq y_i$ **for all** i**. Then** $\log \hat{f}_n = \bar{h}_{y^*}$ **for some** y^*
3. $\int_{\mathbb{R}^d} \hat{f}_n(x)\,dx = 1$**.**

Schematic diagram of MLE on log scale



Computation

Cule, S. and Stewart (2010), Cule, Gramacy and S. (2009)

First attempt: minimise

$$\tau(y) = -\frac{1}{n}\sum_{i=1}^{n} \bar{h}_y(X_i) + \int_{C_n} \exp\{\bar{h}_y(x)\}\, dx.$$

Computation

Cule, S. and Stewart (2010), Cule, Gramacy and S. (2009)

First attempt: minimise

$$\tau(y) = -\frac{1}{n}\sum_{i=1}^{n} \bar{h}_y(X_i) + \int_{C_n} \exp\{\bar{h}_y(x)\}\,dx.$$

Better: minimise

$$\sigma(y) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\{\bar{h}_y(x)\}\,dx.$$

Then σ has a *unique* minimum at y^*, say, $\log \hat{f}_n = \bar{h}_{y^*}$ and σ is *convex* . . .

Computation

Cule, S. and Stewart (2010), Cule, Gramacy and S. (2009)

First attempt: minimise

$$\tau(y) = -\frac{1}{n}\sum_{i=1}^{n} \bar{h}_y(X_i) + \int_{C_n} \exp\{\bar{h}_y(x)\}\,dx.$$

Better: minimise

$$\sigma(y) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\{\bar{h}_y(x)\}\,dx.$$

Then σ has a *unique* minimum at y^*, say, $\log \hat{f}_n = \bar{h}_{y^*}$ and σ is *convex* $\ldots$ but *non-differentiable*!

Log-concave projections

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). ■

One-dimensional characterisation

Dümbgen, S. and Schuhmacher (2011)

Let $P_0 \in \mathcal{P}_1$ **have distribution function** F_0**. Let**

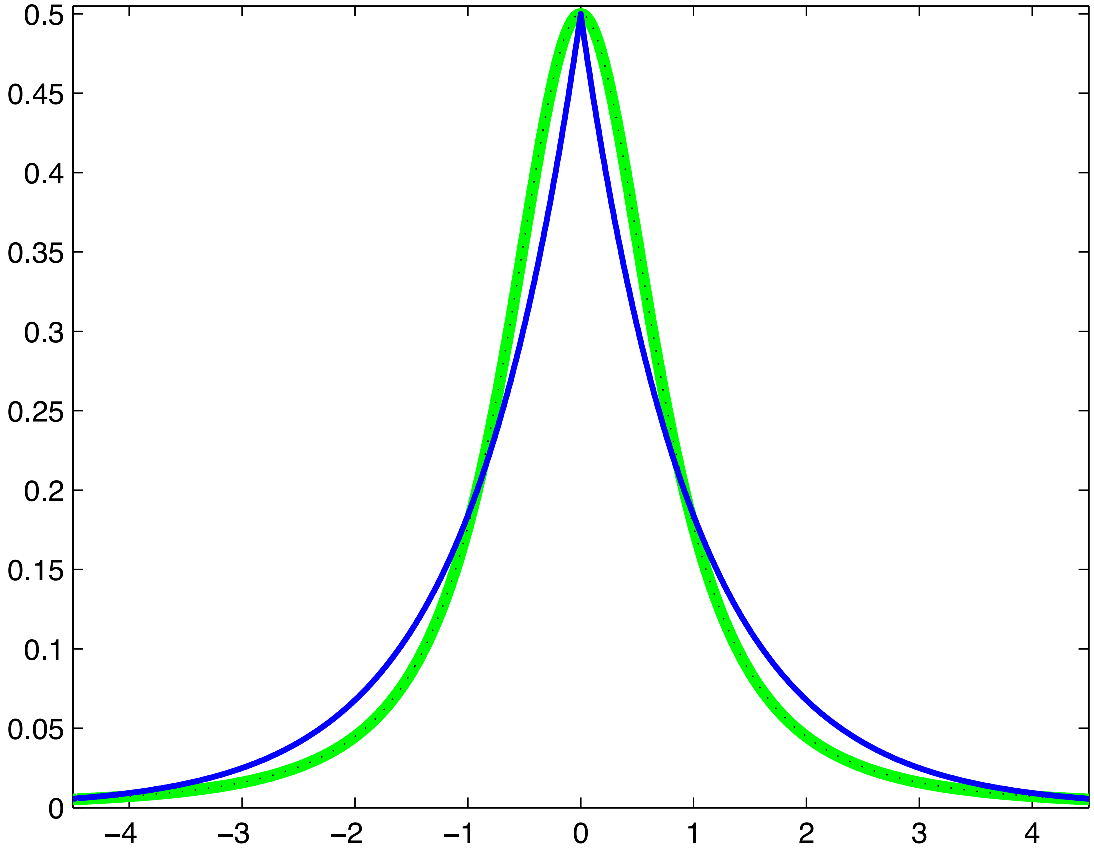
$$S(f^*) = \{x \in \mathbb{R} : \log f^*(x) > \tfrac{1}{2}\log f^*(x-\delta) + \tfrac{1}{2}\log f^*(x+\delta)\ \forall \delta > 0\}.$$

Then the distribution function F^* **of** f^* **is characterised by**

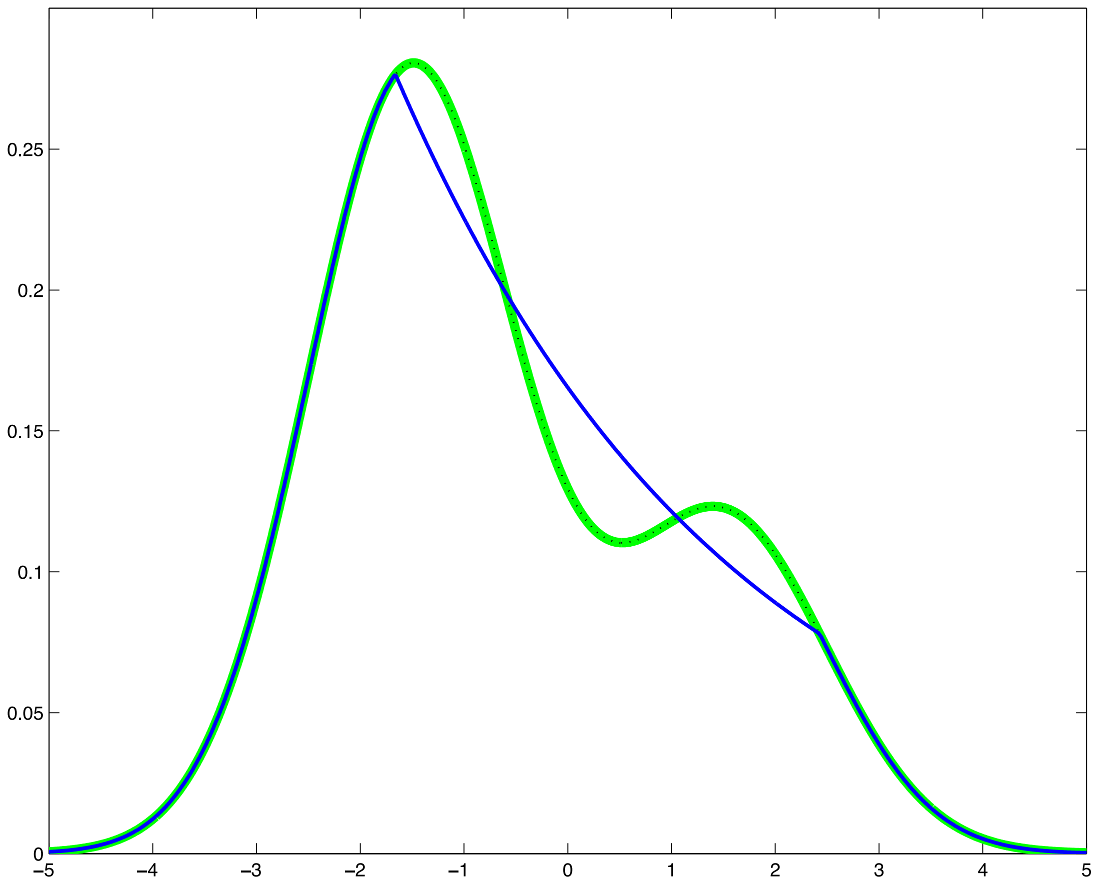
$$\int_{-\infty}^{x} \{F^*(t) - F_0(t)\}\,dt \begin{cases} \leq 0 & \textbf{for all } x \in \mathbb{R} \\ = 0 & \textbf{for all } x \in S(f^*) \cup \{\infty\}. \end{cases}$$

Example 1

Suppose $f_0(x) = \frac{1}{2}(1+x^2)^{-3/2}$**. Then** $f^*(x) = \frac{1}{2}e^{-|x|}$**.**



Example 2



Log-concave projections preserve independence Chen and S. (2012)

Suppose $P \in \mathcal{P}_d$ can be written as $P = P_1 \otimes P_2$, where P_1 and P_2 are probability measures on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, with $d_2 = d - d_1$. If $f^* = \psi^*(P)$ and $f^*_\ell = \psi^*(P_\ell)$ for $\ell = 1, 2$, then

$$f^*(x) = f^*_1(x_1) f^*_2(x_2)$$

for $x = (x_1^T, x_2^T)^T \in \mathbb{R}^d$.

This makes log-concave projections very attractive for independent component analysis (S. and Yuan, 2012).

Convergence of log-concave densities

Cule and S. (2010)

Let (f_n) **be a sequence of log-concave densities on** $\mathbb{R}^d$ **with** $f_n \xrightarrow{d} f$ **for some density** f**. Then:**

(a) f **is log-concave**

(b) $f_n \to f$ **almost everywhere**

(c) Let $a_0 > 0$ **and** $b_0 \in \mathbb{R}$ **be such that** $f(x) \leq e^{-a_0\|x\|+b_0}$**. If** $a < a_0$ **then** $\int e^{a\|x\|}|f_n(x) - f(x)|\,dx \to 0$ **and, if** f **is continuous,** $\sup_x e^{a\|x\|}|f_n(x) - f(x)| \to 0$**.**

Theoretical properties

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

Now let $X_1, \ldots, X_n \stackrel{iid}{\sim} P_0 \in \mathcal{P}_d$**, and let** $f^* = \psi^*(P_0)$**. Taking** $a_0 > 0$ **and** $b_0 \in \mathbb{R}$ **such that** $f^*(x) \leq e^{-a_0\|x\|+b_0}$**, we have for any** $a < a_0$ **that**

$$\int_{\mathbb{R}^d} e^{a\|x\|} |\hat{f}_n(x) - f^*(x)| \, dx \stackrel{a.s.}{\rightarrow} 0,$$

and, if f^* **is continuous,** $\sup_x e^{a\|x\|}|\hat{f}_n(x) - f^*(x)| \stackrel{a.s.}{\rightarrow} 0$**.**

Pointwise asymptotic distribution $(d = 1)$

Balabdaoui, Rufibach and Wellner (2009)

Suppose f_0 is log-concave and let $k \geq 2$ be the smallest integer such that $\phi_0 := \log f_0$ is k times continuously differentiable in a neighbourhood of x_0 with $\phi_0^{(j)}(x_0) = 0$ for $j = 2, \ldots, k-1$ and $\phi_0^{(k)}(x_0) \neq 0$. Then

$$n^{k/(2k+1)}\{\hat{f}_n(x_0) - f_0(x_0)\} \xrightarrow{d} c_k(x_0, \phi_0)H_k^{(2)}(0),$$

where $H_k(t)$ is the 'lower invelope' of an integrated Brownian motion process with drift.

Pointwise minimax lower bound

Seregin and Wellner (2010)

Suppose that $\det(\nabla^2\phi_0(x_0)) < 0$**. Then**

$$\liminf_{n\to\infty} n^{2/(d+4)} \inf_{\tilde{f}_n} \sup_{f\in\mathcal{F}_d} \mathbb{E}|\tilde{f}_n(x_0) - f(x_0)| \geq c(d)\left\{\frac{-\det(\nabla^2\phi_0(x_0))}{f(x_0)^2}\right\}^{1/(d+4)}.$$

Global minimax bounds

Kim and S. (2013)

Let L denote a global loss function (e.g. L_2, L_1, Hellinger, Kullback–Leibler, chi-squared,...). Then

$$\liminf_{n\to\infty} n^{2/(d+4)} \inf_{\tilde{f}_n} \sup_{f\in\mathcal{F}_d} \mathbb{E}\{L(\tilde{f}_n, f)\} > 0.$$

Conversely, for $m > 0$, let $\mathcal{F}(a,b,m,M)$ denote the class of log-concave densities $f : [a,b]^d \to [m,M]$. Then there exists $\tilde{f}_n$ such that

$$\limsup_{n\to\infty} n^{2/(d+4)} \sup_{f\in\mathcal{F}(a,b,m,M)} \mathbb{E}\{L(\tilde{f}_n, f)\} < \infty.$$

Moment (in)equalities

Dümbgen, S. and Schuhmacher (2011)

Let $P \in \mathcal{P}_d$**, let** $f^* = \psi^*(P)$ **and let** $P^*(B) = \int_B f^*$**. Then**

$$\int_{\mathbb{R}^d} x \, dP^*(x) = \int_{\mathbb{R}^d} x \, dP(x)$$

and

$$\int_{\mathbb{R}^d} h \, dP^* \leq \int_{\mathbb{R}^d} h \, dP$$

for all convex $h : \mathbb{R}^d \to (-\infty, \infty]$**.**

Smoothed log-concave density estimator

Dümbgen and Rufibach (2009), Cule, S. and Stewart (2010), Chen and S. (2012)

Let

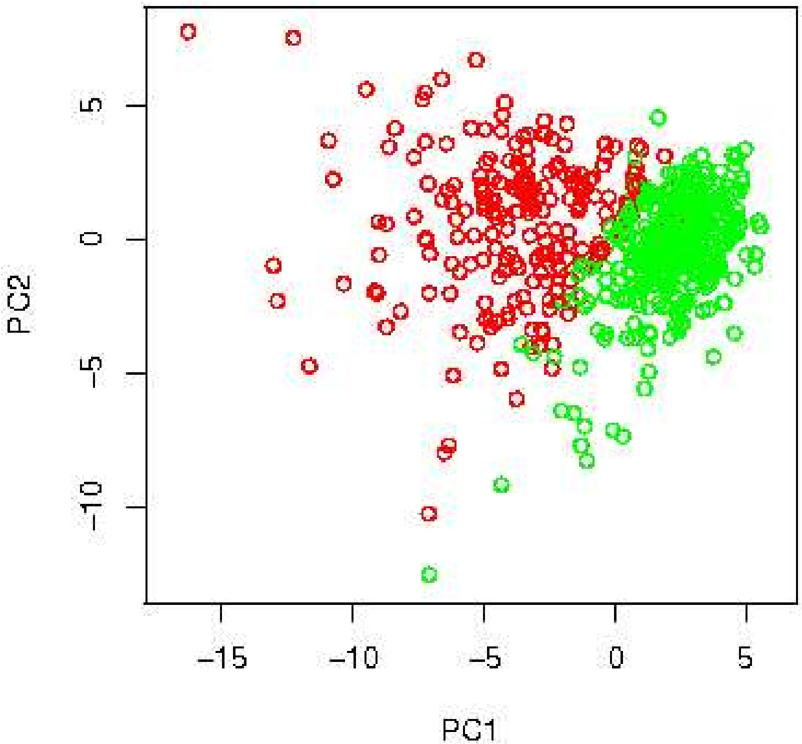
$$\tilde{f}_n = \hat{f}_n * \phi_{\hat{A}},$$

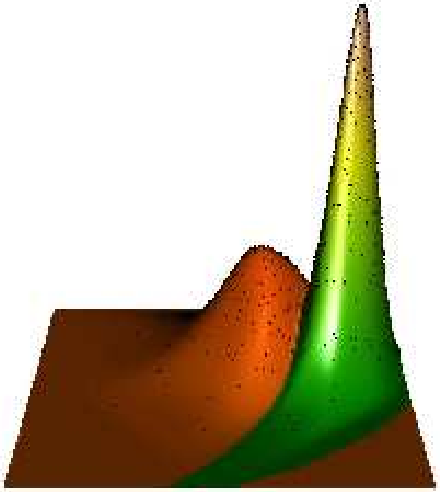
where $\phi_{\hat{A}}$ is a d-dimensional normal density with mean zero and covariance matrix $\hat{A} = \hat{\Sigma} - \tilde{\Sigma}$. Here, $\hat{\Sigma}$ is the sample covariance matrix and $\tilde{\Sigma}$ is the covariance matrix corresponding to $\hat{f}_n$.

Then $\tilde{f}_n$ is a smooth, fully automatic log-concave estimator supported on the whole of $\mathbb{R}^d$ which satisfies the same theoretical properties as $\hat{f}_n$.

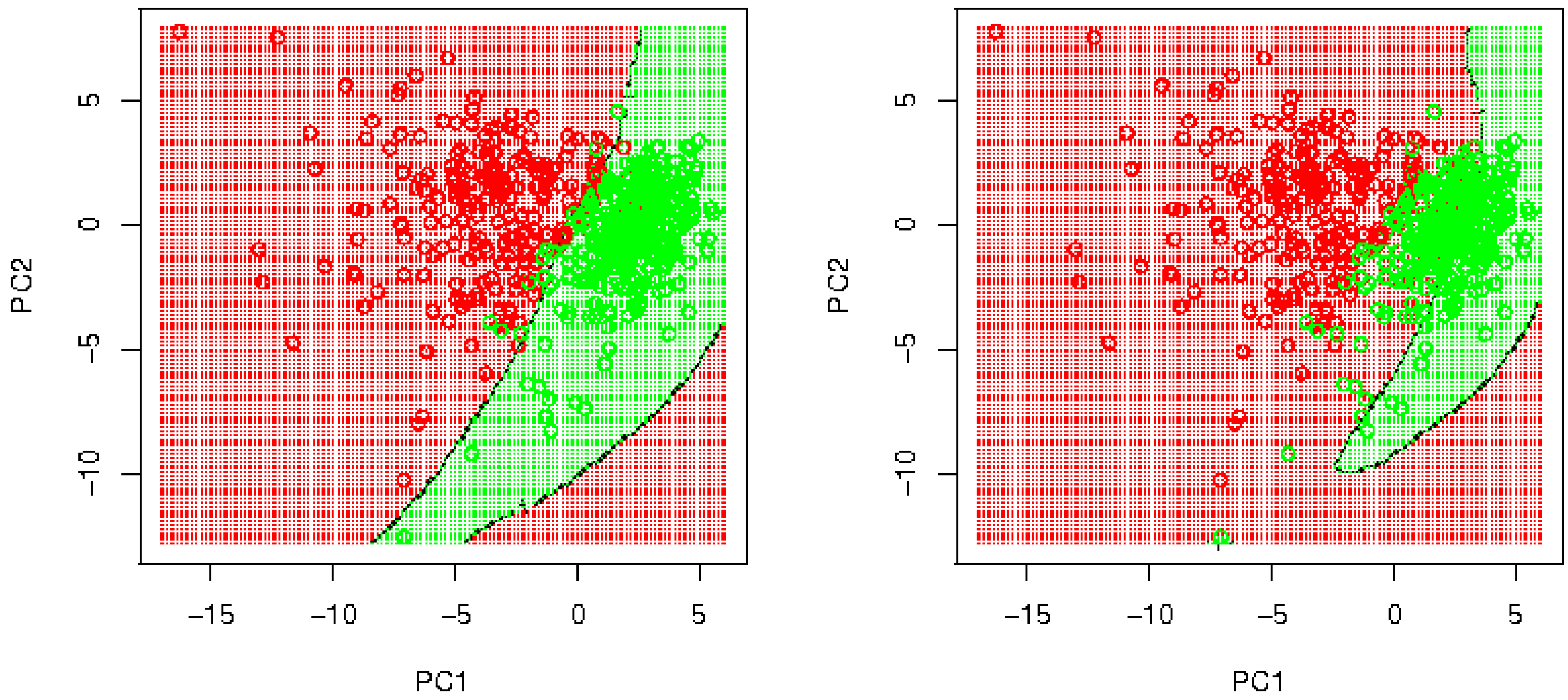
It offers potential improvements for small sample sizes.

Breast cancer data





Classification boundaries



Testing for log-concavity Chen and S. (2012)

Suppose $P_0 \in \mathcal{P}_d$. Then $\mathrm{tr}(A^*) = 0$ if and only if P_0 has a log-concave density.

We can therefore use $\mathrm{tr}(\hat{A})$ as a test statistic, and generate a critical value from bootstrap samples drawn from $\hat{f}_n$.

This test is consistent: if P_0 is not log-concave, then the power converges to 1 as $n \to \infty$.

Regression problems

Dümbgen, S. and Schuhmacher (2011)

Consider the regression model

$$Y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d., log-concave and $\mathbb{E}(\epsilon_i) = 0$. In both of the cases i) μ is linear and ii) μ is isotonic, we can jointly estimate μ and the distribution of ϵ_i.

Significant improvements are obtainable over usual methods when errors are non-normal.

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.

Recap

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**

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

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**

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

If X **has density** $f \in \mathcal{F}_d^{\mathrm{ICA}}$**, then** $w_j^\top X$ **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 $\psi^{**}(P)$ includes all $f \in \mathcal{F}_d^{\mathrm{ICA}}$ that can be represented by an arbitrary orthogonal $W \in \mathcal{W}$ 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

$$
\begin{array}{ccc}
\mathcal{P}_d & \xrightarrow{\psi^*} & \mathcal{F}_d \\
& \overset{\psi^{**}}{\searrow} & \\
\mathcal{P}_d^{\mathrm{ICA}} & \xrightarrow{\psi^{**}|_{\mathcal{P}_d^{\mathrm{ICA}}}} & \mathcal{F}_d^{\mathrm{ICA}}
\end{array}
$$

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**

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

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

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

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

Identifiability

Comon (1994), Eriksson and Koivunen (2004)

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

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

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**

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

$$\biggl\{ \|(\epsilon_j^n)^{-1} w_{\pi^n(j)}^n - w_j\| + \int_{-\infty}^{\infty} \bigl| |\epsilon_j^n| f_{\pi^n(j)}^n(\epsilon_j^n x) - f_j^*(x) \bigr| \, dx \biggr\} \to 0,$$

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 $(W^0, P_1^0, \ldots, P_d^0) \stackrel{\mathrm{ICA}}{\sim} P^0 \in \mathcal{P}_d^{\mathrm{ICA}}$**, and we have data** $\mathbf{x}_1, \ldots, \mathbf{x}_n \stackrel{iid}{\sim} P^0$ **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**

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

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**

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

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

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

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 $w_j^\top \mathbf{x}_1, \ldots, w_j^\top \mathbf{x}_n$ **(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 $T_W SO(d) := \{WY : Y = -Y^\top\}$.

The unique geodesic passing through $W \in SO(d)$ with tangent vector WY (where $Y = -Y^\top$) 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

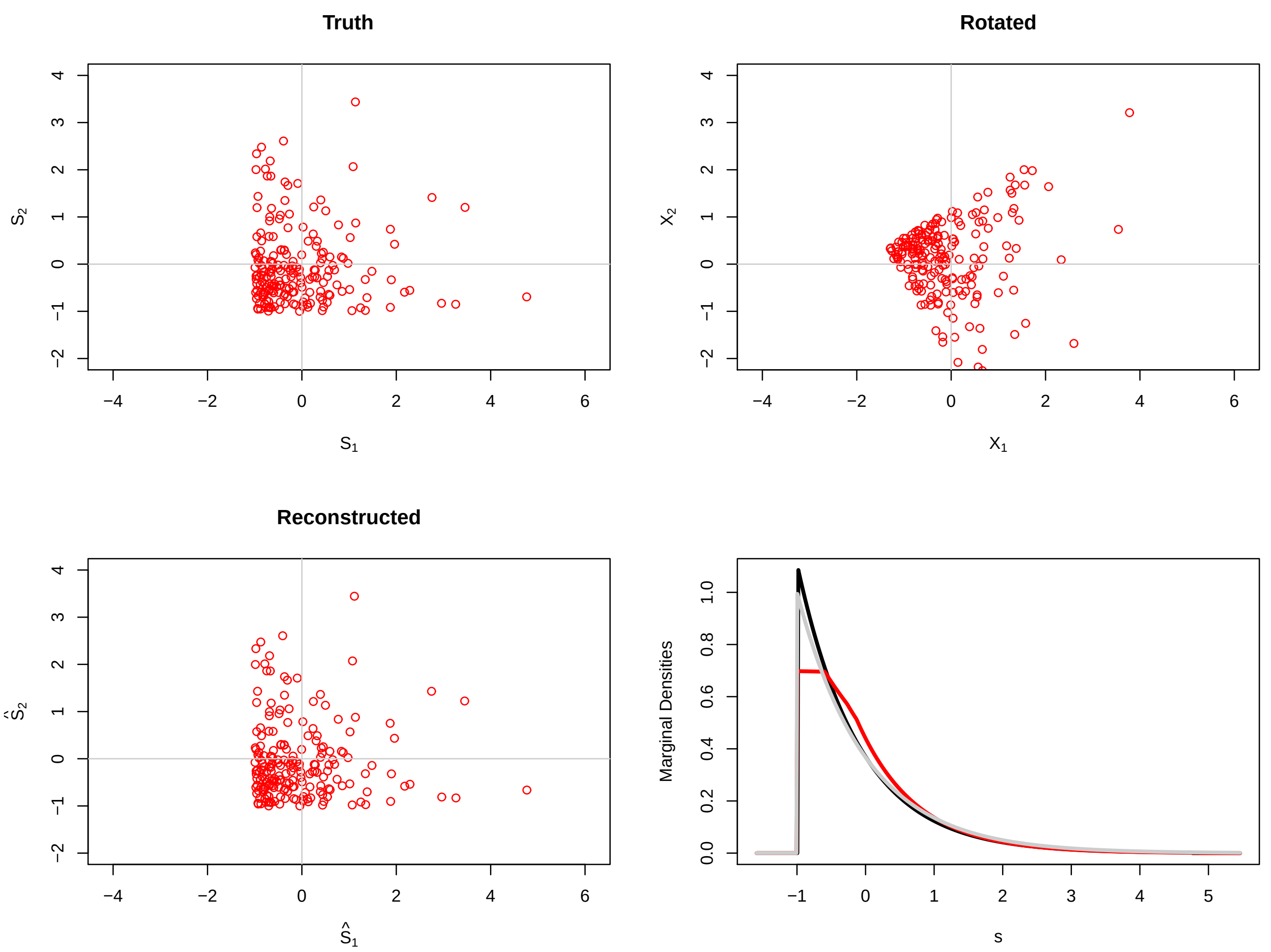
On $[\min(w_j^\top \mathbf{x}_1, \ldots, w_j^\top \mathbf{x}_n), \max(w_j^\top \mathbf{x}_1, \ldots, w_j^\top \mathbf{x}_n)]$**, we have**

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

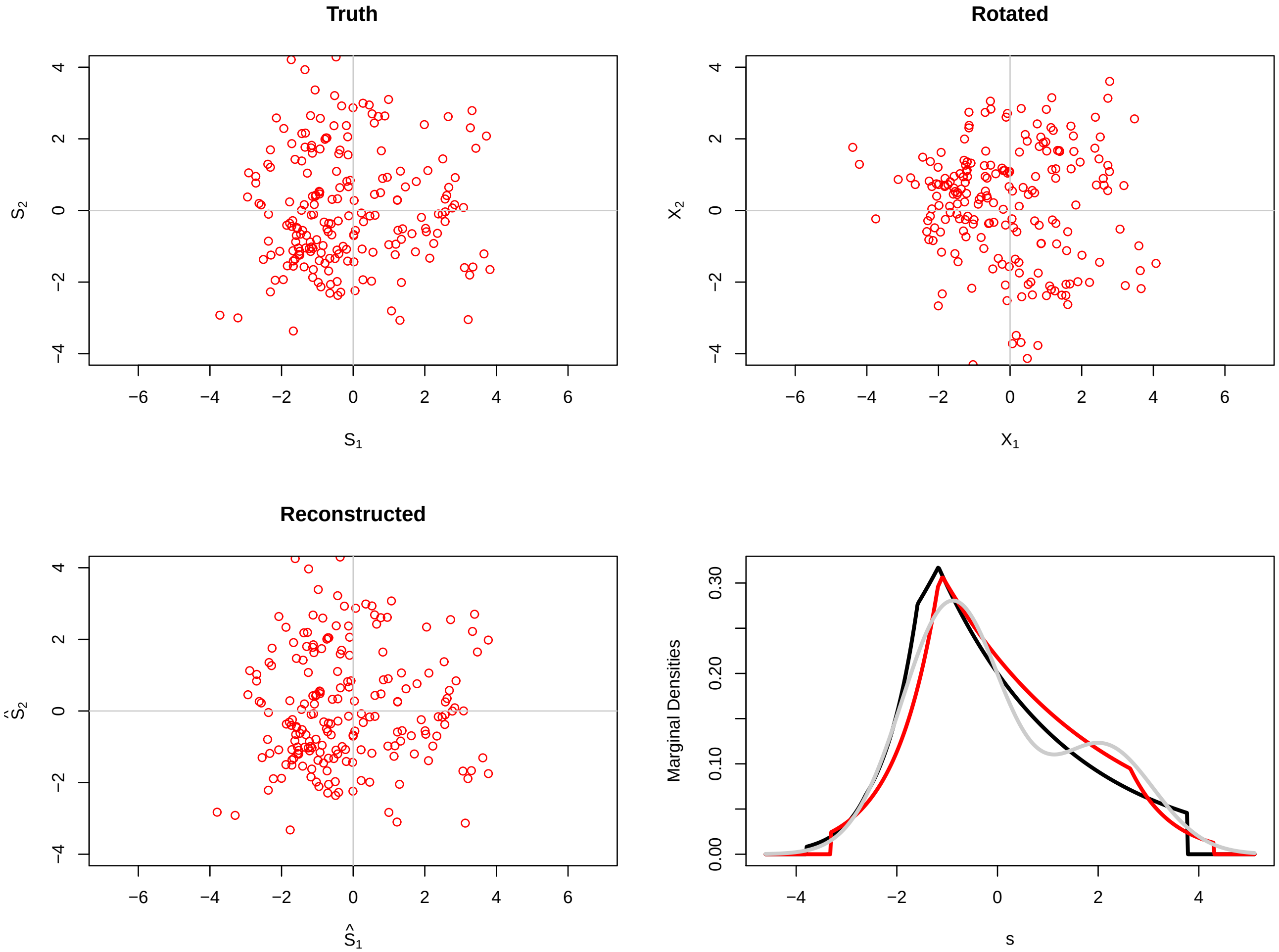
For $1 < s < r < d$**, let** $Y_{r,s}$ **denote the** $d \times d$ **matrix with** $Y_{r,s}(r,s) = 1/\sqrt{2}$**,** $Y_{r,s}(s,r) = -1/\sqrt{2}$ **and zero otherwise. Then** $\mathcal{Y}^+ = \{Y_{r,s} : 1 < s < r < d\}$ **forms an o.n.b. for the skew-symmetric matrices. Let** $\mathcal{Y}^- = \{-Y :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

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

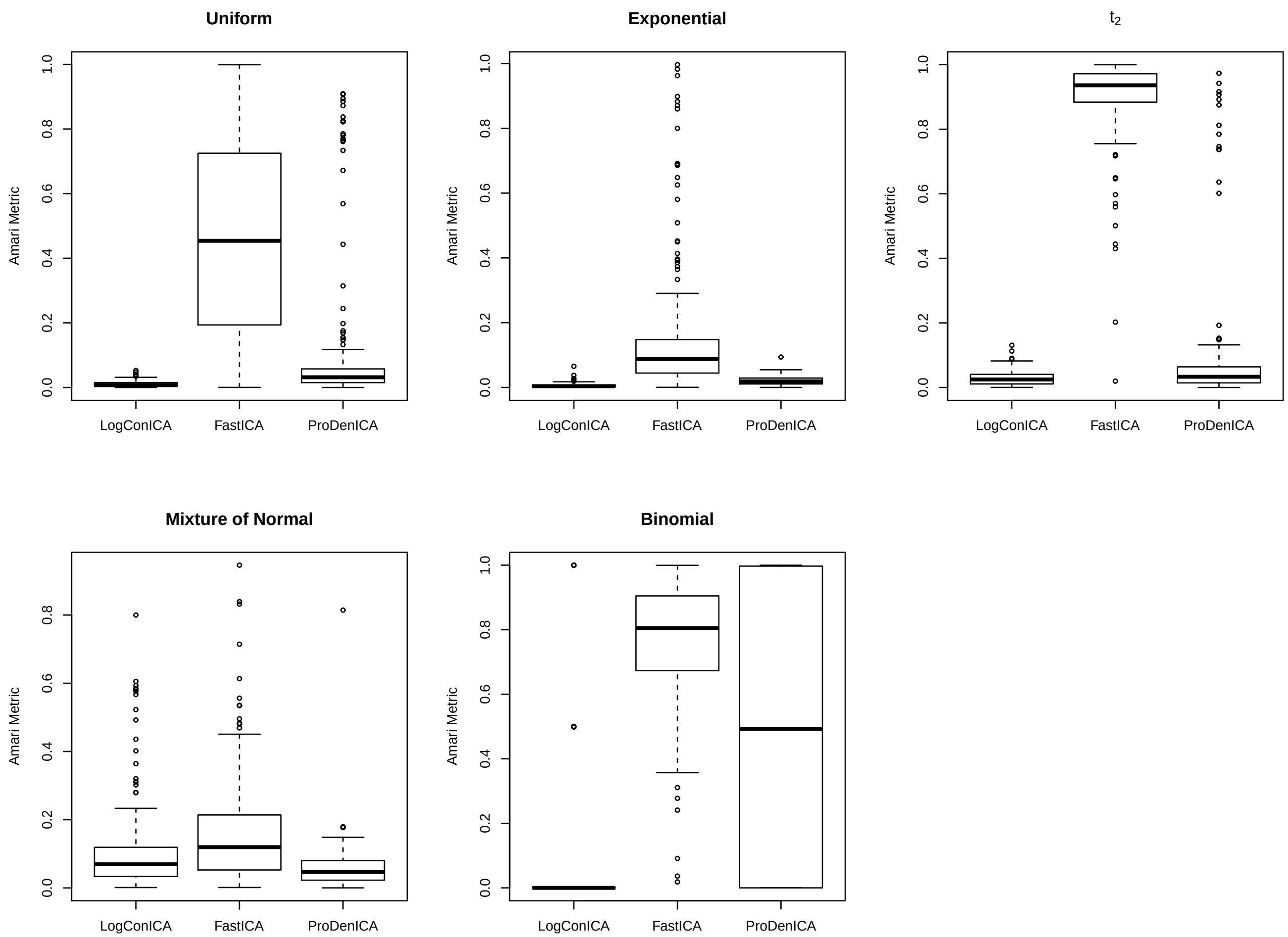
Exp(1)-1


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



Summary

- The log-concave MLE is a fully automatic, nonparametric density estimator
- It has several extensions which can be used in a wide variety of applications, e.g. classification, clustering, functional estimation, regression and Independent Component Analysis problems.
- Many challenges remain: faster algorithms, dependent data, further theoretical results, other applications and constraints,...

References

- Bach, F., Jordan, M. I. (2002) Kernel independent component analysis. *Journal of Machine Learning Research*, 3, 1–48.
- Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009), Limit distribution theory for maximum likelihood estimation of a log-concave density, *Ann. Statist.*, 37, 1299–1331.
- Chen, A. and Bickel, P. J. (2006) Efficient independent component analysis, *The Annals of Statistics*, 34, 2825–2855.
- Chen, Y. and Samworth, R. J. (2012), Smoothed log-concave maximum likelihood estimation with applications, *Statist. Sinica*, to appear.
- Comon, P. (1994) Independent component analysis, A new concept? *Signal Proc.*, 36, 287–314.
- Cule, M., Gramacy, R. and Samworth, R. (2009) LogConcDEAD: an R package for maximum likelihood estimation of a multivariate log-concave density, *J. Statist. Software*, 29, Issue 2.
- Cule, M. and Samworth, R. (2010), Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.*, 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. (2009) Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15, 40–68.

- **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.**
- **Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001) Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.*, 29, 1653–1698.**
- **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.**
- **Kim, A. and Samworth, R. J. (2013) Global minimax bounds for log-concave density estimation. *In preparation.***
- **Koenker, R. and Mizera, I. (2010) Quasi-concave density estimation. *Ann. Statist.*, 38, 2998–3027.**
- **Pal, J., Woodroofe, M. and Meyer, M. (2007) Estimating a Polya frequency function. In *Complex datasets and Inverse problems, Networks and Beyond Tomography*, vol. 54 of *Lecture Notes - Monograph Series*, 239–249. IMS.**
- **Prékopa, A. (1973) On logarithmically concave measures and functions. *Acta Scientarium Mathematicarum*, 34, 335–343.**
- **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, *Ann. Statist.*, to appear.**

- **Schuhmacher, D., Hüsler, A. and Dümbgen, L. (2011) Multivariate log-concave distributions as a nearly parametric model. *Statistics & Risk Modeling*, 28, 277–295.**

- **Seregin, A. and Wellner, J. A. (2010) Nonparametric estimation of convex-transformed densities. *Ann. Statist.*, 38, 3751–3781.**

- **Walther, G. (2002) Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.*, 97, 508–513.**