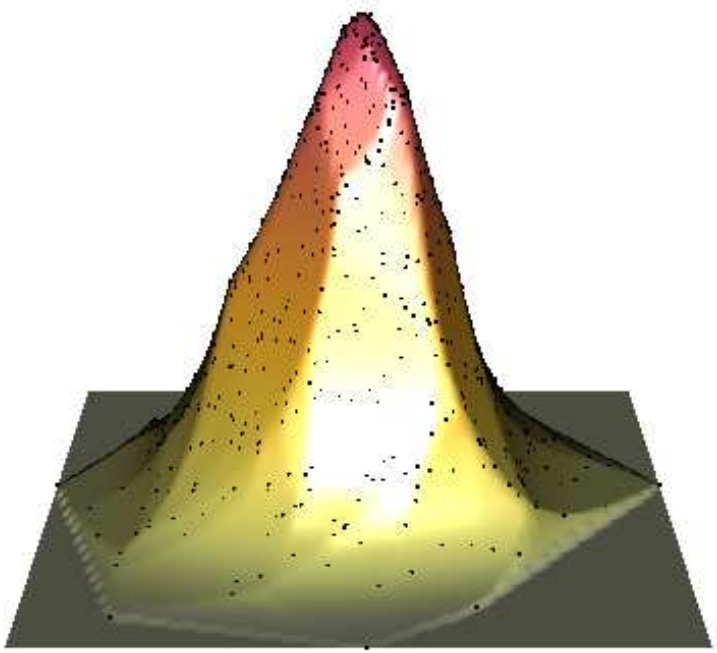
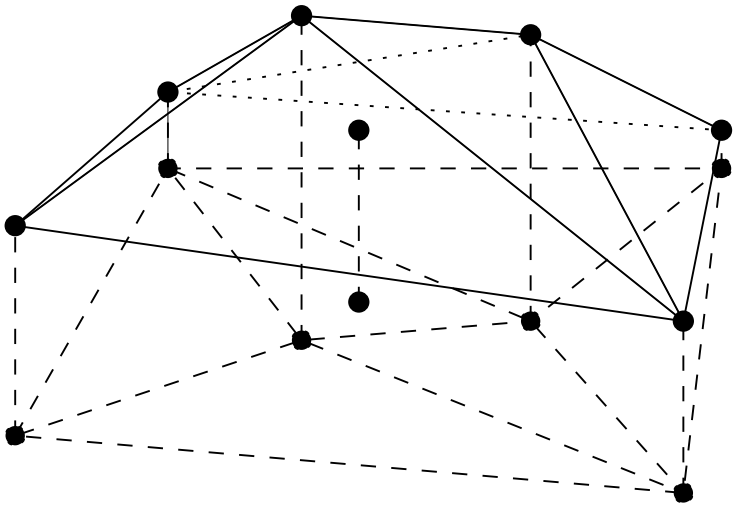
OPTIMISATION CHALLENGES IN MODERN STATISTICS





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

How do optimisation problems arise in Statistics?

Let $X_1, \ldots, X_n$ be independent and identically distributed with $\mathbb{E}(X_1) = \theta$ and $\mathrm{Var}(X_1) < \infty$. Here we can estimate θ using $\hat{\theta} = n^{-1} \sum_{i=1}^n X_i$. But note that

$$\hat{\theta} = \operatorname*{argmin}_{\vartheta \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^{n} (X_i - \vartheta)^2.$$

This perspective is very useful in more complicated problems, where closed-form solutions may not be available.

Incorporating regularisation

Defining a statistical procedure as the solution of an optimisation problem allows us explicitly to incorporate regularisation. Consider the regression model

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

where $a < x_1 < \ldots < x_n < b$**. The natural cubic spline estimator of** g **is**

$$\hat{g}_n = \operatorname*{argmin}_{g \in S_2[a,b]} \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - g(x_i)\}^2 + \lambda \int_a^b g''(x)^2 \, dx,$$

where $S_2[a, b]$ **is the set of real-valued functions on** $[a, b]$ **with absolutely continuous first derivative, and** $\lambda > 0$**.**

Deriving theoretical properties

Consider the potentially high-dimensional linear model

$$\underset{n\times 1}{Y} = \beta_0 \underset{n\times 1}{\mathbf{1}_n} + \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon},$$

where $\epsilon \sim N_n(0, \sigma^2 I)$**. The Lasso estimator (Tibshirani, 1996) of** β **is** $\hat{\beta}_\lambda$**, where** $(\hat{\beta}_0, \hat{\beta}_\lambda)$ **minimises**

$$Q(b_0, b) = \frac{1}{2n}\|Y - b_0\mathbf{1}_n - Xb\|_2^2 + \lambda\|b\|_1.$$

Theory for estimation and prediction accuracy for the Lasso is based on the inequality $Q(\hat{\beta}_0, \hat{\beta}_\lambda) \leq Q(\hat{\beta}_0, \beta)$**.**

Theory for the Lasso

Let $S = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$, let $N = \{1, \ldots, p\} \setminus S$ and let $s = |S|$.

Assume there exists $\phi_0 > 0$ such that for all $b \in \mathbb{R}^p$ with $\|b_N\|_1 \leq 3\|b_S\|_1$, we have

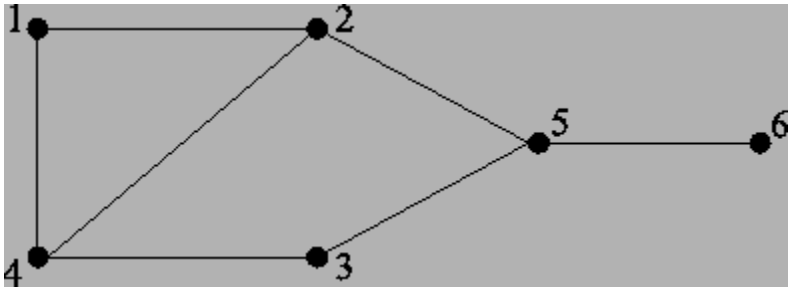
$$\|b_S\|_1^2 \leq \frac{s\|Xb\|_2^2}{n\phi_0^2}.$$

Let $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$ for some $A > 2\sqrt{2}$. Then with probability at least $1 - p^{-(A^2/8-1)}$, we have

$$\frac{1}{n}\|X(\hat{\beta}_\lambda - \beta)\|_2^2 + \lambda\|\hat{\beta}_\lambda - \beta\|_1 \leq \frac{16A^2}{\phi_0^2}\frac{\sigma^2 s \log p}{n}.$$

Graphical models

Graphical models are commonly used to display conditional dependencies in a network.



If we assume $X = (X_1, \ldots, X_6) \sim N_6(0, \Sigma)$ **and let** $\Omega = \Sigma^{-1}$**, then** $\Omega_{ij} = 0$ **iff** X_i **and** X_j **are independent, given** $\{X_k : k \neq i, j\}$**.**

Sparse precision matrix estimation

Graphical models motivate *sparse* precision matrix estimates. The Graphical Lasso (Meinshausen and Bühlmann, 2006; Friedman et al. 2008) minimises the penalised likelihood:

$$Q(\Omega) = -\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda\|\Omega\|_1,$$

over all positive definite matrices Ω, where S is the sample covariance matrix and $\|\cdot\|_1$ denotes the sum of the absolute values of the elements of a matrix.

(Block) coordinate descent algorithms are nowadays very popular.

Non-convex penalties

To avoid bias problems, non-convex penalties are often considered, e.g.

$$Q(b_0, b) = \frac{1}{2n}\|Y - b_0\mathbf{1}_n - Xb\|_2^2 + \sum_{j=1}^{p} p_\lambda(|b_j|),$$

where $p_\lambda : [0, \infty) \to [0, \infty)$ **is concave.**

This means we have to study *local* solutions, often computed by two steps of a local linear approximation algorithm.

Nonparametric density estimation

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

How should we estimate f_0?

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 . . .)**.**

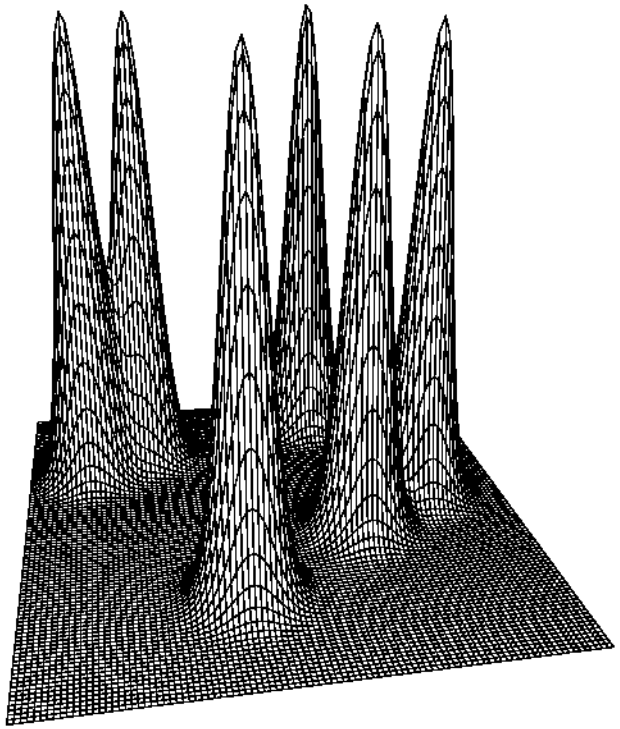
E.g. log-concavity, r-concavity, k-monotonicity, convexity.

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.**

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

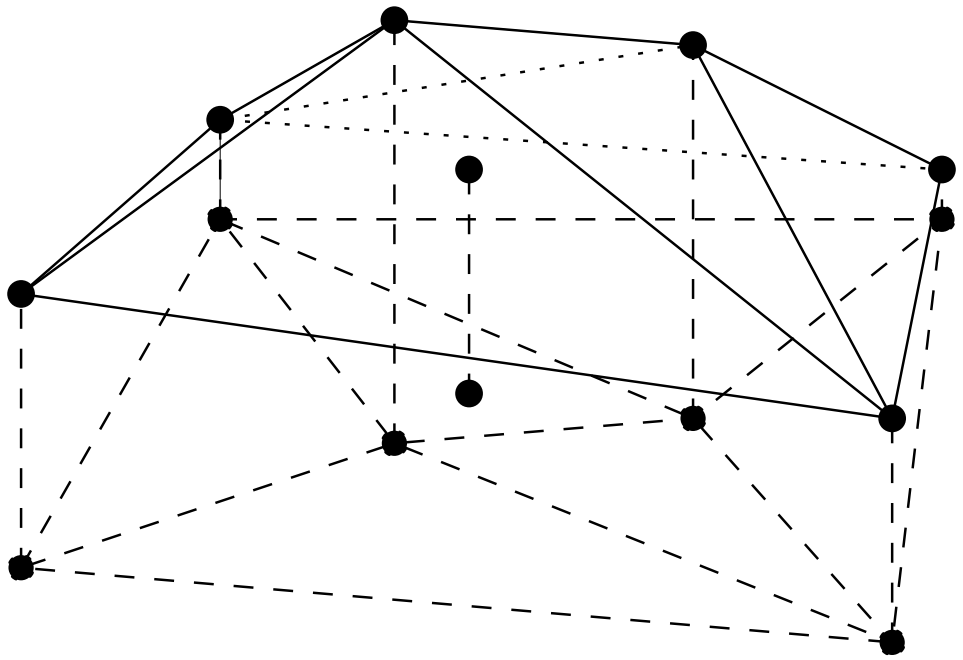
Consider maximising over all log-concave *functions*

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

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* ... but *non-differentiable*!

Smoothed log-concave density estimator

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

Let

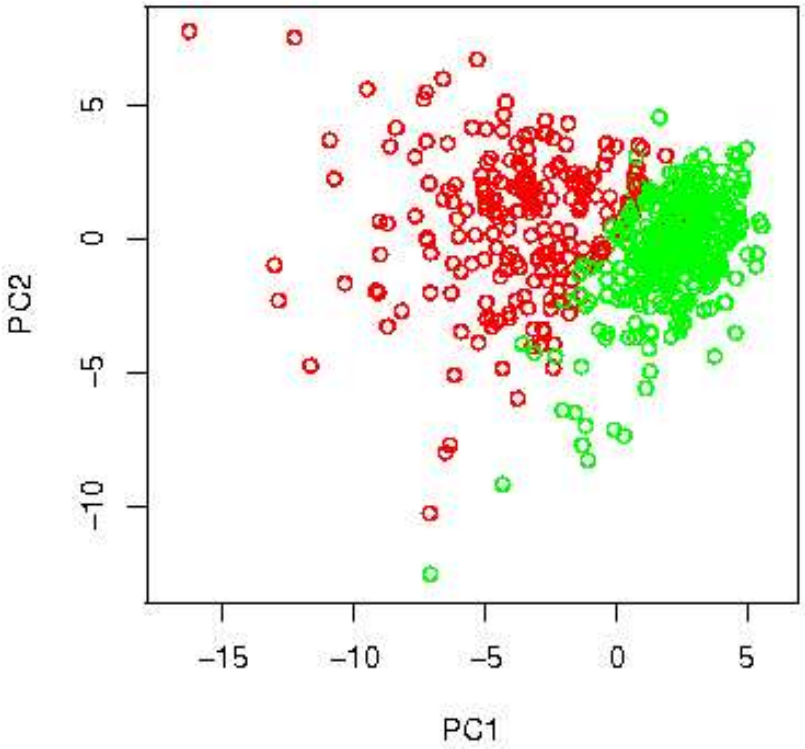
$$\tilde{f}_n = \hat{f}_n * \phi_{\hat{A}},$$

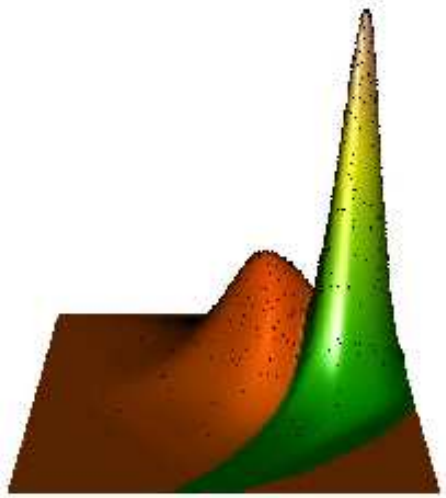
where $\phi_{\hat{A}}$ is a d-dimensional normal density with mean zero and covariance matrix $\hat{A} = \hat{\Sigma} - \tilde{\Sigma}$. Here, $\hat{\Sigma}$ is the sample covariance matrix and $\tilde{\Sigma}$ is the covariance matrix corresponding to $\hat{f}_n$.

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 $\hat{f}_n$.

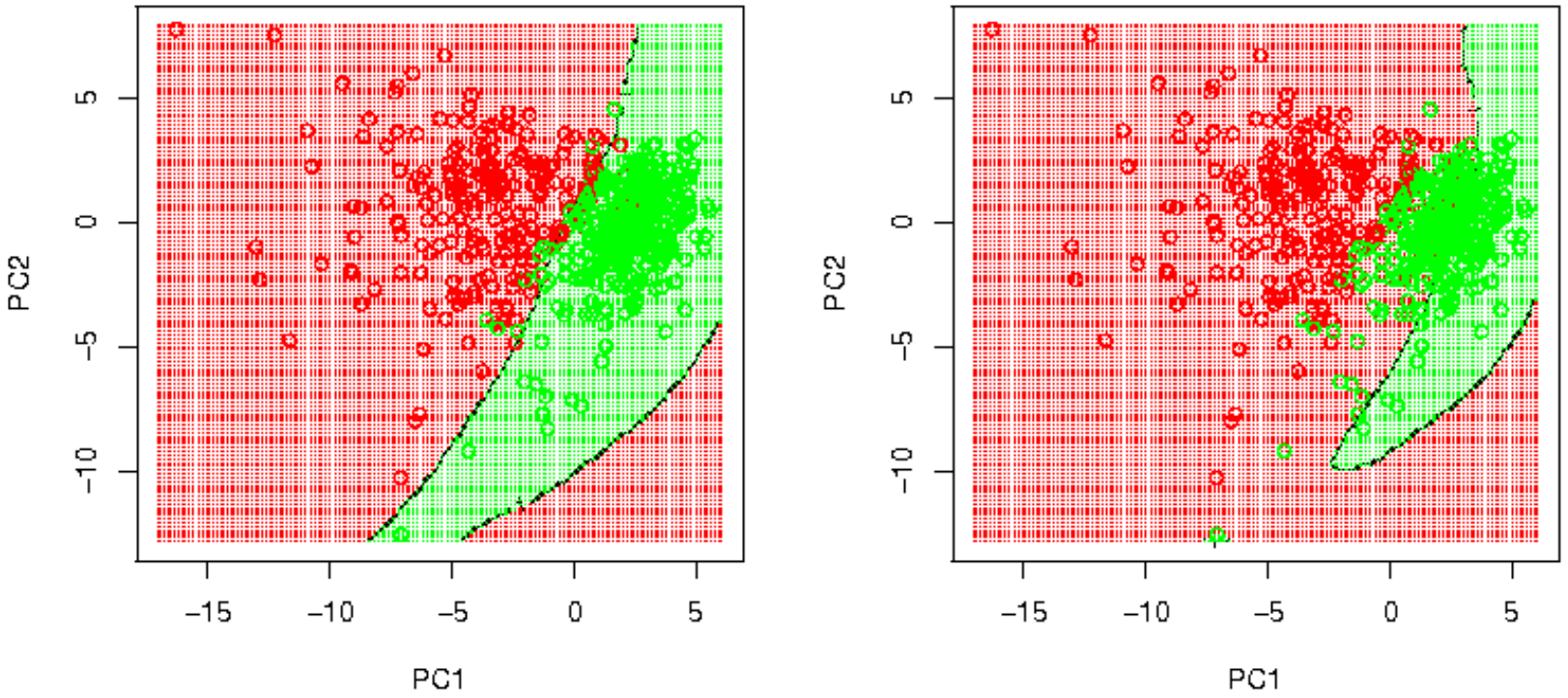
It offers potential improvements for small sample sizes.

Breast cancer data





Classification boundaries



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 $f_1, \ldots, f_d$.

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**.**

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

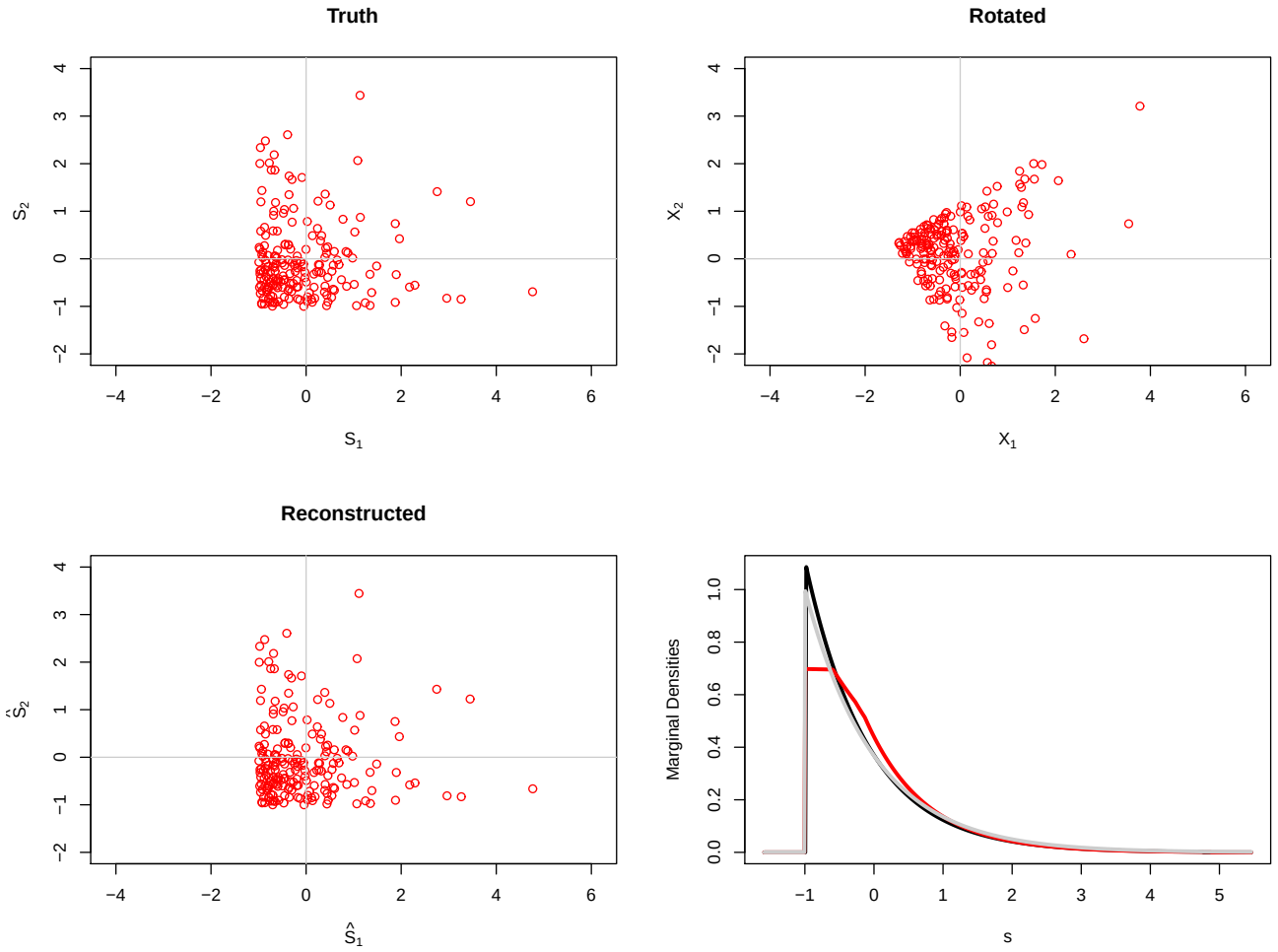
On $[\min(w_j^\mathsf{T}\mathbf{x}_1,\ldots,w_j^\mathsf{T}\mathbf{x}_n), \max(w_j^\mathsf{T}\mathbf{x}_1,\ldots,w_j^\mathsf{T}\mathbf{x}_n)]$**, we have**

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

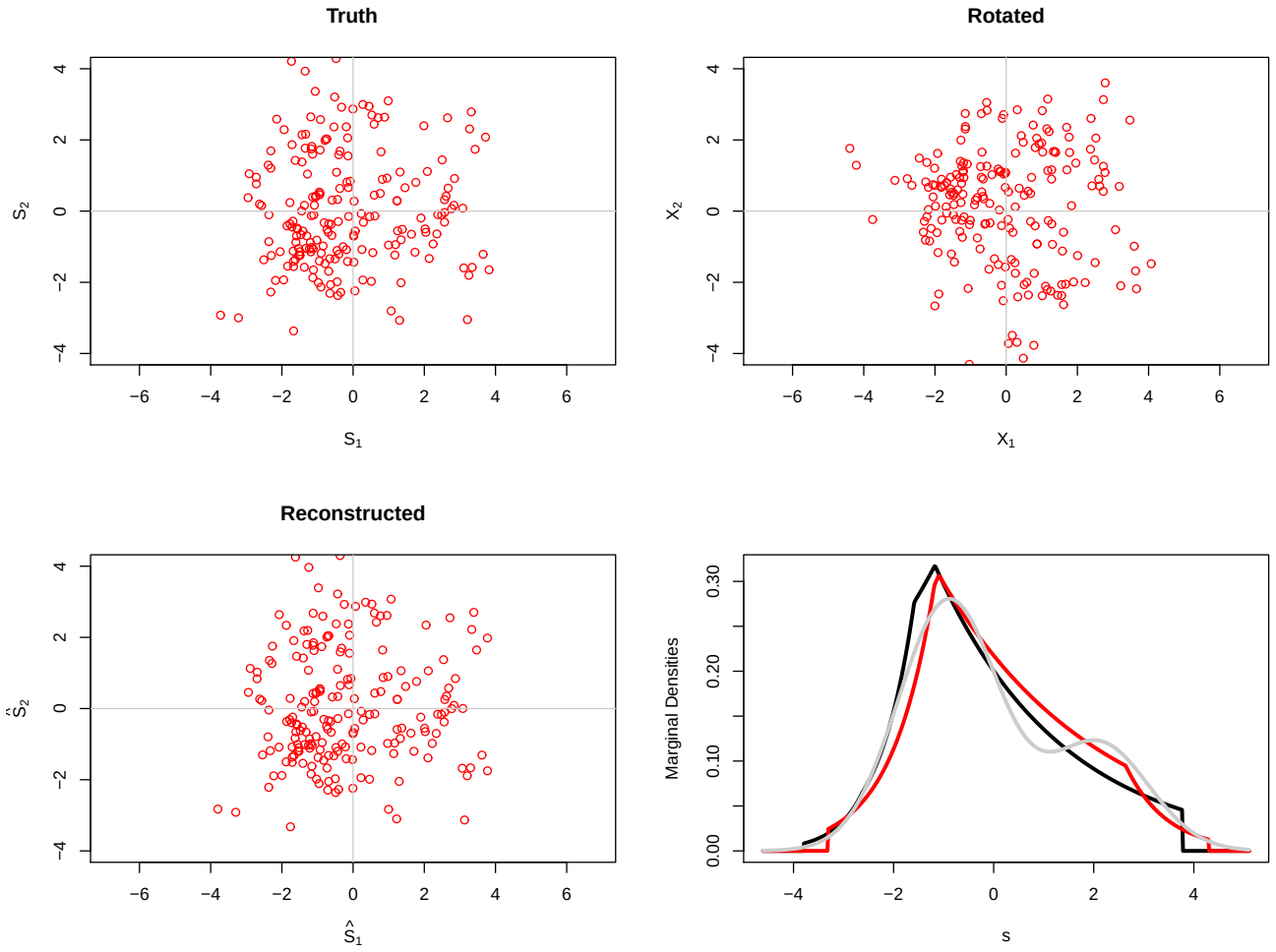
For $1 < s < r < d$**, let** $Y_{r,s}$ **denote the** $d \times d$ **matrix with** $Y_{r,s}(r,s) = 1/\sqrt{2}$**,** $Y_{r,s}(s,r) = -1/\sqrt{2}$ **and zero otherwise. Then** $\mathcal{Y}^+ = \{Y_{r,s} : 1 < s < r < d\}$ **forms an o.n.b. for the skew-symmetric matrices. Let** $\mathcal{Y}^- = \{-Y : Y \in \mathcal{Y}^+\}$**. 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
S2
S1
X2
X1
Ŝ2
Ŝ1
Marginal Densities
s

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



Summary

-
- **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.**
- **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.*, 40, 2973–3002.**
- **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.**