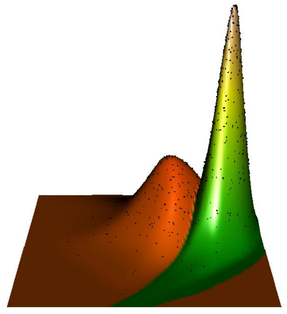


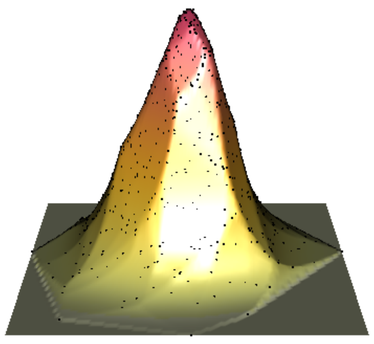
New challenges in high-dimensional statistical inference

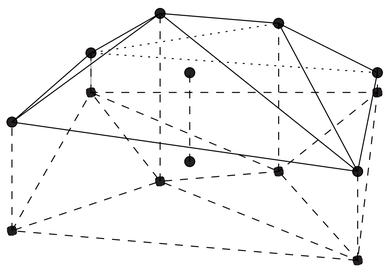
UNIVERSITY OF CAMBRIDGE

Richard J. Samworth (r.samworth@statslab.cam.ac.uk)

Supported by an EPSRC Early Career Fellowship

Shape-constrained inference





The class $\mathcal{F}_d$ of log-concave densities on $\mathbb{R}^d$ is closed under marginalisation, conditioning and convolution, so is very attractive for fully automatic nonparametric density estimation.

New minimax lower bound in squared Hellinger distance:

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{F}_d} \mathbb{E}\{h^2(\tilde{f}_n, f)\} \geq \begin{cases} c_1 n^{-4/5} & \text{if } d = 1 \\ c_d n^{-2/(d+1)} & \text{if } d \geq 2. \end{cases}$$

The problem is therefore *fundamentally harder* than had been anticipated.

The log-concave maximum likelihood estimator $\hat{f}_n$ achieves

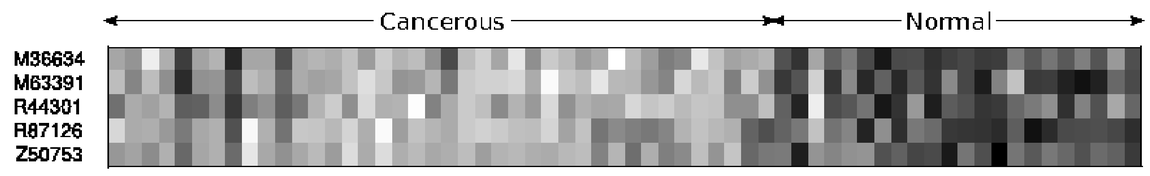
$$\sup_{f \in \mathcal{F}_d} \mathbb{E}\{h^2(\hat{f}_n, f)\} = \begin{cases} O(n^{-4/(d+4)}) & \text{if } d = 1, 2 \\ O(n^{-1/2} \log n) & \text{if } d = 3 \\ O(n^{-1/(d-1)}) & \text{if } d \geq 4. \end{cases}$$

These rates are different from previous conjectures in the literature!

Kim, A. K. H. and Samworth, R. J. (2014) Global rates of convergence in log-concave density estimation. `http://arxiv.org/abs/1404.2298`.

Chen, Y. and Samworth, R. J. (2014) Generalised additive and index models with shape constraints. `http://arxiv.org/abs/1404.2957`.

High-dimensional variable selection



Heat map of genes selected with CPSS.

Complementary Pairs Stability Selection (CPSS) is a new method for improving the performance of any existing variable selection algorithm.

It works by aggregating the results of applying a selection procedure to subsamples of the data.

Let $(A_1, A_2), \ldots, (A_{2B-1}, A_{2B})$ be randomly chosen disjoint pairs of subsets of $\{1, \ldots, n\}$ of size $n/2$.

For $k = 1, \ldots, p$, define $\hat{\Pi}_B(k)$ to be the proportion of subsets on which the base procedure $\hat{S}_{n/2}$ selects variable k.

For some $0 \leq \tau \leq 1$, define the selected variables $\hat{S}^{\mathrm{CPSS}}_{n,\tau}$ to be those variables k for which $\hat{\Pi}_B(k) \geq \tau$.

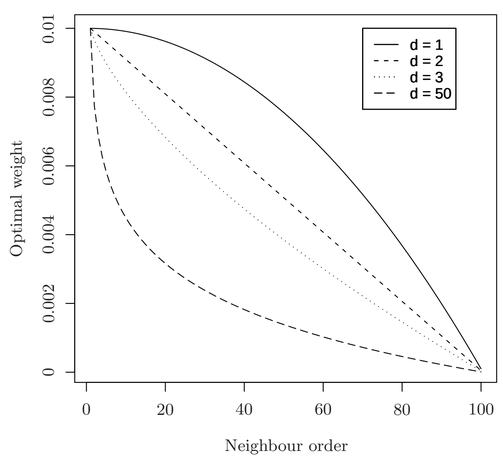
Let $q = \mathbb{E}|\hat{S}_{n/2}|$, $p_{k,n} = \mathbb{P}(k \in \hat{S}_n)$ and $L = \{k : p_{k,n/2} \leq q/p\}$. If $\tau > 1/2$,

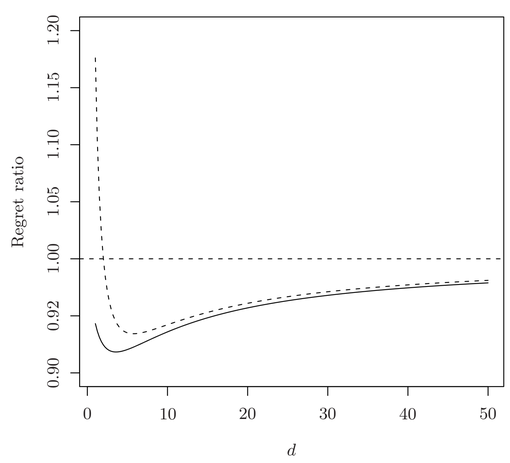
$$\mathbb{E}|\hat{S}^{\mathrm{CPSS}}_{n,\tau} \cap L| \leq \frac{q}{(2\tau - 1)p} \mathbb{E}|\hat{S}_{n/2} \cap L|.$$

No conditions required, and the bound can be further sharpened under unimodality/r-concavity assumptions.

Shah, R. D. and Samworth, R. J. (2013) Variable selection with error control: Another look at Stability Selection, *J. Roy. Statist. Soc., Ser. B*, **75**, 55–80.

Nonparametric classification





Given data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{1, 2\}$ and a new point $x \in \mathbb{R}^d$ to classify, rearrange data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$, where $\|X_{(1)} - x\| \leq \ldots \leq \|X_{(n)} - x\|$.

Weighted nearest neighbour classifier with weights (w_{ni}):

$$\hat{C}^{\mathrm{wnn}}_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}} \geq 1/2 \\ 2 & \text{otherwise.} \end{cases}$$

Optimal weighting scheme: choose $k^* = O(n^{4/(d+4)})$ and

$$w^*_{ni} = \begin{cases} \frac{1}{k^*}\left[1 + \frac{d}{2} - \frac{d}{2(k^*)^{2/d}}\{i^{1+2/d} - (i-1)^{1+2/d}\}\right] & i = 1, \ldots, k^* \\ 0 & i = k^* + 1, \ldots, n. \end{cases}$$

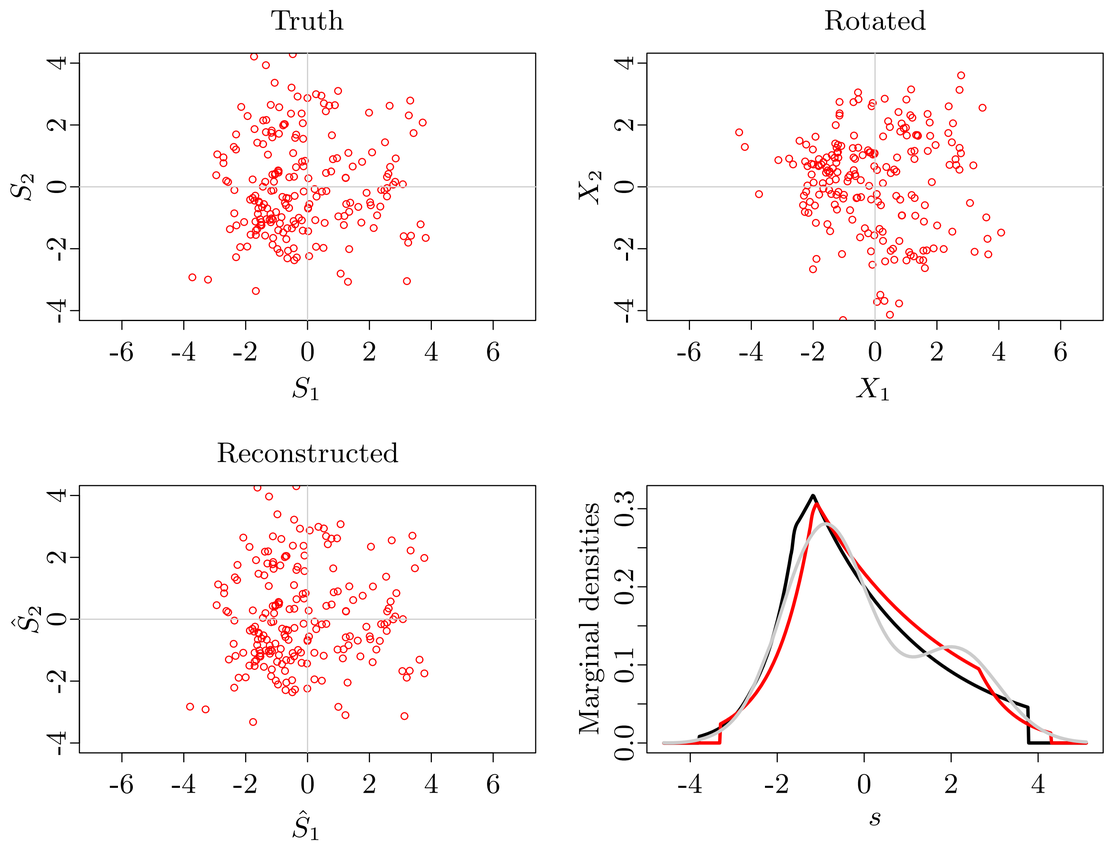
Improvement in risk over unweighted k-nearest neighbours:

$$\frac{R(\hat{C}^{\mathrm{wnn}}_n) - R(C^{\mathrm{Bayes}})}{R(\hat{C}^{\mathrm{knn}}_n) - R(C^{\mathrm{Bayes}})} \to \frac{1}{4^{d/(d+4)}} \left(\frac{2d+4}{d+4}\right)^{(2d+4)/(d+4)}.$$

Asymptotic improvement does not depend on distributions!

Samworth, R. J. (2012) Optimal weighted nearest neighbour classifiers, *Ann. Statist.*, **40**, 2733–2763.

Independent Component Analysis



In ICA, we observe replicates of

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

where A is deterministic and invertible, and S has independent components. We want to estimate the *unmixing matrix* $W = A^{-1}$ and the marginals of S.

New method uses ideas of log-concave projection, and is consistent under very weak conditions — no smoothness conditions or tuning parameters are required!

Samworth, R. J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**, 2973–3002.