Statistics and optimization
Calculation of nonlinear estimators

Sylvain Sardy

Université de Genève

1. Basis Pursuit denoising (Chen, Donoho and Saunders, 1999)

Instead of using an orthonormal set of wavelets (i.e., as many wavelets as observations N) to write

$$\boldsymbol{\mu} = \Phi\boldsymbol{\alpha},$$

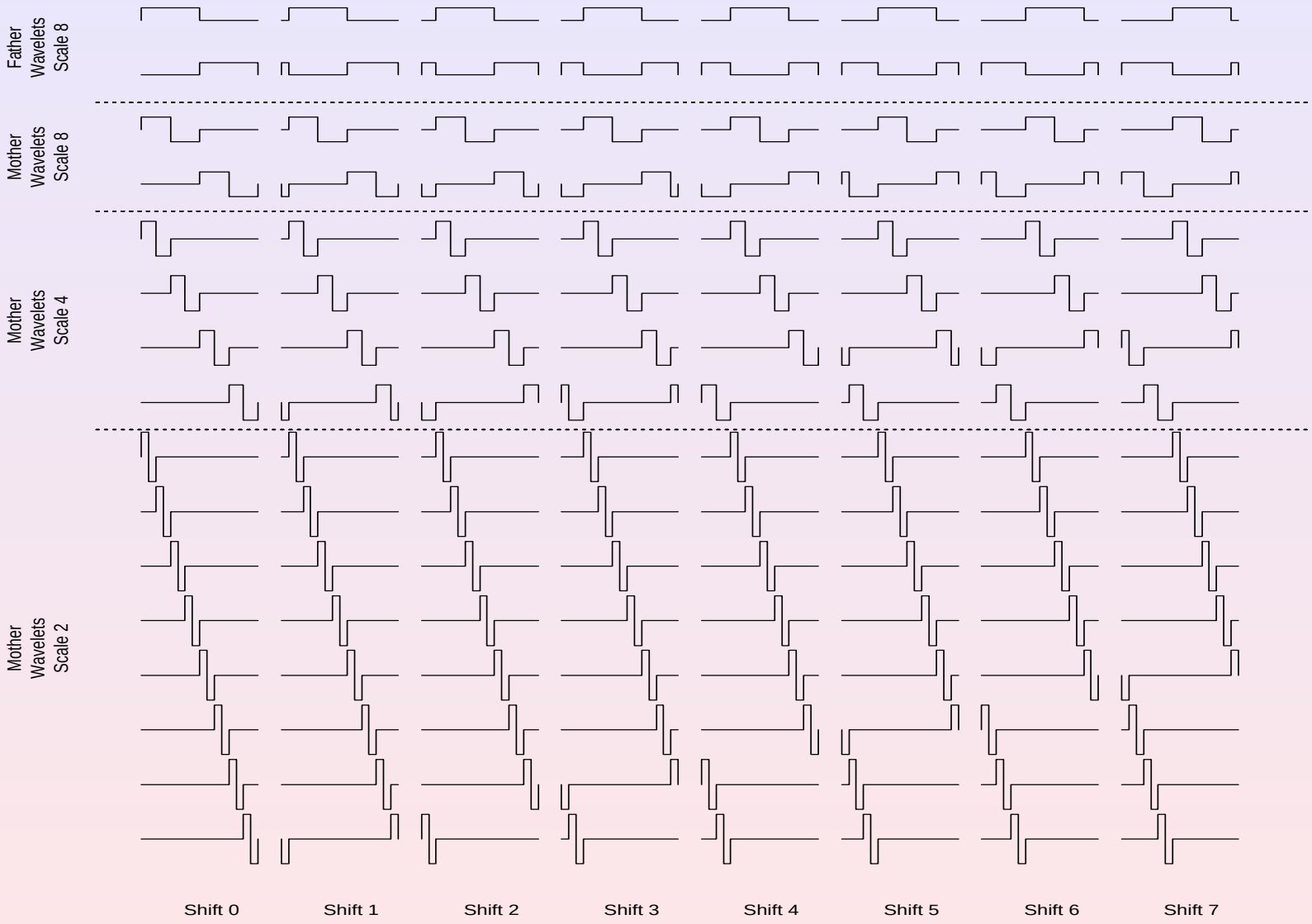
they want to use an *overcomplete* ($P > N$) set of wavelets.

Since Φ is no longer orthonormal, solving

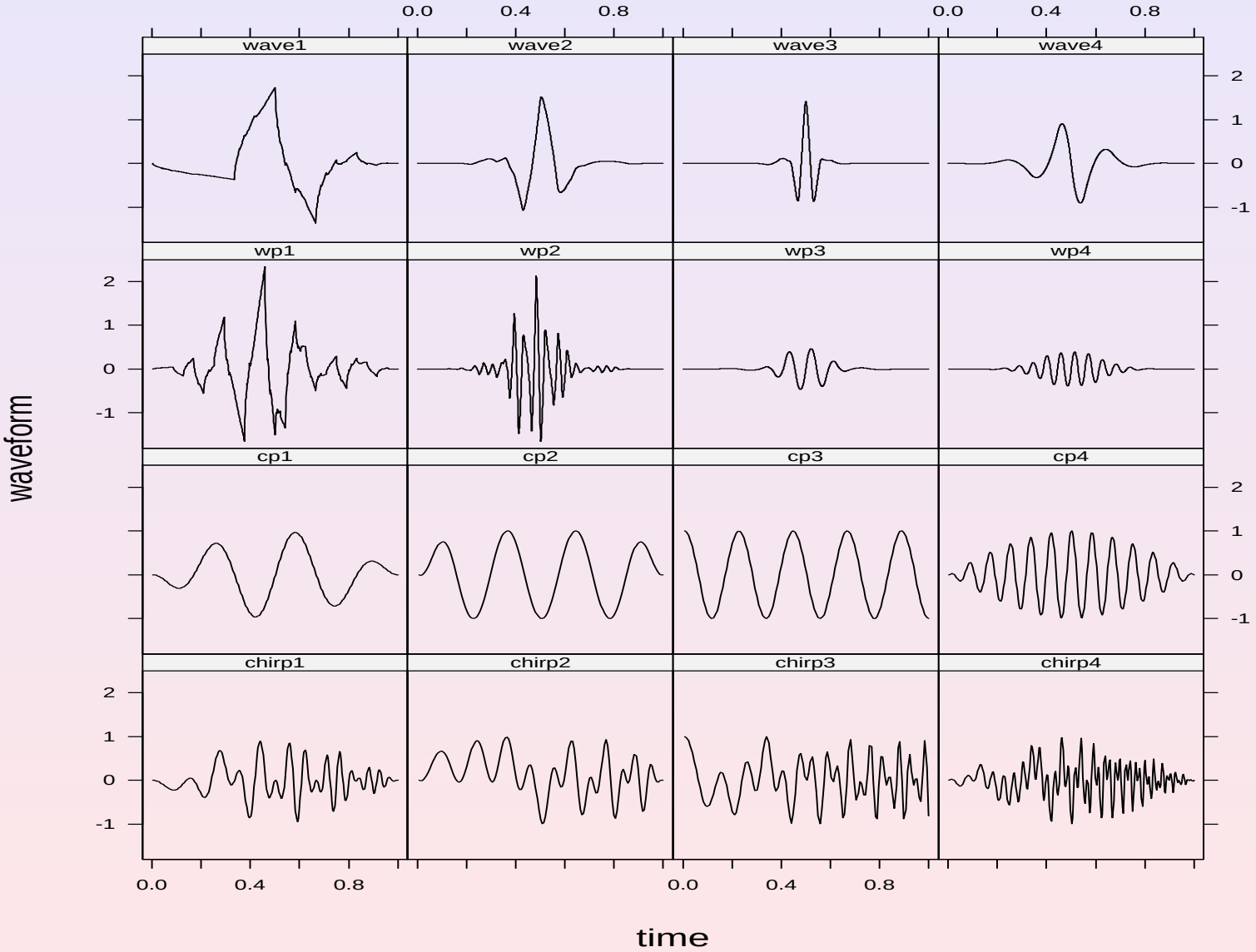
$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \Phi\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

becomes difficult.

Example of overcomplete wavelets: Haar translation invariant.



Example of overcomplete wavelets: Wavelet packets, cosine packets, chirplets, ... (Meyer, Daubechies, ...)



Many overcomplete sets of wavelets are orthonormal union complete matrix:

$$W = \bigcup_j W_j,$$

where W_j orthonormal.

In that case solving Basis pursuit denoising,

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{Y} - \left(\bigcup_j W_j \right) \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{1}$$

with at least two algorithms:

1. Block coordinate relaxation [Sardy, Bruce and Tseng, 2000].
2. Primal-Dual log-barrier interior point algorithm [CDS, 1999].

1. BCR is easy to understand and implement.

- choose a block j
- fixed all entries of $\boldsymbol{\alpha}$ constant except $\boldsymbol{\alpha}_j \in \mathbb{R}^N$
- solve

$$\min_{\boldsymbol{\alpha}_j \in \mathbb{R}^N} \frac{1}{2}\|\mathbf{r}_j - W_j\boldsymbol{\alpha}_j\|_2^2 + \lambda\|\boldsymbol{\alpha}_{-i}\|_1 + \lambda\|\boldsymbol{\alpha}_j\|_1.$$

  with $\mathbf{r}_j = \mathbf{Y} - W_{-j}\boldsymbol{\alpha}_{-j}$
- change j and repeat

A general theorem proves convergence to the global minimum [*Tseng, 2001 "Convergence of block coordinate descent method for nondifferentiable minimization"*]

2. Primal-Dual log-barrier interior point algorithm

CDS make the substitution $\boldsymbol{\alpha} = \mathbf{x}_+ - \mathbf{x}_-$ with $\mathbf{x} = (\mathbf{x}_+, \mathbf{x}_-) \geq 0$; hence

$$\|\boldsymbol{\alpha}\|_1 = \mathbf{1}^{\mathrm{T}}\mathbf{x}.$$

They transform (1) into a quadratic programming problem

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda \mathbf{1}^{\mathrm{T}}\mathbf{x} \quad \text{s.t.} \quad \mathbf{z} = \mathbf{Y} - A\mathbf{x}, \ \mathbf{x} \geq \mathbf{0}$$

with $A = [\Phi, -\Phi]$.

Then, a logarithmic-barrier penalty $-\sum_{p=1}^{2P} \log x_p$ for the nonnegativity constraints on $\mathbf{x}$, multiplied by a barrier parameter $\mu > 0$, is added, yielding the penalized subproblem:

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda \mathbf{1}^{\mathrm{T}}\mathbf{x} - \mu \sum_{p=1}^{2P} \log x_p \quad \text{s.t.} \quad \mathbf{z} = \mathbf{Y} - A\mathbf{x} \tag{2}$$

with implicit inequality constraints.

The KKT conditions are

$$\begin{aligned} \mathbf{0} &= -A^{\mathrm{T}}\mathbf{z} - \mathbf{u} + \lambda \mathbf{1} =: \mathbf{r}_x \\ \mathbf{0} &= \mathbf{Y} - A\mathbf{x} - \mathbf{z} =: \mathbf{r}_y \\ \mathbf{0} &= \mu \mathbf{1} - XU\mathbf{1} =: \mathbf{r}_z, \end{aligned}$$

where $X = \mathrm{diag}(\mathbf{x})$ and $U = \mathrm{diag}(\mathbf{u})$, and $\mathbf{x} > 0$ and $\mathbf{u} > 0$.

CDS solve the corresponding non-linear equations approximately with Newton's method.
Conjugate gradient is employed to solve the corresponding linear system:

$$\begin{aligned} \mathbf{r}_x &= A^{\mathrm{T}}\Delta\mathbf{z} + \Delta\mathbf{u} \\ \mathbf{r}_y &= A\Delta\mathbf{x} + \Delta\mathbf{z} \\ \mathbf{r}_z &= X\Delta\mathbf{u} + U\Delta\mathbf{x}, \end{aligned}$$

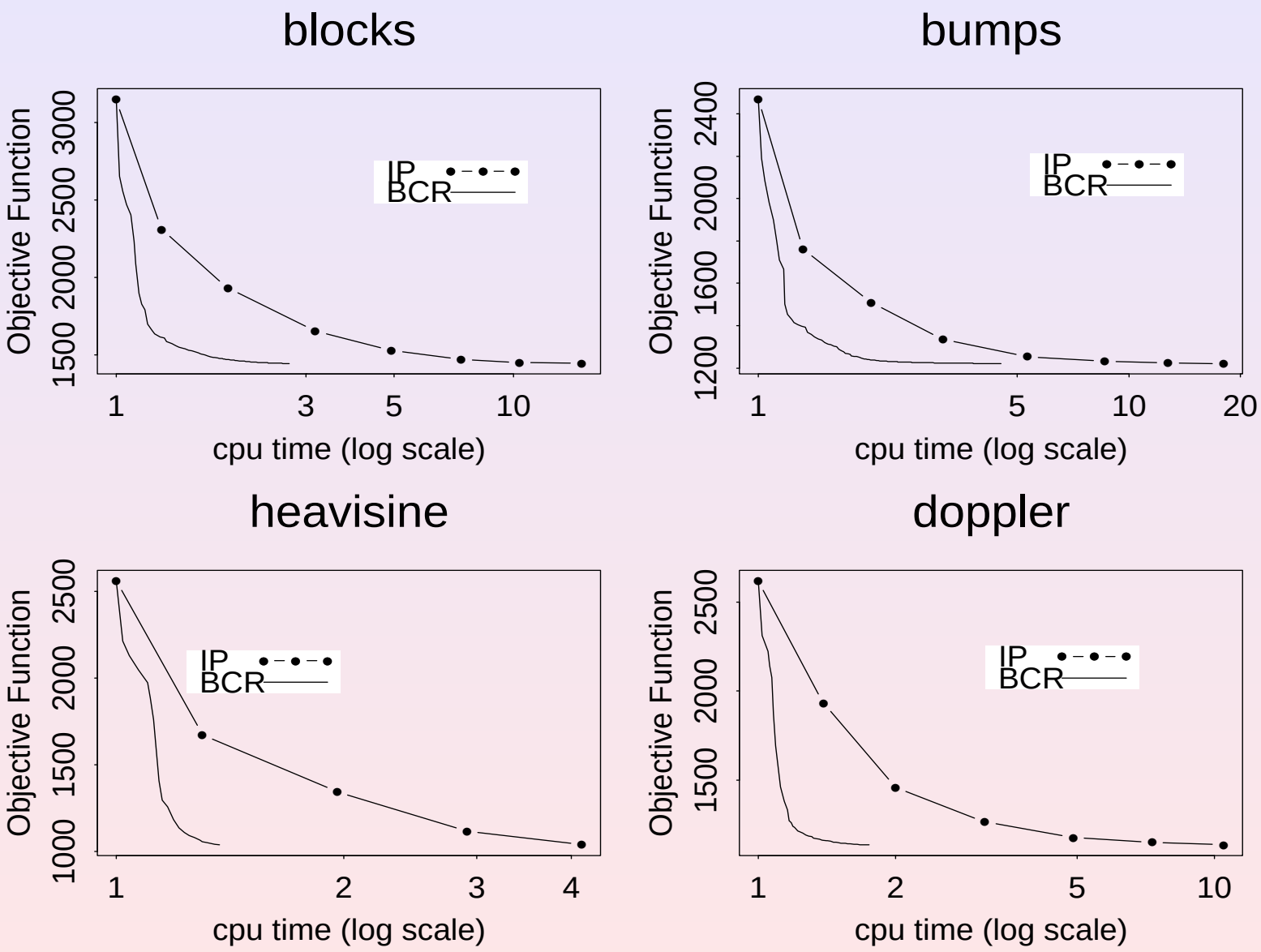
The Newton step in the Newton directions $(\Delta\mathbf{x}, \Delta\mathbf{u}, \Delta\mathbf{z})$ guarantees positivity of $\mathbf{x}$ and $\mathbf{u}$ at each step.

The parameter $\mu \to 0$ is then updated and (2) approximately solved repeatedly.

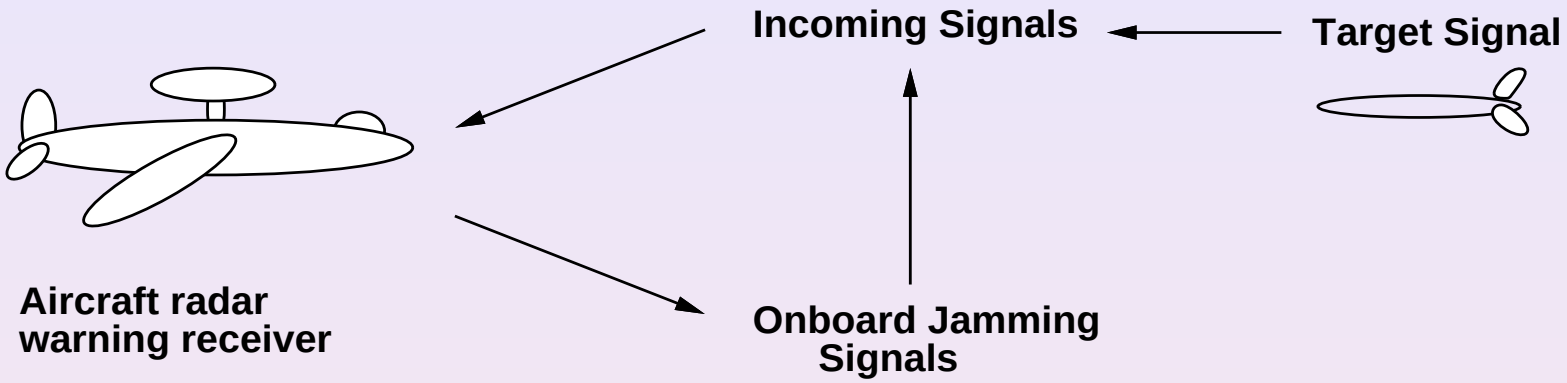
Implementing convex program in `R` with CVX.

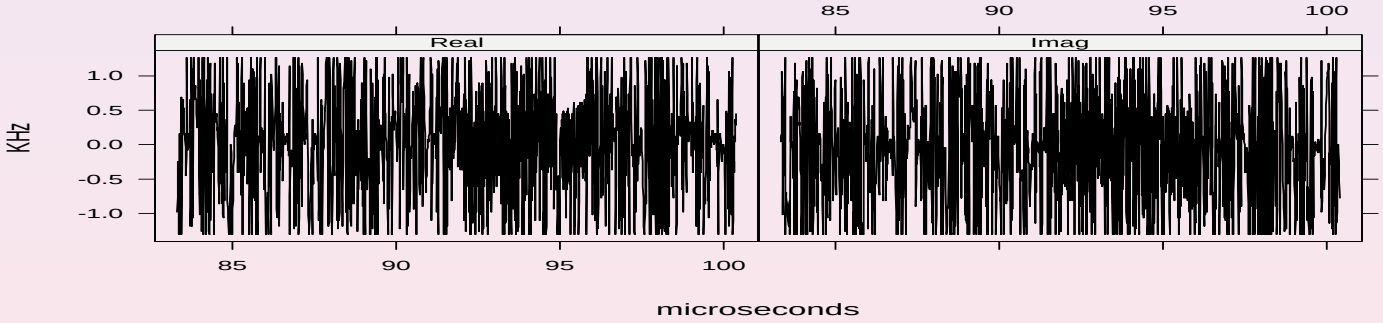
library(CVXfromR) is using interior point methods.

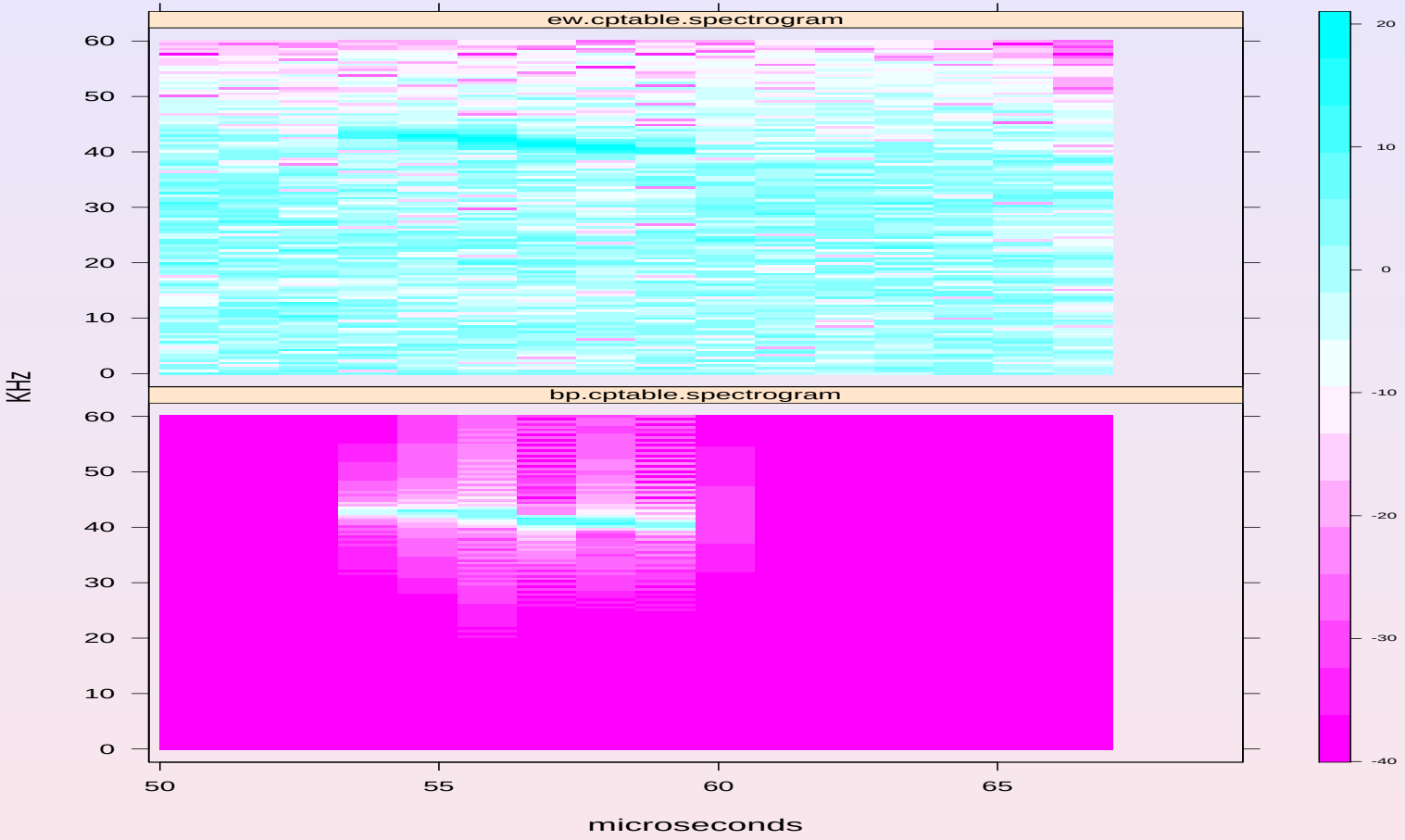
Comparison between CDS's and SBT's algorithms:



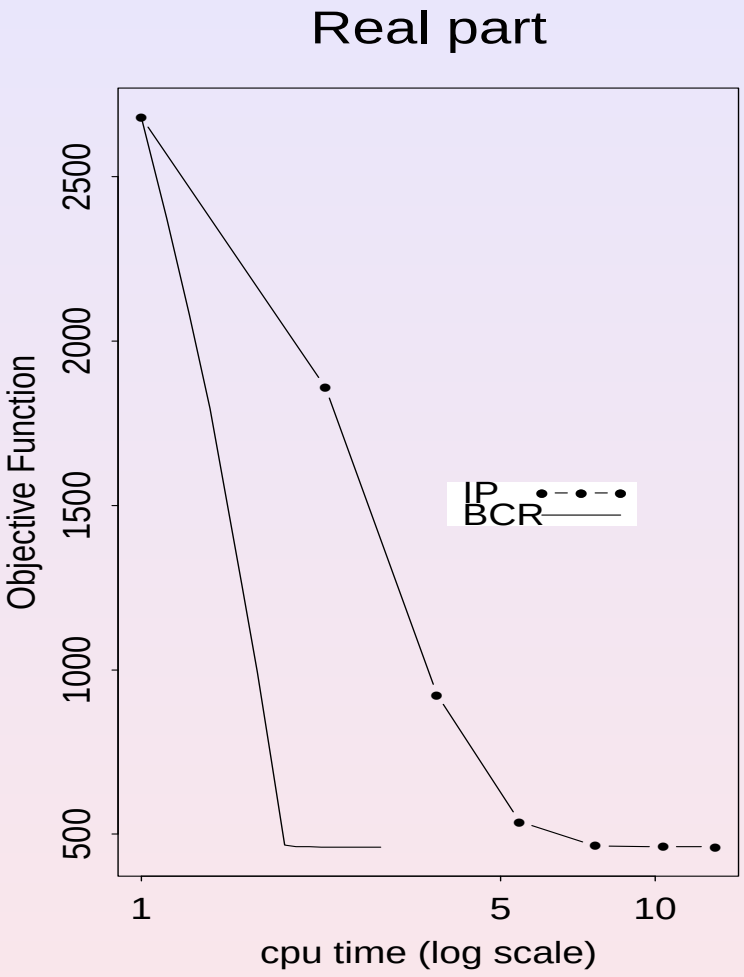
Application

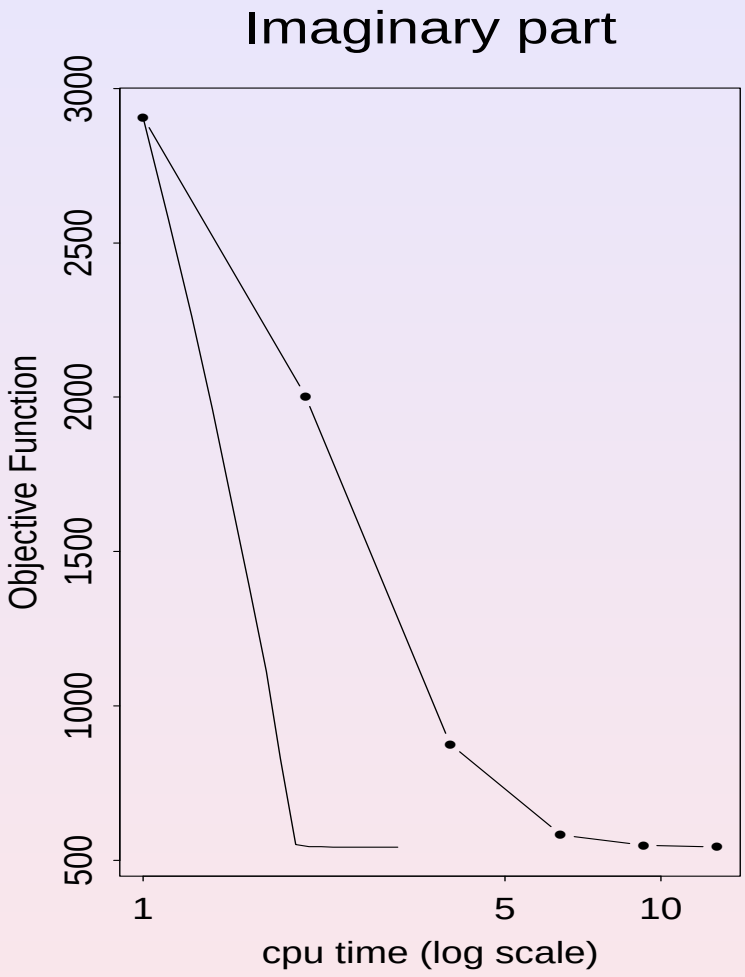






Raw and denoised spectrogram


Real part
Objective Function
500
1000
1500
2000
2500
1
5
10
cpu time (log scale)
IP
BCR


Imaginary part
Objective Function
500
1000
1500
2000
2500
3000
1
5
10
cpu time (log scale)

2. Subbotin waveshrink (Sardy, 2009)

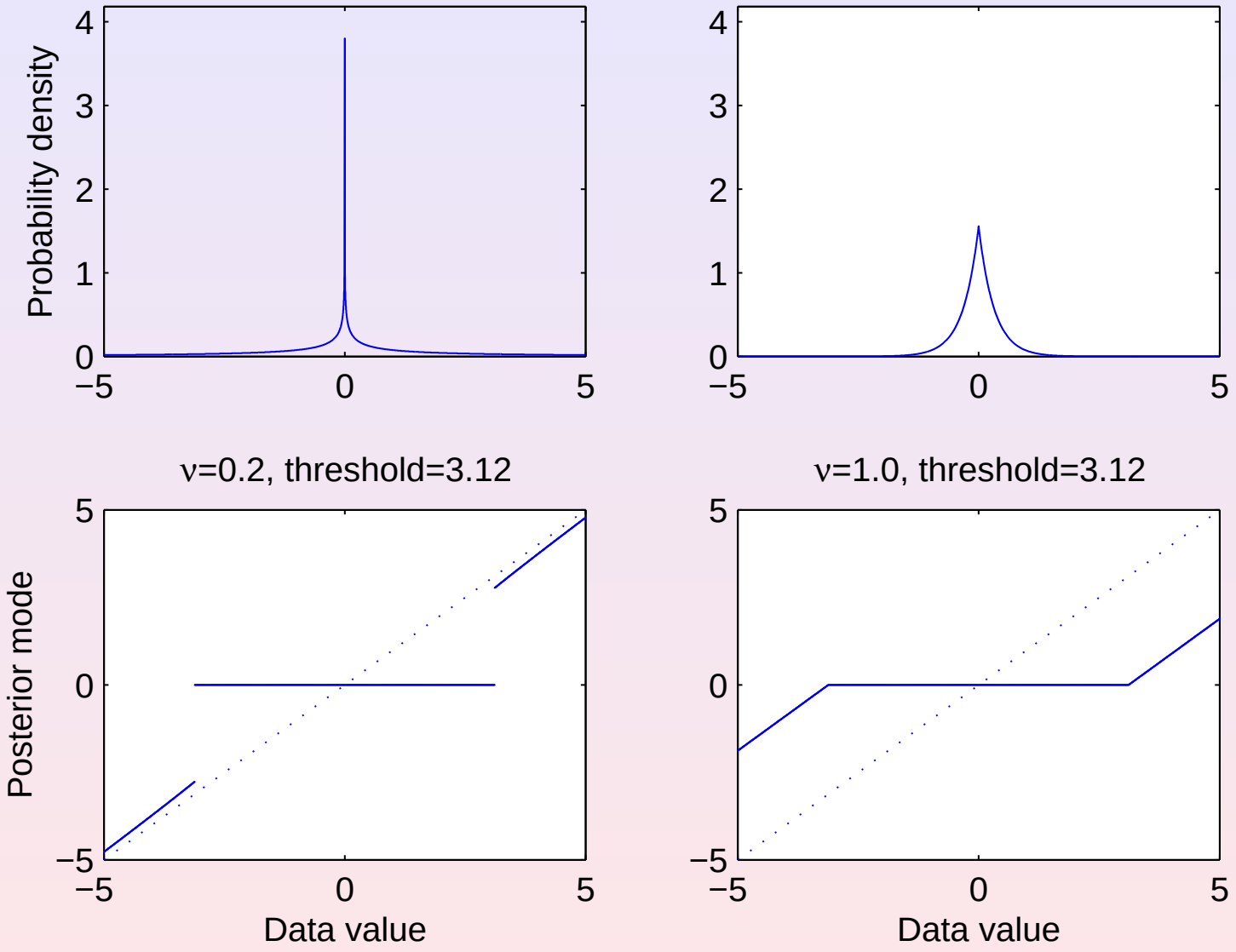
Choosing W orthonormal and $\nu < 1$,

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - W\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu,$$

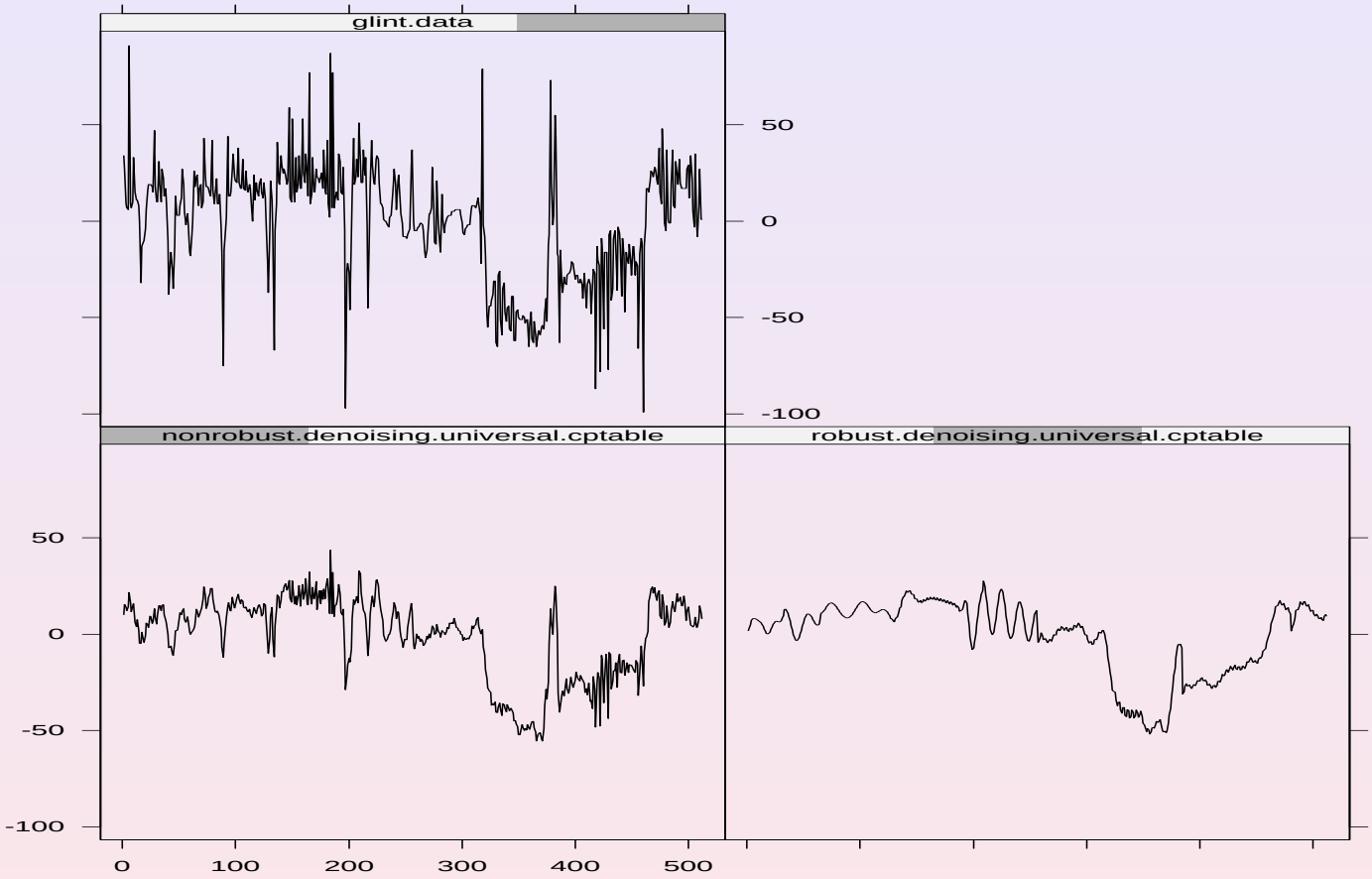
is not a convex problem, but can be solved exactly.

One can show the corresponding thresholding function has a closed form expression for threshold φ and jump κ:

$$\begin{aligned} \varphi(\lambda;\nu) &= (2-\nu)[\lambda\{2(1-\nu)\}^{\nu-1}]^{1/(2-\nu)} \\ \kappa(\lambda;\nu) &= \varphi(\lambda;\nu)\frac{2(1-\nu)}{(2-\nu)}. \end{aligned}$$


Probability density
4
3
2
1
0
−5
0
5
ν=0.2, threshold=3.12
ν=1.0, threshold=3.12
Posterior mode
5
0
−5
Data value

3. Robust waveshrink (Sardy, Bruce and Tseng, 2001)



Removing the glint noise without (left) and with (right) a robust estimator

Peter Huber is one of the main contributors to robust statistics.

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - W\boldsymbol{\alpha}\|_{\text{Huber}} + \lambda\|\boldsymbol{\alpha}\|_1, \tag{3}$$

where $\|\mathbf{r}\|_{\text{Huber}} = \sum_{p=1}^{P} \rho(r_p)$ with:

$$\rho(r) = \begin{cases} r^2/2 & |r| \leq \tau \\ \tau|r| - \tau^2/2 & |r| > \tau \end{cases}.$$

Interestingly, one can show that

$$\rho(r) = \min_{\beta} \frac{1}{2}(r - \beta)^2 + \tau|\beta|$$

leading to a BCR algorithm since (3) amounts to

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - W\boldsymbol{\alpha} - I\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 + \tau\|\boldsymbol{\beta}\|_1.$$

4. Lasso (Tibshirani, 1996)

For a given regularization parameter $\lambda \geq 0$ (chosen by CV or SURE among other methods), solve

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1.$$

The first order necessary (and here sufficient) conditions also known as Karush–Kuhn–Tucker (KKT) conditions are

$$\mathbf{x}_p^{\mathrm{T}}(\mathbf{Y} - X\hat{\boldsymbol{\alpha}}) = \lambda\gamma_p, \quad p = 1, \ldots, P,$$

with

$$\gamma_p \in \begin{cases} \{+1\} & \text{if} \quad \hat{\alpha}_p > 0 \\ \{-1\} & \text{if} \quad \hat{\alpha}_p < 0 \\ [-1, 1] & \text{if} \quad \hat{\alpha}_p = 0 \end{cases}.$$

which amounts to solving $\mathbf{0} \in \partial f(\boldsymbol{\alpha})$ since there are no constraints.

Property: $\boldsymbol{\alpha}_\lambda = \mathbf{0}$ is the lasso solution iff $\lambda \geq \|X^{\mathrm{T}}\mathbf{Y}\|_\infty$ since the KKT conditions tell us that it is necessary and sufficient to have $\mathbf{x}_p^{\mathrm{T}}\mathbf{Y} \in \lambda[-1, 1]$ for all p.

Ways to solve lasso:

- Relaxation: successively choose i and solve

$$\min_{\alpha_i} \frac{1}{2}\|\mathbf{Y} - X_{-i}\boldsymbol{\alpha}_{-i} - \mathbf{x}_i\alpha_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_{-i}\|_1 + \lambda|\alpha_i|$$

  Theorem: Convergence to the global minimum [*Tseng, 2001 "Convergence of block coordinate descent method for nondifferentiable minimization"*]

- LARS: [*Efron, Hastie, Johnstone and Tibshirani, 2004*]

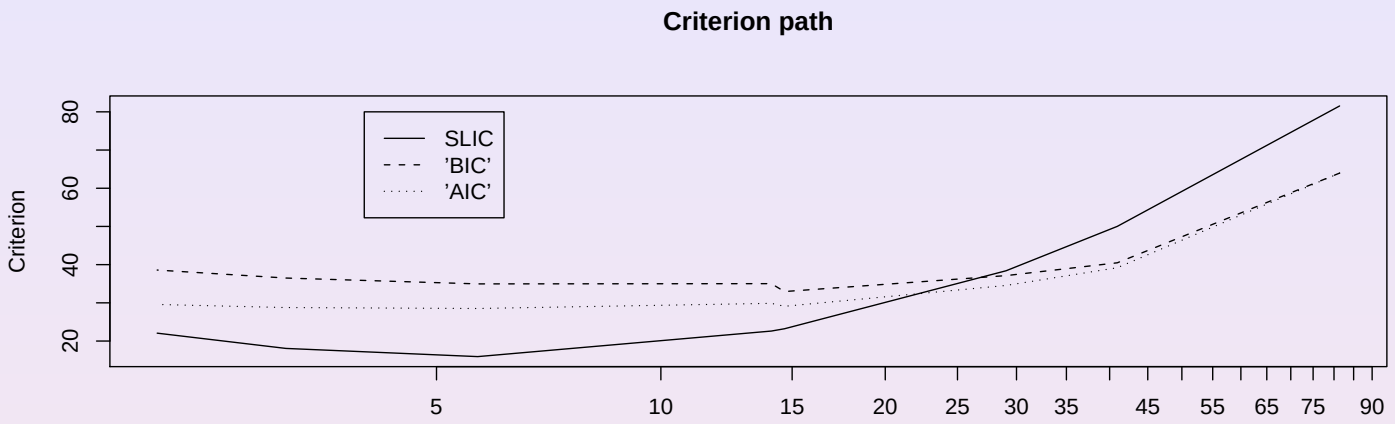
Possibilities to accelerate convergence:

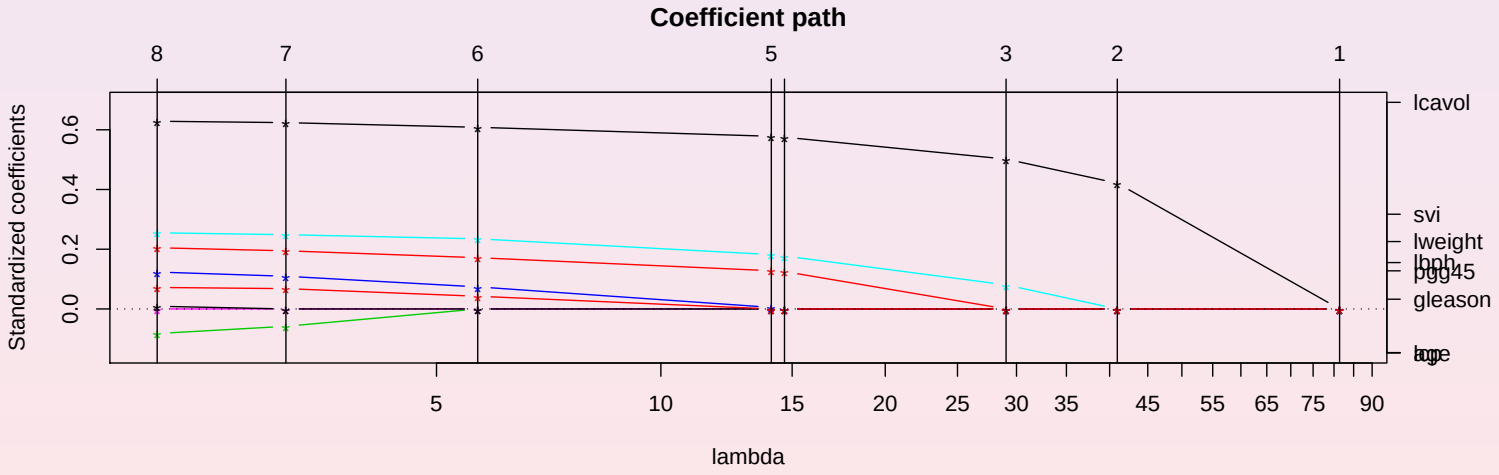
- Instead of cycling for $i = 1, 2, \ldots, P, 1, 2, \ldots$ in a systematic way, we could choose i in an adaptive way.
  The *optimal descent* strategy chooses i as the index of the maximum entry of the subdifferential

$$\left| \partial \{ \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \} \right| .$$

  [*Sardy, Bruce, Tseng, 2000*]
- Warm start!

Typical lasso path:





4. Markov random field/total variation smoothing (Besag 1980's, 90's; Rudin, Osher, Fatemi, 1992)

We measure $(Y_n, x_n)_{n=1,\dots,N}$ according to the model

$$Y_n = f(x_n) + \epsilon_n, \quad \epsilon_n \overset{\text{i.i.d.}}{\sim} (0,1),$$

where f is a univariate function we want to estimate.
Assuming f is a realization of a Markov random field (Bayesian prior), e.g.,

$$\pi(f_n \mid \mathbf{f}) = \pi(f_n \mid \mathbf{f}_{\partial_n}) \propto \exp(-\lambda \sum_{j \in \partial_n} \phi(f_n - f_j)),$$

Hammersley-Clifford's Theorem give the joint prior distribution $\pi_\lambda(\mathbf{f})$.

Bayes Theorem then gives the posterior distribution.
The negative posterior log-likelihood can be minimized to define the maximum a posteriori estimate (MAP)

$$\min_{\mathbf{f}} \|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda \sum_{n=1}^{N} \sum_{j \in \partial_n} \phi(f_n - f_j) \tag{4}$$

Besag (1986) “On the statistical analysis of dirty pictures" *Journal of the Royal Statistical Society* proposed to solve the optimization problem with ICM (Iterative Conditional Mode).

ICM is a coordinate descent algorithm. It converges for differentiable ϕ, e.g., $\phi(\delta) = \delta^2$.

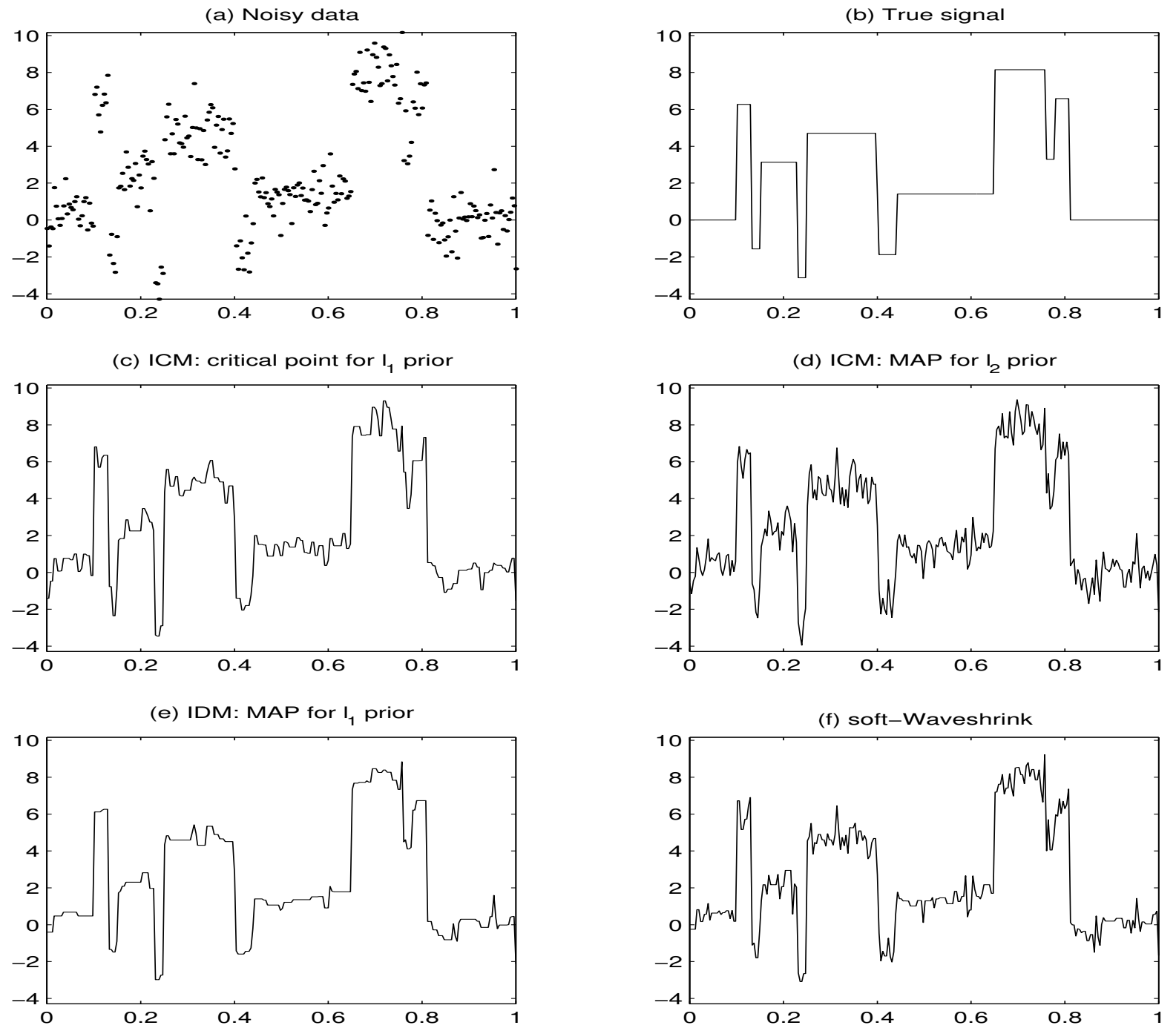
Instead of ℓ_2, we can try ℓ_1 loss with $\phi(\delta) = |\delta|$, leading to total variation smoothing

$$\min_{\mathbf{f} \in \mathbb{R}^N} \frac{1}{2} \|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda \sum_{n=1}^{N-1} |f_{n+1} - f_n|.$$

To calculate the MAP, we considered ICM, and it failed although the paper claims:

"The cost function never increases at each stage and eventual convergence is assured."

This is due to the fact that the non-differentiable penalty is not separable.


(a) Noisy data
(b) True signal
(c) ICM: critical point for l_1 prior
(d) ICM: MAP for l_2 prior
(e) IDM: MAP for l_1 prior
(f) soft−Waveshrink
10
8
6
4
2
0
−2
−4
0
0.2
0.4
0.6
0.8
1

An algorithm that works it to transform

$$\min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda\|B\mathbf{f}\|_1,$$

where $B\mathbf{f} = \sum_{n=1}^{N-1} |f_{n+1} - f_n|$, into

$$\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}_{-1}\|_1,$$

where X is a lower triangular matrix of ones and $\boldsymbol{\alpha}_{-1}$ is the vector $\boldsymbol{\alpha}$ but the first entry (corresponding to the intercept).

We are back to a lasso problem.

Alternative algorithm: Iterative dual mode algorithm (IDM, Sardy, Tseng, 2004).

Consider the equivalent convex problem with linear constraints

$$\min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda\|\mathbf{b}\|_1 \quad \text{s.t.} \quad \mathbf{b} - B\mathbf{f} = \mathbf{0}.$$

and its Lagrange dual function

$$g(\mathbf{v}) = \min_{\mathbf{f},\mathbf{b}} \frac{1}{2}\|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda\|\mathbf{b}\|_1 + \mathbf{v}^{\mathrm{T}}(\mathbf{b} - B\mathbf{f}).$$

By convexity of the cost function, Slater's conditions holds, so one can instead maximize the dual problem

$$\max_{\mathbf{v}} g(\mathbf{v})$$

to find the TV solution (no duality gap).

$$
\begin{aligned}
g^* &= \max_{\mathbf{v}\in\mathbb{R}^{N-1}} \min_{\mathbf{f},\mathbf{b}} \frac{1}{2}\|\mathbf{Y}-\mathbf{f}\|_2^2 + \lambda\|\mathbf{b}\|_1 + \mathbf{v}^{\mathrm{T}}(\mathbf{b}-B\mathbf{f}) \\
&= \max_{\mathbf{v}\in\mathbb{R}^{N-1}} \min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y}-\mathbf{f}\|_2^2 - \mathbf{v}^{\mathrm{T}}B\mathbf{f} + \min_{\mathbf{b}} \lambda\|\mathbf{b}\|_1 + \mathbf{v}^{\mathrm{T}}\mathbf{b} \\
&= \max_{\mathbf{v}\in\mathbb{R}^{N-1}} -\frac{1}{2}\|B^{\mathrm{T}}\mathbf{v}\|_2^2 - \mathbf{Y}^{\mathrm{T}}B^{\mathrm{T}}\mathbf{v} + \begin{cases} 0 & \text{if } \|v\|_\infty \le \lambda \\ -\infty & \text{otherwise} \end{cases} \\
&= \max_{\mathbf{v}\in\mathbb{R}^{N-1}:\ \|v\|_\infty\le\lambda} -\frac{1}{2}\|B^{\mathrm{T}}\mathbf{v}\|_2^2 - \mathbf{Y}^{\mathrm{T}}B^{\mathrm{T}}\mathbf{v}
\end{aligned}
$$

The constraints are box-constraints on $\mathbf{v}$ and therefore separable.

Consequently a relaxation algorithm solves the dual in $\mathbf{v}$.

And

$$\mathbf{f}^* = \mathbf{Y} + B^{\mathrm{T}}\mathbf{v}^* =: \hat{\mathbf{f}}_\lambda,$$

the TV estimate for the given smoothing parameter λ.

5. Total variation density estimation (Sardy, Tseng 2010)

We have a sample $x_1, \ldots, x_N$ from an unknown density f.

Nonparametric MLE: letting $l(f; \mathbf{x}) = \sum_{n=1}^{N} \log f(x_n)$, estimate f as solution to

$$\max_f l(f; \mathbf{x})$$

over the space of density functions. The solution

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(x).$$

is degenerate (not in the space of bounded variation: its total variation in unbounded).

Good and Gaskin (1971) defined the functional estimate $\hat{f}_\lambda$ as the solution to

$$\max_{f \in \mathcal{F}} \; l(f; \mathbf{x}) - \lambda \Phi(f) \quad \text{s.t.} \quad \int f(x) dx = 1, \tag{5}$$

where $\mathcal{F}$ is a class of nonnegative-valued functions on Ω.

Various penalty functionals have been proposed including

$$\Phi(f) = \begin{cases} \int f(x)'^2 / f(x) dx \\ 4 \int \{\nabla \sqrt{f}\}^2 \quad \text{[GG 1971]} \\ \int \{\nabla^3 \log f\}^2 \quad \text{[Silverman 82]} \end{cases} .$$

We propose

$$\Phi_{\text{TV}}(f) = \sup \sum_j |f(u_{j+1}) - f(u_j)|$$

where the supremum is taken over all possible partitions $[u_j, u_{j+1}]$, $j = 1, \ldots, M$, of Ω.

Consider two spaces of functions:

- the space of zeroth-order splines on the **u**-partition,

$$\mathcal{F}_u^0 = \{f : [x_1, x_N] \to \Re \mid f \text{ is constant on } (u_{i-1}, u_i), i = 1, \ldots, N\}.$$

- the space of first-order splines on the **x**-partition,

$$\mathcal{F}_x^1 = \left\{f \in \mathcal{C}^0[x_1, x_N] \mid f \text{ is linear on } [x_{i-1}, x_i], i = 2, \ldots, N\right\}.$$

Property: TV of a monotone univariate function f on a closed interval $[\alpha, \beta]$ is $|f(\beta) - f(\alpha)|$.

The optimization problem on $\mathcal{F}$ becomes a finite-dimensional optimization of the form:

$$\min_{\mathbf{f} \in \mathbb{R}^N} -\sum_{i=1}^{N} \log f_i + \lambda \|B\mathbf{f}\|_1 \quad \text{s.t.} \quad \mathbf{a}^T \mathbf{f} = 1,$$

This optimization problem has a unique solution.

6. Fused lasso (Tibshirani, Saunders, Rosset, Zhu, Knight, 2005)

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^P} \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda_1 \sum_{n=1}^{N-1} |\alpha_{n+1} - \alpha_n| + \lambda_2 \|\boldsymbol{\alpha}\|_1.$$

Possible approaches:

1. TSRZK rewrite is as a quadratic program.
2. FISTA: fast iterative shrinkage-thresholding algorithms (Beck and Teboulle 2009). For instance to solve

$$\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) + g(\boldsymbol{\alpha})$$

with $f(\boldsymbol{\alpha}) = \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2$ and $g(\boldsymbol{\alpha}) = \lambda\|B\boldsymbol{\alpha}\|_1$.
Since $\boldsymbol{\alpha}_{k+1} = \boldsymbol{\alpha}_k - t_k \nabla f(\boldsymbol{\alpha}_k)$ is solution to

$$\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}_k) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_k)^{\mathrm{T}} \nabla f(\boldsymbol{\alpha}_k) + \frac{1}{2t_k}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_k\|_2^2,$$

they solve, for $t_k \in (0, 1/\|X^{\mathrm{T}}X\|)$,

$$\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}_k) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_k)^{\mathrm{T}} \nabla f(\boldsymbol{\alpha}_k) + \frac{1}{2t_k}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_k\|_2^2 + \lambda\|B\boldsymbol{\alpha}\|_1.$$

Take home message

Advices:

- Study EXISTENCE of a solution
- Study uniqueness
- Study the mathematical properties: dimension, constraints, convexity, differentiability, etc.
- Do not use an algorithm as a black box, but read its description, arguments
- TEST your algorithm

Other statistical issues:

- Selection of λ
- Study convex analysis: useful also for statistical properties.