Stein COnsistent Risk Estimator (SCORE) for hard thresholding

Charles Deledalle[1], Gabriel Peyré[2] and Jalal Fadili[3]

[1]IMB, CNRS-Université Bordeaux 1, [2]CEREMADE, CNRS-Université Paris–Dauphine — [3]GREYC, CNRS-ENSICAEN —

Problem statement

Consider the **Hard-Thresholding** (HT) estimator

$$y \mapsto \mathrm{HT}(y,\lambda) \quad \text{and} \quad \mathrm{HT}(y,\lambda)_i = \begin{cases} 0 & \text{if } |y_i| < \lambda , \\ y_i & \text{otherwise} . \end{cases}$$

which aims at recovering x_0 from the observation y of the random variable

$$Y = x_0 + W$$

where we consider

- $x_0 \in \mathbb{R}^n$ the unknown **sparse** vector of interest,
- $y \in \mathbb{R}^n$ the noisy observation of x_0,
- $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$ the noise component,
- $\lambda > 0$ a regularization parameter.

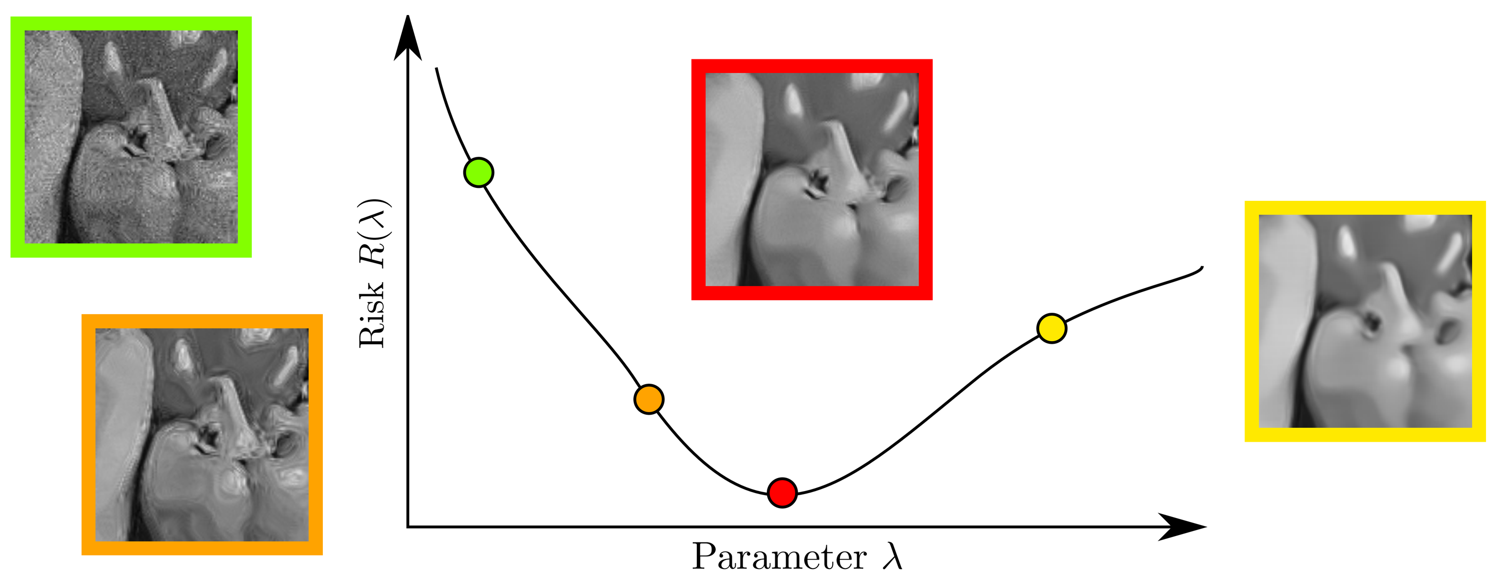
How to choose the value of the parameter λ?

Risk-based selection of λ

- Consider an estimator $y \mapsto x(y,\lambda)$ whith parameter λ.
- Risk associated to λ: measure of the expected quality of $x(y,\lambda)$ wrt x_0,

$$R(\lambda) = \mathbb{E}_W \| x(Y,\lambda) - x_0 \|^2 .$$

- The optimal (theoretical) λ minimizes the risk.



The risk is unknown since it depends on x_0.

Can we estimate the risk solely from $x(y,\lambda)$?

Unbiased risk estimation

Degree of freedom

- Degree Of Freedom (DOF) is defined by [Efron, 1986] as:

$$df\{x\}(x_0,\lambda) \triangleq \sum_{i=1}^n \frac{\mathrm{cov}(Y_i, x(Y_i,\lambda))}{\sigma^2} .$$

- The DOF plays an important role in model/parameter selection.

Risk estimation via SURE

- Assume $y \mapsto x(y,\lambda)$ is weakly differentiable.
- Define the Stein Unbiased Risk Estimator (SURE) as:

$$\mathrm{SURE}\{x\}(Y,\lambda) = \|Y - x(Y,\lambda)\|^2 - n\sigma^2 + 2\sigma^2 \widehat{df}\{x\}(Y,\lambda)$$

where $\widehat{df}\{x\}(y,\lambda) = \mathrm{div}\,(x(y,\lambda))$.

- Steim Lemma [Stein, 1981] implies:

$$\mathbb{E}_W(\widehat{df}\{x\}(Y,\lambda)) = df\{x\}(x_0,\lambda)$$

and $\mathbb{E}_W(\mathrm{SURE}\{x\}(Y,\lambda)) = \mathbb{E}_W(\|x_0 - x(Y,\lambda)\|^2)$.

The Hard-Tresholding is not weakly differentiable.

The DOF cannot be unbiasedly estimated from the divergence.

A biased DOF estimator for hard-thresholding

- Remark that the HT can be written as

$$\mathrm{HT}(y,\lambda) = \mathrm{ST}(y,\lambda) + D(y,\lambda)$$

where $\mathrm{ST}(y,\lambda)_i = \begin{cases} y_i + \lambda & \text{if } y_i < -\lambda \\ 0 & \text{if } -\lambda \leqslant y_i < +\lambda \\ y_i - \lambda & \text{otherwise} \end{cases}$

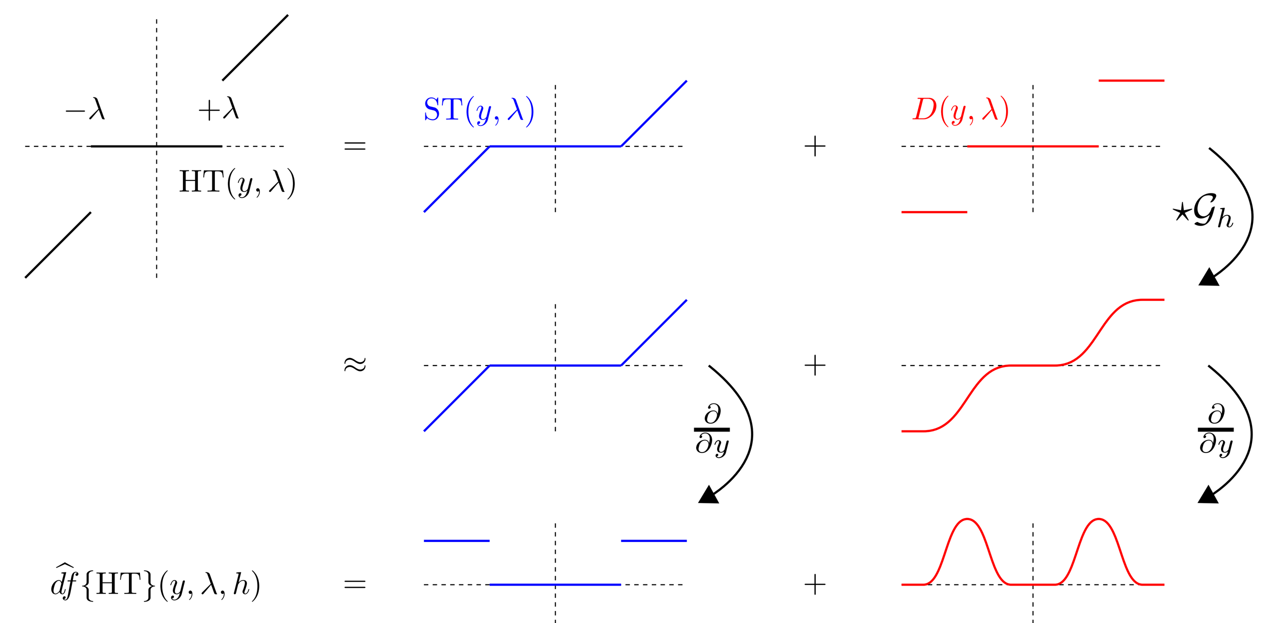
and $D(y,\lambda)_i = \begin{cases} -\lambda & \text{if } y_i < -\lambda \\ 0 & \text{if } -\lambda \leqslant y_i < +\lambda \\ +\lambda & \text{otherwise} \end{cases}$,

where $y \mapsto \mathrm{ST}(y,\lambda)$ is the soft thresholding operator, and:

  - $y \mapsto \mathrm{ST}(y,\lambda)$: Lipschitz continuous $\Rightarrow \widehat{df}\{\mathrm{ST}\}(y,\lambda) = \#\{|y| > \lambda\}$,
  - $y \mapsto D(y,\lambda)$: Piece-wise constant with discontinuities at $\pm\lambda \Rightarrow$ Stein's lemma does not apply.

- Consider a smoothed version $\mathcal{G}_h \star D(.,\lambda)$ where $\mathcal{G}_h$ is a Gaussian kernel of bandwidth h:
  - $y \mapsto (\mathcal{G}_h \star D(.,\lambda))(y)$: $C^\infty \Rightarrow$ Stein's lemma apply leading to:

$$y \mapsto \widehat{df}\{\mathrm{HT}\}(y,\lambda,h) = \#\{|y| > \lambda\} + \frac{\lambda\sqrt{\sigma^2+h^2}}{\sqrt{2\pi}\sigma h} \sum_{i=1}^n \left[\exp\left(-\frac{(y_i+\lambda)^2}{2h^2}\right) + \exp\left(-\frac{(y_i-\lambda)^2}{2h^2}\right) \right] .$$



$\widehat{\mathrm{df}}\{\mathrm{HT}\}(y,\lambda,h)$ is biased. How does its bias evolve w.r.t. n and h ?

Theorem (Stein's Consistent DOF estimator)

Let $Y = x_0 + W$ for $W \sim \mathcal{N}(x_0, \sigma^2 \mathrm{Id}_n)$.
Take $\widehat{h}(n)$ such that $\lim_{n\to\infty} \widehat{h}(n) = 0$ and $\lim_{n\to\infty} n^{-1}\widehat{h}(n)^{-1} = 0$.
Then

$$\mathrm{plim}_{n\to\infty} \frac{1}{n} \left(\widehat{df}\{\mathrm{HT}\}(Y,\lambda,\widehat{h}(n)) - df\{\mathrm{HT}\}(x_0,\lambda) \right) = 0 .$$

In particular

1. $\lim_{n\to\infty} \mathbb{E}_W\left[\frac{1}{n}\widehat{df}\{\mathrm{HT}\}(Y,\lambda,\widehat{h}(n))\right] = \lim_{n\to\infty} \frac{1}{n} df\{\mathrm{HT}\}(x_0,\lambda)$, *and*
2. $\lim_{n\to\infty} \mathbb{V}_W\left[\frac{1}{n}\widehat{df}\{\mathrm{HT}\}(Y,\lambda,\widehat{h}(n))\right] = 0$,

where $\mathbb{V}_W$ is the variance w.r.t. W.

If h decreases slower than $\frac{1}{n}$, the bias vanishes when n increases.

Corollary (Stein's Consistent Risk estimator)

Let $Y = x_0 + W$ for $W \sim \mathcal{N}(x_0, \sigma^2 \mathrm{Id}_n)$, and assume that $\|x_0\|_4 = o(n^{1/2})$.
Take $\widehat{h}(n)$ such that $\lim_{n\to\infty} \widehat{h}(n) = 0$ and $\lim_{n\to\infty} n^{-1}\widehat{h}(n)^{-1} = 0$.
Then, the **Stein COnsistent Risk Estimator (SCORE)** *evaluated at a realization y of Y*

$$\mathrm{SCORE}\{\mathrm{HT}\}(y,\lambda,\widehat{h}(n)) = \sum_{i=1}^n \left(I(|y_i| < \lambda) y_i^2 \right) - n\sigma^2 + 2\sigma^2 \widehat{df}\{\mathrm{HT}\}(y,\lambda,\widehat{h}(n)) ,$$

where $I(\omega)$ is the indicator for an event ω, is such that

$$\mathrm{plim}_{n\to\infty} \frac{1}{n} \left(\mathrm{SCORE}\{\mathrm{HT}\}(Y,\lambda,\widehat{h}(n)) - \mathbb{E}_W \|\mathrm{HT}(Y,\lambda) - x_0\|^2 \right) = 0 .$$

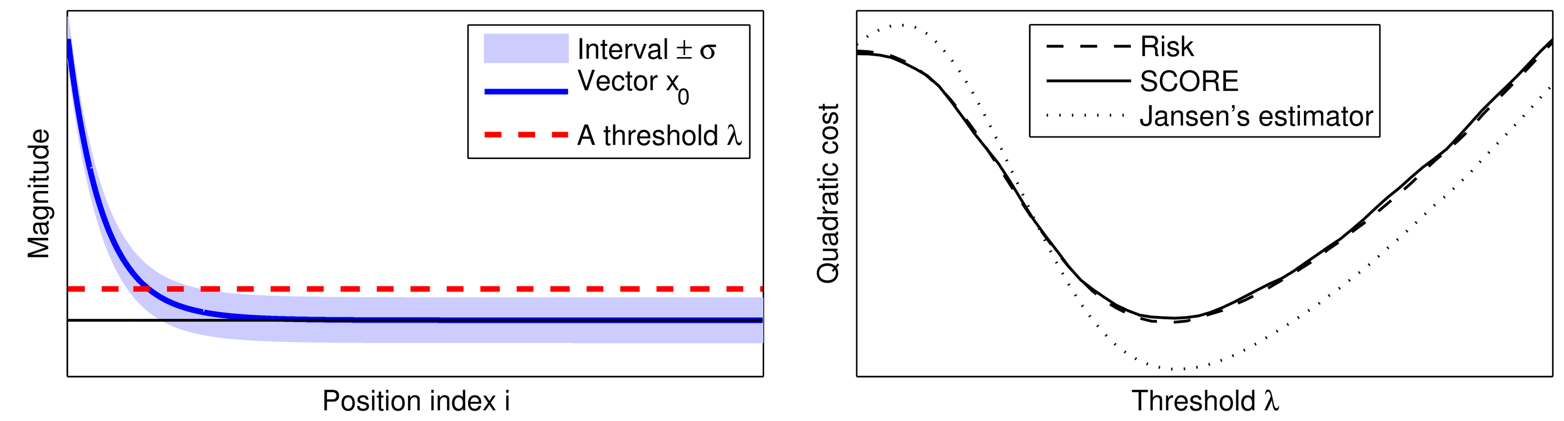
If h decreases slower than $\frac{1}{n}$, the SCORE is consistent.

Numerical example – Recovering of a compressible vector

- Consider x_0 a compressible vector of length $n = 2\mathrm{E}5$ whose sorted values in magnitude decay as

$$|x_0|_{(i)} = 1/i^\gamma \quad \text{for} \quad \gamma > 0$$

- Consider σ chosen such that the SNR of y is of about 5.65dB



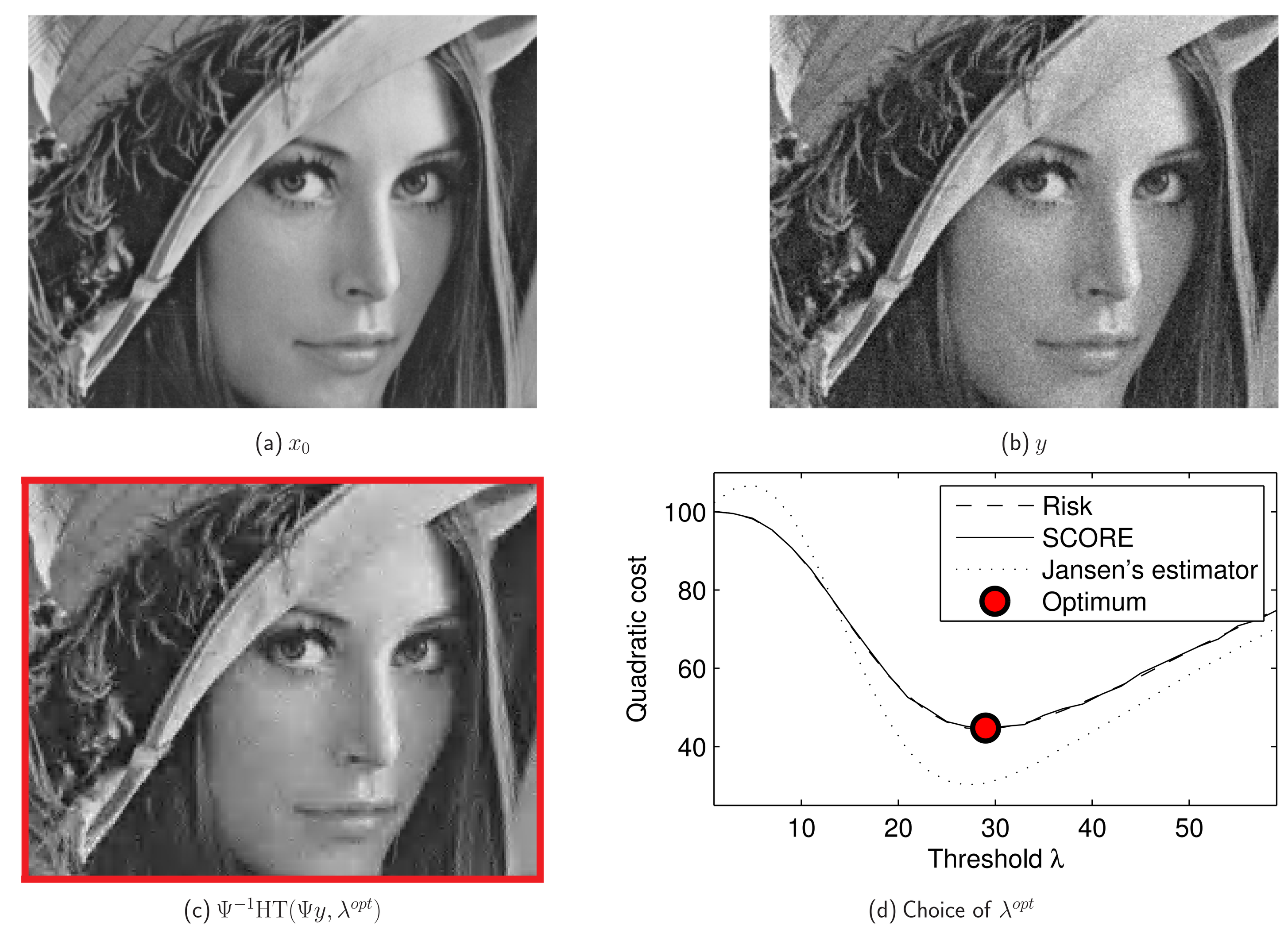
- Compare to Jansen's estimator [Jansen, 2011]

$$\sum_{i=1}^n \left(I(|y_i| < \lambda) y_i^2 \right) - n\sigma^2 + 2\sigma^2 \#\{|y| > \lambda\} + \frac{2\sigma\lambda}{\sqrt{2\pi}} \sum_{i=1}^n \exp\left(-\frac{\lambda^2}{2\sigma^2}\right)$$

Numerical example – Image denoising

- Consider x_0 an image quantified on $[0, 255]$ and $\sigma = 10$.
- Assume Ψx_0 is sparse where Ψ is an orthonormal wavelet basis.
- Since W is white and using Parseval identity:

$$\frac{1}{n}\mathrm{SCORE}\{\mathrm{HT}\}(\Psi Y, \lambda, \hat{h}(n)) \quad \text{consistently estimates} \quad \frac{1}{n}\mathbb{E}_W \|\Psi^{-1}\mathrm{HT}(\Psi Y,\lambda) - x_0\|^2$$



(a) x_0 (b) y (c) $\Psi^{-1}\mathrm{HT}(\Psi y, \lambda^{opt})$ (d) Choice of λ^{opt}

Perspectives

- Deeper investigation of the choice of $\hat{h}(n)$.
- Extend to other non-continuous estimators and inverse problems:
  - Iterative Hard-Thresholding,
  - Ill-conditioned observation operators,
  - Redundant dictionnaries.

References

Efron, B. (1986).
How biased is the apparent error rate of a prediction rule?
Journal of the American Statistical Association, 81(394):461–470.

Jansen, M. (2011).
Information criteria for variable selection under sparsity.
Technical report, Technical report, ULB.

Stein, C. (1981).
Estimation of the mean of a multivariate normal distribution.
The Annals of Statistics, 9(6):1135–1151.