Stein COnsistent Risk Estimator (SCORE) for hard thresholding

Charles-Alban Deledalle
Institut de Mathématiques de Bordeaux
CNRS-Université Bordeaux 1
Bordeaux, France
Email: cdeledal@math.u-bordeaux1.fr

Gabriel Peyré
CEREMADE
CNRS-Université Paris Dauphine
Paris, France
Email: gabriel.peyre@ceremade.dauphine.fr

Jalal Fadili
GREYC
CNRS-ENSICAEN
Caen, France
Email: jalal.fadili@greyc.ensicaen.fr

***Abstract*—In this work, we construct a risk estimator for hard thresholding which can be used as a basis to solve the difficult task of automatically selecting the threshold. As hard thresholding is not even continuous, Stein's lemma cannot be used to get an unbiased estimator of degrees of freedom, hence of the risk. We prove that under a mild condition, our estimator of the degrees of freedom, although biased, is consistent. Numerical evidence shows that our estimator outperforms another biased risk estimator proposed in [1].**

I. Introduction

We observe a realisation $y \in \mathbb{R}^P$ of the normal random vector $Y = x_0 + W$, $W \sim \mathcal{N}(x_0, \sigma^2 \mathrm{Id}_P)$. Given an estimator $y \mapsto x(y, \lambda)$ of x_0 evaluated at y and parameterized by λ, the associated Degree Of Freedom (DOF) is defined as [2]

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

The DOF plays an important role in model/parameter selection. For instance, define the criterion

$$\|Y - x(Y, \lambda))\|^2 - P\sigma^2 + 2\sigma^2 \widehat{df}\{x\}(Y, \lambda) \ . \quad (2)$$

If $x(\cdot, \lambda)$ is weakly differentiable w.r.t. its first argument with an essentially bounded gradient, Stein's lemma [3] implies that $\widehat{df}\{x\}(Y, \lambda) = \mathrm{div}\,(x(Y, \lambda))$ and (2) (the SURE in this case) are respectively unbiased estimates of $df\{x\}(x_0, \lambda)$ and of the risk $\mathbb{E}_W \|x(Y, \lambda) - x_0\|^2$. In practice, (2) relies solely on the realisation y which is useful for selecting λ minimizing (2).

In this paper, we focus on Hard Thresholding (HT)

$$y \mapsto \mathrm{HT}(y, \lambda)_i = \begin{cases} 0 & \text{if } |y_i| < \lambda \ , \\ y_i & \text{otherwise} \ . \end{cases} \quad (3)$$

HT is is not even continuous, and the Stein's lemma does not apply, so that $df\{x\}(x_0, \lambda)$ and the risk cannot be unbiasedly estimated [1]. To overcome this difficulty, we build an estimator that, although biased, turns out to enjoy good asymptotic properties. In turn, this allows efficient selection of the threshold λ.

II. Stein COnsistent Risk Estimator (SCORE)

We define, for $h > 0$, the following DOF formula

$$y \mapsto \widehat{df}\{\mathrm{HT}\}(y, \lambda, h) = \#\{|y| > \lambda\} + \frac{\lambda\sqrt{\sigma^2 + h^2}}{\sqrt{2\pi}\sigma h} \sum_{i=1}^{P} \left[\exp\left(-\frac{(y_i + \lambda)^2}{2h^2}\right) + \exp\left(-\frac{(y_i - \lambda)^2}{2h^2}\right)\right] \quad (4)$$

where $\#\{|y| > \lambda\}$ is the number of entries of $|y|$ greater than λ.

Theorem 1: Let $Y = x_0 + W$ for $W \sim \mathcal{N}(x_0, \sigma^2 \mathrm{Id}_P)$. Take $\widehat{h}(P)$ such that $\lim_{P\to\infty} \widehat{h}(P) = 0$ and $\lim_{P\to\infty} P^{-1}\widehat{h}(P)^{-1} = 0$. Then $\mathrm{plim}_{P\to\infty} \frac{1}{P}\left(\widehat{df}\{\mathrm{HT}\}(Y, \lambda, \widehat{h}(P)) - df\{\mathrm{HT}\}(x_0, \lambda)\right) = 0$. In particular

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

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

The proof is available in the extended version of this abstract [4]. An immediate corollary of Theorem 1, also given in [4], is that (4) and (2) provide together the Stein COnsistent Risk Estimator (SCORE) which is biased but consistent. Fig. 1 summarizes the pseudo-code when applying SCORE to automatically find the optimal threshold λ that minimizes SCORE in a predefined (non-empty) range.

Algorithm Risk estimation for Hard Thresholding

Inputs: observation $y \in \mathbb{R}^P$, threshold $\lambda > 0$
Parameters: noise variance $\sigma^2 > 0$
Output: solution $x^\star$

Initialize $h \leftarrow \widehat{h}(P)$
for all λ in the tested range **do**
  Compute $x \leftarrow \mathrm{HT}(y, \lambda)$ using (3)
  Compute $\widehat{df}\{\mathrm{HT}\}(y, \lambda, h)$ using (4)
  Compute SCORE at y using (2)
end for
return $x^\star \leftarrow x$ that provides the smallest SCORE

Fig. 1. Pseudo-algorithm for HT with SCORE-based threshold optimization.

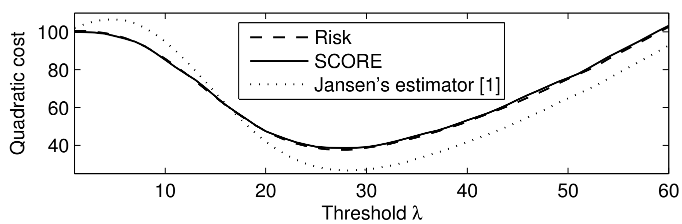

Fig. 2. Risk and its SCORE estimate with respect to the threshold λ.

III. Experiments and Conclusions

Fig. 2 shows the evolution of the true risk, the SCORE and the risk estimator of [1] as a function of λ where x_0 is a compressible vector of length $P = 2\mathrm{E}5$ whose sorted values in magnitude decay as $|x_0|_{(i)} = 1/i^\gamma$ for $\gamma > 0$, and we have chosen σ such that the SNR of y is of about 5.65dB and $\widehat{h}(P) = 6\sigma/P^{1/3} \approx \sigma/10$. The optimal λ is found around the minimum of the true risk.

Future work will concern a deeper investigation of the choice of $\widehat{h}(P)$, comparison with other biased risk estimators, and extensions to other non-continuous estimators and inverse problems.

References

[1] M. Jansen, "Information criteria for variable selection under sparsity," Technical report, ULB, Tech. Rep., 2011.
[2] B. Efron, "How biased is the apparent error rate of a prediction rule?" *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 461–470, 1986.
[3] C. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.
[4] C. Deledalle, G. Peyré, and J. Fadili, "Stein consistent risk estimator (SCORE) for hard thresholding," *Preprint HAL-00776303*, 2013.