Exponential Screening and optimal rates of sparse estimation, a review

Franco Marchesoni & Antoine Van Biesbroeck
MVA Bayesian Machine Learning – Project report

1 Summary

In this work we present a review of the paper titled “Exponential Screening and optimal rates for sparse estimation” ([3]). We frame the problem in a Bayesian way and address some of the shortcomings, proposing a few enhancements. Furthermore, we expose experimental results on the original and the enhanced method, proving the contribution useful.

2 The reference paper in a nutshell

The reference paper's work is around the optimal estimation of sparse coefficients for regression. The work contributions are the following:

- The proposal of the Exponential Screening (ES) estimator along with an algorithm and experimental results of its superiority.
- Theoretical guarantees for the ES estimator to attain the *optimal rate of sparse estimation*: it is the only estimator known that satisfies the minimum expected error upper bound, i.e. the best estimator known in the worst-case.
- Optimal rates of aggregation for the regression model with fixed X.

The authors cite a very similar and parallel work by [1]. The ES estimator is an estimator of coefficients of a linear regression that present some sparsity properties.

3 Description of the model as a decision problem

Let us consider the classical linear regression decision problem defined by state space $\mathcal{S}$ and action space $\mathcal{A}$:

$$\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \qquad \mathcal{A} = \{f_\theta = \sum_{j=1}^{M} \theta_j f_j,\, \theta \in \mathbb{R}^M\}$$

for some $f_1, \ldots, f_M$. A common way to process with this problem is to consider the prior model $Y = X\theta^* + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$ where $X = (f_j(x_i))_{i,j}$ on the observations and the loss function $L(Y, X, \theta) = |Y - X\theta|_2^2$ which leads to the least squares estimator $\hat{\theta} \in \arg\min |Y - X\theta|_2^2$.

The issue of this resolution is that it does not encourage the sparsity of the solution (i.e. minimizing the number of non null coefficients in the estimator). A common approach in the literature to face this issue is to add a term in the loss function : $L(Y, X, \theta) = |Y - X\theta|_2^2 + h(\theta)$, where h penalize non-sparsity. LASSO and BIC estimators are the most popular examples.

In this paper, the idea is different and consists into the introduction of sparsity inside the decision problem's state of actions :

$$\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \qquad \mathcal{A} = \mathbb{R}^m \times \mathcal{P},\, \mathcal{P} = \{0, 1\}^M$$

and a loss function we could consider is $L(Y, X, \theta, p) = \mu|Y - X\theta|_2^2 + \lambda|p|$ for some $\mu, \lambda > 0$. Here we will take $\mu = \frac{1}{4\sigma^2}$ and $\lambda = \frac{1}{2}$.

4 SPA and ES estimators

Minimizing the last loss function presented in previous section would result into an too complex problem. That's why instead we consider a more easy to calculate estimator which is the following:

Definition 1 *From a prior π on $\mathcal{P}$, the sparsity pattern aggregate (SPA) estimator is*

$$\tilde{\theta}^{SPA} = \sum_{p \in \mathcal{P},\, \pi_p \neq 0} \hat{\theta}_p \operatorname{softmax}\left(-L(X, Y, \theta, p) + \log \pi_p\right)$$

where $\hat{\theta}_p$ denote the least squares estimator $\hat{\theta}_p \in \arg\min_{\theta \in \mathbb{R}^p} |Y - X\theta|_2^2$, $R^p = \{\theta \cdot p,\, \theta \in \mathbb{R}^M\}$.

When $\pi_p = \begin{cases} \frac{1}{H}\left(\frac{|p|}{2eM}\right)^{|p|} & \textit{if } |p| < R \\ \frac{1}{2} & \textit{if } |p| = M \\ 0 & \textit{otherwise} \end{cases}$ *we call this estimator the Exponential Screening (ES) estimator.*

The estimator $\tilde{\theta}^{EM}$ is then a barycenter of some sparse least squares estimators $\hat{\theta}_p$ with more weights given to the ones that

1. have a lower square error $|Y - X\hat{\theta}_p|_2^2$
2. are sparser i.e. $|p|$ is low
3. have a higher sparsity probability, i.e. π_p is high

Calculations coming from the expression of $\tilde{\theta}^{ES}$ and π allows us to obtain the following evaluations of the Exponential Screening estimator:

Theorem 1 *For any $\theta \in \mathbb{R}^M$, if we denote $R = \operatorname{rk} X$ we have*

$$\begin{aligned}\mathbb{E}\|f_{\tilde{\theta}^{ES}} - \eta\|^2 &\leq \|f_\theta - \eta\|^2 + \frac{\sigma^2 R}{n} \wedge \frac{9\sigma^2 M(\theta)}{n} \log\left(1 + \frac{eM}{M(\theta) \vee 1}\right) + \frac{8\sigma^2}{n}\log 2 \\ &= \|f_\theta - \eta\|^2 + O\left(Rn^{-1} \wedge |\theta|_0 n^{-1} \log(1 + M) + n^{-1}\right)\end{aligned}$$

The major result we can deduce from this theorem is that if there exists $\theta^* \in \mathbb{R}^M$ such that $\eta = f_{\theta^*}$ then the function $f_{\tilde{\theta}^{ES}}$ converges to η on the dataset when $n \longrightarrow \infty$.

Some more advanced calculations lead the to the theorem bellow which gives a more precise comparison:

Theorem 2 *We assume $\max_j \|f_j\| \leq 1$. Then for any $\theta \in \mathbb{R}^M$ we have*

$$\mathbb{E}\|f_{\tilde{\theta}^{ES}} - \eta\|^2 \leq \|f_\theta - \eta\|^2 + \varphi_{n,M}(\theta) + \frac{\sigma^2}{n}(9\log(1 + eM) + 8\log 2)$$

where $\varphi_{n,M}(\theta)$ can be summarized as

$$\varphi_{n,M}(\theta) = O\left(Rn^{-1} \wedge |\theta|_0 n^{-1} \log(1 + M) \wedge |\theta|_1 \left(n^{-1} \log\left(1 + \frac{M}{|\theta|_1 \sqrt{n}}\right)\right)^{1/2}\right)$$

this way :

$$\begin{aligned}&\mathbb{E}\|f_{\tilde{\theta}^{ES}} - \eta\|^2 \leq \|f_\theta - \eta\|^2 \\ &\qquad + O\left(Rn^{-1} \wedge |\theta|_0 n^{-1} \log(1 + M) \wedge |\theta|_1 \left(n^{-1} \log\left(1 + \frac{M}{|\theta|_1 \sqrt{n}}\right)\right)^{1/2} + n^{-1}\right)\end{aligned}$$

The above theorems make sure that the ES estimator is still viable on high dimension. While $n >> \log(1 + M)$ we still have the convergence of it to η on the dataset if there exists a θ^* such that $f_{\theta^*} = \eta$.

5 Comparison with other sparse estimators

As stated before, some other sparse estimators for linear regression are largely studied and popular in the literature. In this section we will provide a comparison between the ES estimator andthe LASSO estimator defined bellow

Definition 2 *We consider our linear regression model : find θ such that $Y - X\theta = 0$. The LASSO estimator is defined as*

$$\hat{\theta}^L \in \underset{\theta \in \mathbb{R}^M}{\arg\min} \{ \frac{1}{n} |Y - X\theta|_2^2 + \lambda |\theta|_1 \}$$

for some $\lambda > 0$.

This estimator is the solution of an optimization problem whose aim is first to minimize the square error $|Y - X\theta|_2$ and second to minimize a penalty term that encourage sparsity of the solution.

From results demonstrated in [2] and computations we can state the following:

Theorem 3 *We assume $\max_j \|f_j\| \leq 1$. We consider $\hat{\theta}^L$ with $\lambda = A\sigma\sqrt{\frac{\log M}{n}}$, $A > 2\sqrt{2}$. Then we have for any $\theta \in \mathbb{R}^M$*

$$\|f_{\hat{\theta}^L} - \eta\|^2 \leq \|f_\theta - \eta\|^2 + 2A\sigma \frac{|\theta|_1}{\sqrt{n}} \sqrt{\log M}$$

with probability at least $1 - M^{1-A^2/8}$.

This theorem give us a result that is hard to compare with theorem 2 as it states a probability to have an inequality and not a result about the expected value of the square error. However, if we admit the comparison stands we can say that the upper bound in $O\left(\frac{|\theta|_1}{\sqrt{n}}\sqrt{\log M}\right)$ is worse than the one in $O\left(|\theta|_1 \left(n^{-1} \log\left(1 + \frac{M}{|\theta|_1\sqrt{n}}\right)\right)^{1/2}\right)$ that we have for ES estimator.

To go further into the comparison, we can notice that theorem 2 provide a minimax rate for the ES estimator, i.e. a worst possible risk rate compared to any estimator. Actually, this rate is optimal in the sense that it is the best possible rate that any estimator can reach under some assumptions according to following theorem.

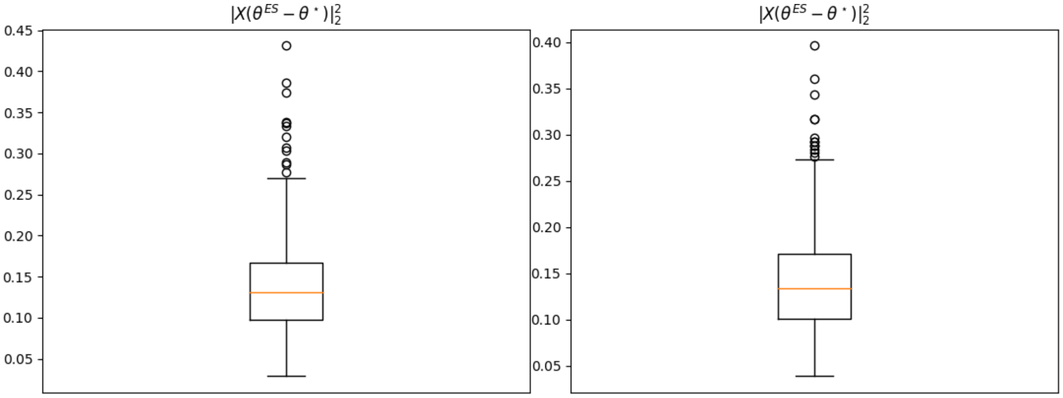
Theorem 4 *Let $\zeta_{n,M,R}(S,\delta) = \frac{\sigma^2 R}{n} \wedge \frac{\sigma^2 S}{n} \log(1 + \frac{eM}{S}) \wedge \frac{\sigma S}{\sqrt{n}} \sqrt{\log\left(1 + \frac{eM\sigma}{\delta\sqrt{n}}\right)} \wedge \delta^2$, then if $X \in \mathcal{D}(S \wedge (m \vee 1), \kappa)$, for any estimator T_n, there exists constants such that*

$$\sup_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_0 \leq S,\, |\theta|_1 \leq \delta}} \sup_\eta \{\mathbb{E}_\eta \|T_n - \eta\|^2 - \|f_\theta - \eta\|^2\} \geq c^* \kappa \zeta_{n,M,R}(S,\delta)$$

As ζ is the rate of the upper bound in theorem 2 the above theorem states the optimality of $\tilde{\theta}^{ES}$ in the sense that it reaches the minimal minimax rate.

6 Numerical simulations

In this section we present three different numerical experiments. Two of them reproduce and check results obtained in the reference paper, while the last one explores an extension that was suggested in the paper but not explored. This extension is furthermore enhanced by better using the information provided by the ES estimator. This algorithm could benefit from a closer-to-metal implementation, but not so much from parallelization, as the Metropolis-Hastings algorithm as described in the paper is iterative. Our implementation is available at `franchesoni/exponential_screening` public repository (click here).



(a) Gaussian. Ref: 0.12 (0.06).

(b) Rademacher. Ref: 0.12 (0.07).

Figure 1: Boxplots of our simulations for $(n.M, S) = (100, 200, 10)$. Values are coherent with reference values expressed as Reference: mean (std).

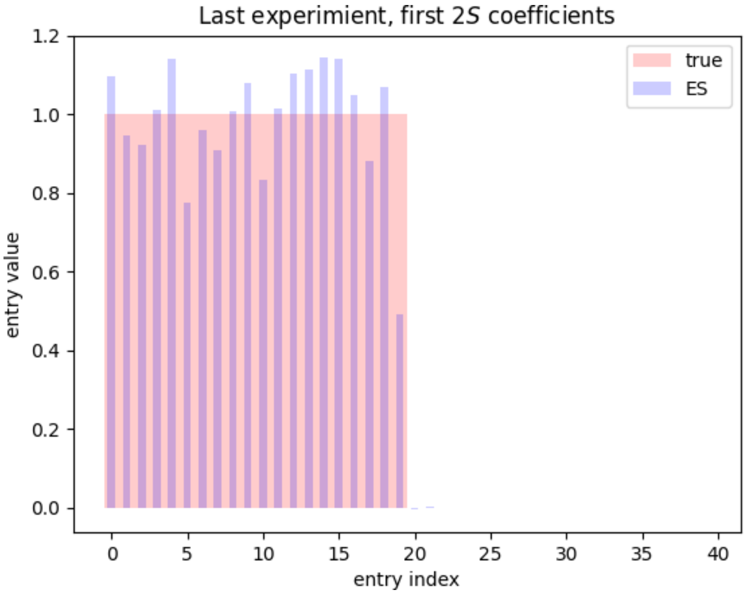


Figure 2: First $2S$ coefficients.

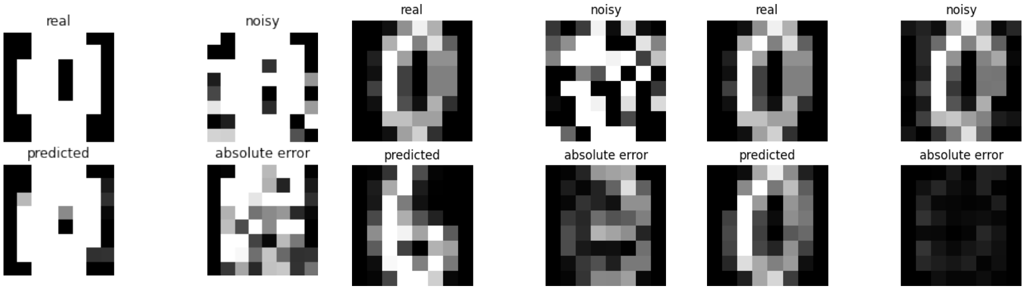
6.1 Synthetic Problem

As in the reference paper, we generate random matrices $X \in \mathrm{R}^{n\times M}$ containing independent random values sampled from a standard gaussian or from a Rademacher distribution. We define $\theta_i^\star = \mathbb{1}_{i\leq S}$ to be a vector with S ones and $M - S$ zeros. Finally, we sample ξ from a standard gaussian and compute $Y = X\theta^\star + \sigma\xi$, with $\sigma = S/9$. The aim is to recover $\theta^\star$ correctly.

As our Python implementation was successful (although not optimized for efficiency), we could reproduce the results. See for instance Figures 1a and 1b. According to the experiments for other estimators in the reference paper, the previous results are better in terms of performance. One can see the first $2S$ estimated coefficients in Figure 2. Recall that the first S should be ones and the rest should be zeros, thus the reconstruction is very good in the sense that it correctly selects the first S indices between M possible indices.

6.2 Digit denoising problem

Although not strictly following the noisy linear regression prior model, one could try to reconstruct the true digit x_0 from a noisy version of the digit image $y_0 = x_0 + \sigma\xi$. The coefficients θ to be



(a) Clipped inputs.

(b) High noise input.

(c) Successful example.

Figure 3: Comparison of digit reconstruction for clipped inputs, high noise inputs and small noise inputs.

learned are the ones that sparsely minimize $|Y - X\theta|$. We would like a sparse recovery selecting those digits between the columns of X corresponding to the same number (i.e. images of a 0s), as we expect that a weighted sum of them can closely approximate x_0. We get different results if we normalize, add too much noise, or not. These results are exposed in Figure 3. We did not use the exact same database, but `scikit-learn` digits dataset.

6.3 Digit classification problem

One can potentially analyse the selected digits and do a majority voting to "classify" the noisy one. A classification problem could be thus thought as setting Y to be a test example and X to be vectors of the training set. Using this formulation we implement two classifiers, one using the suggested majority voting, and one using a weighted voting according to the values in θ. The drawback of this approach is that it takes some substantial time to process each sample in the test set. With an arbitrary $\sigma = 1$, we find the comparison in Table 6.3. Although the proposed weighted voting is better than majority voting, we do not recommend it over other classification approaches (SVM), that are better both in performance and speed.

	Weighted voting				Majority voting			
digit	precision	recall	f1-score	support	precision	recall	f1-score	support
0	0.92	0.95	0.94	88	0.48	0.83	0.61	88
1	0.82	0.91	0.86	91	0.54	0.66	0.59	91
2	0.91	0.87	0.89	86	0.54	0.58	0.56	86
3	0.74	0.78	0.76	91	0.43	0.49	0.46	91
4	0.95	0.86	0.90	92	0.59	0.59	0.59	92
5	0.93	0.84	0.88	91	0.59	0.52	0.55	91
6	0.91	1.00	0.95	91	0.63	0.63	0.63	91
7	0.94	0.91	0.93	89	0.67	0.40	0.50	89
8	0.72	0.70	0.71	88	0.49	0.30	0.37	88
9	0.79	0.79	0.79	92	0.52	0.38	0.44	92
accuracy			0.86	899			0.54	899
macro avg	0.86	0.86	0.86	899	0.55	0.54	0.53	899
weighted avg	0.86	0.86	0.86	899	0.55	0.54	0.53	899

References

[1] Pierre Alquier and Karim Lounici. "PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights". working paper or preprint. Aug. 2010. URL: `https://hal.archives-ouvertes.fr/hal-00465801`.

[2] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. "Simultaneous analysis of lasso and Dantzig selector". In: *The Annals of Statistics* 37.4 (2007), pp. 1705–1732. DOI: `10.1214/00905360600001587`. URL: `https://arxiv.org/abs/0801.1095`.

[3] Philippe Rigollet and Alexandre Tsybakov. "Exponential Screening and optimal rates of sparse estimation". In: *The Annals of Statistics* 39.2 (2011), pp. 731–771. DOI: `10.1214/10-AOS854`. URL: `https://doi.org/10.1214/10-AOS854`.