University of Montpellier

Master 2 Biostatistics

HMMA307 project

Sparse Regularization via Bidualization

Student:
Tanguy Lefort

Teacher:
Joseph salmon

Reproducing parts of the article of
Amir Beck and Yehonathan Refael

2020

Contents

Introduction

The increasing.....

1 Solving problems with regularization

Regularized problems are a very useful way to obtain a solution in many situations. We write the problem as

$$\min_{x\in\mathbb{R}^n} f(x) + \text{pen}(x) \ ,$$

where f only depends of the data collected and pen is a penalty term. It can represent for example a constraint over the number of variables that play a role. In the case of the Tikhonov-ridge regularization, for a matrix $A \in \mathbb{R}^{n\times n}$ made from the data and an observed signal $b \in \mathbb{R}^n$, we write the problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \frac{\lambda}{2}\|x\|_2^2 \ ,$$

where $\lambda \geq 0$ represents how strong the penalty on the l_2 norm of the solution is. Note that for $\lambda = 0$ we retrieve the least square problem, and when $\lambda \to +\infty$, we have $x = 0$. However, when working with an observed signal, we sometimes know in advance the sparsity (or a reasonable estimate) of the original one. So we have an information that is stronger than a constraint on the 2-norm of the vector, we have a constraint on the number of non-zero values, meaning on the l_0 norm.

Definition 1. *Let $x \in \mathbb{R}^n$, the l_0 norm of x noted $\|x\|_0$ is the number of non-zero values:*

$$\|x\|_0 := |\{i \,|\, x_i \neq 0\}| \ .$$

We note C_k the space of the k-sparse vectors: $C_k := \{x \in \mathbb{R}^n \,|\, \|x\|_0 = k\}$.

One of the main advantages of the Tikhonov (l_2) regularization is the differentiability, the convexity and the continuity. With the l_0 norm, we lose them all. This lead to a compromise called the **LASSO** using the l_1 norm to keep the continuity and use sub-differentials instead of regular ones. The **LASSO** method will select some variables (depending on the value of λ) to explain the observed signal. It is very useful in a lot of situations like in genomics, but in situations with highly correlated groups of variables, the **LASSO** tends to only select one from each group and force the others to a null-effect as we can see on Figure 1. This issue lead to the development of the **Elastic net** method which uses a penalizer that is a combination of the Tikhonov and **LASSO** penalty

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda_1\|x_1\| + \frac{\lambda_2}{2}\|x\|_2^2 \ .$$

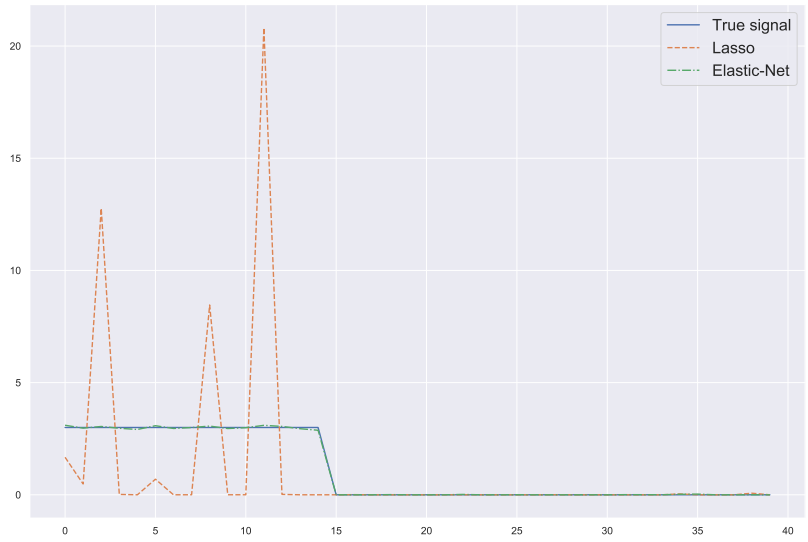

Figure 1: Representation of the grouping effect of the LASSO in a highly correlated situation and how the Elastic net is better is that case.

The proposition of [2] is to keep the l_0 norm for the sparsity penalty and use the l_2 norm in the **Elastic net** for the grouping effect. The penalty is then:

$$s_k(x) = \begin{cases} \frac{1}{2}\|x\|_2^2, \text{ if } \|x\|_0 \leq k \\ \infty, \text{ else} \end{cases} \quad = \frac{1}{2}\|x\|_2^2 + \delta_{C_k} \ .$$

However, the first problem is still there. We don't have neither the continuity nor the convexity. That's why instead of considering s_k, we will consider it's best convex estimator, the sparse envelope:

$$\mathcal{S}_k(x) := s_k^{**}(x) \ ,$$

where, for a function f, we define its convex conjugate (also called Fenchel's conjugate) by

$$f^*(y) = \max_{x \in \mathbb{R}^n} \{\langle x, y\rangle - g(x)\} \ . \tag{1}$$

We now need to find an efficient way to solve the inverse problem with this penalty.

2 Convex biregularization

First, let's note $\left|x_{\langle i \rangle}\right|$ the i^{th} biggest (in absolute value) component of the vector x. The best approximation (w.r.t the l_2 norm) in C_k for a determined $k \in \mathbb{N}^*$ of a signal $x \in \mathbb{R}^n$ is noted $H_k(x)$ and we define its components with

$$\big(H_k(x)\big)_j = \begin{cases} x_j, \text{ if } x_j \geq \left|x_{\langle k \rangle}\right| \\ 0, \text{ else} \end{cases} \quad \forall j = 1, \dots, n \ .$$

Now, we can try to find an expression for $\mathcal{S}_k(x)$.

Proposition 1. *Let $y \in \mathbb{R}^n$ and $k \in \mathbb{N}^*$, then $s_k^*(y) = \frac{1}{2}\|H_k(y)\|_2^2$.*

Proof. Using (1), we know that for $y \in \mathbb{R}^n$,

$$\begin{aligned} s_k^*(y) &= \max_{x\in\mathbb{R}^n} \left[\langle x, y\rangle - s_k(x)\right] = \max_{x\in\mathbb{R}^n} \left[\langle x, y\rangle - \frac{\|x\|_2^2}{2}\right] = \max_{x\in\mathbb{R}^n} \left[-\left(\frac{\|x\|_2^2}{2} - \langle x, y\rangle\right)\right] \\ &= \max_{x\in\mathbb{R}^n} \left[-\frac{\|x-y\|_2^2}{2} + \frac{\|y\|_2^2}{2}\right] \ . \end{aligned}$$

Besides, $H_k(x) \in \arg\min_{y\in C_k} \|x-y\|_2$ and $\|y\|_2^2$ is independent of x, so

$$s_k^*(y) - \frac{1}{2}\|H_k(y) - y\|_2^2 + \frac{1}{2}\|y\|_2^2 \ .$$

And $-\|H_k(y)-y\|_2^2 = -\|H_k(y)\|_2^2 - \|y\|_2^2 + 2\langle H_k(y), y\rangle$. But because $(H_k(y))_j = 0$ for the $n-k$ smallest absolute values of y, $\langle H_k(y), y\rangle = \|H_k(y)\|_2^2$. Thus, we obtain by plugging-in this value:

$$s_k^*(y) = -\frac{1}{2}\|H_k(y)\|_2^2 \ .$$

□

To find the biconjugate of $s_k(y)$, we now only have to find the conjugate of $-\frac{1}{2}\|H_k(y)\|_2^2$. Let's introduce the space $D_k = \{u \in \mathbb{R}^n \mid \sum_{i=1}^n u_i \le k,\ 0 \le u_i \le 1\ \forall i\}$. Then

$$\mathcal{S}_k(x) = \max_{y\in\mathbb{R}^n} \min_{u\in D_k} \left\{\langle x, y\rangle - \frac{1}{2}\sum_{i=1}^n u_i y_i^2\right\} \ . \tag{2}$$

Thanks to Sion's *minimax* theorem [4] thanks to the convexity and concavity, we can inverse the min and max we can switch the two operators and get an expression of $\mathcal{S}_k(x)$.

Proposition 2. *With the same notations we have*

$$\mathcal{S}_k(x) = \frac{1}{2}\min_{u\in D_k} \sum_{i=1}^n \phi(x_i, u_i), \text{ where } \phi(x,u) = \begin{cases} \frac{x^2}{u}, & \text{if } u > 0 \\ 0 & \text{if } u = x = 0 \\ \infty & \text{else} \end{cases} \ . \tag{3}$$

Proof. We only have to consider the inner part of (2). Let $i_0 \in \{1, \dots, n\}$ an index. We want to find the value at the optimum, we can calculate the first order conditions:

$$\frac{\partial}{\partial y_{i_0}} \sum_{i=1}^n x_i y_i - \frac{1}{2}\sum_{i=1}^n u_i y_i^2 = x_{i_0} - \frac{1}{2} 2 u_{i_0} y_{i_0} \ .$$

At the optimum, equalizing to 0, if $u_{i_0} > 0$ we indeed get $\hat{y}_{i_0} = \frac{x_{i_0}}{u_{i_0}}$ and the other values are trivial in the other cases. Plugging-in the obtained value, we finally get the result for $u_i > 0$ (the other cases are direct and the second derivative is indeed negative):

$$\max_{y\in\mathbb{R}^n} \left\{x'y - \frac{1}{2}\sum_{i=1}^n u_i y_i^2\right\} = \sum_{i=1}^n x_i \frac{x_i}{u_i} - \frac{1}{2}\sum_{i=1}^n u_i \frac{x_i^2}{u_i} = \frac{1}{2}\sum_{i=1}^n \frac{x_i^2}{u_i} \ .$$

□

From there, it can be proved [2], that if $\|x\|_0 \leq k$ then $\mathcal{S}_k(x) = \|x\|_2^2/2$. Otherwise, the expression depends on the root of a function defined as the sum of linear monotonous piecewise functions with a single breakpoint:

$$g_x(\eta) = \sum_{i=1}^{n} \min\{|x_i|\,\eta, 1\} - k,\ \eta \geq 0\ .$$

The general expression of $\mathcal{S}_k(x)$ is:

$$S_k(x) = \frac{1}{2}\sum_{i=1}^{N_x} x_{\langle i\rangle}^2 + \frac{1}{2(k-N_x)}\left(\sum_{i=N_x+1}^{n} \left|x_{\langle i\rangle}\right|^2\right)^2\ , \tag{4}$$

where $N_x = \arg\max_{i=1,\dots,n}\left\{\left|x_{\langle i\rangle}\right| \geq \tilde{\eta}^{-1}\right\}$, with $\tilde{\eta}$ the root of g_x.

3 Proximal operator and numerical results

3.1 Compute the proximal operator

In many cases, to improve the computation cost of a function f, we consider the proximal mapping. Fast algorithm already exist to compute these objects such as FISTA [3], the proximal gradient method [5], ...

Definition 2 (Moreau's proximal mapping)**.** *Let f be a real function, its proximal operator (or Moreau's proximal operator)* prox_f *is defined as follows:*

$$\forall x \in \mathbb{E},\ \text{prox}_f(x) = \arg\min_{u\in\mathbb{E}}\left\{f(u) + \frac{1}{2}\|u-x\|_2^2\right\}\ .$$

This can be either a set of multiple elements, an empty set or a singleton. In our case, the following lemma and theorem assure that $\left|\text{prox}_{\mathcal{S}(x)}\right| = 1$. The proof is available in [2].

Lemma 1. *The sparse envelope $\mathcal{S}_k$ is closed* ie *for all $y \in \mathbb{R}$, $\{x \in \mathcal{D}_f \,|\, f(x) \leq y\}$ is closed.*

Theorem 1. *For a proper closed convex function f, $\left|\text{prox}_f(x)\right| = 1$ for all $x \in \mathbb{E}$.*

When solving an inverse problem, as we saw with the ridge penalty or **Elastic net** , there is an hyperparameter multiplying the penalty factor. Let's note it $\lambda_s > 0$ for the sparse envelope. Then 1 still holds and there is an almost explicit formulation for the value of $\text{prox}_{\lambda\mathcal{S}_k}$.

Proposition 3. *From $u = \min_{u\in D_k} \sum_{i=1}^{n} \phi(x_i, \lambda + u_i)$ with $\lambda > 0$, we can compute $w = \text{prox}_{\lambda\mathcal{S}_k}(x)$ defined as:*

$$\forall\ i = 1, \dots, n\ w_i = \frac{x_i u_i}{\lambda + u_i}\ . \tag{5}$$

Proof. Using definition 2 followed by proposition 2,

$$\begin{aligned} w &= \arg\min_{z\in\mathbb{R}^n}\left\{\lambda\mathcal{S}_k(z) + \frac{1}{2}\|z-x\|_2^2\right\} = \arg\min_{z\in\mathbb{R}^n}\left\{\frac{\lambda}{2}\min_{u\in D_k}\sum_{i=1}^{n}\phi(z_i,u_i) + \frac{1}{2}\|z-x\|_2^2\right\} \\ &= \arg\min_{z\in\mathbb{R}^n}\min_{u\in D_k}\underbrace{\left\{\frac{\lambda}{2}\sum_{i=1}^{n}\phi(z_i,u_i) + \frac{1}{2}\|z-x\|_2^2\right\}}_{\varphi}\ . \end{aligned}$$

From there, we need to solve for z the inner minimization. We note $\hat{z}$ the value at the optimum. Suppose that $u_i > 0$ (otherwise $u_i = 0 = \hat{z}_i$), then from the first order condition for all $i_0 = 1, \dots, n$:

$$\frac{\partial \varphi}{\partial z_{i_0}}(x, \hat{z}, u) = 0 \iff \frac{\lambda}{2}\frac{2\hat{z}_{i_0}}{u_{i_0}} + \frac{2}{2}(\hat{z}_{i_0} - u_{i_0})$$
$$\iff \hat{z}_{i_0} = \frac{u_{i_0} z_{i_0}}{\lambda + u_{i_0}} \quad .$$

And plugging-in this result, it is easy to show that $\varphi(x, \hat{z}, u) = \frac{\lambda}{2}\sum_{i=1}^{n} \phi(x_i, \lambda + u_i)$. □

3.2 Algorithms for the sparse envelope

To find the root of a function defined as the sum of piecewise linear functions with a single breakpoint, we can use the *Random Search* algorithm.

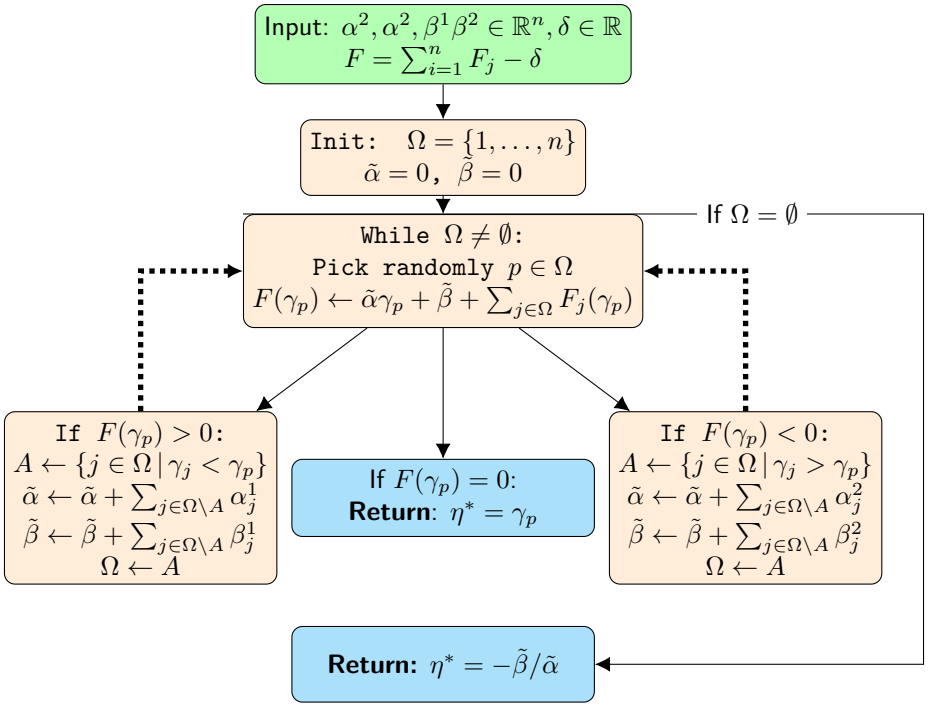

Figure 2: Diagram for the random search algorithm procedure.

Remark 1. *Note that this algorithm exploits the fact that the problem was reduced from finding a vector of dimension n to finding the root of a function in one dimension. Besides, it is efficient with sparse vectors because of the choice of the F_j functions in the sum. In the worse case, we find ourselves with a complexity of $\mathcal{O}(n)$ that does not depend on the size of an initial interval like the bisection method.*

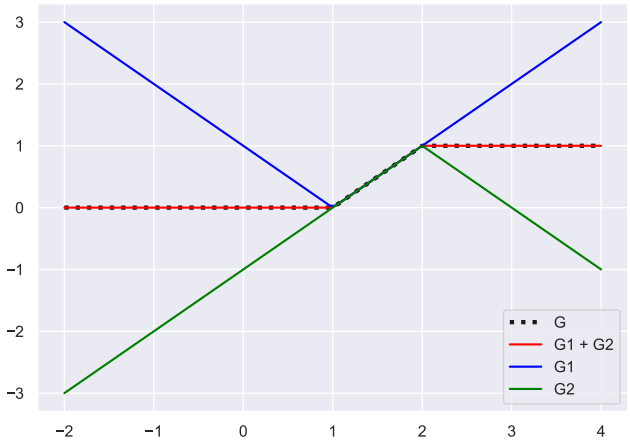

Figure 3: Decomposition of a piecewise linear signal with two breakpoints into the sum of two piecewise learn signals with a single breakpoint each.

3.3 Numerical application

Let's finally compare our regularized problem with the **Elastic net** regularization in an highly correlated situation (the same represented in Figure 1). We simulate a signal with only 3's for the first fifteen components and zeros for the 25 left. The observed signal is

$$b = Ax + \varepsilon \ ,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise. The matrix A is constructed from the algorithm that follows so that we can recreate the grouping effect:

- generate three random vectors $z_1,\ z_2$ and $z_3 \in \mathbb{R}^n$,
- from z_i, compute $Z_i = \big[z_i, z_i, z_i, z_i, z_i\big] \in \mathbb{R}^{n\times 5}$,
- generate $B,\ C,\ D \in \mathbb{R}^{n\times 5}$ and $E \in \mathbb{R}^{n\times 25}$,
- $A = \big[Z_1 + 0.01W_1, Z_2 + 0.01W_2, Z_3 + 0.01W_3, E\big]$.

The sparse-envelope problem is written as:

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \mathcal{S}_k(x) \ ,$$

where λ must be chosen through a train/test procedure and we take $k = 15$ because we know its real value here. Otherwise we would have to search it as well.

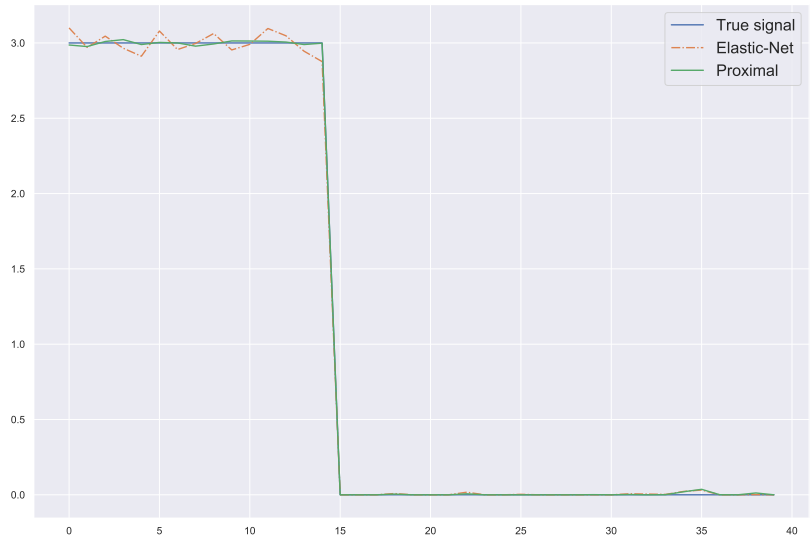

Figure 4: Comparison of the Elastic-net and sparse regularization in a highly correlated situation with $n = 50$ and $\sigma = 0.1$.

As the Figure 4 shows, the proximal envelope is closer to the original signal than the **Elastic net** in this case. We can now compare if this is also the case for some values of $n \in \mathbb{N}^*$ and quantify the gain of using this method instead of the **Elastic net** currently used.

To do so, we call x_{enet} the solution returned by the **Elastic net** method and x_{prox} the one returned by the method we described. The residuals are respectively $\|x - x_{enet}\|_2$ and $\|x - x_{prox}\|_2$ for both methods. Finally, we want to look at the improvement made by the latter against the former. So the percentage of improvement of the residuals is given by:

$$\mathrm{IM}(x_{prox} \,||\, x_{enet}) = \left(\frac{\|x - x_{prox}\|_2}{\|x - x_{enet}\|_2} - 1 \right) \times 100 \ .$$

To see different cases, we tried $n \in \{40, 80\}$ and $\sigma \in \{0.1, 1, 2\}$. We generated new data for A (and therefore b) $J = 50$ times for each experiment, the improvements percentages showed are defined as

$$\overline{\mathrm{IM}}_n(J) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{IM}(x_{prox}^{(j)} \,||\, x_{enet}^{(j)}) \ .$$

Remark 2. *A negative improvement in this case means that the error was reduced from the* ***Elastic net*** *to the sparse envelope method. If the sparse envelope method were perfect and resulted in* $x = x_{prox}$*, then* $\mathrm{IM}(x_{prox} \,||\, x_{enet}) = -100\%$.

In addition to the mean of the improvements, we provide the standard deviation needed for the confidence interval (each simulation is indeed independent of the previous ones and we generate the data with normal distributions of fixed parameters for a chosen σ).

n	σ	$\overline{\mathrm{IM}}_n(50)$	$\hat{\sigma}\left(\mathrm{IM}_n(50)\right)$
40	0.1	-98.55%	0.510
	1.0	-98.29%	0.543
	2.0	-97.75%	0.768
80	0.1	-99.09%	0.308
	1.0	-99.15%	0.361
	2.0	-98.79%	0.580

Table 1: Average improvement obtained comparing for the residuals of the inverse problem using the **Elastic net** method against the sparse envelope.

From Table 1 and Figure 4, we can conclude that in situations involving a grouping effect, the sparse envelope has a better accuracy than the **Elastic net** . However, the bigger the noise, the closer the two method get (especially on the part of the signal where the true signal equals zero).

Conclusion

In conclusion, the biregularization method.....

References

[1] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[2] Amir Beck and Yehonathan Refael. *Sparse Regularization via Bidualization*. 2019.

[3] Amir Beck and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202.

[4] Jürgen Kindler. "A Simple Proof of Sion's Minimax Theorem". In: *The American Mathematical Monthly* 112.4 (2005), pp. 356–358. ISSN: 00029890, 19300972. URL: `http://www.jstor.org/stable/30037472`.

[5] Ernest K Ryu and Wotao Yin. "Proximal-proximal-gradient method". In: *arXiv preprint arXiv:1708.06908* (2017).

[6] Joseph Salmon. *Advanced Linear Models HMMA307*. `http://josephsalmon.eu/HMMA307.html`. 2020.