Debiasing the Elastic Net for models with interactions

Florent Bascou, Sophie Lèbre, Joseph Salmon

Debiasing the Elastic Net for models with interactions

Florent Bascou[1,†] & Sophie Lèbre[1,2,‡] & Joseph Salmon[1,*]

[1] *IMAG, Univ. Montpellier, CNRS Montpellier, France*
[2] *Univ. Paul-Valéry-Montpellier 3, Montpellier, France*
† *florent.bascou@umontpellier.fr*, ‡ *sophie.lebre@umontpellier.fr*,
* *joseph.salmon@umontpellier.fr*

Résumé. Nous présentons un modèle de régression pénalisée et dé-biaisée pour l'estimation, en grande dimension, d'un modèle linéaire parcimonieux avec interactions. L'objectif est d'estimer conjointement le support et les coefficients dé-biaisés associés, en partant d'un estimateur de type Elastic Net. On utilise pour cela un algorithme de descente par coordonnée, qui ne nécessite pas de construire la matrice des interactions. Cette propriété est cruciale sur données réelles sachant que cette matrice peut facilement dépasser les capacités mémoires. Enfin, nous adaptons une méthode de dérivation automatique qui permet d'obtenir simultanément la solution des moindres carrés sur le support, sans avoir à résoudre a posteriori un problème de moindres carrés.

Mots-clés. Lasso, Elastic Net, Interactions, Dé-biasage, Descente par Coordonnée.

Abstract. We present a penalized and de-biased regression model to estimate, in high dimension, sparse linear models with interactions. Our aim is to jointly estimate the support and the associated de-biased coefficients, starting from an Elastic Net type estimator. The main idea is to use a coordinate descent algorithm, which does not require building the interaction matrix. This property is crucial on real data since the design matrix modeling interactions can quickly exceed memory capacities. In addition, we adapt an automatic differentiation method which allows to obtain simultaneously the least squares solution on the support, without having to solve, a posteriori, a least squares problem.

Keywords. Lasso, Elastic Net, Interaction, De-biasing, Coordinate Descent.

1 Introduction

Thanks to their interpretability, linear models are popular for many statistics tasks. Unfortunately, it turns out that the number of variables is frequently larger than the number of samples, so regularization is often required. Sparse regularization techniques leveraging the ℓ_1-norm have led to various popular estimators in the last two decades, including Lasso [Tibshirani, 1996] and Elastic Net [Zou and Hastie, 2005] among the most popular. When targeting feature interactions, such estimator become crucial: even when limited

to quadratic interactions, the number of variables is already (almost) squared, and the number of variables hence created can easily overload computers' memory.

Due to highly correlated variables, we estimate the coefficients using Elastic Net [Zou and Hastie, 2005], which allows to reduce the number of variables thanks to the ℓ_1 penalty, while taking into account the correlation thanks to the ℓ_2 penalty [Tikhonov, 1943, Hoerl and Kennard, 1970]. We adapt a coordinate descent algorithm (popularized by `glmnet` [Friedman et al., 2007, 2010]) so the interaction matrix does not need to be stored.

Finally, it is known that both Lasso and Elastic Net tend to be biased as they shrink large coefficients aggressively. To alleviate this issue, we suggest to compute a de-biased version of the coefficients along with the original coefficients [Deledalle et al., 2017]. We propose an algorithm approaching the LS Elastic Net (Elastic Net followed by a Least Squares step on the support), though in a more stable way as the naive implementation.

2 Elastic Net for interactions

2.1 Model and estimator

In the following, p is the number of features, n the number of samples, and $q = p\,(p+1)/2$ (or $p\,(p-1)/2$ depending whether we include or not the pure quadratic terms) the number of interaction features. The response vector is denoted by $y \in \mathbb{R}^n$. The Elastic Net model now reads :

$$(\mathcal{P}) \qquad \min_{\beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^q} \frac{1}{2n} \left\| y - X\beta - Z\Theta \right\|_2^2 + \alpha_1 \left\| \beta \right\|_1 + \alpha_2 \left\| \Theta \right\|_1 + \frac{\alpha_3}{2} \left\| \beta \right\|_2^2 + \frac{\alpha_4}{2} \left\| \Theta \right\|_2^2 \quad . \tag{1}$$

where $\alpha_1 > 0, \ldots, \alpha_4 > 0$ are tunning parameters. The parameters α_1 and α_2 control the level of ℓ_1 penalization (resp. for the quadratic features), and the sparsity of β and Θ, while α_3 and α_4 control the level of ℓ_2 penalization and how spread out the signal is among active features.

As previously explained, we can not always handle in memory the design matrix Z (Figure 2), which leads us to reformulate the classical coordinate descent algorithm in this context. Let us remind[1] the main step in the coordinate descent algorithm to solve the Elastic Net problem is the coordinates updates for β_{j_0} and $\Theta_{j_0'}$ (for $j_0 \in [\![1, p]\!]$ and $j_0' \in [\![1, q]\!]$). This requires solving one dimensional problems of the form:

$$\arg\min_{\beta_{j_0} \in \mathbb{R}} \frac{1}{2n} \Big(y - \sum_{j=1}^{p} \beta_j x_j - \sum_{j=1}^{q} \Theta_j z_j \Big)^2 + \alpha_1 |\beta_{j_0}| + \frac{\alpha_3}{2} \beta_{j_0}^2 \quad , \tag{2}$$

$$\arg\min_{\Theta_{j_0'} \in \mathbb{R}} \frac{1}{2n} \Big(y - \sum_{j=1}^{p} \beta_j x_j - \sum_{j=1}^{q} \Theta_j z_j \Big)^2 + \alpha_2 |\Theta_{j_0'}| + \frac{\alpha_4}{2} \Theta_{j_0'}^2 \quad . \tag{3}$$

[1] See [Friedman et al., 2010] for details.

Proposition 2.1. *We write $\widehat{\beta}^k$ and $\widehat{\Theta}^k$ for the coefficients computed at the k-th pass over the data by the coordinate descent algorithm, and $r^k = y - X\widehat{\beta}^k - Z\widehat{\Theta}^k$ is the associated residuals. The coordinate update rules for the j_0^{th} and $j_0'^{th}$ coordinate reads*

$$\widehat{\beta}_{j_0}^{k+1} = \frac{1}{\|x_{j_0}\|^2 + n\alpha_3} \mathrm{ST}\left(x_{j_0}^\top \left(r^k + \widehat{\beta}_{j_0}^k x_{j_0}\right), n\alpha_1\right) \quad . \tag{4}$$

$$\widehat{\Theta}_{j_0'}^{k+1} = \frac{1}{\|z_{j_0'}\|^2 + n\alpha_4} \mathrm{ST}\left(z_{j_0'}^\top \left(r^k + \widehat{\Theta}_{j_0'}^k z_{j_0'}\right), n\alpha_2\right) \quad . \tag{5}$$

and ST *representing the soft-thresholding, defined for any $x \in \mathbb{R}$ by:*

$$\mathrm{ST}(x, \alpha) = (|x| - \alpha)_+ \operatorname{sign}(x) \quad . \tag{6}$$

In the previous proposition, we need to compute the $z_{j_0'}$ column of Z, which is made possible by a coordinate descent algorithm (Line 6 of Algorithm 2). Thanks to that, we can handle interactions without explicitly storing the interaction (design) matrix.

2.2 De-biasing

Unfortunately, the Lasso and the Elastic Net coefficients are biased (see [Salmon, 2017]): large coefficients are shrunk toward zero. To reduce this effect, one can perform an Least Squares step on the non-zero coefficients obtain by Elastic Net (Naive-LSEnet). Yet, this approach is limited: the interaction design matrix on the support is needed, which for large datasets could not be stored.

2.2.1 Sequentially de-biasing Elastic Net

Algorithm 1: Naive-LSEnet

Input : $[X, Z]$, y, α

1 $\widehat{\beta}, \widehat{\Theta} \leftarrow$ `Enet`([X, Z], y, α)
`// supp. estimat.`

2 $\mathrm{supp}_{\widehat{\beta}} \leftarrow$ where $(\widehat{\beta} \neq 0)$

3 $\mathrm{supp}_{\widehat{\Theta}} \leftarrow$ where$(\widehat{\Theta} \neq 0)$

4 $\widetilde{\beta}, \widetilde{\Theta} \leftarrow$ `LS`([X, Z], y, $\mathrm{supp}_{\widehat{\beta}}$, $\mathrm{supp}_{\widehat{\Theta}}$)

Output : $\widetilde{\beta}, \widetilde{\Theta}$

To cope with interactions, we instantiate Covariant LEAst-square Refitting (CLEAR) [Deledalle et al., 2017], a framework to simultaneously debias the coefficients along with the algorithm (here, coordinate descent) computing the Elastic Net solution.

Proposition 2.2. *Let us suppose that the coefficients $\widehat{\beta}^k$ and $\widehat{\Theta}^k$ are iteratively updated according to Equation (4) and Equation (5). We define the Jacobian of $\widehat{\beta}^k$ (resp. $\widehat{\Theta}^k$) applied to the residuals as $J_{\widehat{\beta}_j^{k+1} r^k}$ (resp. $J_{\widehat{\Theta}_{jj}^{k+1} r^k}$) and e_j the canonical basis vector :*

$$J_{\widehat{\beta}_j^{k+1} r^{k+1}} = \frac{\left(e_j \|x_j\|_2^2 - X^\top x_j\right)^\top J_{\widehat{\beta}^k} r^k - \left(x_j^\top Z\right)^\top J_{\widehat{\Theta}^k} r^k + x_j^\top r^k}{\|x_j\|^2 + n\alpha_3} \mathbb{1}_{\left\{\left|x_j^\top \left(r^k + \widehat{\beta}_j^k x_j\right)\right| \geqslant n\alpha_1\right\}}$$

$$J_{\widehat{\Theta}_{jj}^{k+1} r^{k+1}} = \frac{\left(e_{jj} \|z_{jj}\|_2^2 - Z^\top z_{jj}\right)^\top J_{\widehat{\Theta}^k} r^k + \left(X^\top z_{jj}\right)^\top J_{\widehat{\beta}^k} r^k + z_{jj}^\top r^k}{\|z_{jj}\|^2 + n\alpha_4} \mathbb{1}_{\left\{\left|z_{jj}^\top \left(r^k + \widehat{\Theta}_{jj}^k z_{jj}\right)\right| \geqslant n\alpha_2\right\}}$$

These updates leads to compute $\rho^{k+1} = \frac{\left\langle [X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top ; r^{k+1} \right\rangle}{\left\| [X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top \right\|_2^2}$.
Considering the problem Equation (1), the CLEAR approach reads :

$$\widetilde{\beta}^{k+1} = \widehat{\beta}^{k+1} + \rho^{k+1} J_{\widehat{\beta}^{k+1}} r^{k+1} \ , \tag{7}$$

$$\widetilde{\Theta}^{k+1} = \widehat{\Theta}^{k+1} + \rho^{k+1} J_{\widehat{\Theta}^{k+1}} r^{k+1} \ . \tag{8}$$

This leads to Algorithm 2 where Lines 4, 8 and 10 are evaluated on the fly without Z being built. We call this method CLEAR Least Squares Elastic Net (CLEAR-Enet). Setting $\rho = 1$ in Lines 11 and 12 recovers the Ridge estimator associated with the Elastic Net support, instead of a Least Squares version, as in Theorem 2.2.

Algorithm 2: Coordinate Descent Epoch for CLEAR-Enet

input : $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha = (\alpha_1, \dots, \alpha_4)^\top$, ...
param. : $\widehat{\beta}(= 0_p)$, $\widehat{\Theta}(= 0_q)$, $J_{\widehat{\beta}} r(= 0_p)$, $J_{\widehat{\Theta}} r(= 0_q)$

1 $jj = 0$; $q = p(p+1)/2$ or $p(p-1)/2$
2 **for** $j_1 = 1, \dots, p$ **do**
3 $\quad \widehat{\beta}_{j_1}^{k+1} = \frac{1}{\|x_{j_1}\|^2 + n\alpha_3} \mathrm{ST}(x_{j_1}^\top (y - r^k + \widehat{\beta}_{j_1}^k x_{j_1}), n\alpha_1)$ // β Elastic Net update
4 $\quad J_{\widehat{\beta}_{j_1}^{k+1}} r^{k+1} = \frac{\left(e_{j_1} \|x_{j_1}\|_2^2 - X^\top x_{j_1}\right)^\top J_{\widehat{\beta}^k} r^k - \left(x_{j_1}^\top Z\right)^\top J_{\widehat{\Theta}^k} r^k + x_{j_1}^\top r^k}{\|x_{j_1}\|^2 + n\alpha_3} \mathbb{1}_{\{|x_{j_1}^\top (r^k + \widehat{\beta}_{j_1}^k x_{j_1})| \geqslant n\alpha_1\}}$
5 $\quad$ **for** $j_2 = 1, \dots, q$ **do**
6 $\quad\quad z_{jj} = x_{j_1} \odot x_{j_2}$ // point-wise multiplication
7 $\quad\quad \widehat{\Theta}_{jj}^{k+1} = \frac{1}{\|z_{jj}\|^2 + n\alpha_4} \mathrm{ST}(z_{jj}^\top (y - r^k + \widehat{\Theta}_{jj}^k z_{jj}), n\alpha_2)$ // Θ Elastic Net update
8 $\quad\quad J_{\widehat{\Theta}_{jj}^{k+1}} r^{k+1} = \frac{\left(e_{jj} \|z_{jj}\|_2^2 - Z^\top z_{jj}\right)^\top J_{\widehat{\Theta}^k} r^k + \left(X^\top z_{jj}\right)^\top J_{\widehat{\beta}^k} r^k + z_{jj}^\top r^k}{\|z_{jj}\|^2 + n\alpha_4} \mathbb{1}_{\{|z_{jj}^\top (r^k + \widehat{\Theta}_{jj}^k z_{jj})| \geqslant n\alpha_2\}}$
9 $\quad\quad jj \mathrel{+}= 1$
10 $\rho^{k+1} = \frac{\langle [X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top \mid r^{k+1} \rangle}{\|[X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top\|_2^2}$
11 $\widetilde{\beta}^{k+1} = \widehat{\beta}^{k+1} + \rho^{k+1} J_{\widehat{\beta}^{k+1}} r^{k+1}$ // β CLEAR-Enet update
12 $\widetilde{\Theta}^{k+1} = \widehat{\Theta}^{k+1} + \rho^{k+1} J_{\widehat{\Theta}^{k+1}} r^{k+1}$ // Θ CLEAR-Enet update

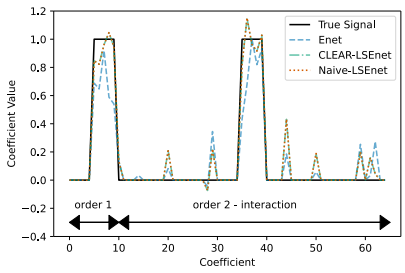
output : $\widehat{\beta}^{k+1}, \widehat{\Theta}^{k+1}\ \widetilde{\beta}^{k+1}, \widetilde{\Theta}^{k+1}$

3 Numerical experiments

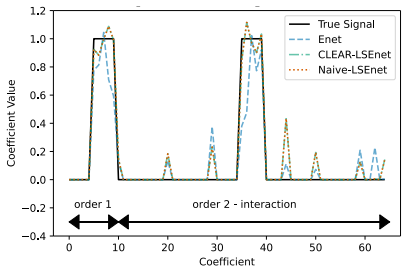
For the Naive-LSEnet, we use the Least Squares solver from `sklearn` [Pedregosa et al., 2011] on the support obtained by Elastic Net. For Figures 1 and 2, we use duality gap as stopping criterion, fixed at 10^{-4} and we set $\alpha_3 = \alpha_4 = 0.001$.

3.1 Artificial datasets

To compare Elastic Net, Naive-LSEnet and CLEAR-Enet, we build an artificial dataset, for which X of size $(n, p) = (60, 10)$ is drawn according to a standard Gaussian distribu-



(a) CV : $\alpha_1 = \alpha_2 \approx 0.21581$.



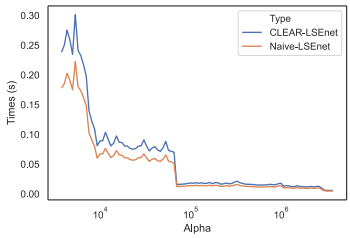
(b) CV (2D grid) : $\alpha_1 \approx 0.159$, $\alpha_2 \approx 0.267$.

Figure 1: Comparison between Elastic Net, CLEAR-Enet and Naive-LSEnet, with cross-validation (condensed CV) on the CLEAR-Enet result.

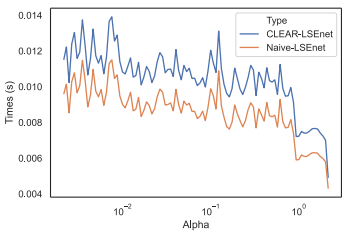
tion, and we set X such that the column x_1, x_2 and x_3 are correlated (we draw x_3 from a Gaussian distribution adding $\frac{1}{2}(x_1 + x_2)$). We include the pure quadratic features leading to 55 interactions features. The true underlying signal has only five non-zero coefficients for the β and five more non-zero coefficients for Θ. Finally, we draw the noise ε from a Gaussian distribution with zero mean and a variance $1/2$. Hence, the response vector $y \in \mathbb{R}^n$ is : $y = X\beta + Z\Theta + \varepsilon$.

In Figure 1, we observe that both CLEAR-Enet and Naive-LSEnet estimator recover better coefficients than the Elastic Net. Indeed, both yield coefficients from the true signal than the Elastic Net on the true support. Outside the true support, CLEAR-Enet and Naive-LSEnet tends to give a larger coefficients than Elastic Net.

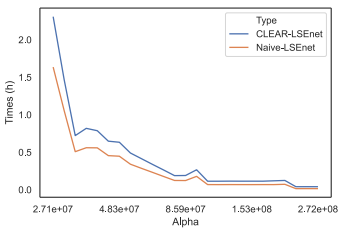
3.2 Real datasets



(a) Boston dataset
size of Z : 368, 4 Kb



(b) Diabetes dataset
size of Z : 194, 5 Kb



(c) Leukemia dataset
size of Z : 14,68 **Gb**

Figure 2: Mean time comparisons : Naive-LSEnet and CLEAR-Enet on real-datasets from `sklearn`. Here, α_1 and α_2 are equal.

We see Figure 2, that CLEAR-Enet is the same order than Naive-LSEnet. We notice that we can do Naive-LSEnet on Leukemia datasets because for those α_1 and α_2, the support is small, so we can build Z on the support. We must note that Elastic Net from

`sklearn` can handle interactions, but it requires to create and store Z, which is not always feasible. For instance with the Leukemia dataset, Z is almost 14Gb, (and possibly does not fit in memory), whereas our method can handle interactions easily here.

4 Conclusion

We presented a penalized and de-biased regression model able handle quadratic interactions in high-dimension. Future work include sensitivity analysis of the tuning parameters and algorithmic speed up, *e.g.,* following the work by Le Morvan and Vert [2018].

References

C.-A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter. CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration. *SIAM J. Imaging Sci.*, 10(1): 243–284, 2017.

J. Friedman, T. J. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

J. Friedman, T. J. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

M. Le Morvan and J.-P. Vert. Whinter: A working set algorithm for high-dimensional sparse second order interaction models. In *ICML*, pages 3632–3641, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

J. Salmon. *On high dimensional regression: computational and statistical perspectives.* Habilitation à diriger des recherches, ENS Paris-Saclay, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39: 176–179, 1943.

H. Zou and T. J. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.