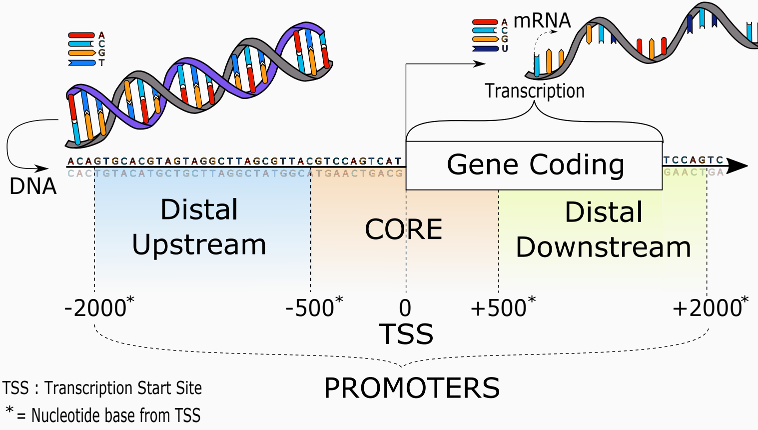
ElasticNet avec gestion des interactions et débiaisage

Florent Bascou[1]
Sophie Lèbre[1,2], Joseph Salmon[1]

January 28, 2021

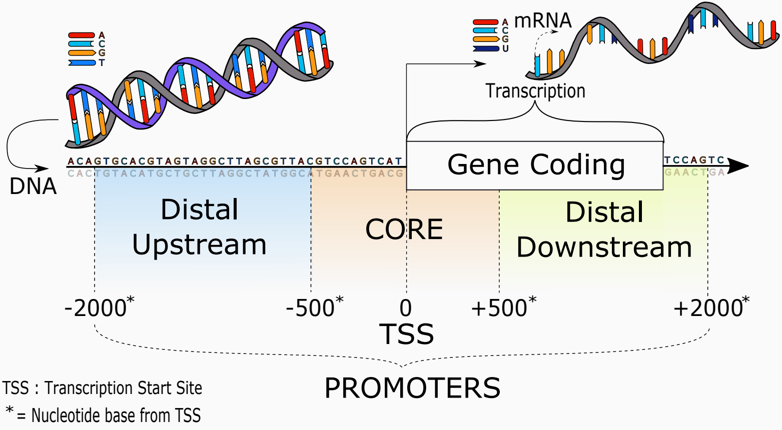
[1]IMAG, Univ. Montpellier, CNRS Montpellier, France
[2]Univ. Paul-Valéry-Montpellier 3, Montpellier, France

Motivation



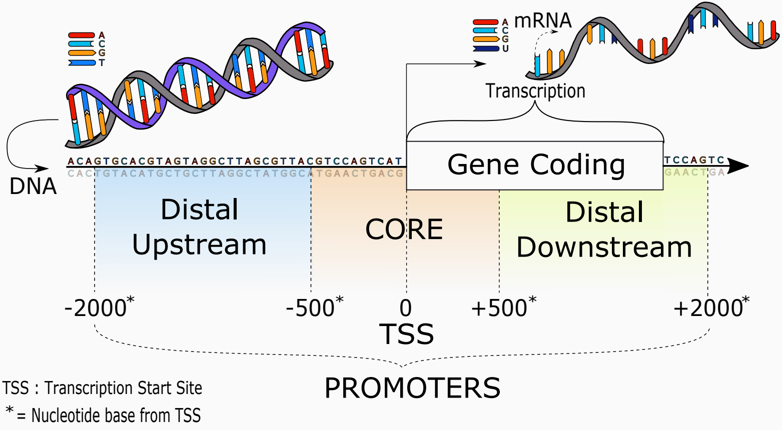
- linear model[1] to explain mRNA count (y) from DNA sequence resume (X)

[1] Chloé Bessière et al. (Jan. 2018). "Probing instructions for expression regulation in gene nucleotide compositions". In: *PLOS Computational Biology* 14.1, pp. 1–28.



- linear model[1] to explain mRNA count (y) from DNA sequence resume (X)
- X : frequency of 1-letter and 2-letter words (AA, AC, AG, . . .)

[1] Chloé Bessière et al. (Jan. 2018). "Probing instructions for expression regulation in gene nucleotide compositions". In: *PLOS Computational Biology* 14.1, pp. 1–28.



- linear model[1] to explain mRNA count (y) from DNA sequence resume (X)
- X : frequency of 1-letter and 2-letter words (AA, AC, AG, . . .)
- Add interaction between variables to improve the model

[1] Chloé Bessière et al. (Jan. 2018). "Probing instructions for expression regulation in gene nucleotide compositions". In: *PLOS Computational Biology* 14.1, pp. 1–28.

Model: sparse linear regression model with interactions :

$$y = X\beta + Z\Theta + \text{ error}$$

where :

- $y \in \mathbb{R}^n$ (response vector): mRNA log transformed counts ;
- $X \in \mathbb{R}^{n\times p}$ design matrix of main variables :

$$X = [x_1, x_2, \ldots, x_p]$$

- $Z \in \mathbb{R}^{n\times q}$ design matrix of interaction variables, $q = \frac{p(p+1)}{2}$:

$$Z = [x_1 \odot x_1, \ldots, x_1 \odot x_p, x_2 \odot x_2, \ldots, x_{p-1} \odot x_p, x_p \odot x_p]$$

Model: sparse linear regression model with interactions :

$$y = X\beta + Z\Theta + \text{ error}$$

where :

- $y \in \mathbb{R}^n$ (response vector): mRNA log transformed counts ;
- $X \in \mathbb{R}^{n \times p}$ design matrix of main variables :

$$X = [x_1, x_2, \ldots, x_p]$$

- $Z \in \mathbb{R}^{n \times q}$ design matrix of interaction variables, $q = \frac{p(p+1)}{2}$:

$$Z = [x_1 \odot x_1, \ldots, x_1 \odot x_p, x_2 \odot x_2, \ldots, x_{p-1} \odot x_p, x_p \odot x_p]$$

- $\beta \in \mathbb{R}^p$ and $\Theta \in \mathbb{R}^q$ sparse : few active features

Model: sparse linear regression model with interactions :

$$y = X\beta + Z\Theta + \text{ error}$$

where :

- $y \in \mathbb{R}^n$ (response vector): mRNA log transformed counts ;
- $X \in \mathbb{R}^{n\times p}$ design matrix of main variables :

$$X = [x_1, x_2, \ldots, x_p]$$

- $Z \in \mathbb{R}^{n\times q}$ design matrix of interaction variables, $q = \frac{p(p+1)}{2}$:

$$Z = [x_1 \odot x_1, \ldots, x_1 \odot x_p, x_2 \odot x_2, \ldots, x_{p-1} \odot x_p, x_p \odot x_p]$$

- $\beta \in \mathbb{R}^p$ and $\Theta \in \mathbb{R}^q$ sparse : few active features

$$\text{Here :} \begin{cases} n \approx 20\ 000 : \text{genes number} \\ p \approx 500 : \text{number of main features} \\ q \approx 140\ 000 : \text{number of quadratic features} \end{cases}$$

ElasticNet with interactions

The usual ElasticNet[2] (shortly Enet) estimator is defined as :

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{\|y - X\beta\|^2}{2n} + \underbrace{\alpha_{1,1}\|\beta\|_1}_{\ell_1 \text{ penalization}} + \underbrace{\alpha_{2,1}\frac{\|\beta\|^2}{2}}_{\ell_2 \text{ penalization}}$$

1. ℓ_1 penalization to enforce sparsity[3]
2. adding ℓ_2 penalization to detect correlated features

[2] H. Zou and T. J. Hastie (2005). "Regularization and variable selection via the elastic net". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2, pp. 301–320.

[3] R. Tibshirani (1996). "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1, pp. 267–288.

[4] J. Friedman et al. (2007). "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2, pp. 302–332.

The usual ElasticNet[2] (shortly Enet) estimator is defined as :

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{\|y - X\beta\|^2}{2n} + \underbrace{\alpha_{1,1}\|\beta\|_1}_{\ell_1 \text{ penalization}} + \underbrace{\alpha_{2,1}\frac{\|\beta\|^2}{2}}_{\ell_2 \text{ penalization}}$$

1. ℓ_1 penalization to enforce sparsity[3]
2. adding ℓ_2 penalization to detect correlated features

The ElasticNet with interactions estimator is defined as :

$$(\widehat{\beta}, \widehat{\Theta}) \in \underset{\beta \in \mathbb{R}^p,\ \Theta \in \mathbb{R}^q}{\arg\min} \frac{\|y - X\beta - Z\Theta\|^2}{2n} + \underbrace{\alpha_{1,1}\|\beta\|_1 + \alpha_{1,2}\|\Theta\|_1}_{\ell_1 \text{ penalization}} + \underbrace{\frac{\alpha_{2,1}}{2}\|\beta\|^2 + \frac{\alpha_{2,2}}{2}\|\Theta\|^2}_{\ell_2 \text{ penalization}}$$

[2] H. Zou and T. J. Hastie (2005). "Regularization and variable selection via the elastic net". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2, pp. 301–320.

[3] R. Tibshirani (1996). "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1, pp. 267–288.

[4] J. Friedman et al. (2007). "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2, pp. 302–332.

The usual ElasticNet[2] (shortly Enet) estimator is defined as :

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{\|y - X\beta\|^2}{2n} + \underbrace{\alpha_{1,1}\|\beta\|_1}_{\ell_1 \text{ penalization}} + \underbrace{\alpha_{2,1}\frac{\|\beta\|^2}{2}}_{\ell_2 \text{ penalization}}$$

1. ℓ_1 penalization to enforce sparsity[3]
2. adding ℓ_2 penalization to detect correlated features

The ElasticNet with interactions estimator is defined as :

$$(\widehat{\beta}, \widehat{\Theta}) \in \underset{\beta \in \mathbb{R}^p,\ \Theta \in \mathbb{R}^q}{\arg\min} \frac{\|y - X\beta - Z\Theta\|^2}{2n} + \underbrace{\alpha_{1,1}\|\beta\|_1 + \alpha_{1,2}\|\Theta\|_1}_{\ell_1 \text{ penalization}} + \underbrace{\frac{\alpha_{2,1}}{2}\|\beta\|^2 + \frac{\alpha_{2,2}}{2}\|\Theta\|^2}_{\ell_2 \text{ penalization}}$$

Q: How to solve it ? A: Coordinate Descent[4] algorithm

[2] H. Zou and T. J. Hastie (2005). "Regularization and variable selection via the elastic net". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2, pp. 301–320.

[3] R. Tibshirani (1996). "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1, pp. 267–288.

[4] J. Friedman et al. (2007). "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2, pp. 302–332.

Algorithm 1: Coordinate Descent for ElasticNet with interactions (1 epoch)

input : $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha = (\alpha_{1,1}, \ldots, \alpha_{2,2})^\top$, $\bar{x}$, $\bar{z}$, $\mathrm{std}(x)$, $\mathrm{std}(z)$ ...

param. : $\widehat{\beta}(= 0_p)$, $\widehat{\Theta}(= 0_q)$

$jj = 0$;

Algorithm 1: Coordinate Descent for ElasticNet with interactions (1 epoch)

input : $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha = (\alpha_{1,1}, \ldots, \alpha_{2,2})^\top$, $\bar{x}$, $\bar{z}$, $\mathrm{std}(x)$, $\mathrm{std}(z)$ ...

param. : $\widehat{\beta}(= 0_p)$, $\widehat{\Theta}(= 0_q)$

$jj = 0$;

for $j_1 = 1, \ldots, p$ **do**

$\quad x = \frac{x_{j_1} - \bar{x}_{j_1}}{\mathrm{std}(x_{j_1})}$

$\quad \widehat{\beta}_{j_1}^{k+1} = \frac{1}{\|x\|^2 + n\alpha_{2,1}} \mathrm{ST}(x^\top(r^k + \widehat{\beta}_{j_1}^k x), n\alpha_{1,1})$; // update β_{j_1}

with ST the Soft-Thresholding operator : $x \in \mathbb{R} : \mathrm{ST}(x, \alpha) = (|x| - \alpha)_+ \,\mathrm{sign}(x)$

Algorithm 1: Coordinate Descent for ElasticNet with interactions (1 epoch)

input : $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha = (\alpha_{1,1}, \dots, \alpha_{2,2})^\top$, $\bar{x}$, $\bar{z}$, $\mathrm{std}(x)$, $\mathrm{std}(z)$ ...

param. : $\widehat{\beta}(= 0_p)$, $\widehat{\Theta}(= 0_q)$

$jj = 0$;

for $j_1 = 1, \dots, p$ **do**

$\quad x = \frac{x_{j_1} - \bar{x}_{j_1}}{\mathrm{std}(x_{j_1})}$

$\quad \widehat{\beta}_{j_1}^{k+1} = \frac{1}{\|x\|^2 + n\alpha_{2,1}} \mathrm{ST}(x^\top (r^k + \widehat{\beta}_{j_1}^k x), n\alpha_{1,1})$; `// update` β_{j_1}

$\quad$ **for** $j_2 = j_1, \dots, p$ **do**

$\quad\quad z_{jj} = \frac{(x_{j_1} \odot x_{j_2}) - \bar{z}_{jj}}{\mathrm{std}(z_{jj})}$; `// creation of` z_{jj}

with ST the Soft-Thresholding operator : $x \in \mathbb{R}$: $\mathrm{ST}(x, \alpha) = (|x| - \alpha)_+ \,\mathrm{sign}(x)$

<u>Advantage</u>: no need to build Z

Algorithm 1: Coordinate Descent for ElasticNet with interactions (1 epoch)

input : $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha = (\alpha_{1,1}, \dots, \alpha_{2,2})^\top$, $\bar{x}$, $\bar{z}$, $\mathrm{std}(x)$, $\mathrm{std}(z)$ ...

param. : $\widehat{\beta}(= 0_p)$, $\widehat{\Theta}(= 0_q)$

$jj = 0$;

for $j_1 = 1, \dots, p$ **do**

$\quad x = \frac{x_{j_1} - \bar{x}_{j_1}}{\mathrm{std}(x_{j_1})}$

$\quad \widehat{\beta}_{j_1}^{k+1} = \frac{1}{\|x\|^2 + n\alpha_{2,1}} \mathrm{ST}(x^\top(r^k + \widehat{\beta}_{j_1}^{k} x), n\alpha_{1,1})$; // update β_{j_1}

$\quad$ **for** $j_2 = j_1, \dots, p$ **do**

$\quad\quad z_{jj} = \frac{(x_{j_1} \odot x_{j_2}) - \bar{z}_{jj}}{\mathrm{std}(z_{jj})}$; // creation of z_{jj}

$\quad\quad \widehat{\Theta}_{jj}^{k+1} = \frac{1}{\|z_{jj}\|^2 + n\alpha_{2,2}} \mathrm{ST}(z_{jj}^\top(r^k + \widehat{\Theta}_{jj}^{k} z_{jj}), n\alpha_{1,2})$; // update Θ_{jj}

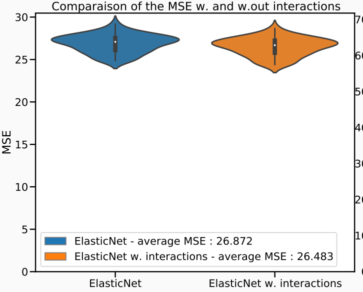
$\quad\quad jj \mathrel{+}= 1$

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

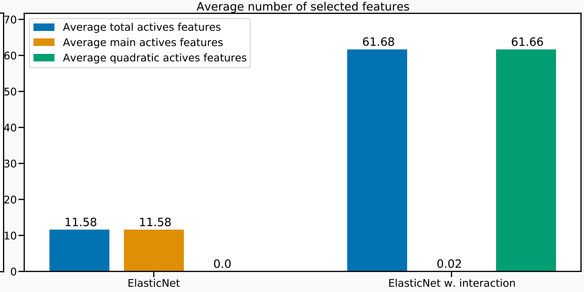
with ST the Soft-Thresholding operator : $x \in \mathbb{R} : \mathrm{ST}(x, \alpha) = (|x| - \alpha)_+ \, \mathrm{sign}(x)$

Advantage: no need to build Z

First real data result[5] : only CORE promoter part



(a) MSE error



(b) Distribution of actives features

[5] Data and parameters : $n = 19393$, $p = 20$, $q = 210$, 50 repetitions of 5-folds CV, `l1_ratio` $\in \{1, 0.99, 0.95, 0.9\}$, duality gap fix to 10^{-4}, $\alpha_{\min} = \frac{\alpha_{\max}}{1000}$ and X, Z standardize

1. Pros
    - Handle quadratic interactions, whatever the datasets
    - Keep ElasticNet benefits taking favoring sparsity and spread signal among correlated features ;

1. Pros
   - Handle quadratic interactions, whatever the datasets
   - Keep ElasticNet benefits taking favoring sparsity and spread signal among correlated features ;
2. Cons
   - ElasticNet shrinks large coefficients toward 0.

Debiasing the ElasticNet

A simple remedy : Naive LS ElasticNet

[6] C.-A. Deledalle et al. (2017). "CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration". In: *SIAM J. Imaging Sci.* 10.1, pp. 243–284.

A simple remedy : Naive LS ElasticNet

1. Fit an ElasticNet to find the set of the actives features.

[6] C.-A. Deledalle et al. (2017). "CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration". In: *SIAM J. Imaging Sci.* 10.1, pp. 243–284.

A simple remedy : Naive LS ElasticNet

1. Fit an ElasticNet to find the set of the actives features.
2. Fit a LeastSquares on the set of the actives features:

$$\hat{\beta}_{\alpha}^{\mathrm{LSEnet}}, \hat{\Theta}_{\alpha}^{\mathrm{LSEnet}} := \underset{\substack{\beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^q \\ \mathrm{supp}(\beta) = \mathrm{supp}(\beta^{\mathrm{Enet}}_{\alpha_{1,2}, \alpha_{2,2}}) \\ \mathrm{supp}(\Theta) = \mathrm{supp}(\Theta^{\mathrm{Enet}}_{\alpha_{1,1}, \alpha_{2,1}})}}{\arg\min} \frac{1}{2n} \|y - X\beta - Z\Theta\|_2^2$$

[6] C.-A. Deledalle et al. (2017). "CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration". In: *SIAM J. Imaging Sci.* 10.1, pp. 243–284.

A simple remedy : Naive LS ElasticNet

1. Fit an ElasticNet to find the set of the actives features.
2. Fit a LeastSquares on the set of the actives features:

$$\hat{\beta}_{\alpha}^{\text{LSEnet}}, \hat{\Theta}_{\alpha}^{\text{LSEnet}} := \underset{\substack{\beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^q \\ \operatorname{supp}(\beta) = \operatorname{supp}(\beta^{\text{Enet}}_{\alpha_{1,2}, \alpha_{2,2}}) \\ \operatorname{supp}(\Theta) = \operatorname{supp}(\Theta^{\text{Enet}}_{\alpha_{1,1}, \alpha_{2,1}})}}{\arg\min} \frac{1}{2n} \|y - X\beta - Z\Theta\|_2^2$$

Problem of Naive LS ElasticNet (w. interactions) :

1. Need LeastSquares solver with interactions for large datasets ;
2. Need to make cross validation on the whole pipeline ;

[6] C.-A. Deledalle et al. (2017). "CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration". In: *SIAM J. Imaging Sci.* 10.1, pp. 243–284.

A simple remedy : Naive LS ElasticNet

1. Fit an ElasticNet to find the set of the actives features.
2. Fit a LeastSquares on the set of the actives features:

$$\hat{\beta}_{\alpha}^{\text{LSEnet}}, \hat{\Theta}_{\alpha}^{\text{LSEnet}} := \underset{\substack{\beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^q \\ \operatorname{supp}(\beta) = \operatorname{supp}(\beta_{\alpha_{1,2},\alpha_{2,2}}^{\text{Enet}}) \\ \operatorname{supp}(\Theta) = \operatorname{supp}(\Theta_{\alpha_{1,1},\alpha_{2,1}}^{\text{Enet}})}}{\arg\min} \frac{1}{2n} \|y - X\beta - Z\Theta\|_2^2$$

Problem of Naive LS ElasticNet (w. interactions) :

1. Need LeastSquares solver with interactions for large datasets ;
2. Need to make cross validation on the whole pipeline ;

We want to debiaise on the fly without build Z : CLEAR[6] .

[6] C.-A. Deledalle et al. (2017). "CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration". In: *SIAM J. Imaging Sci.* 10.1, pp. 243–284.

CLEAR estimator is associated to an differentiable estimator $\mathbb{R}^{n} \ni y \to \widehat{\beta}(y) \in \mathbb{R}^{p}$ is, for all $y \in \mathbb{R}^{n}$, given by :

$$\mathcal{R}_{\widehat{\beta}}(y) := \widehat{\beta}(y) + \rho J \cdot (y - X\widehat{\beta}(y)) \text{ with } \rho := \begin{cases} \frac{\langle XJ\delta | \delta \rangle}{\|XJ\delta\|^{2}} & , \text{ if } XJ\delta \neq 0, \\ 1 & , \text{ otherwise,} \end{cases}$$

where $\delta = y - X\widehat{\beta}(y)$, $J = J_{\widehat{\beta}}(y) \in \mathbb{R}^{p \times n}$ is the **Jacobian** of $\widehat{\beta}(y)$.

CLEAR estimator is associated to an differentiable estimator $\mathbb{R}^n \ni y \to \widehat{\beta}(y) \in \mathbb{R}^p$ is, for all $y \in \mathbb{R}^n$, given by :

$$\mathcal{R}_{\widehat{\beta}}(y) := \widehat{\beta}(y) + \rho J \cdot (y - X\widehat{\beta}(y)) \text{ with } \rho := \begin{cases} \frac{\langle XJ\delta | \delta \rangle}{\|XJ\delta\|^2} & , \text{ if } XJ\delta \neq 0, \\ 1 & , \text{ otherwise,} \end{cases}$$

where $\delta = y - X\widehat{\beta}(y)$, $J = J_{\widehat{\beta}}(y) \in \mathbb{R}^{p \times n}$ is the **Jacobian** of $\widehat{\beta}(y)$.

Rem : we adapt an automatic differentiation scheme to compute J :

CLEAR estimator is associated to an differentiable estimator $\mathbb{R}^n \ni y \to \widehat{\beta}(y) \in \mathbb{R}^p$ is, for all $y \in \mathbb{R}^n$, given by :

$$\mathcal{R}_{\widehat{\beta}}(y) := \widehat{\beta}(y) + \rho J \cdot (y - X\widehat{\beta}(y)) \text{ with } \rho := \begin{cases} \frac{\langle XJ\delta | \delta \rangle}{\|XJ\delta\|^2} & \text{, if } XJ\delta \neq 0, \\ 1 & \text{, otherwise,} \end{cases}$$

where $\delta = y - X\widehat{\beta}(y)$, $J = J_{\widehat{\beta}}(y) \in \mathbb{R}^{p \times n}$ is the **Jacobian** of $\widehat{\beta}(y)$.

Rem : we adapt an automatic differentiation scheme to compute J :

$$(J_{\widehat{\beta}^{k+1}} r^{k+1})_j = \frac{x_j^\top \left((x_j e_j^\top - X) J_{\widehat{\beta}^k} + (\mathrm{Id}_n - Z J_{\widehat{\ominus}^k}) \right) r^k}{\|x_j\|^2 + n\alpha_{2,1}} \mathbb{1}_{\left\{ \left| x_j^\top \left(r^k + \widehat{\beta}_j^k x_j \right) \right| \geqslant n\alpha_{1,1} \right\}}$$

$$(J_{\widehat{\ominus}^{k+1}} r^{k+1})_{jj} = \frac{z_{jj}^\top \left((z_{jj} e_{jj}^\top - Z) J_{\widehat{\ominus}^k} + (\mathrm{Id}_n - X J_{\widehat{\beta}^k}) \right) r^k}{\|z_{jj}\|^2 + n\alpha_{2,2}} \mathbb{1}_{\left\{ \left| z_{jj}^\top \left(r^k + \widehat{\ominus}_{jj}^k z_{jj} \right) \right| \geqslant n\alpha_{1,2} \right\}}$$

Pro : update structure well fitted for coordinate descent.

Simulation Study : $y = X\beta^\star + Z\Theta^\star + \varepsilon$ w. $n = 100, p = 50, q = 1275$

- study different interaction hierarchical[7] cases : strong, weak and random.
- $X \sim \mathcal{N}(0_p, \Sigma_{p\times p})$, $\Sigma_{p\times p}$ produce correlation Toeplitz on X :

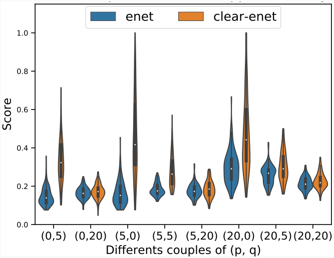
$$\Sigma_{p\times p} = \begin{pmatrix} 1 & 0.9 & 0.9^2 & \dots & 0.9^{p-2} & 0.9^{p-1} \\ 0.9 & 1 & 0.9 & \dots & 0.9^{p-3} & 0.9^{p-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.9^{p-1} & 0.9^{p-2} & 0.9^{p-3} & \dots & 0.9 & 1 \end{pmatrix}$$

- $\beta^\star$ and $\Theta^\star$: coefficients randomly chosen and equals ± 1 ;
- $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 \, \mathrm{Id}_n)$ according to SNR[8] ≈ 16

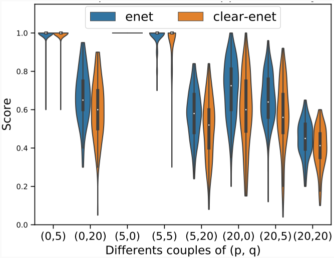
[7] J. Bien, J. Taylor, and R. Tibshirani (2013). "A lasso for hierarchical interactions". In: *Ann. Statist.* 41.3, pp. 1111–1141.

[8] P. Bühlmann and J. Mandozzi (2013). "High-dimensional variable screening and biais in subsequent inference, with an empirical comparison". In:

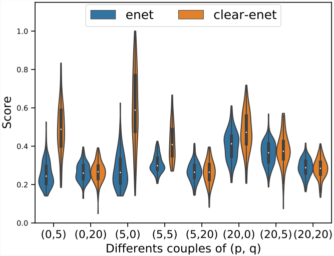
Simulated result in random cases : 100 repetitions of 5-folds CV.[9]



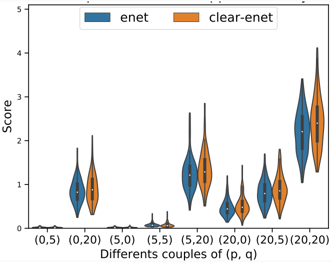
(a) Precision



(b) Recall



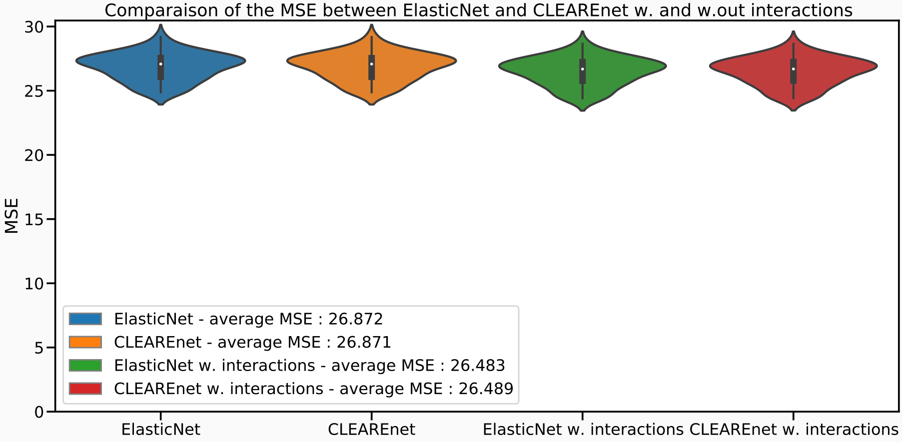
(c) F1 score



(d) MSE

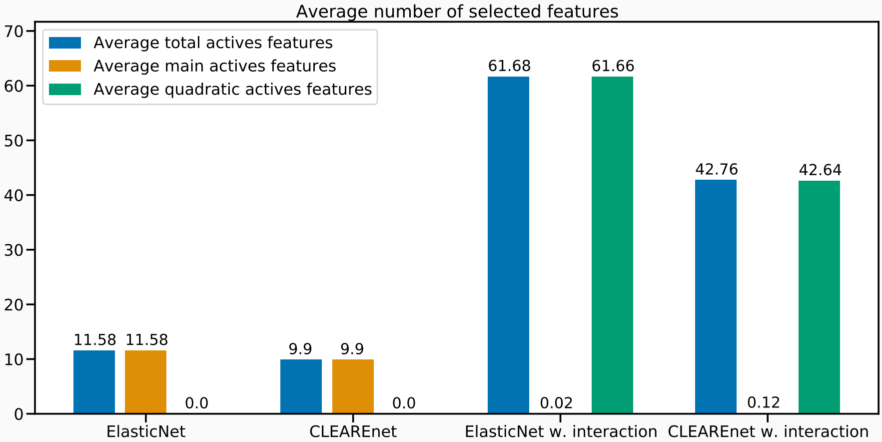
[9] Parameters : `l1_ratio` $\in \{1, 0.99, 0.95, 0.9\}$, duality gap fix to 10^{-4}, $\alpha_{\min} = \dfrac{\alpha_{\max}}{100}$ and X, Z standardize

First real data result[10] : only CORE promoter part



[10] Data and parameters : $n = 19393$, $p = 20$, $q = 210$, 50 repetitions of 5-folds CV, `l1_ratio` $\in \{1, 0.99, 0.95, 0.9\}$, duality gap fix to 10^{-4}, $\alpha_{\min} = \frac{\alpha_{\max}}{1000}$ and X, Z standardize

First real data result[11] : only CORE promoter part



[11] Data and parameters : $n = 19393$, $p = 20$, $q = 210$, 50 repetitions of 5-folds CV, `l1_ratio` $\in \{1, 0.99, 0.95, 0.9\}$, duality gap fix to 10^{-4}, $\alpha_{\min} = \frac{\alpha_{\max}}{1000}$ and X, Z standardize

Conclusion

Conclusion :

1. ElasticNet permits to handle large datasets (no need to build Z) ;
2. CLEAREnet permits to debiaise ElasticNet and reduce the number of actives features ;

---

[12] M. Massias, A. Gramfort, and J. Salmon (2018). "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*.

[13] Q. Bertrand et al. (2020). "Implicit differentiation of Lasso-type models for hyperparameter optimization". In: *ICML*.

Conclusion :

1. ElasticNet permits to handle large datasets (no need to build Z) ;
2. CLEAREnet permits to debiaise ElasticNet and reduce the number of actives features ;

Perspectives :

1. Speed-up the method, e. g. with working set strategies[12] ;
2. Scaling up experiment : both on simulated and real data ;
3. Clever hyperparameters tuning[13] (in $\mathbb{R}^2$ or $\mathbb{R}^4$).

[12] M. Massias, A. Gramfort, and J. Salmon (2018). "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*.

[13] Q. Bertrand et al. (2020). "Implicit differentiation of Lasso-type models for hyperparameter optimization". In: *ICML*.

Conclusion :

1. ElasticNet permits to handle large datasets (no need to build Z) ;
2. CLEAREnet permits to debiaise ElasticNet and reduce the number of actives features ;

Perspectives :

1. Speed-up the method, e. g. with working set strategies[12] ;
2. Scaling up experiment : both on simulated and real data ;
3. Clever hyperparameters tuning[13] (in $\mathbb{R}^2$ or $\mathbb{R}^4$).

Thanks for your attention.

[12] M. Massias, A. Gramfort, and J. Salmon (2018). "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*.

[13] Q. Bertrand et al. (2020). "Implicit differentiation of Lasso-type models for hyperparameter optimization". In: *ICML*.

References

References

Bertrand, Q. et al. (2020). "Implicit differentiation of Lasso-type models for hyperparameter optimization". In: *ICML*.

Bessière, Chloé et al. (Jan. 2018). "Probing instructions for expression regulation in gene nucleotide compositions". In: *PLOS Computational Biology* 14.1, pp. 1–28.

Bien, J., J. Taylor, and R. Tibshirani (2013). "A lasso for hierarchical interactions". In: *Ann. Statist.* 41.3, pp. 1111–1141.

Bühlmann, P. and J. Mandozzi (2013). "High-dimensional variable screening and biais in subsequent inference, with an empirical comparison". In:

Deledalle, C.-A. et al. (2017). "CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration". In: *SIAM J. Imaging Sci.* 10.1, pp. 243–284.

Friedman, J. et al. (2007). "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2, pp. 302–332.

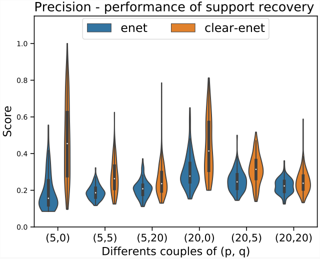
Massias, M., A. Gramfort, and J. Salmon (2018). "Celer: a Fast Solver for the Lasso with Dual Extrapolation". In: *ICML*.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1, pp. 267–288.

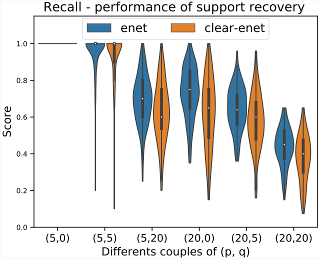
Zou, H. and T. J. Hastie (2005). "Regularization and variable selection via the elastic net". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2, pp. 301–320.

Appendix

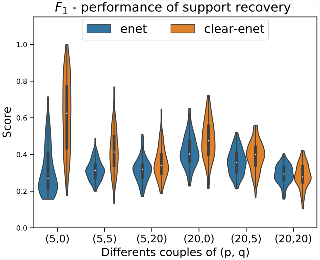
Strong cases : Result after 100 repetition of 5-folds CV.[14]



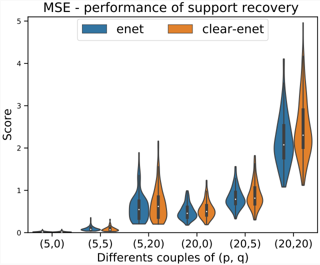
(a) Precision



(b) Recall



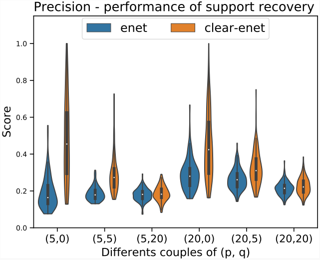
(c) F1 score



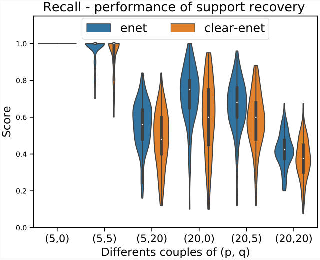
(d) MSE

[14] Others parameters : `l1_ratio` $\in \{1, 0.99, 0.95, 0.9\}$, duality gap fix to 10^{-4}, $\alpha_{\min} = \dfrac{\alpha_{\max}}{100}$ and X, Z standardized.

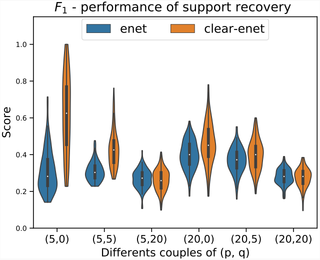
Weak cases : Result after 100 repetition of 5-folds CV.[15]



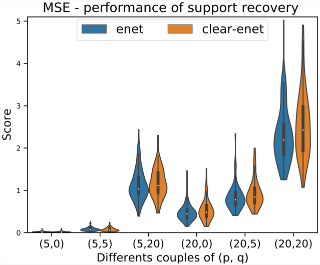
(a) Precision



(b) Recall



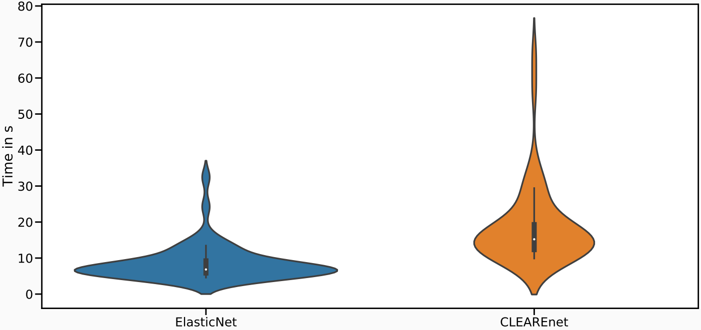
(c) F1 score



(d) MSE

[15] **Others parameters** : `l1_ratio` $\in \{1, 0.99, 0.95, 0.9\}$, **duality gap fix to** 10^{-4}, $\alpha_{\min} = \frac{\alpha_{\max}}{100}$ **and** X, Z **standardized.**

Time comparaison : ElasticNet and CLEAREnet



Correlation Matrix :