Safe elimination of variables

V. Leclère (ENPC), L. El Ghaoui

SESO 2015

June 22, 2015

An introductive regression problem

A LASSO problem, with n features and m datapoints reads

$$\min_{w \in \mathbb{R}^n} \|A^T w - y\|_2 + \lambda \|w\|_1$$

where

- $w \in \mathbb{R}^n$ is the decision variable;
- $A \in \mathbb{R}^{n \times m}$ is the data matrix, each column being one data point, and each line being a feature;
- $y \in \mathbb{R}^m$ being the output value;
- λ is a parameter enforcing sparsity of w.

Why is sparsity desirable ?

- Interpretation
  - gene identification for further research
  - important word of a document
  - portfolio selection / adjustement
- Memory storage of the solution, and computation gain

However, sparsity seems to come at the cost of harder optimization problem. We are going to show that in some cases we can simplify the problem.

Why is sparsity desirable ?

- Interpretation
    - gene identification for further research
    - important word of a document
    - portfolio selection / adjustement
- Memory storage of the solution, and computation gain

However, sparsity seems to come at the cost of harder optimization problem. We are going to show that in some cases we can simplify the problem.

The L^1 penalization

- We would really like to penalized the cardinality of non-zeros in the solution, but it leads to combinatorial problem.
- Penalizing the L^1 norm gives good numerical results, and is computationally faster.
- We can re-weight the L^1 norm to be closer to the cardinality penalization.
- There are some theoretical results on the L^1 penalization
  - "Robust uncertainty principles: Exact recovery from highly incomplete Fourier information", Emmanuel Candes, Justin Romberg, and Terence Tao, 2006.
  - "Stable signal recovery from incomplete and inaccurate measurements", Emmanuel Candes, Justin Romberg, and Terence Tao, 2006.
  - ...

Presentation Outline

1. SAFE elimination for LASSO
   - Working out the dual
   - Exploiting bounds
   - Robust LASSO
2. SAFE elimination in learning setting
   - Machine Learning Setting
   - Safe Elimination in Machine Learning
   - Numerical results
3. SAFE elimination in constrained optimization
   - Re-allocation problem formulation
   - Duality approach
   - Penalization

Contents

Dual representation of norm

We start with a very simple remark

$$\forall w \in \mathbb{R}^n, \qquad \|w\|_p = \max \quad \{\alpha^T w \quad | \quad \|\alpha\|_{p'} \leq 1\},$$

with

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

In particular

$$\|A^T w - y\|_2 + \lambda\|w\|_1 = \max_{\substack{\|\alpha\|_2 \leq 1 \\ \|\beta\|_\infty \leq \lambda}} \alpha^T(A^T w - y) + \beta^T w$$

Dual representation of norm

We start with a very simple remark

$$\forall w \in \mathbb{R}^n, \qquad \|w\|_p = \max \quad \{\alpha^T w \quad | \quad \|\alpha\|_{p'} \leq 1\},$$

with

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

In particular

$$\|A^T w - y\|_2 + \lambda \|w\|_1 = \max_{\substack{\|\alpha\|_2 \leq 1 \\ \|\beta\|_\infty \leq \lambda}} \alpha^T (A^T w - y) + \beta^T w$$

Dual formulation of LASSO

The LASSO problem

$$\min_{w \in \mathbb{R}^n} \|A^T w - y\|_2 + \lambda \|w\|_1$$

is equivalent to

$$\max_{\substack{\|\alpha\|_2 \leq 1 \\ \|\beta\|_\infty \leq \lambda}} \quad \min_{w \in \mathbb{R}^n} \quad (A\alpha + \beta)^T w - \alpha^T y$$

or

$$\begin{aligned} \max_{\|\alpha\|_2 \leq 1} \quad & -\alpha^T y \\ s.t. \quad & |\beta_i| \leq \lambda, \qquad \forall i \\ & a_i^T \alpha + \beta_i = 0, \qquad \forall i \end{aligned}$$

SAFE elimination for LASSO

Finally the dual LASSO problem reads

$$\max_{\|\alpha\|_2 \leq 1} \quad -\alpha^T y$$
$$s.t. \quad |a_i^T \alpha| \leq \lambda, \qquad \forall i$$

In particular if

$$(T1) \qquad \max_{\|\alpha\|_2 \leq 1} \alpha^T a_i = \|a_i\|_2 < \lambda$$

then we can suppress the feature i from the regression.

Contents

Improving SAFE with bounds

Assume that we know some bounds $0 \leq \underline{\gamma} < \overline{\gamma} \leq +\infty$ on the value of the LASSO problem. Then we have the following elimination test

$$\max_{\substack{\|\alpha\|_2 \leq 1 \\ \underline{\gamma} \leq -\alpha^T y \leq \overline{\gamma}}} \alpha^T a_i < \lambda.$$

And we can obtain an explicit test of feature i only requiring the computation of $a_i^T y$ and $\|a_i\|_2$. We can note that the lower bound is the most significant.

Improving SAFE with bounds

- If $X_i^T y = 0$, and

$$\|X_i\|_2 \sqrt{\|y\|_2^2 - \underline{\gamma}^2} \qquad < \lambda \|y\|_2$$

  then $w_i^\sharp = 0$.
- If $|X_i^T y| / \|X_i\|_2 > \|y\|_2^2 / \underline{\gamma}$, and

$$|X_i^T y| \underline{\gamma} + \sqrt{\|y\|_2^2 \|X_i\|_2^2 - |X_i^T y|^2} \sqrt{\|y\|^2 - \underline{\gamma}^2} \qquad < \lambda \|y\|_2^2,$$

  then $w_i^\sharp = 0$.
- If $|X_i^T y| / \|X_i\|_2 < \|y\|_2^2 / \bar{\gamma}$, and

$$\sqrt{\|y\|_2^2 - |X_i^T y|^2} \sqrt{\|y\|_2^2 - \bar{\gamma}^2} + |X_i^T y| \bar{\gamma} \qquad < \lambda \|y\|_2^2,$$

  and

$$\sqrt{\|y\|_2^2 - |X_i^T y|^2} \sqrt{\|y\|_2^2 - \underline{\gamma}^2} - |X_i^T y| \underline{\gamma} \qquad < \lambda \|y\|_2^2,$$

Obtaining bounds

- 0 is always a lower bound, $\|y\|_2$ is always an upper bound.
- An upper bound is given by any primal solution w.
- A lower bound is given by any admissible dual solution α.
- Assume that we have solved $\mathcal{P}(\lambda)$, and want to solve $\mathcal{P}(\lambda')$.
    - If $\lambda > \lambda'$, optimal primal solution can be used to compute upper bound, dual solution can be scaled to obtain lower bound.
    - If $\lambda < \lambda'$, optimal primal solution can be used to compute upper bound, optimal value gives a lower bound, optimal dual solution can be scaled to obtain lower bound.
    - We can also exploit the fact that the value function (in function of λ) is concave to obtain a lower bound.

Obtaining bounds

- 0 is always a lower bound, $\|y\|_2$ is always an upper bound.
- An upper bound is given by any primal solution w.
- A lower bound is given by any admissible dual solution α.
- Assume that we have solved $\mathcal{P}(\lambda)$, and want to solve $\mathcal{P}(\lambda')$.
  - If $\lambda > \lambda'$, optimal primal solution can be used to compute upper bound, dual solution can be scaled to obtain lower bound.
  - If $\lambda < \lambda'$, optimal primal solution can be used to compute upper bound, optimal value gives a lower bound, optimal dual solution can be scaled to obtain lower bound.
  - We can also exploit the fact that the value function (in function of λ) is concave to obtain a lower bound.

Obtaining bounds

- 0 is always a lower bound, $\|y\|_2$ is always an upper bound.
- An upper bound is given by any primal solution w.
- A lower bound is given by any admissible dual solution α.
- Assume that we have solved $\mathcal{P}(\lambda)$, and want to solve $\mathcal{P}(\lambda')$.
    - If $\lambda > \lambda'$, optimal primal solution can be used to compute upper bound, dual solution can be scaled to obtain lower bound.
    - If $\lambda < \lambda'$, optimal primal solution can be used to compute upper bound, optimal value gives a lower bound, optimal dual solution can be scaled to obtain lower bound.
    - We can also exploit the fact that the value function (in function of λ) is concave to obtain a lower bound.

Contents

Low-rank approximation of data

- Low rank approximation of data $X = \hat{X} + \Delta$, with controlled error $\|\Delta\| \leq \varepsilon$. Example : partial SVD decomposition.
- Worst case on the approximation :

$$\min_{w \in \mathbb{R}^n} \max_{\|\Delta\| \leq \varepsilon} \|(\hat{A} + \Delta)^T w - y\|_2 + \lambda \|w\|_1$$

- Equivalent to

$$\min_{w \in \mathbb{R}^n} \|\hat{A}^T w - y\|_2 + \varepsilon \|w\|_2 + \lambda \|w\|_1.$$

- New SAFE test :

$$\|a_i\|_2 < \lambda - \varepsilon$$

Contents

El Ghaoui, L., Viallon, V. and Rabbani, T. (2012).
Safe feature elimination for the lasso and sparse supervised learning problems.
Pacific Journal of Optimization

Generic Machine Learning problem

We consider a generic sparse supervised learning process

$$\mathcal{P}(\lambda) = \varphi(\lambda) = \min_{v,w} \sum_{j=1}^{m} f(a_j^T w + b_j v + c_j) + \lambda \|w\|_1,$$

where f is a l.s.c. convex loss function where $a_j \in \mathbb{R}^n$ are datapoints. They are column of the data matrix A whose lines are x_i.

The Fenchel conjugate of f is given by

$$f^*(\theta) = \max_{\xi} \{\theta\xi - f(\xi)\}.$$

As f is l.s.c. convex we have

$$f(x_j) = f^{**}(x_j) = \max_{\theta_j} \{\theta_j x_j - f^*(\theta_j)\}.$$

Generic Machine Learning problem

We consider a generic sparse supervised learning process

$$\mathcal{P}(\lambda) = \varphi(\lambda) = \min_{v,w} \sum_{j=1}^{m} f(a_j^T w + b_j v + c_j) + \lambda \|w\|_1,$$

where f is a l.s.c. convex loss function where $a_j \in \mathbb{R}^n$ are datapoints. They are column of the data matrix A whose lines are x_i.

The Fenchel conjugate of f is given by

$$f^*(\theta) = \max_{\xi} \{\theta\xi - f(\xi)\}.$$

As f is l.s.c. convex we have

$$f(x_j) = f^{**}(x_j) = \max_{\theta_j} \{\theta_j x_j - f^*(\theta_j)\}.$$

Contents

A dual formulation

Using the Fenchel representation as well as the dual representation of the L_1-norm, $\mathcal{P}(\lambda)$ reads

$$\min_{w,v} \sum_{j=1}^{m} \max_{\theta_j} \{\theta_j(a_j^T w + b_j v + c_j) - f^*(\theta_i)\} + \max_{\|\alpha\|_\infty \leq \lambda} \alpha^T w$$

Rearranging variables yields

$$\min_{w,v} \max_{\theta, \|\alpha\|_\infty \leq \lambda} \underbrace{c^T\theta - \sum_{j=1}^{m} f^*(\theta_j)}_{G(\theta)} + \theta^T bv + (\sum_{j=1}^{m} \theta_j a_j + \alpha)^T w$$

Interverting min and max, and minimizing over v and w gives

$$\mathcal{D}(\lambda): \quad \max \quad \left\{ G(\theta) \quad | \quad \theta^T b = 0, \quad \forall i, |x_i^T \theta| \leq \lambda \right\}$$

SAFE - test

Due to optimality conditions if $|\theta^T x_k| < \lambda$ then $w_k^\sharp = 0$.
If we have a lower bound $0 \le \underline{\gamma} \le \sum_i f(c_i)$ on our optimization problem, we deduce the following test

$$\lambda > T(\underline{\gamma}, x_k) := \max \left\{ |\theta^T x_k| \mid G(\theta) \ge \underline{\gamma}, \quad \theta^T b = 0 \right\} \Longrightarrow w_k^\sharp = 0$$

Note that $T(\underline{\gamma}, x_k) = \max\{P(\underline{\gamma}, x_k), P(\underline{\gamma}, -x_k)\}$, where

$$P(\underline{\gamma}, x) = \max\{\theta^T x \mid G(\theta) \ge \underline{\gamma}, \quad \theta^T b = 0\}$$

Or again

$$P(\underline{\gamma}, x) = \min_{\mu \ge 0, \nu} -\underline{\gamma}\mu + \mu \sum_{j=1}^{m} f\left(\frac{x_j + \mu c_j + \nu b_j}{\mu}\right).$$

SAFE - test

Due to optimality conditions if $|\theta^T x_k| < \lambda$ then $w_k^\sharp = 0$.
If we have a lower bound $0 \leq \underline{\gamma} \leq \sum_i f(c_i)$ on our optimization problem, we deduce the following test

$$\lambda > T(\gamma, x_k) := \max \{|\theta^T x_k| \mid G(\theta) \geq \underline{\gamma}, \quad \theta^T b = 0\} \Longrightarrow w_k^\sharp = 0$$

Note that $T(\underline{\gamma}, x_k) = \max\{P(\underline{\gamma}, x_k), P(\underline{\gamma}, -x_k)\}$, where

$$P(\underline{\gamma}, x) = \max\{\theta^T x \mid G(\theta) \geq \underline{\gamma}, \quad \theta^T b = 0\}$$

Or again

$$P(\underline{\gamma}, x) = \min_{\mu \geq 0, \nu} -\underline{\gamma}\nu + \mu \sum_{j=1}^{m} f\left(\frac{x_j + \mu c_j + \nu b_j}{\mu}\right).$$

Obtaining lower bound

A lower bound is given by any admissible point θ_0 of the dual problem $\mathcal{D}(\lambda)$, by $\underline{\gamma} = G(\theta_0)$.

In fact if we choose θ such that $\theta^T b = 0$, and define $\lambda_0 = \|X\theta_0\|_\infty$. Then we need a scaling factor s, such that $\theta = s\theta_0$ is an admissible point for $\mathcal{D}(\lambda)$. The best scaling factor is given by

$$s \in \arg\max\{G(s\theta_0) \quad | \quad |s| \leq \frac{\lambda}{\lambda_0}\},$$

which is a one dimensional convex optimization problem.

Remark : this scaling procedure can be used to obtain lower bounds from solution obtained from higher coefficients λ.

Contents

A numerical experiment

- We consider text data in a bag-of-word format where stop words have been eliminated and without capitalization.
- The data consists in the headlines from the New York Times (1985 - 2007).
- Number of features $n = 159'943$, number of documents $m = 3'241'260$, about 90 non-zero per features.

Numerical results

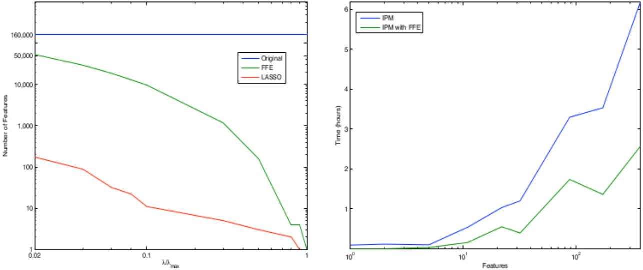

Figure 4: *Left panel:* In blue, the original number of features. In red, the number of features obtained via IPM-LASSO, as a function of of the ratio $\lambda/\lambda_{\max}$. In green, the number of features not eliminated by SAFE1. *Right panel:* computational time for solving the IPM-LASSO before and after SAFE1 as function of the number of active features at the optimum

Contents

Re-allocation problem

- Consider a problem of production where we have multiple unit of production and need to satisfy an offer equal demand equality.
- On day $D - 1$ we have a prediction of the demand and define a planning of production for day D.
- On day D we have a new prediction of demand and have to modify the planning of production.
- Problem : find the best new planning with a small number of modification.

Re-allocation problem formulation

$$
\begin{aligned}
\min_{X^{(1)}} \quad & \sum_{j=1}^{m} c_j^{(1)} \cdot X_j^{(1)} & \\
s.t. \quad & AX^{(1)} = d & \text{new production } = \text{ demand} \\
& LB \leq HX^{(1)} \leq UB & \text{Bounds on new production} \\
& Y^T(X^{(1)} - X^{(0)}) = 0 & \text{Modification iff } Y_j = 1 \\
& Y \in \{0,1\}^m & \\
& \sum_{i=1}^{n} Y_i \leq n & \text{No more than } n \text{ modifications}
\end{aligned}
$$

Re-allocation problem relaxation

We propose a L_1 relaxation of the cardinality constraint on the L_∞ norm of the modification.

$$\begin{aligned} \min_{X^{(1)}} \quad & \sum_{j=1}^{m} c_j^{(1)} \cdot X_j^{(1)} + \lambda \sum_{j=1}^{m} \delta_j \|X_j^{(1)} - X_j^{(0)}\|_\infty \\ s.t. \quad & AX^{(1)} = d \\ & LB \leq HX^{(1)} \leq UB \end{aligned}$$

Contents

Re-allocation problem relaxation II

With a slight change of notation and simplification we obtain

$$
\begin{aligned}
\min_{\tilde{X}} \quad & c \cdot \tilde{X} + \lambda \sum_{j=1}^{m} \|\tilde{X}_j\|_\infty \\
s.t. \quad & \sum_{j=1}^{m} A_j \tilde{X}_j = b \\
& \underline{X}_j \leq \tilde{X} \leq \overline{X}_j \qquad \forall j
\end{aligned}
$$

Dualizing the equality constraint we obtain

$$
\max_{\mu} \quad -\mu \cdot b + \sum_{j=1}^{m} \min_{\underline{X}_j \leq \tilde{X}_j \leq \overline{X}_j} \quad (c_j + A_j^T \mu) \cdot \tilde{X}_j + \lambda \|\tilde{X}_j\|_\infty,
$$

Re-allocation problem relaxation II

With a slight change of notation and simplification we obtain

$$\min_{\tilde{X}} \quad c \cdot \tilde{X} + \lambda \sum_{j=1}^{m} \|\tilde{X}_j\|_\infty$$

$$s.t. \quad \sum_{j=1}^{m} A_j \tilde{X}_j = b$$

$$\underline{X}_j \leq \tilde{X} \leq \overline{X}_j \qquad \forall j$$

Dualizing the equality constraint we obtain

$$\max_{\mu} \quad -\mu \cdot b + \sum_{j=1}^{m} \min_{\underline{X}_j \leq \tilde{X}_j \leq \overline{X}_j} \quad (c_j + A_j^T \mu) \cdot \tilde{X}_j + \lambda \|\tilde{X}_j\|_\infty,$$

Test for the reoptimization problem

for any given multiplier μ

$$(c_j + A_j^T \mu) \cdot \tilde{X}_j + \lambda \|\tilde{X}_j\|_\infty \geq \left(\lambda - \|c_j + A_j^T \mu\|_1\right) \|\tilde{X}_j\|_\infty .$$

Which leads to a first safe test (given that $\tilde{X} = 0$ is admissible – or without bound on $\tilde{X}$)

$$(T1) \qquad \|c_j + A_j^T \mu\|_1 \leq \lambda \qquad \Longrightarrow \qquad X_j^\sharp(\mu) = X_j^{(0)} .$$

A generic idea

Consider the almost decomposable problem

$$\min_{x} \quad \{\sum_{j=1}^{m} f_j(x_j) + \lambda \|X\|_1 \quad | \quad \sum_{j=1}^{m} g_j(x_j) = 0\}$$

The dual problem reads

$$\max_{\mu} \quad \sum_{j=1}^{m} \underbrace{\min_{x_j} f_j(x_j) + \mu^T g_j(x_j) + \lambda |x_j|}_{\mathcal{P}_j(\lambda,\mu)}$$

This problem can be solved by iteration over μ (classical decomposition method).
For a given μ, if 0 is an optimal solution to $\mathcal{P}_j(\lambda,\mu)$ then unit j is not modified.

A generic idea

Consider the almost decomposable problem

$$\min_{x} \quad \{\sum_{j=1}^{m} f_j(x_j) + \lambda\|X\|_1 \quad | \quad \sum_{j=1}^{m} g_j(x_j) = 0\}$$

The dual problem reads

$$\max_{\mu} \quad \sum_{j=1}^{m} \quad \underbrace{\min_{x_j} f_j(x_j) + \mu^T g_j(x_j) + \lambda|x_j|}_{\mathcal{P}_j(\lambda,\mu)}$$

This problem can be solved by iteration over μ (classical decomposition method).
For a given μ, if 0 is an optimal solution to $\mathcal{P}_j(\lambda,\mu)$ then unit j is not modified.

Contents

A penalized version

Instead of dualizing the constraint we look at its penalized version

$$\min_{\tilde{X}} \quad \sum_{j=1}^{m} c_j^T \tilde{X}_j + \lambda \|\tilde{X}_j\|_\infty + \varepsilon \| \sum_{i=1}^{m} A_j \tilde{X}_j - \tilde{b} \|_2$$

Dual formulation of the norm leads to

$$\begin{aligned} \min_{\tilde{X}} \quad \max_{\alpha,\beta} \quad & \sum_{j=1}^{m} \Big(c_j^T \tilde{X}_j + \alpha_j^T \tilde{X}_j + \beta^T A_j \tilde{X}_j \Big) - \beta^T \tilde{b} \\ s.t. \quad & \|\alpha_j\|_1 \leq \lambda \qquad \forall j \\ & \|\beta\|_2 \leq \varepsilon \end{aligned}$$

Obtaining a safe elimination test

Interverting max and min operators and minimizing over $\tilde{X}_j$ gives

$$
\begin{aligned}
\max_{\alpha,\beta} \quad & -\beta^T \tilde{b} \\
s.t. \quad & \|\alpha_j\|_1 \le \lambda \qquad \forall j \\
& \|\beta\|_2 \le \varepsilon \\
& c_j + \alpha_j + A_j^T \beta = 0
\end{aligned}
$$

Combining the constraints we see that

$$
(T2) \qquad \max_{\|\beta\|_2 \le 1} \|c_j + \varepsilon A_j^T \beta\|_1 < \lambda \qquad \Longrightarrow \qquad \tilde{X}_j^\sharp = 0
$$

then unit j is not modified.

Practical safe elimination tests

$$(T2) \qquad \max_{\|\beta\|_2 \leq 1} \|c_j + \varepsilon A_j^T \beta\|_1 < \lambda \qquad \Longrightarrow \qquad \tilde{X}_j^\sharp = 0$$

This test is hard to check, however we can derive the following more restrictive test

$$(T3) \qquad \|c_j\|_1 + \varepsilon \| \sum_{k=1}^{m} |a_k| \|_2 < \lambda \qquad \Longrightarrow \qquad \tilde{X}_i^\sharp = 0,$$

where $A_j = [a_1 \cdots a_m]$.
Furthermore if A_j is either diagonal or with positive coefficients both tests are equivalent.

Conclusion

- In some cases sparsity is desirable.
- One way of enforcing sparsity consists in penalizing the L^1 norm, with good numerical results and some theoretical results.
- We can design specific methods for this type of non-differentiable problem.
- Safe test consists in knowing the optimal value of some variable before solving the problem.
  - Reduce computation time.
  - Reduce memory size limits.

The end

Thank you for your attention !