Exponentially weighted aggregation
Laplace prior for linear regression

Arnak Dalalyan, Edwin Grappin & Quentin Paris

CREST ENSAE ParisTech université PARIS-SACLAY

edwin.grappin@ensae.fr

JPS - Les Houches - 2016

Goals & settings

We observe n **labels** $(Y_i)_{i\in\{1,\dots,n\}}$ and there is a linear relation between the label and the p **features** $(X_{i,j})_{j\in\{1,\dots,p\}}$ such that:

$$Y = X\beta^\star + \xi,$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta^\star \in \mathbb{R}^p$ and $\xi \in \mathbb{R}^n$ a random variable such that ξ_i is $\mathcal{N}(0, \sigma^2)$.

Goals & settings

We observe n **labels** $(Y_i)_{i\in\{1,\dots,n\}}$ and there is a linear relation between the label and the p **features** $(X_{i,j})_{j\in\{1,\dots,p\}}$ such that:

$$Y = X\beta^\star + \xi,$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta^\star \in \mathbb{R}^p$ and $\xi \in \mathbb{R}^n$ a random variable such that ξ_i is $\mathcal{N}(0, \sigma^2)$.

Our interests are:

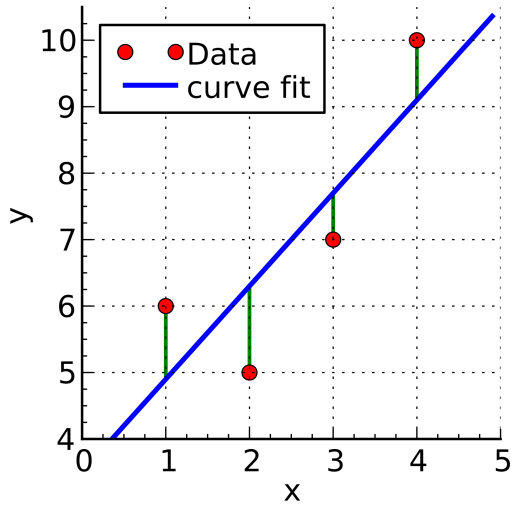
- **Low prediction loss**: $\|X(\beta^\star - \hat{\beta})\|_2^2$ (fitting $\beta^\star$ is less important),
- Good quality when p is **large** ($p >> n$),
- Efficient use of **sparsity** property of $\beta^\star$ ($\beta^\star$ is s-sparse if at most s elements are non null).

Least squares method

Ordinary least squares (OLS) estimator is defined by:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

OLS minimizes the sum of the squares of the residuals.

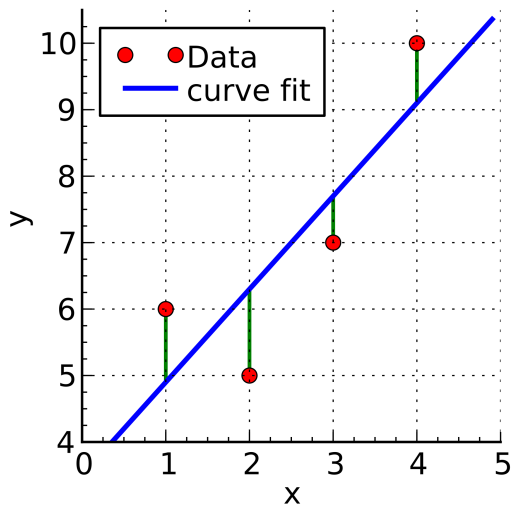


Least squares method

Ordinary least squares (OLS) estimator is defined by:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

OLS minimizes the sum of the squares of the residuals.



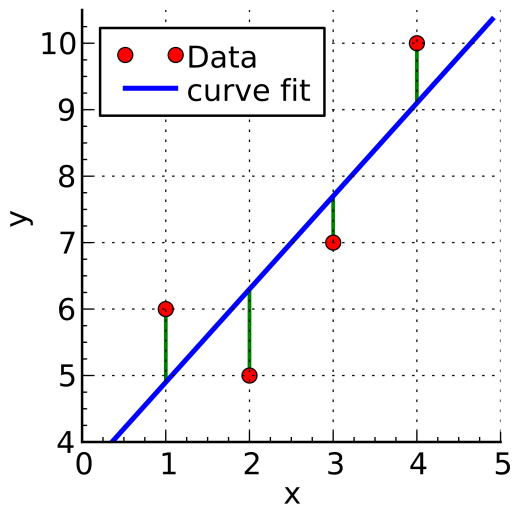
Overfitting. If p is very large, OLS has poor prediction results:

Least squares method

Ordinary least squares (OLS) estimator is defined by:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

OLS minimizes the sum of the squares of the residuals.



Overfitting. If p is very large, OLS has poor prediction results:

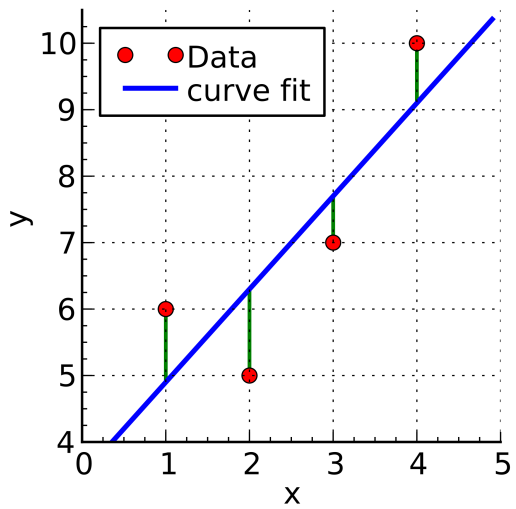
- There is not a unique solution when $p > n$,

Least squares method

Ordinary least squares (OLS) estimator is defined by:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

OLS minimizes the sum of the squares of the residuals.



Overfitting. If p is very large, OLS has poor prediction results:

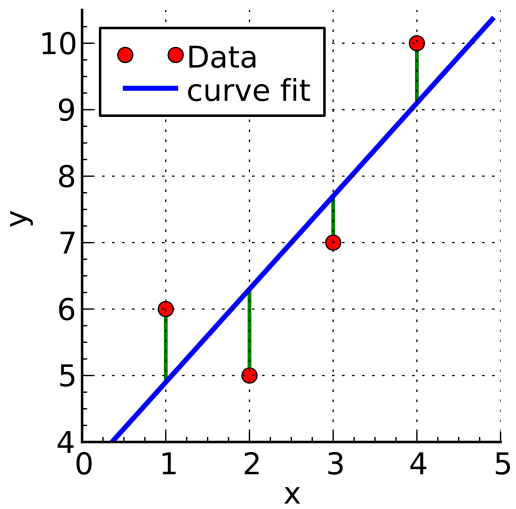
- There is not a unique solution when $p > n$,
- Does not detect meaningful features among all features,

Least squares method

Ordinary least squares (OLS) estimator is defined by:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

OLS minimizes the sum of the squares of the residuals.



Overfitting. If p is very large, OLS has poor prediction results:

- There is not a unique solution when $p > n$,
- Does not detect meaningful features among all features,
- Performance is focus on fitting the data not predicting labels.

Penalized regression

In our case, a good estimator has the following properties:

- Guarantees on prediction results,
- Use sparsity assumption to manage $p > n$,
- Computationnaly fast (of paramount importance when p is large).

Penalized regression

In our case, a good estimator has the following properties:

- Guarantees on prediction results,
- Use sparsity assumption to manage $p > n$,
- Computationnaly fast (of paramount importance when p is large).

Penalized regression is a method that combines the usual fitting term with a penalty term :

$$\hat{\beta}_{pen} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda P(\beta) \right),$$

P is the penalty function and $\lambda \geq 0$ controls the trade off between the two terms.

Subset selection with a ℓ_0 penalization

An intuitive candidate would be a penalization based on ℓ_0 pseudo-norm (the sparsity level):

$$\|\beta\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\beta_i \neq 0}.$$

$$\hat{\beta}_{\ell_0} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right).$$

Subset selection with a ℓ_0 penalization

An intuitive candidate would be a penalization based on ℓ_0 pseudo-norm (the sparsity level):

$$\|\beta\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\beta_i \neq 0}.$$

$$\hat{\beta}_{\ell_0} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right).$$

The penalty forces many elements of $\hat{\beta}$ to be null. It chooses the most important features.

Subset selection with a ℓ_0 penalization

An intuitive candidate would be a penalization based on ℓ_0 pseudo-norm (the sparsity level):

$$\|\beta\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\beta_i \neq 0}.$$

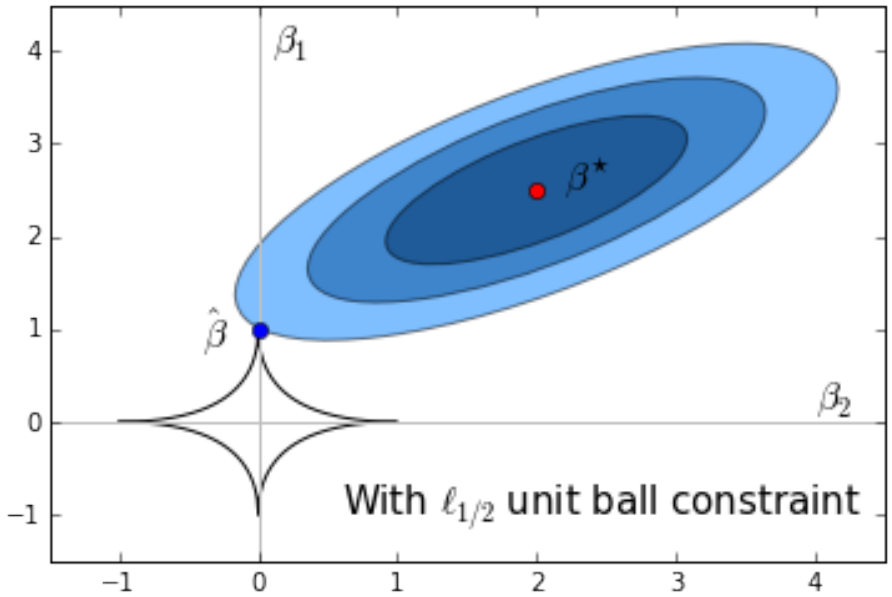
$$\hat{\beta}_{\ell_0} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right).$$

The penalty forces many elements of $\hat{\beta}$ to be null. It chooses the most important features.
Due to the ℓ_0 pseudo-norm, the objective function is nonconvex.
Hence, computational time grows exponentially with p.

Choice of the penalization term

Let $q > 0$, we consider the estimators

$$\hat{\beta}_q = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right).$$

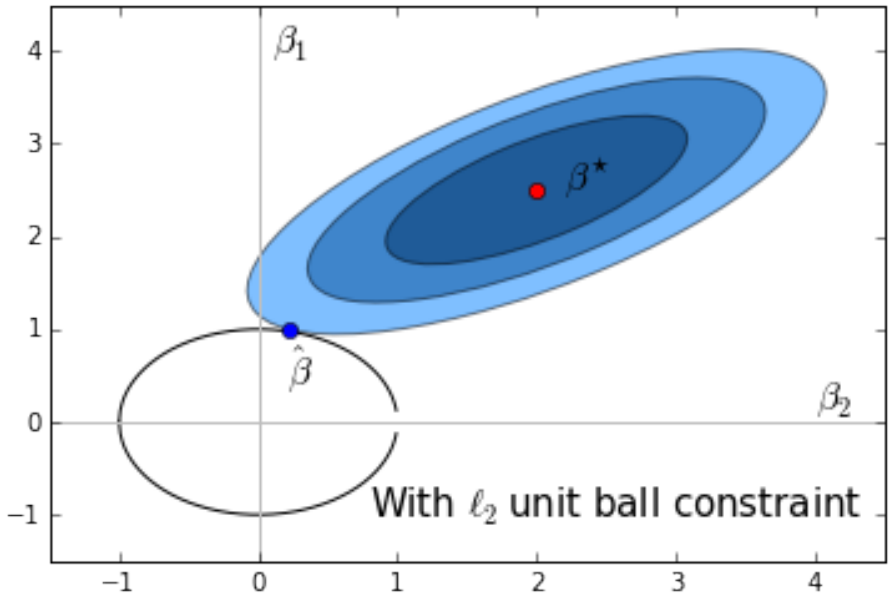


If $\mathbf{q < 1}$, the solution is **sparse** but the problem is **nonconvex**.

Choice of the penalization term

Let $q > 0$, we consider the estimators

$$\hat{\beta}_q = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right).$$

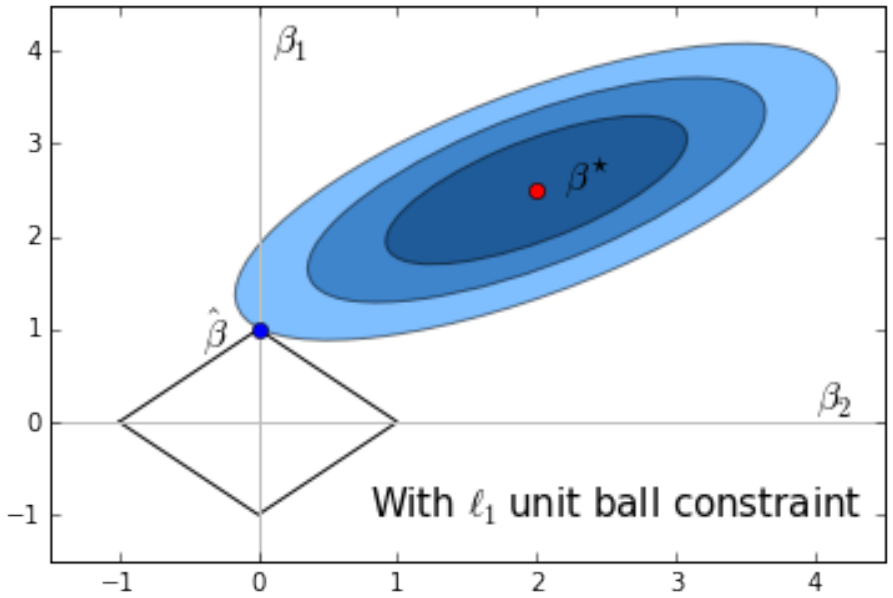


If $\mathbf{q > 1}$, the problem is **convex** but the solution is **not sparse**.

Choice of the penalization term

Let $q > 0$, we consider the estimators

$$\hat{\beta}_q = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right).$$



If $\mathbf{q = 1}$, the solution is **sparse** and the problem is **convex**.

Lasso, the ℓ_1 norm

The Lasso estimator is defined by:

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{\|Y - X\beta\|_2^2}{2n} + \lambda \|\beta\|_1 \right).$$

Lasso, the ℓ_1 norm

The Lasso estimator is defined by:

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{\|Y - X\beta\|_2^2}{2n} + \lambda \|\beta\|_1 \right).$$

Theorem

Dalalyan & al. (2014). *On the Prediction Performance of the Lasso*

Let $\lambda = 2\sigma\sqrt{\frac{2\log(p/\delta)}{n}}$. Then, with probability at least $1-\delta$,

$$\frac{\|X(\beta^\star - \hat{\beta}_L)\|_2^2}{n} \leq \inf_{\substack{\beta \in \mathbb{R}^p \\ s-sparse}} \left(\frac{\|X(\beta^\star - \beta)\|_2^2}{n} + \frac{10\ s\ \sigma^2 \log(p/\delta)}{n\ \kappa} \right),$$

where κ is a constant depending on the design of X.

EWA: definition

Lasso estimator is a maximum a posteriori estimator with Laplace prior :

$$\hat{\beta}_L = \arg\min_{\beta\in\mathbb{R}^p} \left(\frac{\|Y - X\beta\|_2^2}{2n} + \lambda\|\beta\|_1 \right)$$

$$= \arg\max_{\beta\in\mathbb{R}^p} \left[\underbrace{\exp\left(-\frac{1}{2}\frac{\|Y - X\beta\|_2^2}{\sigma^2} \right)}_{\propto \mathcal{N}(X\beta,\sigma^2 I_n)} \underbrace{\exp\left(-\frac{\lambda n}{\sigma^2}\|\beta\|_1 \right)}_{\propto \pi_0(\beta)\text{:Laplace prior}} \right]$$

EWA: definition

Lasso estimator is a maximum a posteriori estimator with Laplace prior :

$$\hat{\beta}_L = \arg\min_{\beta\in\mathbb{R}^p} \left(\frac{\|Y - X\beta\|_2^2}{2n} + \lambda\|\beta\|_1 \right)$$

$$= \arg\max_{\beta\in\mathbb{R}^p} \left[\underbrace{\exp\left(-\frac{1}{2}\frac{\|Y - X\beta\|_2^2}{\sigma^2} \right)}_{\propto \mathcal{N}(X\beta,\sigma^2 I_n)} \underbrace{\exp\left(-\frac{\lambda n}{\sigma^2}\|\beta\|_1 \right)}_{\propto \pi_0(\beta)\text{:Laplace prior}} \right]$$

Let, $V(\beta) = \frac{1}{2\sigma^2}\|Y - X\beta\|_2^2 + \frac{\lambda n}{\sigma^2}\|\beta\|_1$, and $\hat{\pi}_T(\beta) \propto \exp\left\{ -\frac{V(\beta)}{T} \right\}$.

EWA: definition

Lasso estimator is a maximum a posteriori estimator with Laplace prior :

$$\hat{\beta}_L = \arg\min_{\beta\in\mathbb{R}^p} \left(\frac{\|Y - X\beta\|_2^2}{2n} + \lambda\|\beta\|_1 \right)$$

$$= \arg\max_{\beta\in\mathbb{R}^p} \left[\underbrace{\exp\left(-\frac{1}{2}\frac{\|Y - X\beta\|_2^2}{\sigma^2} \right)}_{\propto \mathcal{N}(X\beta,\sigma^2 I_n)} \underbrace{\exp\left(-\frac{\lambda n}{\sigma^2}\|\beta\|_1 \right)}_{\propto \pi_0(\beta)\text{:Laplace prior}} \right]$$

Let, $V(\beta) = \frac{1}{2\sigma^2}\|Y - X\beta\|_2^2 + \frac{\lambda n}{\sigma^2}\|\beta\|_1$, and $\hat{\pi}_T(\beta) \propto \exp\left\{ -\frac{V(\beta)}{T} \right\}$. We define the exponentially weighted average (EWA) estimator with Laplace prior by

$$\hat{\beta}_{EWA} = \int_{\mathbb{R}^p} \beta\hat{\pi}_T(\beta)\mathrm{d}\beta.$$

Results

Theorem

Let $\lambda = 2\sigma\sqrt{\frac{2\log(p/\delta)}{n}}$, then with probability at least $1-\delta$,

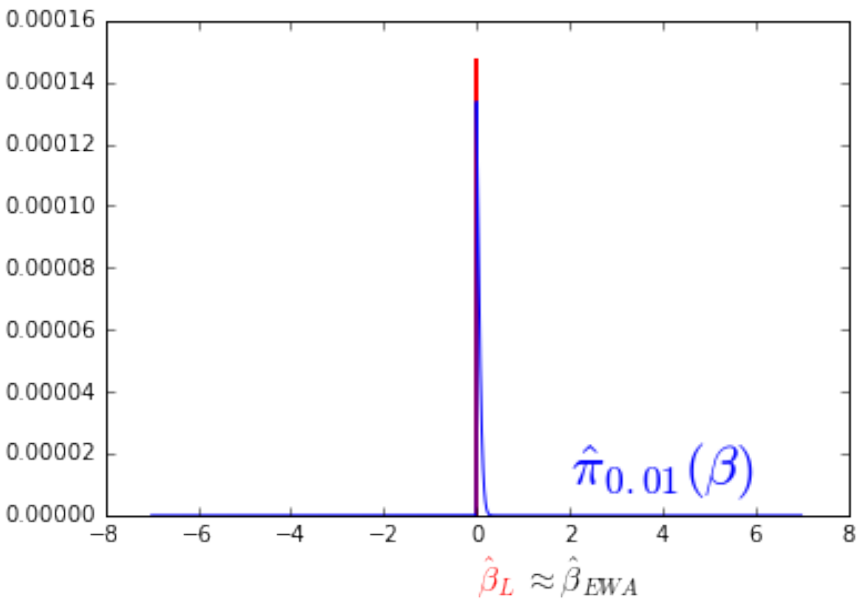
$$\frac{\|X(\beta^\star - \hat{\beta}_{EWA})\|_2^2}{n} \leq \inf_{\substack{\beta\in\mathbb{R}^p \\ s-sparse}} \left(\frac{\|X(\beta^\star-\beta)\|_2^2}{n} + \frac{10s\sigma^2\log(p/\delta)}{n\kappa}\right) + 2H(T).$$

Where

$$H(T) = pT - \int G(\beta)\hat{\pi}_T(\beta)\mathrm{d}\beta + G(\hat{\beta}_{EWA}),$$

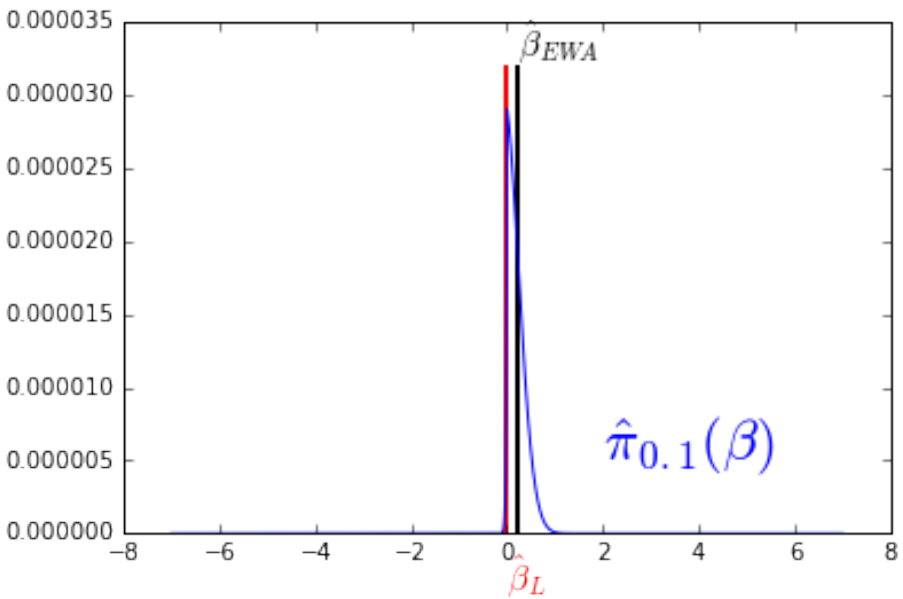
and $G(\beta) = \frac{1}{n}\|\mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$. G is convex, hence $H(T) \leq pT$.

The choice of T



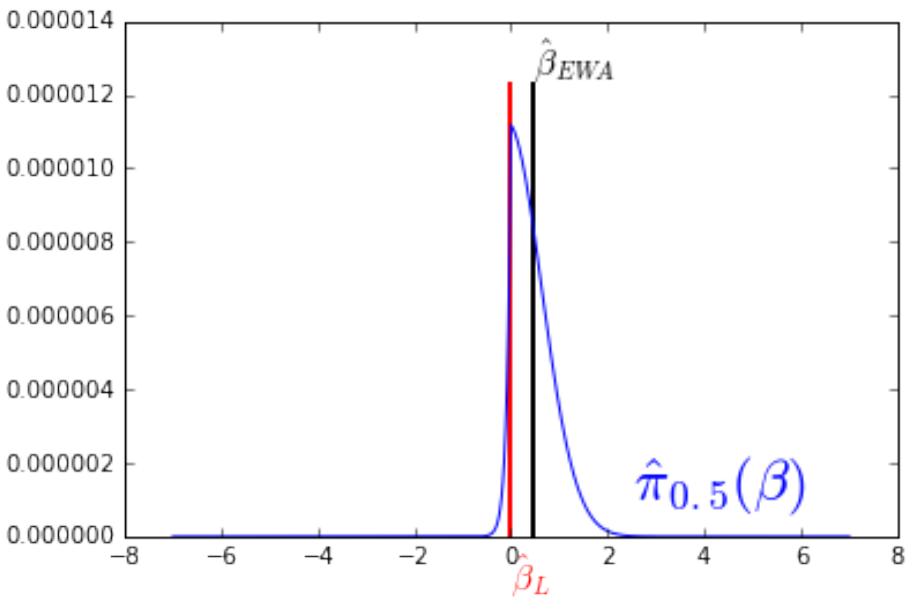
If $T = 0$, $\hat{\beta}_L = \hat{\beta}_{EWA}$.

The choice of T



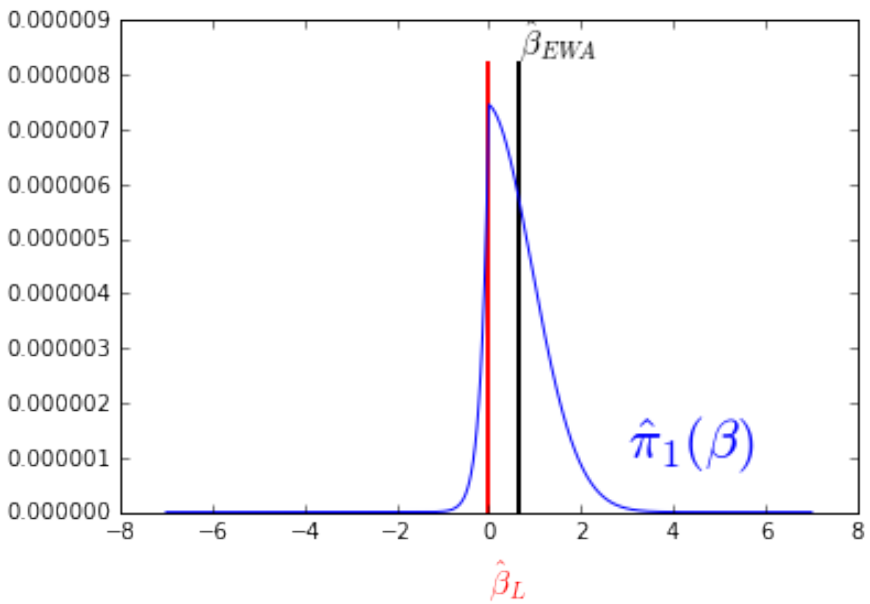
If $T = 0$, $\hat{\beta}_L = \hat{\beta}_{EWA}$.
We are interested in $T < 1/p$,
remember: $H(T) \leq pT$.

The choice of T



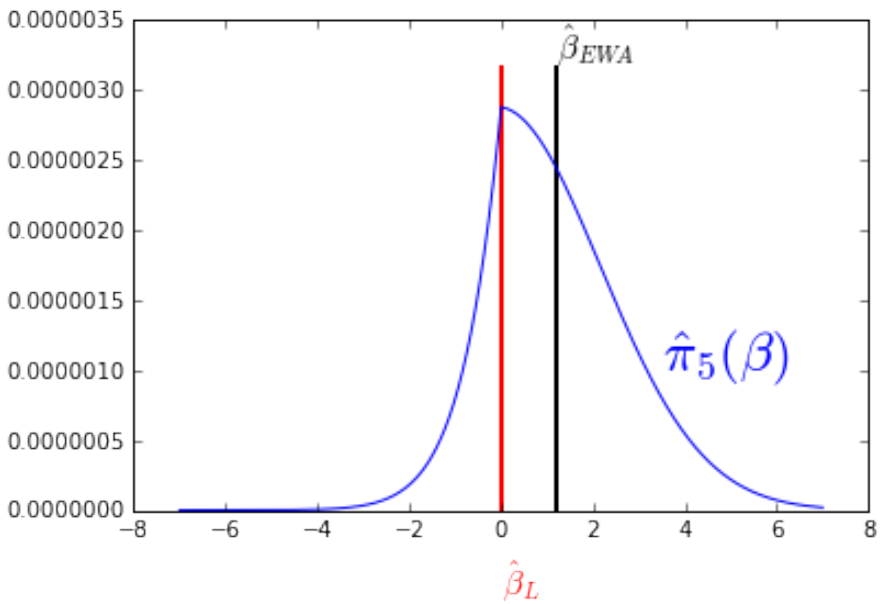
If $T = 0$, $\hat{\beta}_L = \hat{\beta}_{EWA}$.
We are interested in $T < 1/p$,
remember: $H(T) \leq pT$.
The larger T is, the larger is the variance of the posterior.

The choice of T



If $T = 0$, $\hat{\beta}_L = \hat{\beta}_{EWA}$.
We are interested in $T < 1/p$,
remember: $H(T) \leq pT$.
The larger T is, the larger is the variance of the posterior.
We believe that variance brings robustness to the choice of λ.

The choice of T



If $T = 0$, $\hat{\beta}_L = \hat{\beta}_{EWA}$.
We are interested in $T < 1/p$,
remember: $H(T) \leq pT$.
The larger T is, the larger is the variance of the posterior.
We believe that variance brings robustness to the choice of λ.

Conclusion & questions

Results:

- EWA with Laplace prior is a family of estimator that includes the Lasso.
- There is a sharp oracle inequality for this family of estimator.

Conclusion & questions

Results:

- EWA with Laplace prior is a family of estimator that includes the Lasso.
- There is a sharp oracle inequality for this family of estimator.

Questions:

- What is a good value of T?
- Can we prove a result on the robustness of λ?
- Can we compute efficiently this estimator?

Conclusion & questions

Results:

- EWA with Laplace prior is a family of estimator that includes the Lasso.
- There is a sharp oracle inequality for this family of estimator.

Questions:

- What is a good value of T?
- Can we prove a result on the robustness of λ?
- Can we compute efficiently this estimator?

Thank you!