Some reminders about the Lasso estimator.

Marius Kwemou [1,2]
Supervised by: M-L. Taupin [1] And A.K. Diongue [2]

[1] **Université d'Evry val d'Essonne (France)**
Laboratoire de Statistique et génome.

[2] **Université Gaston Berger de Saint-Louis (Sénégal)**
LERSTAD.

February, 2014

Outline

Outline

linear regression model

- One observes n independent and identically distributed (i.i.d) couples (z_1, Y_1),..,(z_n, Y_n) from the joint distribution of $(Z, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that:

$$y_i = z_i^T \beta_0 + \epsilon_i \tag{1}$$

- β_0 is the unknown parameter to estimate.
- Matrix notation, let $(y, X) \in \mathbb{R}^n \times \mathcal{M}_{n,d}$

$$y = X\beta_0 + \epsilon \tag{2}$$

linear regression model

- One observes n independent and identically distributed (i.i.d) couples (z_1, Y_1),..,(z_n, Y_n) from the joint distribution of $(Z, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that:

$$y_i = z_i^T \beta_0 + \epsilon_i \tag{1}$$

- β_0 is the unknown parameter to estimate.
- Matrix notation, let $(y, X) \in \mathbb{R}^n \times \mathcal{M}_{n,d}$

$$y = X\beta_0 + \epsilon \tag{2}$$

Estimation

- If $n > d$, the estimator $\hat{\beta}$ of β_0 is defined as follows

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^d}{\operatorname{argmax}} \ \ell(\beta),$$

where ℓ is the likelihood function.

$\rightarrow$ if $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, $\hat{\beta}$ is the well known OLS estimator.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Outline

Penalization

If $d >> n$, direct maximization of likelihood can lead to:

- Overfitting: the classifier can only behave well in training set, and can be bad in test set.
- *Unstable*: since empirical risk is data dependent, hence random, small change in the data can lead to very different estimators.

$\rightarrow$ Penalization

$$\underbrace{\hat{L}(\beta)}_{\text{empirical risk}} + \underbrace{\text{pen}(\beta)}_{\text{penalty}}$$

Penalization

If $d >> n$, direct maximization of likelihood can lead to:

- Overfitting: the classifier can only behave well in training set, and can be bad in test set.
- *Unstable*: since empirical risk is data dependent, hence random, small change in the data can lead to very different estimators.

$\rightarrow$ Penalization

$$\underbrace{\hat{L}(\beta)}_{\text{empirical risk}} + \underbrace{\text{pen}(\beta)}_{\text{penalty}}$$

ℓ_0-penalization

- Akaike information criterion (AIC)

$$AIC(\beta) = 2\hat{L}(\beta) + 2\|\beta\|_0,$$

- Bayesian information criterion (BIC)

$$BIC(\beta) = 2\hat{L}(\beta) + 2\log(n)\|\beta\|_0,$$

where $\|\beta\|_0 = card\{j \in 1, \ldots, d, \beta_j \neq 0\}$.

$\star$ Produce interpretable models.

$\rightsquigarrow$ Non-convex optimization.

- $\mathrm{card}(\mathcal{P}(\{1, \ldots, d\})) = 2^d$ models,
- stepwise selection.

ℓ_0-penalization

- Akaike information criterion (AIC)

$$AIC(\beta) = 2\hat{L}(\beta) + 2\|\beta\|_0,$$

- Bayesian information criterion (BIC)

$$BIC(\beta) = 2\hat{L}(\beta) + 2\log(n)\|\beta\|_0,$$

where $\|\beta\|_0 = card\{j \in 1, \ldots, d, \beta_j \neq 0\}$.

$\star$ Produce interpretable models.

$\rightsquigarrow$ Non-convex optimization.

- card$(\mathcal{P}(\{1, \ldots, d\})) = 2^d$ models,
- stepwise selection.

ℓ_1−penalization

Lasso [Tibshirani, 1996]

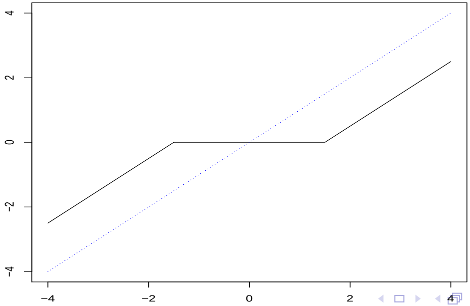
$$\hat{\beta}(\lambda) = \arg\min_{\beta \in R^d} \left\{ \hat{L}(\beta) + \lambda \sum_{j=1}^{d} |\beta_j| \right\}$$

- ⋆ Regularization technique for simultaneous estimation and selection, $\hat{\beta}_{L,j}(\lambda) = 0$ for $j \notin K(\hat{\beta}_L) \subset \{1, \ldots, d\}$.
- ⋆ High dimension $d >> n$.
- ⋆ Convex optimization.

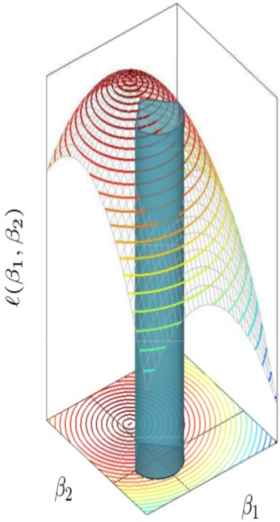
ℓ_1-penalization

- If $\epsilon \sim \mathcal{N}(0, I)$ and $X^T X = I$, for $j \in \{1, \ldots, d\}$,

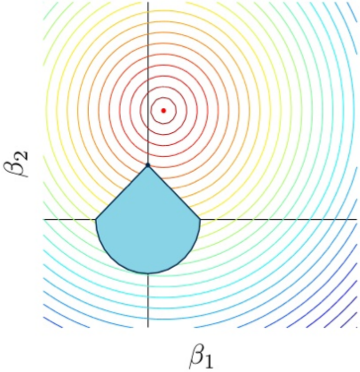
$$\beta_j(\lambda) = \begin{cases} \hat{\beta}_j^{0LS} - \lambda/2 & \textit{if } \hat{\beta}_j^{0LS} > \lambda/2 \\ \hat{\beta}_j^{0LS} + \lambda/2 & \textit{if } \hat{\beta}_j^{0LS} < -\lambda/2 \\ 0 & \textit{otherwise} \end{cases}$$



Geometric view of penalization



$$\max_{\beta} \ell(\beta) \quad s.t. \quad \text{pen}(\beta) \leq \xi$$



Geometry of the Lasso

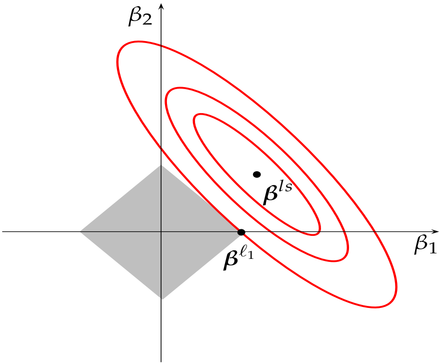


Figure : *Lasso solution.*

Choice of λ

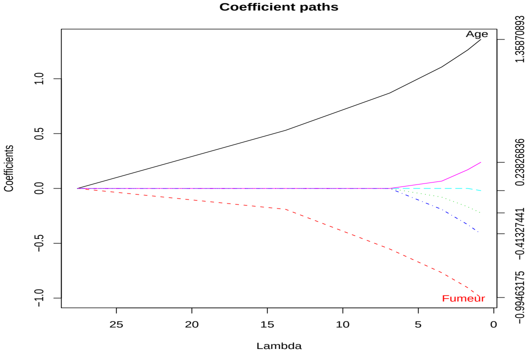


Figure : *Regularization path.*

Choice of λ

- Cross validation which is recommended when the goal is to minimize the prediction error. It can be computationally slow.
- Information criteria such as AIC or BIC can be defined for the Lasso

$$\begin{aligned} AIC(\lambda) &= L_n(\hat{\beta}(\lambda)) + \frac{2}{n} df(\lambda), \\ BIC(\lambda) &= L_n(\hat{\beta}(\lambda)) + \frac{\log n}{n} df(\lambda), \end{aligned}$$

where $df(\lambda)$ is the degrees of freedom of the Lasso for a given parameter λ.

Choice of λ

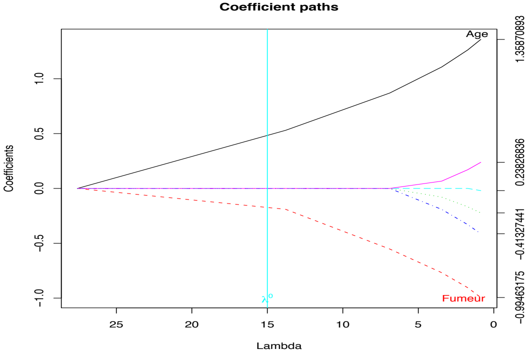


Figure : *Regularization path*

Properties of the Lasso

- Prediction
  Does the Lasso provide a good approximation of $X\beta_0$?
- Estimation
  Does the Lasso provide a good approximation of β_0?
- Selection
  Does the Lasso select the right covariates?
- Sign recovery
  Does the Lasso select the right covariates and identify correctly the signs of their coefficients?

Outline

- Model :

$$y_i = f_0(x_i) + \epsilon_i$$

-

$$\Gamma \subseteq \left\{ \beta \in \mathbb{R}^d, \quad f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.) \right\}.$$

- Oracle inequality

$$R(f_{\hat{\beta}}, f_0) \leq C \times \inf_{\beta \in \Gamma} \left\{ R(f_\beta, f_0) + \Delta_n \right\}. \tag{3}$$

- Model :

$$y_i = f_0(x_i) + \epsilon_i$$

- $$\Gamma \subseteq \left\{ \beta \in \mathbb{R}^d, \quad f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.) \right\}.$$

- Oracle inequality

$$R(f_{\hat{\beta}}, f_0) \leq C \times \inf_{\beta \in \Gamma} \left\{ R(f_\beta, f_0) + \Delta_n \right\}. \tag{3}$$

- Model :

$$y_i = f_0(x_i) + \epsilon_i$$

-

$$\Gamma \subseteq \left\{ \beta \in \mathbb{R}^d, \quad f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.) \right\}.$$

- Oracle inequality

$$R(f_{\hat{\beta}}, f_0) \leq C \times \inf_{\beta \in \Gamma} \left\{ R(f_\beta, f_0) + \Delta_n \right\}. \tag{3}$$

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