Regression shrinkage and selection via the Lasso. Robert Tibshirani, 1996.

François Caron

Department of Statistics, Oxford

October 14, 2014

- Linear regression problem

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, N$$

  - Standardized predictors x_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, p$
  - Centered response variable y_i, $i = 1, \ldots, N$

- Standard approaches
  - Ordinary least square estimate: low bias/high variance, non-interpretable estimates
  - Ridge shrinkage: prediction accuracy but non sparse estimates
  - Subset selection: interpretable but unstable results

- Lasso estimator: achieves both shrinkage (least absolute shrinkage) and sparsity (selection operator)
- Minimize

$$\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq t \qquad (1)$$

or

$$\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (2)$$

- Convex optimization problem

- Orthonormal design case
- X is a $n \times p$ design matrix with $X^T X = I$
- Minimizing

$$\frac{1}{2}(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1$$

equivalent to minimizing

$$\frac{1}{2}(\beta - \widehat{\beta}^0)^T(\beta - \widehat{\beta}^0) + \lambda||\beta||_1$$

where $\widehat{\beta}^0 = X^T y$ is the OLS estimate

- For $j = 1, \ldots, p$

$$\begin{aligned}\beta_j &= \arg\min \frac{1}{2}(\beta_j - \widehat{\beta}_j^0)^2 + \lambda|\beta_j| \\ &= \text{sign}(\widehat{\beta}_j^0)\max(|\widehat{\beta}_j^0| - \lambda, 0)\end{aligned}$$

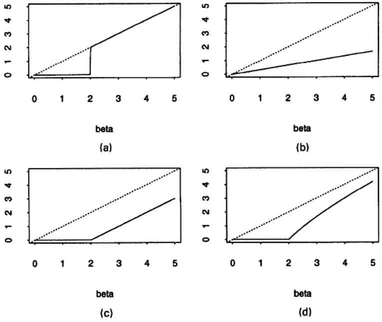


Fig. 1. (a) Subset regression, (b) ridge regression, (c) the lasso and (d) the garotte: ———, form of coefficient shrinkage in the orthonormal design case; ··········, 45°-line for reference

- Geometry of the lasso
- Minimize

$$(\beta - \widehat{\beta}^0)^T X^T X (\beta - \widehat{\beta}^0) \text{ subject to } ||\beta||_1 \leq t$$

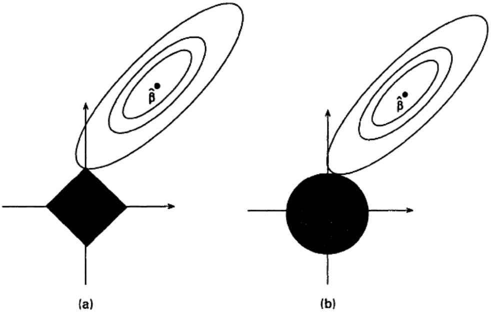
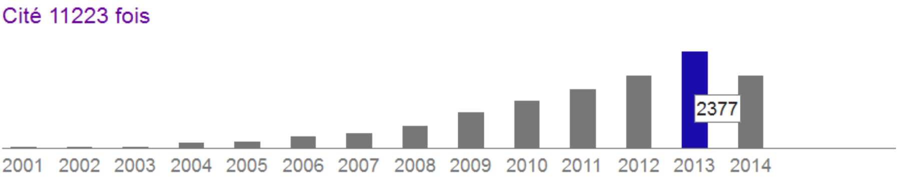


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression



- Enormous influence
- High dimensional problems (large p small n)
- Compressed sensing
- Various extensions: generalized linear models, sparse graphs, group/fused lasso, matrix completion...

Related work

- Non-negative garotte by Breiman (1993)
- Bridge regression by Frank and Friedman (1993)
- Basis pursuit by Chen, Donoho, Saunders (1998)

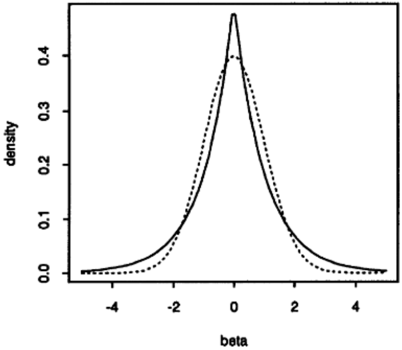
Algorithm

- Quadratic program solver
- Does not scale very well
- LARS algorithm (Efron et al. 2002) provides an efficient way of solving the lasso problem

Bayesian interpretation

- Maximum a posteriori estimate under a Laplace prior

$$p(\beta_j) = \lambda \exp(-\lambda|\beta_j|)$$



Bayesian interpretation

- Laplace distribution is a scale mixture of Gaussians

$$\begin{aligned} \beta_j | \tau_j &\sim \mathcal{N}(0, \tau_j) \\ \tau_j &\sim \text{Exp}(\lambda^2/2) \end{aligned}$$

- Suggests iterative Expectation-Maximization algorithm for solving lasso
- Repeat until convergence

E step: $V^{(k)} = \text{diag}\left(\frac{\lambda}{|\beta_1^{(k-1)}|}, \dots, \frac{\lambda}{|\beta_p^{(k-1)}|}\right)$

M step: $\beta^{(k)} = (V^{(k)} + X^T X)^{-1} X^T y$

Proposed project

- Code the EM algorithm to solve the Lasso problem
- Reproduce the lasso results on the prostate data (available in R)