Spline LASSO with thresholding in high-dim regression

Bing-Yi Jing

June 10, 2013

Outline

High Dimensional Problems with Ordered Features

- Small n and large p problem: $p \gg n$
- Sparsity:
  - Number of influential features k is small.
- Feathers are ordered, and correlated.
- Examples:
  - protein mass spectroscopy data
  - gene expression data

Objectives in model selection:

- Model sparsity
- Feathers (variables) selection
- Prediction power

Linear regression model

- Given $(Y_i, x_{i1}, ..., x_{ip})$, $i = 1, ..., n$, assume

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n.$$

- In matrix form,

$$Y = X\beta + \epsilon$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & a_{np} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

One Example: protein mass spectroscopy data, Adam et al (2003)

For each blood serum sample i,

- x_{ij}: intensity for many *time-of-flight* value t_j
- *time-of-flight*: related to mass over charge ratio m/z
- $p = 48538$ m/z-sites
- $n_1 = 157$ healthy patients, $n_2 = 167$ with cancer.

Objective:

- find m/z-sites discriminating between 2 groups.

Least square estimate (LSE)

- LSE:
$$\hat{\beta}^{lse} = \{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} = (X^T X)^{-1} XY$$
- ill-posed if $p > n$

Ridge regression

- Ridge Regression: (Hoerl and Kennard, Technometrics, 1970)

$$\begin{aligned}\hat{\beta}^{ridge} &= argmin\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= (X^T X + \lambda I)^{-1} XY\end{aligned}$$

- When $X^T X = I$, then

$$\hat{\beta}^{ridge} = \frac{\hat{\beta}^{lse}}{1+\lambda}$$

- L_1 penalty does shrinkage of LSE,
- but no variable selection.

LASSO (least absolute shrinkage and selection operator)

- Lasso (Tibshirani (1996), JRSSB):

$$\hat{\beta}^{lasso} = argmin\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- When $X^T X = I$, then

$$\hat{\beta}^{lasso} = sgn(\hat{\beta}^{lse}) \left(|\hat{\beta}^{lse}| - \lambda\right)^+$$

- L_1 penalty does shrinkage and variable selection simultaneously.
- It works for $p > n$ as well.
- Computation: LARS (Efron et al, 2004), very efficient

Variants of LASSO

Method	Reference	Detail
Elastic Net	Zou and Hastie(2005)	$\lambda \sum \beta_j^2$
Fused Lasso	Tibshirani *et al*.(2005)	$\lambda \sum \lvert\beta_{j+1} - \beta_j\rvert$
Adaptive Lasso	Zou(2006)	$\lambda \sum w_j \lvert\beta_j\rvert$
Grouped Lasso	Yuan and Lin(2007)	$\sum_g \lvert\lvert\beta_g\rvert\rvert_2$
Dantzig selector	Candes and Tao(2007)	$min\{\lvert\lvert X^T(y - X\beta)\rvert\rvert_\infty\}\lvert\lvert\beta\rvert\rvert_1 < t$

Fused LASSO

- Fused lasso (Tibshirani *et al*.(2005)):

$$\hat{\beta} = argmin\{\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|\}$$

- Equivalently,

$$\hat{\beta} = argmin\{\sum_{i}(y_i - \sum_{j}x_{ij}\beta_j)^2\}$$

subject to $\sum_{j=1}^{p}|\beta_j| \leq s_1$ *and* $\sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \leq s_2$.

- 1st penalty $\Longrightarrow$ sparse β_j, while 2nd penalty $\Longrightarrow$ flatness of β_j. (i.e., maintain grouping effects and sparsity of the coefficients).
- Fussed lasso picks grouped ordered features.

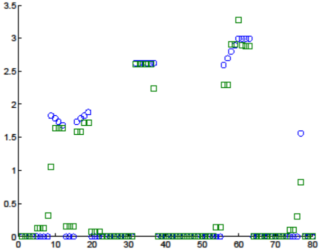

Figure: $N = 50, p = 80, X \in N(0, \Sigma)$ where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.9^{|i-j|}$.

Drawbacks:

- Hard to keep shape of β_j's within the same group.
- Computation:
  - very intensive for large p,
  - not fully automatic

Smooth LASSO (Hebiri, M. and Van de Geer, S.(2011))

$$\hat{\beta} = argmin\{\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2\}$$

Advantages:

- Capture the smooth change in features in a group
- Computation efficient

Disadvantages:

- Cannot capture changes in curvature.
- Less prediction power for large p,

Spline LASSO

$$\hat{\beta} = argmin\{\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p-1}(\beta_{j-1} - 2\beta_j + \beta_{j+1})^2\}$$

- The first penalty encourages sparse solution.
- The second penalty mimics cubic spline, i.e. penalizing large second-order derivatives of coefficients.

Computation

Spline LASSO can be solved by LARS.

Proposition

Given dataset (Y, X) *and* (λ_1, λ_2), *define an artificial data set* (Y^*, X^*) *by*

$$X^*_{(N+p-2)\times p} = \left(X,\ \sqrt{\lambda_2}L\right)^T,\ Y^*_{(N+p-2)} = (Y,\ \mathbf{0})^T.$$

where L *is a* $(p-2)\times p$ *matrix with* $L_{i,i} = L_{i,i+2} = 1, L_{i,i+1} = -2$ *and* $L_{i,j} = 0$ *otherwise. Then the spline lasso optimization can be written as*

$$(Y^* - X^*\beta)^T(Y^* - X^*\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j|,$$

which is an equivalent lasso problem and can be solved efficiently.

Simulations

- Model:
$$Y = X\beta + Z,$$
where $(X_1, \ldots, X_p)^T \in N(0, \Sigma)$, and $Z \in N(0, \sigma)$ where $\sigma \in [0.1, 1]$.
- $p > n$.
- β's are generated from a continuous function with k non-zero terms, ($k > n$ or $k < n$).
- Comparison in variable selection and prediction:
  - fused-lasso,
  - smooth-lasso,
  - spline,
  - spline-lasso

Case 1: $N = 60, p = 200, k = 50, \sigma = \frac{1}{10}$ and neighboring correlation is high

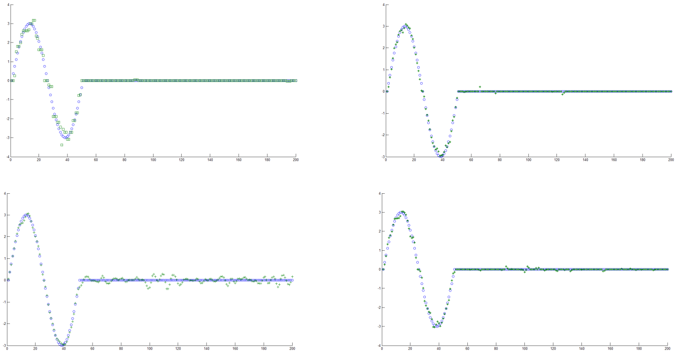

Figure: $X \sim N(0, \Sigma)$ where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.9^{|i-j|}$.

Case 2: $N = 60, p = 300, k = 70, \sigma = \frac{1}{5}$ and neighboring correlation is median

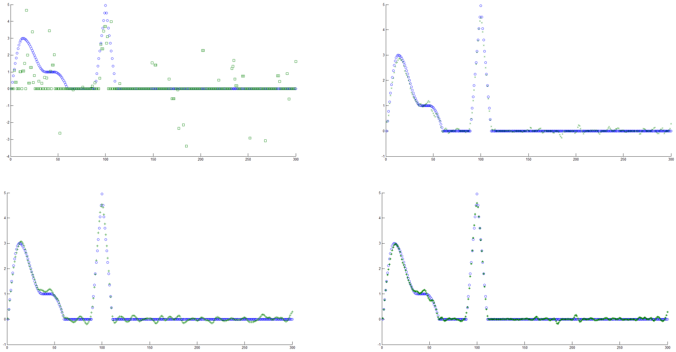

Figure: $X \sim N(0, \Sigma)$ where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.5^{|i-j|}$.

Case 3: $N = 60, p = 500, k = 70, \sigma = 1$ and all features are i.i.d.

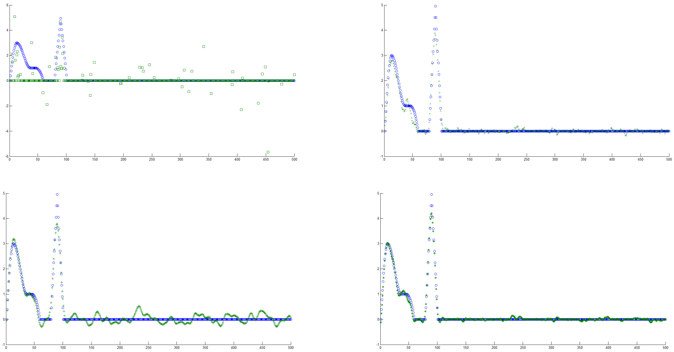

Figure: $X \sim N(0, I)$.

Prediction Accuracy

- We generated 1000 testing data to compare the prediction accuracy of the above estimates and a summary is given as follow:

MSE	Fused Lasso	Smooth Lasso	Spline OLS	Spline Lasso
Case 1	73.9222	52.7745	64.8594	50.2912
Case 2	201.9074	63.1969	59.2204	48.2225
Case 3	258.1292	106.0806	133.276	63.4307

Estimation Accuracy

- We also summarize the L_2 norm of the difference between these estimates and the true β:

$\|\|\hat{\beta} - \beta\|\|_2$	Fused Lasso	Smooth Lasso	Spline OLS	Spline Lasso
Case 1	2.1621	1.3238	1.7614	1.2262
Case 2	15.9208	1.7069	1.5989	1.1407
Case 3	17.5091	3.1576	4.0761	1.747

Why Threshold?

- Spline LASSO is good at prediction, but not quite sparse.

Specificity	Fused Lasso	Smooth Lasso	Spline OLS	Spline Lasso
Case 1	0.9603	0.7881	0	0.7483
Case 2	0.8371	0.4208	0	0.1674
Case 3	0.9121	0.3682	0	0.1378

- WHY? Lasso penalty (λ_1) is often quite small.
- Solution: Applying thresholding after estimation.
- Thresholding can also be applied to smooth lasso and spline OLS, etc.

Level of thresholding?

- thresholding level $\pm\hat{\sigma}\sqrt{2\log(N)}$, where $\hat{\sigma}$ is standard error of small coefficients.
- In wavelet, Donoho and Johnstone (1998)
- Theoretical study is under investigation.

Applying the Threshold

The specificity improved a lot while sensitivity was barely affected.

Specificity	Smooth Lasso	Smooth Lasso Thr	Spline OLS	Spline OLS Thr	Spline Lasso	Spline Lasso Thr
Case 1	0.7881	0.9603	0	1	0.7483	0.9669
Case 2	0.4208	0.9683	0	0.9864	0.1674	0.9729
Case 3	0.3682	0.9549	0	0.9786	0.1378	0.9501
Sensitivity	Smooth Lasso	Smooth Lasso Thr	Spline OLS	Spline OLS Thr	Spline Lasso	Spline Lasso Thr
Case 1	0.9592	0.9388	1	0.9184	0.9796	0.9592
Case 2	0.9873	0.9747	1	0.9873	1	0.9873
Case 3	0.9873	0.962	1	0.9241	1	0.9747

Prediction and estimation also improved:

MSE	Smooth Lasso	Smooth Lasso Thr	Spline OLS	Spline OLS Thr	Spline Lasso	Spline Lasso Thr
Case 1	52.7745	52.7739	64.8594	42.3359	50.2912	49.698
Case 2	63.1969	59.1765	59.2204	47.7498	48.2225	45.1844
Case 3	106.0806	104.24	133.276	96.0668	63.4307	61.1429

$\|\|\hat{\beta} - \beta\|\|_2$	Smooth Lasso	Smooth Lasso Thr	Spline OLS	Spline OLS Thr	Spline Lasso	Spline Lasso Thr
Case 1	1.3238	1.3219	1.7614	0.8796	1.2264	1.1983
Case 2	1.7069	1.546	1.5989	1.12	1.1407	1.0048
Case 3	3.1576	3.07	4.0761	2.7968	1.747	1.6269

Summary

- Spline lasso captures smooth-changing feathers
- It works well in high dimensional settings with ordered features.
- Prediction is better than existing methods.
- Thresholding improves variable selection, prediction and estimation.

Reference

- Candes, E. and Tao, T. (2007) The dantzig selector statistical estimation when *p* is much larger than *n*. Ann. Statist., 35, 2313-2351.
- David L. Donoho and Iain M. Johnstone. (1998) Minimax Estimation via Wavelet Shrinkage. Ann. Statist., 26, 879-921.
- Hebiri, M. and Van de Geer, S. (2011) The Smooth-Lasso and Other $\ell_1 + \ell_2-$penalized methods. Electron. J. Statist. 5, 1184-1226.
- Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. J. R. Statist. Soc. B, 58, 267-288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. J. R. Statist. Soc. B, 67, 91-108.

Reference

- Yuan, M. and Lin, Y. (2007) Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B, 68, 49-67.
- Zou, H. (2006) The adaptive lasso and its oracle properties. J. Am. Statist. Ass., 101, 1418-1429.
- Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. J. R. Statist. Soc. B, 67, 301-320.