Learning from sparsity

Can Yang

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

It is my great honor to share my reading with you. Most of the materials are from Stanford.

Outline

1. Introduction
2. Beyond ℓ_1 regularization
   - Generalized Shrinkage operators
   - Comparison study between ℓ_1 and ℓ_0
3. Application: outlier detection
   - Outlier detection in regression
   - Outlier detection in low-rank representations
4. Other Lasso-type problems
5. Reference

High dimensional data: $p >> n$

High dimensional data are coming ...

- Genome-wide association studies: $p = 500K$ SNPs, $n = 5000$ case-control subjects.
- Microarray studies: $p = 40K$ genes, $n = 100$ subjects.
- Image sequence analysis: $p = 60K$ pixels, $n = 100$ frames.
- Proteomics: ...
- Social networks: ...

A brief history of ℓ_1 regularization

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

- Wavelet shrinkage: Donoho and Johnstone (1994).
- Lasso for linear regression: in statistics (Tibshirani,1995); motivated by nonnegative garotte (Breiman, 1994).
- Basis Pursuit: in signal processing (Chen, Donoho and Saunders 1996).
- Extension to generalized linear models: e.g., logistic regression and so on.
- Structured sparsity: e.g., fused-Lasso, group-Lasso, elastic net, graphical Lasso and so on.
- Compressed Sensing: near exact recovery of sparse signals in very high dimensions (Donoho 2004, Candes and Tao 2005) – ℓ_1 is a good surrogate for ℓ_0 in many cases.
- Low-rank approximation: from vectors to matrices.

History of algorithm for ℓ_1 regularization

Solution Path for $\beta(\lambda)$

$$\min_{\boldsymbol{\beta}} \sum_{i}^{n} \left(y_i - \beta_0 - \sum_{j}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j}^{p} |\beta_j| \tag{2}$$

- Least angle regression (LARS): (Efron et al., 2002).
- Coordinate Decent algorithm (CD): (Friedman et al, 2006).
- Alternating Direction Method of Multipliers (ADMM) (Boyd et al, 2010): ADMM is equivalent or closely related to many other algorithms, such as dual decomposition, augmented Lagrangian multipliers, (Split) Bregman iterative algorithms, proximal methods.
- Gradient methods such as Nesterov's method (optimal convergence rate), gradient projection etc.

Citation counts from ISI Web of Knowledge

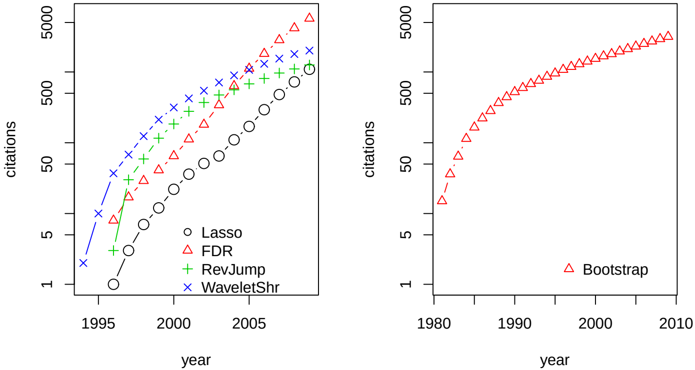


Figure: Left: Lasso, False discovery rate, Reversible jump MCMC, Wavelet shrinkage. Right: Bootstrap. See P. Buhlmann, Regression shrinkage and selection via the Lasso: a retrospective 2011. Also see `http://sciencewatch.com/dr/tt/2009/09-octtt-COM/`

Other regularization

Regularization makes fitting linear models ($p > n$) well-posed.

- Forward stepwise regression: It adds variables one at a time, refit the model with current variables and stops when overfitting is detected. This is a greedy algorithm. In signal processing, it is known as "Orthogonal Match Pursuit" (OMP).
- Forward stagewise regression: It adds variables one at a time without refitting the model, and stops when overfitting is detected. It is known as "Match Pursuit" (MP) in signal processing while known as "Boosting" in statistics.
- Best-subset regression: exhaustive search all subsets (can only be applied when p is small).
- Ridge regression:

$$\min_{\boldsymbol{\beta}} \sum_{i}^{n} \left(y_i - \beta_0 - \sum_{j}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j}^{p} \beta_j^2 \tag{3}$$

Lasso vs. Ridge regression

Lasso significantly outperforms Ridge regression in sparse settings.

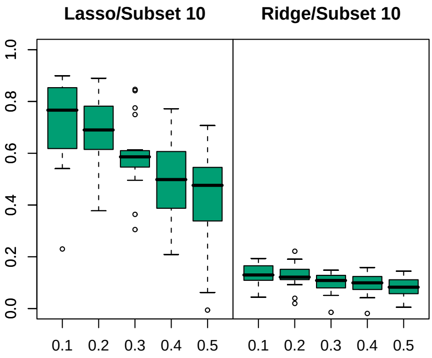


Figure: The design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is given by independent gaussian variables. Here $n = 50, p = 300$. The response variable $\mathbf{y}$ is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\beta}$ has only ten nonzero coefficients and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma)$. The x-axis is Noise-to-Signal Ratio (NSR = $\sigma / \sqrt{\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}}$) and the y-axis is variance explained (evaluated on independent data with large samples).

The “bet on sparsity” principle

Friedman et al, 2004

Using a procedure that does well in sparse problems, since no procedure does well in dense problems.

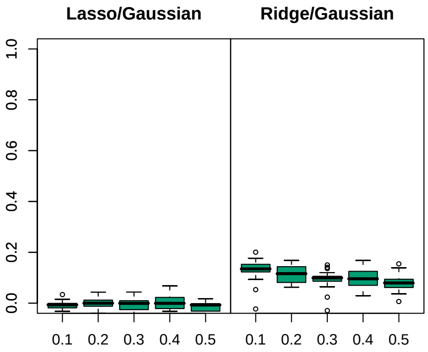


Figure: Simulation similar with previous one except that all coefficients in β are nonzero.

Theoretical results

- $K = \{k : \beta_k \neq 0\}$ indexes the set of relevant variables.
- $N = \{1, \ldots, p\} \backslash K$ indexes the irrelevant variables.

The irrepresentable condition (Zhao and Yu, 2006)

$$\| \underbrace{\mathbf{X}_N^T \mathbf{X}_K (\mathbf{X}_K^T \mathbf{X}_K)^{-1}}_{\in \mathbb{R}^{|N| \times |K|}} \operatorname{sign}(\beta_K) \|_\infty < 1. \quad (4)$$

When the signs are unknown, we need

$$\max_{j \in N} \| \mathbf{X}_j^T \mathbf{X}_K (\mathbf{X}_K^T \mathbf{X}_K)^{-1} \|_1 < 1. \quad (5)$$

- The above condition says that the least squares coefficients for the columns of $\mathbf{X}_N$ on $\mathbf{X}_K$ are not large, that is, the relevant variables are not too highly correlated with the irrelevant variables.

Failure of Lasso as a variable selector

When the representable condition is not satisfied, Lasso can fail as a variable selector.

A simple example (Zhao and Yu, 2006)

$$\begin{aligned} &\mathbf{X}_1 \sim N(0,1); \mathbf{X}_2 \sim N(0,1); \mathbf{e} \sim N(0,1) \\ &\mathbf{X}_3 = \frac{2}{3}\mathbf{X}_1 + \frac{2}{3}\mathbf{X}_2 + \frac{1}{3}\mathbf{e}; \\ &\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon; \epsilon \sim N(0,1) \end{aligned} \tag{6}$$

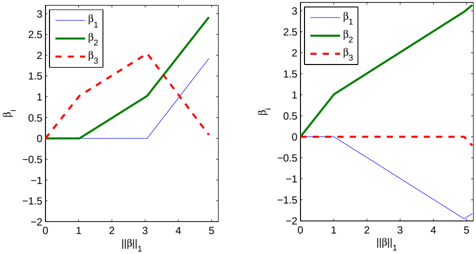
$$n^{-1}\mathbf{X}_K^T\mathbf{X}_K = \mathbf{I};$$

$$n^{-1}\mathbf{X}_N^T\mathbf{X}_K = [\frac{2}{3}\ \frac{2}{3}];$$

$$\mathbf{X}_N^T\mathbf{X}_K(\mathbf{X}_K^T\mathbf{X}_K)^{-1} = [\frac{2}{3}\ \frac{2}{3}].$$

$$(a)\mathbf{X}_N^T\mathbf{X}_K(\mathbf{X}_K^T\mathbf{X}_K)^{-1}[2\ 3]^T = \frac{10}{3} > 1;$$

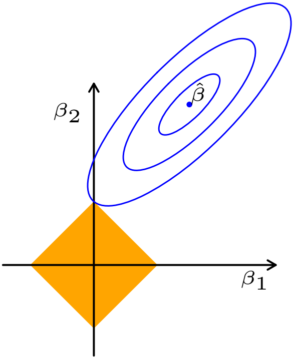
$$(b)\mathbf{X}_N^T\mathbf{X}_K(\mathbf{X}_K^T\mathbf{X}_K)^{-1}[-2\ 3]^T = \frac{2}{3} < 1.$$



(a) $\beta_1 = 2, \beta_2 = 3$ (b) $\beta_1 = -2, \beta_2 = 3$

Outline

From ℓ_1 to non-convex penalties



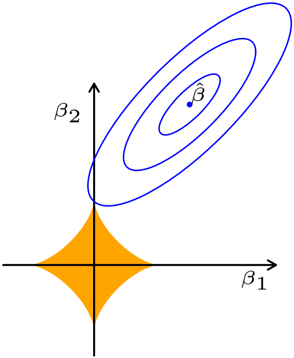


Figure: Shown are ℓ_1 (Lasso) and ℓ_γ penalty $\sum_j |\beta_j|^\gamma \leq t$ with $\gamma = 0.7$. Note that ℓ_0 regularization corresponds to best-subset selection.

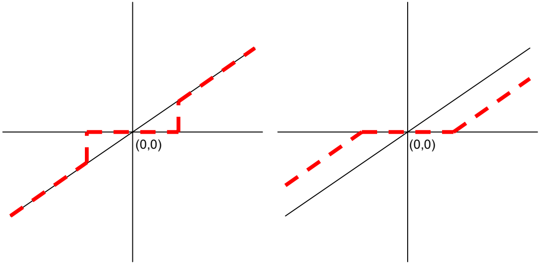
Thresholding operators

The hard thresholding operator $S_H(\cdot;\lambda)$

$$\beta = S_H(\tilde{\beta},\lambda) = \arg\min\left(\frac{1}{2}(\tilde{\beta}-\beta)^2 + \frac{1}{2}\lambda^2|\beta|_0\right) = \tilde{\beta}\,\mathbb{I}(|\tilde{\beta}| > \lambda). \qquad (7)$$

The soft thresholding operator $S(\cdot;\lambda)$

$$\beta = S(\tilde{\beta},\lambda) = \arg\min\left(\frac{1}{2}(\tilde{\beta}-\beta)^2 + \lambda|\beta|\right) = sign(\tilde{\beta})(|\tilde{\beta}|-\lambda)_+. \qquad (8)$$



Coordinate Decent for Lasso. I

Suppose $\mathbf{y}$ has been centered and $\mathbf{X}_j$ has been standardized (i.e., $\sum_i x_{ij}^2 = 1$).

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{9}$$

Suppose we are solving β_j and other $\beta_{k \neq j}$ are fixed. Rearrange the above equation:

$$\min R(\beta_j) = \min_{\beta_j} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(\underbrace{y_i - \sum_{k \neq j}^{p} x_{ik}\beta_k}_{\tilde{y}_i^{(j)}} - x_{ij}\beta_j \right)^2 + \lambda \sum_{k \neq j}^{p} |\beta_k| + \lambda |\beta_j| \right\} \tag{10}$$

Coordinate Decent for Lasso. II

Minimizing R w.r.t β_j

$$\frac{\partial R}{\partial \beta_j} = \sum_i \left(-x_{ij}(y_i - \tilde{y}_i^{(j)} - x_{ij}\beta_j) \right) + \lambda \text{sign}(\beta_j) = 0.$$

we have closed-form solution given by soft thresholding:

$$\beta_j = S(\sum_i x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda) \tag{11}$$

Algorithm

Initialize all $\beta_j = 0$. Cycle over $j = 1, 2, \ldots, p, 1, 2, \ldots$ till convergence:

- Compute $\tilde{y}_i^{(j)} = \sum_{k \neq j}^{p} x_{ik}\beta_k$;
- $\beta_j \leftarrow S(\sum_i x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda)$.

Speed of coordinate decent – real data sets

data sets	n	p	CD	An interior method
		Dense		
Cancer	144	16,063	2.5 mins	NA
Leukemia	72	3571	2.50s	55.0s
		Sparse		
Internet Ad	2359	1430	5.0s	20.9s
Newsgroup	11,314	777,811	2 mins	3.5 hrs

Table: The interior method uses $\ell_1 -$ *logreg* designed by Prof. Boyd's group. For Cancer, Leukemia and Internet-Ad, times are for ten-fold cross-validation over 100 λ values; for Newsgroup time is for a single run with 100 values of λ.

When coordinate decent works?

Consider

$$f(\beta_1, \ldots, \beta_p) = g(\beta_1, \ldots, \beta_p) + \sum_{j=1}^{p} h_j(\beta_j) \tag{12}$$

Coordinate decent converges to the global optimum if

- $g(\cdot)$ is differentiable and convex.
- $h_j(\beta_j)$ is convex.
- Here each β_j can be a vector, but β_j and β_k cannot have any overlapping members.

For example, CD does not work for $\sum_i \frac{1}{2}(y_i - \beta_i)^2 + \lambda \sum_i |\beta_i - \beta_{i-1}|$.

J. Friedman et al.
Pathwise coordinate decent.
Annals of applied statistics, 2007.

Generalized thresholding operators (GTO)

Generalized thresholding operators (Definition)

$$S_\gamma(\beta, \lambda) = \arg\min_\beta Q(\beta) = \frac{1}{2}(\beta - \tilde{\beta})^2 + \lambda P(|\beta|, \lambda, \gamma) \tag{13}$$

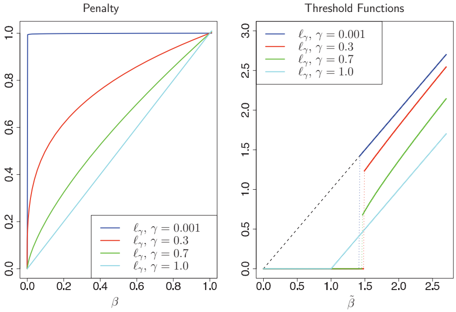
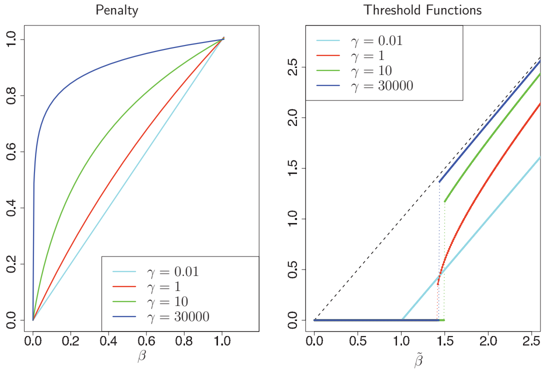


Figure: The ℓ_γ penalty: $\lambda P(t, \lambda, \gamma) = \lambda |t|^\gamma$

The log penalty

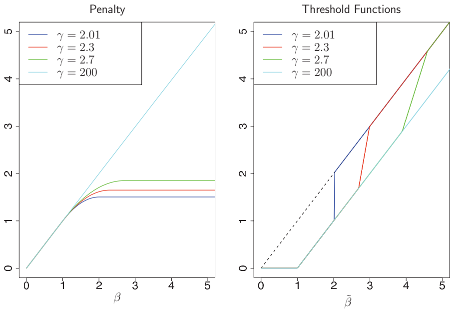
$$\lambda P(t, \lambda, \gamma) = \frac{\lambda}{\log(\gamma + 1)} \log(\gamma|\beta| + 1), \gamma > 0 \qquad (14)$$

- ℓ_1 ($\gamma \to 0+$); ℓ_0 ($\gamma \to +\infty$).



The SCAD penalty

$$
\begin{aligned}
\frac{d}{dt}P(t,\lambda,\gamma) &= \mathbb{I}(t \leq \lambda) + \frac{(\gamma\lambda - t)_+}{(\gamma-1)\lambda}\mathbb{I}(t > \lambda) \text{ for } t > 0, \gamma > 2 \\
P(t,\lambda,\gamma) &= P(-t,\lambda,\gamma) \\
P(0,\lambda,\gamma) &= 0
\end{aligned}
\tag{15}
$$



The MC+ penalty

$$\begin{aligned}\lambda P(t,\lambda,\gamma) &= \lambda \int_0^{|t|} (1 - \frac{x}{\gamma\lambda})_+ dx \\ &= \lambda(|t| - \frac{t^2}{2\lambda\gamma} \mathbb{I}(|t| < \lambda\gamma) + \frac{\lambda^2\gamma}{2} \mathbb{I}(|t| \geq \lambda\gamma).\end{aligned} \quad (16)$$

- $\gamma \to \infty$ (Soft threshold operator); $\gamma \to 1+$ (Hard threshold operator).

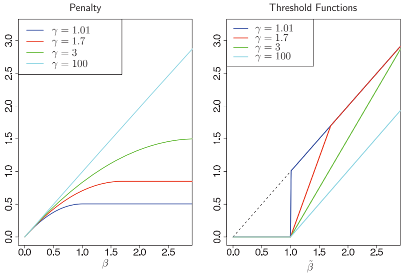


Figure: known as “firm shrinkage” in signal processing (Gao and Bruce, 1997).

Desirable properties for a family of threshold operators

The family of threshold operators

$S_\gamma(\cdot, \lambda) : \mathbb{R} \rightarrow \mathbb{R} \quad \gamma \in (\gamma_0, \gamma_1)$.

- $\gamma \in (\gamma_0, \gamma_1)$ should bridge the gap between soft and hard threholding.
- The map $\tilde{\beta} \rightarrow S_\lambda(\tilde{\beta}, \lambda)$ should be continuous (Strict convexity of $Q(\beta)$ implies this).
- The function $\gamma \rightarrow S_\lambda(\tilde{\beta}, \lambda)$ should be continuous on $\gamma \in (\gamma_0, \gamma_1)$.

Illustration via the MC+ penalty

$$Q(\beta) = \frac{1}{2}(\beta - \tilde{\beta})^2 + \lambda \int_0^{|t|} (1 - \frac{x}{\gamma\lambda})_+ dx \tag{17}$$

The thresholding function is given by

$$S_\gamma(\tilde{\beta}, \lambda) = \left\{ \begin{array}{ll} 0, & \text{if } |\tilde{\beta}| \leq \lambda; \\ \text{sign}(\tilde{\beta}) \left(\frac{|\tilde{\beta}| - \lambda}{1 - \frac{1}{\gamma}} \right), & \text{if } \lambda < |\tilde{\beta}| \leq \lambda\gamma \\ \tilde{\beta} & \text{if } |\tilde{\beta}| > \lambda\gamma \end{array} \right. . \tag{18}$$

We have

$$\begin{aligned} \gamma \to 1+, &\quad S_\gamma(\tilde{\beta}, \gamma) \to S_H(\tilde{\beta}, \lambda), \\ \gamma \to \infty, &\quad S_\gamma(\tilde{\beta}, \gamma) \to S(\tilde{\beta}, \lambda). \end{aligned} \tag{19}$$

Coordinate Decent using GTO (SparseNet)

- Input a grid of increasing λ values $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$, and a grid of increasing γ values $\Gamma = \{\gamma_1, \ldots, \gamma_K\}$, where γ_K indexes the Lasso penalty. Define λ_{L+1} such that $\boldsymbol{\beta}_{\gamma_K, \lambda_{L+1}} = \mathbf{0}$.
- For each value of $l \in \{L, L-1, \ldots, 1\}$ repeat the following
  - Initialize $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\gamma_K, \lambda_{l+1}}$.
  - For each value of $k \in \{K, K-1, \ldots, 1\}$ repeat the following
    - Cycle through $j = 1, \ldots, p, 1, \ldots, p, \ldots$

$$\tilde{\beta}_j = S_{\gamma_k}\Big(\sum_i x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda_l\Big) \qquad (20)$$

      where $\tilde{y}_i^{(j)} = \sum_{k \neq j} x_{ik} \tilde{\beta}_k$, until the updates converge to $\boldsymbol{\beta}^*$.
    - $\hat{\boldsymbol{\beta}}_{\gamma_k, \lambda_l} \leftarrow \boldsymbol{\beta}^*$.
  - $k \leftarrow k - 1$.
- $l \leftarrow l - 1$.
- Return the 2D solution surface $\boldsymbol{\beta}_{\lambda,\gamma}, \quad (\lambda, \gamma) \in \Lambda \times \Gamma$.

Outline

1. Introduction
2. Beyond ℓ_1 regularization
   - Generalized Shrinkage operators
   - Comparison study between ℓ_1 and ℓ_0
3. Application: outlier detection
   - Outlier detection in regression
   - Outlier detection in low-rank representations
4. Other Lasso-type problems
5. Reference

Simulation studies of linear Regression

Settings

- Linear model: $\mathbf{y} = \mathbf{X}\beta + \epsilon$, $\epsilon \sim N(0, \sigma)$.
- Covariance matrix $\Sigma(\rho; m)$: a $m \times m$ matrix with 1's on the diagonal, and ρ's on the off-diagonal.
- SNR (signal-to-noise-ratio): $\mathrm{SNR} = \frac{\sqrt{\beta^T \Sigma \beta}}{\sigma}$
- Prediction Error: $\frac{E(\mathbf{X}\beta - \mathbf{X}\hat{\beta})^2}{\sigma^2}$.
- Percentage of non-zeros coefficients.
- the mis-identification of the true non-zero coefficients (0-1 Error)

Small p

- p1: $n = 35, p = 30, \Sigma(0.4; p)$ and
  $\beta = (-0.033, -0.067, -0.1, \mathbf{0}_{1\times 23}, -0.9, 0.93, -0.97, 0)$.
- p2: $n = 35, p = 30, \Sigma(0.4; p)$ and
  $\beta = (0.033, 0.067, 0.1, \mathbf{0}_{1\times 23}, 0.9, 0.93, 0.97, 0)$.

Comparison among Stepwise regression (st), best subset regression (bs), Lasso (las) and SparseNet for different SNRs.

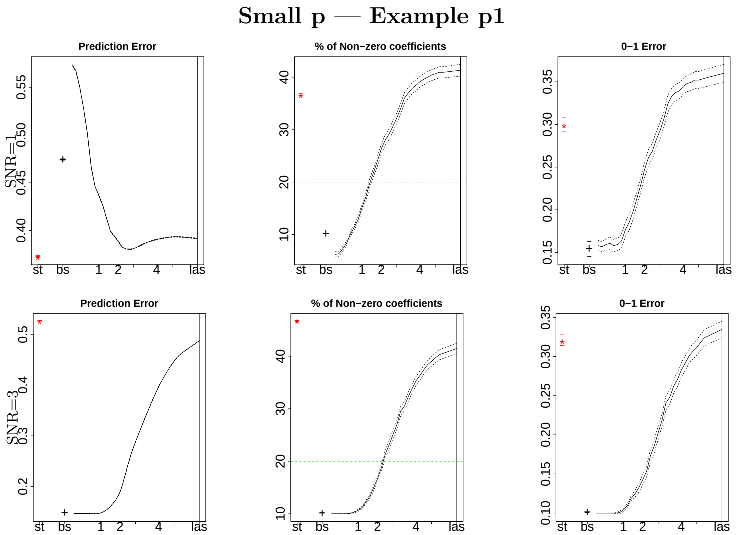
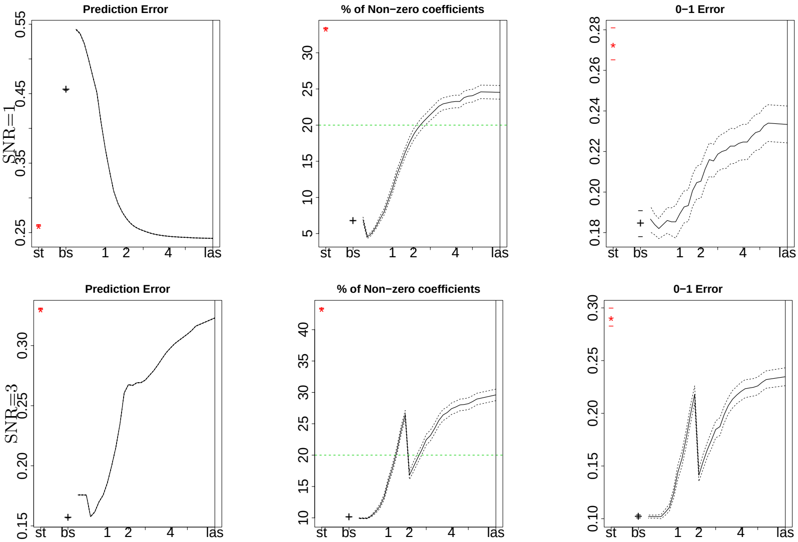


Figure: The x-axis is γ shown on the log scale.

Small p — Example p2

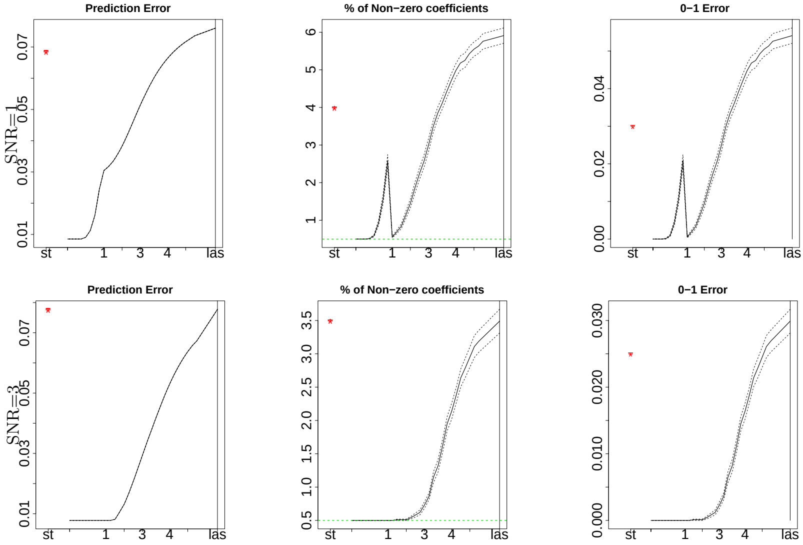


Large p

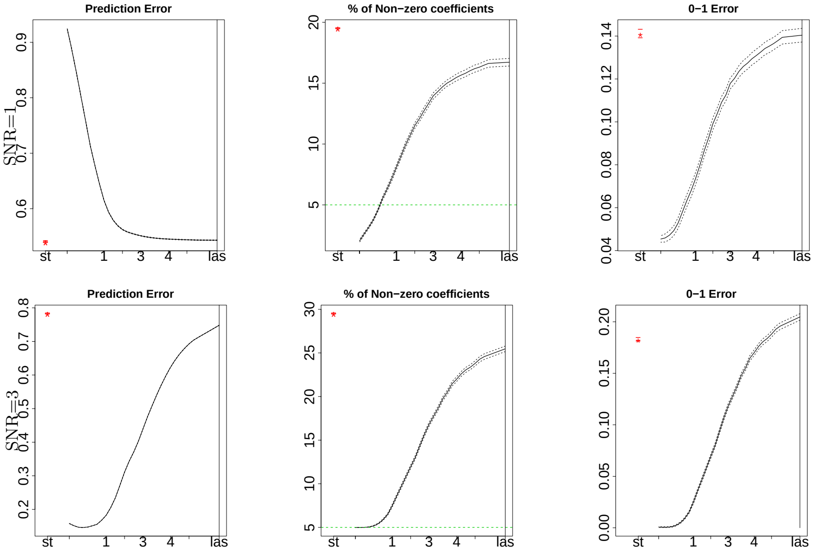
- P1: $n = 100, p = 200, \Sigma(0.004; p)$ and $\boldsymbol{\beta} = (.1, \mathbf{0}_{1\times 199})$.
- P2: $n = 100, p = 200; \Sigma^{\rho}_{p\times p} = ((0.7^{|i-j|}))1 \le i, j \le p$ and β has 10 non-zeros such that $\beta_{(20\times i)+1} = 1, i = 0, 1, \ldots, 9$; and $\beta_i = 0$ otherwise.
- P3: $n = 100, p = 200, \Sigma(0.5; p)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{10}, \mathbf{0}_{1\times 190})$, $\beta_1, \beta_2, \ldots, \beta_{10}$ form an equi-spaced grid on [0,0.5].
- P4: $n = 100, p = 200, \Sigma(0; p)$ and $\boldsymbol{\beta} = (\mathbf{1}_{1\times 20}, \mathbf{0}_{1\times 180})$.

Comparison among Stepwise regression (st), Lasso (las) and SparseNet for different SNRs.

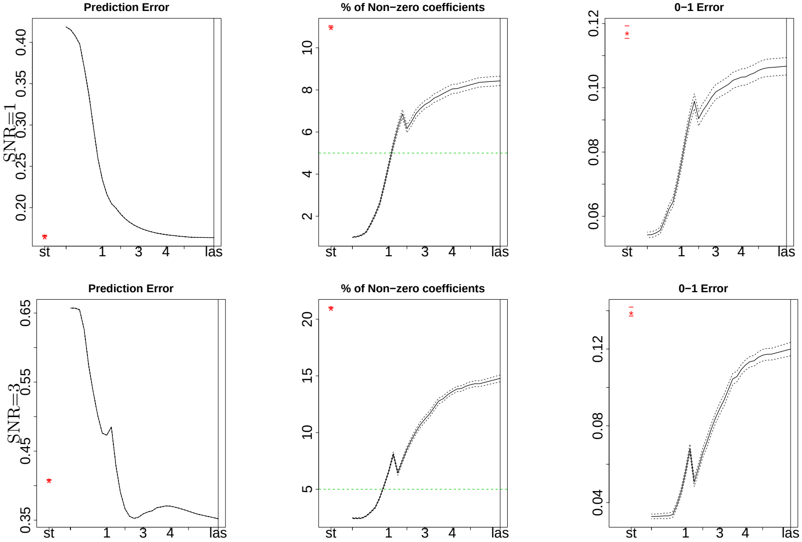
Large p — Example P1



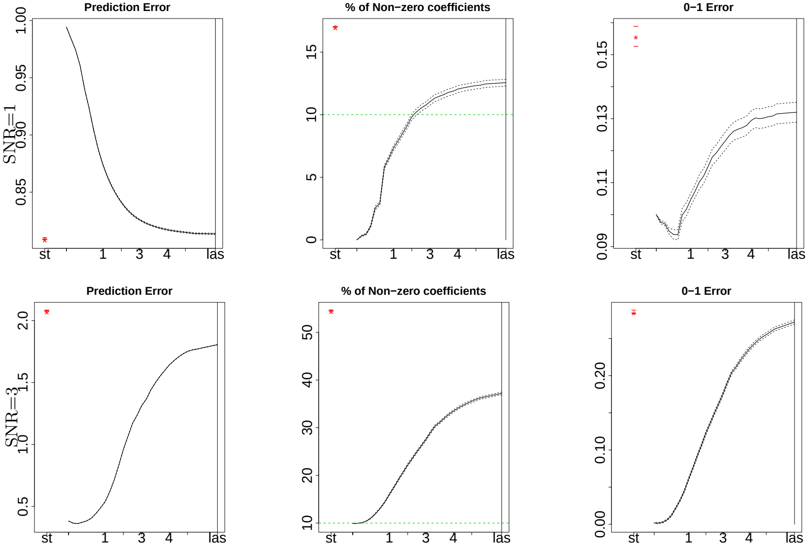
Large p — Example P2



large p — Example P3



large p — Example P4



Summary

In terms of prediction accuracy

- ℓ_0 often performs better when SNR is high.
- ℓ_1 often performs better when SNR is low.
- ℓ_0 can be better than ℓ_1 when the signal is extremely sparse (see large p: P1).
- The performances of ℓ_0 and ℓ_1 also depend on Σ.

In terms of identifying interesting variables

- ℓ_0 always lead to less variables selected in the model.
- ℓ_0 always has less 0-1 Errors than ℓ_1.

Matrix completion

$$\min_Z \frac{1}{2}\|P_\Omega(X-Z)\|_F^2 + \lambda||Z||_* \tag{21}$$

where Ω indexes the observed entries and

$$P_\Omega(Y)(i,j) = \begin{cases} Y_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega \end{cases} \tag{22}$$

$P_\Omega(Y) + P_{\Omega^\perp}(Y) = Y$.

Lemma 1

Suppose the matrix $W_{m\times n}$ has rank r. The solution to the optimization problem

$$\min_Z \frac{1}{2}||W-Z||_F^2 + \lambda||Z||_* \tag{23}$$

is given by $\hat{Z} = \mathbf{S}_\lambda(W)$ where

$$\mathbf{S}_\lambda(W) = UD_\lambda V^T \textit{with } D_\lambda = \textit{diag}[(d_1-\lambda)_+, \ldots, (d_r-\lambda)_+], \tag{24}$$

An algorithm for matrix completion

$$
\begin{aligned}
&\frac{1}{2}\|P_\Omega(X) - P_\Omega(Z)\|_F^2 + \lambda||Z||_* \\
&= \frac{1}{2}\|P_\Omega(X) - [Z - P_{\Omega^\perp}(Z)]\,\|_F^2 + \lambda||Z||_* \\
&= \frac{1}{2}\|\,[P_\Omega(X) + P_{\Omega^\perp}(Z)] - Z\|_F^2 + \lambda||Z||_*
\end{aligned}
\tag{25}
$$

We can iteratively update Z using

$$
Z \leftarrow \mathbf{S}_\lambda(P_\Omega(X) + P_\Omega^\perp(Z)) \tag{26}
$$

$\mathbf{S}_\lambda$ can be replaced with hard thresholding (approximately solve $\lambda\text{rank}(Z)$).

Simulation I

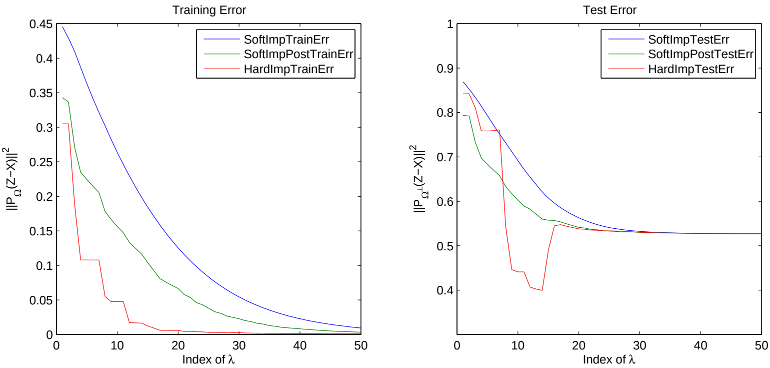


Figure: SNR=5, 80% entries missed. $\mathrm{rank}(X^*) = 5$, $X \in \mathbb{R}^{50\times 50}$.

Simulation II

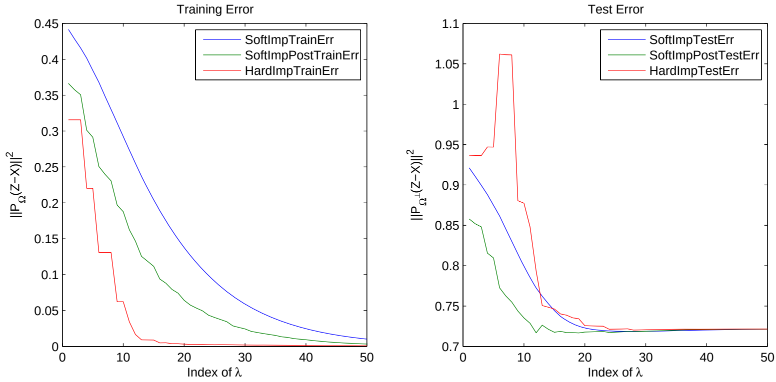


Figure: SNR=2, 80% entries missed. rank$(X^*) = 5$, $X \in \mathbb{R}^{50\times 50}$.

Summary

- ℓ_0 (rank) often performs better when SNR is high.
- ℓ_1 (nuclear norm) often performs better when SNR is low.
- ℓ_0 needs more iterations to solve even with warm starts.
- Discontinuity: The solution of ℓ_0 may change suddenly.

Outline

The shift model for outlier detection

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} + \epsilon \tag{27}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{E} \in \mathbb{R}^{n \times 1}$ is the shift component caused by outliers and $\epsilon \sim N(\mathbf{0}, \mathbf{I}\sigma)$. For simplicity, here we assume $n > p$.

Outlier detection using sparsity

$$\min_{\boldsymbol{\beta}, \mathbf{E}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\mathbf{E}\|_1 \tag{28}$$

Outlier detection using sparsity

- Let $\mathbf{X} = \mathbf{UDV}$ be the SVD of $\mathbf{X}$, $\mathbf{U} \in \mathbb{R}^{n \times n}$, $\mathbf{D} \in \mathbb{R}^{n \times p}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$.
- Define an index set $c = \{i : D_{ii} = 0\}$ and Let $\mathbf{U}_c$ be the corresponding columns of $\mathbf{U}$: $\mathbf{U}_c \in \mathbb{R}^{n \times (n-p)}$.

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\beta + \mathbf{E} + \epsilon \\
\mathbf{U}_c^T \mathbf{y} &= \underbrace{\mathbf{U}_c^T \mathbf{X}\beta}_{\mathbf{0}} + \mathbf{U}_c^T \mathbf{E} + \mathbf{U}_c^T \epsilon \\
\tilde{\mathbf{y}} &= \tilde{\mathbf{X}}\mathbf{E} + \tilde{\epsilon}
\end{aligned} \tag{29}
$$

Here $\tilde{\mathbf{y}} \in \mathbb{R}^{(n-p)\times 1}$, $\tilde{\mathbf{X}} \in \mathbb{R}^{(n-p)\times n}$, $\mathbf{E} \in \mathbb{R}^{n \times 1}$ and $\tilde{\epsilon} \in \mathbb{R}^{(n-p)\times 1}$.

Outlier detection using sparsity

$$
\min_{\mathbf{E}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{E}\|^2 + \lambda \|\mathbf{E}\|_1 \tag{30}
$$

The irrepresentable condition in outlier detection

Recall that

The irrepresentable condition (Zhao and Yu, 2006)

$$\|\mathbf{X}_N^T \mathbf{X}_K (\mathbf{X}_K^T \mathbf{X}_K)^{-1} \text{sign}(\beta_K)\|_\infty < 1.$$

- Here $\mathbf{X}$ and β_K should be replaced with $\tilde{\mathbf{X}}$ and $\mathbf{E}_K$, respectively.
- If the irrepresentable condition is not satisfied, ℓ_1 can't correctly detect outliers, while ℓ_0 works.

Outline

Outlier detection in low-rank representations

$$\mathbf{Y} = \mathbf{X} + \mathbf{E} + \epsilon \tag{31}$$

where $\mathbf{Y} \in \mathbb{R}^{m \times n}$ is the observed matrix, $\mathbf{X}$ is a low-rank matrix with rank($\mathbf{X}$)=r, $\mathbf{E}$ is the shift caused by outliers, and ϵ denotes Gaussian noises.

- Let $\mathbf{X} = \mathbf{UDV}^T$.
- Define an index set $c = \{i : D_{ii} = 0\}$ and Let $\mathbf{U}_c$ be the corresponding columns of $\mathbf{U} : \mathbf{U}_c \in \mathbb{R}^{m \times (n-r)}$.
- Notice that $\mathbf{X}$ is unknown and the SVD can only be done when $\mathbf{X}$ is known, e.g., simulation.

Outlier detection in low-rank representations

Using the similar trick as in regression, we have

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} + \mathbf{E} + \epsilon \\ \mathbf{U}_c^T\mathbf{Y} &= \underbrace{\mathbf{U}_c^T\mathbf{X}}_{\mathbf{0}} + \mathbf{U}_c^T\mathbf{E} + \mathbf{U}_c^T\epsilon \\ \tilde{\mathbf{Y}} &= \tilde{\mathbf{X}}\mathbf{E} + \tilde{\epsilon} \end{aligned} \tag{32}$$

Here $\tilde{\mathbf{Y}} = \mathbf{U}_c^T\mathbf{Y} \in \mathbb{R}^{(n-r)\times n}$, $\tilde{\mathbf{X}} = \mathbf{U}_c^T \in \mathbb{R}^{(n-r)\times m}$, $\mathbf{E} \in \mathbb{R}^{m\times n}$ and $\tilde{\epsilon} \in \mathbb{R}^{(n-r)\times n}$.
Let $\tilde{\mathbf{Y}}_j$ and $\mathbf{E}_j$ denote the j-th column of $\tilde{\mathbf{Y}}$ and $\mathbf{E}$, respectively.

$$\sum_{j=1}^{n} \|\tilde{\mathbf{Y}}_j - \tilde{\mathbf{X}}\mathbf{E}_j\|^2 + \sum_{j=1}^{n} \|\mathbf{E}_j\|_1 \tag{33}$$

When $\tilde{\mathbf{X}}$ satisfies the irrepresentable condition for all j, ℓ_1 works.

ℓ_1/ℓ_γ regularization

The penalty becomes

$$\lambda \sum_{j \in G} \|\boldsymbol{\beta}_j\|_\gamma \tag{34}$$

- $\gamma = 2$: Group Lasso ((Yuan and Lin, 2007).
- $1 \leq \gamma \leq \infty$: (Zhao and Yu, 2009).
- Other variants include sparse Group Lasso (Friedman et al. 2010) and overlapping Group Lasso (Jacob et al. 2009).

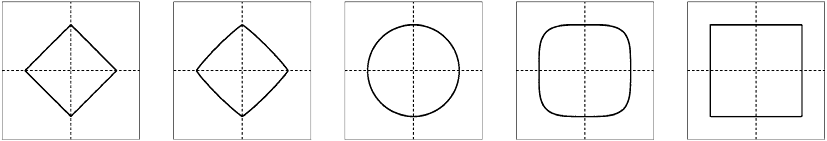
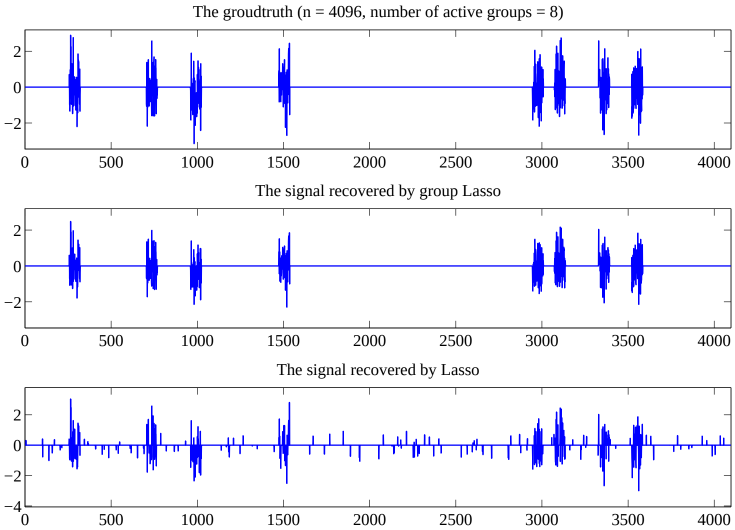


Figure: $\|(\beta_{j1}, \beta_{j2})\|_\gamma$. From Left to Right: $\gamma = 1$,$\gamma = 1.1$,$\gamma = 2$,$\gamma = 4$,$\gamma = \infty$.

Group Lasso vs. Lasso



Graph-regularized Lasso

The penalty becomes

$$\lambda_1 \|\beta\|_1 + \lambda_2 \beta^T \mathbf{L} \beta \tag{35}$$

- $\mathbf{L} = \mathbf{I}$: Elastic net (Zou and Hastie, 2005).
- $\mathbf{L}$ can be a Laplacian matrix of a graph.
- Coordinate Decent also works here.

H. Zou and T. Hastie.
Regularization and Variable Selection via the Elastic Net.
JRSSB, 2005.

Generalized Lasso (Tibshirani, 2010)

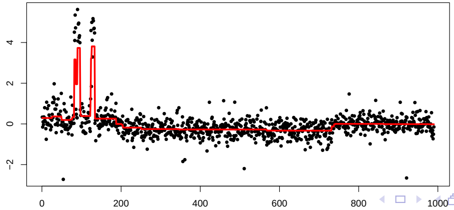
The penalty becomes

$$\lambda \|\mathbf{T}\beta\|_1 \tag{36}$$

- Fused Lasso (piecewise constant):

$$\mathbf{T} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ & & & \dots & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \tag{37}$$

- Coordinate Descent can't be directly applied here.
- ADMM and Nesterov's method work here.



Generalized Lasso

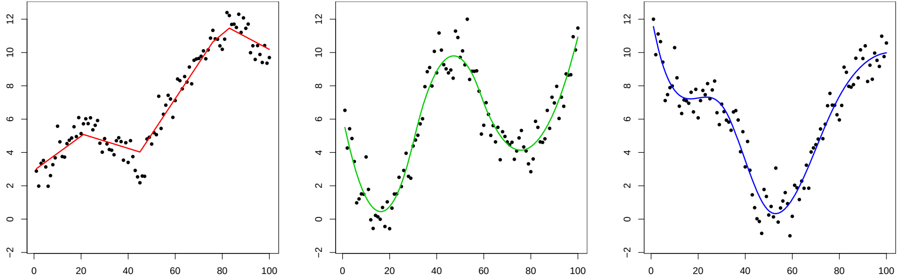
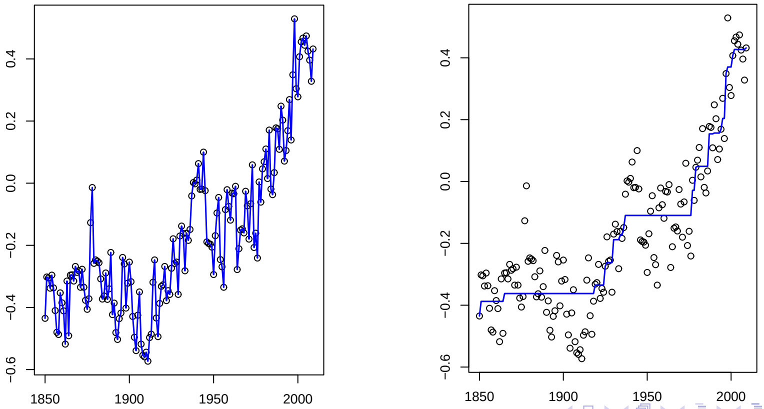


Figure: Left: piecewise linear. Middle: piecewise quadratic. Right: piecewise cubic.

R. Tibshirani and J. Taylor.
The Solution Path of the Generalized Lasso.
Annals of Statistics, 2010.

Nearly-isotonic fitting

$$\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_i)^2+\lambda\sum_{i=1}^{n-1}(\beta_i-\beta_{i+1})_+ \tag{38}$$



Graphical Lasso

Covariance estimation in Gaussian family

$$L(\mathbf{\Theta}) = \log\det\mathbf{\Theta} - \mathrm{trace}\mathbf{S}\mathbf{\Theta} \tag{39}$$

where $\mathbf{S}$ is the empirical covariance matrix. The maximum likelihood estimate of $L(\mathbf{\Theta})$ is

$$\hat{\mathbf{\Theta}} = \arg\max L(\mathbf{\Theta}) = \mathbf{S}^{-1}. \tag{40}$$

It becomes an ill-posed problem when $p > n$ ($\mathbf{S}$ becomes singular). A well-posed formulation is

$$\max_{\mathbf{\Theta}} \quad \log\det\mathbf{\Theta} - \mathrm{trace}\mathbf{S}\mathbf{\Theta} - \lambda\|\mathbf{\Theta}\|_1 \tag{41}$$

which is known as “Graphical Lasso” (Friedman et al., 2008).

J. Friedman, T. Hastie and R. Tibshirani
Sparse inverse covariance estimation with the lasso.
Biostatistics, 2008.

Graphical Lasso

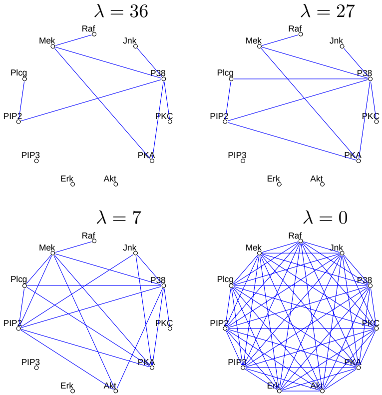


Figure: Four different Graphical-Lasso solutions. Coordinate Decent can be used here (Friedman et al., 2008).

Graphical Lasso for time-varying networks

Graphical Lasso for time-varying networks

$$\max_{\boldsymbol{\Theta}^t, t \in \{1,\ldots,T\}} \quad \sum_{t=1}^{T} \left(\log\det \boldsymbol{\Theta}^t - \mathrm{trace} \mathbf{S}^t \boldsymbol{\Theta}^t \right) - \lambda_1 \sum_{t=1}^{T} \|\boldsymbol{\Theta}^t\|_1 - \lambda_2 \sum_{t=2}^{T} \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1}\|_1$$

where t is the time index. The fused term $\|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1}\|_1$ assumes that $\boldsymbol{\Theta}^t$ can change with time but in a piecewise constant way.

- This model is proposed by myself. It may be useful in system biology.
- To my knowledge, there is no particularly designed algorithm for this convex optimization problem. How to solve it in an efficiently way is an open question.
- Probably, ADMM or Nesterov's method will work.

Conclusion

- We have discussed sparsity in various situations, such as sparse coefficients in regression, low-rank representation and sparse edge of a graph.
- Structured sparsity can be further explored in a specific application, such as fused Lasso for neighboring information (e.g., neighboring pixels, neighboring SNPs), group Lasso for grouping information (e.g., pathway information, peptides belonging to the same protein form a group).
- Empirical comparison between ℓ_1 and ℓ_0 enables us to know more about sparsity.

Opportunity and challenges

Albert Einstein "As far as the laws of mathematics refer to reality, they are not certain, as far as they are certain, they do not refer to reality."

(Loading DECOLOR) (Loading DECOLOR)

Figure: X. Zhou et al. (2011). A lot of applications ...

Stephen Boyd "God knows the last thing we need is another algorithm for the Lasso." (Sept 28, 2010, known from Tibshrani's talk).

Reference I

T. Haste, R. Tibshirani, J. Friedman
The elements of statistical learning (2nd).
Springer, 2009.

R. Mazumder, J. Friedman and T. Hastie.
SparseNet: Coordinate Descent with Non-Convex Penalties.
Journal of Machine Learning Research, Stanford University, 2010.

R. Mazumder, T. Hastie.
Spectral Regularization Algorithms for Learning Large Incomplete Matrices.
Journal of Machine Learning Research, Stanford University, 2010.

S. Boyd, N. Parikh, E., Chu, B. Peleato and J. Eckstein
Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers.
Tech. Report, Stanford University, Oct., 2010.

L. Breiman
Better Subset Regression Using the Nonnegative Garrote.
Technometrics, 37:373-384, 1995.

P. Zhao and B. Yu
On model selection consistency of Lasso.
The Journal of Machine Learning Research, 2006.

Reference II

D. Donoho and I. Johostone
Ideal spatial adaption by wavelet shinkage.
Biometrika, 81:425-455, 1994.

S. Chen, D. Donoho and M. Saunders
Atomic decomposition by basis pursuit.
SIAM Journal on Scientific computing, 1998.

R. Tibshirani,
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society: Series B, 1996.

J. Friedman, T. Hastie, H. Hoefling and R. Tibshirani
Pathwise coordinate optimization.
Annals of applied statistics, 2007.

D. Donoho
Compressed sensing.
IEEE Tran. on Information Theory, 2006.

E. Candes and T. Tao
Near-optimal signal recovery from random projections: universal encoding strategies.
IEEE Tran. on Information Theory, 2006.

Reference III

- B. Efron, T. Hastie, I. Johnstone and R. Tibshirani
  Least angle regression.
  Annals of statistics, 2004.
- H. Gao and A. Bruce
  Waveshink with firm shrinkage.
  Statistica Sinica, 1997.
- C. Zhang
  Nearly unbiased variable selection under minimax concave penalty.
  Annals of statistics, 2011.
- J. Friedman, T. Hastie, S. Rosset, R. Tibshirani and J Zhu
  Discussion of three Boosting papers.
  Annals of statistics, 2004.
- P. Buhlmann
  Invited Discussion on “Regression shrinkage and selection via the Lasso: a retrospective (Tibshirani)”.
  Journal of the Royal Statistical Society: Series B, 2011.
- J. Friedman, T. Hastie and R. Tibshirani
  A Note on the Group Lasso and a Sparse Group Lasso.
  Technique report, Stanford University, 2010.

Reference IV

R. Tibshrani et al.
Strong Rules for Discarding Predictors in Lasso-type Problems.
Tech. report, Stanford University, 2010.

J. Friedman, T. Hastie and R. Tibshirani
Applications of the lasso and grouped lasso to the estimation of sparse graphical models.
Technique report, Stanford University, 2010.

J. Fan and R. Li
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American Statistical Association, 2001.

J. Fan and J. Lv
Sure independence screening for ultrahigh dimensional feature space.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008.