ST509 Computational Statistics

Lecture 6: Extension of LASSO

Seung Jun Shin

Department of Statistics
Korea University

E-mail: sjshin@korea.ac.kr

LASSO-penalized GLM I

- Lasso-penalized GLM solves

$$\min -\frac{1}{n}\sum L(\boldsymbol{\beta};\mathbf{y},\mathbf{X}) + \lambda\|\boldsymbol{\beta}\|_1$$

ex. Logistic regression:

$$\min_{\boldsymbol{\beta}} \; -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\beta_0+\boldsymbol{\beta}^T\mathbf{x}_i) - \log(1+e^{\beta_0+\boldsymbol{\beta}^T\mathbf{x}_i})\right\} + \lambda\|\boldsymbol{\beta}\|_1$$

- When y_i is coded as $\{-1, 1\}$,

$$\min_{\boldsymbol{\beta}} \; \frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i f(\mathbf{x}_i)}\right) + \lambda\|\boldsymbol{\beta}\|_1$$

where $f(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i$.

LASSO-penalized GLM II

- Recall that the unpenalized LR iteratively solves

$$\boldsymbol{\beta}^{(t+1)} = \min_{\boldsymbol{\beta}} \frac{1}{n}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}),$$

  where $\mathbf{X} = \mathbf{W}_{(t)}^{1/2}\mathbf{X}$ and $\mathbf{W}_{(t)}^{1/2}\mathbf{z}_{(t)}$ with $\mathbf{z}_{(t)} = \mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{W}_{(t)}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{(t)})$.
- Lasso penalized version solves

$$\boldsymbol{\beta}_{\text{lasso}}^{(t+1)} = \min_{\boldsymbol{\beta}} \frac{1}{n}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|,$$

  which can be readily implemented by CD algorithm.
- Notice that various GLM with Lasso penalty can be implemented by changing $\tilde{\mathbf{y}}$, $\tilde{\mathbf{X}}$, and related quantities accordingly.

Elastic Net I

- Lasso does not handle highly correlated predictors well.
- Toy example Suppose $Z_1, Z_2 \overset{iid}{\sim} N(0,1)$ and the true model is

$$Y = 3Z_1 - 1.5Z_2 + 2\epsilon, \text{with } \epsilon \sim N(0,1)$$

However, we observe $X_1, \cdots, X_6$ where

$$X_j = Z_1 + \xi_j/5 \text{ with } \xi_j \sim N(0,1) \text{ for } j = 1,2,3; \text{and}$$
$$X_j = Z_2 + \xi_j/5 \text{ with } \xi_j \sim N(0,1) \text{ for } j = 4,5,6.$$

Elastic Net II

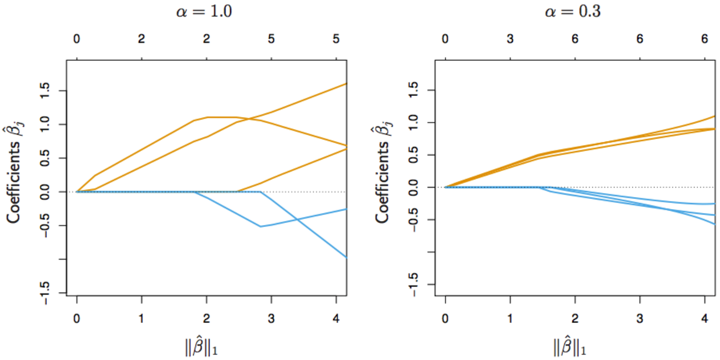

Figure 4.1 *Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the correlated groups are pulled together.*

Figure: From SLS.

Elastic Net III

- Elastic net (Zhu and Hastie, 2005) solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda \left[\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1 \right]$$

  where $\alpha \in [0, 1]$.
- Elastic net penalty is a hybrid version of lasso and ridge penalty.
- It is not difficult to show that the one-dimensional solution for orthogonal regression problem with elastic net penalty is

$$\hat{\beta}_j = \frac{S_{\lambda\alpha}\left(\frac{1}{n}\mathbf{y}^T\mathbf{x}\right)}{1 + \lambda(1-\alpha)}, \quad j = 1, \cdots, p.$$

- CD algorithm can be readily applied. (`glmnet` package in R)

Elastic Net IV

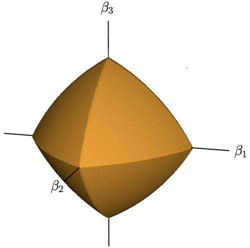

Figure 4.2 *The elastic-net ball with* $\alpha = 0.7$ *(left panel) in* $\mathbb{R}^3$*, compared to the* ℓ_1 *ball (right panel). The curved contours encourage strongly correlated variables to share coefficients (see Exercise 4.2 for details).*

Group LASSO I

- Assume we have a group structure on $\mathbf{X}$:

$$\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_J) \text{ and } \boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_J^T)$$

  with $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}, j = 1, \cdots, p$; and $\sum_{j=1}^J p_j = p$.
- WLOG we assume $\mathbf{X}_j$ is orthonormalized to $\mathbf{Z}_j$ (i.e., $\mathbf{Z}_j^T \mathbf{Z}_j = \mathbf{I}_{p_j}$), group LASSO solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^J \mathbf{Z}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^J \|\boldsymbol{\beta}_j\|_2$$

  where $\|\boldsymbol{\beta}\|_2 = \sqrt{\beta_1^2 + \cdots + \beta_p^2}$

Group LASSO II

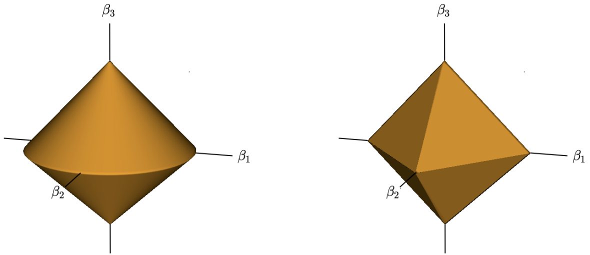

Figure 4.3 *The group lasso ball (left panel) in* $\mathbb{R}^3$*, compared to the* ℓ_1 *ball (right panel). In this case, there are two groups with coefficients* $\theta_1 = (\beta_1, \beta_2) \in \mathbb{R}^2$ *and* $\theta_2 = \beta_3 \in \mathbb{R}^1$.

Group LASSO III

- Subgradient equation is

$$-\mathbf{Z}_j^T(\mathbf{y} - \sum_{j=1}^{J} \mathbf{Z}_j \hat{\boldsymbol{\beta}}_j) + \lambda \hat{\mathbf{s}}_j = 0, \qquad \text{for } j = 1, \cdots, J,$$

where

$$\hat{\mathbf{s}}_j = \begin{cases} \hat{\boldsymbol{\beta}}_j / \|\hat{\boldsymbol{\beta}}_j\|_2, & \text{when } \boldsymbol{\beta}_j \neq \mathbf{0} \\ \text{any vector with } \|\hat{\mathbf{s}}_j\| \leq 1, & \text{when } \boldsymbol{\beta}_j = \mathbf{0} \end{cases}$$

- With all $\{\hat{\boldsymbol{\beta}}_k, k \neq j\}$ fixed, we write

$$-\mathbf{Z}_j^T(\mathbf{r}_j - \mathbf{Z}_j \boldsymbol{\beta}_j) + \lambda \hat{\mathbf{s}}_j = \mathbf{0}$$

where the jth partial residual $\mathbf{r}_j$ is

$$\mathbf{r}_j = \mathbf{y} - \sum_{k \neq j} \mathbf{Z}_k \hat{\boldsymbol{\beta}}_k$$

Group LASSO IV

- We can update

$$\hat{\boldsymbol{\beta}}_j \quad \leftarrow \quad \left(1 - \frac{\lambda}{\|\mathbf{Z}_j^T \mathbf{r}_j\|_2}\right)_+ \mathbf{Z}_j^T \mathbf{r}_j$$

Group LASSO V

- Consider the eigenvalue decomposition of $\mathbf{X}_j^T \mathbf{X}_j$:

$$\mathbf{X}_j^T \mathbf{X}_j = \mathbf{Q}_j \mathbf{\Lambda}_j \mathbf{Q}_j^T$$

- Then we have the following transformation of $\mathbf{X}_j$

$$\mathbf{Z}_j = \mathbf{X}_j \mathbf{Q}_j \mathbf{\Lambda}_j^{-1/2}$$

with

$$\mathbf{Z}_j^T \mathbf{Z}_j = \mathbf{I} \quad \text{and} \quad \mathbf{Z}_j \tilde{\boldsymbol{\beta}}_j = \mathbf{X}_j (\mathbf{Q}_j \mathbf{\Lambda}_j^{-1/2} \tilde{\boldsymbol{\beta}}_j)$$

where $\tilde{\boldsymbol{\beta}}_j$ is the solution on the orthonormalized scale.

Group LASSO VI

1. For centered $\mathbf{X}$ and $\mathbf{y}$, then we orthonormalize $\mathbf{X}_j$ by computing $\mathbf{Z}_j = \mathbf{X}_j \mathbf{Q}_j \mathbf{\Lambda}_j^{-1/2}$ where $\mathbf{Q}_j$ and $\mathbf{\Lambda}_j$ are eigenvectors and eigenvalues of $\mathbf{X}_j^T \mathbf{X}_j$.
2. Initialize $\boldsymbol{\beta}_j, j = 1, \cdots, J$, and compute full residuals $\mathbf{r} = \mathbf{y} - \sum_{j=1}^J \mathbf{Z}_j \boldsymbol{\beta}_j$.
3. Repeat for $j = 1, \cdots, J$ until convergence
   3.1 Compute $\mathbf{r}_j$ (partial residual)
   $$\mathbf{r}_j \quad \leftarrow \quad \mathbf{r} + \mathbf{Z}_j \hat{\boldsymbol{\beta}}_j$$
   3.2 Update $\boldsymbol{\beta}_j$
   $$\hat{\boldsymbol{\beta}}_j \quad \leftarrow \quad S_{\lambda_j}(\|\mathbf{Z}_j^T \mathbf{r}_j\|_2) \frac{\mathbf{Z}_j^T \mathbf{r}_j}{\|\mathbf{Z}_j^T \mathbf{r}_j\|_2}$$
   where $\lambda_j = \lambda / \sum_{k \in j\text{th group}} (\ell_k)$ with ℓ denotes an eigenvalue of $\mathbf{X}^T \mathbf{X}$.
   3.3 Update $\mathbf{r}$ (full residual)
   $$\mathbf{r} \quad \leftarrow \quad \mathbf{r}_j - \mathbf{Z}_j \hat{\boldsymbol{\beta}}_j$$
4. back-transformation:
   $$\hat{\boldsymbol{\beta}}_j \quad \leftarrow \quad \mathbf{Q}_j \mathbf{\Lambda}^{-1/2} \hat{\boldsymbol{\beta}}_j$$

Fused LASSO I

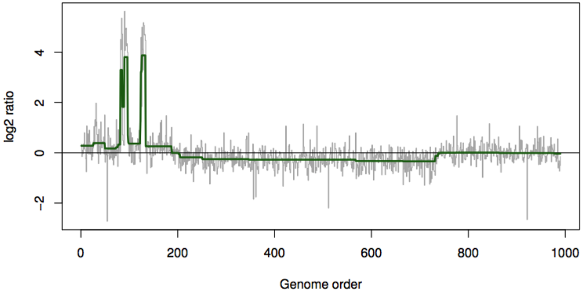

Figure 4.8 *Fused lasso applied to CGH data. Each spike represents the copy number of a gene in a tumor sample, relative to that of a control (on the log base-2 scale). The piecewise-constant green curve is the fused lasso estimate.*

Figure: Example of Fused LASSO: CHG data for copy number detection.

Fused LASSO II

- Fused LASSO signal approximator solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=2}^{n} |\beta_j - \beta_{j-1}|$$

  that can solve the change point detection problem illustrated in Figure 2.

- We have

$$\hat{\beta}_i(\lambda_1, \lambda_2) = S_{\lambda_1}\left(\hat{\beta}_i(0, \lambda_2)\right)$$

  where $S_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+$ denotes the soft-threshold operator.

- Thus, if we solve the fused lasso with λ_0, all other solutions can be obtained immediately.

- That is, it suffices to focus on solving

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum (y_i - \beta_i)^2 + \lambda_2 \sum_{i=2}^{n} |\beta_j - \beta_{j-1}| \tag{1}$$

Fused LASSO III

- One simple approach is to consider $\boldsymbol{\gamma} = \mathbf{M}\boldsymbol{\beta}$ such that

$$\gamma_1 = \theta_1 \quad \text{and} \quad \gamma_i = \theta_i - \theta_{i-1} \;\; \text{for } i = 2, \cdots, N.$$

- (1) is equivalently rewritten as the LASSO problem.

$$\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|^2 + \lambda\|\boldsymbol{\gamma}\|_1, \qquad \text{with } \mathbf{X} = \mathbf{M}^{-1}.$$

- One generalization of Fused LASSO is (ℓ_1-)trend filtering which solves

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda\|\mathbf{D}^{(k)}\boldsymbol{\beta}\|_1$$

where $\mathbf{D}^{(k)}$ is a matrix that computes discrete difference of order k.

Graphical LASSO I

- Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a p- dimensional Gaussian distribution.

$$f(\mathbf{x}) = \left\{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\right\}^{-1} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix.
- It can be shown that

$$\boldsymbol{\Theta}_{ij} = 0 \qquad \Rightarrow \qquad X_i \perp X_j \mid \mathbf{X}_{-ij}$$

Graphical LASSO II

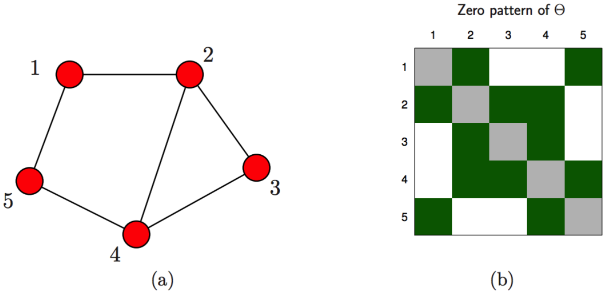

Figure 9.3 *(a) An undirected graph G on five vertices. (b) Associated sparsity pattern of the precision matrix $\boldsymbol{\Theta}$. White squares correspond to zero entries.*

Figure: From SLS

Graphical LASSO III

- WLOG assume $\mu = 0$, the log-density is

$$\log f(\mathbf{x}_i; \boldsymbol{\Theta}) = -\frac{1}{2}\log\{\det(\boldsymbol{\Theta})/(2\pi)\} - \frac{1}{2}\mathbf{x}_i^T \boldsymbol{\Theta}\mathbf{x}$$

  and hence the (scaled) log-likelihood is

$$\frac{1}{n}\sum_{i=1}^{n} \log f(\mathbf{x}_i; \boldsymbol{\Theta}) \propto \log\det(\boldsymbol{\Theta}) - \mathrm{trace}(\mathbf{S}\boldsymbol{\Theta})$$

  where $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T$ is the sample covariance estimator.
- MLE of $\boldsymbol{\Theta}$ is $\mathbf{S}^{-1}$ assuming the non-singularity of $\mathbf{S}$.
- In high-dimensional case with $p > n$, however, $\mathbf{S}$ is singular.
- Sparse structure of $\boldsymbol{\Theta}$ is often assumed.

Graphical LASSO IV

- To identify the sparsity structure of $\mathbf{\Theta}$, we can solve

$$\max_{\mathbf{\Theta} \succeq \mathbf{0}} \log \det(\mathbf{\Theta}) - \text{trace}(\mathbf{S}\mathbf{\Theta}) - \lambda \|\mathbf{\Theta}\|_1 \tag{2}$$

  where $\|\mathbf{\Theta}\|_1 = \sum_{s \neq t} |\theta_{st}|$
- The log-determinant function is defined on the space of symmetric matrices as

$$\log \det(\mathbf{\Theta}) = \begin{cases} \sum_{j=1}^{p} \log\{\lambda_j(\mathbf{\Theta})\} & \text{if } \mathbf{\Theta} \succ \mathbf{0} \\ -\infty; & \text{otherwise,} \end{cases}$$

  where λ_j denotes the jth leading eigenvalues of $\mathbf{\Theta}$.

Graphical LASSO V

- Taking derivative of (2) w.r.t $\mathbf{\Theta}$ yields

$$\mathbf{\Theta}^{-1} - \mathbf{S} - \lambda \mathbf{\Phi} = \mathbf{0} \tag{3}$$

  where $\mathbf{\Phi} = \text{sign}(\mathbf{\Theta})$ with $\text{sign}(\theta) \in [-1, 1]$ if $\theta = 0$.

- Let $\mathbf{W}$ denote the current working version of $\mathbf{\Theta}^{-1}$.

- We can use

$$\begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & w_{22} \end{bmatrix} \begin{bmatrix} \mathbf{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^T & \theta_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}$$

  and the upper-right part gives

$$\mathbf{W}_{11}\mathbf{\Theta}_{11} + \mathbf{w}_{12}\theta_{22} = \mathbf{0} \quad \Rightarrow \mathbf{w}_{12} = \mathbf{W}_{11}\boldsymbol{\beta}$$

  where $\boldsymbol{\beta} = -\boldsymbol{\theta}_{12}/\theta_{22}$.

Graphical LASSO VI

- The upper-right part of (3) yields

$$\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} - \lambda\boldsymbol{\psi}_{12} = 0 \tag{4}$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{bmatrix}, \text{ and } \boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\psi}_{12} \\ \boldsymbol{\psi}_{12}^T & \psi_{22} \end{bmatrix}$$

- It turns out to be (4) is identical to

$$\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} + \lambda \cdot \text{sign}(\boldsymbol{\beta}) = \mathbf{0} \tag{5}$$

- This because

$$\boldsymbol{\theta}_{12} = -\theta_{22}\mathbf{W}_{11}^{-1}\mathbf{w}_{12}, \quad \text{and} \quad \theta_{22} > 0$$

and therefore

$$\text{sign}(\boldsymbol{\theta}_{12}) = \text{sign}(-\mathbf{W}_{11}^{-1}\mathbf{w}_{12}) = \text{sign}(-\boldsymbol{\beta}).$$

Graphical LASSO VII

- Recall that the lasso minimizes

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

- Its stationary equations are

$$\frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \frac{1}{n}\mathbf{X}^T\mathbf{y} + \lambda \cdot \mathrm{sign}(\boldsymbol{\beta}) = \mathbf{0}.$$

- You can realize its similarity to (5).
- (5) is the subgradient equation for

$$\min_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{W}_{11}^{1/2}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{W}_{11}^{1/2}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \tag{6}$$

where $\mathbf{y} = \mathbf{W}_{11}^{-1/2}\mathbf{s}_{12}$.

Graphical LASSO VIII

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda \mathbf{I}$. Note that the diagonal of $\mathbf{W}$ is unchanged in what follows.
2. Repeat for $j = 1, 2, \cdots, p$ until convergence:
   2.1 Compute $\mathbf{W}_{11} =$ `W[-j,-j]`, $\mathbf{W}_{11}^{1/2}$, and $\mathbf{W}_{11}^{-1/2}$. t
   2.2 Solve (6) using CD algorithm which gives $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p-1}$.
   2.3 Update $\mathbf{w}_{12} = \mathbf{W}_{11} \hat{\boldsymbol{\beta}}$.
3. In the final cycle (for each j) update $\hat{\boldsymbol{\theta}}_{12} = -\hat{\boldsymbol{\beta}} \cdot \hat{\theta}_{22}$ with $1/\hat{\theta}_{22} = w_{22} - \mathbf{w}_{12}^T \hat{\boldsymbol{\beta}}$.

Algorithm 1: Graphical Lasso Algorithm

Reference

- Friedman, Hastie, & Tibshirani (2009) The Elements of Statistical Learning; 2nd edition, Springer.
- Tibshirani, Wainwright, & Hastie (2015) Statistical Learning with Sparsity: the lasso and generalizations Chapman and Hall/CRC.
- Zou & Hastie (2005) Regularization and variable selection via the elastic net JRSSb, 67(2), 301-320.
- Yuan & Lin (2006) Model selection and estimation in regression with grouped variables JRSSb, 68(1), 49-67.
- Tibshirani, Saunders, Rosset, Zhu, and Knight (2005) Sparsity and smoothness via the fused lasso JRSSb, 67(1), 91-108.
- Friedman, Hastie, & Tibshirani (2008) Sparse inverse covariance estimation with the graphical lasso Biostatistics, 9(3), 432-441.