Generalizations of the Lasso Penalty

Yan Liu, Lilian Müller

1 October 2018

Content

1 Introduction and Elastic Net
- Introduction
- Elastic Net

2 The Group Lasso
- Introduction: Group Lasso
- Sparse Group Lasso
- Overlap Group Lasso

3 Sparse Additive Models and the Group Lasso
- Additive Models and Backfitting
- SPAM Sparse Additive Models and Backfitting
- Optimization and the Group Lasso
- Multiple Penalization for Sparse Additive Models

4 Theoretical Results for Lasso
- Introduction: Support Recovery in Linear Regression
- Theorem: Variable-selection Consistency
- Numerical studies

5 Conclusion

Introduction

Recall: Basic Model

The Lasso Estimator in Lagrangian form is

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1 \right\}, \qquad \lambda \geq 0.$$

Our Goal

Expand the scope of the basic lasso:
investigate variations/generalizations of the basic lasso l_1-penalty: $\lambda||\beta||_1$

Why Generalized Lasso?

- Classical lasso does not perform well with highly correlated variables.

  e.g. In studies of microarray, genes operating in the same biological pathway express(or not) together

  $\Rightarrow$ correlated variables are selected (or not) together and corresponding variables share coefficients

  $\star$ **Elastic Net**: combining a squared l_2-penalty with the l_1-penalty

Why Generalized Lasso?

- Classical lasso does not perform well with highly correlated variables.

  e.g. In studies of microarray, genes operating in the same biological pathway express(or not) together

  $\Rightarrow$ correlated variables are selected (or not) together and corresponding variables share coefficients

  $\star$ **Elastic Net**: combining a squared l_2-penalty with the l_1-penalty

- Classical lasso does not perform well with structurally grouped variables.

  e.g. In regression problems, covariates have a natural group structure

  $\Rightarrow$all coefficients within one group become zero(nonzero) simultaneously

  $\star$ **Group Lasso, Overlap Group Lasso**: using sums of un-squared l_2-penalties
  $\star$ **Nonparametric smoothing methods**: COSSO, SPAM

Elastic Net

Definition

The elastic net solves the convex problem

$$\min_{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p} \left\{ \frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda[\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1] \right\}, \qquad \alpha \in [0,1]$$

Elastic Net

Definition

The elastic net solves the convex problem

$$\min_{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p} \left\{ \frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda[\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1] \right\}, \qquad \alpha \in [0,1]$$

Remark

- penalty (omit λ) applied to each coefficient:

$$\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|$$

  is a compromise between ridge (squared l_2-penalty) and lasso (l_1-penalty)
- $\alpha = 1$: lasso penalty
  $\alpha = 0$: ridge penalty
- α: high-level parameter, determined subjectively or by cross-validation

Constraint Region

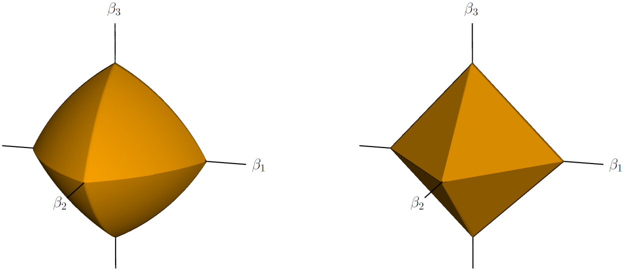


Abbildung: Elastic net ball with $\alpha = 0.7$ (left) versus l_1- ball(right)

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 58

Elastic net ball shares attributes of both l_1- ball and l_2- ball :

- sharp corners and edges $\rightarrow$ selection
- curved contours $\rightarrow$ sharing of coefficients

Example: Comparison of Lasso and Elastic Net on highly correlated variables

Simulation settings:

- 2 sets of 3 variables, pairwise correlations around 0.97 in each group
- Sample size: N=100
- data simulated as follows:

$$
\begin{aligned}
Z_1, Z_2 &\sim N(0,1) \qquad \text{independent} \\
Y &= 3Z_1 - 1.5Z_2 + 2\epsilon, \quad \epsilon \sim N(0,1) \\
X_j(j = 1,2,3) &= Z_1 + \xi/5, \quad \xi_j \sim N(0,1) \\
X_j(j = 4,5,6) &= Z_2 + \xi/5, \quad \xi_j \sim N(0,1)
\end{aligned}
$$

Example: Implementation of lasso and elastic net in R

```r

## ----------------------------------
##  SIMULATION LASSO VS. ELASTIC NET
## ----------------------------------

N = 100
set.seed(1234)
Z1 <- rnorm(N)
Z2 <- rnorm(N)
epsilon <- rnorm(N)

Y <- 3*Z1 - 1.5*Z2 + 2*epsilon
X1 <- Z1 + rnorm(N)/5
X2 <- Z1 + rnorm(N)/5
X3 <- Z1 + rnorm(N)/5
X4 <- Z2 + rnorm(N)/5
X5 <- Z2 + rnorm(N)/5
X6 <- Z2 + rnorm(N)/5

X <- as.matrix(cbind(X1,X2,X3,X4,X5,X6))

library(glmnet)
## Fit classical Lasso:
fit_lasso <- glmnet(X,Y, alpha=1)
## Plot classical Lasso fit:
plot(fit_lasso)

## Fit elastic net:
fit_elnet <- glmnet(X,Y, alpha=0.3)
## Plot elastic net fit:
plot(fit_elnet)
```

Result analysis

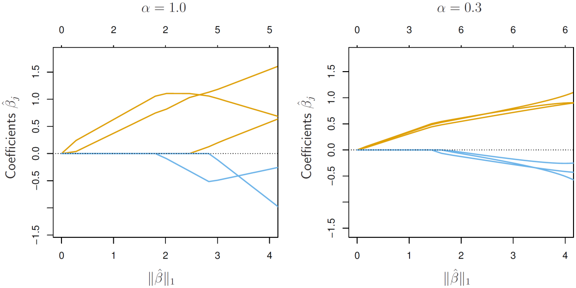


Abbildung: lasso estimates(left) versus elastic net(right)

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 56

- lasso estimates exhibit erratic behavior as λ varies: one variable is excluded and the correlations among variables are not clear
- elastic net includes all variables and correlated groups are pulled together, sharing values approximately equally

Result Analysis

Conclusion

In practice, group structure may not be as evident as the previous 'ideal' model, this example does capture the main idea of elastic net:

by adding ridge penalty to lasso penalty, elastic net automatically controls for strong within-group correlations

Introduction: Group Lasso

- Groups of covariates be selected into or out of a model together
- Desirable to have all coefficients within a group become nonzero (or zero) simultaneously

Introduction: Group Lasso

- Groups of covariates be selected into or out of a model together
- Desirable to have all coefficients within a group become nonzero (or zero) simultaneously

We use group lasso penalty for such situations

Introduction: Group Lasso

- Groups of covariates be selected into or out of a model together
- Desirable to have all coefficients within a group become nonzero (or zero) simultaneously

We use group lasso penalty for such situations

Example:

Genes and proteins often lie in known pathways, an investigator may be more interested in which pathway are related to an outcome than whether particular individual genes are.

Group Lasso: The model

- Consider linear regression model involving J groups of covariates, where $j = 1, \ldots, J$
- Vector $Z_j \in \mathbb{R}^{p_j}$ represents the covariates in group j
- **Goal:** predict real-valued response $Y \in \mathbb{R}$ based on collection of covariates $(Z_1, \ldots, Z_J)$

Group Lasso: The model

- Consider linear regression model involving J groups of covariates, where $j = 1, \dots, J$
- Vector $Z_j \in \mathbb{R}^{p_j}$ represents the covariates in group j
- **Goal:** predict real-valued response $Y \in \mathbb{R}$ based on collection of covariates $(Z_1, \dots, Z_J)$

linear model for the regression function $\mathbb{E}(Y|Z)$

linear model takes the form:

$$\theta_0 + \sum_{j=1}^{J} Z_j^T \theta_j$$

where $\theta_j \in \mathbb{R}^{p_j}$

Solution to group lasso problem

Given a collection of N samples $\{(y_i, z_{i1}, z_{i2}, \ldots, z_{iJ})\}_{i=1}^{N}$ the group lasso solves the convex problem:

$$\underset{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^{p_j}}{minimize} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \theta_0 - \sum_{j=1}^{J} z_{ij}^T \theta_j)^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\}$$

Where $\|\theta_j\|_2$ is the Euclidean norm.

Solution to group lasso problem

Given a collection of N samples $\{(y_i, z_{i1}, z_{i2}, \dots, z_{iJ})\}_{i=1}^N$ the group lasso solves the convex problem:

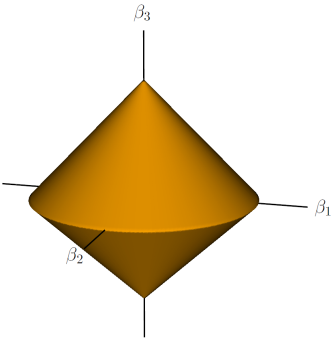
$$\underset{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^{p_j}}{minimize} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \theta_0 - \sum_{j=1}^{J} z_{ij}^T \theta_j)^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\}$$

Where $\|\theta_j\|_2$ is the Euclidean norm.
This is group generalization of the lasso with properties:

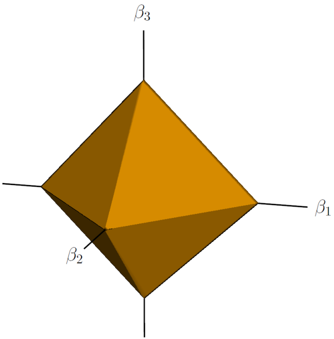
- Depending on $\lambda \geq 0$ either the entire vector $\hat{\theta}_j$ will be zero, or all its elements will be nonzero
- When $p_j = 1$ then we have $\|\theta_j\|_2 = |\theta_j|$, so reduces to the ordinary lasso

Constraint region

group-lasso



lasso



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 59

Two groups with coefficients $\theta_1 = \{\beta_1, \beta_2\} \in \mathbb{R}^2$ and $\theta_2 = \beta_3 \in \mathbb{R}^1$

Example: Regression with multilevel factors 1

- A predictor variable can be a multilevel factor
- Include separate coefficient for each level of the factor

Take continuous predictor X and a three-level factor G, with levels g_1, g_2, g_3. Linear model for mean:

$$\mathbb{E}(Y|X, G) = X\beta + \sum_{k=1}^{3} \theta_k \mathbb{1}_k [G]$$

event $\{G = g_k\}$

- Introduce vector $Z = (Z_1, Z_2, Z_3)$ of dummy variables
- $Z_k = \mathbb{1}_k [G]$

Can write this model as a standard linear regression

$$\mathbb{E}(Y|X, G) = \mathbb{E}(Y|X, Z) = X\beta + Z^T\theta$$

$\theta = (\theta_1, \theta_2, \theta_3)$

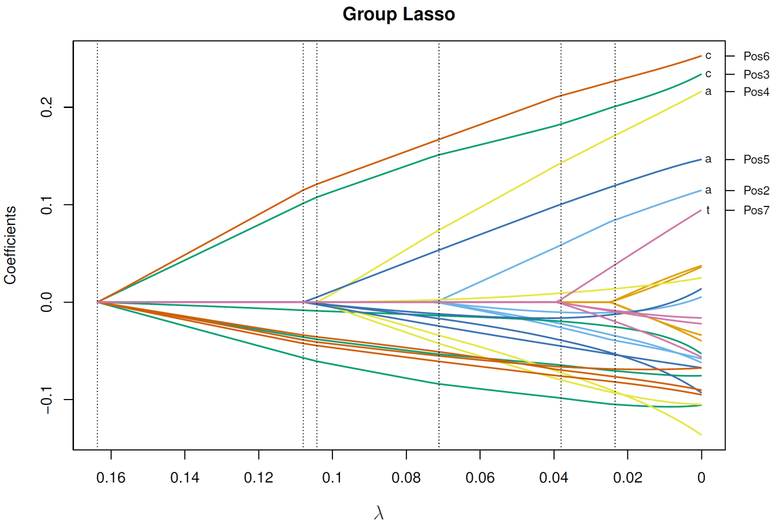
Example: Regression with multilevel factors 2

- Z is a group variable that represents the single factor G
- If the Variable G has no predictive power, then the full vector θ should be zero.
- When G is useful for prediction, then we expect that all coefficients of θ are likely nonzero.

We can have a number of such single and group variables, and so have models of the form:

$$\mathbb{E}(Y|X, G_1, \ldots, G_J) = \beta_0 + X^T\beta + \sum_{j=1}^{J} Z_j^T \theta_j$$

Coefficient path for a group-lasso fit



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 61

Sparse Group Lasso

When a group is included in a group-lasso fit, all the coefficients in that group are nonzero.

Sparse Group Lasso

When a group is included in a group-lasso fit, all the coefficients in that group are nonzero.

We want sparsity both with respect to which groups are selected, and which coefficients are nonzero within a group.

Short overview of Example:

Although a biological pathway may be implicated in the progression of a particular type of cancer, not all gene in the pathway need be active.

Sparse Group Lasso 2

In order to achieve within-group sparsity, augment with additional ℓ_1-penalty, leading to the convex program:

$$\underset{\{\theta_j \in \mathbb{R}^{p_j}\}_{j=1}^{J}}{\textit{minimize}} \left\{ \frac{1}{2} \|\mathbf{y} - \sum_{j=1}^{J} \mathbf{Z}_j \theta_j\|_2^2 + \lambda \sum_{j=1}^{J} \left[(1-\alpha) \|\theta_j\|_2 + \alpha \|\theta_j\|_1 \right] \right\}$$

$\alpha \in [0, 1]$

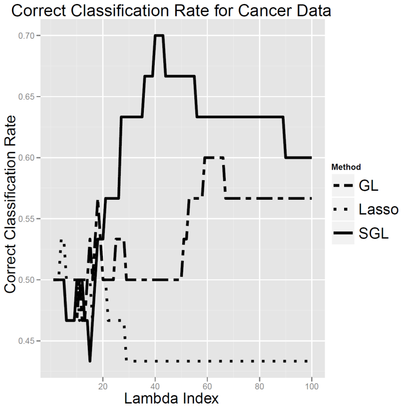
- $\alpha = 0$ group lasso
- $\alpha = 1$ lasso

Example breast cancer

- Dataset contains gene expression values from 60 patients with estrogen positive breast cancer
- Treated with medication for 5 years
- 28 recurrences
- Gene expression values were run
- Significant missing data
- First pass genes with more than 50% missingness were removed. Only 12071 of 22575 genes left
- Grouped genes by position data
- Final design matrix 4989 genes in 270 pathways
- 30 patient chosen at random, used $\alpha = 0.05$ for sparse-group lasso

Example breast cancer

- Sparse group lasso outperforms lasso and group lasso
- SGL includes 54 genes from 11 bands, GL selects all 74 genes form 15 bands, lasso selects 3 genes from separate bands

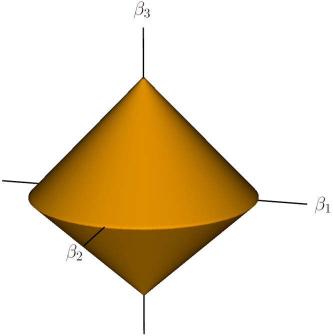


Source: https://web.stanford.edu/ hastie/Papers/SGLpaper.pdf, 5. Applications

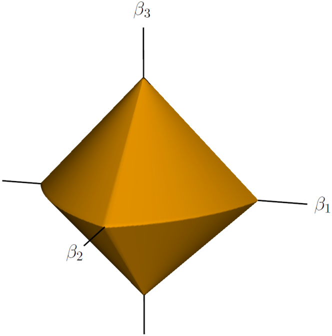
Constraint region

In two horizontal axes, constraint region resembles that of elastic net.

The group-lasso ball



The sparse group-lasso ball



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 64

$\alpha = 0.5$. Two groups with coefficients $\theta_1 = \{\beta_1, \beta_2\} \in \mathbb{R}^2$ and $\theta_2 = \beta_3 \in \mathbb{R}^1$

Overlap Group Lasso: Basic idea

Sometimes variables can belong to more than one group.

Short example:

Genes can belong to more than one biological pathway.

Overlap Group Lasso: Basic idea

Sometimes variables can belong to more than one group.

Short example:

Genes can belong to more than one biological pathway.

Example:

Case $p = 5$

$$Z_1 = (X_1, X_2, X_3) \quad Z_2 = (X_3, X_4, X_5)$$

Overlap Group Lasso: Basic idea

Sometimes variables can belong to more than one group.

Short example:

Genes can belong to more than one biological pathway.

Example:

Case $p = 5$

$$Z_1 = (X_1, X_2, X_3) \quad Z_2 = (X_3, X_4, X_5)$$

1. Replicate coefficients
2. Replicate variables

1. Replicate coefficients

For the problem before with

$$X = (X_1, \ldots, X_5) \text{ and } \beta = (\beta_1, \ldots, \beta_5)$$

we define:

$$\theta_1 = (\beta_1, \beta_2, \beta_3) \text{ and } \theta_2 = (\beta_3, \beta_4, \beta_5)$$

- Group lasso penalty $\|\theta_1\|_2 + \|\theta_2\|_2$

1. Replicate coefficients

For the problem before with

$$X = (X_1, \dots, X_5) \text{ and } \beta = (\beta_1, \dots, \beta_5)$$

we define:

$$\theta_1 = (\beta_1, \beta_2, \beta_3) \text{ and } \theta_2 = (\beta_3, \beta_4, \beta_5)$$

- Group lasso penalty $\|\theta_1\|_2 + \|\theta_2\|_2$
- Whenever $\hat{\theta}_1 = 0$ in any optimal solution, then we must have $\hat{\beta}_3 = 0$ in both groups
- Only possible sets of nonzero coefficients are: $\{1, 2\}$, $\{4, 5\}$ and $\{1, 2, 3, 4, 5\}$. Original groups are not a possibility

2. Replicate variables

- Replicates a variable in whatever group it appears, and then fits the ordinary group lasso as said.
- Variable X_3 replicated, and fit coefficient vectors $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ and $\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23})$
- Using group penalty $\|\theta_1\|_2 + \|\theta_2\|_2$

2. Replicate variables

- Replicates a variable in whatever group it appears, and then fits the ordinary group lasso as said.
- Variable X_3 replicated, and fit coefficient vectors $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ and $\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23})$
- Using group penalty $\|\theta_1\|_2 + \|\theta_2\|_2$

In terms of original variables, coefficient $\hat{\beta}_3$ of X_3 given by sum:

$$\hat{\beta}_3 = \hat{\theta}_{13} + \hat{\theta}_{21}$$

Variable X_3 has a better chance of being included in the model that other variables.

Solution overlap group lasso

So now the possible sets of nonzero coefficients for the overlap group lasso are:

$$\{1, 2, 3\}, \{3, 4, 5\} \text{ and } \{1, 2, 3, 4, 5\}$$

In general the sets of possible nonzero coefficients correspond to groups or the union of groups.

Solution overlap group lasso

So now the possible sets of nonzero coefficients for the overlap group lasso are:

$$\{1,2,3\}, \{3,4,5\} \text{ and } \{1,2,3,4,5\}$$

In general the sets of possible nonzero coefficients correspond to groups or the union of groups.

- $\nu_j \in \mathbb{R}^p$ is a vector which is zero everywhere except in those positions corresponding to member of the group j.
- $\mathcal{V}_j \subseteq \mathbb{R}^p$ subspace of possible vectors.
- For $X = (X_1, \ldots, X_p)$ the coefficient vector is given by $\beta = \sum_{j=1}^{J} \nu_j$

The overlap group lasso solves the problem:

$$\underset{\nu_j \in \mathcal{V}_j, j=1,\ldots,J}{\mathit{minimize}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}(\sum_{j=1}^{J} \nu_j)\|_2^2 + \lambda \sum_{j=1}^{J} \|\nu_j\|_2 \right\}$$

Example breast cancer

- Uses breast cancer gene expression dataset, consist of 8141 genes in 295 breast cancer tumors (78 metastatic 271 non-metastatic)
- Organize genes into overlapping gene sets (groups): pathways and edges

Pathways:

- 637 gene groups, average number of genes in each group is 23.7, largest gene group has 213 gene
- 3510 genes appear in the 637 groups with average appearance frequency of 4

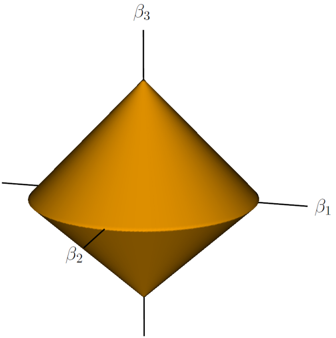
Edges:

- 42594 edges from the network
- 42594 overlapping gene sets of size 2
- All 8141 genes appear in the 42594 groups with an average appearance frequency of 10

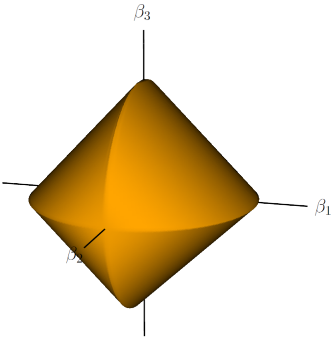
Source: http://fodava.gatech.edu/files/reports/FODAVA-11-27.pdf, 4.2 Gene Expression Data

Constraint region

The group-lasso ball



The overlap-group-lasso ball



Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 67

Two groups: $\{X_1, X_2\}$ and X_3

Two groups: $\{X_1, X_2\}$ and $\{X_2, X_3\}$

Sparse Additive Models: Introduction

Model given by:

- Zero-mean response variable $Y \in \mathbb{R}$
- Vector of predictors $X \in \mathbb{R}^J$
- Interested in estimating the regression function $f(x) = \mathbb{E}(Y|X = x)$

Sparse Additive Models: Introduction

Model given by:

- Zero-mean response variable $Y \in \mathbb{R}$
- Vector of predictors $X \in \mathbb{R}^J$
- Interested in estimating the regression function $f(x) = \mathbb{E}(Y|X = x)$

The class of sparse additive models limits these approximations further, by encouraging many of the components to be zero.

Sparse Additive Models: Introduction

Model given by:

- Zero-mean response variable $Y \in \mathbb{R}$
- Vector of predictors $X \in \mathbb{R}^J$
- Interested in estimating the regression function $f(x) = \mathbb{E}(Y|X = x)$

The class of sparse additive models limits these approximations further, by encouraging many of the components to be zero.

1. Backfitting
2. SPAM
3. COSSO
4. Combinations

Additive Models and Backfitting

Additive models are based on approximating the regression function by sums of the form:

$$f(x) = f(x_1, \ldots, x_J) \approx \sum_{j=1}^{J} f_j(x_j), \quad f_j \in \mathcal{F}_j, \quad j = 1, \ldots, J$$

- $\mathcal{F}_j$ are fixes set of univariate function classes
- Each $\mathcal{F}_j$ assumed to be a subset of $L^2(\mathbb{P}_j)$
- $\mathbb{P}_j$ is the distribution of covariate X_j equipped with squared $L^2(\mathbb{P}_j)$ norm $\|f_j\|_2^2 := \mathbb{E}\left[f_j^2(X_j)\right]$

Optimal Solutions

Best additive approximation to regression function $\mathbb{E}(Y|X = x)$ solves problem:

$$\underset{f_j \in \mathcal{F}_j j=1,\dots,J}{\text{minimize}} \ \mathbb{E}\left[\left(Y - \sum_{j=1}^{J} f_j(X_j)\right)^2\right]$$

The optimal solution $(\tilde{f}_1, \dots, \tilde{f}_J)$ is characterized by the backfitting equations:

$$\tilde{f}_j(x_j) = \mathbb{E}\left[Y - \sum_{k \neq j} \tilde{f}_k(X_k) | X_j = x_j\right], \text{ for } j = 1, \dots, J$$

SPAM Sparse Additive Model

- Extension of basic additive model is sparse additive Model
- Assume there is subset $S \subset \{1, 2, \ldots, J\}$
- $f(x) = \mathbb{E}(Y|X = x) \approx \sum_{j \in S} f_j(x_j)$

For given sparsity level $k \in \{1, \ldots, J\}$ best k-sparse approximation to regression function is given by:

$$\underset{|S|=k, f_j \in \mathcal{F}_j, j=1,\ldots,J}{\text{minimize}} \mathbb{E}\left[\left(Y - \sum_{j \in S} f_j(X_j)\right)^2\right]$$

Nonconvex and computationally intractable!

SPAM Problem

- Instead measure the sparsity of an additive approximation $f = \sum_{j=1}^{J} f_j$ via the sum $\sum_{j=1}^{J} \|f_j\|_2$
- $\|f_j\|_2 = \sqrt{\mathbb{E}\left[f_j^2(X_j)\right]}$

For $\lambda \geq 0$ type of best sparse approximation:

$$\underset{f_j \in \mathcal{F}_j j=1,\dots,J}{\text{minimize}} \left\{ \mathbb{E}\left[\left(Y - \sum_{j \in S} f_j(X_j)\right)^2\right] + \lambda \sum_{j=1}^{J} \|f_j\|_2 \right\}$$

Convex function of $(f_1, \dots, f_J)$

- SPAM combines ideas from sparse linear modeling and additive nonparametric regression
- Can obtain effective fit even when the number of covariates is larger than the sample size

Additive smoothing-spline model

Form of an additive smoothing-spline model, obtained form the optimization of a penalized objective function:

$$\underset{f_j \in \mathcal{H}_j j=1,\dots,J}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{J} f_j(x_{ij}))^2 + \lambda \sum_{j=1}^{J} \frac{1}{\gamma_j} \|f_j\|^2_{\mathcal{H}_j} \right\}$$

$\|f_j\|_{\mathcal{H}_j}$ is an appropriate Hilbert-space norm for the j^{th} coordinate.

COSSO

The COSSO (Component Selection and Smoothing Operator) method is based on the objective function:

$$\underset{f_j \in \mathcal{H}_j j=1,\dots,J}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{J} f_j(x_{ij}))^2 + \tau \sum_{j=1}^{J} \|f_j\|_{\mathcal{H}_j} \right\}$$

Multiple Penalization

- Multiple ways of enforcing sparsity for a nonparametric problem. (SPAM backfitting, COSSO)
- SPAM backfitting base on a combination of ℓ_1-norm: $\|f\|_{N,1} := \sum_{j=1}^{J} \|f_j\|_N$ with $\|f_j\|_N^2 := \frac{1}{N} \sum_{j=1}^{J} f_j^2(x_{ij})$
- COSSO method uses combination of the ℓ_1-norm with the Hilbert norm: $\|f\|_{\mathcal{H},1} := \sum_{j=1}^{J} \|f_j\|_{\mathcal{H}}$

Multiple Penalization

- Multiple ways of enforcing sparsity for a nonparametric problem. (SPAM backfitting, COSSO)
- SPAM backfitting base on a combination of ℓ_1-norm: $\|f\|_{N,1} := \sum_{j=1}^{J} \|f_j\|_N$ with $\|f_j\|_N^2 := \frac{1}{N}\sum_{j=1}^{J} f_j^2(x_{ij})$
- COSSO method uses combination of the ℓ_1-norm with the Hilbert norm: $\|f\|_{\mathcal{H},1} := \sum_{j=1}^{J} \|f_j\|_{\mathcal{H}}$

General family of estimator:

$$\min_{f_j \in \mathcal{H}_j j=1,\ldots,J} \left\{ \frac{1}{N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{J} f_j(x_{ij}))^2 + \lambda_{\mathcal{H}} \sum_{j=1}^{J} \|f_j\|_{\mathcal{H}_j} + \lambda_N \sum_{j=1}^{J} \|f_j\|_N \right\}$$

Why better?

- Yields an estimator that is minmax-optimal
- Therefore its convergence rate (as a function of sample size, problem dimension and sparsity) is the fastest possible

Theoretical Results for Lasso: Variable-Selection Consistency

Consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\beta^* + w,$$

and the corresponding lasso

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2N} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda_N ||\beta||_1 \qquad (\star)$$

- $\mathbf{X} \in \mathbb{R}^{N \times p}$: design matrix
- $\mathbf{w} \in \mathbb{R}^N$: noise, i.i.d. with $N(0, \sigma^2)$
- $\beta^* \in \mathbb{R}^p$: unknown coefficient
- λ_N: Lagrange multiplier

Support Recovery

Whether or not a lasso estimate $\hat{\beta}$ has nonzero entries in the same positions as the true regression vector β^*

Background

Definition($k-$sparse)

A vector β is $k-$sparse, if it is supported on a subset $S = S(\beta)$ of cardinality $k = |S|$.

Now, the previous question becomes:

Question

Given an optimal lasso solution $\hat{\beta}$ with support set $\hat{S}$, when is $\hat{S} = S$?

S: support of true regression vector β^*

Variable-Selection Consistency(Sparsistency)

Variable-Selection Consistency: Some Conditions

Let **X** be the design matrix.

- Column Normalization Condition

$$\max_{j=1,\cdots,p} \frac{||\mathbf{x}_j||_2}{\sqrt{N}} \leq K,$$

- Eigenvalue Condition

$$\lambda_{\min}\left(\frac{\mathbf{X}_S^T \mathbf{X}_S}{N}\right) \geq C_{\min},$$

- Irrepresentability Condition
  There exists some $\gamma > 0$ such that

$$\max_{j \in S^c} ||(\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{x}_j||_1 \leq 1 - \gamma$$

$\lambda_{\min}(A)$: smallest eigenvalue of matrix A
K, C_{min}: constants

Interpretation of These Conditions

Column Normalization Condition

$$\max_{j=1,\cdots,p} \frac{||\mathbf{x}_j||_2}{\sqrt{N}} \leq K$$

Make sure the design matrix $\mathbf{X}$ has normalized columns

Eigenvalue Condition

$$\lambda_{\min}\left(\frac{X_S^T X_S}{N}\right) \geq C_{\min}$$

Make sure the submatrix $\mathbf{X}_S$ is well behaved: if this condition were violated, columns of $\mathbf{X}_S$ would be linearly dependent $\Rightarrow$ impossible to estimate β^* even when the support set S is known

Interpretation of These Conditions

Irrepresentability

$$\max_{j \in S^c} ||(\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{x}_j||_1 \leq 1 - \gamma$$

- $\mathbf{X}_S \in \mathbb{R}^{N \times k}$: subset of covariates in the support set
- For each $j \in S^c$, the $k-$vector $(\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{x}_j$ is a regression coefficient of $\mathbf{x}_j$ on $\mathbf{X}_S$, a measure of how $\mathbf{x}_j$ aligns with the columns of submatrix $\mathbf{X}_S$
- Desirable case: $\mathbf{x}_j, j \in S^c$ orthogonal to columns of $\mathbf{X}_S \Rightarrow \gamma = 1$
- In high dimensional settings ($p \gg N$), complete orthogonality is impossible, we hope for a type of near orthogonality to hold

Variable-Selection Consistency

Theorem(Variable-Selection Consistency)

Suppose X satisfies all 3 conditions above, consider the lasso $(\star)$ with:

$$\lambda_N \geq \frac{8K\sigma}{\gamma}\sqrt{\frac{\log p}{N}}, \qquad \lambda_N : \text{Lagrange multiplier}$$

Then the following properties hold with probability greater than $1 - c_1 e^{-c_2 N \lambda_N^2}$:

(a) Uniqueness: Optimal solution $\hat{\beta}$ is unique

(b) No false inclusion: $S(\hat{\beta}) \subset S(\beta^*)$

(c) l_∞−bounds:
The error $\hat{\beta} - \beta^*$ satisfies the l_∞ bound

$$||\hat{\beta}_S - \beta^*_S||_\infty \leq \underbrace{\lambda_N[\frac{4\sigma}{\sqrt{C_{\min}}} + ||(X_S^T X_S/N)^{-1}||_\infty]}_{B(\lambda_N, \sigma; \mathbf{X})}$$

(d) No false exclusion: for all $j \in S(\beta^*)$ such that $|\beta^*_j| > B(\lambda_N, \sigma; \mathbf{X})$, $j \in S(\hat{\beta})$

Interpretation of claims in the theorem

- Uniqueness claim (a) allows us to talk unambiguously about the support of the lasso estimate $\hat{\beta}$
- (b) guarantees the lasso does not falsely include variables that are not in the true support of β^*
- (c) guarantees $\hat{\beta}_S$ is uniformly close to β^*_S in the sense of the $l_\infty -$norm
- (d) : consequence of (b)+(c): for any $j \in S(\beta^*)$, as long as the value of $|\beta^*_j|$ is not too small, lasso will also include the variable associated with the index j, i.e., lasso is variable-selection consistent in the full sense

How do the 3 conditions influence the results of the theorem?

Brief sketch of the proof:

- Based on a construction procedure: **primal-dual witness method(PDW)**
- When this procedure succeeds, it constructs an optimal primal-dual pair $(\hat{\beta}, \hat{z}) \in \mathbb{R}^p \times \mathbb{R}^p$ that acts as a witness for the fact that lasso has a unique optimal solution with correct signed support
- Construction procedure:
  - Set $\hat{\beta}_{S^c} = 0$
  - Determine $\hat{\beta}_S, \hat{z}_S$ by solving subproblem

$$\hat{\beta}_S \in \arg \min_{\beta_S \in \mathbb{R}^k} \left\{ \frac{1}{2N} ||\mathbf{y} - \mathbf{X}_S \beta_S||_2^2 + \lambda_N ||\beta_S||_1 \right\}$$

    $\hat{z}_S$: a subdifferential of $||\hat{\beta}_S||_1$
  - Solve for $\hat{z}_{S^c}$ via zero-subgradient condition:

$$\frac{1}{N} \mathbf{X}^T (\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda_N \hat{z} = 0$$

    and check whether the **strict dual feasibility condition** $||\hat{z}_{S^c}||_\infty < 1$ holds
  - If this condition holds
    $\Rightarrow$ construction succeeds
    $\Rightarrow$ variable-selection consistency is certified

How do the 3 conditions influence the results of the theorem?

- Eigenvalue Condition
  - $\Rightarrow$ the subproblem is strictly convex $\Rightarrow$ has a unique minimizer $\Rightarrow (\hat{\beta}_S, 0) \in \mathbb{R}^p$: unique optimal solution of the lasso (Uniqueness) (p.307)
  - $\Rightarrow$ invertibility of $\mathbf{X}_S^T \mathbf{X}_S \Rightarrow$ solve for explicit expression of $\hat{\beta}_S - \beta_S^* \Rightarrow$ establishing l_∞ bounds of $||\hat{\beta}_S - \beta_S^*||_\infty$ (p.308-p.309)
- Uniqueness and No False Inclusion $\Leftarrow$ Establishing strict dual feasibility : $||\hat{z}_{S^c}||_\infty < 1$

$$\hat{z}_{S^c} = \underbrace{\mathbf{X}_{S^c}^T \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathrm{sign}(\beta_S^*)}_{\mu} + \underbrace{\mathbf{X}_{S^c}^T [\mathbf{I} - \mathbf{X}_S (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T] (\frac{\mathbf{w}}{\lambda_N N})}_{V_{S^c}}$$

Apply triangle inequality $\Rightarrow$

$$||\hat{z}_{S^c}||_\infty \leq ||\mu||_\infty + ||V_{S^c}||_\infty$$

  - Irrepresentability Condition $\Rightarrow ||\mu||_\infty \leq 1 - \gamma$
  - V_{S^c}: zero-mean Gaussian random vector, using Column Normalization Condition
    $\Rightarrow V_j$: zero-mean with variance $\leq \sigma^2 K^2 / (\lambda_N^2 N)$
    $\Rightarrow \mathbb{P}[||V_{S^c}||_\infty \geq \gamma/2]$ vanishes at rate $2e^{-\lambda_N^2 N}$ for λ_N given in the theorem statement
    $\Rightarrow ||V_{S^c}||_\infty < \gamma$ holds with high probability $\Rightarrow ||\hat{z}_{S^c}||_\infty < 1$

Numerical Studies

In order to learn about the impact of the 3 conditions on the results of the theorem in practice, we take the irrepresentability condition as an example, and ran a small simulation to examine how this condition influences the lasso solution

Simulation Settings

- $p = 500$, $N = 1000$, with $k = 15$ having nonzero coefficients
- Generate a range of p variables i.i.d. standard Gaussian variates, with k of them in the support set S
- For each $j \in S$, randomly choose a predictor $l \in S^c$, set $\mathbf{x}_l \leftarrow \mathbf{x}_l + c \cdot \mathbf{x}_j$, with c chosen s.t. corr($\mathbf{x}_j, \mathbf{x}_l$)=$\rho$
- $\mathbf{x}_j$: true predictor,$\mathbf{x}_l$: null predictor partner of $\mathbf{x}_j$
- Response $\mathbf{y} = \mathbf{X}_S\beta_S + \mathbf{w}$, with elements of $\mathbf{w}$ i.i.d. $N(0, 1)$
- All the nonzero regression coefficients in β_S chosen to be 0.25 with randomly selected signs
- λ_N : chosen in an optimal way in each run i.e., use the value yielding the correct number of nonzero coefficients

Impacts of Irrespresentability Condition on lasso solution

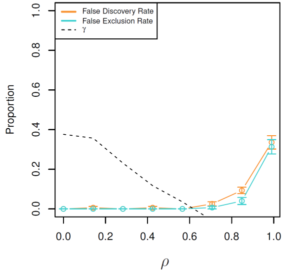


Abbildung: Average false discovery and false exclusion rates from simulations with $p = 500$ variables

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 306

- Average false discovery and false exclusion probabilities are zero until ρ is greater than about 0.6, after this point, value of γ drops below 0
  $\Rightarrow$ irrepresentability condition does not hold
  $\Rightarrow$ lasso starts to include false variables and exclude good ones due to high correlation between signal and noise variables

Conclusion

- For different datasets we have different models
- Penalty terms depend on l_2- or l_1- norms or combination of both
- For different penalties the constraint region varies
- In linear regression problem, the lasso solution is unique and approximates the true solution with high accuracy under certain conditions

Questions