Variable selection in regression analysis

ISyE 8813 - Lecture slides

Loosely based on Chapter 3 of *The Elements of Statistical Learning* by Hastie et al. (2009)

Overview of regression

Linear model

We assume the following **linear model** for data:

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

where:

- $\mathbf{y} = (y_1, \cdots, y_n) \in \mathbb{R}^n$ is the vector of observed responses,
  - Assume $\mathbf{y}$ is centered, i.e., $\sum_{i=1}^{n} y_i = 0$, hence no intercept
- $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj}) \in \mathbb{R}^n$ is the j-th covariate vector,
- $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ is the model matrix,
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p) \in \mathbb{R}^p$ is the coefficient vector,
- $\{\epsilon_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ is the observation noise.

Linear model

Why **linear**?

- After transformations, often a reasonable approximation for many applications
- Efficiency in variable selection

What **inputs** can be modeled?

- Quantitative inputs
- Basis expansions
  - e.g., polynomial, spline, wavelet
- Numeric or “dummy” coding of qualitative inputs
  - e.g., five-level factor coded as $1, \cdots, 5$)
- Interactions between inputs

Least-squares estimation

- Want to select and estimate coefficients $\boldsymbol{\beta}$ using data $(\mathbf{X}, \mathbf{y})$.
- Most popular estimation method is **least-squares estimation (LSE)**, which minimizes the residual-sum-squares (RSS):

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

  - Intuition: Obtaining the hyperplane-of-best-fit to data
- If noise is Gaussian, same as maximum-likelihood estimation

Least-squares estimation

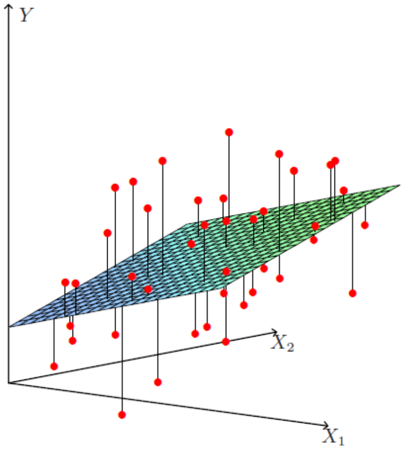


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Least-squares estimation

The minimization can be performed in closed-form:

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \nabla_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) &= -2\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \overset{\mathrm{set}}{=} 0 \\ \hat{\boldsymbol{\beta}}_{\mathrm{LS}} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}. \end{aligned}$$

Using this estimator, the fitted values at the training inputs are:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.$$

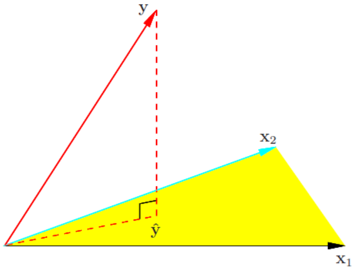


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors.*

Gauss-Markov Theorem

Theorem (Gauss-Markov)

For any linear unbiased estimator $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$ *with* $\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\mathit{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) \preceq \mathit{Var}(\hat{\boldsymbol{\beta}})$.

- In other words, the variance from the LSE estimator is optimal among all linear estimators of $\boldsymbol{\beta}$
- But is this enough?

А. А. Марковъ (1886).

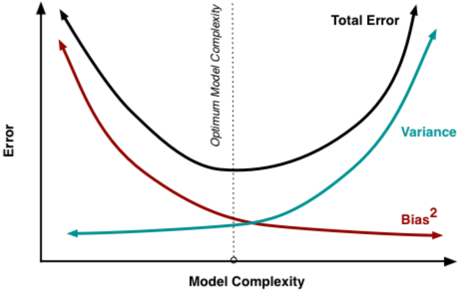
Bias-variance trade-off

- Let $\mathbf{x}_{new} \in \mathbb{R}^p$ be a new input setting, with Y its observation from the linear model
- Consider the following decomposition of prediction error:

$$\begin{aligned}\mathbb{E}\left[\left(Y - \mathbf{x}_{new}^T \hat{\beta}\right)^2\right] &= \left(\mathbf{x}_{new}^T \mathbb{E}\left[\hat{\beta} - \beta\right]\right)^2 \\ &\quad + \mathbb{E}\left[\left(\mathbf{x}_{new}^T \hat{\beta} - \mathbf{x}_{new}^T \beta\right)^2\right] + \sigma^2 \\ &= \text{Bias}^2 + \text{Variance} + \text{Observation Error}\end{aligned}$$

- The estimator $\hat{\beta}$ should jointly reduce prediction bias and variance. However, a decrease in one often leads to an increase in the other; this is the **bias-variance trade-off**

Model selection



This motivates the need for **model selection**:

- Selecting which variables are active provides a way to control the bias-variance trade-off, which leads to better predictions
- When many variables are considered, model selection provides a more interpretable model using a small subset of variables

Convex penalties

Penalized selection

Penalized selection optimizes the following problem:

$$\min_{\beta} \left[\mathrm{RSS}(\beta) + \lambda \sum_{j=1}^{p} P(\beta_j) \right],$$

where $P(\beta) \geqslant 0$ is a penalty function.

- P should have the increasing property:

$$P(\beta) \geqslant P(\beta') \text{ for } |\beta| \geqslant |\beta'|$$

- This forces the optimization to set most coefficients in β to 0, thereby performing selection

LASSO

Tibshirani (1996) proposed the popular **LASSO** method (least absolute shrinkage and selection operator), which optimizes:

$$\hat{\boldsymbol{\beta}}_n(\lambda) \equiv \min_{\boldsymbol{\beta}} \left[\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

This can be stated in the equivalent primal form:

$$\hat{\boldsymbol{\beta}}_n(t) \equiv \min_{\boldsymbol{\beta}} \left\{ \mathrm{RSS}(\boldsymbol{\beta}) \ : \ \sum_{j=1}^{p} |\beta_j| \leqslant t \right\},$$

which can be viewed as the tightest convex relaxation of the desired (discrete) model selection problem:

$$\min_{\boldsymbol{\beta}} \left\{ \mathrm{RSS}(\boldsymbol{\beta}) \ : \ \sum_{j=1}^{p} 1\{\beta_j \neq 0\} \leqslant t \right\},$$

LASSO: Motivation

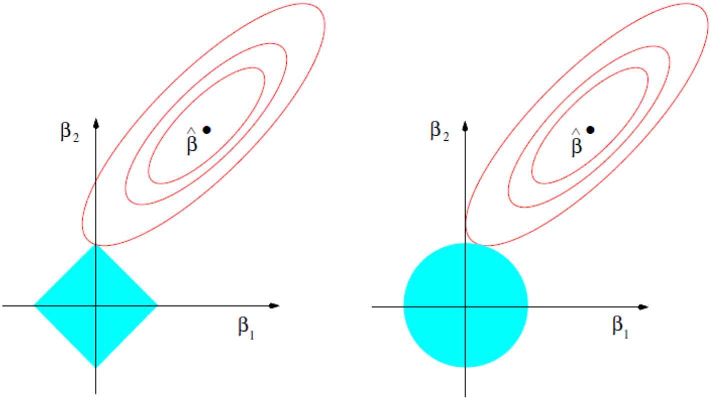


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

LASSO: Theoretical properties

We want a method which selects the correct model as the number of observations $n \to \infty$, i.e.:

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\beta}_n(\lambda) =_s \beta\right) = 1,$$

where $=_s$ denotes sign equality.

This is indeed true for LASSO:

Theorem (Zhao and Yu, 2006)

Under regularity conditions on $\mathbf{X}$, LASSO is **selection consistent** *if the penalty parameter λ_n satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{(1+c)/2} \to \infty$ for all $0 \leqslant c < 1$.*

LASSO: Application to prostate dataset

Consider the prostate cancer study by Stamey et al. (1989):

- **Response:** Prostate-specific antigen levels
- **Predictors:**
  - Log cancer volume (`lcavol`)
  - Log prostate weight (`lweight`)
  - age
  - Log benign prostatic hyperplasia (`lbph`)
  - Seminal vesicle invasion (`svi`)
  - Log capsular penetration (`lcp`)
  - Gleason score (`gleason`)
  - % of Gleason scores 4 or 5 (pgg45)

LASSO: Application to prostate dataset

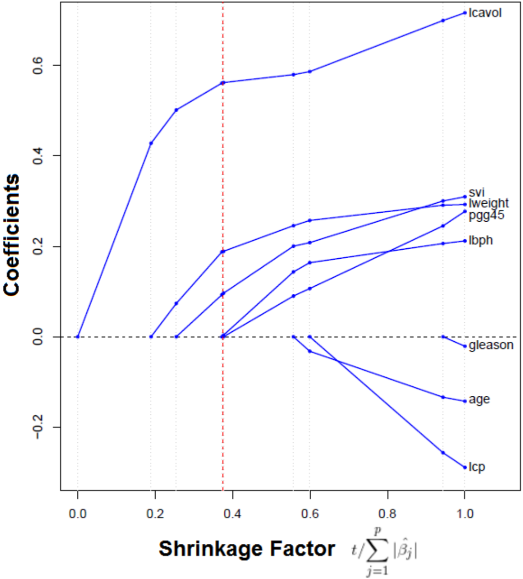
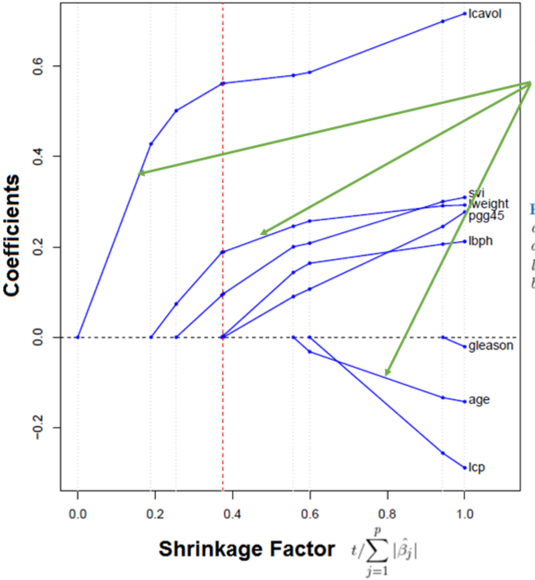


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation.*

LASSO: Application to prostate dataset



Observation:

LASSO path is piecewise linear and continuous in t!

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p|\hat{\beta}_j|$. *A vertical line is drawn at* $s = 0.36$, *the value chosen by cross-validation.*

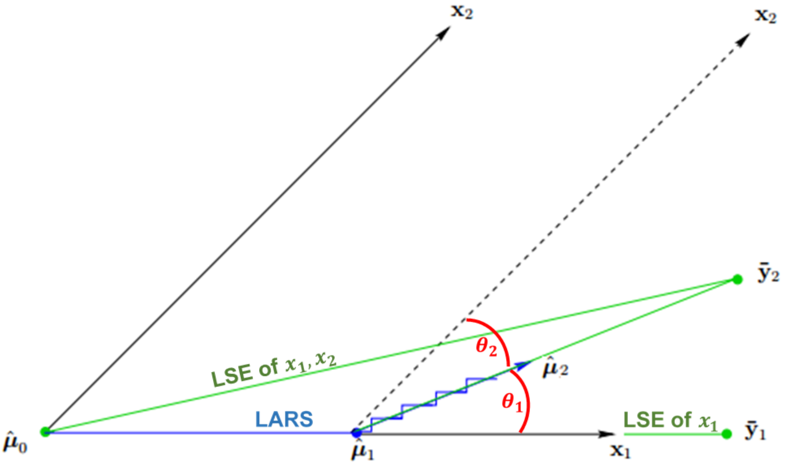
LASSO: Optimization using LARS

Least-angle regression (LARS, Efron et al., 2004) is an efficient way for solving the LASSO path $\mathcal{P} \equiv \{\hat{\beta}_n(t), t > 0\}$:

- Motivated by piecewise linearity and continuity of $\mathcal{P}$
- Algorithm:
  1. Begin with empty active set $\mathcal{A} = \emptyset$ and residual $\mathbf{r} = \mathbf{y}$
  2. Add to $\mathcal{A}$ the variable j with smallest angle $\cos\left\{|\mathbf{x}_j^T \mathbf{r}|/\|\mathbf{r}\|\right\}$, i.e., the variable with largest correlation $|\mathbf{x}_j^T \mathbf{r}|^2/\|\mathbf{r}\|^2$
  3. Move LARS solution in the direction of the LSE for $\mathcal{A}$, and update residual $\mathbf{r}$.
  4. Stop when a non-active variable has smallest angle with $\mathbf{r}$, and go to Step 2.
- See Algorithm 3.2 in Hastie et al. (2009) for details

LASSO: Optimization using LARS

Visualization:



LASSO: Optimization using LARS

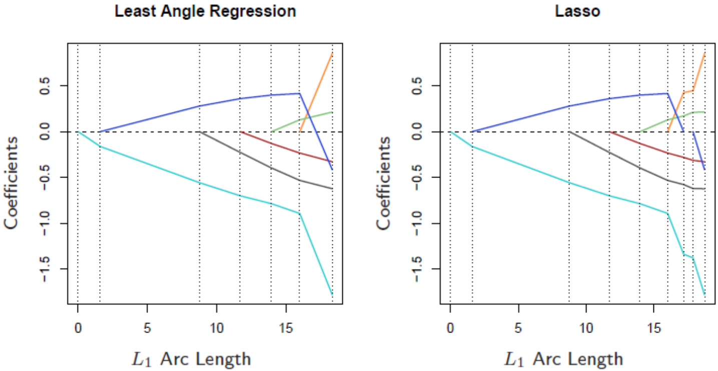


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.*

LASSO: Optimization using coordinate descent

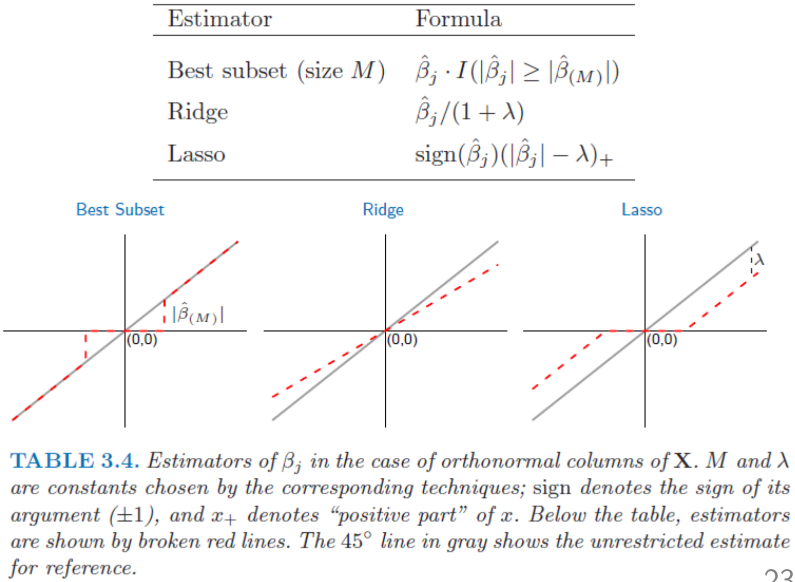
When many variables are considered ($p \gg 1$), LARS can be computationally expensive. State-of-the-art algorithms employ a technique called **coordinate descent**:

- Idea dates back to the Gauss-Seidel method from 1823
- Iteratively optimizes each coefficient β_k with other coefficients $\{\beta_k\}_{k \neq j}$ fixed.
- For LASSO, this coordinate optimization for β_j has closed-form minimizer:

$$S\left\{\mathbf{x}_j^T \left(\mathbf{y} - \sum_{k=1, k\neq j}^{n} \mathbf{x}_k \beta_k\right); \lambda\right\},$$

where $S\{z; \lambda\} = \text{sgn}(z)(|z| - \lambda)_+$ is the soft-thresholding operator in Donoho (1995).

LASSO: Optimization using coordinate descent

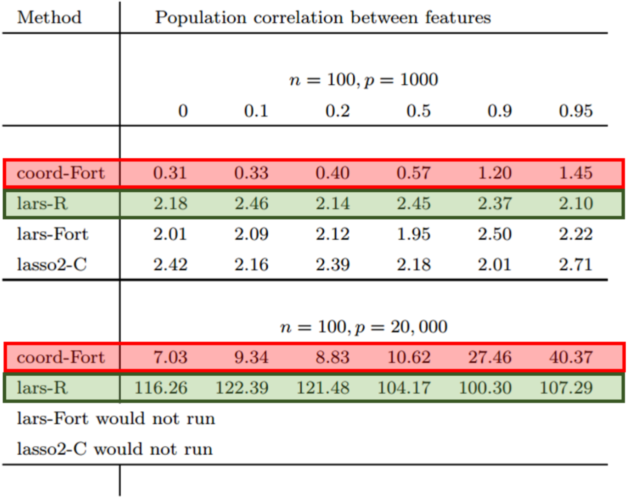


Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

Best Subset
$\lvert\hat{\beta}_{(M)}\rvert$
(0,0)
Ridge
(0,0)
Lasso
λ
(0,0)

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of $\mathbf{X}$. M and λ are constants chosen by the corresponding techniques; sign denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

LASSO: Optimization using coordinate descent



Method	Population correlation between features					
	$n = 100, p = 1000$					
	0	0.1	0.2	0.5	0.9	0.95
coord-Fort	0.31	0.33	0.40	0.57	1.20	1.45
lars-R	2.18	2.46	2.14	2.45	2.37	2.10
lars-Fort	2.01	2.09	2.12	1.95	2.50	2.22
lasso2-C	2.42	2.16	2.39	2.18	2.01	2.71
	$n = 100, p = 20,000$					
coord-Fort	7.03	9.34	8.83	10.62	27.46	40.37
lars-R	116.26	122.39	121.48	104.17	100.30	107.29
lars-Fort would not run						
lasso2-C would not run						

- Observation: Coordinate descent much faster than LARS for $n, p \gg 1$!

Non-negative garrote (NNG)

- Brieman (1995) proposed the **non-negative garrote** (NNG), which optimizes:

$$\hat{\mathbf{d}} = \min_{\mathbf{d}} \left[\mathrm{RSS}(\hat{\beta}_{LS} \odot \mathbf{d}) + \lambda \sum_{j=1}^{p} d_j, \; d_j \geqslant 0 \; \forall j = 1, \cdots, p \right],$$

  where $\odot$ is the Hadamard (element-wise) product.
- The resulting estimator for NNG is $\hat{\beta}(\lambda) = \hat{\beta}_{LS} \odot \hat{\mathbf{d}}$

NNG: Comparison with LASSO

Advantages:

- Stable selection method, often outperforming LASSO when $n \geqslant p$ (# observations $\geqslant$ # variables)
- For small p, efficient optimization using **quadratic programming** (QP)

Disadvantages:

- Performs poorly when $n < p$ (# observations $<$ # variables), due to reliance on LSE
- QPs are computationally expensive for large p

NNG: Optimization using QP

When $n \geqslant p$, the NNG problem can be reformulated as a QP (try as exercise), which has general form:

$$\min_{x} \left[\frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x}, \ \mathbf{A}\mathbf{x} \leqslant \mathbf{b} \right].$$

QPs can be solved efficiently using:

- interior point methods,
- active set optimization,
- augmented Lagrangian penalization,
- extensions of the simplex algorithm

See Nocedal and Wright (2006) for details.

NNG: Optimization using LARS

Two drawbacks of NNG are that:

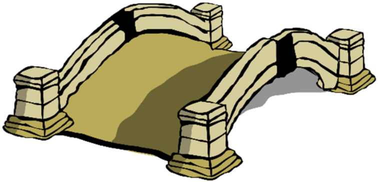
- it performs poorly for $n < p$,
- QPs are computationally expensive for large p.

These two problems are addressed in Yuan and Lin (2007), who:

- used LASSO to generate initial estimates for NNG,
- proposed a LARS-like algorithm to efficiently generate the NNG path,
- proved the resulting estimator is both estimation- and selection-consistent.

Non-convex penalties

Non-convex penalties: Motivation



- Selection consistency of LASSO relies on the **irrepresentability condition** (Zhao and Yu, 2006), which prevents variables from being "too correlated"
- But observational data are often highly correlated in practice, particularly in biology and social sciences!
- Non-convex penalties address this by bridging the gap between the l_1-norm relaxation in LASSO and the l_0-norm desired for selection.

Non-convex penalties

Many flavors proposed in the literature:

- **Bridge (power) penalty** (Frank & Friedman, 1993):

$$P(\beta_j) = |\beta_j|^\gamma, \quad \gamma \in (0, 1],$$

- **SCAD penalty** (Fan & Li, 2001):

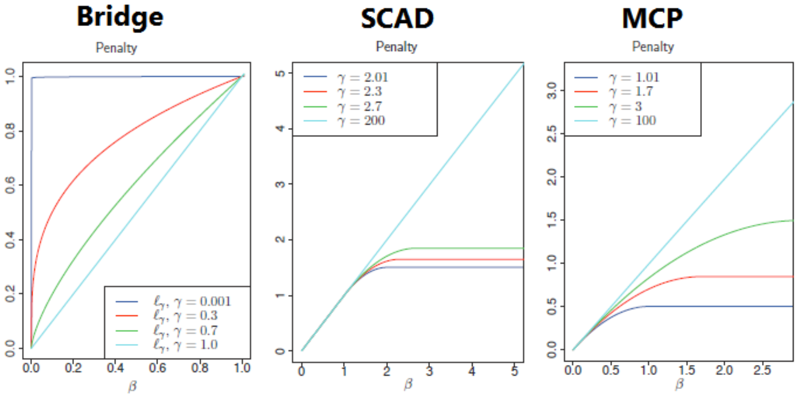
$$P(\beta_j) = \int_0^{|\beta_j|} \min\left\{1, \frac{(\gamma - t/\lambda)_+}{\gamma - 1}\right\} dt, \quad \gamma > 2,$$

- **Minimax concave penalty (MCP)** (Zhang, 2010):

$$P(\beta_j) = \int_0^{|\beta_j|} (1 - t/(\gamma\lambda))_+ \ dt, \quad \gamma > 1.$$

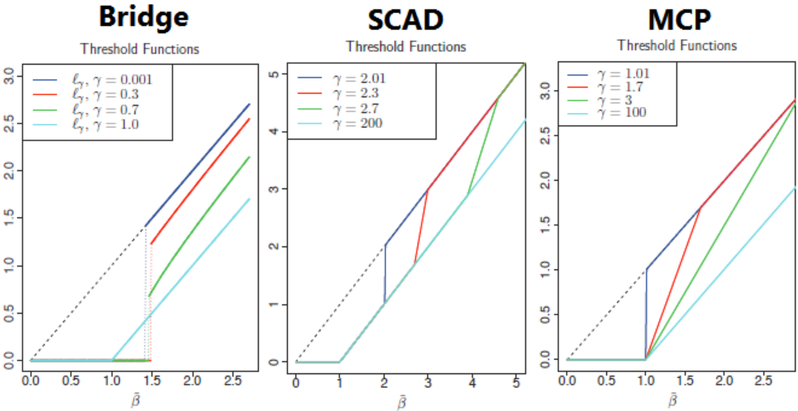
Non-convex penalties

Visualization of non-convex penalties:



Non-convex penalties: Coordinate descent

Similar to LASSO, the coordinate optimization for these non-convex penalties have closed-form minimizers (called **threshold functions**):



See Mazumder et al. (2011) for details.

Hierarchy and heredity

Hierarchy and heredity: Motivation

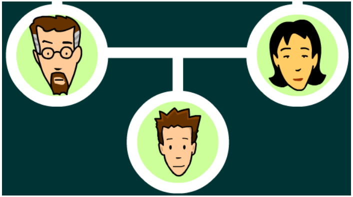
Until now, we considered only the general setting where there is no relationships between variables.

In practice, variables often have an innate structure which can be further exploited (see Wu and Hamada, 2009):

- **Hierarchy:** Some variables are more important than others, forming a hierarchy,
- **Heredity:** Some variables can be active only when other variables (called parent effects) are active
  - e.g., a two-factor interaction is active only if one or both of its parent main effects are also active.

Both occur naturally in engineering, in genetics and more generally, in designed experiments.

Hierarchy and heredity: designed experiments



For designed experiments, Yuan, Joseph and Lin (2007) generalized LARS to incorporate hierarchy and heredity:

- For a variable j, define its dependency set $\mathcal{D}_j$ as its parent effects,
- Instead of the variable with highest correlation, the modified LARS picks the variable j with the highest average correlation:

$$\frac{1}{1+\#\{\mathcal{D}_j\}}\|\mathbf{X}_{j\cup\mathcal{D}_j}^T\mathbf{r}\|^2/\|\mathbf{r}\|^2,$$

where the columns of $\mathbf{X}_{j\cup\mathcal{D}_j}$ correspond to j and $\mathcal{D}_j$.

Hierarchy and heredity: observational data

From this, several approaches have been proposed for incorporating hierarchy and heredity in the model selection of observational data:

- Zhao, Rocha and Yu (2009): Uses composite absolute penalties to select hierarchical variables,
- Bien, Taylor and Tibshirani (2013): Selects hierarchical interactions using a convex-constrained LASSO,
- Lim and Hastie (2013): Selects hierarchical interactions using a group-LASSO formulation.

Summary

- **Model selection** is necessary for two reasons:
  - to reduce prediction error in the bias-variance trade-off,
  - to obtain a more interpretable model.
- LASSO provides a convex relaxation of this selection problem, and can be solved via LARS or coordinate descent
- NNG works well in practice when paired with LASSO
- Non-convex penalties are necessary when variables are highly correlated
- More elaborate selection methods are needed when variables have known structures, such as hierarchy or heredity