Biostatistics-Lecture 16 Model Selection

Ruibin Xi

Peking University

School of Mathematical Sciences

Motivating example1

- Interested in factors related to the life expectancy (50 US states,1969-71)
  - Per capita income (1974)
  - Illiteracy (1970,percent of population)
  - Murder rate per 100,000 population
  - Percent high-school graduates
  - Mean number of days with min temperature < 30 degree
  - Land area in square mile

Motivating Example2

- The role of microRNA on regulating gene expression
  - Response: standard deviation of a gene expression
  - Covariates:
    - mean gene expression
    - length of the 3’-UTRs
    - number of microRNA targets in the 3’-UTRs
    - mean target score of the microRNA targets
    - number of common SNPs in the 3’-UTR
    - mean of minor allele frequencies of common SNPs in 3’-UTRs

Motivating Example2

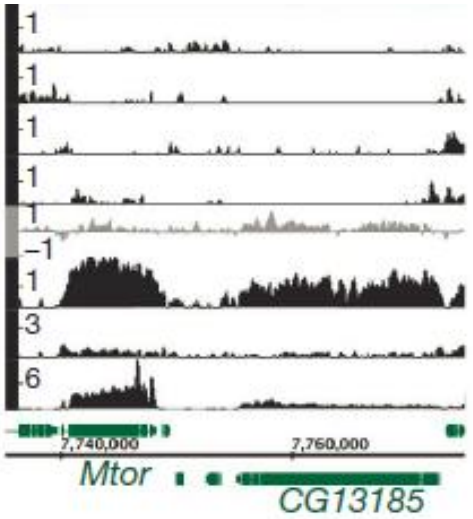
	Expression SD	Expression Mean	Gene Length	Length of 3' UTR	Number of miRNA targets	Mean target score	Number of SNPs	Mean MAF
Expression SD	1	0.952031	0.084030	0.068084	0.146080	0.135781	0.035820	0.033893
Expression Mean	0.952031	1	0.157620	0.120004	0.165147	0.156963	0.055338	0.051066
Gene Length	0.084030	0.157620	1	0.471435	0.406605	0.311641	0.216913	0.173899
Length of 3' UTR	0.068084	0.120004	0.471435	1	0.246593	0.206723	0.227899	0.197424
Number of miRNA targets	0.146080	0.165147	0.406605	0.246593	1	0.849602	0.185446	0.142214
Mean target score	0.135781	0.156963	0.311641	0.206723	0.849602	1	0.151916	0.128167
Number of SNPs	0.035820	0.055338	0.216913	0.227899	0.185446	0.151916	1	0.947236
Mean MAF	0.033893	0.051066	0.173899	0.197424	0.142214	0.128167	0.947236	1

Motivating Example3

- Communities and Crime
  - Response: total number of violent crimes per 100K population
  - Covariates (128):
    - population for community
    - percentage of population that is caucasian
    - percentage of population that is african american
    - median household income
    - per capita income for african americans
    - percentage of kids born to never married
    - number of vacant households
    - number of sworn full time police officers
    - ….
  - http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

Motivating Example4

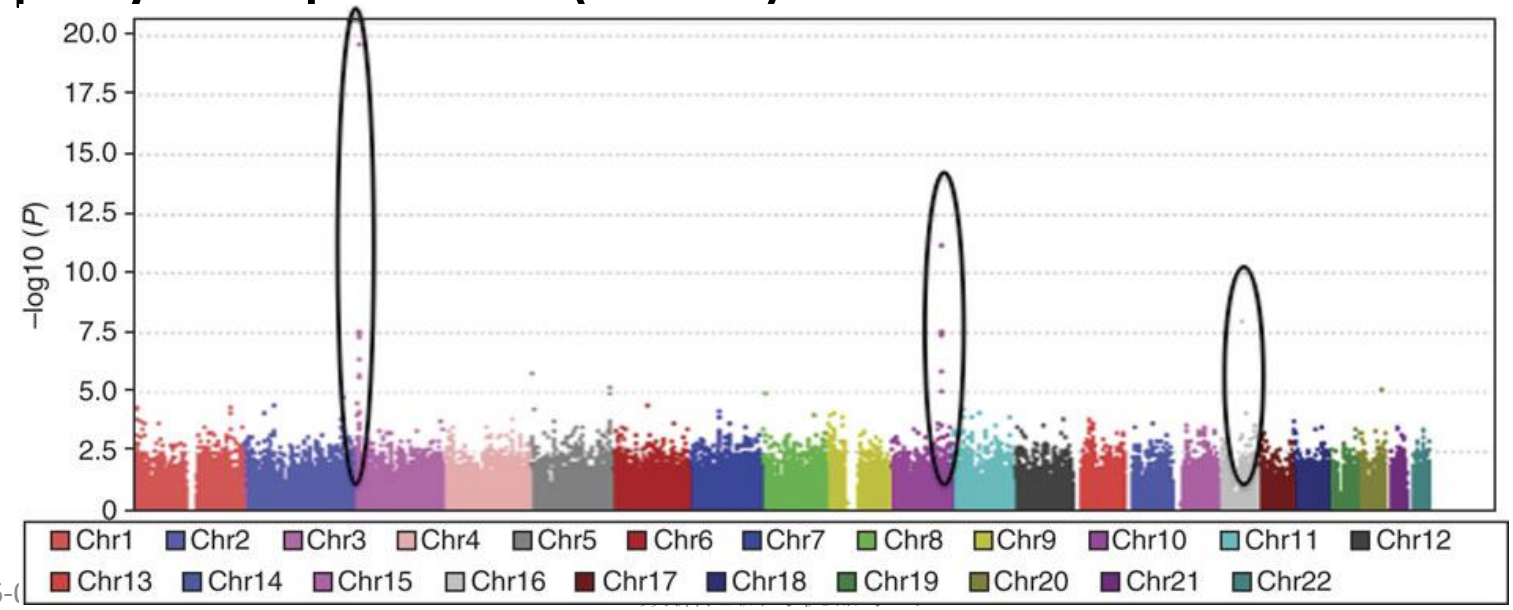
- Motif finding



- Response : univariate response measuring binding intensity (ChIP-seq or ChIP-chip data)
- Covariates (~200): abundant score of candidate motifs

Motivating example 5

- Genome-wide association studies
- Response: disease or not
- Covariates (~10^6): single nucleotide polymorphisms (SNPs)



Motivating example 6

- Expression quantitative trait loci (eQTL) studies
  - Response (~ 20,000): gene expression
  - Covariates(~10^6): SNPs

Motivating example 6

- Exp
  stuc
  - R
  - C

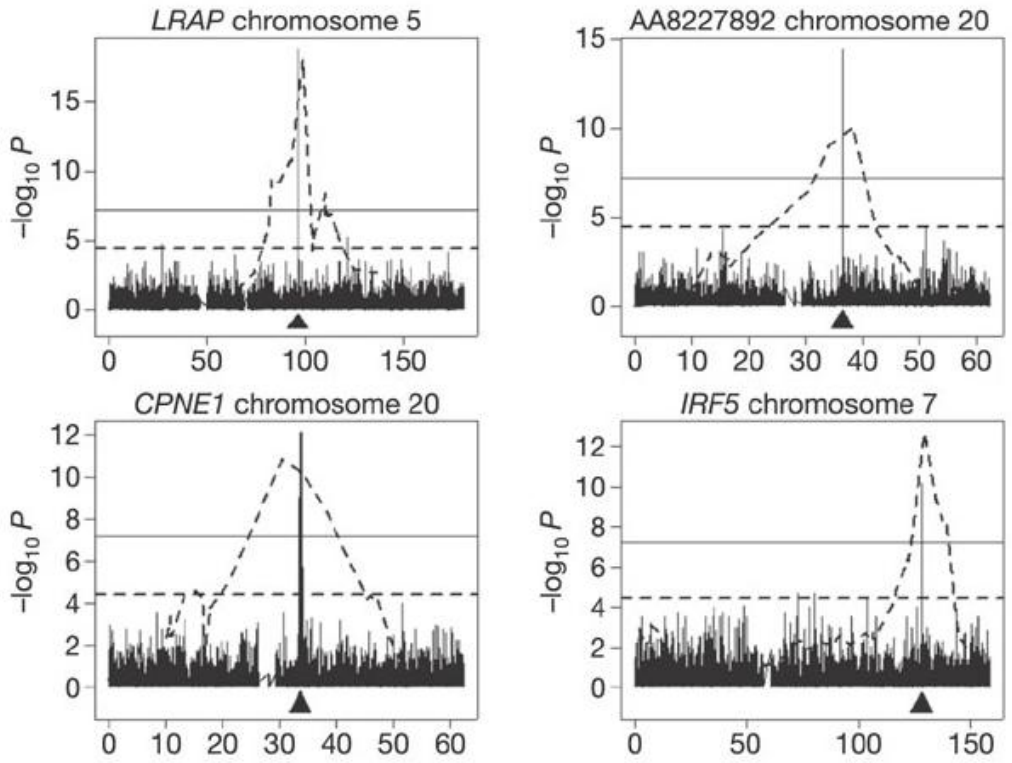


Figure 17: Figure adopted from Cheung et al. Figure 1

General framework

general framework:

$Z_1, \ldots, Z_n$ (with some "i.i.d. components")

$\dim(Z_i) \gg n$

for example:

$Z_i = (X_i, Y_i)$, $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$, $Y_i \in \mathcal{Y} \subseteq\in \mathbb{R}$: regression with $p \gg n$

$Z_i = (X_i, Y_i)$, $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$, $Y_i \in \{0, 1\}$: classification for $p \gg n$

General framework

$$Y_i = \sum_{j=1}^{p} \beta_j^0 X_i^{(j)} + \varepsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

goals:

- prediction, e.g. w.r.t. squared prediction error
- estimation of β^0, e.g. w.r.t. $\|\hat{\beta} - \beta^0\|_q$ $(q = 1, 2)$
- variable selection
  i.e. estimating the active set with the effective variables
  (having corresponding coefficient $\neq 0$)

Stepwise selection

- Backward Elimination

1. Start with all the predictors in the model
2. Remove the predictor with highest p-value greater than α_{crit}
3. Refit the model and goto 2
4. Stop when all p-values are less than α_{crit}.

Stepwise selection

- Forward Selection

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit}.
3. Continue until no new predictors can be added.

Drawbacks of stepwise selection

- One-at-a-time: may miss optimal
- P-values of the remaining predictors tends to be overstated
  - Multiple testing
- Model tends to be smaller than desirable for prediction purpose
- Variable not in the model may still be correlated with the response

Stepwise selection—An example

- Interested in factors related to the life expectancy (50 US states,1969-71)
  - Per capita income (1974)
  - Illiteracy (1970,percent of population)
  - Murder rate per 100,000 population
  - Percent high-school graduates
  - Mean number of days with min temperature < 30 degree
  - Land area in square mile

Bias, Variance, and Model Complexity

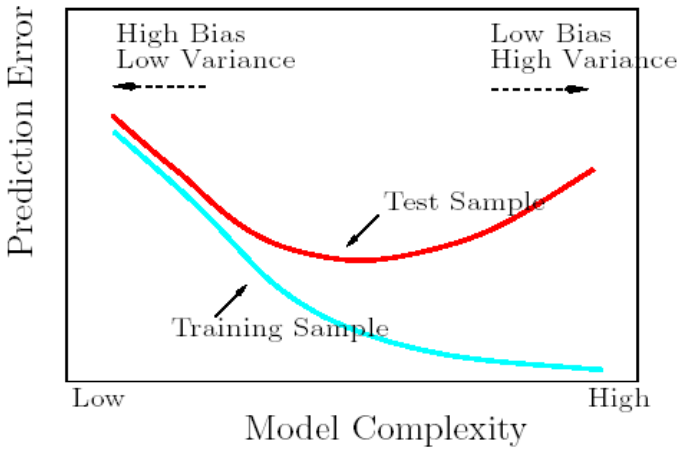


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

- Bias-Variance trade-off again
- Generalization: test sample vs. training sample performance
  - Training data usually monotonically increasing performance with model complexity

Measuring Performance

- target variable Y
- Vector of inputs X
- Prediction model $\hat{f}(X)$

- Typical Choices of Loss function

$$L\left(Y, \hat{f}(X)\right) = \begin{cases} \left(Y - \hat{f}(X)\right)^2 & \textit{squared error} \\ \left|Y - \hat{f}(X)\right| & \textit{absolute error} \end{cases}$$

Generalization Error

- Test error. Generalization error

$$Err = E\left[L\left(Y, \hat{f}\left(X\right)\right)\right]$$

- Note: This expectation averages anything that is random, including the randomness in the training sample that it produced
- Training error
- $$\overline{err} = \frac{1}{N}\sum_{i=1}^{n} L\left(y_i, \hat{f}\left(x_i\right)\right)$$
  - average loss over training sample
  - not a good estimate of test error (next slide)

Training Error

- Training error - Overfitting
  - not a good estimate of test error
  - consistently decreases with model complexity
  - drops to zero with high enough complexity

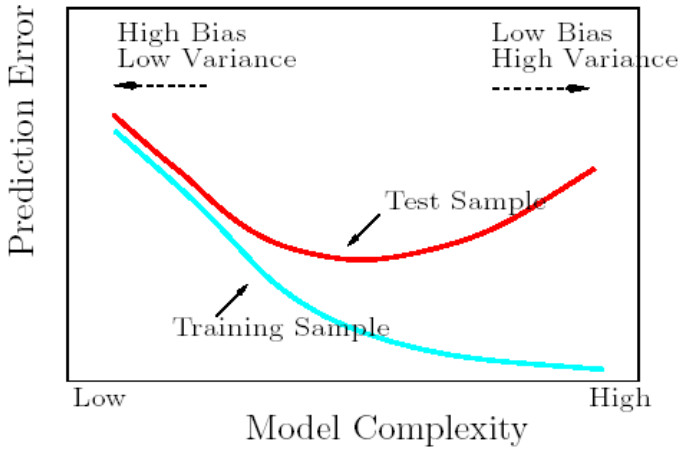


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

Two separate goals

- Model selection:
  - Estimating the performance of different models in order to choose the (approximate) best one
- Model assessment:
  - Having chosen a final model, estimating its prediction error (generalization error) on new data

- Ideal situation: split data into the 3 parts for *training*, *validation (est. prediction error+select model)*, and *testing (assess model)*
- Typical split*: 50% / 25% / 25%*

Bias-Variance Decomposition

$$Y = f(X) + \varepsilon, \quad E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma_\varepsilon^2$$

- Then for an input point $X = x_0$ using unit-square loss and regression fit:

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}.
\end{aligned}$$

Irreducible Error	Bias^2	Variance
variance of the target around the true mean	Amount by which average estimate differs from the true mean	Expected deviation of f^ around its mean

Bias-Variance Decomposition

Linear Model Fit: $\hat{f}_p(x) = \hat{\beta}^T x$

$$Err(x_0) = \sigma_\varepsilon^2 + \left[f(x_o) - E\hat{f}_p(x_0) \right]^2 + \left\| h(x_0) \right\|^2 \sigma_\varepsilon^2$$

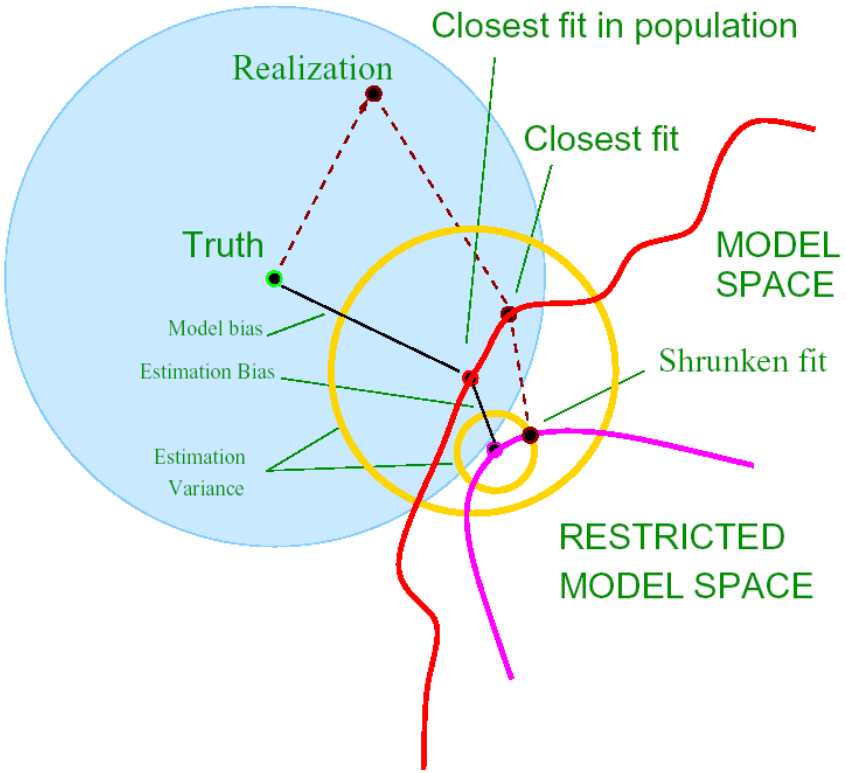
$$\mathbf{h}(x_0) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$$

average over sample values x_i :

$$\frac{1}{N}\sum_{i=1}^{N} Err(x_i) = \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N}\left[f(x_i) - E\hat{f}(x_i) \right]^2 + \frac{p}{N}\sigma_\varepsilon^2$$ ... in-sample error

Model complexity is directly related to the number of parameters p

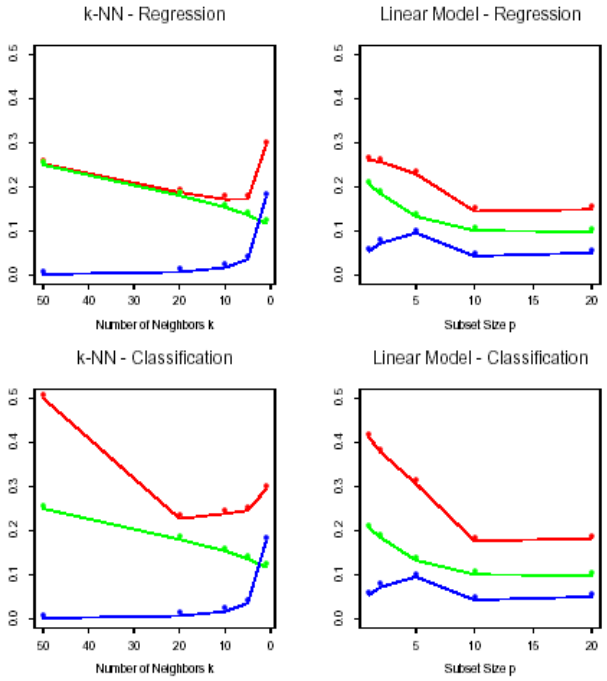
Bias-Variance Decomposition



Taken from Elements of Statistical Learning

Bias-Variance Decomposition - Example

- 50 observations. 20 predictors. Uniform in $[0,1]^{20}$



Left panels:

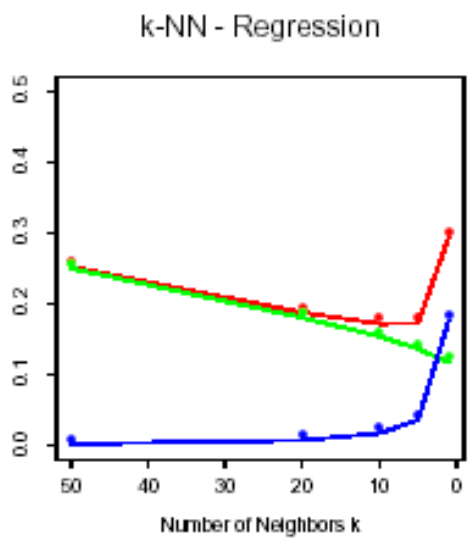
Y is 0 if $X_1 \leq \frac{1}{2}$ and 1 if $X_1 > \frac{1}{2}$, and we apply kNN

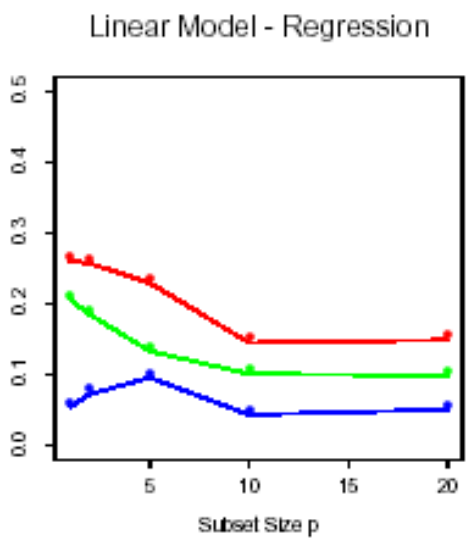
Right panels

Y is 1 if $\sum_{j=1}^{10} X_j > 5$ and 0 otherwise, and we use the best subset linear regression of size p

Example, continued

Regression with squared error loss



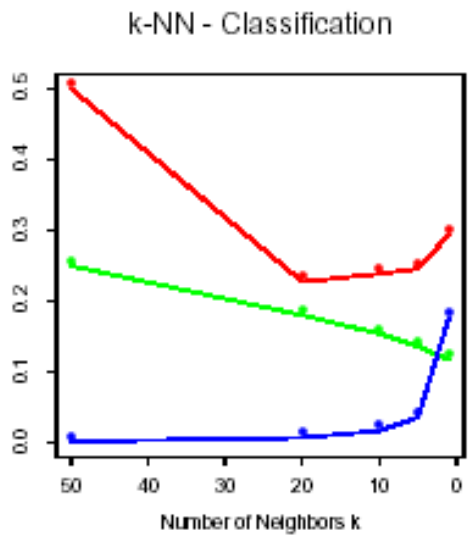


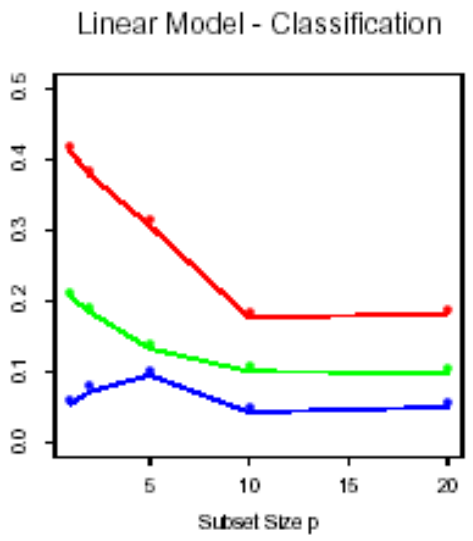
Prediction error

Squared bias

Variance

Classification with 0-1 loss





Optimism of the Training Error Rate

- Typically: training error rate < true error
- (same data is being used to fit the method and assess its error)

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{n} L\left(y_i, \hat{f}\left(x_i\right)\right) \quad < \quad Err = E\left[L\left(Y, \hat{f}\left(X\right)\right)\right]$$

overly optimistic

Optimism of the Training Error Rate

Err … kind of extra-sample error: test features don't need to coincide with training feature vectors

Focus on in-sample error:

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_Y E_{Y^{new}} L\left(Y_i^{new}, \hat{f}\left(x_i\right)\right)$$

Y^{new} … observe N **new** response values at each of training points x_i, i=1, 2, …,N

$$\text{optimism: } op \equiv Err_{in} - E_y\left(\overline{err}\right)$$

for squared error 0-1 and other loss functions:

$$op = \frac{2}{N}\sum_{i=1}^{N} Cov\left(\hat{y}_i, y_i\right)$$

Optimism of the Training Error Rate

Summary:

$$Err_{in} = E_y\left(\overline{err}\right) + \frac{2}{N}\sum_{i=1}^{N} Cov\left(\hat{y}_i, y_i\right)$$

The harder we fit the data, the greater $Cov\left(\hat{y}_i, y_i\right)$ will be, thereby increasing the optimism.

- For linear fit with d indep covariates:

$$Err_{in} = E_y\left(\overline{err}\right) + \frac{2}{N} d\sigma_\varepsilon^2$$

  - optimism ↑ linearly with # d of covariates
  - Optimism ↓ as training sample size ↑

Optimism of the Training Error Rate

- Ways to estimate prediction error:
  - Estimate optimism and then add it to training error rate
    - AIC, BIC, and others work this way, for a special class of estimates that are linear in their parameters
  - Direct estimates of the sample error
    - Cross-validation, bootstrap
    - Can be used with any loss function, and with nonlinear, adaptive fitting techniques

Estimates of In-Sample Prediction Error

- General form of the in-sample estimate:

$$\hat{\mathrm{Err}}_{in} = \overline{err} + \hat{o}p$$

with estimate of optimism

- For linear fit and with $Err_{in} = E_y\left(\overline{err}\right) + \frac{2}{N} d\sigma_\varepsilon^2$:

$$C_p = \overline{err} + \frac{2d}{N}\hat{\sigma}_\varepsilon^2, \text{ so called } C_p \text{ statistic}$$

$\hat{\sigma}_\varepsilon^2$ ... estimate of noise variance, from mean-squared error of low-bias model

d... # of basis functions

N... training sample size

Estimates of In-Sample Prediction Error

- Similarly: Akaike Information Criterion (AIC)
  - More applicable estimate of Err_{in}, when log-likelihood function is used

$$\text{For } N \to \infty: \quad -2E\left[\log \Pr_{\hat{\theta}}(Y)\right] \approx -\frac{2}{N} E\left[\log \text{lik}\right] + 2\frac{d}{N}$$

$\Pr_{\theta}(Y)$... family density for Y (containing the true density)

$\hat{\theta}$... ML estimate of θ

$$\text{loglik} = \sum_{i=1}^{N} \log \Pr_{\hat{\theta}}(y_i)$$

Maximized log-likelihood due to ML estimate of theta

AIC

$$\text{For } N \to \infty: \quad -2E\left[\log \Pr_{\hat{\theta}}(Y)\right] \approx -\frac{2}{N} E\left[\log \text{lik}\right] + 2\frac{d}{N}$$

For example, for logistic regression model, using binomial log-likelihood:

$$AIC = -\frac{2}{N} \cdot \text{loglik} + 2 \cdot \frac{d}{N}$$

To use AIC for model selection: choose the model giving smallest AIC over the set of models considered.

$$AIC(\alpha) = \overline{err}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}_{\varepsilon}^{2}$$

$f_{\hat{\alpha}}(x)$... set of models, α... tuning parameter

$\overline{\text{err}}(\alpha)$... training error, $d(\alpha)$... # parameters

Effective Number of Parameters

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$ Vector of Outcomes, similarly for predicitons

$\hat{y} = Sy$ Linear fit (e.g. linear regression, quadratic shrinkage – ridge, splines)

S... $\mathrm{N} \times \mathrm{N}$ matrix, depends on input vector x_i but not on y_i

effective number of parameters: $\mathrm{d}(\mathrm{S}) = trace(S)$ c.f. $Cov(\hat{y}, y)$

$d(s)$ is the correct d for C_p

$$C_p = \overline{err} + \frac{2d}{N}\hat{\sigma}_\varepsilon^2$$

Bayesian Approach and BIC

- Like AIC used in when fitting by max log-likelihood

Bayesian Information Criterion (BIC):

$$BIC = -2\,\text{loglik} + (\log N)d$$

$$\text{Assuming Gaussian model : } \sigma_\varepsilon^2 \text{ known,}$$

$$-2 \cdot \text{loglik} \approx \sum_i (y_i - \hat{f}(x_i))^2 / \sigma_\varepsilon^2 = N \cdot \overline{err} / \sigma_\varepsilon^2$$

$$\text{then } BIC = \frac{N}{\sigma_\varepsilon^2}[\overline{err} + (\log N) \cdot \frac{d}{N} \sigma_\varepsilon^2]$$

BIC proportional to AIC except for log(N) rather than factor of 2. For $N > e^2$ (approx 7.4), BIC penalizes complex models more heavily.

BIC Motivation

- Given a set of candidate models $\mathbf{M}_m, m = 1 \ldots M$ and model parameters θ_{m}
- Posterior probability of a given model: $\Pr(\mathbf{M}_m \mid \mathbf{Z}) \propto \Pr(\mathbf{M}_m) \cdot \Pr(\mathbf{Z} \mid \mathbf{M}_m)$
- Where $\mathbf{Z}$ represents the training data $\{x_i, y_i\}_1^N$
- To compare two models, form the posterior odds:

$$\frac{\Pr(\mathbf{M}_m \mid \mathbf{Z})}{\Pr(\mathbf{M}_l \mid \mathbf{Z})} = \frac{\Pr(\mathbf{M}_m)}{\Pr(\mathbf{M}_l)} \cdot \frac{\Pr(\mathbf{Z} \mid \mathbf{M}_m)}{\Pr(\mathbf{Z} \mid \mathbf{M}_l)}$$

- If odds > 1, then choose model m. Prior over models (left half) considered constant. Right half, contribution of data (Z) to posterior odds, is called the Bayes factor BF(Z).
- Need to approximate $Pr(Z|M_m)$.
- Can est. posterior from BIC and compare relative merits of models.

General framework

$$Y_i = \sum_{j=1}^{p} \beta_j^0 X_i^{(j)} + \varepsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

goals:

- prediction, e.g. w.r.t. squared prediction error
- estimation of β^0, e.g. w.r.t. $\|\hat{\beta} - \beta^0\|_q$ $(q = 1, 2)$
- variable selection
  i.e. estimating the active set with the effective variables
  (having corresponding coefficient $\neq 0$)

Penalty based methods

if true β^0 is sparse w.r.t.

- $\|\beta^0\|_0^0 =$ number of non-zero coefficients
  $\rightsquigarrow$ regularize with the $\|\cdot\|_0$-penalty:
  $\text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_0^0)$, e.g. AIC, BIC
  $\rightsquigarrow$ computationally infeasible if p is large (2^p sub-models)

- $\|\beta^0\|_1 = \sum_{j=1}^p |\beta_j^0|$
  $\rightsquigarrow$ penalize with the $\|\cdot\|_1$-norm, i.e. Lasso:
  $\text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_1)$
  $\rightsquigarrow$ convex optimization:
  computationally feasible and very fast for large p

The Lasso

Lasso for linear models

$$\hat{\beta}(\lambda) = \text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \underbrace{\lambda}_{\geq 0} \underbrace{\|\beta\|_1}_{\sum_{j=1}^p |\beta_j|})$$

$\rightsquigarrow$ convex optimization problem

- Lasso does variable selection
  some of the $\hat{\beta}_j(\lambda) = 0$
  (because of "ℓ_1-geometry")
- $\hat{\beta}(\lambda)$ is a shrunken LS-estimate

The Lasso

equivalence to primal problem

$$\hat{\beta}_{\text{primal}}(R) = \text{argmin}_{\beta;\|\beta\|_1 \le R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n,$$

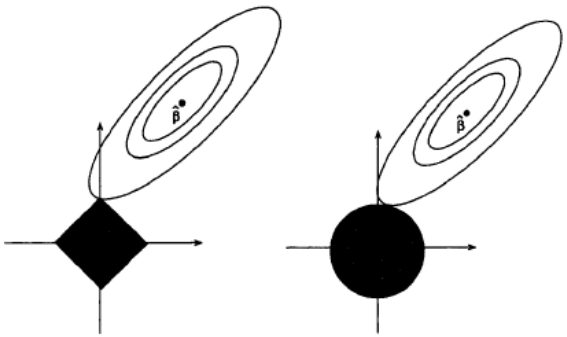
with a correspondence between λ and R which depends on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$

since

- $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ is convex in β
- convex constraint $\|\beta\|_1 \le R$

The Lasso and the Ridge Regression

p=2



left: ℓ_1-"world"
residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the ℓ_1-ball in its corner
$\rightsquigarrow \hat{\beta}_1 = 0$

The Lasso and the Ridge Regression

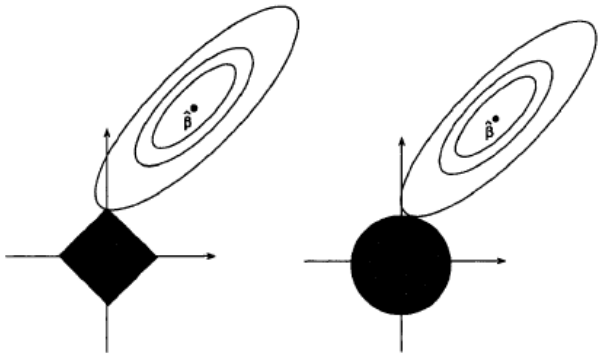
Ridge regression,

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \text{argmin}_{\beta}\left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_2^2 \right),$$

equivalent primal equivalent solution

$$\hat{\beta}_{\text{Ridge;primal}}(R) = \text{argmin}_{\beta;\|\beta\|_2 \le R}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n,$$

with a one-to-one correspondence between λ and R



Relationship with Bayesian methods

model:

$$\beta_1, \ldots, \beta_p \text{ i.i.d. } \sim p(\beta)d\beta,$$
$$\text{given } \beta: \ \mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I_n) \text{ with density } f(\mathbf{y}|\sigma^2, \beta)$$

posterior density:

$$p(\beta|\mathbf{Y}, \sigma^2) = \frac{f(\mathbf{Y}|\beta, \sigma^2)p(\beta)}{\int f(\mathbf{Y}|\beta, \sigma^2)p(\beta)d\beta} \propto f(\mathbf{Y}|\beta, \sigma^2)p(\beta)$$

and hence for the MAP (Maximum A-Posteriori) estimator:

$$\begin{aligned}\hat{\beta}_{\text{MAP}} &= \text{argmax}_\beta p(\beta|\mathbf{Y}, \sigma^2) = \text{argmin}_\beta - \log\left(f(\mathbf{Y}|\beta, \sigma^2)p(\beta)\right) \\ &= \text{argmin}_\beta \left(\frac{1}{2\sigma^2}\|\mathbf{Y} - X\beta\|_2^2 - \sum_{j=1}^{p} \log(p(\beta_j))\right)\end{aligned}$$

Relationship with Bayesian methods

examples:

1. Double-Exponential prior DExp(ξ):

$p(\beta) = \frac{\tau}{2} \exp(-\tau\beta)$

$\rightsquigarrow \hat{\beta}_{\text{MAP}}$ equals the Lasso with penalty parameter $\lambda = n^{-1} 2\sigma^2\tau$

2. Gaussian prior $\mathcal{N}(0, \tau^2)$:

$p(\beta) = \frac{1}{\sqrt{2\pi\tau}} \exp(-\beta^2/(2\tau^2))$

$\rightsquigarrow \hat{\beta}_{\text{MAP}}$ equals the Ridge estimator with penalty parameter $\lambda = n^{-1}\sigma^2/\tau^2$

but we will argue that Lasso (i.e., the MAP estimator) is also good if the truth is sparse with respect to $\|\beta^0\|_0^0$, e.g. if prior is (much) more spiky around zero than Double-Exponential distribution

Lasso for orthogonal design

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad n^{-1}\mathbf{X}^T\mathbf{X} = I$$

Lasso = soft-thresholding estimator

$$\hat{\beta}_j(\lambda) = \mathrm{sign}(Z_j)(|Z_j| - \lambda/2)_+, \quad \underbrace{Z_j}_{=\mathrm{OLS}} = (n^{-1}\mathbf{X}^T\mathbf{Y})_j,$$

$$\hat{\beta}_j(\lambda) = g_{\mathrm{soft}}(Z_j),$$

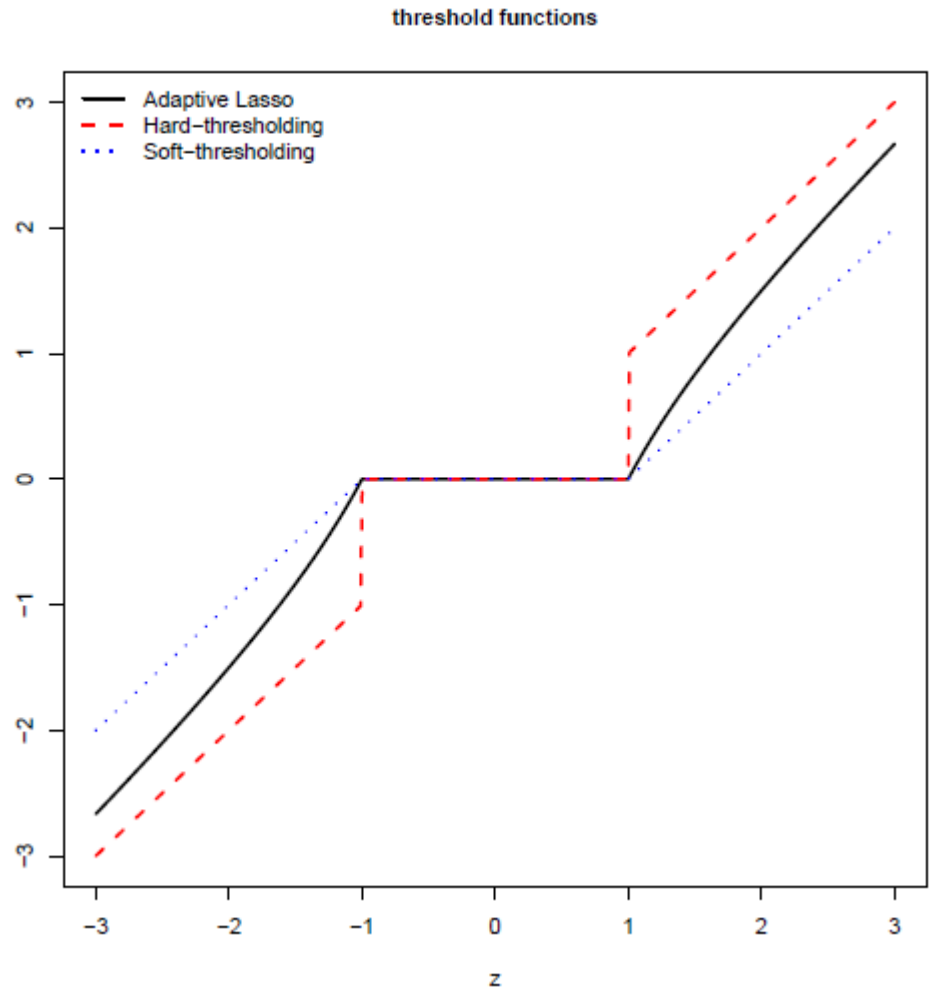
Lasso for orthogonal design

Lass

$\hat{\beta}_j(\lambda$

$\hat{\beta}_j(\lambda$

$\mathbf{X}^T\mathbf{Y})_j,$



Estimation of regression coefficients

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon,\ p \gg n$$

with fixed (deterministic) design $\mathbf{X}$

problem of identifiability:
for $p > n$: $\mathbf{X}\beta^0 = X\theta$
for any $\theta = \beta^0 + \xi$, ξ in the null-space of $\mathbf{X}$

$\rightsquigarrow$ cannot say anything about $\|\hat{\beta} - \beta^0\|$ without further assumptions!
$\rightsquigarrow$ we will work with the compatibility assumption (see later) and we will explain: under compatibility condition

$$\|\hat{\beta} - \beta^0\|_1 \leq C\frac{s_0}{\phi_0^2}\sqrt{\log(p)/n},$$

$$s_0 = |\mathrm{supp}(\beta^0)| = |\{j;\ \beta_j^0 \neq 0\}|$$

Asymptotic Results-preview

for (fixed design) linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$ with
active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$
two key assumptions

1. neighborhood stability condition for design $\mathbf{X}$
   $\Leftrightarrow$ irrepresentable condition for design $\mathbf{X}$
2. beta-min condition

$$\min_{j \in S_0} |\beta_j^0| \quad \geq \quad C\sqrt{s_0 \log(p)/n}, \quad C \text{ suitably large}$$

both conditions are sufficient and "essentially" necessary for

$$\hat{S}(\lambda) = S_0 \text{ with high probability}, \quad \lambda \gg \quad \sqrt{\log(p)/n}$$

Asymptotic Results

neighborhood stability condition $\Leftrightarrow$ irrepresentable condition

$$n^{-1}\mathbf{X}^T\mathbf{X} = \hat{\Sigma}$$

active set $S_0 = \{j;\ \beta_j \neq 0\} = \{1, \ldots, s_0\}$ consists of the first s_0 variables; partition

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{S_0,S_0} & \hat{\Sigma}_{S_0,S_0^c} \\ \hat{\Sigma}_{S_0^c,S_0} & \hat{\Sigma}_{S_0^c,S_0^c} \end{pmatrix}$$

irrep. condition : $\|\hat{\Sigma}_{S_0^c,S_0}\hat{\Sigma}_{S_0,S_0}^{-1}\mathrm{sign}(\beta_1^0, \ldots, \beta_{s_0}^0)^T\|_\infty < 1$

Parameter Tuning

choice of λ: $\hat{\lambda}_{CV}$ from cross-validation
empirical and theoretical indications (Meinshausen & PB, 2006) that

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

moreover

$$|\hat{S}(\hat{\lambda}_{CV})| \leq \min(n, p) (= n \text{ if } p \gg n)$$

Parameter Tuning

recall:

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

and we would then use a second-stage to reduce the number of false positive selections

$\rightsquigarrow$ re-estimation on much smaller model with variables from $\hat{S}$

- OLS on $\hat{S}$ with e.g. BIC variable selection
- thresholding coefficients and OLS re-estimation
- adaptive Lasso (Zou, 2006)
- ...

Adaptive Lasso

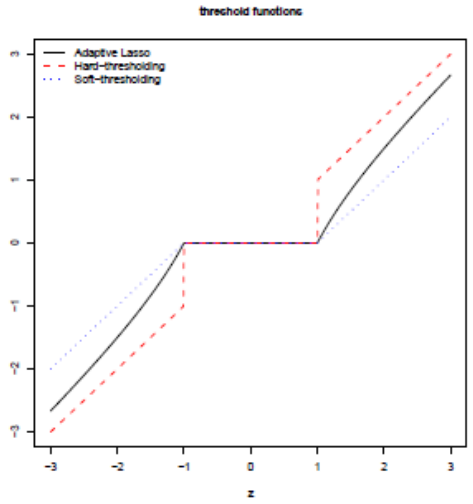
re-weighting the penalty function

$$\hat{\beta} = \text{argmin}_\beta(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|}),$$

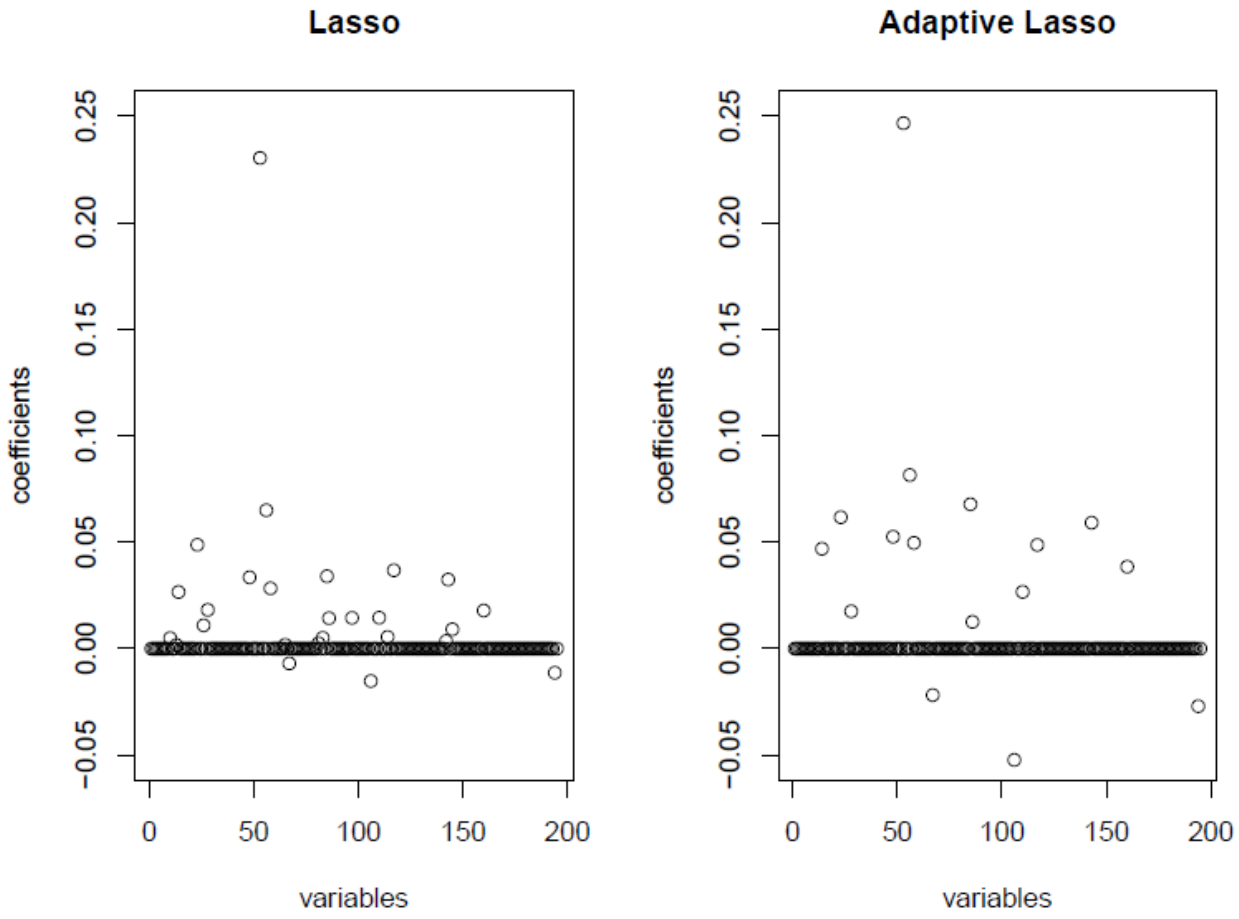
$\hat{\beta}_{init,j}$ from Lasso in first stage $\underbrace{\text{(or OLS if } p < n\text{)}}$

Zou (2006)

for orthogonal design,
if $\hat{\beta}_{init}$ = OLS:
Adaptive Lasso = NN-garrote
$\rightsquigarrow$ less bias than Lasso



Adaptive Lasso



KKT conditions and Computation

characterization of solution(s) $\hat{\beta}$ as minimizer of the criterion function

$$Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1$$

since $Q_\lambda(\cdot)$ is a convex function:
necessary and sufficient that subdifferential of $\partial Q_\lambda(\beta)/\partial\beta$ at $\hat{\beta}$ contains the zero element

Lemma
denote by $G(\beta) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)/n$ the gradient vector of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$
Then: $\hat{\beta}$ is a solution if and only if

$$G_j(\hat{\beta}) = -\text{sign}(\hat{\beta}_j)\lambda \text{ if } \hat{\beta}_j \neq 0,$$
$$|G_j(\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0$$

Coordinate descent algorithm for computation

general idea is to compute a solution $\hat{\beta}(\lambda_{\text{grid},k})$ and use it as a starting value for the computation of $\hat{\beta}(\underbrace{\lambda_{\text{grid},k-1}}_{<\lambda_{\text{grid},k}})$

$\beta^{(0)} \in \mathbb{R}^p$ an initial parameter vector. Set $m = 0$.

REPEAT:
Increase m by one: $m \leftarrow m + 1$.
For $j = 1, \ldots, p$:

$$\text{if } |G_j(\beta_{-j}^{(m-1)})| \leq \lambda : \text{ set } \beta_j^{(m)} = 0,$$
$$\text{otherwise: } \beta_j^{(m)} = \text{argmin}_{\beta_j} Q_\lambda(\beta_{+j}^{(m-1)}),$$

β_{-j}: parameter vector setting jth component to zero
$\beta_{+j}^{(m-1)}$: parameter vector which equals $\beta^{(m-1)}$ except for jth component equalling β_j
UNTIL numerical convergence

Coordinate descent algorithm for computation

For linear regression

$$G_j(\beta) = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta)/n$$

$$\beta_j^{(m)} = \frac{\text{sign}(Z_j)(|Z_j| - \lambda/2)_+}{\hat{\Sigma}_{jj}},$$

$$Z_j = \mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta_{-j})/n, \quad \hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}.$$

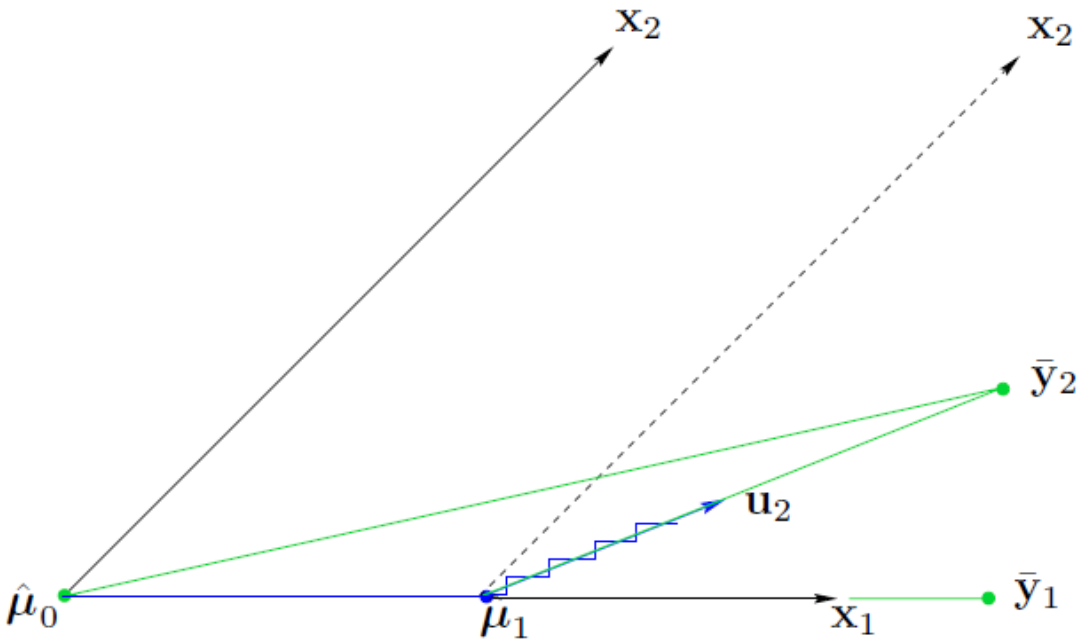
$\rightsquigarrow$ componentwise soft-thresholding

glmnet: R-package

Least Angle Regression-- LAR

1. Start with $r = y$, $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p = 0$. Assume x_j standardized.
2. Find predictor x_j most correlated with r.
3. Increase β_j in the direction of sign$\langle r, x_j \rangle$ until some other competitor x_k has as much correlation with current residual as does x_j.
4. Move $(\hat{\beta}_j, \hat{\beta}_k)$ in the joint least squares direction for (x_j, x_k) until some other competitor x_ℓ has as much correlation with the current residual
5. Continue in this way until all predictors have been entered. Stop when $\langle r, x_j \rangle = 0 \; \forall \; j$, i.e. OLS solution.

Least Angle Regression-- LAR



The LAR direction $\mathbf{u}_2$ at step 2 makes an equal angle with $\mathbf{x}_1$ and $\mathbf{x}_2$.

Least Angle Regression-- LAR

For each iteration, we have:

- Active set $\mathcal{A}_k$ at the beginning of the kth step
- Coefficient vector $\beta_{\mathcal{A}_k}$ at this step
- $k-1$ nonzero values, the one just entered the model has coefficient 0.

Then we do:

- Compute current residual $\mathbf{r}_k = \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$;
- Compute direction $\delta_k = (\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k$;
- Evolve the coefficient $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$ until some $\mathbf{x}_l$ has as much correlation with the current residual.
- Add $\mathbf{x}_l$ to the active set $\mathcal{A}_{k+1}$

Least Angle Regression-- LAR

Algorithm 3.2a *Least Angle Regression: Lasso Modification.*

4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

Least Angle Regression-- LAR

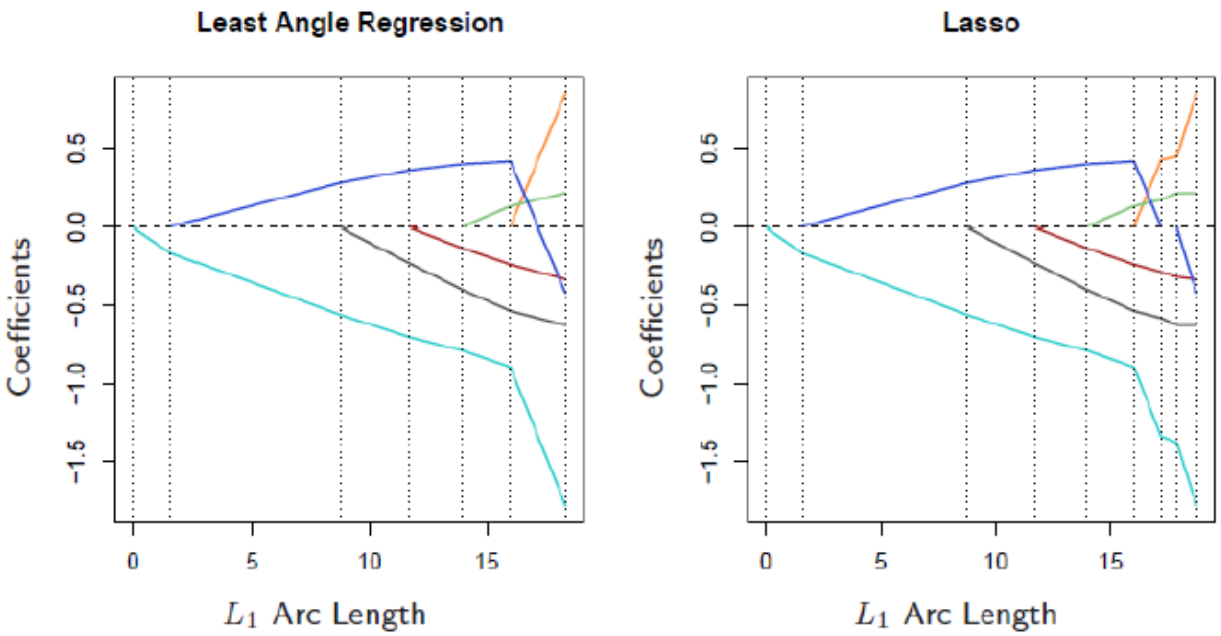


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about* 18.

Generalization of Lasso and Ridge regression

Consider:

$$\tilde{\beta} = \underset{\beta}{\text{Argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

- $|\beta_j|^q$ prior distribution for β_j,
- $q = 1$ smallest q that constraint is still convex,
- Typically $q = 1, 2$ (Lasso and Ridge regression).
- can be determined by data, but not worth the effort.

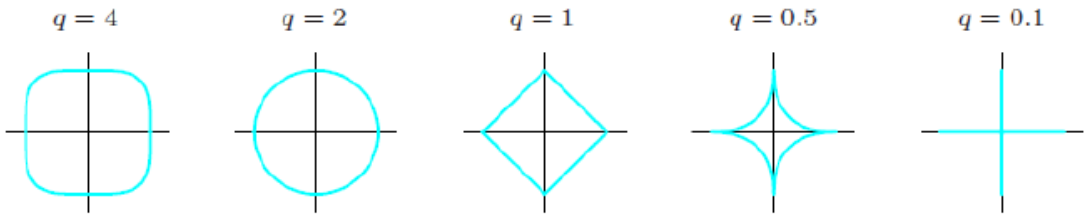


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

The Dantzig Selector

Candes and Tao (2007) proposed the Dantzig Selector (DS):

$$\min_{\beta} \|\beta\|_1 \text{ subject to } \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)\|_\infty \leq s$$

It can also be written as:

$$\min_{\beta} \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)\|_\infty \text{ subject to } \|\beta\|_1 \leq t$$

The grouped Lasso

Suppose the p predictors are divided into L groups, with p_ℓ the number in group ℓ. The grouped-Lasso solves the minimization problem:

$$\min_{\beta \in \mathbb{R}^p} \left(\|\mathbf{y} - \beta_0 \mathbf{1} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \beta_\ell\|_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \|\beta_\ell\|_2 \right)$$

where $\mathbf{X}_\ell$ is the predictors corresponding to the ℓth group, with corresponding coefficient vector β_ℓ.

Sure Independent Screening

In Linear regression:

The true sparse model

$$\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$$

$$\omega = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\}$$

$$\gamma \in (0, 1)$$

Need to show

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1 \qquad \text{as } n \to \infty$$

Sure Independent Screening

Rationale of correlation learning:

When p > n, the OLS estimator

$$\hat{\beta}_{\mathrm{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}\mathbf{X}^{\mathrm{T}}\mathbf{y} \text{ is noisy}$$

$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}$ is the Moore-Penrose generalized inverse

The ridge regression

$$\omega^{\lambda} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\omega^{\lambda} \to \hat{\beta}_{\mathrm{LS}} \qquad \text{as } \lambda \to 0,$$

$$\lambda\omega^{\lambda} \to \omega \qquad \text{as } \lambda \to \infty.$$

Sure Independent Screening

Iteratively thresholded ridge regression screener (ITRRS)

$$\mathcal{M}^1_{\delta,\lambda} = \{1 \leqslant i \leqslant p : |\omega_i^\lambda| \text{ is among the first } [\delta p] \text{ largest of all}\}. \tag{8}$$

(a) First, carry out the procedure in submodel (8) to the full model $\{1, \ldots, p\}$ and obtain a submodel $\mathcal{M}^1_{\delta,\lambda}$ with size $[\delta p]$.
(b) Then, apply a similar procedure to the model $\mathcal{M}^1_{\delta,\lambda}$ and again obtain a submodel $\mathcal{M}^2_{\delta,\lambda} \subset \mathcal{M}^1_{\delta,\lambda}$ with size $[\delta^2 p]$, and so on.
(c) Finally, obtain a submodel $\mathcal{M}_{\delta,\lambda} = \mathcal{M}^k_{\delta,\lambda}$ with size $d = [\delta^k p] < n$, where $[\delta^{k-1} p] \geqslant n$.

Sure Independent Screening

Theorem 1 (accuracy of SIS). Under conditions 1–4, if $2\kappa + \tau < 1$ then there is some $\theta < 1 - 2\kappa - \tau$ such that, when $\gamma \sim cn^{-\theta}$ with $c > 0$, we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

Theorem 2 (asymptotic sure screening). Under conditions 1–4, if $2\kappa + \tau < 1$, $\lambda(p^{3/2}n)^{-1} \to \infty$, and $\delta n^{1-2\kappa-\tau} \to \infty$ as $n \to \infty$, then we have, for some $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}^1_{\delta,\lambda}) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}].$$

Sure Independent Screening

Condition 1. $p > n$ and $\log(p) = O(n^{\xi})$ for some $\xi \in (0, 1-2\kappa)$, where κ is given by condition 3.

Condition 2. $\mathbf{z}$ has a spherically symmetric distribution and property C. Also, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

Condition 3. $\operatorname{var}(Y) = O(1)$ and, for some $\kappa \geqslant 0$ and $c_2, c_3 > 0$,

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geqslant \frac{c_2}{n^{\kappa}} \quad \text{and} \quad \min_{i \in \mathcal{M}_*} |\operatorname{cov}(\beta_i^{-1} Y, X_i)| \geqslant c_3.$$

As seen later, κ controls the rate of probability error in recovering the true sparse model. Although $b = \min_{i \in \mathcal{M}_*} |\operatorname{cov}(\beta_i^{-1} Y, X_i)|$ is assumed here to be bounded away from 0, our asymptotic study applies as well to the case where $b \to 0$ as $n \to \infty$. In particular, when the variables in $\mathcal{M}_*$ are uncorrelated, $b = 1$. This condition rules out the situation in which an important variable is marginally uncorrelated with Y, but jointly correlated with Y.

Condition 4. There are some $\tau \geqslant 0$ and $c_4 > 0$ such that

$$\lambda_{\max}(\boldsymbol{\Sigma}) \leqslant c_4 n^{\tau}.$$

$\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{x})$

Sure Independent Screening

S=8, 18

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	62.5	15	37	27	34
	(1.381)	(0.895)	(0.374)	(0.795)	(0.614)	(1.269)
20000	—	—	37	119	60.5	99
	—	—	(0.288)	(0.732)	(0.372)	(1.014)

Feature Screening by Distance Correlation

- Szekely, Rizzo and Bakirov (2007) proposed the distance correlation

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = \int_{R^{d_u+d_v}} \|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - \phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})\|^2 w(\mathbf{t}, \mathbf{s})\, d\mathbf{t}\, d\mathbf{s},$$

$\phi_{\mathbf{u}}(\mathbf{t})$ $\phi_{\mathbf{v}}(\mathbf{s})$ $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s})$ characteristic functions of two random vectors

$$w(\mathbf{t}, \mathbf{s}) = \left\{ c_{d_u} c_{d_v} \|\mathbf{t}\|_{d_u}^{1+d_u} \|\mathbf{s}\|_{d_v}^{1+d_v} \right\}^{-1}$$

$$\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u}, \mathbf{u})\mathrm{dcov}(\mathbf{v}, \mathbf{v})}}$$

Feature Screening by Distance Correlation

For two normal random variables

$$\operatorname{dcorr}(U, V) = \left\{ \frac{\rho \arcsin(\rho) + \sqrt{1-\rho^2} - \rho \arcsin(\rho/2) - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}} \right\}^{1/2}$$

Distance correlation is strictly increasing in $|\rho|$

Feature Screening by Distance Correlation

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = S_1 + S_2 - 2S_3.$$

$$S_1 = E\{\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{d_u} \|\mathbf{v} - \widetilde{\mathbf{v}}\|_{d_v}\},$$
$$S_2 = E\{\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{d_u}\} E\{\|\mathbf{v} - \widetilde{\mathbf{v}}\|_{d_v}\}, \quad \text{and}$$
$$S_3 = E\{E(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{d_u} \mid \mathbf{u}) E(\|\mathbf{v} - \widetilde{\mathbf{v}}\|_{d_v} \mid \mathbf{v})\},$$

$(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$ is an independent copy of $(\mathbf{u}, \mathbf{v})$.

Feature Screening by Distance Correlation

$$\widehat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i - \mathbf{u}_j\|_{d_u}\|\mathbf{v}_i - \mathbf{v}_j\|_{d_v},$$

$$\widehat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i - \mathbf{u}_j\|_{d_u}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{v}_i - \mathbf{v}_j\|_{d_v},$$

$$\widehat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{i=1}^{n}\sum_{l=1}^{n}\|\mathbf{u}_i - \mathbf{u}_l\|_{d_u}\|\mathbf{v}_j - \mathbf{v}_l\|_{d_v}.$$

$$\widehat{\mathrm{dcov}}^2(\mathbf{u},\mathbf{v}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3.$$

$$\widehat{\mathrm{dcorr}}(\mathbf{u},\mathbf{v}) = \frac{\widehat{\mathrm{dcov}}(\mathbf{u},\mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u},\mathbf{u})\mathrm{dcov}(\mathbf{v},\mathbf{v})}}.$$

Feature Screening by Distance Correlation

$$\mathcal{D} = \{k : F(\mathbf{y} \mid \mathbf{x}) \text{ functionally depends on } X_k \text{ for some } \mathbf{y} \in \Psi_y\},$$

$$\mathcal{I} = \{k : F(\mathbf{y} \mid \mathbf{x}) \text{ does not functionally depend on } X_k \text{ for any } \mathbf{y} \in \Psi_y\}. \quad (2.5)$$

$$\omega_k = \text{dcorr}^2(X_k, \mathbf{y}), \quad \text{and} \quad \widehat{\omega}_k = \widehat{\text{dcorr}}^2(X_k, \mathbf{y})$$

$$\widehat{\mathcal{D}}^\star = \{k : \widehat{\omega}_k \geq cn^{-\kappa}, \text{ for } 1 \leq k \leq p\}$$

Li et al. (2012) JASA

Feature Screening by distance correlation

Theorem 1. Under Condition (C1), for any $0 < \gamma < 1/2 - \kappa$, there exist positive constants $c_1 > 0$ and $c_2 > 0$ such that

$$\Pr\left(\max_{1\le k\le p} |\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\right) \le O(p[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]). \quad (2.6)$$

Under Conditions (C1) and (C2), we have that

$$\Pr\left(\mathcal{D} \subseteq \widehat{\mathcal{D}}^{\star}\right) \ge 1 - O(s_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]),$$

Feature Screening by distance correlation

(C1) Both $\mathbf{x}$ and $\mathbf{y}$ satisfy the subexponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that for all $0 < s \leq 2s_0$,

$$\sup_p \max_{1 \leq k \leq p} E\left\{\exp\left(s\|X_k\|_1^2\right)\right\} < \infty, \quad \text{and}$$

$$E\left\{\exp\left(s\|\mathbf{y}\|_q^2\right)\right\} < \infty.$$

(C2) The minimum DC of active predictors satisfies

$$\min_{k \in \mathcal{D}} \omega_k \geq 2cn^{-\kappa},$$

for some constants $c > 0$ and $0 \leq \kappa < 1/2$.

Feature Screening by distance correlation

$$
\begin{aligned}
\text{(1.a):}\quad & Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) \\
& \quad + c_4\beta_4 X_{22} + \varepsilon, \\
\text{(1.b):}\quad & Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) + c_4\beta_3 X_{22} + \varepsilon, \\
\text{(1.c):}\quad & Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) X_{22} + \varepsilon, \\
\text{(1.d):}\quad & Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) \\
& \quad + \exp(c_4|X_{22}|)\varepsilon,
\end{aligned}
$$

Feature Screening by distance correlation

Table 1. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 1

$\mathcal{S}$	SIS					SIRS					DC-SIS				
Model	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
						Case 1: $p = 2000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$									
(1.a)	4.0	4.0	5.0	7.0	21.2	4.0	4.0	5.0	7.0	45.1	4.0	4.0	4.0	6.0	18.0
(1.b)	68.0	578.5	1180.5	1634.5	1938.0	232.9	871.5	1386.0	1725.2	1942.4	5.0	9.0	24.5	73.0	345.1
(1.c)	395.9	1037.2	1438.0	1745.0	1945.1	238.5	805.0	1320.0	1697.0	1946.0	6.0	10.0	22.0	59.0	324.1
(1.d)	130.5	611.2	1166.0	1637.0	1936.5	42.0	304.2	797.0	1432.2	1846.1	4.0	5.0	9.0	41.0	336.2
						Case 2: $p = 2000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$									
(1.a)	5.0	9.0	16.0	97.0	729.4	5.0	9.0	18.0	112.8	957.1	4.0	7.0	11.0	31.2	507.2
(1.b)	26.0	283.2	852.0	1541.2	1919.0	103.9	603.0	1174.0	1699.2	1968.0	5.0	8.0	11.0	17.0	98.0
(1.c)	224.5	775.2	1249.5	1670.0	1951.1	118.6	573.2	1201.5	1685.2	1955.0	7.0	10.0	15.0	38.0	198.3
(1.d)	79.0	583.8	1107.5	1626.2	1930.0	50.9	300.5	728.0	1368.2	1900.1	4.0	7.0	17.0	73.2	653.1
						Case 3: $p = 5000$ and $\sigma_{ij} = 0.5^{\|i-j\|}$									
(1.a)	4.0	4.0	5.0	6.0	59.0	4.0	4.0	5.0	7.0	88.4	4.0	4.0	4.0	6.0	34.1
(1.b)	165.1	1112.5	2729.0	3997.2	4851.5	560.8	1913.0	3249.0	4329.0	4869.1	5.0	11.8	45.0	168.8	956.7
(1.c)	1183.7	2712.0	3604.5	4380.2	4885.0	440.4	1949.0	3205.5	4242.8	4883.1	7.0	17.0	53.0	179.5	732.0
(1.d)	259.9	1338.5	2808.5	3990.8	4764.9	118.7	823.2	1833.5	3314.5	4706.1	4.0	5.0	15.0	77.2	848.2
						Case 4: $p = 5000$ and $\sigma_{ij} = 0.8^{\|i-j\|}$									
(1.a)	5.0	10.0	26.5	251.5	2522.7	5.0	10.0	28.0	324.8	3246.4	5.0	8.0	14.0	69.0	1455.1
(1.b)	40.7	639.8	2072.0	3803.8	4801.7	215.7	1677.8	3010.0	4352.2	4934.1	5.0	8.0	11.0	21.0	162.0
(1.c)	479.2	1884.8	3347.5	4298.5	4875.2	297.7	1359.2	2738.5	4072.5	4877.6	8.0	12.0	22.0	83.0	657.9
(1.d)	307.0	1544.0	2832.5	4026.2	4785.2	148.2	672.0	1874.0	3330.0	4665.2	4.0	7.0	21.0	165.2	1330.0

Bayesian Methods

See Blackboard