# 统计学习

贾金柱

September 12, 2017

# 讲课老师

- 主讲人： 贾金柱
  - Email: jzjia@math.pku.edu.cn
- 助教： 肖一君
  - Email: xiaoyijun1994@126.com

# 先修课程

- 数学分析
- 线性代数
- 数理统计
- 概率论
- R
- Python

# 主要参考文献


- Hastie, Trevor, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. The elements of statistical learning :. Springer, 2001:192-192. http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Ian Goodfellow, Yoshua Bengio and Aaron Courville. Deep Learning. http://www.deeplearningbook.org

# 作业与考核

- 作业一般两周交一次（周五交）
- 考核为作业＋竞赛＋报告
- 其中作业占 50%，竞赛和报告占 50%
  https://www.kaggle.com/competitions
- 作业若迟交，助教可不批改，本次作业分为0
- 课程不易，勿无故缺课

# 目录

# 第一章 Overview of Statistical Learning (1/2)

- Supervised learning
  - Regression:
    - Linear Regression
    - Non-Linear Regression
      - spline, TREES like CART, Random Forest, Deep learning
  - Classification:
    - LDA, QDA
    - Logist Regression, SVM
    - Naive Bayes

# Overview of Statistical Learning (2/2)

- Unsupervised learning
  - Dimension Reduction (e.g. PCA)
  - Clustering
  - Graphical Model
  - Causal Network Learning

# 第二章 Linear Models

- 课程目标
  - 线性回归
  - 变量选择
  - 正则化方法
  - R/Python 调用回归分析包
  - R/Python 做数据分析

# 线性回归

- 简单，容易理解
- 避免 overfitting
- 是其它非线性方法的基础：很多方法都可以转到线性回归

# Best prediction (1/2)

考虑这样一个问题：

假设我们要用predictors $X = (X_1, X_2, \ldots, X_p)$ 去预测 $Y \in \mathbb{R}$.
请问最好的预测函数 $f(X)$ 是什么？

- 这是一个统计决策问题
- 定义Loss: $L(f(X), Y) := [f(X) - Y]^2$
- Goal: $\min_f E[L(f(X), Y)]$

# Best prediction (2/2)

$$E[L(f(X), Y)] = E[f^2(X) - 2f(X)Y + Y^2]$$

$$= E[(f^2(X) - 2f(X)Y + Y^2|X)]$$

$$= E[f^2(X) - 2f(X)E(Y|X)] + E(Y^2)$$

$$= E([f(X) - E(Y|X)]^2) + E(Y^2) - E(E^2(Y|X))$$

- 当 $f(X) = E(Y|X)$ 时， 平方损失达到最小。
- $E(Y|X)$ 称为最佳预测！

# 如何估计最佳预测?

- 非参数方法：

$$ave(Y|X \in (x-\delta, x+\delta)) \to E(Y|X)$$

- 参数建模方法：

$$Y = f(X) + \epsilon,$$

$$\epsilon\text{与}X\text{独立}$$

$$E(\epsilon) = 0$$

  - 如果假定 $f(X) = X\beta$, 这就是线性回归。

# 线性回归中的参数估计

- 极大似然估计
  - 优点：效率较高
  - 缺点：假设较多
- 最小二乘估计
  - 优点：对模型的假设较少
  - 优点：计算速度快

# 极大似然估计

假设误差项 i.i.d. 服从正态分布 $\epsilon \sim N(0, \sigma^2 I_n)$

似然函数是：

$$\prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - X_i^T \beta)^2}{2\sigma^2}} \right\}$$

对数似然函数是：

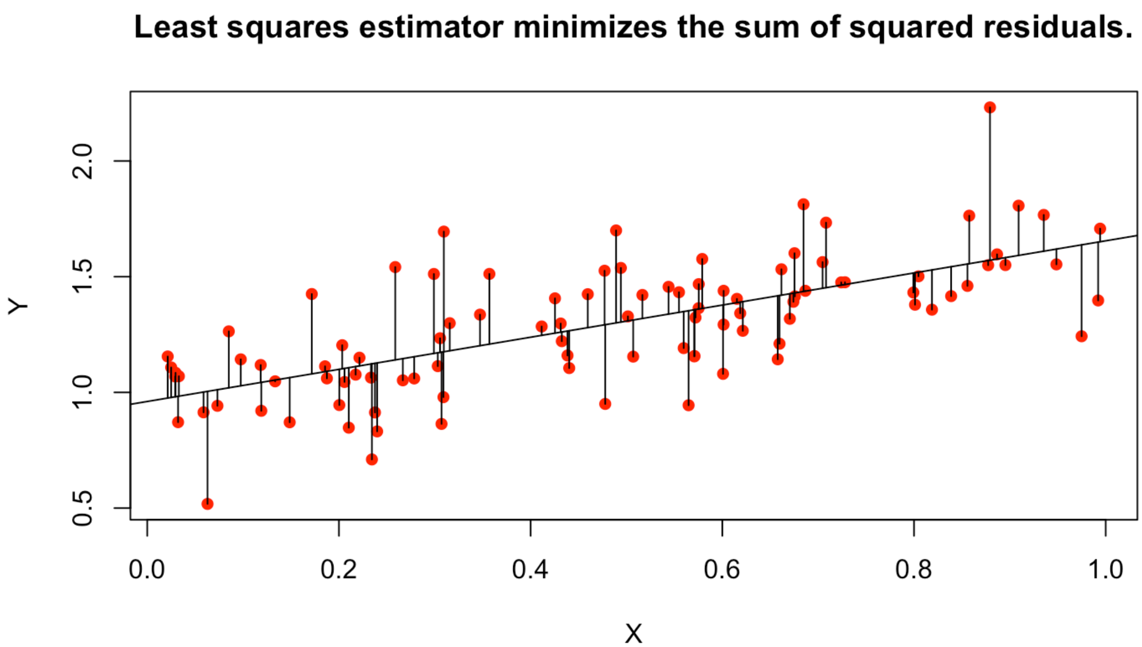
$$\sum_{i=1}^{n} \left\{ -1/2 \log(\sigma^2) - \frac{(y_i - X_i^T \beta)^2}{2\sigma^2} \right\} + C$$

求 $\beta$ 的MLE，等价于求如下的最小二乘：

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 := \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

# 最小二乘估计

$\|Y - X\beta\|_2^2 = (Y - X\beta)^T(Y - X\beta)$ 又称为残差平方和 (residual sum-of-squares).



# 最小二乘估计推导

$$\frac{\partial \|Y - X\beta\|_2^2}{\partial \beta} = -2X^T(Y - X\beta)$$

$$\frac{\partial^2 \|Y - X\beta\|_2^2}{\partial \beta \partial \beta^T} = 2X^T X$$

首先假设 $X^T X$ 是正定的。

Let $\frac{\partial \|Y - X\beta\|_2^2}{\partial \beta} = 0$, we have

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

# 最小二乘估计的性质

- 几何解释：
  - $(Y - X\hat{\beta}) \perp X_j, \ \ \forall j = 1, 2, \ldots, p.$
  - $\hat{Y} = X\hat{\beta}$ 是 $Y$ 在 $X$ 列空间的投影。

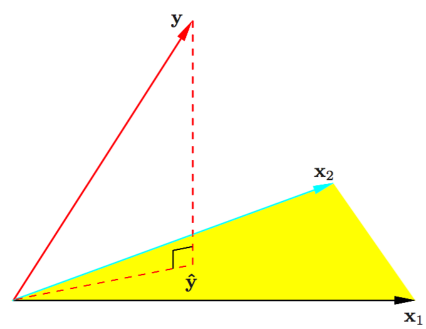



图 1: 最小二乘的几何解释

  - $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY := HY$

# 最小二乘估计的性质

- 无偏性

$$E(\hat{\beta}) = E[(X^TX)^{-1}X'Y] = E[(X^TX)^{-1}X'(X\beta + \epsilon)] = \beta$$

- 方差：

$$var(\hat{\beta}) = \sigma^2(X^TX)^{-1}X'X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

- 方差的估计：

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

证明：

$$E(\hat{\sigma}^2) = \sigma^2.$$

# 参数的区间估计与假设检验

- 为衡量点估计的不确定性，可以考察其分布
- 假设 $\epsilon \sim N(0, \sigma^2 I_n)$
- $\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$
- $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$
- $\hat{\beta}$ 与 $\hat{\sigma}^2$ 相互独立

# 置信区间

- 构造枢轴量（pivatol）
  - 当 $\sigma^2$ 已知时

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \sim N(0,1), P\Big(\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \in [-1.96, 1.96]\Big) = 0.95$$

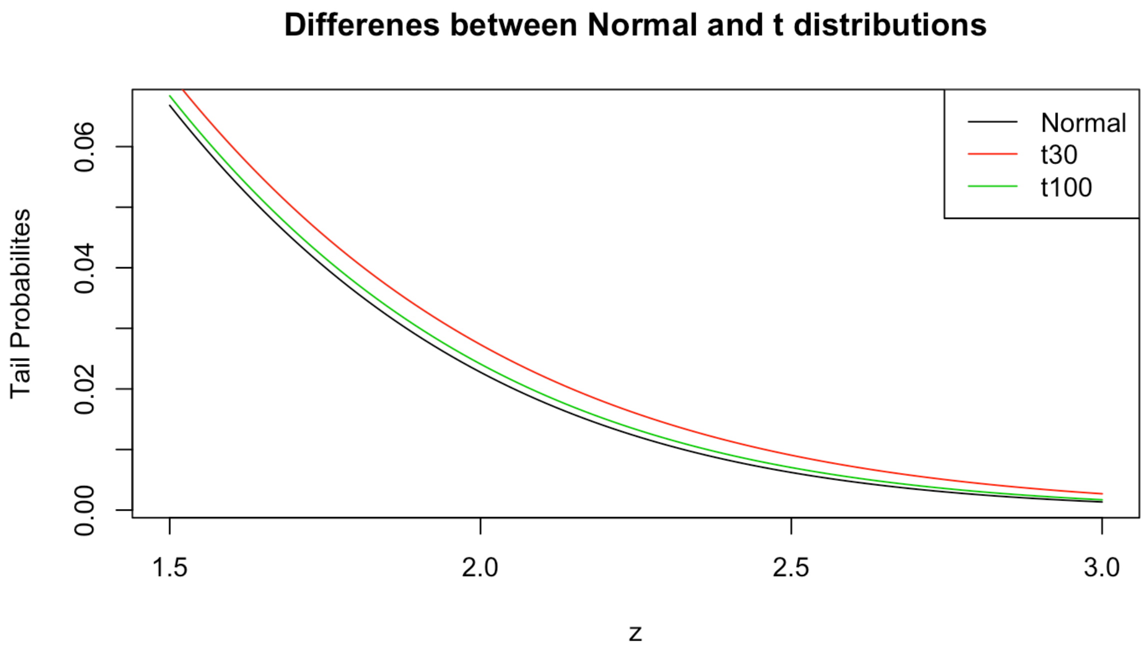
  - 当 $\sigma^2$ 未知时

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} / \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \sim t(n-p)$$

$$P\Big(\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}} \in [-t_{1-\alpha/2}, t_{1-\alpha/2}]\Big) = 1 - \alpha$$

# t分布和正态分布

当自由度很大时，比如>100, 两者的差距很小



# 多元参数的置信区域

- 构造枢轴量

$$\frac{(\hat{\beta}-\beta)^T(X'X)^{-1}(\hat{\beta}-\beta)}{\sigma^2} \sim \chi^2(p)$$

- 置信区域：

$$\left\{\beta \middle| \frac{(\hat{\beta}-\beta)^T(X'X)^{-1}(\hat{\beta}-\beta)}{\sigma^2} \le \chi^2_{1-\alpha}(p)\right\}$$

# 假设检验

- 单个参数的假设检验
  - $H_0 : \beta_j = 0$
  - 检验统计量

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}} \sim t(n-p)$$

  - 拒绝域

$$\{data : |T(data)| \geq t_{1-\alpha/2}\}$$

  - p-value: 统计量出现极端值的概率

$$P(|t(n-p)| > |T|)$$

# 假设检验

- 检验一组随机变量是否同时为0（变量选择）
  - 例：$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$
  - 一般地， $H_0 : Model_0 \ has \ p_0 \ varibales$ V.S. $H_1 : Model_1 \ has \ p_1 \ variables$； Model1 包含 Model0
  - 检验统计量

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1)} \sim F(p_1 - p_0, n - p_1)$$

  - t 检验是一个特殊的F检验

# 例子 (Linear regression in R)

```
data = read.table('./data/prostate.data')
head(data)
```

```
##       lcavol  lweight age      lbph svi       lcp gleason pgg45
## 1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0 -0
## 2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0 -0
## 3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20 -0
## 4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0 -0
## 5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0  0
## 6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0  0
##   train
## 1  TRUE
## 2  TRUE
## 3  TRUE
## 4  TRUE
## 5  TRUE
## 6  TRUE
```

# 例子 (Linear regression in R)

```
data <- read.table('./data/prostate.data')
Xtrain <- data[data$train == TRUE,]   ## this is equivalent to the ne
Xtrain <- data[data$train,]           ## this is equivalent to the ab
Xtest <- data[!data$train,]
cat(dim(Xtrain))
```

```
## 67 10
```

```
cat(dim(Xtest))
```

```
## 30 10
```

```
cat(dim(data))
```

```
## 97 10
```

# 例子 (Linear regression in R)

```
cor(data[,1:8])
```

```
##            lcavol    lweight       age         lbph         svi
## lcavol  1.0000000 0.28052138 0.2249999  0.027349703  0.53884500
## lweight 0.2805214 1.00000000 0.3479691  0.442264399  0.15538490
## age     0.2249999 0.34796911 1.0000000  0.350185896  0.11765804
## lbph    0.0273497 0.44226440 0.3501859  1.000000000 -0.08584324
## svi     0.5388450 0.15538490 0.1176580 -0.085843238  1.00000000
## lcp     0.6753105 0.16453714 0.1276678 -0.006999431  0.67311118
## gleason 0.4324171 0.05688209 0.2688916  0.077820447  0.32041222
## pgg45   0.4336522 0.10735379 0.2761124  0.078460018  0.45764762
##                  lcp    gleason      pgg45
## lcavol   0.675310484 0.43241706 0.43365225
## lweight  0.164537142 0.05688209 0.10735379
## age      0.127667752 0.26889160 0.27611245
## lbph    -0.006999431 0.07782045 0.07846002
## svi      0.673111185 0.32041222 0.45764762
## lcp      1.000000000 0.51483006 0.63152825
## gleason  0.514830063 1.00000000 0.75190451
## pgg45    0.631528246 0.75190451 1.00000000
```

# 例子 (Linear regression in R)

```
Xtrain = Xtrain[,-10]
obj = lm(lpsa ~ .,data = Xtrain )
summary(obj)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = Xtrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64870 -0.34147 -0.05424  0.44941  1.48675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.429170   1.553588   0.276  0.78334
## lcavol       0.576543   0.107438   5.366 1.47e-06 ***
## lweight      0.614020   0.223216   2.751  0.00792 **
## age         -0.019001   0.013612  -1.396  0.16806
## lbph         0.144848   0.070457   2.056  0.04431 *
## svi          0.737209   0.298555   2.469  0.01651 *
## lcp         -0.206324   0.110516  -1.867  0.06697 .
## gleason     -0.029503   0.201136  -0.147  0.88389
## pgg45        0.009465   0.005447   1.738  0.08755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7123 on 58 degrees of freedom
## Multiple R-squared:  0.6944, Adjusted R-squared:  0.6522
## F-statistic: 16.47 on 8 and 58 DF,  p-value: 2.042e-12
```

# R square and adjusted R square

- 平方和分解公式 （勾股定理）

$Y = X\hat{\beta} + residual$, $X$ 与 residual 垂直

所以 $\|Y\|_2^2 = \|X\hat{\beta}\|_2^2 + \|Residule\|_2^2$

每一项都减去均值，有

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{\hat{y}})^2 + \sum_i (r_i - \bar{r})^2$$

$$SS_{tot} = SS_{reg} + SS_{res}$$

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

# Adjusted R square

- 注意到 Least squares estimator 致力于 minimize $SS_{res}$
- 因此，加入的变量越多， $SS_{res}$ 越小，于是 $R^2$ 越大。
- 因此，要调整。
- 加入自由度！

$$\bar{R}^2 = 1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t} = 1 - \frac{SS_{res}/(n-p)}{SS_{tot}/(n-1)}$$

- 对比 $R^2$,

$$R^2 = 1 - \frac{SS_{res}/n}{SS_{tot}/n}$$

# Gauss–Markov Theorem

Statement: For i.i.d. errors, Least squares estimate is the best linear unbiased estimate (BLUE).

$\hat{\beta} = (X^T X)^{-1} X^T Y$

$\hat{\theta} = L^T Y, E(\hat{\theta}) = \beta$, then $Var(\hat{\beta}) \leq Var(\hat{\theta})$.

# Proof for Gauss–Markov Theorem

Let $L = X(X^T X)^{-1} + \Delta$, then

$$\beta = E(L^T Y) = \beta + \Delta^T X\beta$$

So, $\Delta^T X$ = 0

$$var(\hat{\theta}) = \sigma^2 L^T L$$

$$= \sigma^2 (X^T X^{-1} + \Delta^T \Delta) \geq \sigma^2 (X^T X^{-1})$$

# Mean squared error (MSE)

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

$$= E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2$$

$$= E(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2$$

$$= var(\hat{\theta}) + bias^2(\hat{\theta})$$

# Stein Estimator

- MSE/LSE 在无偏估计中是最好的
- 考虑到方差＋偏差的分解公式，MLE/LSE 或许不是最好的估计
- Stein Estimator 是一个例子

考虑一个简单的例子，$p$维观测 $x_1, x_2, \ldots, x_p$, 其中 $x_i$ independent from $N(\theta_i, 1)$. 问题：如何估计参数 $\theta_i$?

LSE：$\hat{\theta}_i^{(LSE)} = x_i$.

- 无偏
- $var(\hat{\theta}_i^{(LSE)}) = 1$

# Stein Estimator

现在考虑一种Bayes 方法， 假设 $\theta_i$, i.i.d from $N(0, \sigma^2)$.

$$p(\theta_i|x_i) \propto p(x_i|\theta_i)p(\theta_i)$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta_i)^2}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\theta_i^2}{2\sigma^2}}$$

So,

$$\theta_i|x_i \sim N((1 - \frac{1}{1+\sigma^2})x_i, 1 - \frac{1}{1+\sigma^2})$$

Under square loss, the best estimator is $E(\theta_i|x_i) = (1 - \frac{1}{1+\sigma^2})x_i$.

The loss is $1 - \frac{1}{1+\sigma^2} < 1$.

# Stein Estimator

- $\sigma^2$ 未知
- 可以从 data 去估计: $\sum x_i^2 \sim (1+\sigma^2)\chi^2(p)$
- Stein Estimator 给出一个具有更小的MSE的估计：

$$(1 - \frac{p-2}{\sum_i x_i^2})x_i$$

-Referecne:

Efron, B.; Morris, C. (1973). "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach". Journal of the American Statistical Association. American Statistical Association. 68 (341): 117–130. doi:10.2307/2284155

# 第一次作业

- due： 9月22日，周五

1. $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. 证明： $E(\hat{\sigma}^2) = \sigma^2$.

2. 假设 $\epsilon \sim N(0, \sigma^2 I_n)$， 证明:

   2.1 $\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$

   2.2 $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$

   2.3 $\hat{\beta}$ 与 $\hat{\sigma}^2$ 相互独立

3.证明:

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n-p_1)} \sim F(p_1 - p_0, n - p_1)$$

# 线性回归的缺点

- variance is big
- 在高维环境中，解释性不强
- 针对这两个特点，我们考虑变量选择

# 最佳子集选择 (Best subset selction)

- 对于所有变量集合$\{1, 2, ..., p\}$ 的任意子集，可以训练一个模型
- 规定模型的大小（有多少变量被选进模型），可以选出最好的子集
- 怎样选择最好的模型?

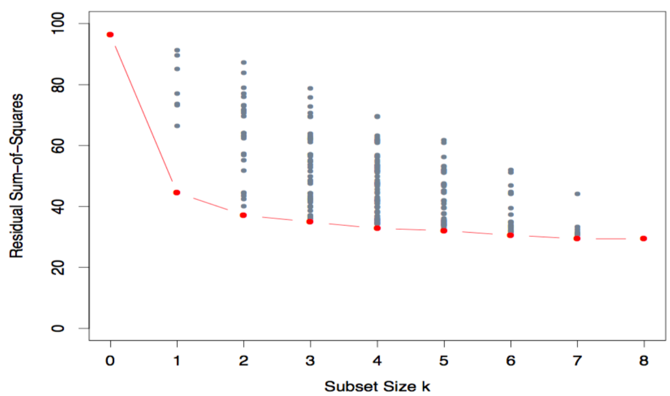



**FIGURE 3.5.** *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

图 2: 最佳子集选择

# 前进法(Forward Selection)

- 考虑一种贪婪的方法：逐渐增加变量
- 和最佳自己选择一样，要确定模型的大小
- 如何确定每次选择哪个变量？（Excersize 3.9）
- Forward Stage-wise Selection

# 后退法(Backward Selection)

- 从 Full model出发，每次删除一个变量
- 只针对 $n > p$ 的model

# 正则化方法（regularized method）

- Ridge Regression

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

- Solution

$$\hat{\beta} = (X^T X + \lambda)^{-1} X^T Y$$

- 如何选择 $\lambda$?
  - CV (Cross Validation)
  - Stein method?

# 正则化方法与 Bayes 之间的联系

- OLS V.S. Likelihood
- Ridge V.S. $\log(Likelihood \times e^{-\frac{\|\beta\|_2^2}{2\sigma^2}})$
- Lasso V.S. $\log(Likelihood \times e^{-\lambda\|\beta\|_1})$

# Lasso

- Least absolute shrinkage and selection operator

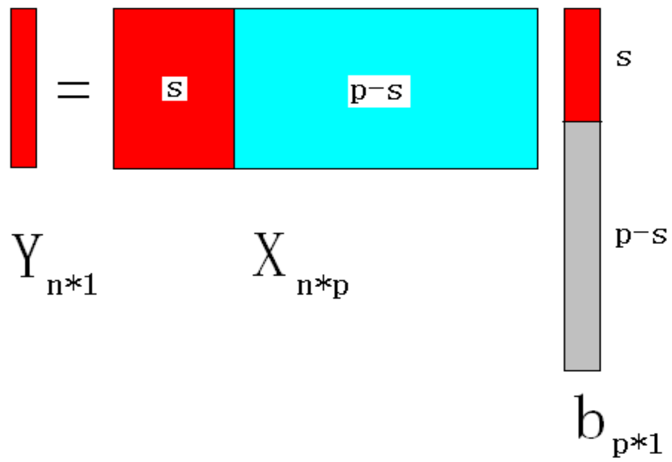



图 3: Lasso

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

# Compressed Sensing （压缩感知）

- 图像压缩基本原理

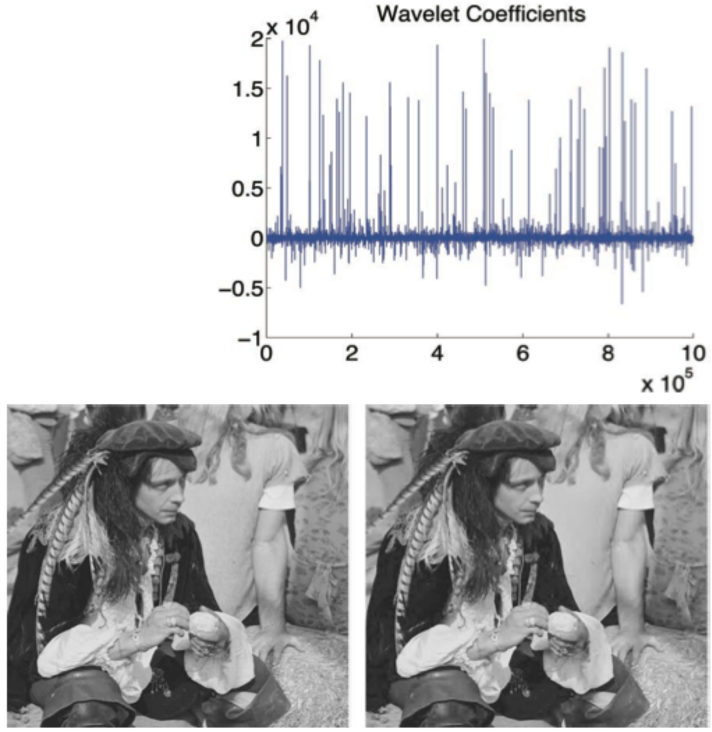



图 4: 图像压缩原理（From Dana Mackenzie (2009)）

# Compressed Sensing （压缩感知）

- Signal is sparse for some dictionary (say, wavelet space)

$$X_{N\times 1} = \Phi\beta.$$

- Signal X can be recovered by very few samples $Y$.

$$Y_{n\times 1} = \psi_{n\times N}X = \psi_{n\times N}\Phi\beta,$$

  where $\psi \in \mathbb{R}^{n\times N}$ is some sampling scheme.

- The recovery of the true signal $X$ depends on the fact that $\beta$ is sparse.
- The recovery process can be viewed as follows: (see Donoho, 2004; Candes and Tao, 2004; Candes and Tao, 2005)

# Compressed Sensing （压缩感知）

$$\min \|\beta\|_0 \quad s.t.\, Y = \psi\Phi\beta$$

Convex relaxation:

$$\min \|\beta\|_1 \quad s.t.\, Y = \psi\Phi\beta$$

# Compressed Sensing （压缩感知,一个应用）

- 单像素相机

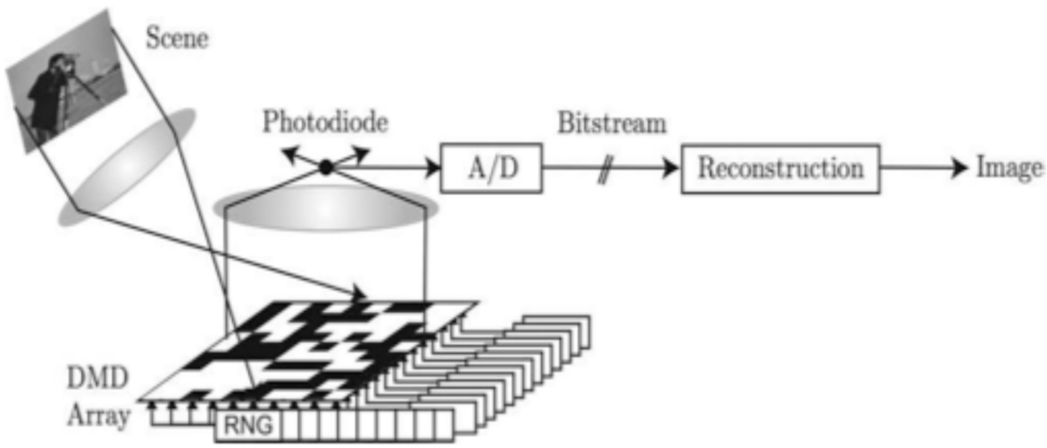



图 5: 单像素相机（From Dana Mackenzie (2009)）

# Compressed Sensing （压缩感知,一个应用）

- 单像素相机

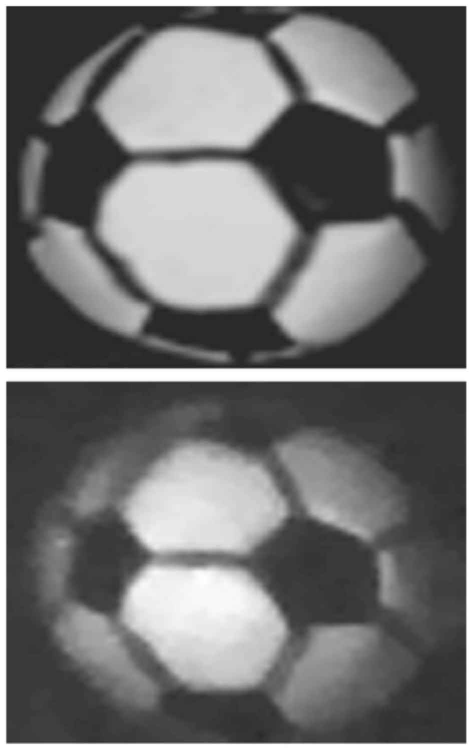


图 6: 单像素相机照片（From Dana Mackenzie (2009)）

# Lasso 求解

先看一个简单例子： $X^TX = I.$
$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 = \frac{1}{2}\beta^T\beta - Y^TX\beta + \lambda\|\beta\|_1$

易知，

$$\hat{\beta}_j = \begin{cases} X_j^TY - \lambda, & if \;\; X_j^TY > \lambda \\ 0, & if \;\; |X_j^TY| \le \lambda \\ X_j^TY + \lambda, & if \;\; X_j^TY < -\lambda \end{cases}$$

- soft thresholding

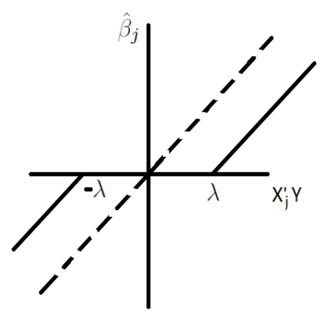



图 7: Soft thresholding

# Lasso 求解

- glmnet
- LARS
  - 安装： install.pacakges('lars')
  - 使用： library(lars)

# Lasso 求解 （Simulations）

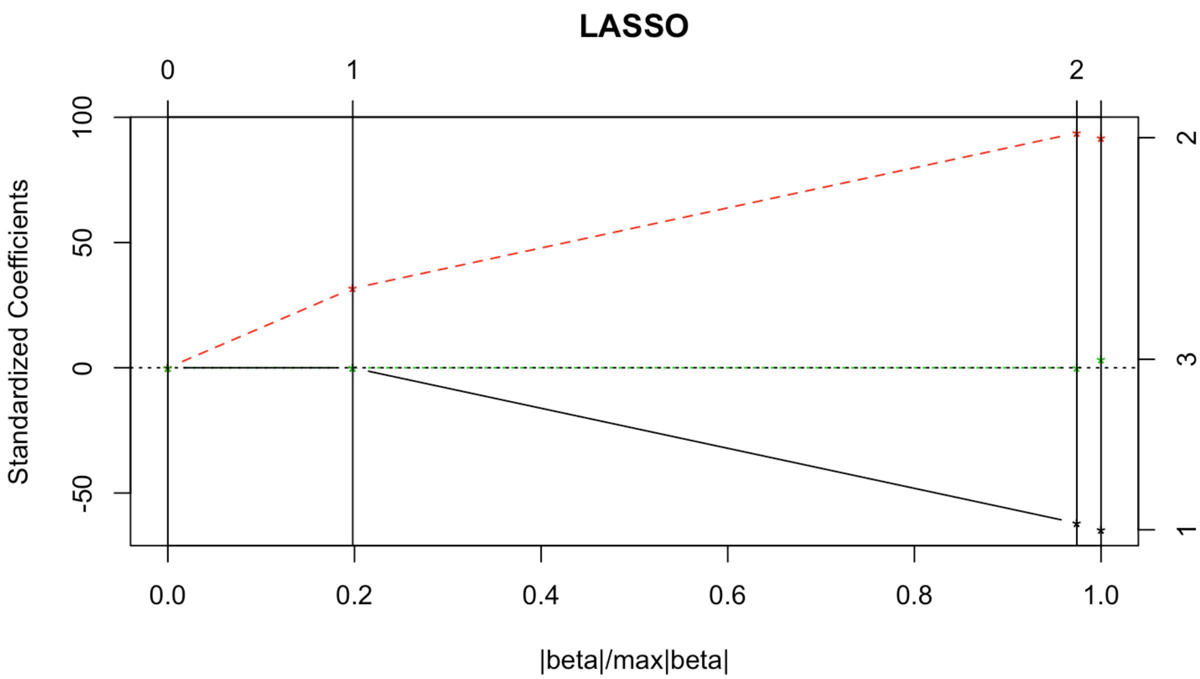
```
library(lars)
```

```
## Loaded lars 1.2
```

```
n = 1000
p = 3
beta1 = -2
beta2 = 3
X1 = rnorm(n)
X2 = rnorm(n)
e  = rnorm(n)
X3 =  2/3 *X1 + 2/3*X2 + 1/3*e
epsilon = rnorm(n)
Y = X1*beta1 + X2*beta2 + epsilon
X = cbind(X1,X2,X3)
obj = lars(X,Y)
```

# Lasso 求解 （Simulations）

```
plot(obj)
```



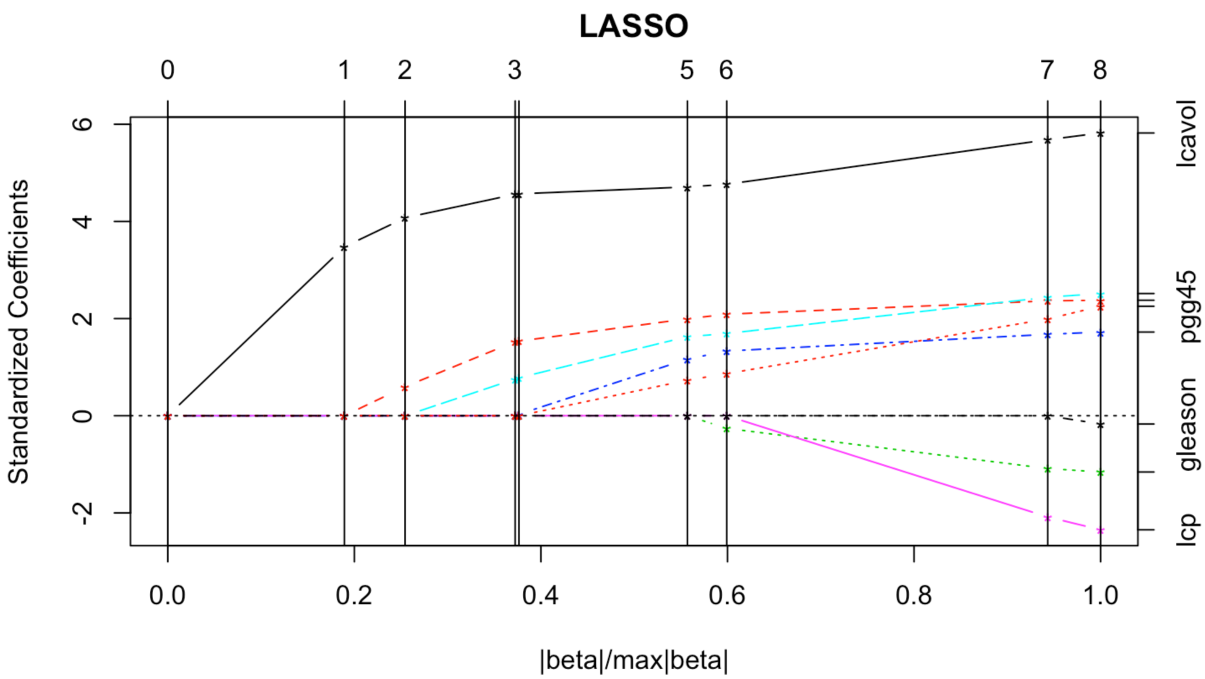
# Real data

```
data <- read.table('./data/prostate.data')
Xtrain <- data[data$train == TRUE,]   ## this is equivalent to the ne
Xtrain <- data[data$train,]           ## this is equivalent to the ab
Xtest <- data[!data$train,]
#cat(dim(Xtrain))
#cat(dim(Xtest))
#cat(dim(data))

X = as.matrix(Xtrain[,1:8])
Y = Xtrain[,9]
obj = lars(X,Y)
```

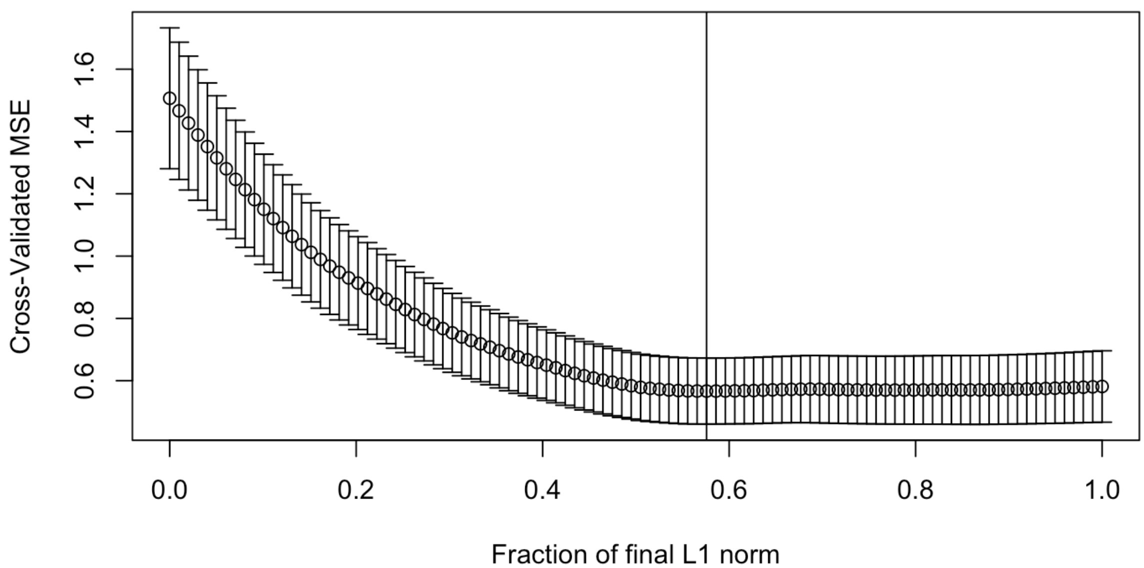
# Real data

```
myplot(obj)
```



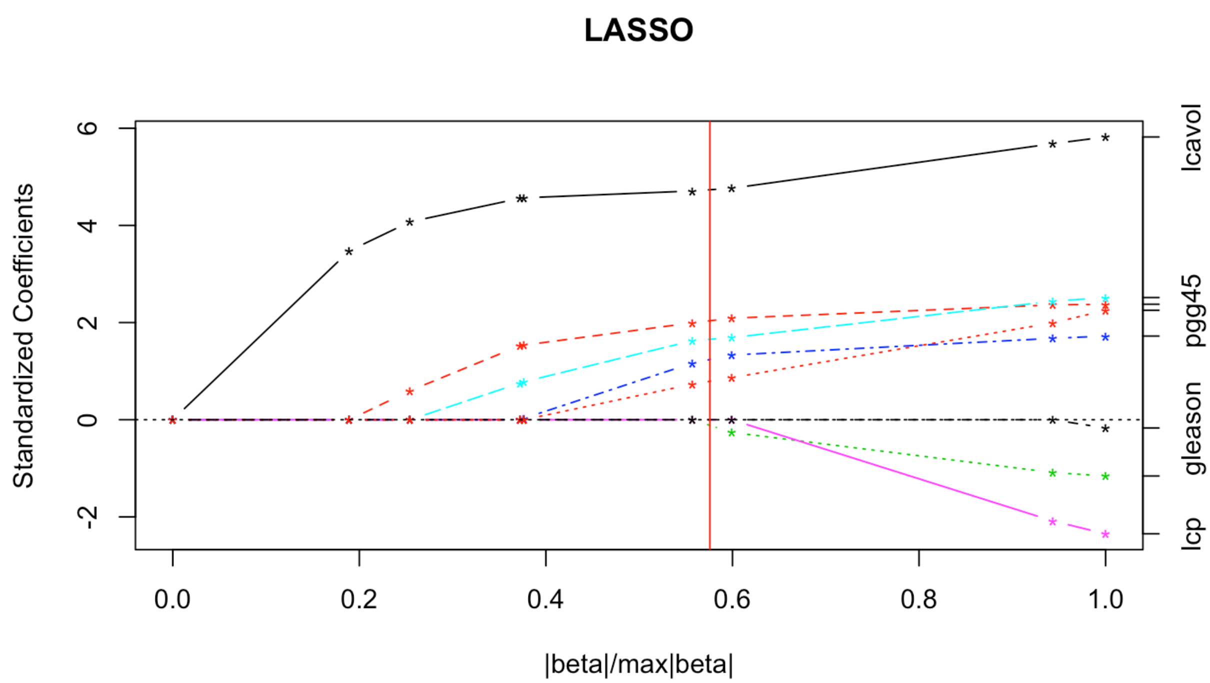
# Real data

```
cvobj = cv.lars(X,Y,K=10)
#cat(min(cvobj$cv))
k = which(cvobj$cv == min(cvobj$cv))
s = cvobj$index[k]
#cat(s)
abline(v=s)
```



# Real data

```
myplot(obj,breaks = FALSE)
abline(v=s,col='red')
```



# Real data

```r
coef(obj,s=s,mode = 'fraction')
```

```
##      lcavol     lweight         age        lbph         svi 
##  0.468927962  0.526173624 -0.001898395  0.104218278  0.485294925 
##          lcp     gleason       pgg45 
##  0.000000000  0.000000000  0.003325212
```

# 第二次作业

- due: OCT 6+

Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data[TABLE 3.3].

| Term | LS | Best Subset | Ridge | Lasso |
|---|---|---|---|---|
| Intercept | | | | |
| lcavol | | | | |
| ..... | | | | |
| Test Error | | | | |
| Std Error | | | | |

# Statistical Properties of the Lasso

**引理** 假设 $Y = X\beta^* + \epsilon$, 并假设$X_S^T X_S$ 是一个可逆矩阵。那么，Lasso的解 $\hat{\beta}(\lambda)$ 和 $\beta^*$ 有相同的符号，记作 $\hat{\beta}(\lambda) =_s \beta^*$，如果下面两个条件成立：

- （1）

$$\left| X_{S^c} X_S (X_S^T X_S)^{-1} \left[ \frac{1}{n} X_S^T \epsilon - \lambda sign(\beta^*(S) \right] - \frac{1}{n} X_{S^c}^T \epsilon \right| < \lambda$$

- （2）

$$\beta^*(S) + (\frac{1}{n} X_S^T X_S)^{-1} \left[ \frac{1}{n} X_S^T \epsilon - \lambda sign(\beta^*(S) \right] =_s \beta^*(S)$$

# 引理的证明

$\hat{\beta}$ 是 Lasso 的解 当且仅当，存在 $\vec{s} = (s_1, s_2, \ldots, s_p)$, 它满足

$$s_j = \begin{cases} 1, & if \ \ \hat{\beta}_j > 0 \\ [-1, 1], & if \ \ \hat{\beta}_j = 0 \\ -1, & if \ \ \hat{\beta}_j < 0 \end{cases}$$

且

$$-\frac{1}{n} X^T (Y - X\hat{\beta}) + \lambda \vec{s} = 0.$$

# 引理的证明 (续)

接下来我们要证明的是如果引理中的两个条件满足，我们可以构造一个LASSO 的解，它满足 $\hat{\beta}(\lambda) =_s \beta^*$。

特别地，这两个条件满足的时候，Lasso 的解唯一。

这样我们就可以证明该引理。

# 引理的证明（续）

我们这样构造一个$\hat{\beta}(\lambda)$: 它的分量分为两个部分 $\hat{\beta}(S)$ 和 $\hat{\beta}(S^c)$.
其
中，$\hat{\beta}(S^c) = 0$，$\hat{\beta}(S) := \beta^*(S) + (\frac{1}{n}X_S^T X_S)^{-1}\left[\frac{1}{n}X_S^T \epsilon - \lambda sign(\beta^*(S)\right]$

注意到引理中的条件（2）保证了 $sign(\hat{\beta}(S)) =_s sign(\beta^*(S))$.

我们同时定义$\vec{s}$, 它也相应地分为两个部分 $\vec{s}(S) := sign(\hat{\beta}(S))$,
$\vec{s}(S^c) := \frac{X_{S^c} X_S (X_S^T X_S)^{-1}\left[\frac{1}{n}X_S^T \epsilon - \lambda sign(\beta^*(S)\right] - \frac{1}{n}X_{S^c}^T \epsilon}{\lambda}$ 引理的条件（1）
保证了 $|\vec{s}(S)| < 1$.

易知，这样构造的$\hat{\beta}(\lambda)$是Lasso的解。 最后我们再证明: 引理
条件下，Lasso解的唯一性。

# 引理的证明 （续）

现证明唯一性。 为证明唯一性。 我们使用下面这一引理：

**引理 2** 如果 $\beta^+$ 是Lasso 的解， $\vec{s}$ 是满足Lasso的解的subgradient。 再假设$\beta^-$也是Lasso的解， 则$\vec{s}^T\beta^- = \|\beta^-\|_1$.

利用引理2，立即有 如果 $\tilde{\beta}$ 是另外一个解，则 $\vec{s}^T\tilde{\beta} = \|\tilde{\beta}\|_1$。由前面$\hat{\beta}(\lambda)$的构造知， $\vec{s}(S^c) < 1$. 于是 $\tilde{\beta}_{S^C} = 0$. 如果$X_S^T X_S$可逆，则最优化问题 $\min_\beta \|Y - X_S\beta\|_2^2 + \lambda\|\beta\|_1$的解唯一，因为此时目标函数是一个严格凸的函数。

# 引理2的证明

令 $f_0(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2$. 因为 $\beta^+$ 和 $\beta^-$ 都是Lasso问题的解，于是

$$f_0(\beta^+) + \lambda\|\beta^+\|_1 = f_0(\beta^-) + \lambda\|\beta^-\|_1$$

注意到 $\vec{s}^T\beta^+ = \|\|\beta^+\|_1\|_1$，上式写作

$$f_0(\beta^+) + \lambda\vec{s}^T\beta^+ = f_0(\beta^-) + \lambda\|\beta^-\|_1$$

两边都减去 $\lambda\vec{s}^T\beta^-$, 有

$$f_0(\beta^+) + \lambda\vec{s}^T(\beta^+ - \beta^-) = f_0(\beta^-) + \lambda(\|\beta^-\|_1 - \vec{s}^T\beta^-)$$

# 引理2的证明(续)

$$f_0(\beta^+) - f_0(\beta^-) + \lambda\vec{s}^T(\beta^+ - \beta^-) = \lambda(\|\beta^-\|_1 - \vec{s}^T\beta^-)$$

再注意 $\nabla f_0(\beta^+) = -\lambda\vec{s}$, 上式写作

$$f_0(\beta^+) - \nabla f_0(\beta^+)(\beta^+ - \beta^-) - f_0(\beta^-) = \lambda(\|\beta^-\|_1 - \vec{s}^T\beta^-)$$

利用 $f_0(\beta)$ 的凸函数性质

$$f_0(\beta^-) \geq f_0(\beta^+) + \nabla f_0(\beta^+)(\beta^- - \beta^+),$$

于是

$$\|\beta^-\|_1 \leq \vec{s}^T\beta^- \leq \max_s(\vec{s}^T\beta^-) = \|\beta^-\|_1,$$

即

$$\|\beta^-\|_1 = \vec{s}^T\beta^-.$$

# 进一步的分析 (续)

现定义两个量

$$V_j = X_j^T \left\{ X_S (X_S^T X_S)^{-1} \lambda \vec{s} - [X_S (X_S^T X_S)^{-1} X_S^T - I] \frac{\epsilon}{n} \right\}$$

$$U_i = e_i^T (\frac{1}{n} X_S^T X_S)^{-1} [\frac{1}{n} X_S^T \epsilon - \lambda \vec{s}]$$

易知，

$\max_{j \in S^c} |V_j| < \lambda$ 等价于 引理中的条件 （1）

$\max_{i \in S} |U_i| < min_{j \in S} |\beta_j|$ 可以推出 引理中的条件 （2）

# Sign consistency

定义两个随机事件

$$\mathcal{M}(V) = \left\{ \max_{j \in S^c} |V_j| < \lambda \right\}$$

$$\mathcal{M}(U) = \left\{ \max_{i \in S} |U_i| < min_{j \in S} |\beta_j| \right\}$$

如果我们能证明

$$P\Big( \mathcal{M}(V) \bigcap \mathcal{M}(U) \Big) \to 1,$$

我们就证明了Lasso的解是sign consistent 的。

# Sign consistency 的必要条件

注意到 $\hat{\beta} =_s \beta^*$ 可以轻松得到：

$$\left| X_{S^c} X_S (X_S^T X_S)^{-1} \left[ \frac{1}{n} X_S^T \epsilon - \lambda sign(\beta^*(S) \right] - \frac{1}{n} X_{S^c}^T \epsilon \right| \leq \lambda$$

即

$$X_j^T \left\{ X_S (X_S^T X_S)^{-1} \lambda \vec{s} - [X_S (X_S^T X_S)^{-1} X_S^T - I] \frac{\epsilon}{n} \right\} \leq \lambda, \forall j \in S^c$$

由此我们可以证明 如果存在$j$, 使得 $|X_j^T X_S (X_S^T X_S)^{-1} \vec{s}| > 1$, 那么

$$P(\hat{\beta} =_s \beta^*) \leq 1/2.$$

# Sign consistency 的必要条件

**定理** 假设线性模型成立 $Y = X\beta^* + \epsilon$, 其中 $\epsilon \sim N(0, \Sigma_\epsilon)$. 如果存在$j \in S^c$, 使得

$$\left| X_j^T X_S (X_S^T X_S)^{-1} \vec{s} \right| > 1,$$

那么

$$P(\hat{\beta} =_s \beta^*) \leq 1/2.$$

# Sign consistency 的必要条件

证明： 不妨令 $X_j^T X_S(X_S^T X_S)^{-1}\vec{s} = 1 + \zeta$.
则$V_j := X_j^T \left\{X_S(X_S^T X_S)^{-1}\lambda\vec{s} - [X_S(X_S^T X_S)^{-1}X_S^T - I]\frac{\epsilon}{n}\right\}$ 可以写成 $V_j = \lambda(1+\zeta) + \tilde{V}_j$, 其中 $\tilde{V}_j$ 服从均值为0 的正态分布。 于是 $P(\tilde{V}_j > 0) = 1/2$. 所以

$$P(V_j > \lambda) \geq 1/2.$$

Finally, we have

$$P(\hat{\beta} =_s \beta^*) \leq P(|V_j| \leq \lambda) \leq 1 - P(V_j \geq \lambda) \leq 1/2.$$

# Sign consistency 的充分条件

1. Irrepresentable condition:

$$\max_{j \in S^c} \left| X_j^T X_S (X_S^T X_S)^{-1} \vec{s} \right| \le 1 - \eta$$

   A stronger version:

$$\max_{j \in S^c} \left\| X_j^T X_S (X_S^T X_S)^{-1} \right\|_1 \le 1 - \eta$$

2. relationship between $n, p, s$:

$$n \gg s \log(p + 1)$$

3. signal to noise ratio:

$$\frac{n \min_{j \in S} \beta_j^*}{\sigma^2} \gg s \log(p + 1)$$

# Sign consistency 的充分条件

证明sign consistency 的主要工具是：分析 Gaussian (或者 subGaussian) variable。

**引理** 对于任意的均值为0的正态随机向量 $(X_1, X_2, \ldots, X_n)$, 对于任意的正数 $t > 0$, 我们有

$$P(\max_{1\leq i\leq n} |X_i| \geq t) \leq 2n \exp\Big\{\frac{-t^2}{2\max_i E(X_i^2)}\Big\}.$$

证明： $P(\max_{1\leq i\leq n} |X_i| \geq t) \leq nP(|X_i| \geq t)$
$\leq 2nP(X_i \geq t) \leq 2n \exp\Big\{\frac{-t^2}{2\max_i E(X_i^2)}\Big\}.$

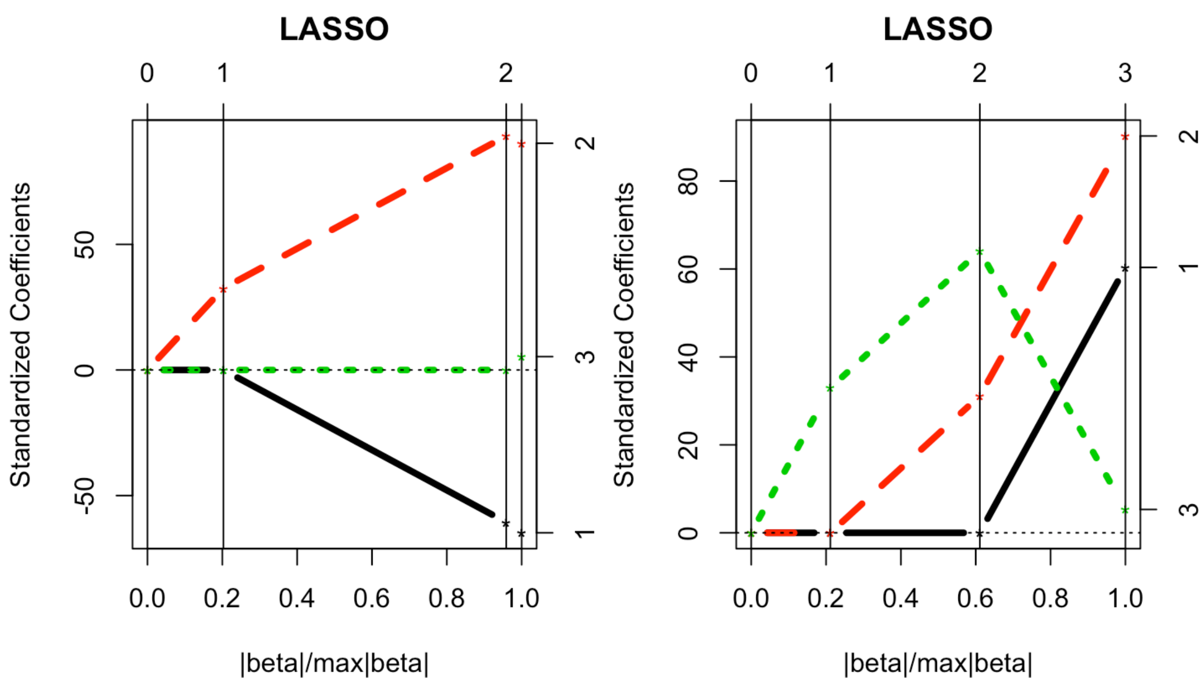
# 一个例子

```
library(lars)
n = 1000
p = 3
beta1 = -2
beta2 = 3
X1 = rnorm(n)
X2 = rnorm(n)
e  = rnorm(n)
X3 =  2/3 *X1 + 2/3*X2 + 1/3*e
epsilon = rnorm(n)
Y = X1*beta1 + X2*beta2 + epsilon
X = cbind(X1,X2,X3)
obj1 = lars(X,Y)

beta1 = 2
beta2 = 3
Y = X1*beta1 + X2*beta2 + epsilon
X = cbind(X1,X2,X3)
obj2 = lars(X,Y)
```

# 一个例子(续)

```
par(mfrow = c(1,2))
plot(obj1,lwd = 4)
plot(obj2,lwd = 4)
```



# L2 consistency

注意Lasso的解定义如下：

$$\hat{\beta} = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 := L(\beta) + \lambda\|\beta\|_1$$

为研究 $\hat{\beta}$ 与 $\beta^*$ 是不是很近，通常研究目标函数在这两点的取值之差。为保证研究目标函数在这两点的取值之差接近就能得到$\hat{\beta}$ 与 $\beta^*$ 很近， 需要假设Hessian 矩阵比较好。

以Least squares 为例，

$$L(\beta^* + \Delta) - L(\beta^*) - \langle \Delta, \frac{dL(\beta)}{d\beta} \rangle\Big|_{\beta=\beta^*} \geq \gamma\|\Delta^2\|$$

$$\frac{1}{n}\Delta^T(X^TX)\Delta \geq 2\gamma\|\Delta^2\|.$$

# L2 consistency

**定义** 如果存在$\gamma > 0$, $\frac{1}{n}\Delta^T(X^TX)\Delta \geq \gamma\|\Delta^2\|$, 对于任意的$\Delta$ 成立，称 $X$ 满足（正定）特征值条件。

**定义** 如果存在$\gamma > 0$, $\frac{1}{n}\Delta^T(X^TX)\Delta \geq \gamma\|\Delta^2\|$, 对于任意的$\Delta \in someset$ 成立，称 $X$ 满足限制特征值条件。

# L2 consistency

**引理3**： 对于Lasso的解 $\hat{\beta}$，令 $\Delta = \hat{\beta} - \beta^*$如果选择的$\lambda$ 满足 $\left|\frac{1}{n}X^T(Y - X\beta^*)\right| \leq \frac{1}{2}\lambda$, 则

$$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1.$$

**证明：** 注意到$\hat{\beta}$ 最优化 $\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$. 因此，

$$\frac{1}{2n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1$$

# L2 consistency

- 证明（续）

$$
\begin{aligned}
0 &\geq \left[\frac{1}{2n}\|Y - X\beta^* - X\Delta\|^2 - \frac{1}{2n}\|Y - X\beta^*\|^2\right] + [\lambda\|\beta^* + \Delta\|_1 - \lambda\|\beta^*\|_1] \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle\Delta, X^T(Y - X\beta^*)\rangle + \lambda(\|\beta^* + \Delta\|_1 - \|\beta^*\|_1) \\
&\geq \frac{1}{2n}\|X\Delta\|_2^2 - \left|\langle\Delta, \frac{1}{n}X^T(Y - X\beta^*)\rangle\right| + \lambda(\|\beta_S^* + \Delta_S + \Delta_{S^c}\|_1 - \|\beta_S^*\|_1) \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - (\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1)\max|\frac{1}{n}X^T(Y - X\beta^*)| + \lambda(\|\beta_S^* + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 - \|\beta_S^*| \\
&\geq \frac{1}{2n}\|X\Delta\|_2^2 - (\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1)\frac{\lambda}{2} + \lambda(\|\Delta_{S^c}\|_1 - \|\Delta_S\|_1) \\
&= \frac{1}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2}(\|\Delta_{S^c}\|_1 - 3\|\Delta_S\|_1) \\
&\geq \frac{\lambda}{2}(\|\Delta_{S^c}\|_1 - 3\|\Delta_S\|_1).
\end{aligned}
$$

# Restricted Eigenvalue Condition

**定义** 如果存在$\gamma > 0$, $\frac{1}{n}\Delta^T(X^TX)\Delta \geq \gamma\|\Delta^2\|$, 对于任意的$\Delta \in \{\Delta : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$ 成立，称 $X$ 满足限制特征值条件 $RE(S,3)$.

**引理4** 令 $\Delta = \hat{\beta} - \beta^*$. 如果 $RE(S,3)$ 成立，且选择的$\lambda$ 满足 $\left|\frac{1}{n}X^T(Y - X\beta^*)\right| \leq \frac{1}{2}\lambda$, 则

- $\|\Delta\|_2 \leq \frac{3\lambda\sqrt{s}}{\gamma}$
- $\frac{1}{n}\|X\Delta\|_2^2 \leq \frac{9\lambda^2 s}{\gamma}$
- $\|\Delta\|_1 \leq 4\sqrt{s}\|\Delta\|_2 \leq \frac{12\lambda s}{\gamma}$

# 引理4的证明

由前面引理3的证明，可知

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{3\lambda}{2}(\|\Delta_S\|_1)$$

条件$RE(S,3)$表明

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \gamma\|\Delta\|_2^2.$$

于是，我们立即有

$$\gamma\|\Delta\|_2^2 \leq 3\lambda\|\Delta_S\|_1$$

所以，

$$\|\Delta\|_2^2 \leq \frac{3\lambda}{\gamma}\|\Delta_S\|_1 \leq \sqrt{s}\frac{3\lambda}{\gamma}\|\Delta_S\|_2 \leq \sqrt{s}\frac{3\lambda}{\gamma}\|\Delta\|_2$$

# 引理4的证明(续)

即

$$\|\Delta\|_2 \le \frac{3\lambda\sqrt{s}}{\gamma}$$

再利用

$$\frac{1}{2n}\|X\Delta\|_2^2 \le \frac{3\lambda}{2}(\|\Delta_S\|_1)$$

有

$$\frac{1}{n}\|X\Delta\|_2^2 \le 3\lambda(\|\Delta_S\|_1) \le 3\sqrt{s}\lambda\|\Delta\|_2 \le \frac{9\lambda^2 s}{\gamma}$$

# 引理4的证明(续)

最后

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{s}\|\Delta\|_2 \leq \frac{12\lambda s}{\gamma}.$$

# L2 consistency of the Lasso

**定理** 假定线性模型成立 $Y = X\beta^* + \epsilon$ with $\epsilon_i \sim N(0, \sigma^2)$. 再假设数据的每一列都是归一化的：$\frac{X_j^T X_j}{n} = 1$. 如果restricted eigenvalue condition $RE(S, 3)$成立，则通过取 $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$, 其中 $A > 2\sqrt{2}$, 则下列三个事件发生的概率大于 $1 - 2p^{1-A^2/8}$.

- $\|\Delta\|_2 \le \frac{3A\sigma}{\gamma}\sqrt{\frac{s \log p}{n}}$
- $\frac{1}{n}\|X\Delta\|_2^2 \le \frac{9A^2\sigma^2 s \log p}{n\gamma}$
- $\|\Delta\|_1 \le 4\sqrt{s}\|\Delta\|_2 \le \frac{12sA\sigma}{\gamma}\sqrt{\frac{\log p}{n}}$

# 定理的证明

由引理4，我们只需要证明 当 $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$ 时，

$$P\Big(\Big|\frac{1}{n}X^T(Y - X\beta^*)\Big| \leq \frac{1}{2}\lambda\Big) \geq 1 - 2p^{1-A^2/8}$$

即可。

易知 $\Big|\frac{1}{n}X^T(Y - X\beta^*)$ 的每一个分量 $\frac{1}{n}X_j^T\epsilon$ 服从正态分布，$N(0, \frac{1}{n}\sigma^2)$. 于是，

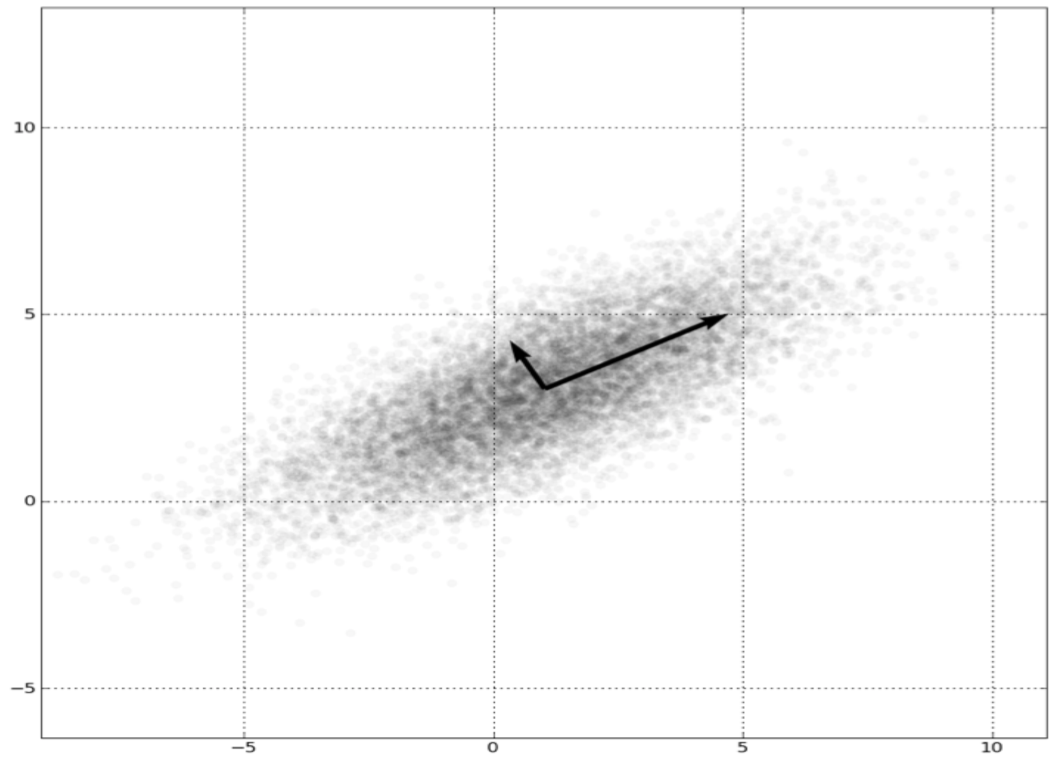
$$P\Big(\max_j |\frac{1}{n}X_j^T\epsilon| \geq \frac{1}{2}\lambda\Big) \leq 2pe^{-\frac{n\lambda^2}{2\times 4\sigma^2}} = 2p^{1-A^2/8}.$$

# 其他的降维方法

- PCA
- LDA
- Partial Least Squares

# PCA (主成分分析)

- PCA 示意图



# PCA (主成分分析)

- 降维
- $X = [X_1, X_2, \ldots, X_p]$
- 投影 data 至一维直线
- $Z = a_1X_1 + a_2X_2 + \cdots + a_pX_p$
- 怎样确定这些系数 $a$?
- 目标： $\max var(Z) = a^T\Sigma a$
- $a$ 就是协方差矩阵的最大特征根对应的特征向量
- $Z$ 称为第一主成分

# PCA

- 第 $k$ 主成分，其系数 $u_k$ 与前面的第 $1, 2, \ldots, k-1$主成分之间的关系：
  - $u_k$ 与前面的向量都垂直
  - $u_k = \arg\max_u var(u^T X) = \arg\max_u u^T \Sigma u$
  - $u_k$ 是协方差矩阵第$k$大特征根对应的特征向量
- 在实际数据分析中，通常用样本协方差代替总体协方差。
- 记$X$是中心化的数据[每列的数据减去均值]。 各主成分的系数是矩阵 $X^T X$ 的特征向量。

# 主成分分析的步骤

- $X = UDV^T$
- $X^TX = VDV^T$
- 第一主成分的系数为 $V[:,1]$
- 第 $k$ 主成分的系数为 $V[:,k]$
- 第一主成分为 $XV[:,1]$
- 注意到 $XV = UD$
- 所以第一主成分为 $UD$的第一列，即$U[:,1] \times D_1$

# PCA 的简单性质

主成分分析一个性质：寻找一个矩阵$Z \in R^{n\times p}$ 它的秩是 $K$,且最接近 $X$, 则 $Z$ 就是$X$ 的前$K$ 个主成分。

因此， 主成分分析可以看成是对数据降噪的一种处理。

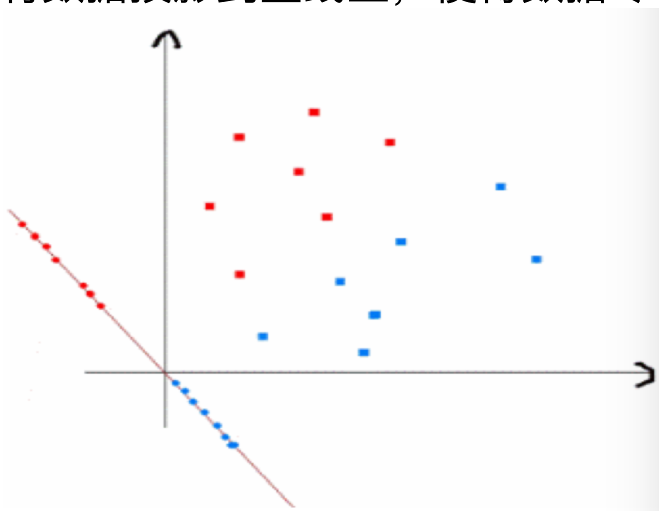
# 主成分回归

用前$K$个主成分当作新的回归变量做回归分析，就称为主成分回归。

- 主成分回归的优点：
  - 降维
  - 减少共线性
- 缺点：
  - 没有使用 $Y$ 的信息
  - 只用线性 （没有非线性部分）

# LDA （Linear Discriminant Analysis）

- 线性分类器:将数据投影到直线上，使得数据尽可能分开!



# LDA

- 怎样寻找这条直线?
- 规则:最大化组间方差，最小化组内方差。
- 组间方差：

$$var([w^T x^{(1)}, w^T x^{(2)}, \dots, w^T x^{(K)}])$$

- 组内方差

$$\sum_{i=1}^{K} var(w^T x^{(i)})$$

- 最大化目标函数：

$$\frac{var([w^T x^{(1)}, w^T x^{(2)}, \dots, w^T x^{(K)}])}{\sum_{i=1}^{K} var(w^T x^{(i)})}$$

# LDA

- 组间方差的计算

$$w^T \frac{1}{K} (\bar{x}^{(i)} - \bar{x})(\bar{x}^{(i)} - \bar{x})^T w$$

- 组内方差的计算

$$w^T \frac{1}{K} \Big( \sum_{i=1}^{K} \hat{\Sigma}_i \Big) w$$

- 目标：寻找 $w$, 使之,

$$\max_w \frac{w^T S_B w}{w^T S_W w}.$$

# LDA

- $w$ 是最大广义特征根对应的特征向量。也就是 $S_W^{-1}S_B$的特征向量。
- 如果是多类问题，显然一个方向的投影是不够的，需要多个投影
- 可以像PCA那样，得到更多的投影方向

# 两类问题的LDA

- 目标函数为

$$\frac{\sigma_{between}^2}{\sigma_{within}^2} = \frac{(w^T\mu_1 - w^T\mu_0)^2}{w^T\Sigma_1 w + w^T\Sigma_0 w}$$

- 可以证明 当 $w \propto (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)$ 时，目标函数达到最大值。

# 两类问题的LDA

用似然比（极大似然、Bayes）等思想，可以知道如果两组正态数据之间的最好的分割线是

$$
\begin{aligned}
&(x-\mu_0)^T \Sigma_0 (x-\mu_0) + \log|\Sigma_0| \\
&- (x-\mu_1)^T \Sigma_1 (x-\mu_1) - \log|\Sigma_1| < T
\end{aligned}
$$

假设$\Sigma_1 = \Sigma_0 = \Sigma$ 则判别是一个线性判别，判别条件是

$$
w^T x + c > T,
$$

其中 $w = \Sigma^{-1}(\mu_1 - \mu_0)$,

$c = \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 + \mu_1^T \Sigma^{-1} \mu_1)$.

# 作业

- due: Nov 2

1. 证明 Lasso 的sign consistency
2. 证明 Danzig Selector 的L2 consistency。 Danzig Selector 定义如下：

$$
\begin{aligned}
\hat{\beta} &= \arg\min \|\beta\|_1 \\
s.t. &\left|\frac{1}{n}X^T(Y - X\beta)\right| \leq \lambda
\end{aligned}
$$

3. 设 $X \in R^{n\times p}$, 寻找一个矩阵$Z \in R^{n\times K}$ 它的秩是 $K$,且最接近 $X$, 则 $Z$ 就是$X$ 的前$K$ 个主成分。即下面最优化的解是 $X$ 的前 $K$ 主成分。

$$
\min_Z \|XX^T - ZZ^T\|_F
$$

# Partial Least Squares

- 和 PCA 、 LDA 一样，这个方法也是将数据投影到一些正交的方向
- 和 PCA 一样，也是希望数据投影过来之后，尽可能能的分散
- 和 PCA 不一样的是，这个PLS方法要用到$Y$ 的信息。

注意，计算PCA的第 $m$ 个主成分的系数$v_l$时，是解下面的最优化问题：

$$\max_{\alpha} Var(X\alpha)$$
$$s.t. \|\alpha\|_2 = 1, \alpha^T S v_l = 0, l = 1, 2, \ldots, m-1,$$

其中，$S = X^T X$ the sample covariance matrix.

# Partial Least Squares (续)

PLS 用到了$y$ 的信息求解这些方向：计算PLS的第$m$ 个方向$\phi_m$的定义如下：

$$
\begin{aligned}
& \max_{\alpha} Corr^2(y, X\alpha)Var(X\alpha) \\
& s.t. \|\alpha\|_2 = 1, \alpha^T S\phi_l = 0, l = 1, 2, \ldots, m-1,
\end{aligned}
$$

其中，$S = X^T X$ the sample covariance matrix.

# 一些思考题

1. 两类问题的LDA 和 Least square是等价的 （这个已知）
2. 两类问题的QDA 是否也可以转化为Least square? （未知）
3. LDA 和 Partial Least Square 是否可以合并为一个最优化问题? （未知）