平成24年度　東京大学理学部情報科学科

月曜2限10:30～12:00

理7-214

知識処理論

2011年5月14日(月)

井元清哉
東京大学医科学研究所
ヒトゲノム解析センター
DNA情報解析分野
http://bonsai.hgc.jp/~imoto
imoto@ims.u-tokyo.ac.jp

線形重回帰モデルの変数選択

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{22} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \ldots, \beta_p)^T$$

- $\boldsymbol{\beta}^{M_1} = (\beta_1, 0, 0, 0, \ldots, 0)^T$: Model M_1
- $\boldsymbol{\beta}^{M_2} = (\beta_1, \beta_2, 0, 0, \ldots, 0)^T$: Model M_2
- …

赤池情報量規準(AIC; Akaike, 1973)

$$\mathrm{AIC}(M_k) = n \log \hat{\sigma}^2_{M_k} + 2|M_k|$$

$\boldsymbol{\beta}^{M_k}$ の非ゼロ成分の数

Note: 定数項は除いています

$$\hat{M} = \arg\min_k \mathrm{AIC}(M_k)$$

線形重回帰モデル

データ: $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$

線形重回帰モデル:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{22} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

y を説明(予測)するための変数が p 個ある。
どれが必要なものであろうか?

統計的モデル選択

線形重回帰モデルの変数選択

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{22} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \ldots, \beta_p)^T$$

- $\boldsymbol{\beta}^{M_1} = (\beta_1, 0, 0, 0, \ldots, 0)^T$: Model M_1
- $\boldsymbol{\beta}^{M_2} = (\beta_1, \beta_2, 0, 0, \ldots, 0)^T$: Model M_2
- …

赤池情報量規準(AIC; Akaike, 1973)

$$\mathrm{AIC}(M_k) = n \log \hat{\sigma}^2_{M_k} + 2|M_k|$$

Note: 定数項は除いています

$\boldsymbol{\beta}^{M_k}$ の非ゼロ成分の数

$$\hat{M} = \arg\min_k \mathrm{AIC}(M_k)$$

候補モデルの数

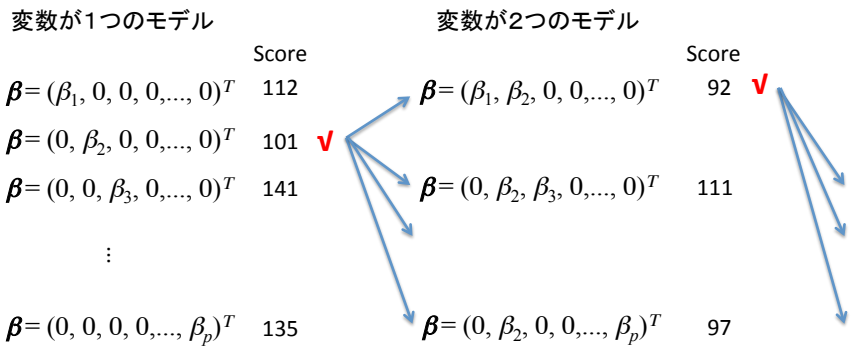
$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \ldots, \beta_p)^T$$

それぞれのパラメータが 0 か否か ⇒ 2^p

10 程度の p なら完全探索が出来るが、100 だと？

変数増加法

A Greedy Search

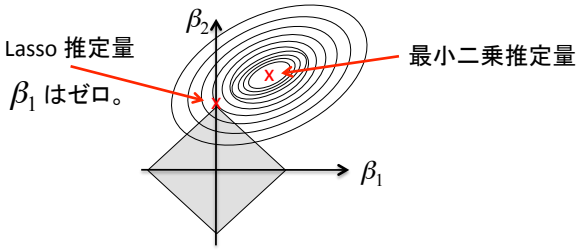


Lasso (Tibshirani, JRSS B, 1996)

$S_{\mathrm{L}}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$ を最小にする $\boldsymbol{\beta}$ を求める

ラグランジュの未定係数法

$\sum_{j=1}^{p}|\beta_j| < t$ の下で $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ の最小化



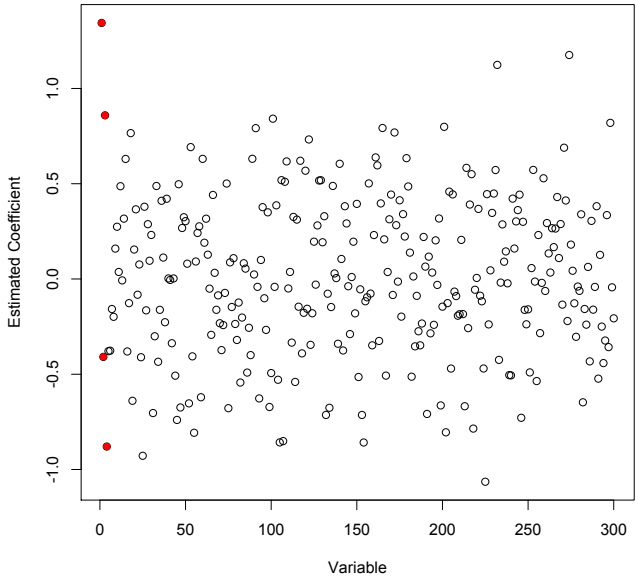
簡単な数値実験

$X_1, X_2, \ldots, X_{300} \sim N(0,1)$

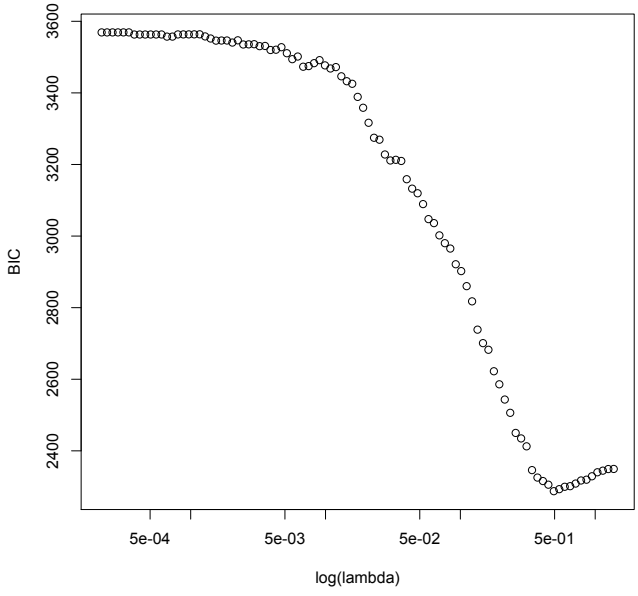
$Y = x_1 - x_2 + x_3 - x_4 + \varepsilon, \quad \varepsilon \sim N(0,4)$

データ: $\{(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i); i = 1, \ldots, 400\}$

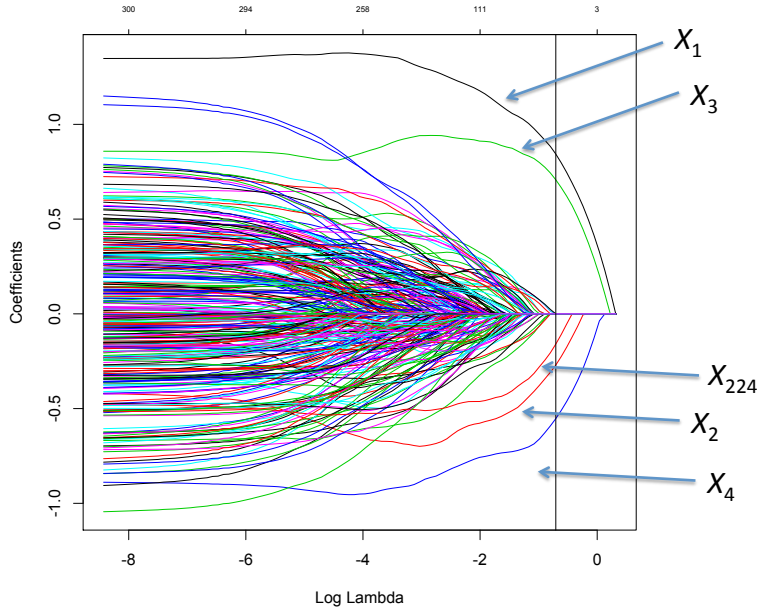
最小2乗法で推定された300個のパラメータ値



Lasso パラメータの選択



各 λ におけるパラメータの推定値



ベイズ型モデル評価基準 BIC
Bayesian Information Criterion

データ: $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$

モデル: $M_j : f_j(x \mid \boldsymbol{\theta}_j), \quad \boldsymbol{\theta}_j \in \Theta_j \subset R^{p_j}$

モデル M_j の事後確率: $\Pr(M_j \mid \boldsymbol{x}_n) \propto \Pr(M_j) p(\boldsymbol{x}_n \mid M_j)$

周辺尤度: $p(\boldsymbol{x}_n \mid M_j) = \int f_j(\boldsymbol{x}_n \mid \boldsymbol{\theta}_j) \pi_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j$

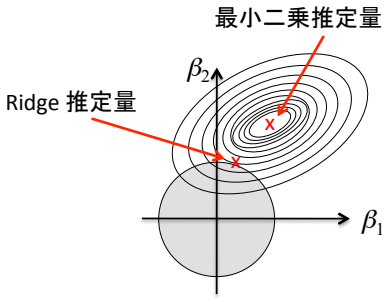
近似: $-2 \log p(\boldsymbol{x}_n \mid M_j) \approx -2 \log f_j(\boldsymbol{x}_n \mid \hat{\boldsymbol{\theta}}_j) + 2 p_j \log n$

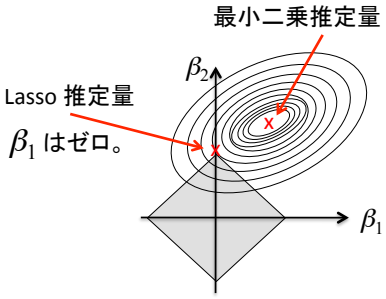
$$\hat{\boldsymbol{\theta}}_j = \arg\max_{\boldsymbol{\theta}_j} \{\log f_j(\boldsymbol{x}_n \mid \boldsymbol{\theta}_j)\}$$

$$\mathrm{BIC}(M_j) = -2 \log f_j(\boldsymbol{x}_n \mid \hat{\boldsymbol{\theta}}_j) + 2 p_j \log n$$

Ridge 推定量　と　Lasso 推定量

$$S_\lambda(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \qquad S_{\mathrm{L}}(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$$





Ridge 推定量再考

$\hat{\boldsymbol{\beta}}_{\mathrm{P}} = \min_{\boldsymbol{\beta}} S_\lambda(\boldsymbol{\beta})$ は次式で与えられる:

$$\hat{\boldsymbol{\beta}}_{\mathrm{P}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

今、$\boldsymbol{X} = (\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(p)})$ において、各 $\boldsymbol{x}_{(i)} = (x_{1i},\ldots,x_{ni})^T$ が平均 0 にシフトされている

このとき、$\boldsymbol{X}^T\boldsymbol{X}$ は

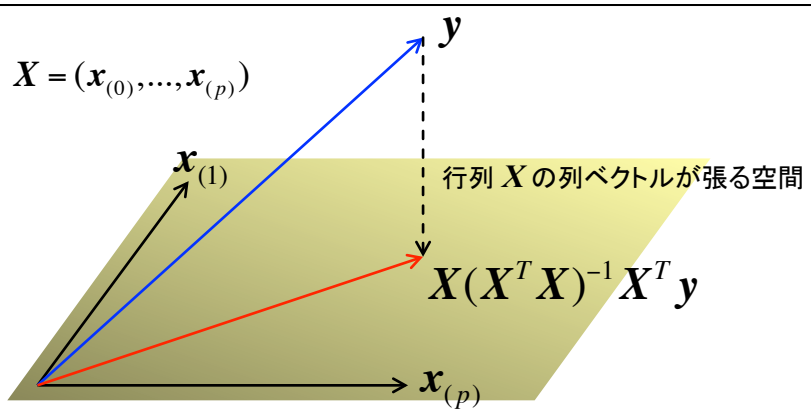
(i, i) 成分　$\displaystyle\sum_{k=1}^{n} x_{ki}^2 = \sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2 = n\cdot\mathrm{Var}(X_i)$

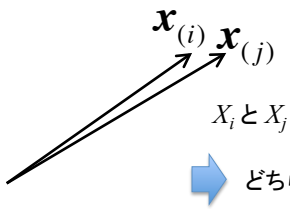
(i, j) 成分　$\displaystyle\sum_{k=1}^{n} x_{ki}x_{kj} = \sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j) = n\cdot\mathrm{Cov}(X_i, X_j)$

最小2乗推定量　　Ridge 推定量

$$\frac{(i,j)\text{ 成分}}{(i,i)\text{ 成分}} \quad \frac{\mathrm{Cov}(X_i,X_j)}{\mathrm{Var}(X_i)} > \frac{\mathrm{Cov}(X_i,X_j)}{\mathrm{Var}(X_i)+n\cdot\lambda}$$

$X_i\,(X_j)$ の分散を大きく見せることで、相対的に
X_i と X_j の共分散を小さく見せている





X_i と X_j は同じような情報を持っている

どちらかで良いんじゃない？

ケチの原理

Parsimonious Principle

$$\mathrm{AIC}(M_k) = n\log\hat{\sigma}^2_{M_k} + 2|M_k|$$

尤度の項 = モデルのデータへの当てはまりの程度を表す

もし、X_i と X_j は同じような情報を持っているならば、X_i と X_j の両方を入れたモデル M と片方だけのモデル M' はデータへの当てはまりはあまり変わらない

$$\hat{\sigma}^2_M \approx \hat{\sigma}^2_{M'}$$

ならば、第2項のパラメータ数が少ないモデルが良いモデル

データ数とパラメータ数の関係にも注意

モデルと解釈

データから構築された統計モデルは、データを取得した分野における専門的な観点から評価・解釈される。

X_i と X_j は同じような変数なのに、なんで X_i だけ採用されて X_j は採用されないのか？

もし、X_i と X_j が本当は同じ情報を持っていたとしても、データ数は有限であるため、たまたまの計測ノイズの乗り方でどちらが選ばれるかが決まってしまう。

推定されたモデルの解釈が困難になる。

Ridge 推定量は、変数間の共分散（相関）を相対的に小さく見せることによってそれらの変数を積極的にモデルに取り込む方法とも見なすことが出来る。

Lasso で選ばれる変数

Lasso には、Ridge 推定量のように変数間の共分散を小さく見せるといった
トリックは入っていない。

つまり、Lasso で選ばれる変数は、最小2乗推定量＋情報量規準による
モデル選択と同じ性質を持つ。

類似性の高い（グループを形成するような）変数から1つの変数を選ぶのではなく
それらの変数の多くをモデルに残し、かつ、変数選択が Lasso のように自動的に
出来るような方法。

Elastic net という方法がある。

Elastic net (Zou and Hastie, JRSS B, 2005)

$$S_{\mathrm{E}}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j| + \frac{\nu}{2}\sum_{j=1}^{p}\beta_j^2$$ を最小にする $\boldsymbol{\beta}$ を求める

Lasso　　Ridge

変数選択　　X の共分散構造を考慮