- SAS共同企画セッション -

一世を風靡したRandom Forest (Random Woods)がSAS(IMSTAT)で使えるようになったので縦長/横長データに適用してみる

塩野義製薬株式会社

木口 亮, 北西 由武, 都地 昭夫, 渡辺 秀章

- SAS joint planning session -

Make an attempt to apply Random Forest (Random Woods) available in IMSTAT to huge records' / super multi-dimensional data

Shionogi & Co., Ltd.

Ryo Kiguchi, Yoshitake Kitanishi, Akio Tsuji, Hideaki Watanabe

要旨：

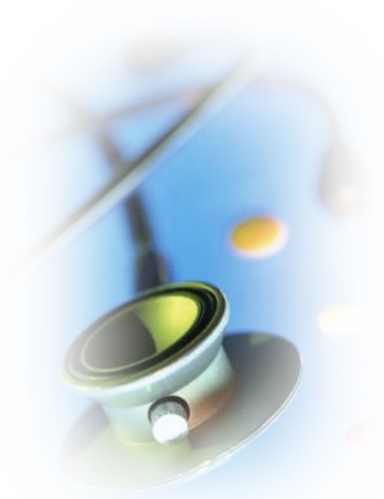
縦長/横長の仮想データに対して，ロバストな変数選択の手法の一つであるRandom ForestをIMSTATで適用し，その性能を見る．また，その他の変数選択の手法との比較も行う．

キーワード：IMSTAT　RANDOMWOOD　GLMSELECT　Lasso　Elastic Net

Outline

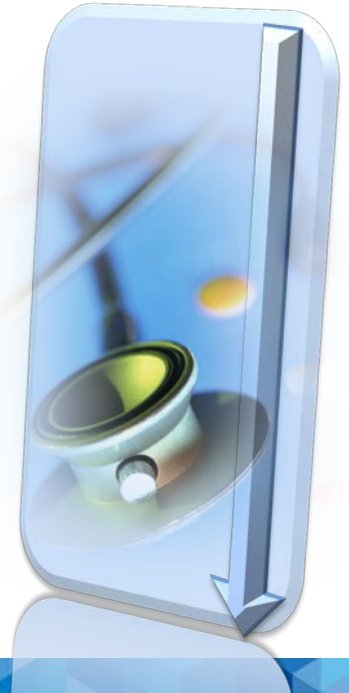
- 縦長データと横長データ
- 重要な変数の選択
  - ① Random Forest
  - ② Lasso / Elastic net
- シミュレーション
- まとめ
- 課題

■ 縦長データと横長データ



縦長データ　EX.) 医療ビッグデータ

- 医師の診療行為から生まれるビッグデータに，カルテ，レセプト（診療報酬明細）などをソースとして得られるものがある
- 患者さん一人ひとりの薬剤処方歴，手術歴，診断歴，入院歴といった情報が蓄積されたデータ
- Real World Data（RWD) と呼ばれることも多い
- 一般に**超多サンプル少変数**のデータ構造の場合が多い



横長データ　EX.) ゲノムビッグデータ

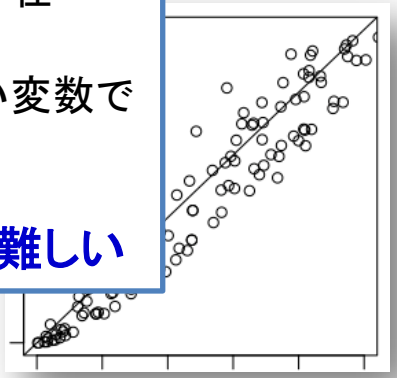
- 従来から扱っている臨床試験データは少サンプル多変数
- 遺伝子データに至っては，

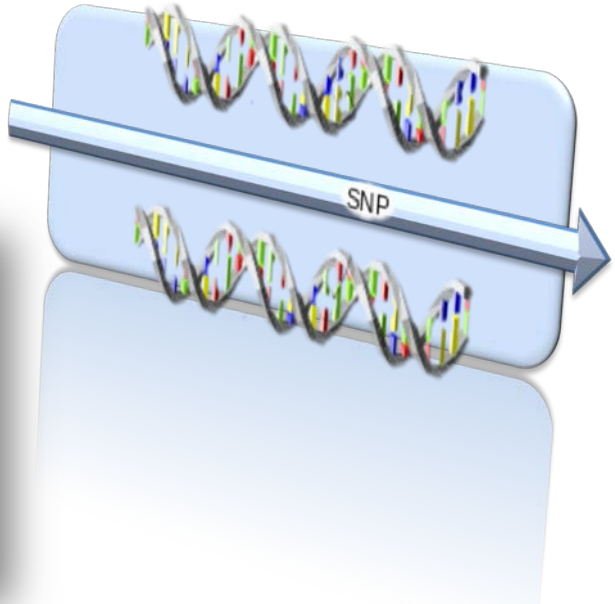
少サンプル超多変数のデータ構造をとることが多い

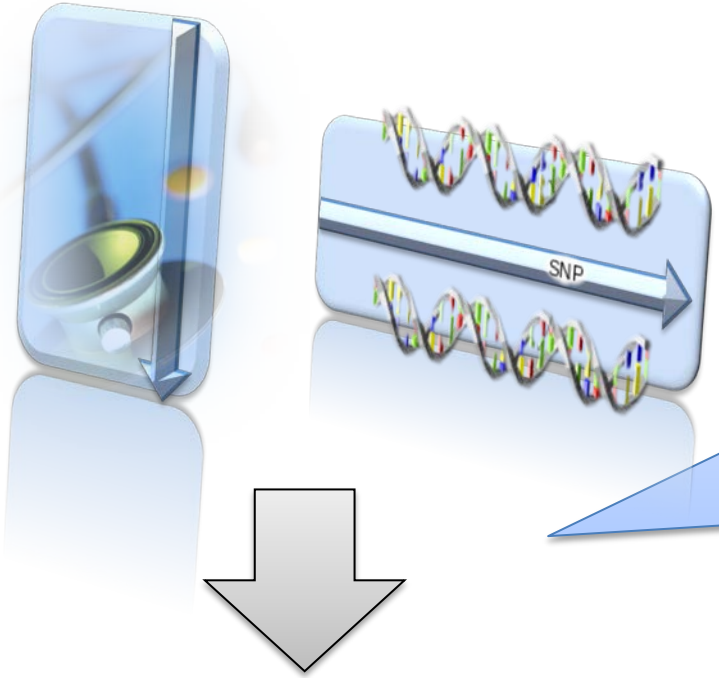
データの特徴

- 説明変数同士の相関が強いグループが存在
- 少サンプルなので，グループに含まれない変数でも相関が強くなる可能性がある

→ 偶然相関が強くなった変数の排除が難しい



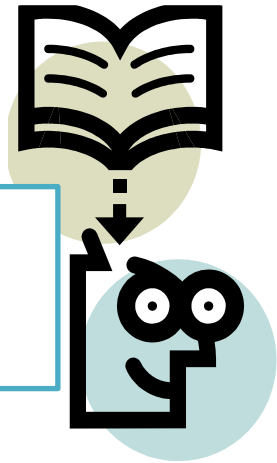




データ構造に依らず，関心のあるアウトカムに対する予測モデルに適切な解釈を与えるために，説明変数を予め選択してモデル構築をすることは重要

モデル構築

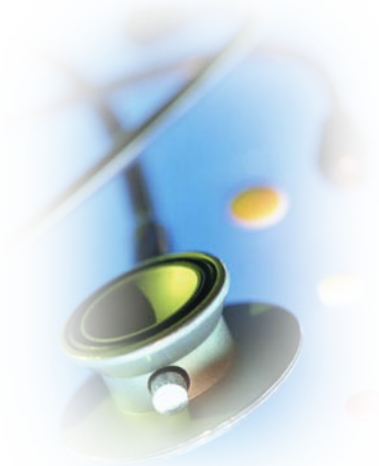
応答変数に影響を与える
重要な変数を選択したい！



■重要な変数の選択

① Random Forest

② Lasso / Elastic net



重要な変数の選択①

Random Forestの「重要度」をもとに変数選択する

Random Forest

- ランダムサンプリングされたトレーニングデータによって学習した多数の**決定木or回帰木**を使用する機械学習アルゴリズム

IMSTATのRANDOMWOODSステートメントでRandom Forestを縦長データ/横長データに実施し，変数の重要度を測定してその性能をみる

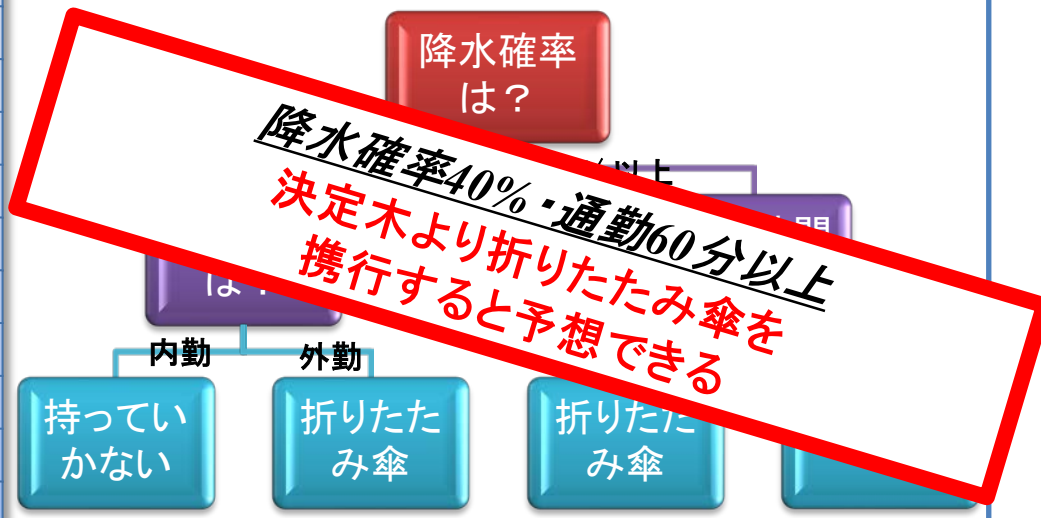
重要な変数の選択①

Random Forestの「重要度」をもとに変数選択する

決定木と回帰木

- 明示的な関数を用いず，一連の手順に沿ってデータを分岐させることで，予測や判別を行う手法を決定木（応答変数：カテゴリ変数）または回帰木（応答変数：連続変数）

傘の有無・種類	降水確率	勤務形態	通勤時間
持っていかない	0%	内勤	20分
持っていかない	30%	内勤	75分
折りたたみ傘	10%	外勤	55分
折りたたみ傘	30%	外勤	90分
折りたたみ傘	30%	外勤	30分
折りたたみ傘	50%	外勤	45分
折りたたみ傘	70%	内勤	80分
折りたたみ傘	70%	外勤	100分
長傘	40%	内勤	15分
長傘	70%	内勤	40分
長傘	80%	内勤	35分
長傘	100%	外勤	50分

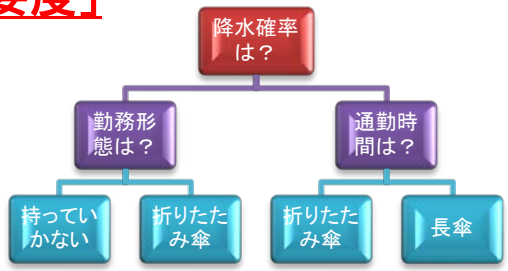


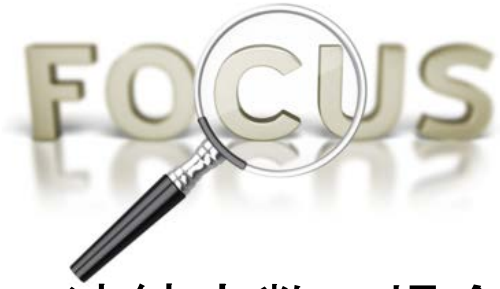
重要な変数の選択①

Random Forestの「重要度」をもとに変数選択する

決定木と回帰木

- 利点は？
  - ✓ 仮定が不要
    - ➢ 回帰分析：モデルや正規性の仮定の問題
    - ➢ データの素性がわからない状況で適用しやすい
  - ✓ アルゴリズムが容易であり、解釈しやすい
    - ➢ ノード内の不均一性の尺度 $i(t)$に基づき，分割し，決定木/回帰木を求める
    - ➢ この尺度$i(t)$を用いて，Random Forestで「重要度」を算出する



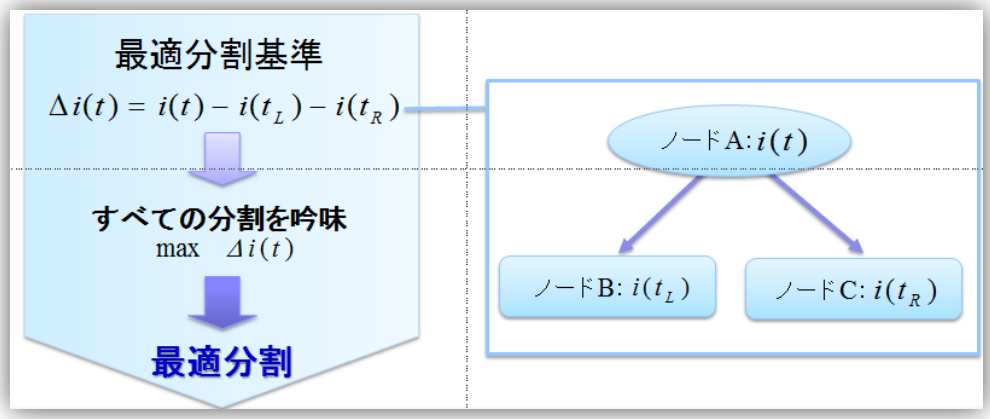


ノード内の不均一性の尺度 $i(t)$

- 連続変数の場合
  - 残差平方和: $i(t) = \sum_{i \in I}(y_i - \bar{y}(t))^2$
- カテゴリ変数の場合
  - ジニ不純度: $i(t) = 1 - \sum_k p_k(t)^2$
  - エントロピー不純度: $i(t) = -\sum_k p_k(t) \log p_k(t)$

※ $p_k(t)$はノードt でのクラスkの割合

新たに変数を木に追加したときに減少する統計量が，最大になる変数を選ぶ



重要な変数の選択①

Random Forestの「重要度」をもとに変数選択する

Random Forest

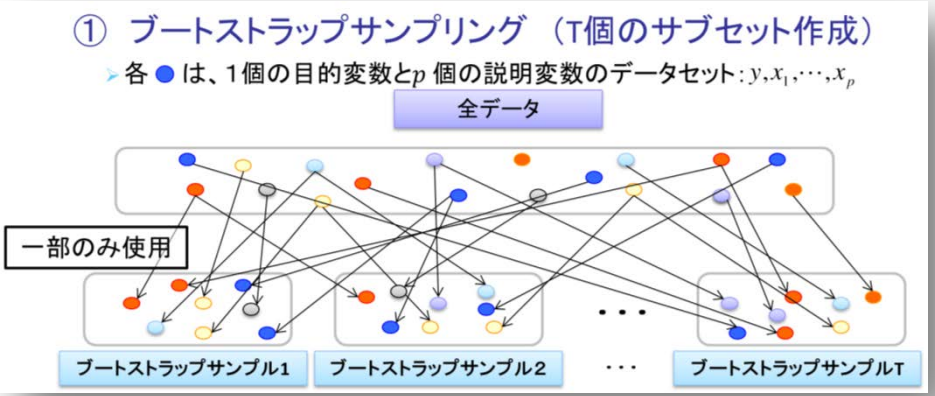
- Random Forestの発想は決定木 / 回帰木 ＋ブートストラップである
- Random Forestの出力は決定木/回帰木の出力結果の平均値などであり，木を作ることが目的ではないことに注意！

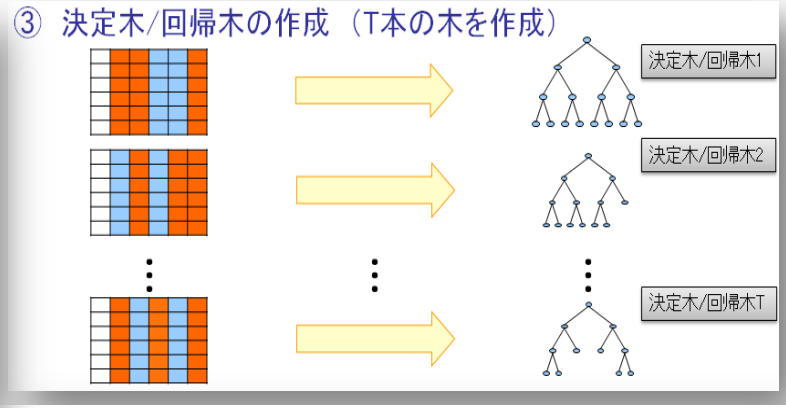
- 利点は？
  - ✓ 特徴量である「重要度」が学習とともに計算できる
  - ✓ 学習が早い
  - ✓ 過学習が起こりにくい
  - ✓ 適用範囲が広い
- 欠点は？
  - ✓ パラメータが多い
  - ✓ 学習データが少ないとうまく学習できない

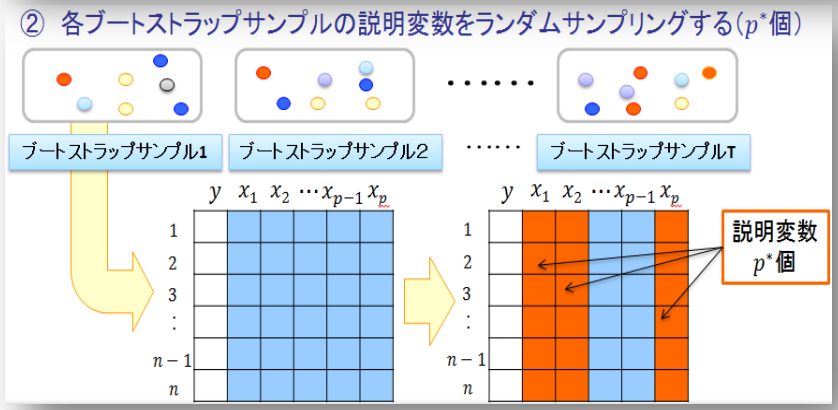
重要な変数の選択①

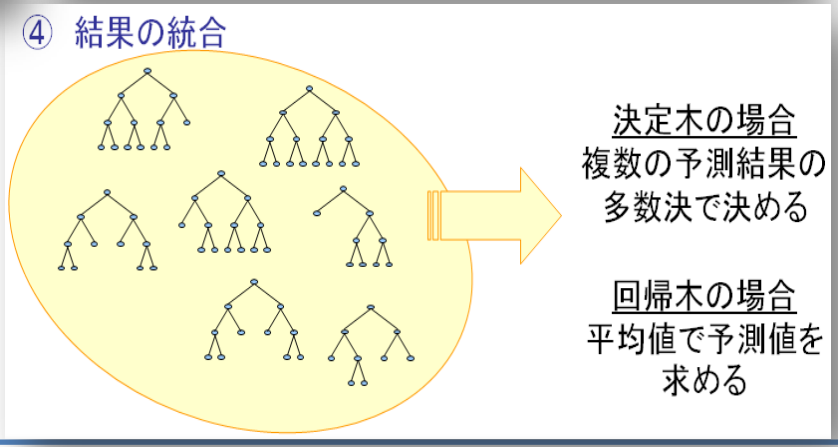
Random Forestの「重要度」をもとに変数選択する

Random Forestの手順









重要度

重要度：新たに変数を木に追加したときに減少する統計量を元に算出する

ノード内の不均一性の尺度 $i(t)$から
算出される$max\Delta i(t)$

- 算出手順
  1. ある変数が追加されるごとに**減少する統計量**を算出する
     - 同じ変数が何度か用いられる場合には、減少する統計量の合計
  2. 木が複数本ある場合，各木で求めた“**減少する統計量”の平均**が変数の**重要度**

Sample Code

```
proc imstat DATA=LASRLIB.Dataset;
        RANDOMWOODS OUTCOME/                                    /* 応答変数 */
                INPUT=(COL1 COL2 COL3 COL4 COL5 COL6)           /* 説明変数 */
                NOMINAL=(COL3 COL5 COL6)                        /* カテゴリー変数 */
                M=4                                             /* 説明変数のサンプリング数*/
                LEAFSIZE=5                                      /* 葉の枚数 */
                MAXBRANCH=2                                     /* 枝の最大数 */
                MAXLEVEL=10                                     /* 木の深さ */
                BOOTSTRAP=0.8                                   /* Default:1-exp(-1) */
                NTREE=3000;                                     /* 決定木の本数 */
run;
quit;
```

重要な変数の選択②

Lasso / Elastic net で変数選択する

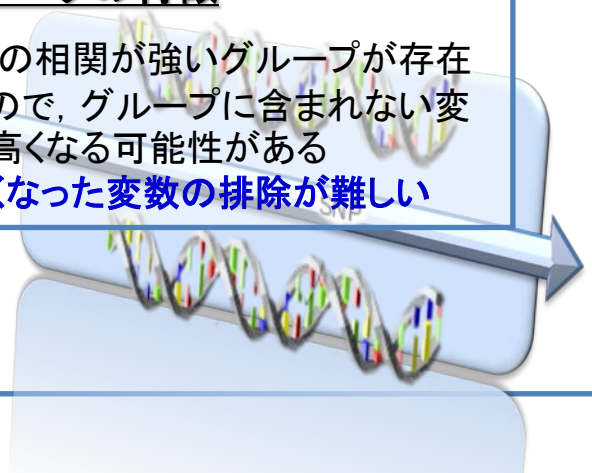
遺伝子データの変数選択を考える

- 超多変数であり，ノイズとなる変数を多く含むので，**モデルの予測精度を高める変数選択を行いたい**
- 単に「変数」に注目するのではなく，**説明変数同士の相関が強い「グループ」に注目して変数選択すること**がデータの特徴上，重要

データの特徴

- 説明変数同士の相関が強いグループが存在
- 少サンプルなので，グループに含まれない変数でも相関が高くなる可能性がある

→ **偶然相関が高くなった変数の排除が難しい**



重要な変数の識別をする**Lasso**，加えて，相関性を考慮することが可能な**Elastic net**に注目した

重要な変数の選択②

Lasso / Elastic net で変数選択する

- 一般的な線形回帰モデルで，p個の目的変数$x_1, \cdots, x_p$が与えられ，応答変数が以下のように予測されたとする.

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \cdots + x_p\hat{\beta}_p = X\hat{\beta}$$

最小二乗推定量(OLS推定量)

$$\hat{\beta} = \mathrm{argmin}_{\beta}\{(y - X\beta)^2\}$$

- 推定したモデルの評価基準
  - モデルの予測精度
  - モデルの解釈

2つの側面に対して，OLS推定量は優れていない

→この改善のために***「罰則」***の考え方がある.

重要な変数の選択②

Lasso / Elastic net で変数選択する

「罰則」の例：リッジ回帰

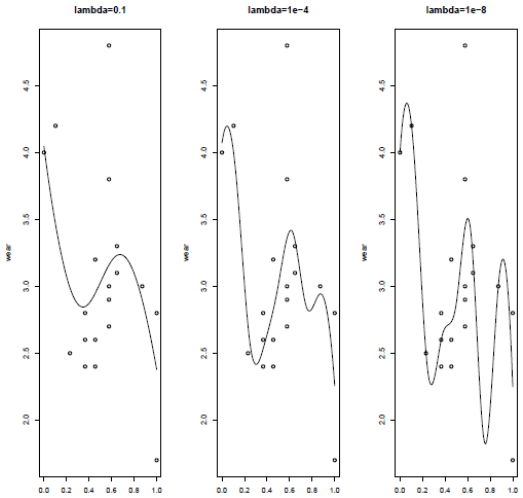
- 一般的な線形回帰モデルの推定量

罰則項！

$$\hat{\beta} = \mathrm{argmin}_{\beta}\{(y - X\beta)^2 + \lambda|\beta|^2\}$$

チューニング
パラメータ

- モデルの予測精度は向上！



課題

モデルに取り込む変数を選択できず，
すべての変数をモデルに組み込む…

これに対して考案されたのが**Lasso**

重要な変数の選択②

Lasso / Elastic net で変数選択する

Lasso

- 一般的な線形回帰モデルの推定量

$$\hat{\beta} = \mathrm{argmin}_{\beta}(y - X\beta)^2 + \lambda|\beta|\}$$

罰則項！

- 特徴
  - 連続的に縮小推定を行い，かつ変数選択が可能
  - 予測の性能はその他（リッジ回帰など）の罰則より優れる
- Lassoが機能するために必要な制約
  - ✓ p>n の場合，Lassoでは高々n個の説明変数しか選択できない
  - ✓ 説明変数同士の相関が強い場合，それらの変数をグループと呼ぶとすると，変数選択する際にそのグループの中から1つの変数のみをモデルに組み込み，それ以外を無視する傾向がある

遺伝子データに不向きな制約

重要な変数の選択②

Lasso / Elastic net で変数選択する

Elastic Net

- Lassoの特性である，変数選択と連続的な縮小推定に加えて，変数間の相関によるグループ効果を考慮することができる

$$\hat{\beta} = (1 + \lambda_2) \times \mathrm{argmin}_{\beta}\{(y - X\beta)^2 + \lambda_2 \|\beta\|^2 + \lambda_1 |\beta|_1\}$$

グループ効果を表すパラメータ
(チューニングパラメータ)

$$\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2 , |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

Sample Code

Lasso

```
proc glmselect data=work.Data plots=all ;
        model OUTCOME=COL1-COL10
          /  selection=lasso(steps=1000 choose=AIC) ;
run ;
```

停止条件

モデル選択の基準

Elastic net

```
proc glmselect data=work.Data plots(stepaxis=normb)=coefficients ;
        model OUTCOME=COL1-COL10
          / selection=elasticnet(steps=1000 L2=0.1  choose=AIC) ;
run ;
```

チューニングパラメータ

シミュレーション

当日公開

まとめ

当日公開

課題

当日公開

参考文献

- Robert Tibshirani (2011). Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society, Series B **73**(**3**), 273-282.

- Hui Zou and Trevor Hastie(2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B **67**(**2**), 301-320.

- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). The Elements of Statistical Learning.

- Simon N. Wood(2006). Generalized Additive Models: an introduction with R. Chapman & Hall/CRC.

End of Slide