GAMとその周辺

○田中 祐輔* 伊庭 克拓** 辻谷 将明***
*大阪電気通信大学大学院 工学研究科 情報工学専攻
**東京CRO株式会社 DM統計本部 統計解析部
***大阪電気通信大学 情報通信工学部 情報工学科

GAM and Its Related Problems

Yusuke Tanaka* Katsuhiro Iba** Masaaki Tsujitani***
*Osaka Electro-Communication University, Graduate School of Engineering
**TOKYO CRO, Inc.
***Osaka Electro-Communication University

SASユーザー総会 アカデミア/テクノロジー&ソリューション セッション 2009

発表内容

- はじめに
- 一般化加法モデル(GAM)
  - 平滑化スプライン
  - 薄板平滑化スプライン
- 適用例
  - 脊柱後湾症データ
  - 糖尿病網膜症データ
- シミュレーション実験
- まとめ

はじめに

- 非線形性をもつ統計的多変量解析の発展
  - 薄板平滑化スプライン ⇒PROC TPSPLINE
  - 局所回帰 ⇒PROC LOESS
  - 一般化加法モデル ⇒PROC GAM
    (Generalized Addtive Models : GAM)

平滑化スプラインによるロジスティック判別の紹介と適用

ロジット変換

- 2値応答：Y

$$\Pr(Y=1)=\pi,\ \Pr(Y=0)=1-\pi$$

ロジット変換$(-\infty,+\infty)$

$$\ln\left(\frac{\pi}{1-\pi}\right)=c_0+c_1x_1+\cdots+c_kx_k$$

$Y \sim Bin(1,\pi)$ ：ベルヌーイ分布

$$E[Y]=\pi, V[Y]=\pi(1-\pi)$$

ロジスティック判別

- ロジット逆変換

$$\pi = \frac{1}{1+\exp\left\{-\left(c_0 + c_1 x_1 + \cdots + c_k x_k\right)\right\}}$$

- 判別方法

$$\text{2値応答 } y = \begin{cases} 1: \text{第1群} \\ 0: \text{第2群} \end{cases}$$

観測値が第1群に属する$(y=1)$確率：π

観測値が第2群に属する$(y=0)$確率：$1-\pi$

4

ロジスティック判別

- ロジット逆変換

$$\pi = \frac{1}{1+\exp\left\{-\left(c_0 + c_1 x_1 + \cdots + c_k x_k\right)\right\}}$$

- 判別方法

$$\pi \geq 0.5 \Rightarrow \text{第1群} \left(y=1\right)$$

$$\pi < 0.5 \Rightarrow \text{第2群} \left(y=0\right)$$

5

一般化加法モデル
(Generalized Addtive Models : GAM)

- 応答変数に指数分布族を仮定
- 平滑化関数$s(x)$の加法モデル

GAM :

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + s(x_1) + \cdots + s(x_k)$$

ロジット変換

平滑化スプライン

ペナルティ付き残差平方和

平滑化パラメータ

$$\sum_{i=1}^{n}\left[y_i - s(x_i)\right]^2 + \lambda \int \left\{ s''(x) \right\}^2 dx$$

$s(x)$の曲率

小さいほどモデルの当てはまりは良い

小さいほど滑らかな曲線（曲げ弾性エネルギー）

⇒ 最小にするスプライン関数

平滑化スプライン

$$y = s(x) = c_0 + c_1 x + \frac{1}{12}\sum_{d=1}^{n} \theta_d \left| x - x_d \right|^3$$

自由度と平滑化パラメータ

- ハット行列

  応答$\boldsymbol{y}$の予測値: $\hat{\boldsymbol{y}} = \boldsymbol{H}_{\lambda}\boldsymbol{y}$

- モデルの自由度
  - 実効自由度(=有効パラメータ数)

    $df = tr\left(\boldsymbol{H}_{\lambda}\right)$
  - 平滑化パラメータλの決定

    ⇔ 自由度dfの決定

クロス・バリデーション(CV)

- 1例消去(leaving-one-out)CV法

初期標本: $\boldsymbol{X} = \left\{\boldsymbol{X}^{\langle 1\rangle}, \boldsymbol{X}^{\langle 2\rangle}, \ldots, \boldsymbol{X}^{\langle n\rangle}\right\}$

⇓ d番目を除去: $\boldsymbol{X}^{\langle d\rangle} = \left\{x_1^{\langle d\rangle}, \ldots x_I^{\langle d\rangle}; y^{\langle d\rangle}\right\}$

$$\boldsymbol{X}_{[d]} = \left\{\boldsymbol{X}^{\langle 1\rangle}, \boldsymbol{X}^{\langle 2\rangle}, \ldots, \boldsymbol{X}^{\langle d-1\rangle}, \boldsymbol{X}^{\langle d+1\rangle}, \ldots, \boldsymbol{X}^{\langle n\rangle}\right\}$$

$$CV = \frac{1}{n}\sum_{d=1}^{n}\left\{y^{\langle d\rangle} - \hat{y}_{[d]}^{\langle d\rangle}\right\}$$

$\boldsymbol{X}_{[d]}$で構築したモデルの, $\boldsymbol{X}^{\langle d\rangle}$の予測値

⇒ 最小にする平滑化パラメータλが最適

クロス・バリデーション(CV)

- 1例消去(leaving-one-out)CV法

初期標本: $\boldsymbol{X} = \left\{ \boldsymbol{X}^{\langle 1 \rangle}, \boldsymbol{X}^{\langle 2 \rangle}, \ldots, \boldsymbol{X}^{\langle n \rangle} \right\}$

d番目を除去: $\boldsymbol{X}^{\langle d \rangle} = \left\{ x_1^{\langle d \rangle}, \ldots x_I^{\langle d \rangle}; y^{\langle d \rangle} \right\}$

$$\boldsymbol{X}_{[d]} = \left\{ \boldsymbol{X}^{\langle 1 \rangle}, \boldsymbol{X}^{\langle 2 \rangle}, \ldots, \boldsymbol{X}^{\langle d-1 \rangle}, \boldsymbol{X}^{\langle d+1 \rangle}, \ldots, \boldsymbol{X}^{\langle n \rangle} \right\}$$

$$CV = -2 \sum_{d=1}^{n} \left\{ y^{\langle d \rangle} \ln \hat{\pi}_{[d]}^{\langle d \rangle} + \left(1 - y^{\langle d \rangle}\right) \ln \left(1 - \hat{\pi}_{[d]}^{\langle d \rangle}\right) \right\}$$

$\boldsymbol{X}_{[d]}$で構築したモデルの, $\boldsymbol{X}^{\langle d \rangle}$の予測値

最小にする平滑化パラメータλが最適

適用例1

- 脊柱後湾症データ

応答変数

y : 術後の症状の有無

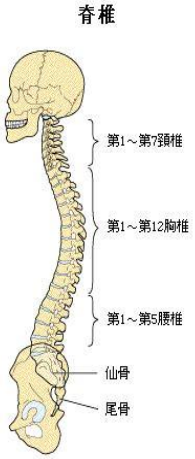
1 : 症状が残る
0 : 症状なし

予測変数

x_1 : *age* 手術時の月齢

x_2 : *start* 何番目の脊椎から先を手術したか

x_3 : *num* 手術した脊椎の個数



PROC GAMの文法

```
proc gam data=kyphosis( where=(id^=&d) ) ;
model kyp = param( start num ) spline( age , df=&df )
   ①          ②                  / link = logit dist = binomial ;
run;
```

① model文の左辺は応答変数，右辺はモデル式

② 線形項 param(*variable* ・・・)

LOESS loess(*variable* ,df=*number*)

平滑化スプライン spline(*variable* ,df=*number*)

薄板平滑化スプライン spline2(*variable1*,*variable2* ,df=*number*)

PROC GAMの文法

```
proc gam data=kyphosis( where=(id^=&d) ) ;   ③
model kyp = param( start num ) spline( age , df=&df )
   ①          ②                  / link = logit dist = binomial ;
run;                                                          ④
```

③ df値を指定しないとき、デフォルトdf=4

④ ③の場合，/の後にmethod=gcvを指定するとGCV規準に基づいて最適な平滑化パラメータの探索

自由度の最適選択

- PROC GAMにおける最適自由度の決定
  - グリッド検索による平滑化パラメータの探索
  - method=gcvを指定した場合，モデルが収束しないことが起こり得る
- 提案法

主効果スプラインモデル：

$$\ln\left(\frac{\pi}{1-\pi}\right) = s(age) + start + num$$

  - クロス・バリデーション(CV)規準に基づいて，各項の最適な自由度を探索

自由度の最適選択

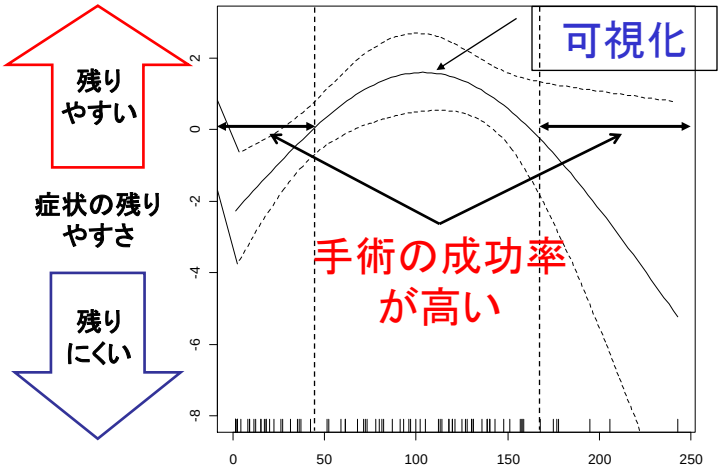
- 最適選択の結果

	CV	逸脱度	*df*(*age*)	誤判別率
1例消去CV法	65.92	56.39	2	0.133

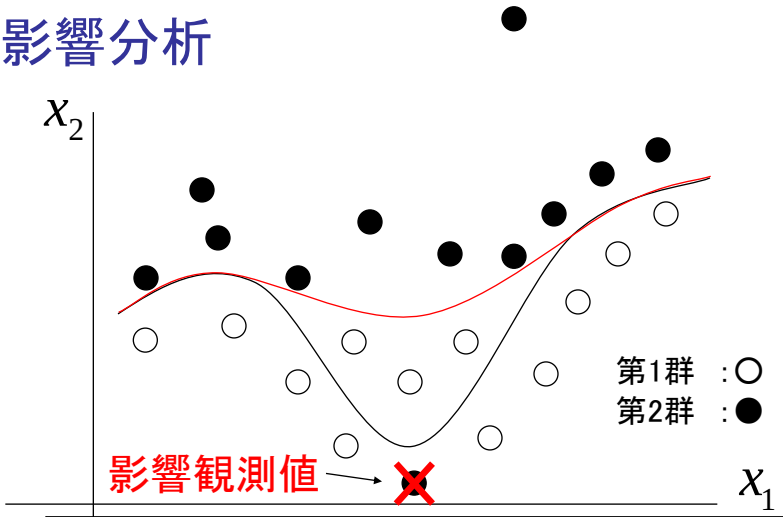
- 他の判別手法との比較

誤判別率＝
- 0.133 ：GAM
- 0.205 ：ロジスティック判別
- 0.193 ：線形判別

平滑化スプラインの予測値

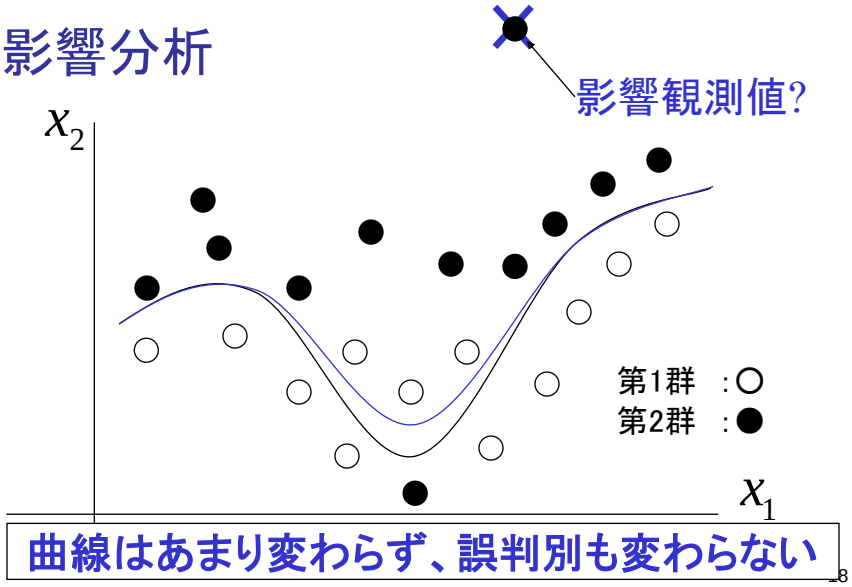


影響分析



曲線は簡単になり、誤判別は少なくなる

影響分析



曲線はあまり変わらず、誤判別も変わらない

DIFDEV

$$\Delta Dev_{[d]} = Dev - Dev_{[d]} \geq 0$$

Dev : すべての個体を用いたときの逸脱度

$Dev_{[d]}$: d番目の個体を取り除いたときの逸脱度

- **検定の方法**

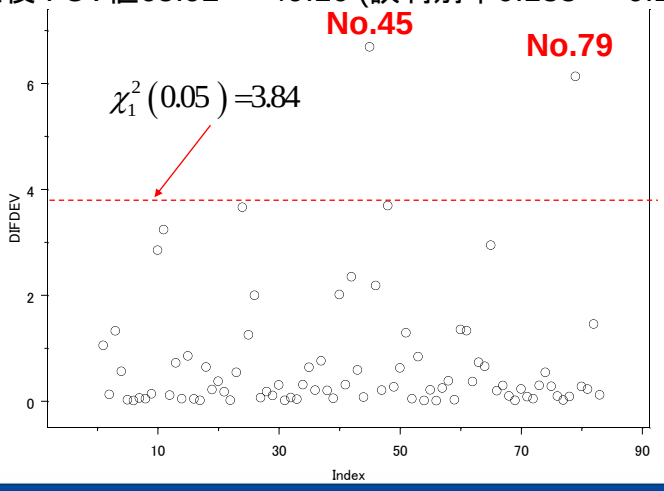
自由度1のχ^2分布に基づき、有意水準0.05で検定

$$\Delta Dev_{[d]} \geq \chi^2(0.05) = 3.84$$

ならば、d番目の個体を除去

DIFDEV

除去後 : CV値65.92 ⇒ 49.10 (誤判別率0.133 ⇒ 0.111)



適用例2

- 糖尿病網膜症
  - 網膜内の血管障害に伴う病気
  - 669例の糖尿病患者における網膜症の進行の有無を調査

y : ret　糖尿病網膜症の進行　{ 1 : 進行あり / 0 : 進行なし

x_1 : dur　糖尿病の罹病期間 ; year

x_2 : gly　糖化ヘモグロビン ; %

x_3 : bmi　肥満度 ; kg/m^2

交互作用

薄板平滑化スプライン

- 罹病期間とBMIとには交互作用あり

➡ 薄板平滑化スプラインの利用

交互作用スプラインモデル :

$$\ln\left(\frac{\pi}{1-\pi}\right) = gly + s(dur, bmi)$$

```
proc gam data=WESDR( where=(id^=&d) ) ;
model ret = param( gly ) spline2( dur , bmi , df=&df )
                                    / link = logit dist = binomial ;
run;
```

22

自由度の最適選択

- 最適選択の結果

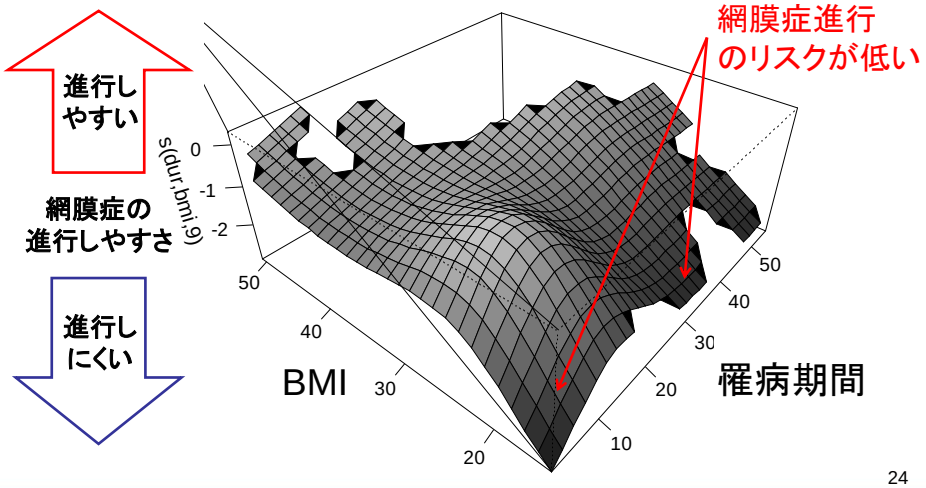
	CV	逸脱度	*df*	誤判別率
1例消去CV法	763.49	743.47	9	0.272
SAS自動選択	764.25	744.81	8.3	0.274

- 他の判別手法との比較

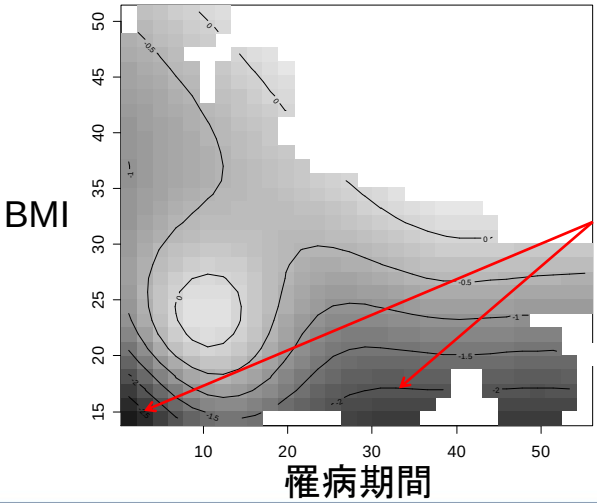
誤判別率＝
- 0.272 : GAM
- 0.306 : ニューラルネット
- 0.293 : ロジスティック判別
- 0.318 : 線形判別

23

薄板平滑化スプラインの予測値



薄板平滑化スプラインの予測値



シミュレーション・データの生成

手順1：予測変数として，一様乱数

$$x_1, x_2 \in [-1, +1]$$

を生成

手順2： x_1, x_2 を

$$f(x_1, x_2) = \sin(2 \times 3.14 \times x_1) + \alpha \times x_1 \times x_2 + \sin(2 \times 3.14 \times x_2)$$

へ代入

交互作用項

交互作用の調整用

$\alpha = 1,5$

シミュレーション・データの生成

手順3： p を求める

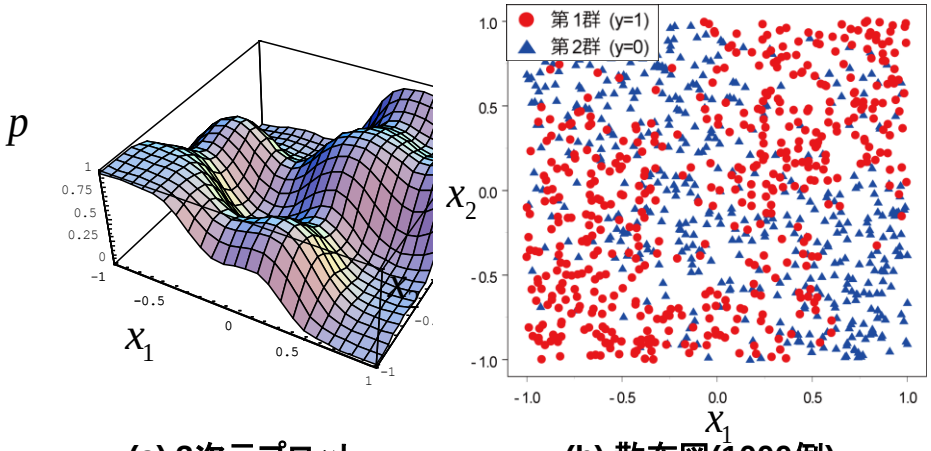
$$p = \frac{1}{1 + \exp\{-f(x_1, x_2)\}}, 0 \le p \le 1$$

手順4：手順3の p を用いて，ベルヌーイ乱数

$$y \sim Bin(1, p) \Rightarrow y = \begin{cases} 1(\text{第1群}) \\ 0(\text{第2群}) \end{cases}$$

を発生

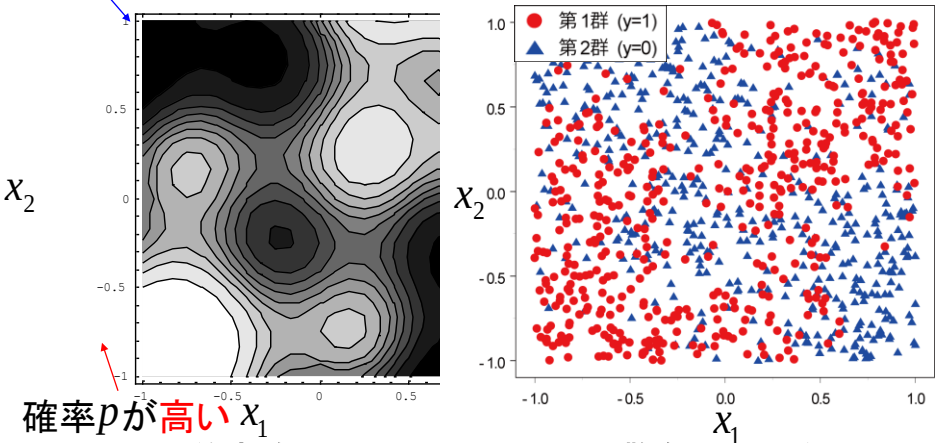
シミュレーション・データ(α=5)



(a) 3次元プロット　(b) 散布図(1000例)

シミュレーション・データ(α=5)

確率pが低い



確率pが高い

(a) 等高線図　(b) 散布図(1000例)

シミュレーションの設定

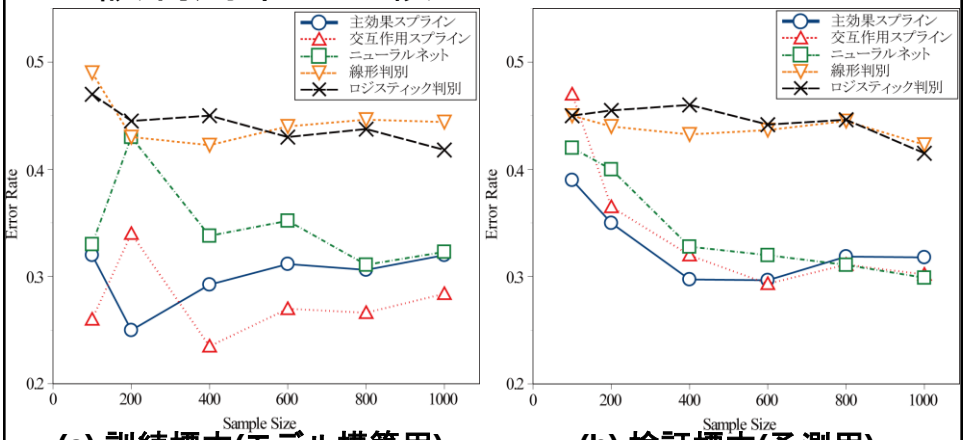
- 観測値の構成
  - 予測変数 x_1, x_2 、2値応答 y
- シミュレーション・データ
  - 訓練標本(モデル構築用)
  - 検証標本(未知のデータ)
- サンプル・サイズ
  - 100, 200, 400, 600, 800, 1000例

シミュレーションの設定

- 判別モデル
  - 主効果スプライン $\ln\left(\frac{p}{1-p}\right) = s(x_1) + s(x_2)$
  - 交互作用(薄板)スプライン $\ln\left(\frac{p}{1-p}\right) = s(x_1, x_2)$
  - ニューラルネットワーク
    - 隠れユニット数 : 100,200例=5個,
      400,600,800,1000例=6個
  - 線形判別
  - ロジスティック判別

性能評価(α=1)

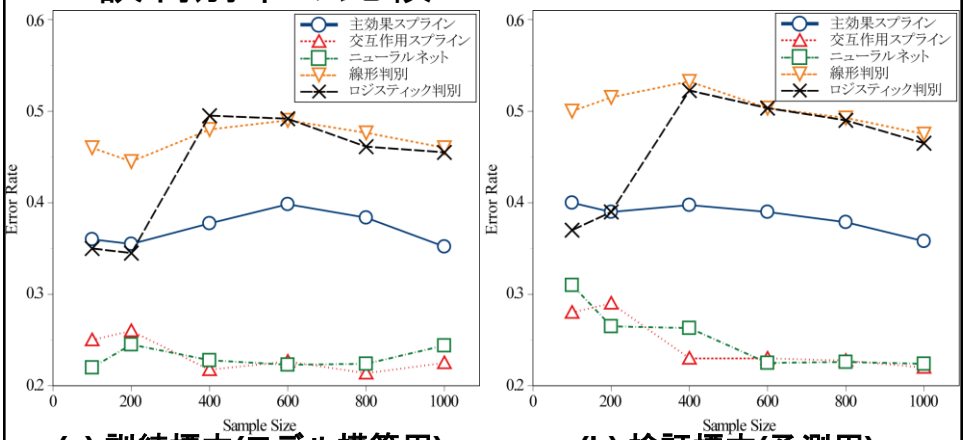
- 誤判別率の比較



(a) 訓練標本(モデル構築用)　(b) 検証標本(予測用)

性能評価(α=5)

- 誤判別率の比較



(a) 訓練標本(モデル構築用)　(b) 検証標本(予測用)

まとめ

- 最適な自由度の決定が必要
- 共変量の非線形性の可視化
  - ➡ ニューラルネットでは不可能
- 交互作用効果の考慮