

ニューラルネットにおける 影響分析に関する研究

大阪電気通信大学院 工学研究科 情報工学専攻
竹植 久勝

大阪電気通信大学 情報通信工学部 情報工学科
辻谷 将明

研究背景

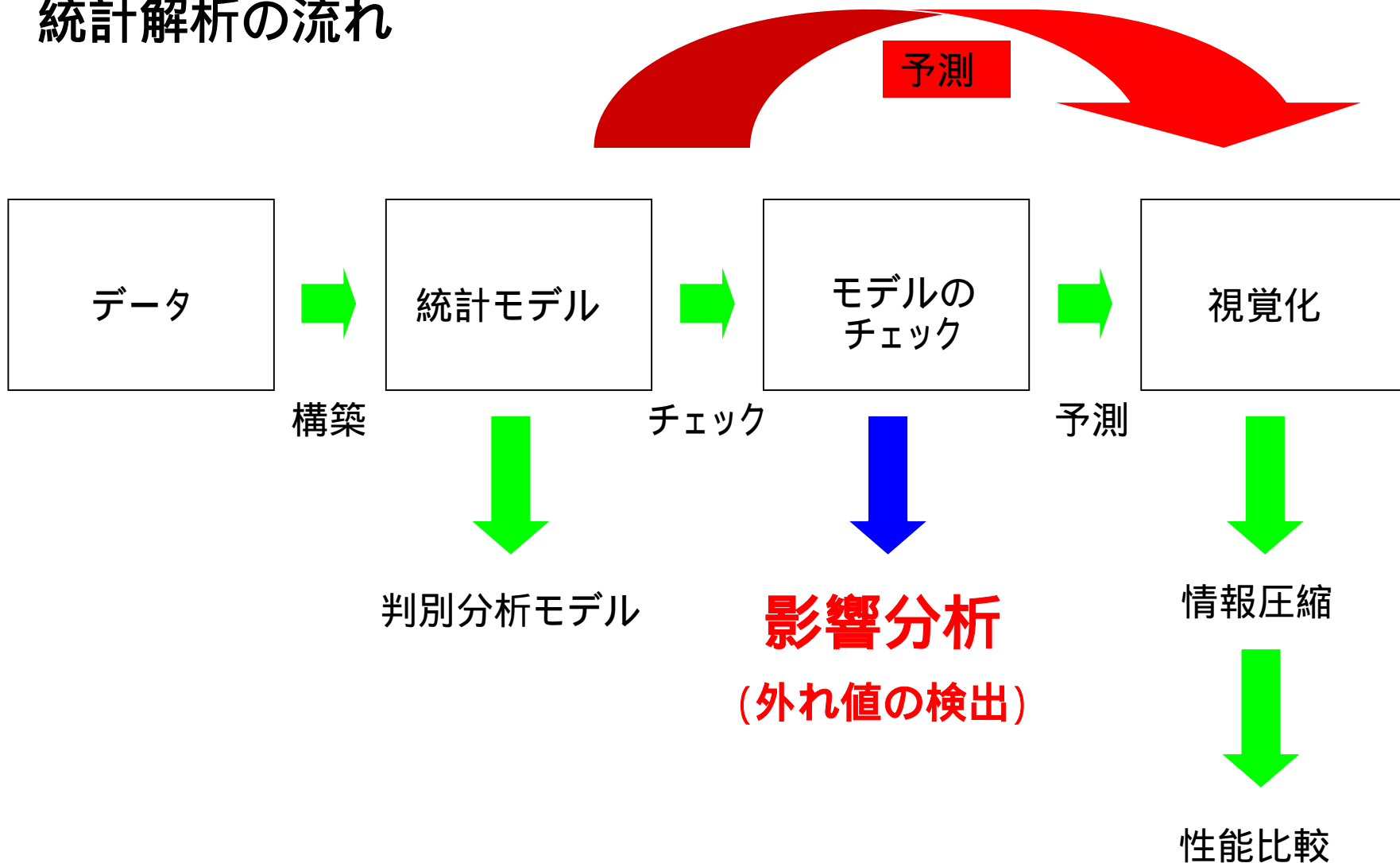
- ・近年、情報化社会において、得られたデータから有用な情報を取り出せるかということが注目を浴びている。
- ・データの情報化の手法として様々なものが提案されている。

1. はじめに

影響分析とは・・・

データの中に解析結果に影響を与える観測値の検出.

統計解析の流れ



首都圏マンションの各駅周辺のデータ

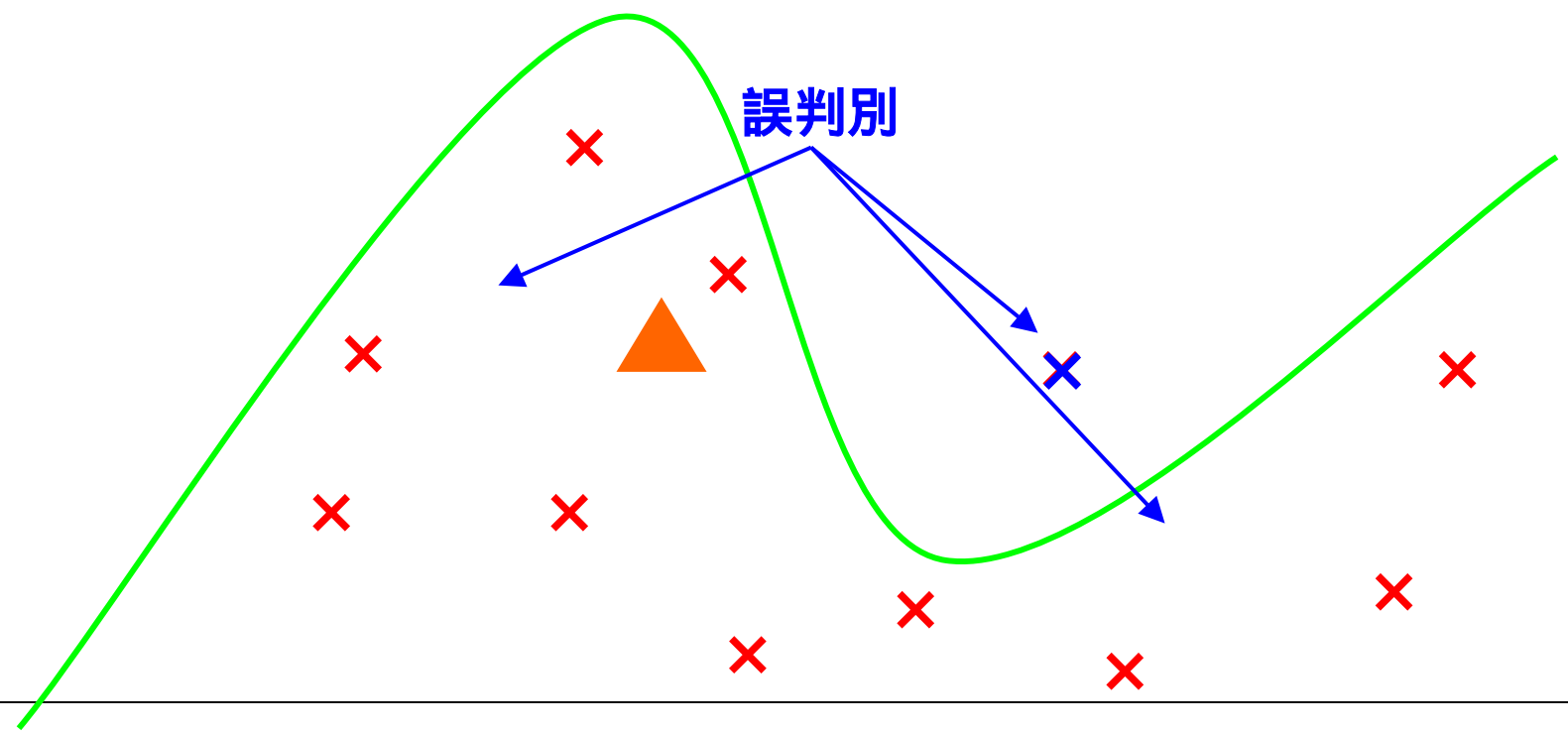
ID	駅名	分譲平均 (万円)	賃貸平均 (万円)	利回り (%)	駅力 (点)	ランク
1	新宿	5850	25.2	5.2	88	(極めて高い)
・	・	・	・	・	・	
・	・	・	・	・	・	
53	横浜	3740	18.4	5.9	100	
54	浅草橋	4310	19.3	5.4	84	(非常に高い)
・	・	・	・	・	・	
・	・	・	・	・	・	
173	関内	4080	18.0	5.3	68	
174	南柏	2890	9.9	4.1	80	(高い)
・	・	・	・	・	・	
・	・	・	・	・	・	
501	大宮	3500	13.5	4.6	72	
502	北柏	2810	9.6	4.1	80	(現状維持)
・	・	・	・	・	・	
・	・	・	・	・	・	
806	千葉	3120	11.1	4.3	76	

x_2

判別分析

第1群	:
第2群	: ×

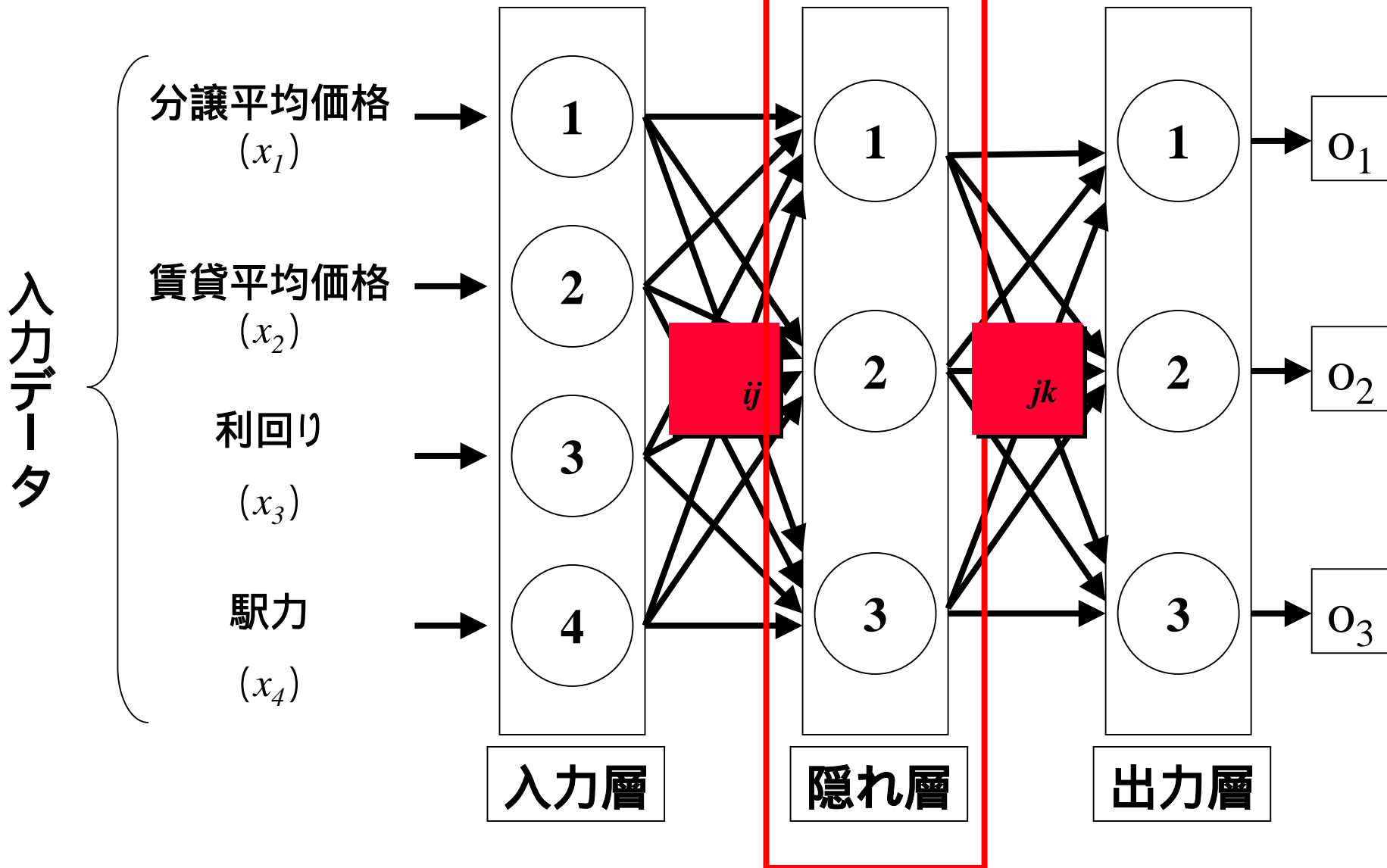
グループのわからないサンプル



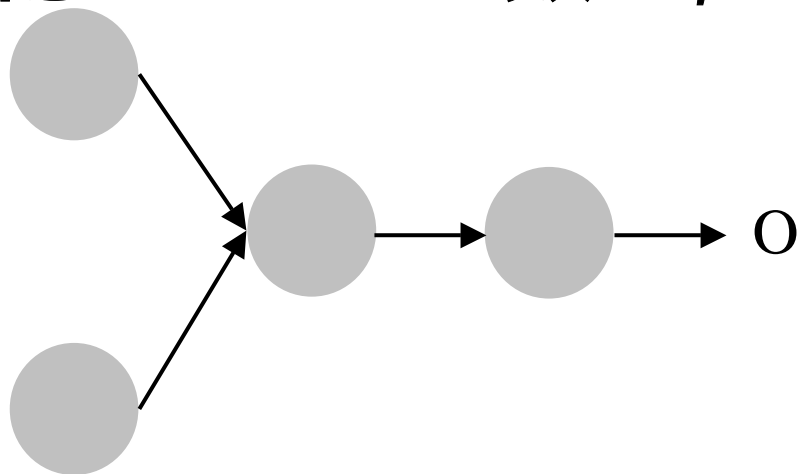
誤判別

 x_1

2. ニューロ判別モデル

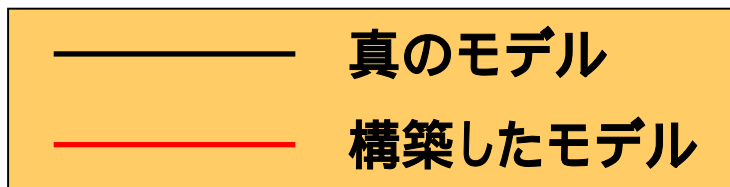


隠れユニット数と、モデルの性能の関係

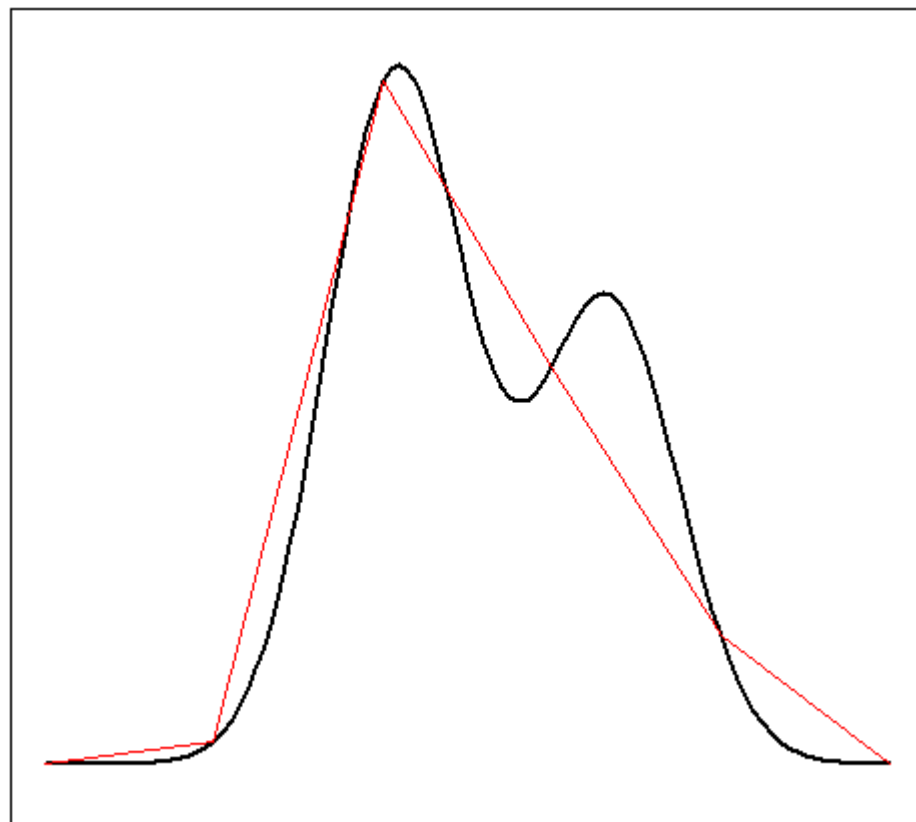


[隠れユニット数が少なすぎる場合]

うまく近似できない



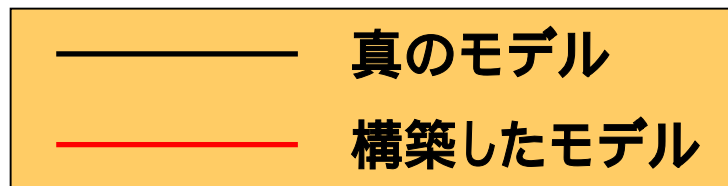
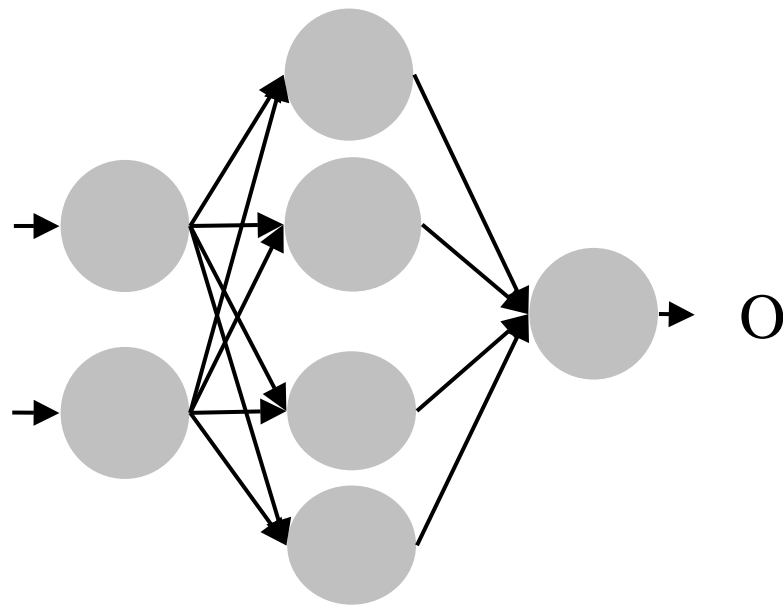
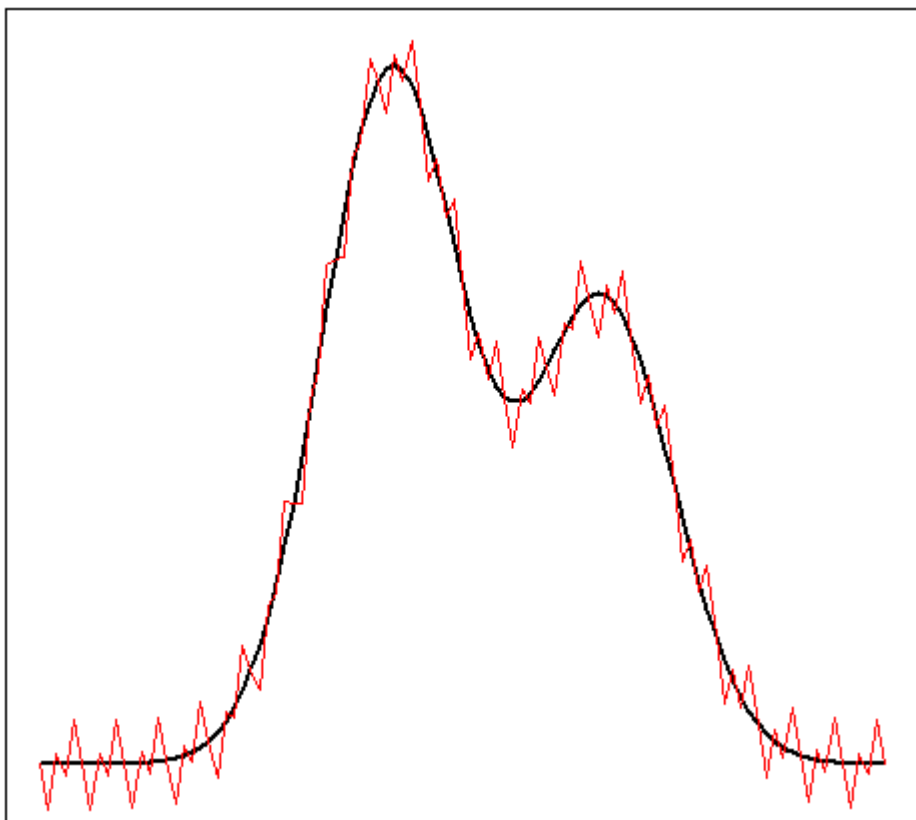
データ	当てはまり
訓練(現在のデータ)	悪い



隠れユニット数と、モデルの性能の関係

[隠れユニット数が多すぎる場合]

構築に使ったデータの誤差
まで近似

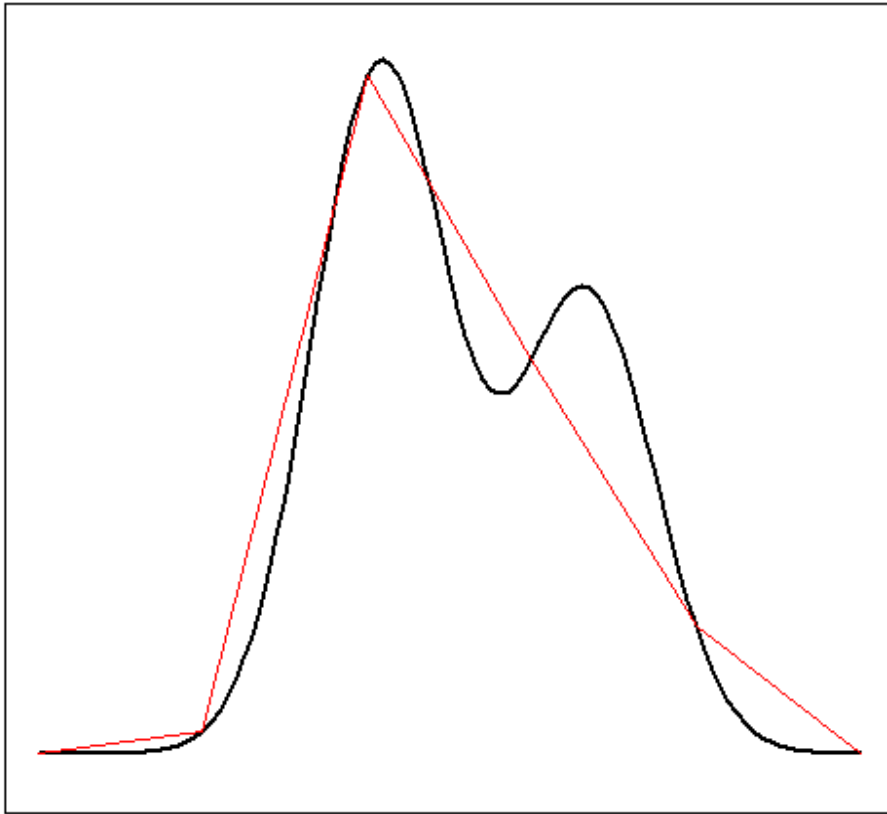


データ	当てはまり
訓練(現在のデータ)	良い(様に見える)

隠れユニット数とモデルの性能との関係

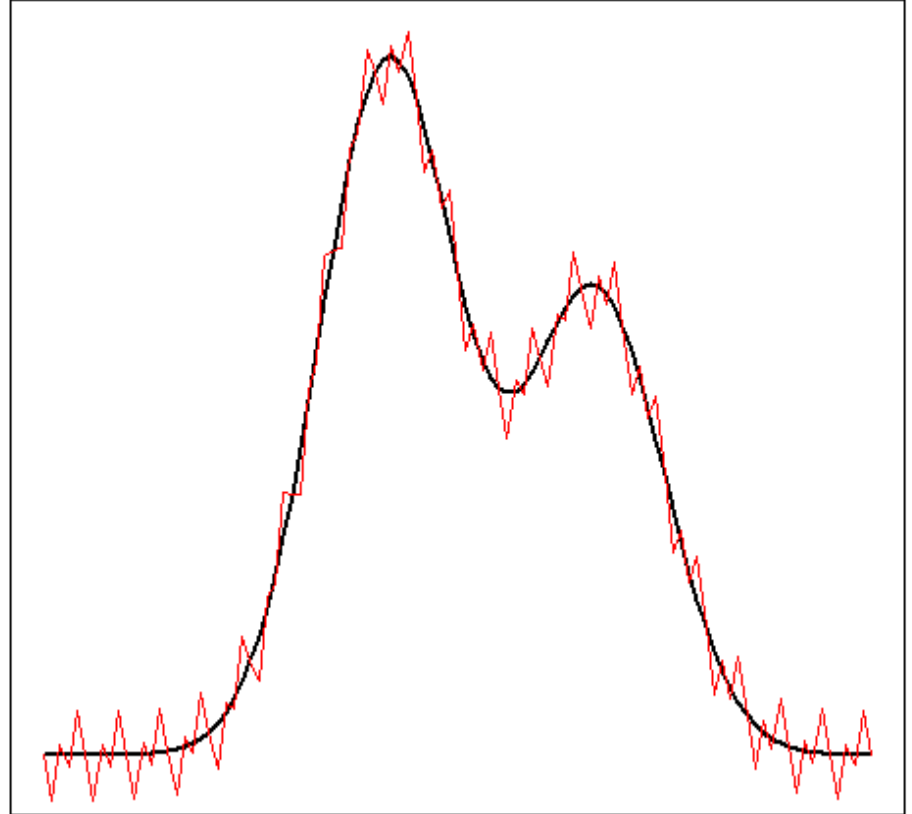
[隠れユニット数が少なすぎる場合]

うまく近似できない



[隠れユニット数が多すぎる場合]

構築に使ったデータの誤差まで近似



ほど良い近似が良いモデル

隠れユニット数の決定

EIC (***E***xtended ***I***nformation ***C***riterion)

$$EIC = -2 \ln L(X; \hat{\theta}(X)) + 2C^*$$

当てはまりの良さ

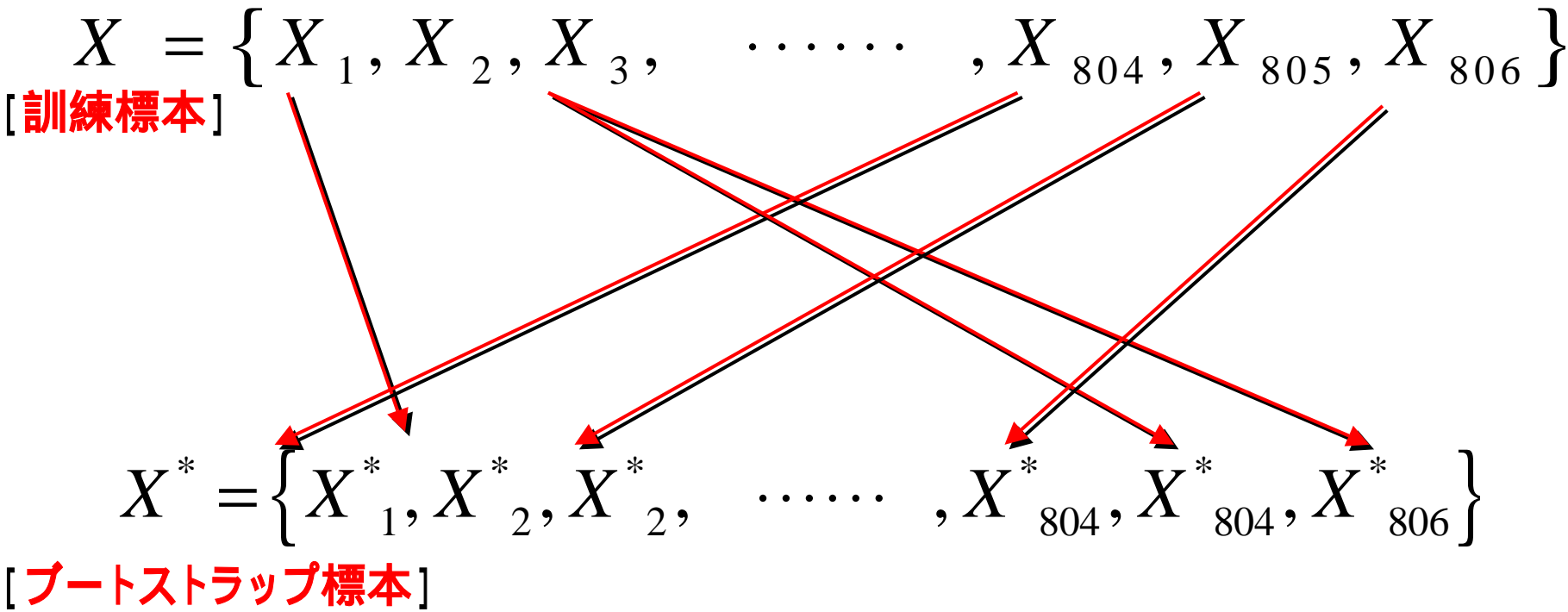
バイアスの
ブートストラップ推定

C^* = ブートストラップ・バイアス推定量

(モデルのバイアスを数値的に算出)

ブートストラップ法による、バイアスの推定法

手順1 訓練標本から、リサンプリングにより**ブートストラップ標本**を生成.



ブートストラップ標本より,ニューロ判別モデルの構築.

手順2 対数尤度の算出.

$$\ln L(X^*; \hat{\theta}(X^*))$$

~ ブートストラップ標本より構築したニューラルネットワークの対数尤度.

$$\ln L(X; \hat{\theta}(X^*))$$

~ ブートストラップ標本より構築したニューラルネットワークに元の訓練標本を当てはめたときの対数尤度.

手順3 手順1, 2を必要回数繰り返す. (本実験ではB=400回)

手順4 手順3で得られた値より、バイアスのブートストラップ推定.

$$\text{バイアスの平均 } C^* \approx \frac{1}{B} \sum_{b=1}^B \left\{ \ln L(X_b^*; \hat{\theta}(X_b^*)) - \ln L(X; \hat{\theta}(X_b^*)) \right\}$$

ズレ(バイアス)

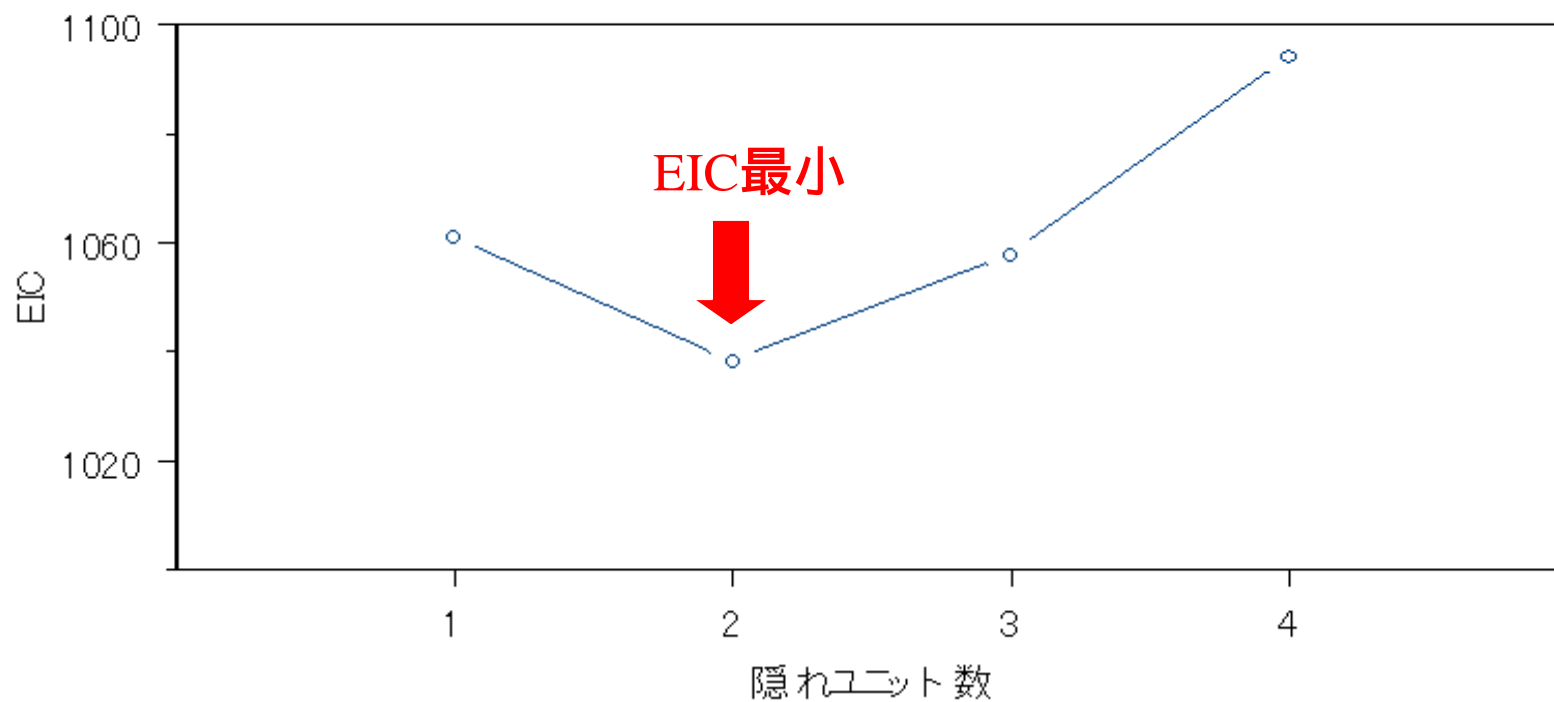
ブートストラップ法に基づく, 情報量規準.

$$EIC = -2 \ln L(X; \hat{\theta}(X)) + 2C^* \quad \text{最小化}$$

もとの対数尤度

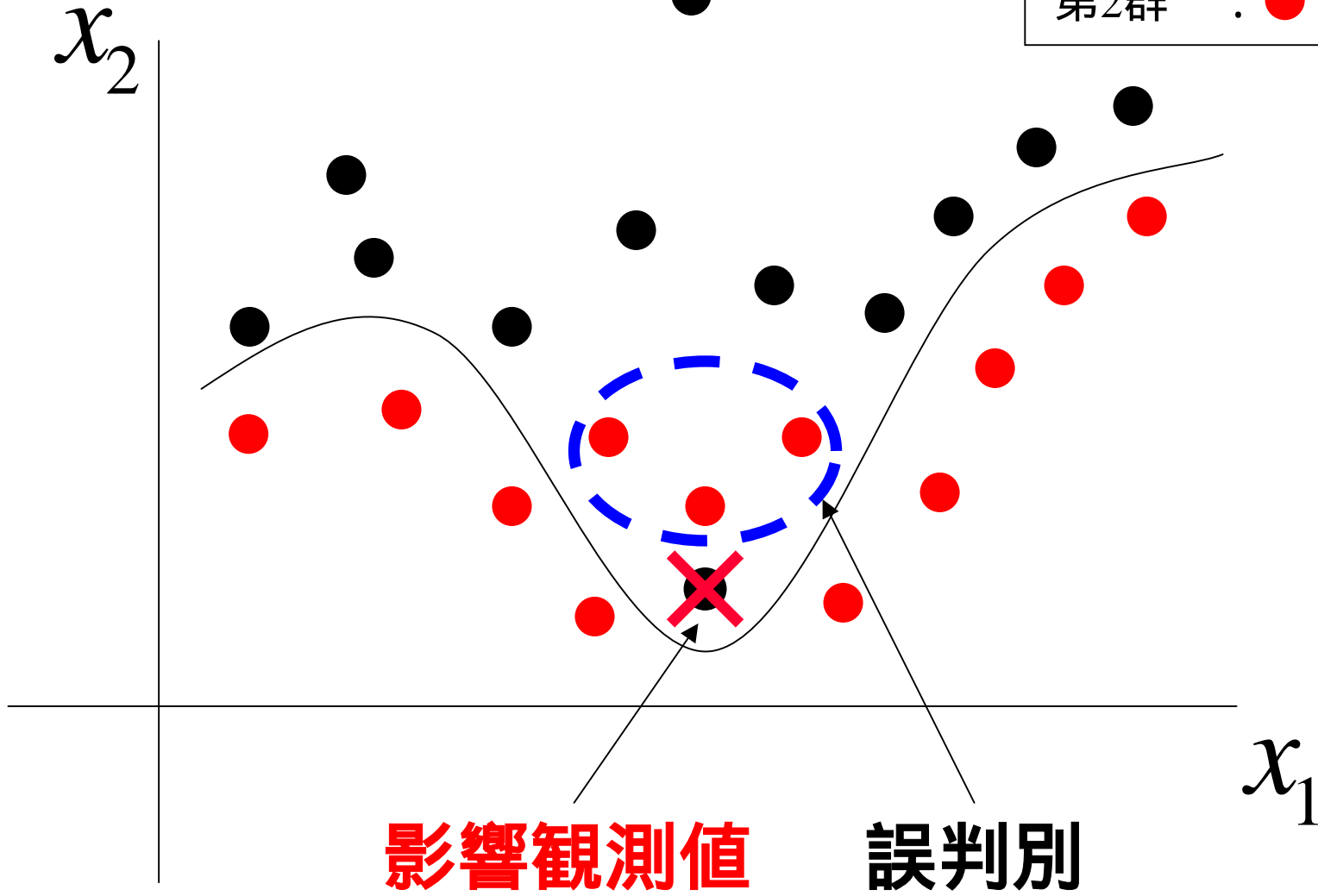
バイアスのブートストラップ推定

隠れユニット数	1	2	3	4
<i>EIC</i>	1060.90	1037.96	1057.70	1093.92

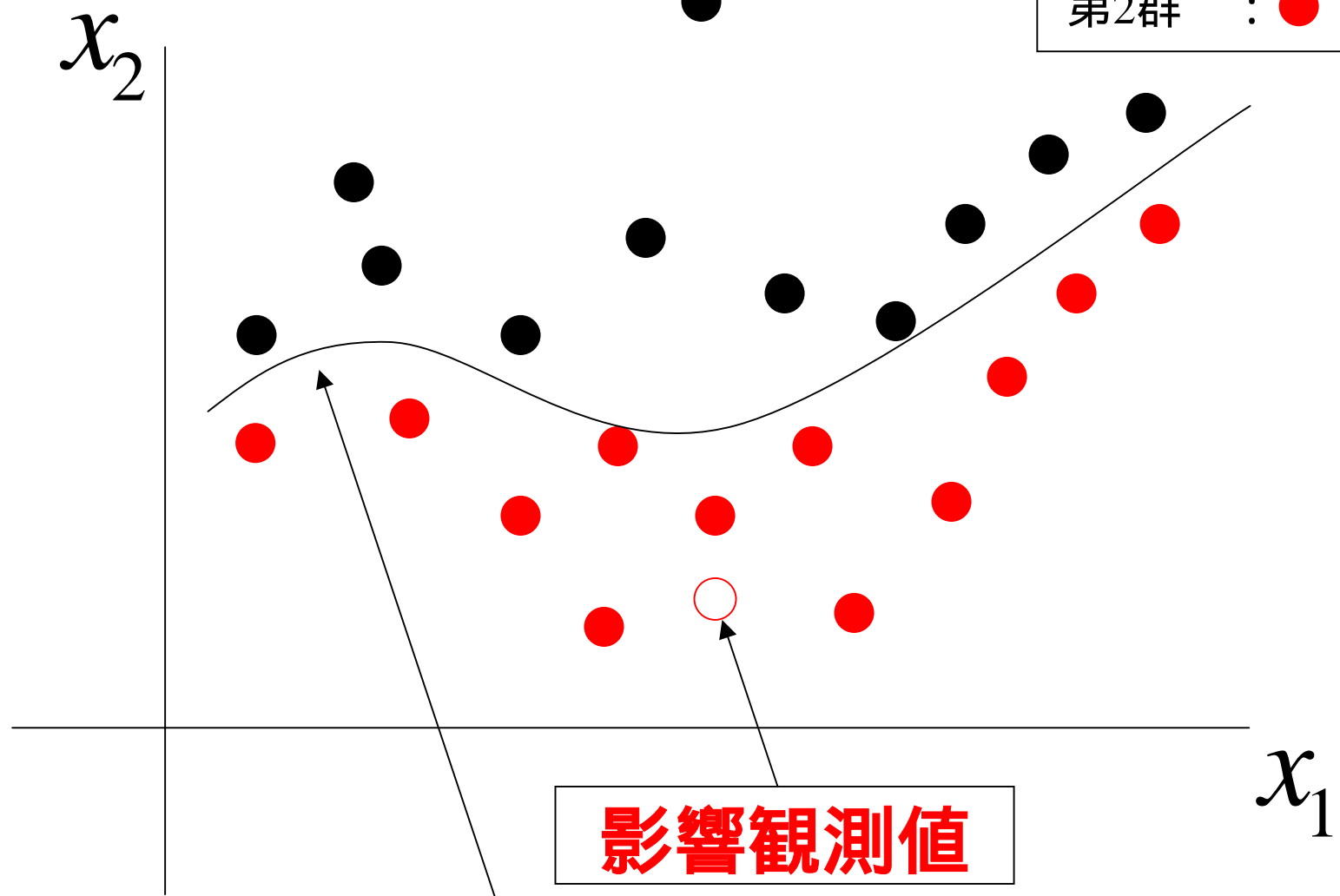


3. 影響分析

第1群 :
第2群 : ●



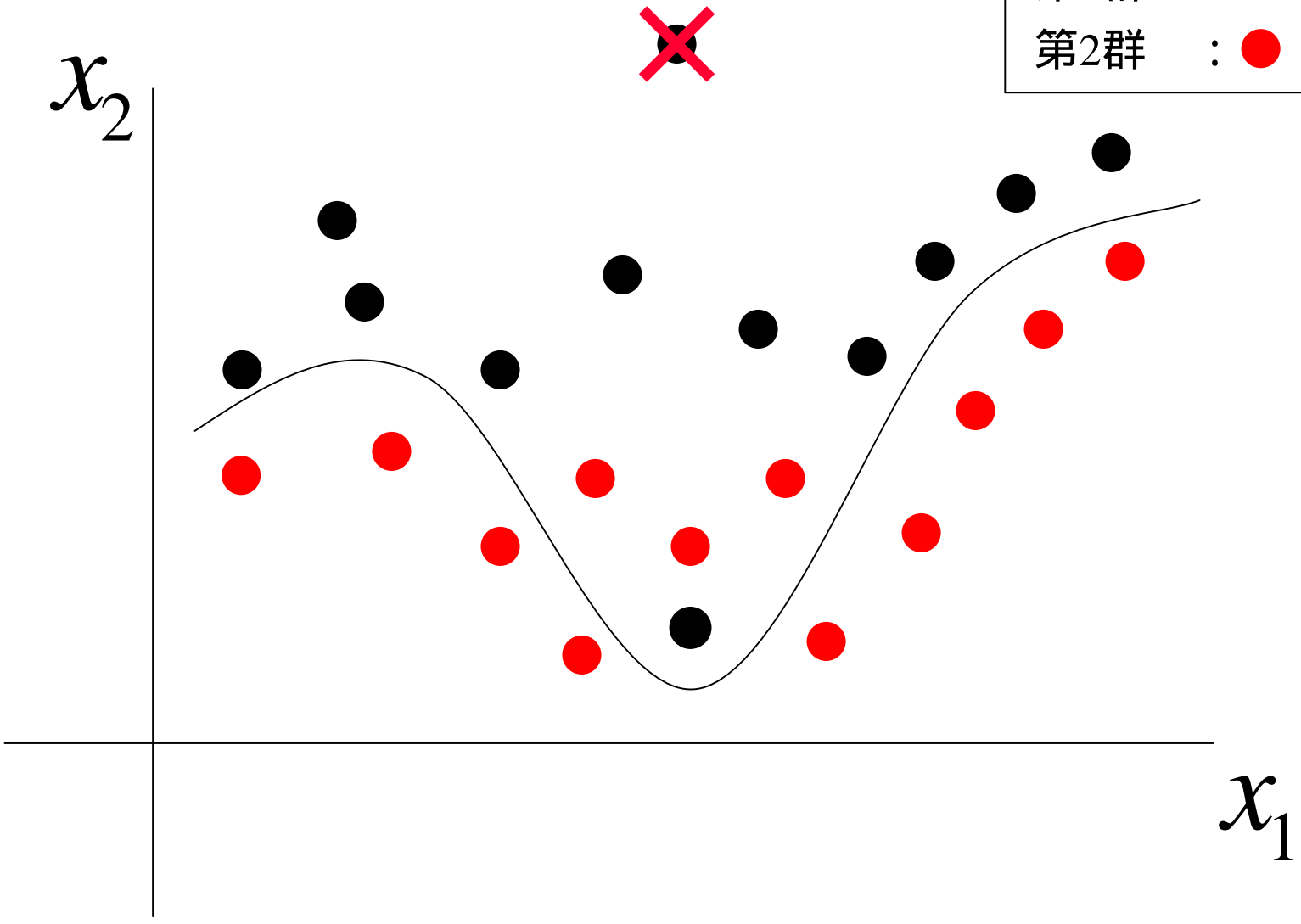
第1群 : ●
第2群 : ●



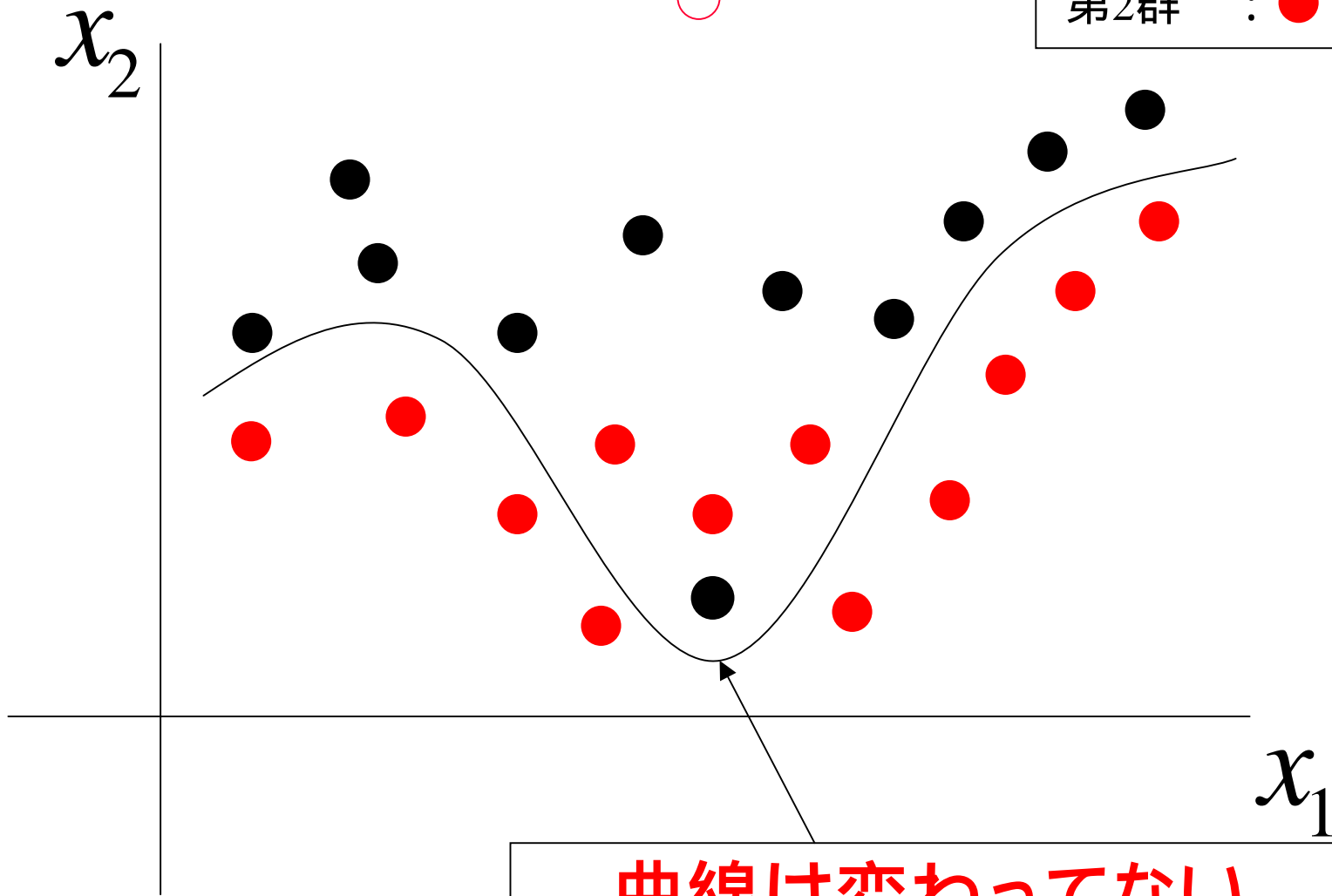
影響観測値

簡単な関数に変わり、誤判別が無くなる

第1群 :
第2群 : ●



第1群 :
第2群 : ●



曲線は変わってない

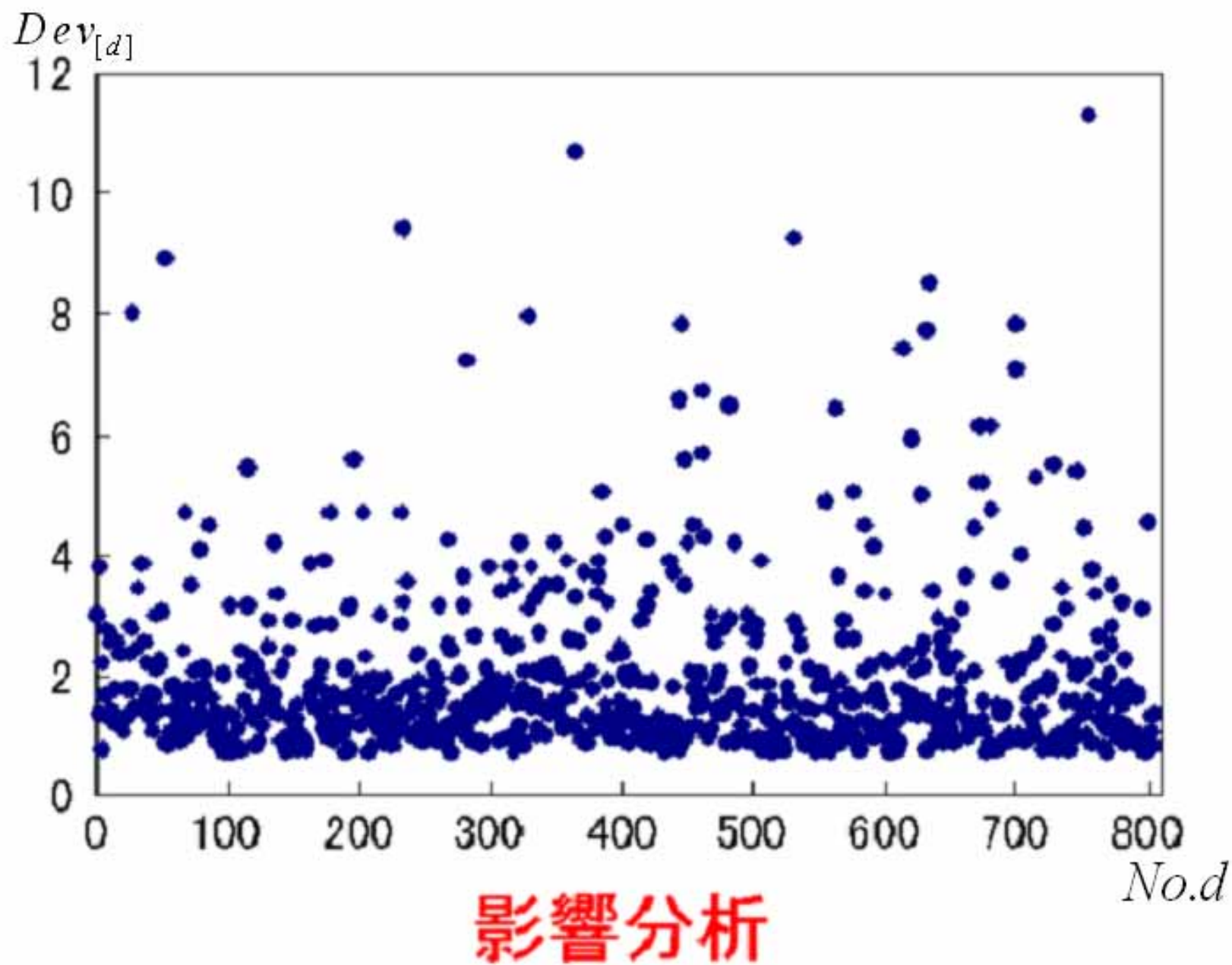
(1) DIFDEV

第d番目のマンションデータが解析において
有意であるか否かを検定。

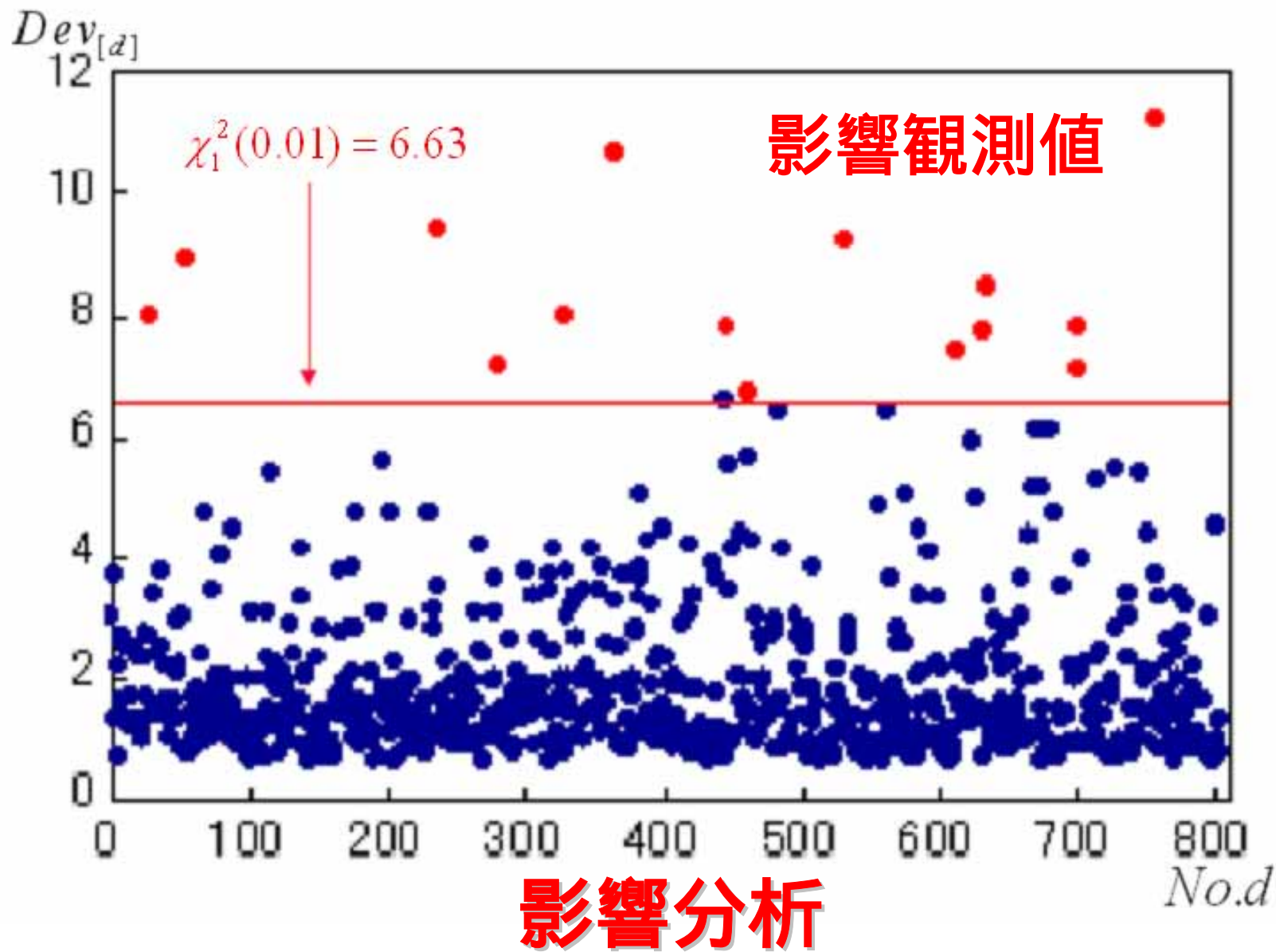
$$\Delta Dev_{[d]} = Dev - Dev_{[d]} \sim \chi_1^2$$
$$Dev = -2 \ln L \{ X; \hat{\theta}(X) \}$$

*Dev*は対数尤度の-2倍

小さい程望ましい



外れ値の統計的根拠



影響分析の結果

	逸脱度	誤判別率
全ての観測値を用いた場合	976.68	0.259
影響観測値15個を削除した場合	856.51	0.241

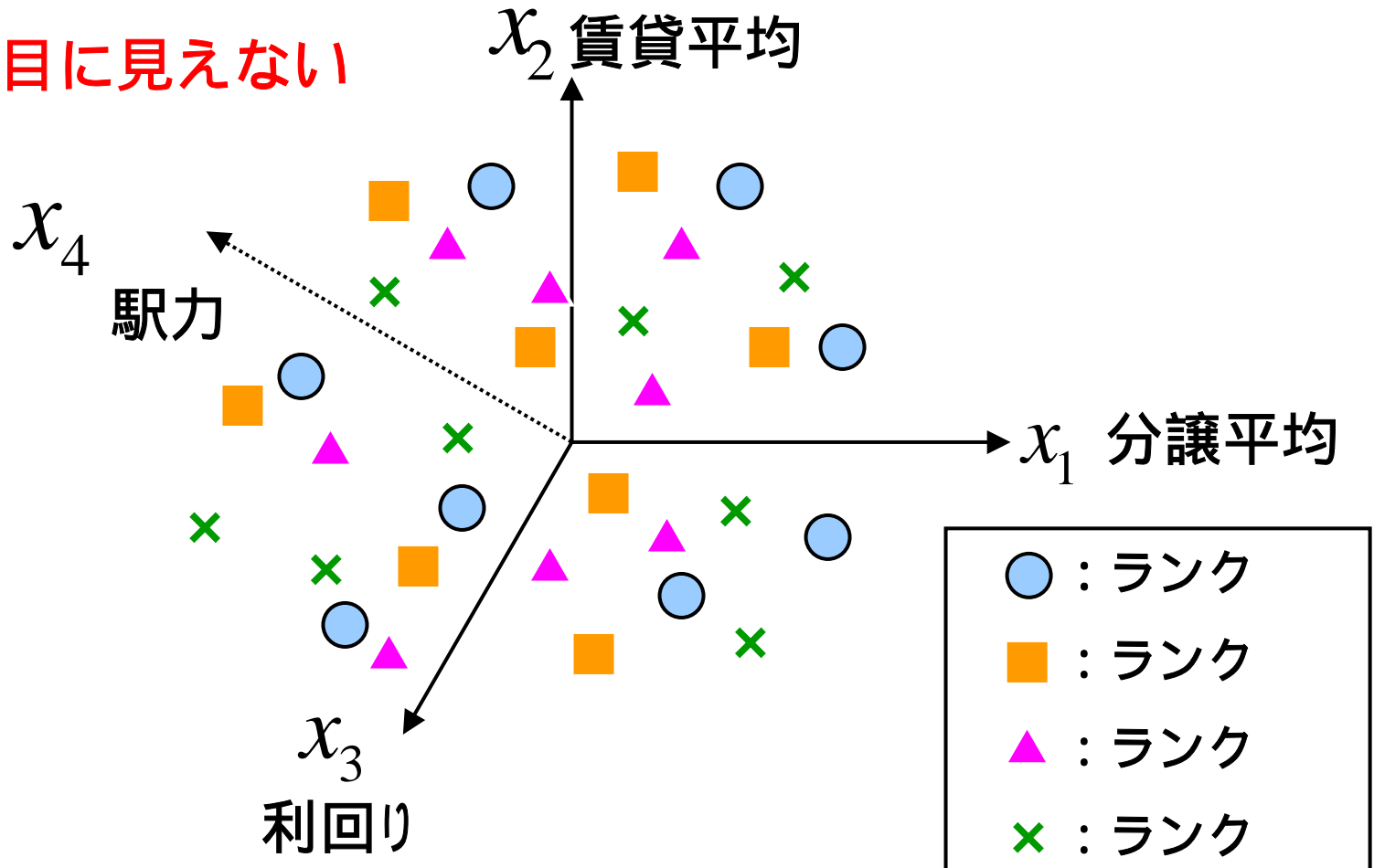
4. 情報圧縮

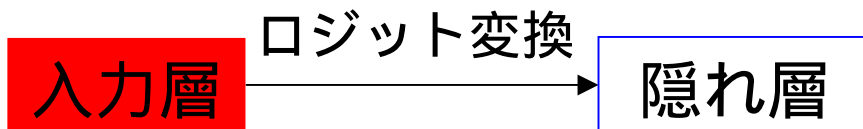
情報圧縮



多次元データを2次元に
視覚化する

点線は目に見えない





$$x_i \Rightarrow u_j = \sum_{i=0}^I \alpha_{ij} x_i \Rightarrow y_i = \frac{1}{1 + \exp(-u_j)} \quad : \text{ロジット変換}$$

圧縮スコア : $\hat{u}_j = \sum_{i=1}^I \hat{\alpha}_{ij} x_i$

\Rightarrow 2次元プロット (\hat{u}_1, \hat{u}_2)

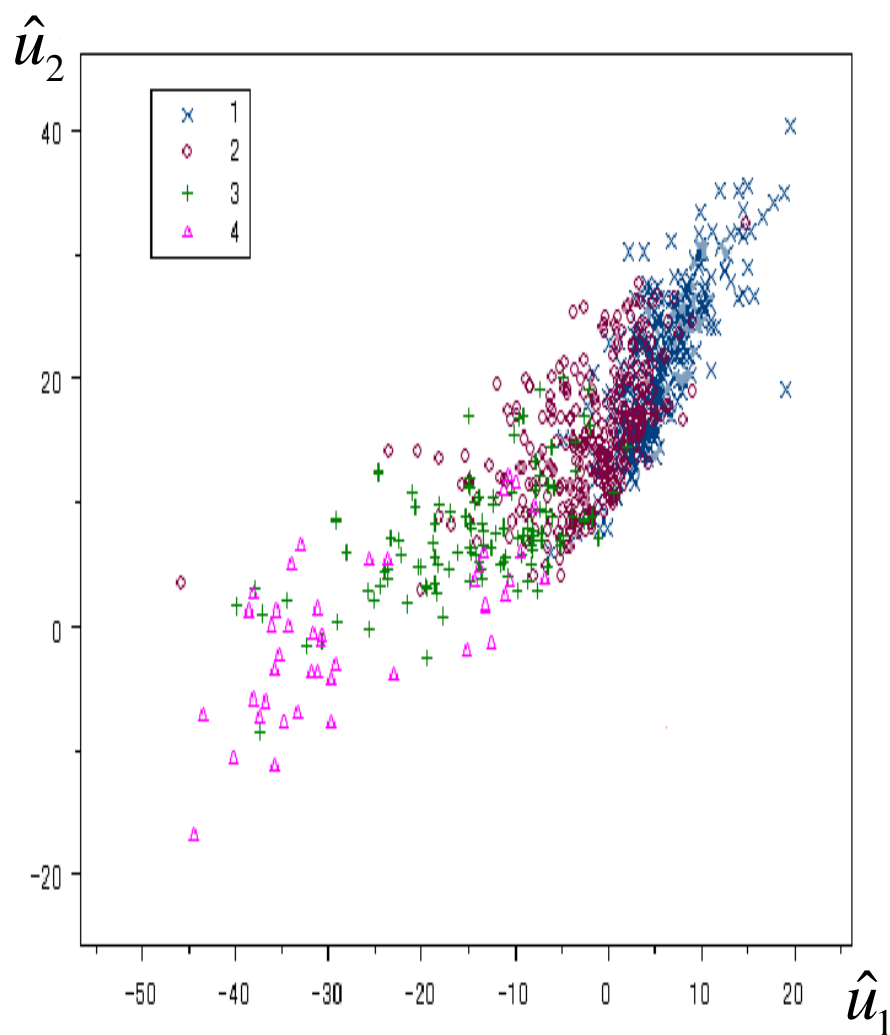
2次元 \downarrow

4次元 $\swarrow \downarrow \searrow$

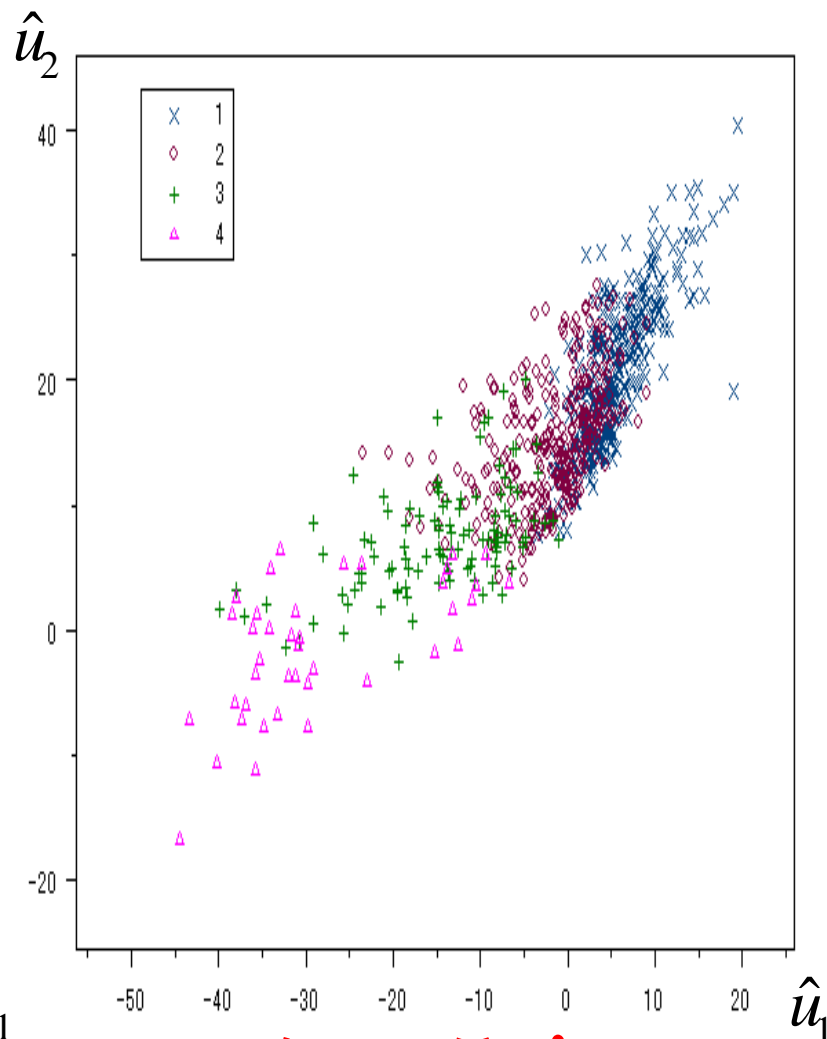
$$\begin{cases} \hat{u}_1 = \hat{\alpha}_{01} + \hat{\alpha}_{11} x_1 + \hat{\alpha}_{21} x_2 + \hat{\alpha}_{31} x_3 + \hat{\alpha}_{41} x_4 \\ \hat{u}_2 = \hat{\alpha}_{02} + \hat{\alpha}_{12} x_1 + \hat{\alpha}_{22} x_2 + \hat{\alpha}_{32} x_3 + \hat{\alpha}_{42} x_4 \end{cases}$$

4次元 (x_1, x_2, x_3, x_4) \Rightarrow 2次元 (\hat{u}_1, \hat{u}_2) でプロット

情報圧縮(隠れユニット2個)

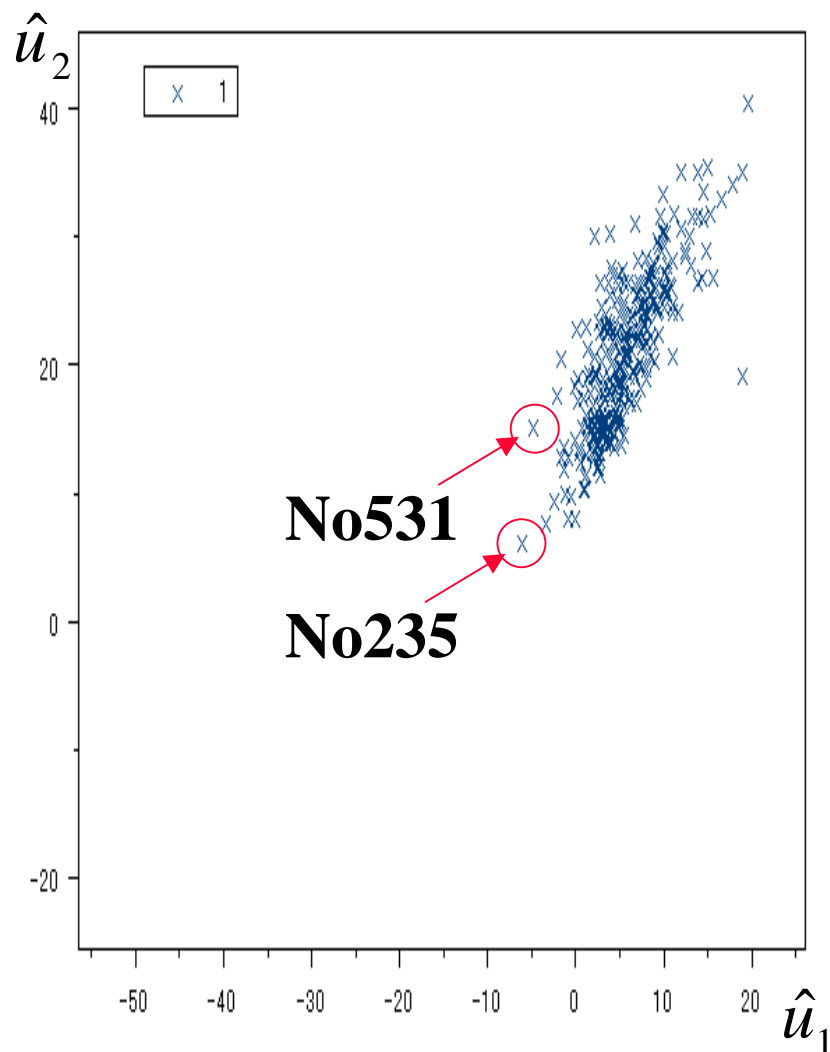


2次元圧縮プロット(全データ)

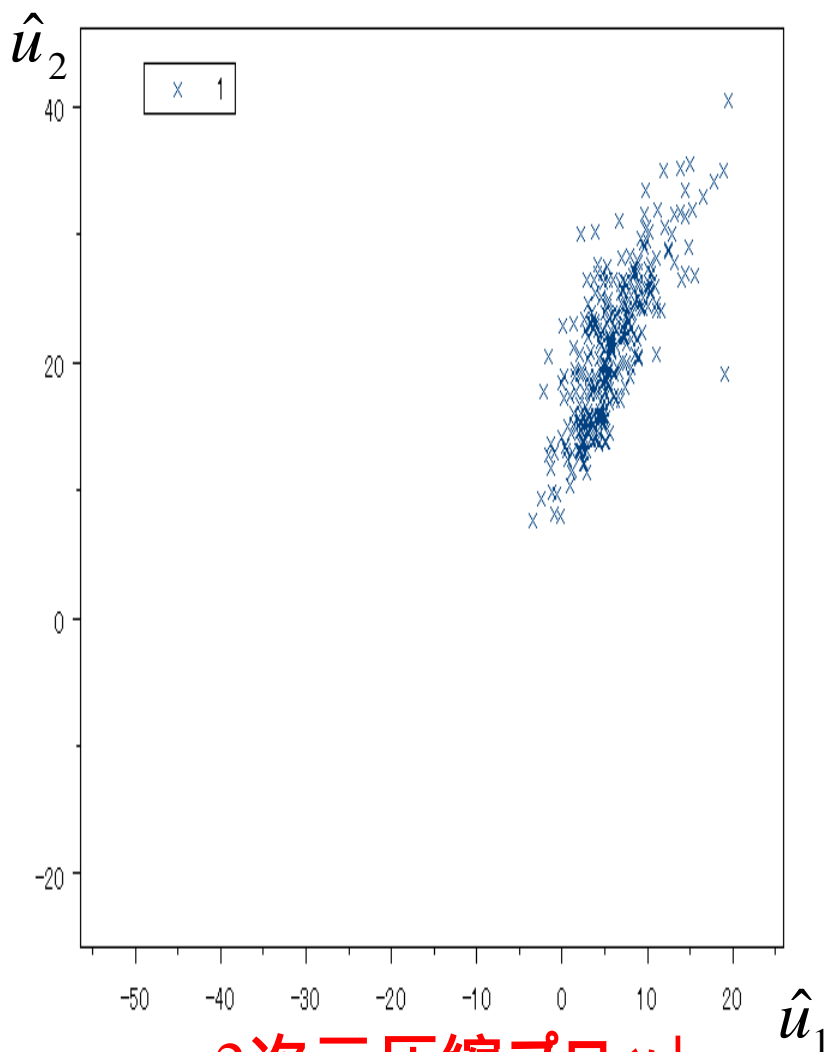


2次元圧縮プロット
(影響観測値除去後)

情報圧縮 (隠れユニット2個)

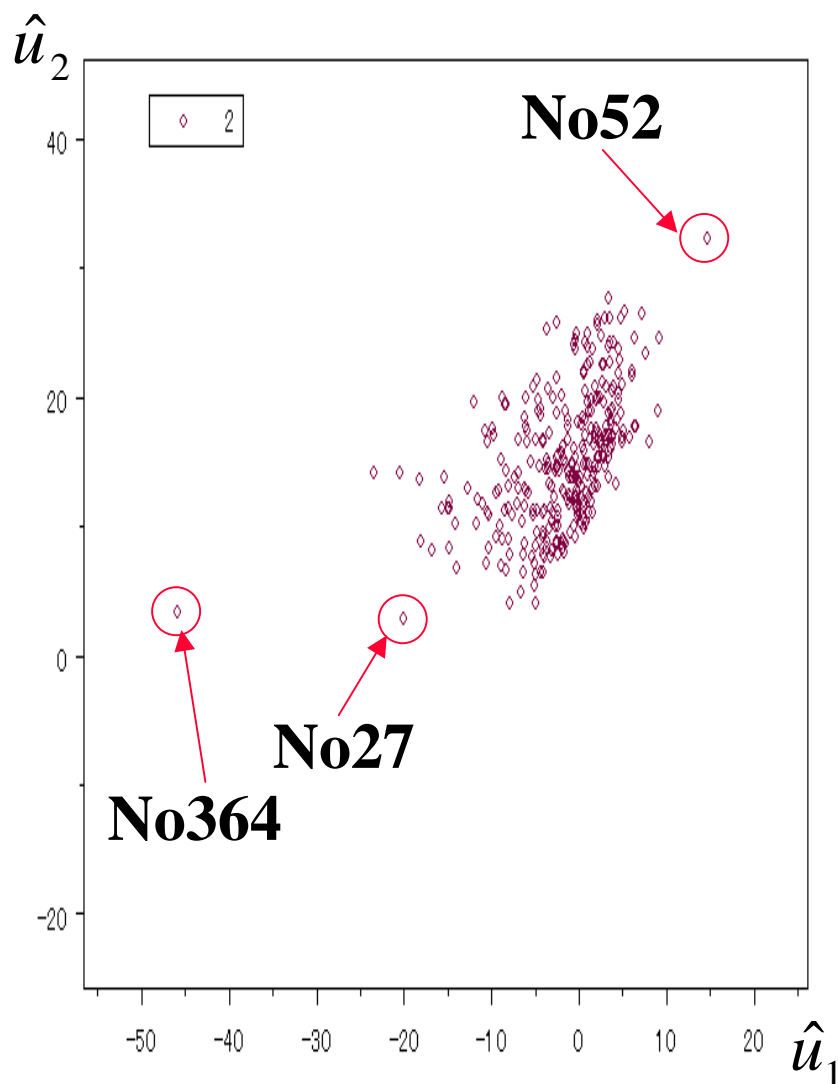


2次元圧縮プロット(全データ)

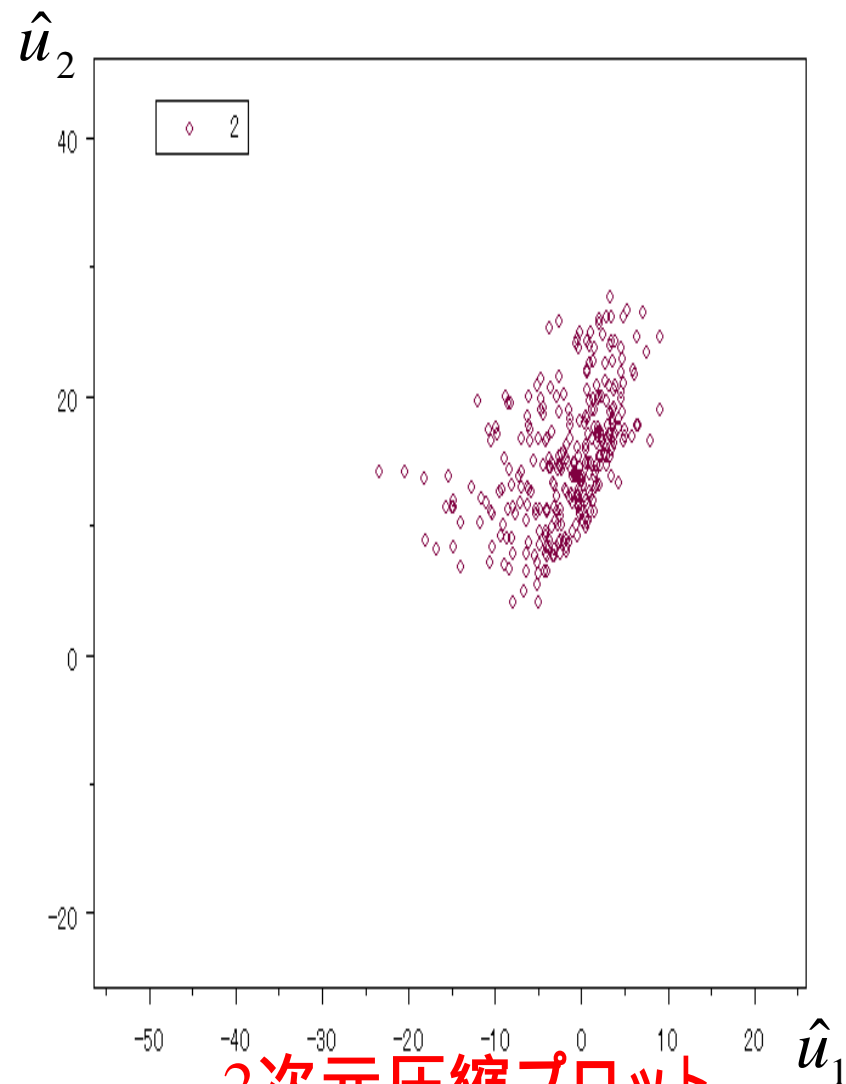


2次元圧縮プロット
(影響観測値除去後)

情報圧縮(隠れユニット2個)

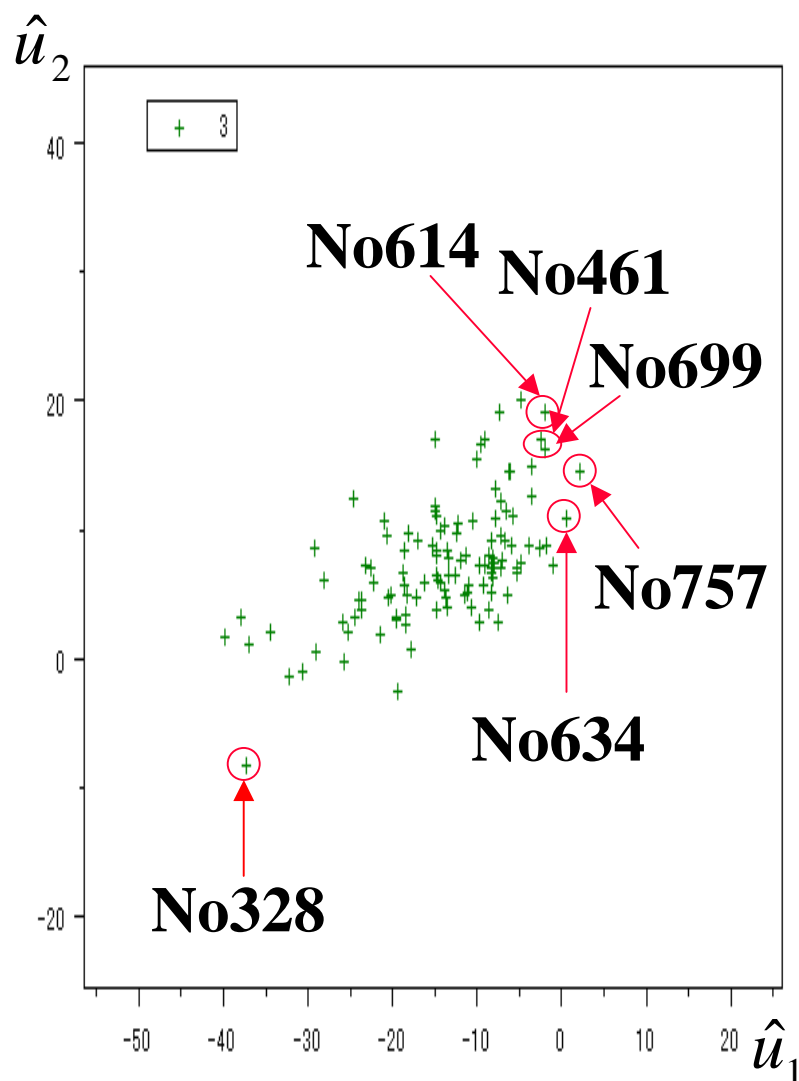


2次元圧縮プロット(全データ)

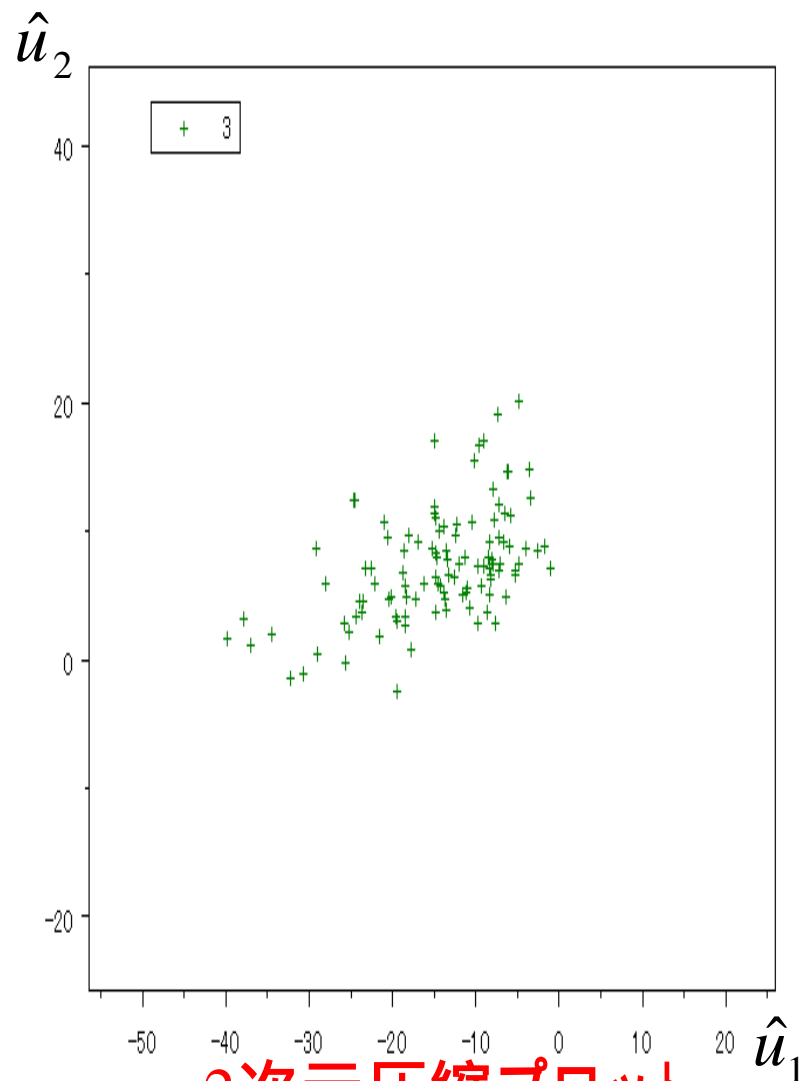


2次元圧縮プロット
(影響観測値除去後)

情報圧縮(隠れユニット2個)

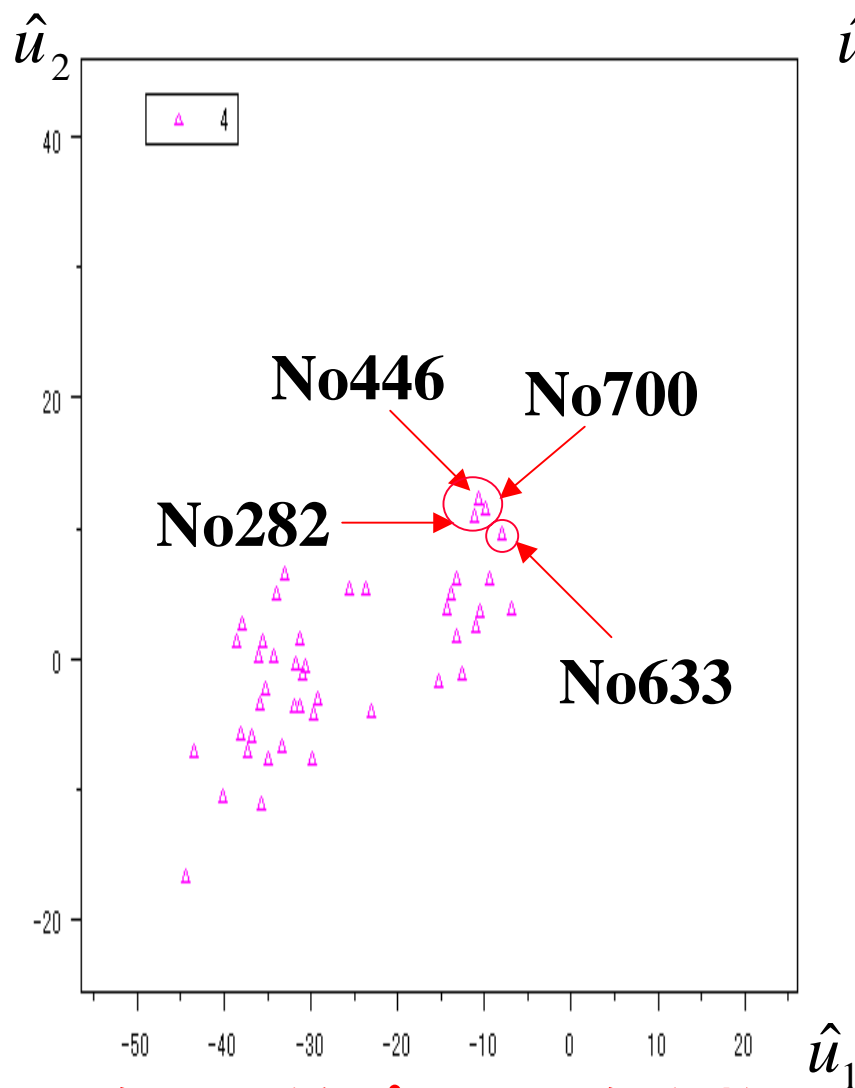


2次元圧縮プロット(全データ)

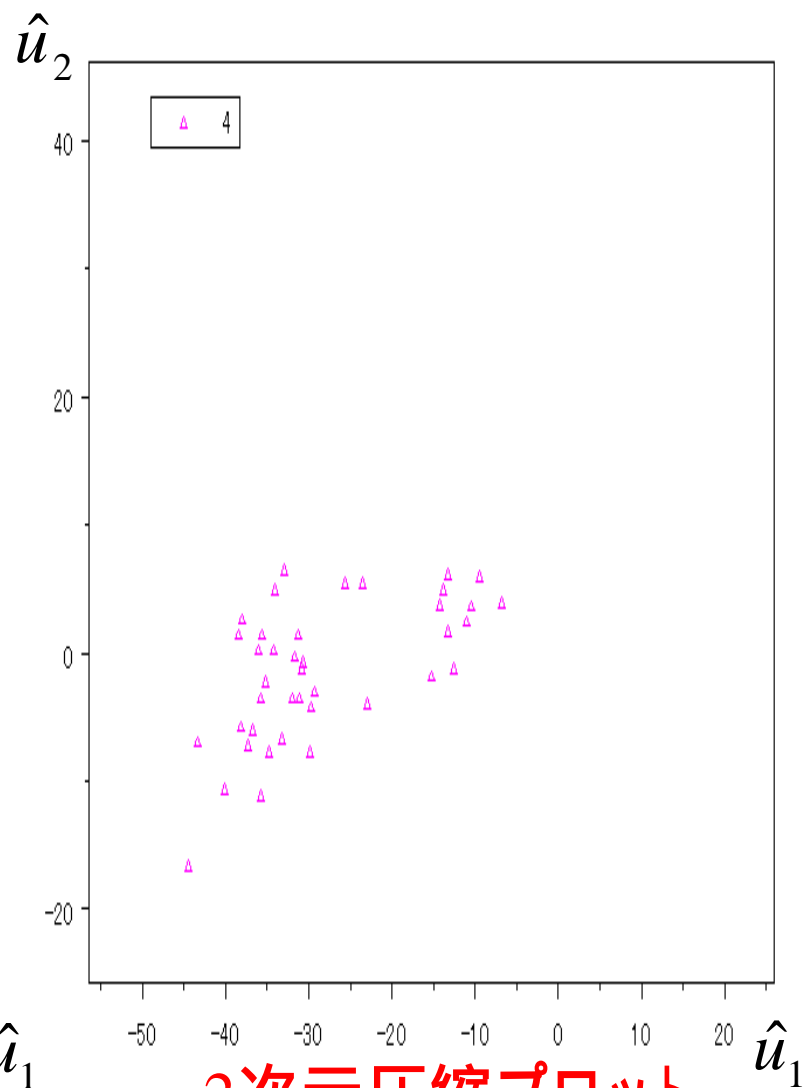


2次元圧縮プロット
(影響観測値除去後)

情報圧縮 (隠れユニット2個)



2次元圧縮プロット (全変数)



2次元圧縮プロット
(影響観測値除去後)

5.性能比較

判別方法	誤判別率	
	もとのデータ	影響観測値 15個削除後
線形判別	0.319	0.295
2次判別	0.288	0.278
多群ロジスティック判別	0.268	0.245
ニューロ判別	0.259	0.241

6. まとめ

(1) ニューラルネットワークモデルの構築

隠れユニット数の決定 (EIC)

(2) DIFDEVを用いた影響分析 (15個削除)

逸脱度, 誤判別率が減少

(3) 情報圧縮

4次元データを2次元に視覚化

(4) 性能比較

ニューロ判別が、もっと良い