SAS ユーザー総会 2017

GLMSELECTプロシジャにおけるLassoの有用性に関する検討

○川﨑勝己[1]、有光導徳[2]、新城博子[1]

([1]エイツーヘルスケア株式会社

ベルメディカル開発本部　データサイエンス部　臨床薬理グループ、[2]同　東京臨床統計グループ)

A study on the usefulness of Lasso using GLMSELECT procedure

Katsumi Kawasaki[1]　Michinori Arimitsu[2]　Hiroko Shinjo[1]

[1]Clinical Pharmacology group, A2 Healthcare Corporation

[2]Tokyo Biostatistics Group, A2 Healthcare Corporation

要旨:

モデル構築の際の変数選択の手法の1つであるLassoに関して、いくつかのデータパターンを想定したGLMSELECTプロシジャによるシミュレーションを実行し、その有用性について検討した。

キーワード:Lasso、GLMSELECT

発表構成

- 1. はじめに
- 2. 背景
- 3.Lassoについて
- 4. シミュレーション方法
- 5. シミュレーション結果
- 6. 結果のまとめ

はじめに

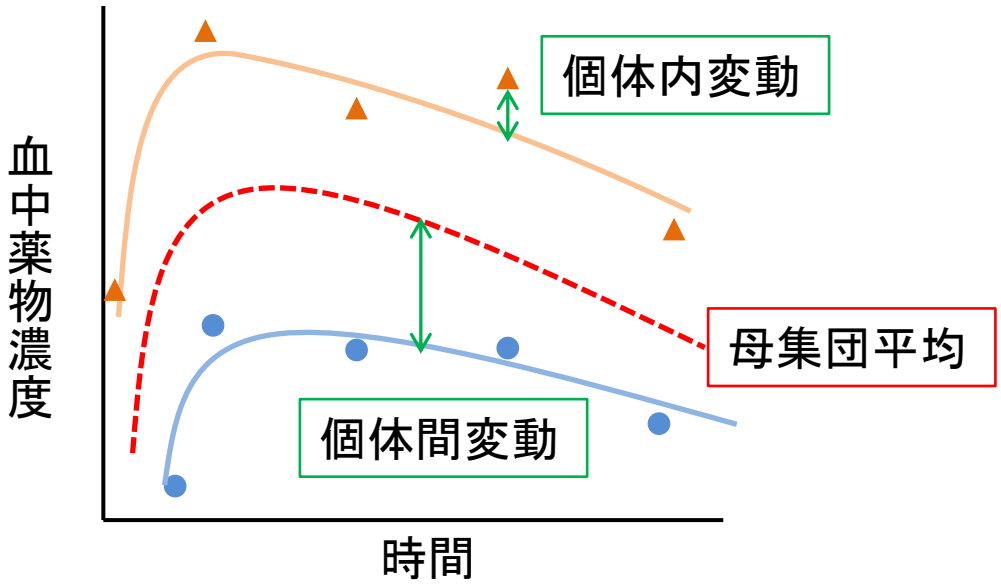
- 臨床薬理領域で母集団解析によりモデル構築を行う際、
  データの状況によってはモデル構築が困難となる場合もある。
  一方、最近Lassoを用いた変数選択が広い分野で注目を集めている。そこで、本発表ではGLMSELECTプロシジャを用いて、どのようなデータに対してLassoが有用性を発揮するか検討した。

背景[1] - 用語の説明 -

- 母集団薬物動態（Population Pharmacokinetics, PPK）解析

医薬品開発における薬物動態の検討として使われる解析手法

非線型混合効果モデル（Nonlinear mixed-effect model）に基づく



背景[2]- PPK解析の特徴 -

① 集団として十分な被験者数であれば、1被験者あたりの採血点数が少なくても解析可能

② 薬物動態に影響を与える共変量を定量的に評価可能

⇒ 濃度点数がスパースなデータを用いて、共変量選択しモデルを構築することも少なくない。

背景[3]- 変数選択法 -

・Stepwise法：汎用される変数選択法の1つ。PPK解析においても利用される方法。

⇒近年、機械学習の分野等で、変数の数が比較的多いデータに対する変数選択法として、Lassoをはじめとしたスパース性を利用した手法が注目を集めている。

本検討の目的

今回は、線形モデルを用いたシミュレーションを実施し、どのようなデータに対して、Lassoが有用性を発揮するかを検討した。

Lasso

- Lasso (Least Absolute Shrinkage and Selection Operator)
  - Robert Tibshirani(1996)により提案された、パラメータの推定と変数選択を同時に行う手法
  - 一部のパラメータを完全に0として推定することができる
  - 多数の説明変数からなる大規模モデルの変数選択手段として研究が進められている

Lasso

$$\sum_{i=1}^{n}\left\{y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right\}^2 \quad \text{subject to} \quad \sum_{j=1}^{p}\left|\beta_j\right| \le t$$

損失関数

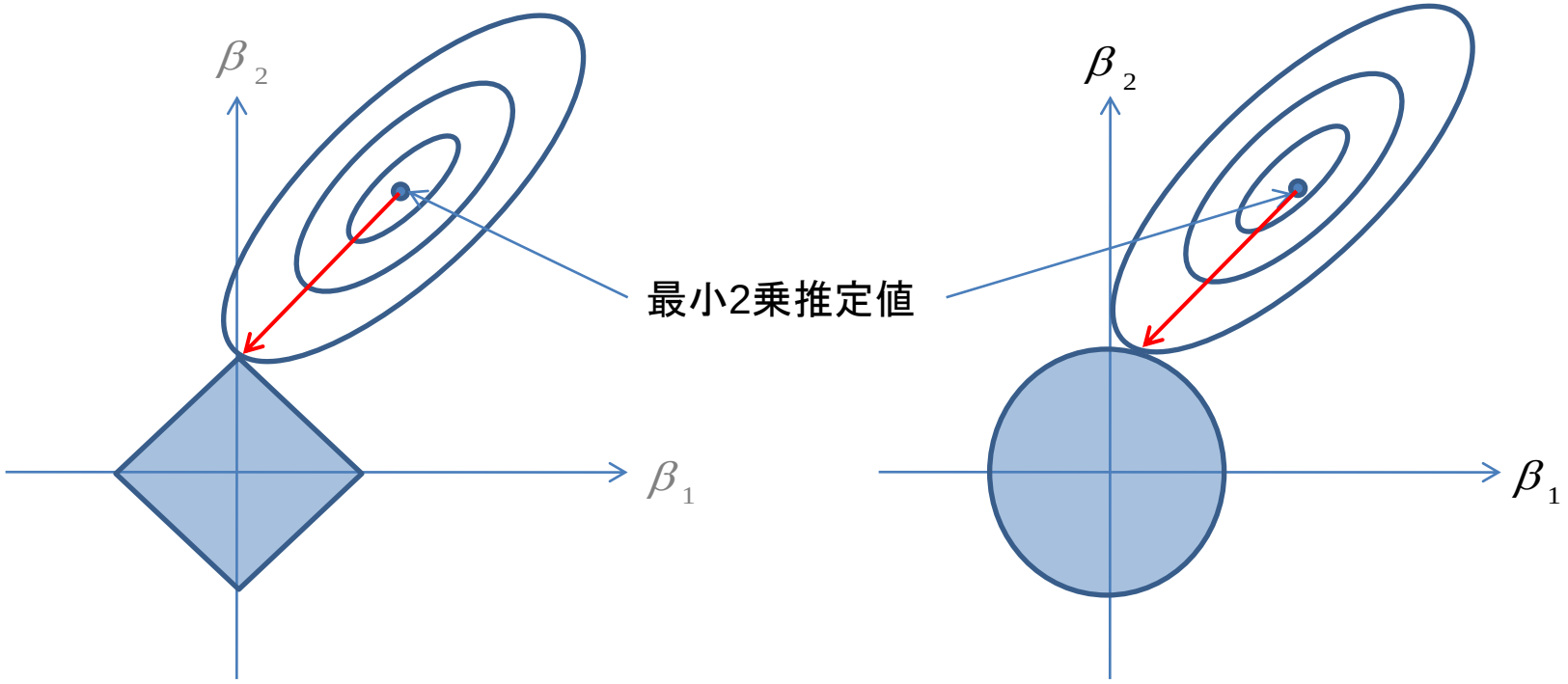
制約条件

※　変数は中心化

– 損失関数をパラメータの絶対値を含む制約（L1ノルム正則化）をつけて最小化する。

Lasso

p = 2の場合



$$|\beta_1| + |\beta_2| \leq t$$

$$\beta_1^{\,2} + \beta_2^{\,2} \leq t^2$$

Lasso

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right\}^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$

- Lagrangeの未定乗数法を用いた最適化問題として考える。
- パラメータにバイアスを加えることで、パラメータの分散を小さくし、結果的に予測精度を上げる。
- λ（>0）（正則化パラメータ）はパラメータの縮小度合いを制御する。
- λ を大きくすると、制約の性質上0となるパラメータが多くなり、ある種の変数選択が可能となる。また、パラメータの分散を小さくすることが出来る。
- ただし、λ が大きくなりすぎるとパラメータのバイアスが大きくなり予測精度が下がってしまう。
- cross-validation等で適切な λ を選択する必要がある。

オラクル性

- 変数選択の一致性
  - サンプルサイズnが大きくなると、0でない係数（$\beta_j \neq 0$）を持つ説明変数が正しく選択される確率が1に収束する。

- 漸近正規性
  - 0でない係数（$\beta_j \neq 0$）の推定量は、漸近不偏性、漸近正規性を持つ。

Lassoはオラクル性を持たない。

Addaptive Lasso

$$\sum_{i=1}^{n}\left\{ y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right\}^2 \qquad \text{subject to} \quad \sum_{j=1}^{p} \left| w_j \beta_j \right| \le t$$

損失関数

制約条件

※ 変数は中心化

- Zou(2006)により考案された、パラメータにオラクル性を持たせたLassoの改良型。
- Lassoのパラメータ制約条件に重み $w_j = 1/\left|\hat{\beta}_j\right|^{\gamma}$ を考慮している。

シミュレーション法[1] -条件設定-

- 検討した変数選択法
  - ✓Stepwise Selection
  - ✓Lasso Selection
  - ✓Adaptive LASSO Selection
- 検討のための条件設定
  - ✓候補変数間の関連[ρ]（低←→高）
  - ✓候補変数の数[P]（少←→多）
  - ✓オブザベーション数[N]（小←→大）

シミュレーション法[2] -モデル設定-

- シミュレーションデータの設定

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + N(\mathbf{0}, \boldsymbol{\sigma}^2)$$

$\mathbf{y}$: 従属変数

$\mathbf{X}$: 変数間の関連を考慮した多次元正規分布に従う乱数により発生させた候補変数

$\boldsymbol{\beta}$: $\mathbf{X}$ の係数

- 想定したモデルの設定

✓共変量と見立てた変数を任意に選択し、係数の値を設定した。

シミュレーション法[3] -評価方法-

- 評価方法
  - GLMSELECTプロシジャのModelAverageステートメントの（Bootstrap法）を用いた結果より以下の内容を検討した。サンプリング法は無作為抽出（復元抽出）を用いた。
    - ✓想定したモデルを選んだ確率（%）
    - ✓パラメータとして設定した係数の推定値の平均
    - ✓Non-zero / Zeroの個数
    - ✓選択されたモデルのNon-zeroの個数

シミュレーション法[4] -設定まとめ-

- 条件設定

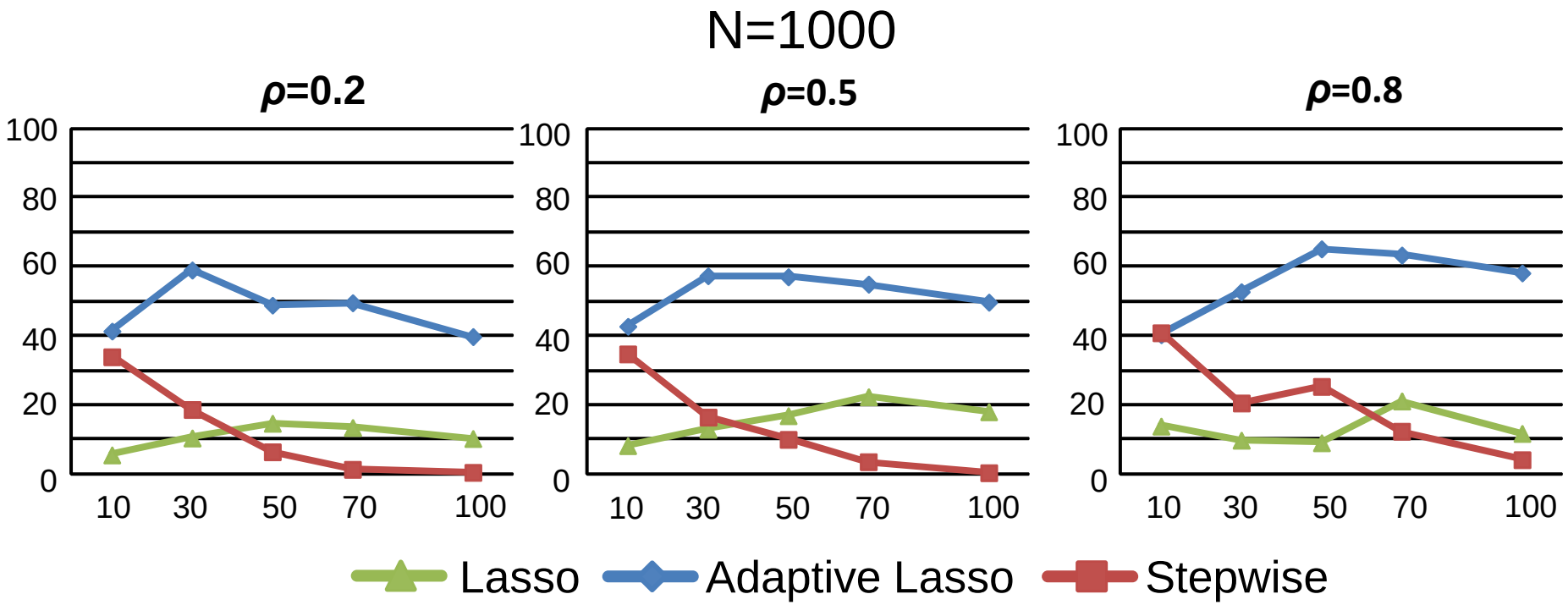
シナリオ	オブザベーション数 [N]	候補変数の数 [P]	変数間の相関 [ρ]	Non-zeroの係数
1	1000	10,30,50,70,100	0.2	{β1,β2,β3,β4,β5,β6,β7,β8, …,βp} ={3, 1.5, 0, 0, 2, 2.5, 3.2, 0,…,0} Non-zeroの個数:5個
2			0.5	
3			0.8	
4	500	10,30,50,70,100	0.2	
5			0.5	
6			0.8	

- Bootstrapによる試行回数:1000回

シミュレーション結果[1]

●想定したモデルを選択した確率（変数の個数毎の推移）

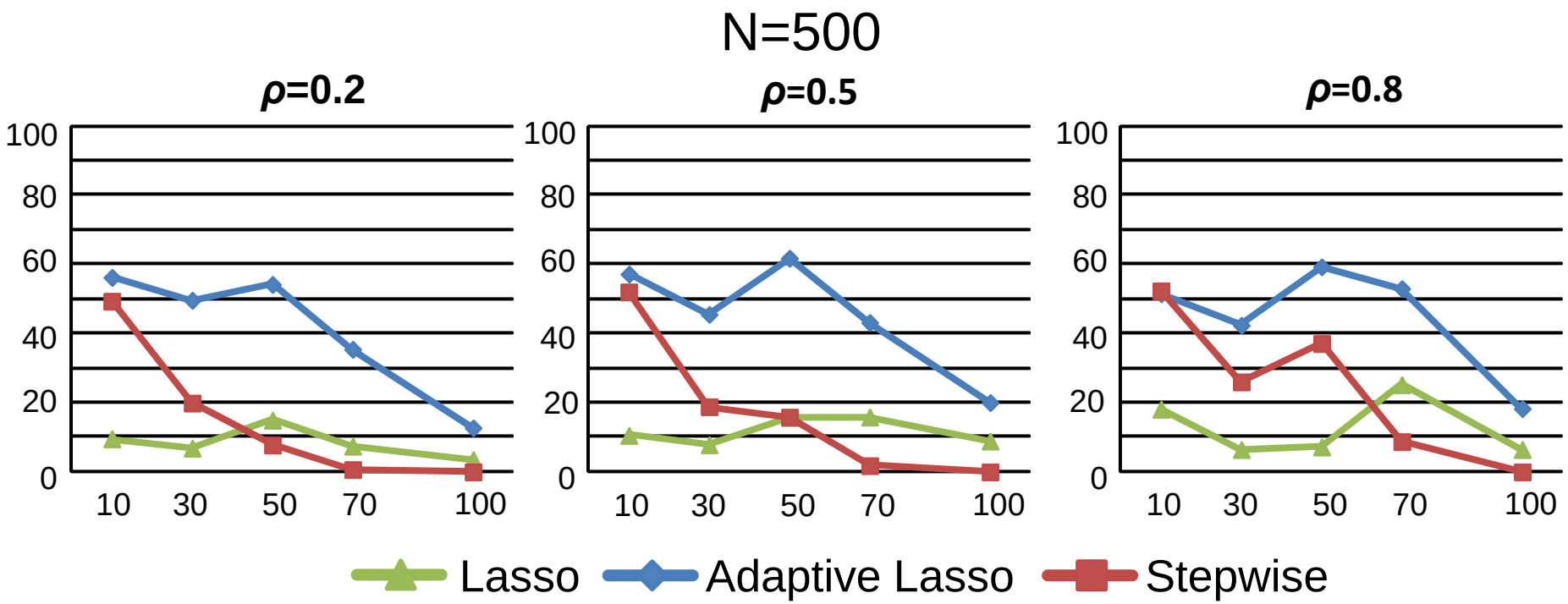
縦軸：想定したモデルを選択した確率（%）、横軸：変数の個数



シミュレーション結果[2]

●想定したモデルを選択した確率（変数の個数毎の推移）

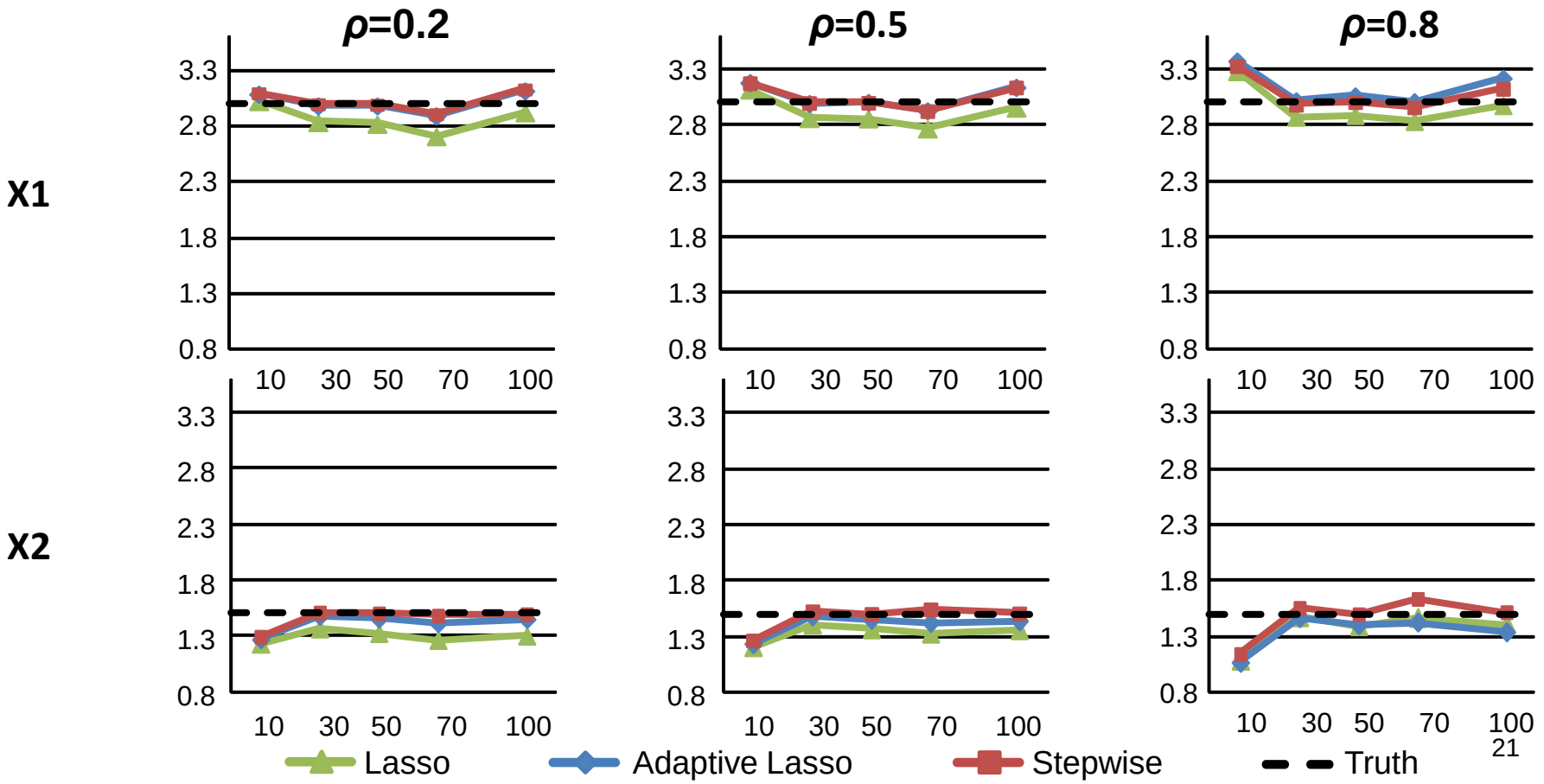
縦軸：想定したモデルを選択した確率（%）、横軸：変数の個数



シミュレーション結果[3-1]

●設定した係数の推定値の平均の推移(N=1000)

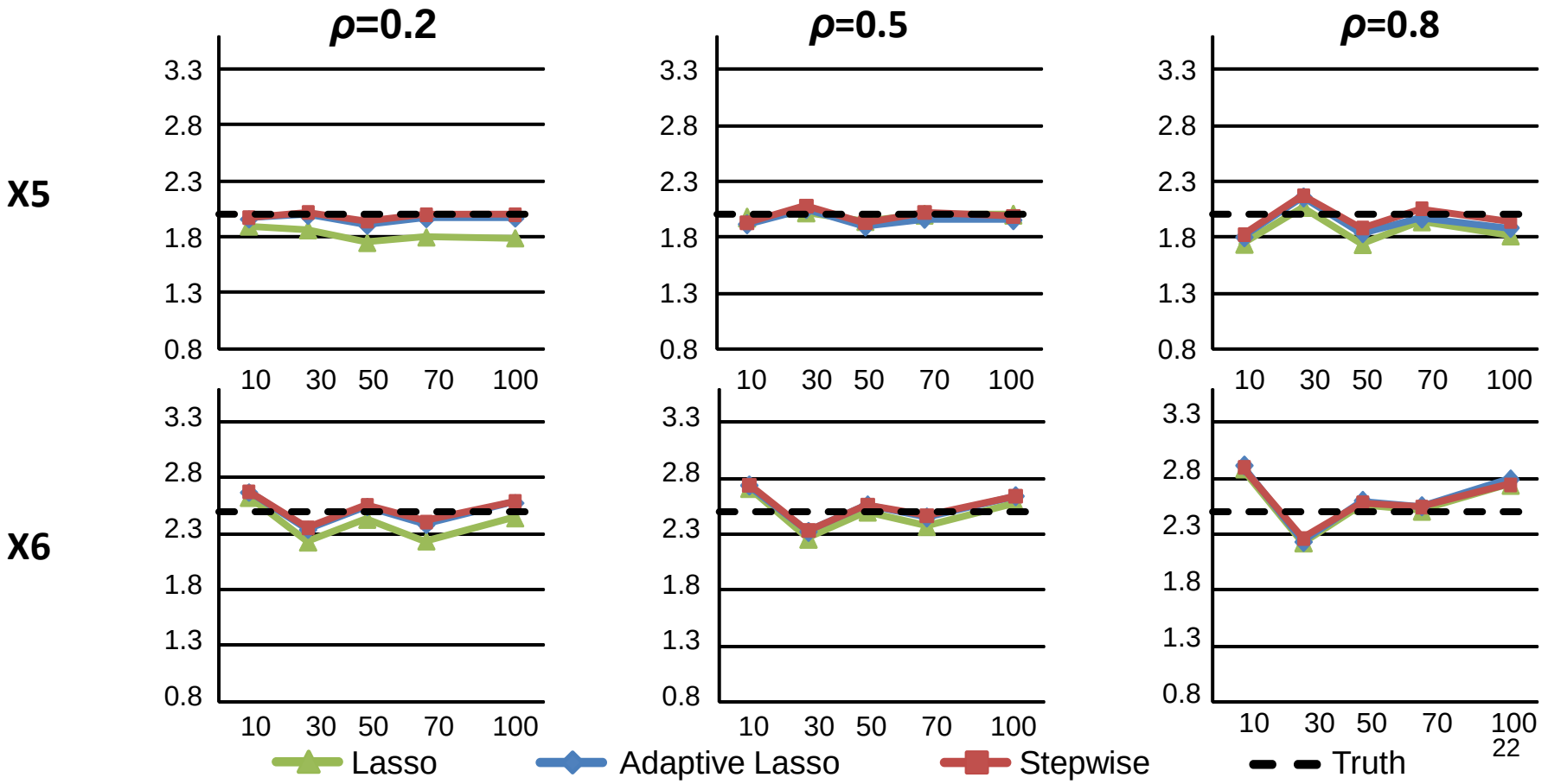
縦軸：推定値の平均、横軸：変数の個数



シミュレーション結果[3-2]

●設定した係数の推定値の平均の推移(N=1000)

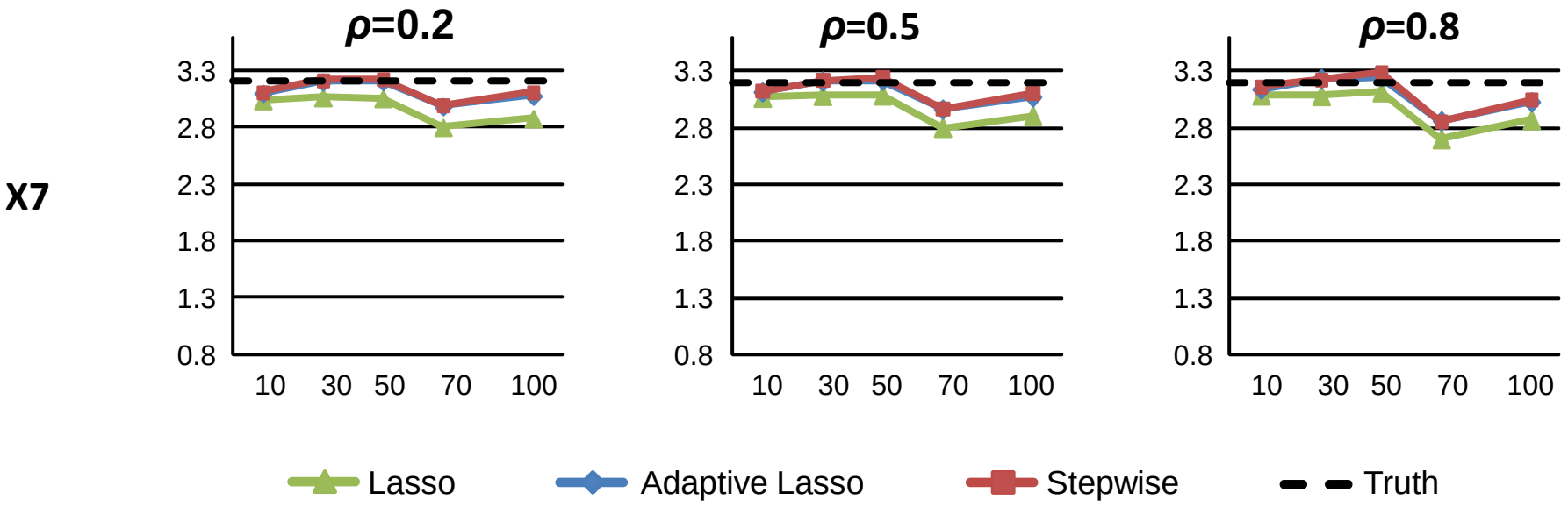
縦軸：推定値の平均、横軸：変数の個数



シミュレーション結果[3-3]

●設定した係数の推定値の平均の推移(N=1000)

縦軸：推定値の平均、横軸：変数の個数



シミュレーション結果[4]

Non-zero/Zeroの個数

N=1000

		候補変数の数									
		10		30		50		70		100	
		Non-Zero	Zero	Non-Zero	Zero	Non-Zero	Zero	Non-Zero	Zero	Non-Zero	Zero
	Truth	5	5	5	25	5	45	5	65	5	95
ρ =0.2	Lasso	5(5, 5)	2(0, 4)	5(5, 5)	21(13,24)	5(5, 5)	41(29,44)	5(5, 5)	61(43,64)	5(5, 5)	90(75,94)
	Adaptive Lasso	5(5, 5)	2(0, 4)	5(5, 5)	23(16,24)	5(5, 5)	42(33,44)	5(5, 5)	59(45,60)	5(5, 5)	92(77,93)
	Stepwise	5(5, 5)	2(0, 4)	5(5, 5)	22(18,25)	5(5, 5)	41(35,44)	5(5, 5)	60(54,64)	5(5, 5)	87(79,93)
ρ =0.5	Lasso	5(5, 5)	2(0, 5)	5(5, 5)	21(14,25)	5(5, 5)	42(30,45)	5(5, 5)	61(48,64)	5(5, 5)	91(75,94)
	Adaptive Lasso	5(5, 5)	2(0, 4)	5(5, 5)	23(16,24)	5(5, 5)	41(32,42)	5(5, 5)	61(46,62)	5(5, 5)	86(68,87)
	Stepwise	5(5, 5)	2(0, 5)	5(5, 5)	22(18,25)	5(5, 5)	41(34,44)	5(5, 5)	61(53,64)	5(5, 5)	88(78,93)
ρ =0.8	Lasso	5(5, 5)	3(0, 5)	5(5, 5)	22(16,25)	5(5, 5)	42(33,44)	5(5, 5)	58(49,61)	5(5, 5)	80(69,83)
	Adaptive Lasso	5(5, 5)	2(0, 4)	5(5, 5)	23(16,24)	5(5, 5)	42(26,43)	5(5, 5)	59(48,60)	5(5, 5)	80(62,81)
	Stepwise	5(5, 5)	2(0, 4)	5(5, 5)	22(18,25)	5(5, 5)	42(35,45)	5(5, 5)	62(55,64)	5(5, 5)	90(81,94)

Median (min, max)

シミュレーション結果[5]

Non-zero/Zeroの個数

N=500

		候補変数の数									
		10		30		50		70		100	
		Non-Zero	Zero	Non-Zero	Zero	Non-Zero	Zero	Non-Zero	Zero	Non-Zero	Zero
	Truth	5	5	5	25	5	45	5	65	5	95
ρ =0.2	Lasso	5(5, 5)	2(0, 5)	5(5, 5)	21(11,25)	5(5, 5)	41(29,45)	5(5, 5)	60(36,64)	5(5, 5)	87(50,93)
	Adaptive Lasso	5(5, 5)	3(1, 5)	5(5, 5)	23(16,24)	5(5, 5)	43(33,44)	5(5, 5)	62(40,64)	5(5, 5)	88(51,93)
	Stepwise	5(5, 5)	3(1, 5)	5(5, 5)	22(18,25)	5(5, 5)	41(35,44)	5(5, 5)	59(48,64)	5(5, 5)	82(71,90)
ρ =0.5	Lasso	5(5, 5)	3(0, 5)	5(5, 5)	21(12,25)	5(5, 5)	42(31,45)	5(5, 5)	61(43,64)	5(5, 5)	87(55,92)
	Adaptive Lasso	5(5, 5)	3(1, 5)	5(5, 5)	22(16,25)	5(5, 5)	43(31,44)	5(5, 5)	62(43,64)	5(5, 5)	90(52,93)
	Stepwise	5(5, 5)	3(1, 5)	5(5, 5)	22(18,25)	5(5, 5)	41(35,45)	5(5, 5)	60(52,64)	5(5, 5)	84(70,91)
ρ =0.8	Lasso	5(5, 5)	3(1, 5)	5(5, 5)	22(14,25)	5(5, 5)	43(34,45)	5(5, 5)	62(51,64)	5(5, 5)	89(71,94)
	Adaptive Lasso	5(4, 5)	3(1, 5)	5(4, 5)	22(15,24)	5(4, 5)	43(35,44)	5(5, 5)	62(49,62)	5(4, 5)	90(54,93)
	Stepwise	5(4, 5)	3(1, 5)	5(4, 5)	22(18,25)	5(4, 5)	42(36,45)	5(5, 5)	61(55,65)	5(4, 5)	87(76,93)

Median (min, max)

シミュレーション結果[6]

選択されたモデルのNon-zeroの個数　N=1000

		候補変数の数				
		10	30	50	70	100
	Truth	5	5	5	5	5
ρ=0.2	Lasso	7(5,10)	7(5,16)	7(5,20)	8(5,27)	8(5,23)
	Adaptive Lasso	6(5, 8)	5(5,15)	6(5,15)	6(5,21)	6(5,21)
	Stepwise	6(5, 8)	6(5,10)	7(5,13)	9(5,16)	11(5,19)
ρ=0.5	Lasso	7(5,10)	7(5,15)	7(5,20)	7(5,22)	7(5,24)
	Adaptive Lasso	6(5, 9)	5(5,15)	5(5,14)	5(5,22)	6(5,24)
	Stepwise	6(5, 8)	6(5,10)	7(5,14)	8(5,16)	10(5,20)
ρ=0.8	Lasso	7(5,10)	7(5,13)	7(5,16)	7(5,16)	7(5,17)
	Adaptive Lasso	6(5, 9)	5(5,12)	5(5,21)	5(5,18)	5(5,23)
	Stepwise	6(5, 8)	6(5,10)	6(5,13)	7(5,14)	8(5,17)

Median (min, max)

シミュレーション結果[7]

選択されたモデルのNon-zeroの個数　N=500

		候補変数の数				
		10	30	50	70	100
	Truth	5	5	5	5	5
ρ=0.2	Lasso	7(5,10)	8(5,19)	7(5,19)	8(5,33)	11(5,48)
	Adaptive Lasso	5(5, 9)	6(5,14)	5(5,16)	6(5,30)	10(5,49)
	Stepwise	6(5, 8)	6(5,10)	7(5,13)	10(5,20)	16(8,27)
ρ=0.5	Lasso	7(5,10)	8(5,18)	7(5,17)	7(5,25)	9(5,41)
	Adaptive Lasso	5(5,10)	6(5,14)	5(5,18)	6(5,27)	8(5,46)
	Stepwise	5(5, 8)	6(5,10)	7(5,13)	9(5,17)	14(7,28)
ρ=0.8	Lasso	6(5,10)	7(5,15)	7(5,16)	6(5,17)	9(5,28)
	Adaptive Lasso	5(4,10)	6(4,15)	5(4,15)	5(5,18)	8(4,44)
	Stepwise	5(4, 9)	6(5,10)	6(5,12)	8(5,14)	11(5,22)

Median (min, max)

結果のまとめ[1]

- 想定したモデルを選択した確率について
  - Stepwiseは、変数の個数（以下、P）が少ないときは高い確率を示したが、Pが増えるにつれ低下した。
  - Lassoは、Pが少ない場合においても一番低い確率を示したが、Pが増えるとstepwiseを上回る確率を示した。
  - Adaptive Lassoが他の2つの選択方法に比べ安定した高い結果を示した。一方、PがNに近づくにつれ低くなる傾向が見られた。

結果のまとめ[2]

- 設定した係数の推定値の平均の推移について
  - いずれの方法においても、また、Pの大きさと関係なく推定値がぶれることはなかった。
- 選択されたモデルのNon-Zeroの個数について
  - 中央値が真の値の5となるのはAdaptive Lassoがほとんどであった。
  - Nが500の場合は、1000のときに比べPが大きくなるにつれNon-Zeroの個数がいずれの選択法においても増加する傾向が見られた。

参考文献

1. SAS//STAT 13.2 User's Guide
2. Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B, 58, 267–288.
3. Vasquez et al. (2016), Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application, BMC Medical Research Methodology (2016) 16:154
4. Zou, H. (2006). The Adaptive Lasso And Its Oracle Properties, Journal of the American Statistical Association, 101(476), 1418-1429.
5. 荒木孝治 (2013), 罰則付き回帰とデータ解析環境R, オペレーションズ・リサーチ 2013年5月号, 261-266
6. 小西貞則 (2010), 多変量解析入門, 岩波書店
7. 廣瀬慧 (2016), スパースモデリングとモデル選択, 電子情報通信学会誌 Vol. 99, No. 5, 2016 , 392-398
8. 渡部亮, Elastic netによるスパースロジスティック回帰と判別, 中央大学大学院研究年報理工学研究科篇, 第45号/2015

謝辞

本発表を纏めるにあたり、数々のご助言をいただきました以下のメンバーに感謝申し上げます

エイツーヘルスケア株式会社

開発戦略本部　生物統計部

畑山 知慶氏、渡部 亮氏