数理システム　学生研究奨励賞　応募研究

スパース多変量重回帰分析の新たな方法の開発

山下直人
東京工業大学　社会理工学研究科　修士2年

本研究の要旨

これまでの多変量重回帰分析

複数の**従属変数**の個人差を，複数の**独立変数**によって説明・予測する方法．たとえば，入社後の成績(営業成績・事務能力…)を，高校時代の成績(国語・数学…)で予測するなど．

解は

(**従属変数の個数**) × (**独立変数の個数**)の**偏回帰係数行列**

	営業	事務	交渉
国	0.8	0.2	0.5
英	0.2	0.4	−0.4
数	0.9	0.5	0.9
理	0.4	0.7	−0.5
社	0.9	0.5	0.7

(解の一例)

特に，**変数の数が多いとき**，解釈が困難

例えば従属変数100個，独立変数200個なら

解釈する要素は100 × 200 = **20000個**！

数字を1つ1つ読む途方も無い作業．

解釈の容易な解が得られるような多変量重回帰分析の開発

本研究の要旨

開発手法:スパース多変量重回帰分析

本研究では**スパース**な(= 解に0の要素が多い), 解釈のしやすい解を得る方法を新たに開発する.
開発手法では, **解釈すべき要素の数を抑えることができる**.

0.0でない要素だけ読めば良い. 既存手法では5×3 = 15個あった.

	営業	事務	交渉
国	1.8	0.0	0.0
英	1.4	0.0	0.0
数	0.0	1.3	0.0
理	0.0	0.0	−0.8
社	0.0	0.0	1.2

(得られる解の一例)

得られる解では
独立変数(行)と
従属変数(列)が
一対一に対応

→ **解釈**が非常に容易になる

統計手法を「作る」研究
→SPLUSは主に数値計算に利用

数値シミュレーションと適用例で
適切な動作を確認済み
+ 拡張的方法も開発

本発表の構成

1. はじめに
2. 開発手法
3. 数値シミュレーションと適用例
4. 開発手法の拡張
5. 要旨と今後の課題

1. はじめに

1.1 多変量重回帰分析とは?

1.2 解の「スパースさ」の追求

1.3 既存手法の紹介

1.4 本研究の目的

1.1 多変量重回帰分析とは？

従属変数(Q個)の個人差を，**独立変数**(P個)によって説明するための方法　→ 従属変数が複数の場合の，重回帰分析.

$$\min.\ f(\mathbf{W}) = \|\underset{N\times Q}{\mathbf{Y}} - \underset{N\times P}{\mathbf{X}}\underset{P\times Q}{\mathbf{W}}\|^2 \text{ over } \mathbf{W} \quad (1)$$

ここで

- $\mathbf{X}$: N(個体)×P(変数) の**独立変数**行列
- $\mathbf{Y}$: N(個体)×Q(変数) の**従属変数**行列
- $\mathbf{W}$: P(変数)×Q(変数) の**偏回帰係数**行列

解(最小二乗解)は

$$\mathbf{W} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2)$$

この行列の要素をすべて解釈するのは大変！

1.2 解の「スパースさ」の追求

特に変数が増えると，解釈すべき要素の数が爆発的に増えてしまい，それらすべての**解釈は困難**となってしまう．

もしも，**偏回帰係数行列の要素の幾つかが0ならば**，解釈すべき要素の数が**大幅に減る**ため，**解釈が容易**になる．

研究のモチベーション

スパース(sparse)な解が得られる
多変量重回帰分析

※スパースさ　＝　**0要素の多さのこと**

1.3 既存手法の紹介

スパースな解を得るための方法として，**ペナルティ項**を加えた多変量重回帰分析がいくつか開発されている(Izenman, 2008).

$$f_{\mathrm{SMR}}(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|^2 + \lambda(\text{ペナルティ項}) \text{ の最小化} \quad (3)$$

λ：ペナルティの強さを決めるパラメータ

ペナルティ項を加えることで，**値の小さな要素を0に近づける方法**．ペナルティ項の種類によって，様々な方法が開発されている．例えばLASSO(Tibshirani, 1996b)など．

ただし，λの値をどう決定するか？など，問題は多い．

1.4 本研究の目的

そこで本研究では

ペナルティパラメータ等の設定を必要としない
新たなスパース多変量重回帰分析の開発

および

開発手法の有用性の評価・検証

を目的とする.

2. 開発手法

2.1 開発手法

2.2 解の推定アルゴリズム

2.3 アルゴリズム：値ステップ

2.4 アルゴリズム：位置ステップ

2.5 局所解を避けるための方法

2.1 開発手法

本研究では，以下で定式化される，スパース多変量重回帰分析を開発する．

$$f(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|^2 \text{ を} \mathbf{W} \text{で最小化} \qquad (4)$$

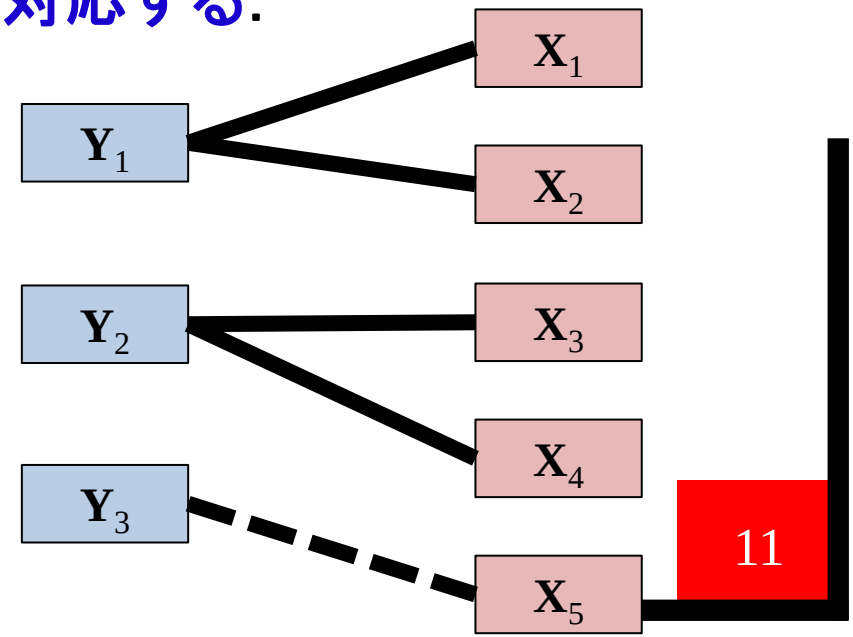
制約：**W**の各行において，非ゼロ要素がただ1つ存在し，それ以外の要素は0に等しい

つまり，**X**(行)と**Y** (列)が**一対一に対応する**．

1.8	0.0	0.0
1.4	0.0	0.0
0.0	1.3	0.0
0.0	0.0	−0.8
0.0	0.0	1.2

(得られる解の一例)

パス図表現



2.2 解の推定アルゴリズム

開発手法の数学的定式化

$f(\mathbf{W}) = \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|^2$ を$\mathbf{W}$で最小化

制約:$\mathbf{W}$の各行において，非ゼロ要素がただ1つ存在し，
それ以外の要素は0に等しい

要は，非ゼロ要素の**位置**と，その**値**を推定すれば良い．

推定すべきパラメータは

各行で非ゼロ要素が位置する**列番号**：$k(1), k(2),\ldots,k(P)$

非ゼロ要素の**値**：$w_{1,k(1)}, w_{2,k(2)},\cdots,w_{P,k(P)}$

2.2 解の推定アルゴリズム

パラメータの推定には，以下の繰り返しアルゴリズムを用いる．

SPLUSでは，for関数，if関数で簡単に実装可能
SMR関数として，SPLUSで動作確認済．

2.3 アルゴリズム：値ステップ

値ステップ：$w_{1,k(1)},\cdots,w_{p,k(p)},\cdots,w_{P,k(P)}$ の更新

$f(\mathbf{W})$を$w_{p,k(p)}$について偏微分し，それを0とおくと

$$
\begin{aligned}
& x_{1p}\{y_{1k(p)} - (x_{1p}w_{p,k(p)} + \textstyle\sum_{m\neq p}(x_{1m}w_{m,k(m)}))\} \qquad (6)\\
& \quad +\ldots+ x_{Np}\{y_{Nk(p)} - (x_{Np}w_{p,k(p)} + \textstyle\sum_{m\neq p}(x_{Nm}w_{m,k(m)}))\} = 0\\
\Leftrightarrow\ & \mathbf{x}_i'\mathbf{y}_{k(p)} - \|\mathbf{x}_i\|^2 w_{p,k(p)} + \textstyle\sum_n\sum_{m\neq p}(x_{nm}w_{m,k(m)}) = 0
\end{aligned}
$$

が得られる．従って，$w_{p,k(p)}$の更新式は

$$
w_{p,k(p)} = \left\{\mathbf{x}_i'\mathbf{y}_{k(p)} - \sum_{n}^{N}\sum_{m\neq p}^{P}(x_{nm}w_{m,k(m)})\right\} \Big/ \|\mathbf{x}_i\|^2 \qquad (7)
$$

で与えられる．これをすべてのpについて繰り返す．

SPLUSでは，for関数，sum関数で実装可能

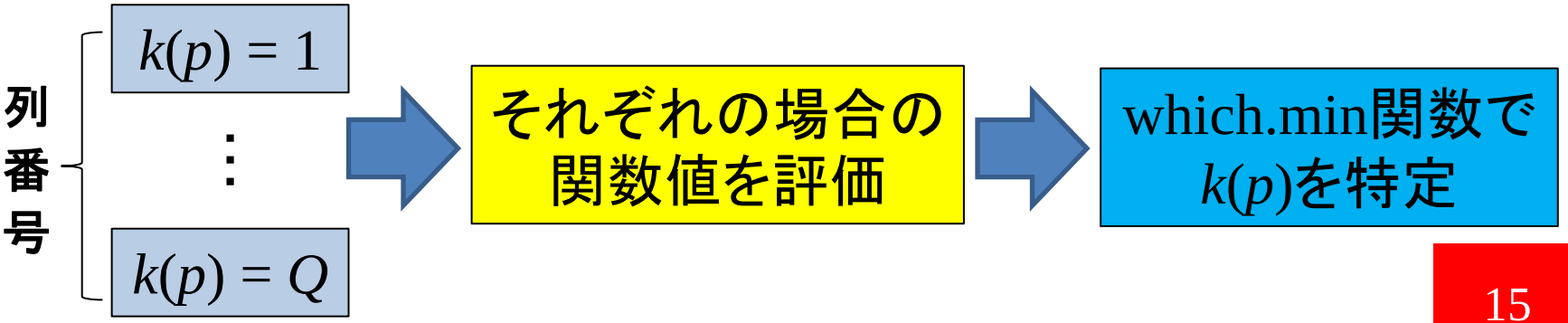
2.4 アルゴリズム：位置ステップ

位置ステップ：$k(1),\ldots,k(p),\ldots,k(P)$ の更新

Wの第$(p, k(p))$要素を(5)の$w_{p,k(p)}$で置き換えた行列を$\mathbf{W}^{\#}(p, k(p))$と表せば，$k(p)$は次のように更新できる．

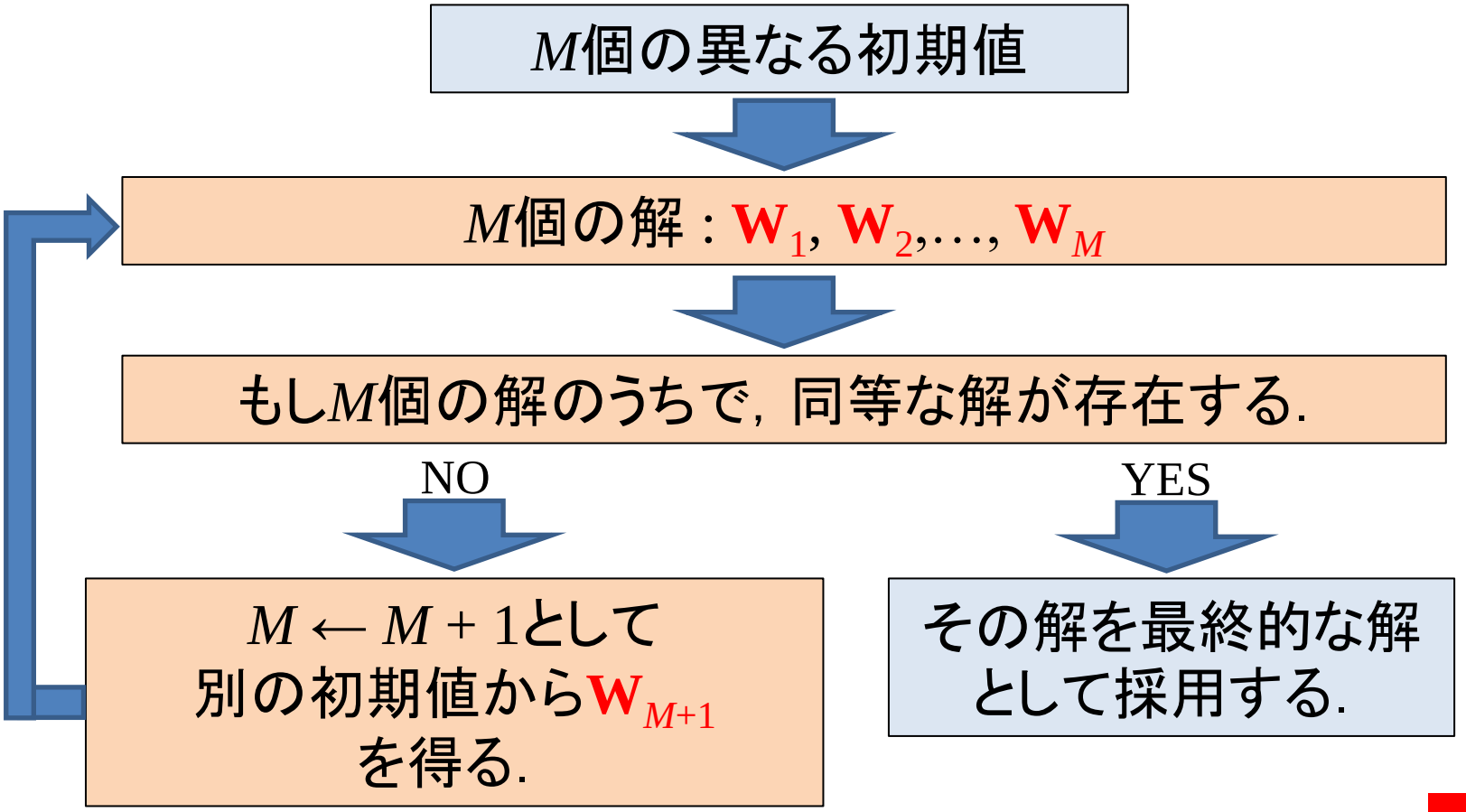
$$k(p) = \underset{k(p)=1,\ldots Q}{\arg\min}\, f(\mathbf{W}^{\#}(p, k(p))) \qquad (8)$$

つまり，第p行で，すべての列番号を試し，その中で最も目的関数を小さくする，「**総当り**」によって位置を推定する．



2.5 局所解を避けるための方法

開発手法では，**W**が多くの0を含むため，局所解(不適当な解)が頻発することが予想される．そこで，次の方法を用いた．



2つの同等な解が得られるまで，反復を繰り返す方法．

3. 数値シミュレーションと適用例

3.1 シミュレーションのデザイン

3.2 結果[1] 解の復元度

3.3 結果[2] 局所解の頻度

3.4 実データへの適用例

3.1 シミュレーションのデザイン

まず，解**W**の真値$\mathbf{W}_T$，独立変数行列**X**，誤差行列**E**をランダムに生成した上で，(9)で**Y**を構成した．

rnorm関数でランダムに生成

$$\underset{30\times 8}{\mathbf{Y}} = \underset{30\times 10}{\mathbf{X}}\underset{\times 8}{\mathbf{W}_T} + \theta(\rho)\underset{30\times 8}{\mathbf{E}} \qquad (9)$$

誤差の大きさを決める関数(Adachi, 2011)　**ρは分散説明率(0 %~ 100%)**
ρ = 100%のとき，データに誤差が含まれない．

つまり，**答え**($\mathbf{W}_T$)からデータ(**X**と**Y**)を発生させる．**X**，**Y**に開発手法を適用し，得られた解**W**が$\mathbf{W}_T$と十分近ければ，開発手法はうまくいっていると判断して良い．

W(推定値)と$\mathbf{W}_T$(真値)の差異を評価

3.1 シミュレーションのデザイン

結果の評価では，次の指標を用いた．

指標1：非ゼロ要素の「値の隔たり」(≥ 0)

Wの非ゼロ要素 ($\omega = W1$) ⇔ **W_Tの非ゼロ要素 ($\omega_T = W_T 1$)** ⇒ $\|\omega - \omega_T\|^2$ (10)

値の隔たり(二乗誤差)

指標2：ゼロ要素の「位置の類似度」(0 ~ 1)

Wの非ゼロ要素の位置 ($W^{\#}$) ⇔ **W_Tの非ゼロ要素の位置 ($W_T{}^{\#}$)** ⇒ $\mathrm{tr}W^{\#}W_T{}^{\#\prime}/P$ (11)

位置の類似度
(位置が一致する割合)

SPLUSでは簡単な行列演算で計算可能

3.1 シミュレーションのデザイン

分散説明率(ρ)として

100% 誤差無し　　**85% 誤差中**　　**0.75% 誤差大**

の3水準を設け，各水準で100個の真値W_TとデータX, Yを生成して，それぞれのデータに開発手法を適用すれば，各水準で100組の(W，W_T)が得られる．そこで，各組に対して

得られたW(推定値)とW_T(真値)の差異を
「値の隔たり」と「位置の類似度」で評価
→**真値の再現度**を確かめる

⇒ 各水準での平均・パーセンタイル値を計算

＋

推定で局所解が発生した頻度を数える
→**計算効率の良さ**を評価する

⇒ 各水準での平均を計算

この手続きにより，開発手法の有用性を見極めることができる．

3.2 結果[1] 解の復元度

各水準における，評価指標のパーセンタイル値，平均値は次の通り.

		ρ(分散説明率)		
		100%	85%	70%
値の隔たり 二乗誤差	25%	0.00	0.00	0.01
	50%	0.00	0.01	0.05
	75%	0.00	0.04	0.14
	平均	**0.00**	**0.05**	**0.13**
位置の再現度 0 ~ 1(完全再現)	25%	1.00	0.86	0.71
	50%	1.00	0.86	0.86
	75%	1.00	1.00	0.86
	平均	**1.00**	**0.89**	**0.81**

値の隔たり : **ほぼゼロに近い**
位置の再現度 : **ほぼ完璧に近い**
解の高い再現度が確認された

3.3 結果[2] 局所解の頻度

各水準での計算で局所解発生数の平均値(上限10回)は次の通り.

		ρ(分散説明率)		
		100%	85%	70%
局所解発生数	平均	0.00回	0.39回	0.42回

局所解の頻度は，誤差が大きくなるにつれて，増加してゆくものの，解の再現度は極めて良好.

局所解の発生は実用上問題なし

3.4 実データへの適用例

タバコの葉の化学組成に関するデータ(Izenman, 2008)に対して，開発手法を適用して得られるスパース解，および，通常の最小二乗解は次の通り．

glarsパッケージを利用

	開発手法			最小二乗解			LASSO		
従属変数	燃焼速度	砂糖	ニコチン	燃焼速度	砂糖	ニコチン	燃焼速度	砂糖	ニコチン
独立変数	解釈しやすい			解釈しづらい			解釈しづらい		
窒素	0.000	**−0.686**	0.000	0.103	−0.580	0.290	0.000	0.000	0.339
塩素	**−0.611**	0.000	0.000	−0.581	0.390	−0.322	−0.410	−0.437	0.000
カリウム	**0.624**	0.000	0.000	0.449	0.198	0.106	0.252	0.208	0.000
リン	0.000	**0.167**	0.000	−0.129	0.222	−0.045	−0.053	0.000	0.000
カルシウム	**0.330**	0.000	0.000	0.408	0.106	0.243	0.000	0.000	0.013
マグネシウム	0.000	0.000	**0.734**	−0.324	−0.215	0.483	0.000	−0.092	0.248

解釈容易な解が確かに得られた

4. 開発手法の拡張

4.1 スパース主成分分析

4.2 スパース主成分分析の適用例

4.1 スパース主成分分析

スパース主成分分析：**スパースな主成分負荷量行列(A)**を得る方法．近年開発が盛ん(Jolliffe, et.al., 2003; Shen & Huang, 2008; Zou, et.al., 2006)．一般的な定式化として

$$\underset{N\times P}{\mathbf{Z}} - \underset{N\times r\times P}{\mathbf{FA'}}\|^2 + \gamma(\text{ペナルティ項}) \text{ の最小化} \quad (12)$$

LASSOのように罰則項をつけたものが多い．ただ，ペナルティパラメータγの設定の問題がある．そこで，開発手法を拡張して，スパースなAを求める方法を考える．ただし制約を少しだけ修正することが必要．

制約条件

Aの各行において，非ゼロ要素がただ1つ存在し，それ以外の$Q-1$個の要素は0に等しい (13)

↓

A′の各列において，非ゼロ要素がただ1つ存在し，それ以外の$P-1$個の要素は0に等しい

4.1 スパース主成分分析

まず，**F**に関する最小化は，**ZA**の特異値分解**ZA** = **UΛV′**により

$$\mathbf{F} = N^{-1/2}\mathbf{UV}' \quad (14)$$

← svd関数が利用可

さらに，**A**に関する最小化は，

独立変数行列を$\mathbf{F}$**，従属変数行列を**$\mathbf{Z}$

としたうえで，**最適化アルゴリズムにおいて行と列の役割を入れ替えた**開発手法と同等であるから，そのような改変を施した開発手法を適用すれば，スパースな**A**を得ることができる．

$$\|\mathbf{Z} - \mathbf{FA}'\|^2 \quad (15)$$

← 開発手法(SMR関数)の改造版

従属変数　　独立変数

以上の2ステップを収束まで繰り返せば良い．

SPLUSには，SPCA関数として実装した．

4.2 スパース主成分分析の適用例

先ほどのスパースPCAを，野球の成績に関するデータ(足立，2006)に適用すると，次のような結果が得られた．

	スパース主成分分析			通常の主成分分析			右のバリマックス回転後		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
打率	0.000	0.000	**−0.818**	0.429	**0.650**	**−0.524**	0.330	0.151	**−0.866**
本塁打	**0.974**	0.000	0.000	**0.957**	−0.051	0.040	**0.951**	0.041	−0.039
打点	**0.976**	0.000	0.000	**0.965**	−0.059	−0.084	**0.961**	−0.048	−0.126
得点	0.000	**0.809**	0.000	**0.543**	**0.680**	0.407	0.481	**0.805**	−0.210
三振	0.000	0.000	**0.775**	**0.527**	−0.472	**0.609**	**0.603**	0.090	**0.708**
盗塁	0.000	**0.810**	0.000	−0.481	**0.610**	**0.571**	**−0.521**	**0.809**	0.068

スパース主成分分析で得られた解は
各変数が単一の主成分と対応して
非常に解釈がしやすい．

5. 要旨と今後の課題

5.1 本研究の要旨

5.2 今後の課題

5.3 SPLUSの利点

5.1 本研究の要旨

本研究では

スパースで解釈のしやすい偏回帰係数行列を推定でき
なおかつ**ペナルティパラメータの設定も不要**な
スパース多変量重回帰分析の新たなアルゴリズム と
そのSPLUS用関数(SMR関数)を開発

数値シミュレーション，および，実データへの適用例から

[1] 開発手法により真のパラメータが**復元できる**こと.
[2] 実データに対しても，**解釈のしやすい解**が得られること

が確認された．また，開発手法の応用例として**スパース主成分分析**を開発.

5.2 今後の課題

今後の課題として

1) スパース多変量重回帰分析について

- 罰則化法を主とした既存手法との比較
- より多彩な実データへの適用

2) スパース主成分分析について

- 既存アルゴリズムとの比較

3) 様々な手法への拡張

- 多変量重回帰分析は，様々な多変量解析法を包含
- 他の手法の解の「スパース化」は極めて容易

5.3 SPLUSの利点

SPLUSを使ってみた感想として

1) 行列演算との親和性

- 導出を行列で行えば，簡単に実装できる.
- 「行列使い」にうってつけの計算ツール.

2) 多くのパッケージが利用できる

- LASSO等，既存のアルゴリズムを気軽に利用できる.
- 若干の手直しが必要だが，Rのコードも併用可.

3) 豊富なGUI

- グラフィックスの出力・一般的な解析が簡単.
- 主成分分析・因子分析も簡単に実行できる.
- 今後はそのような機能を積極的に使っていきたい.
  (たとえば，スパース解の図的表示など)

References

Adachi, K. (2009). Joint Procrustes analysis for simultaneous nonsingular transformation of component score and loading matrices. *Psychometrika*, **74**, 667-683.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer, New York.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**, 531-547.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**, 1015 – 1034.

Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

Zou, H., Hastie, T., & Tibshirani, R (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15, 265-286.

足立浩平 (2006). *多変量データ解析法 ―心理・教育・社会系のための入門―*. ナカニシヤ出版.