研究論文

部分的最小二乗法における質的データの扱い

西山ちとせ（岡山理科大学大学院総合情報研究科）
片山浩子（岡山理科大学留学生別科）
森　裕一（岡山理科大学）

要約：

分析したい目的が量的データに適用する手法で実現されるような場合に，質的変数が混在したデータをどう扱うかを検討する。分析手法として，部分的最小二乗法（PLS：Partial least squares）を適用する場面をあげ，質的変数の数量化には最適尺度法を用いる。数量化では，説明変数のみを用いて質的変数を数量化して PLS に適用する方法と目的変数も含めたデータ全体から質的変数を数量化して PLS に適用する方法の 2 つを提案し，数値例により，その性能を評価する。その結果，数量化データを用いて PLS が実行できること，および数量化を行ってから PLS を適用する方が予測効率がよくなることを確認した。

キーワード：

質的データ，部分的最小二乗法，最適尺度法，PRINCIPALS，数量化

１．はじめに

マーケティングや教育の分野で扱う検査や調査において，量的データと質的データが混在したデータが得られた場合，分析の目的が回帰分析のような量的データに適用する手法により達成される場合，質的データ部分の扱いに困る。たとえば，選択肢番号の 1,…,4 をそのまま数値として分析にかけることが行われるが，実際には，選択肢間の距離が等間隔であるという保証はない。したがって，その距離を正しく推定するためには，数量化を行う必要がある。すなわち，質的データを収集した場合でも分析や処理ができるようにすることを考える。

質的データを数量化する手法に交互最小二乗法（ALS：Alternating Least Squires）を利用した最適尺度法がある。主成分分析の文脈では，PRINCIPALS（Young et al，1978）や

PRINCALS（Gifi，1990）といったアルゴリズムが提案されている。これを利用することで，質的データの問題が解決できる可能性が出てくる。そこで，この数量化を，個体数よりも調査項目数の方が多い場合や多重共線性がある場合に適用できる 1966 年に Herman Wold によって開発された部分的最小二乗法（PLS：Partial least squares，Wold，1966）への統合を考える。PLS は，その計算において，主成分分析の考えが含まれており，主成分分析を背景とする PRINCIPALS や PRINCALS を使用する適切性も示唆される。その評価には，質的データのままで PLS による予測を行う場合と，数量化してから PLS で予測を行うことを比較し，性能を確認する。

以下，2 節で PLS と最適尺度法のアルゴリズム PRINCIPALS を概観した後，3 節で PLS に PRINCIPALS を統合した手法を述べる。4 節では，数値例により，数量化してから PLS を適用する方が予測効率がよいことを示す。そこでは，説明変数のみを用いて質的データを数量化する方法（Type1）と目的変数（量的データ）も含めて数量化する方法（Type2）の 2 つを検討する。最後に結論を述べる。

２．部分的最小二乗法と最適尺度法のアルゴリズム

２．１．部分的最小二乗法

PLS は，説明変数 $\mathbf{X}$ の主成分 $\mathbf{t}$ と，目的変数 $\mathbf{Y}$ との関連性が大きい主成分 $\mathbf{t}$ を抽出することである（Wold，1966）。PLS が使用される分野として，計量化学，教育学，心理学，経済学，マーケティング分野などがあり，説明変数と目的変数の両方から推定（線形結合）される変数（潜在変数）で考察したい場合に役立ち，可視化，変数の考察，多重共線性に対応できることなどが特徴である。PLS には，部分的最小二乗回帰（PLS-R：Partial least squares regression），部分的最小二乗構造方程式モデル（PLS-SEM：Partial least squares - Structural Equation Modeling）などがある。ここでは，PLS-R に注目する。

PLS-R は，モデルを次のように設定する（Vinzi and Russolillo，2013）。

$$\begin{aligned}\mathbf{X}&=\mathbf{T}_H\mathbf{P}^T{}_H+\mathbf{E}_H\\ \mathbf{Y}&=\mathbf{T}_H\mathbf{C}^T{}_H+\mathbf{F}_H\end{aligned} \qquad (1)$$

$\mathbf{X}$ は $n\times p$ 予測行列（説明変数行列），$\mathbf{Y}$ は $n\times r$ 応答行列（目的変数行列），$\mathbf{T}_H$ は主成分行列，$\mathbf{P}_H$ と $\mathbf{C}_H$ は負荷行列，$\mathbf{E}_H$ と $\mathbf{F}_H$ は，それぞれ残差行列，H は主成分数を示す（$1 \leq H \leq \min(p, r)$）。

PLS-R のアルゴリズムは，成分ごとに反復ループが実行される。まず，初期値 $\mathbf{E}_0=\mathbf{X}$,

$\mathbf{F}_0=\mathbf{Y}$ とする。$\mathbf{W}$ を $\mathbf{X}$ の重みベクトル，$\mathbf{C}$ を $\mathbf{Y}$ の負荷量，$\mathbf{T}$ を主成分，$\mathbf{U}$ を $\mathbf{Y}$ の成分，$\mathbf{P}$ を $\mathbf{X}$ の負荷量を示す各行列とすると，PLS-R は次のように計算を行う（Wold，1966）。

Step0：$\mathbf{Y}$ の最初の成分 $\mathbf{u}_1$ の初期値を指定する。

Step1：h=1,…,H に対して $\mathbf{w}_h$ が収束するまで以下を繰り返す。

Step1.1　$\mathbf{w}_h=\mathbf{E}^{\mathrm{T}}{}_{h-1}\,\mathbf{u}_h / \|\mathbf{E}^{\mathrm{T}}{}_{h-1}\,\mathbf{u}_h\|$

Step1.2　$\mathbf{t}_h=\mathbf{E}_{h-1}\,\mathbf{w}_h / (\mathbf{w}^{\mathrm{T}}{}_h\mathbf{w}_h)$

Step1.3　$\mathbf{c}_h=\mathbf{F}_{h-1}\,\mathbf{t}_h / (\mathbf{t}^{\mathrm{T}}{}_h\mathbf{t}_h)$

Step1.4　$\mathbf{u}_h=\mathbf{F}_{h-1}\,\mathbf{c}_h / (\mathbf{c}^{\mathrm{T}}{}_h\mathbf{c}_h)$

Step2～Step4：$\mathbf{Y}$，$\mathbf{X}$ から求まる $\mathbf{t}$ の値が等しくなるまで繰り返す。

Step2　$\mathbf{p}_h=\mathbf{E}^{\mathrm{T}}{}_{h-1}\mathbf{t}_h/(\mathbf{t}^{\mathrm{T}}{}_h\mathbf{t}_h)$

Step3　$\mathbf{E}_h=\mathbf{E}_{h-1}-\mathbf{t}_h\mathbf{p}^{\mathrm{T}}{}_h$

Step4　$\mathbf{F}_h=\mathbf{F}_{h-1}-\mathbf{t}_h\mathbf{c}^{\mathrm{T}}{}_h$

これによって主成分 $\mathbf{t}$ が求められる。

２．２．最適尺度法

最適尺度法とは，非計量データのカテゴリースコアを求めることで，質的データを数量化する手法である。非線形主成分分析（NLPCA：Nonlinear Principal Component Analysis）の中で使われており，そのアルゴリズムとして，PRINCIPALS（Young et al，1978）や PRINCALS（Gifi，1990）がある。ここでは，PRINCIPALS に注目する。

PRINCIPALS は，交互最小二乗法に基づいており，そのアルゴリズムは，以下の通りである（森 他，2017）。

n 個体×p 変数のデータ行列 $\mathbf{X}$（$\mathbf{Q}$）を陽に表現するため，新たなベクトル $\mathbf{q}_j$ と行列 $\mathbf{G}_j$ を導入する。$\mathbf{Q}_j$ は j 番目の変数のカテゴリースコアベクトルで，$\mathbf{G}_j$ は j 番目の変数に対するダミー変数から構成される指標行列である。$\mathbf{Z}$ を主成分行列，$\mathbf{a}_j$ を j 番目の固有ベクトルとすると，目的関数

$$\mathrm{LS}(\mathbf{X}(\mathbf{Q}),\mathbf{M}(\boldsymbol{\Theta}))=\sum_{j=1}^{p}\left\|\mathbf{G}_j\mathbf{q}_j-\mathbf{Z}\mathbf{a}_j\right\|^2 \tag{2}$$

を最小化する $\mathbf{q}_j$ を求めるために ALS を適用する。解を一意に定めるために，

$$\frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}=\mathbf{I}_h,\ \mathbf{A}^{\mathrm{T}}\mathbf{A}\text{ は対角要素が降順の対角行列} \tag{3}$$

$$\mathbf{1}_n^{\mathrm{T}}\mathbf{G}_j\mathbf{q}_j=0,\ \frac{1}{n}\mathbf{q}_j^{\mathrm{T}}\mathbf{G}_j^{\mathrm{T}}\boldsymbol{q}_j=1\quad(j=1,\cdots,\mathrm{p}) \tag{4}$$

を制約条件として与える。$\mathbf{I}_h$は，$h\times h$ 単位行列であり，$\mathbf{1}_n$は，要素がすべて **1** の $n\times\mathbf{1}$ ベクトルである。

$\mathbf{Q}^{[t]}$，$\mathbf{Z}^{[t]}$，$\mathbf{A}^{[t]}$などを反復計算における t 回目の推定値とするとき，まず，**Z** と **A** の初期値 $\mathbf{Z}^{[0]}$と $\mathbf{A}^{[0]}$を設定する。このとき，PRINCIPALS は次の 2 つのステップを収束するまで繰り返す。

［Q ステップ］制約条件（3）式のもとで，$\mathbf{Q}^{[t+1]}=[\mathbf{q}_1^{[t+1]},\cdots,\mathbf{q}_p^{[t+1]}]$を

$$\mathbf{q}_j^{[t+1]}=\left(\mathbf{G}_j^{\mathrm{T}}\mathbf{G}_j\right)^{-1}\mathbf{G}_j^{\mathrm{T}}\mathbf{Z}^{[t]}\mathbf{a}_j^{[t]}$$

により計算する。ただし，変数 jのカテゴリーに順序制約があるとき，単調回帰法により $\mathbf{q}_j^{[t+1]}$を計算する。

［Θステップ］$\frac{1}{n}\mathbf{X}\left(\mathbf{Q}^{[t+1]}\right)^{\mathrm{T}}\mathbf{X}\left(\mathbf{Q}^{[t+1]}\right)$の固有値分解，あるいは$\mathbf{X}\left(\mathbf{Q}^{[t+1]}\right)^{\mathrm{T}}$の特異値分解により，$\mathbf{Z}^{[t+1]}$と$\mathbf{A}^{[t+1]}$を求める。

これにより最終的に得られたデータが数量化データである。

３．提案手法

以上より，PLS において，質的データを扱う手法を提案する。ここで，数量化の対象とする行列によって 2 つの方法 Type1 と Type2 が考えられる。ただし，**y** は目的変数（量的データ），**X** は説明変数（質的データを含む）のデータである。また **X*** は **X** の数量化されたデータとする。

（1）Type1

Type1 は，説明変数の行列に対して数量化を行ってから PLS-R を適用する方法である。つまり，**X** の質的データを数量化した **X*** を用いて分析する方法が Type1 である。

（2）Type2

Type2 は，説明変数と目的変数の両方を合わせた行列に数量化を行ってから PLS-R を適用する方法である。つまり，目的変数 **y** と説明変数 **X** の両方の情報を用いて数量化を行って得られた **X*** を用いて分析する方法が Type2 である。すなわち，**X** の数量化に **y** の情報を用いたかどうかが 2 つのタイプの違いである。

４．数値例

４．１．使用データと評価方法について

使用するデータは，文部科学省の全国学力・学習状況調査の調査結果の統計的性質を一

部保存した上で，集計表の統計量から乱数を発生させて作成した擬似データである（文部科学省，2015）。

全国学力・学習状況調査は，教科に関する学力調査と質問紙調査が行われている。学力調査は，国語，算数，理科の 3 教科で出題された A 問題，B 問題の成績（正答数）である。質問紙調査は，児童自身が学校での勉強や生活の様子について 4 件法（1：している，2：どちらかといえばしている，3：あまりしていない，4：全くしていない）で回答するもので，「朝食を毎日食べているか」や「学校の決まりを守っているか」などを質問している。また，学校質問紙では，学校長が学校の様子について，4 件法（1：その通りだと思う，2：どちらかといえばそう思う，3：どちらかといえばそう思わない，4：そう思わない）で回答するものである。擬似データは，このような内容でとられた調査結果を基に作成されている。このデータから，各教科の正答数を合計した総正答数（score）と，児童質問紙回答 5 問，学校質問紙回答 5 問を取り出したデータを用い，児童および学校質問紙回答の質的データ 10 問（Q1～Q10）から総正答数の予測を行うことを試みる。

以下では，このデータ 100 人分のデータを既知のデータとして PLS-R を行い，正答数のわかっている新たな 10 人の 10 質問紙の回答から，正答数を予測してみる。表 4.1 はその既知データである。表 4.2 が予測を行う 10 人のデータで，この Q1～Q10 を使って，その score を予測する。その際，以下の 3 つを実行し，その結果を比較する。

表 4.1　既知データ

児童番号	score	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	38	1	1	1	1	3	3	2	3	2	2
2	28	2	2	2	2	1	2	2	2	3	2
3	66	1	1	1	2	3	2	2	1	2	1
4	65	1	2	1	2	3	2	1	2	2	2
5	44	1	1	1	2	3	2	2	1	3	1
6	50	1	3	4	1	3	1	2	2	2	2
7	54	1	2	1	1	2	2	2	3	3	2
8	51	2	1	2	1	2	2	2	3	3	3
9	48	1	3	3	2	3	1	1	1	1	1
					・						
					・						
					・						
					・						
					・						
91	61	1	1	1	2	2	1	3	2	2	2
92	60	2	2	3	3	2	1	1	1	2	1
93	44	2	2	3	1	2	2	2	2	2	2
94	53	1	1	2	1	1	1	1	2	1	2
95	46	1	2	3	2	1	1	1	1	3	2
96	61	1	2	2	1	2	2	1	2	2	2
97	29	1	3	3	1	2	2	1	2	3	2
98	38	2	1	3	1	3	2	2	2	1	2
99	51	1	2	1	1	1	2	1	1	3	3
100	41	2	2	1	2	2	1	2	2	2	2

表 4.2　予測を行う 10 人分のデータ

児童番号	score	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
101	58	2	3	2	1	3	3	2	3	2	2
102	35	1	1	1	2	3	2	2	2	2	2
103	38	1	1	1	1	2	1	1	1	2	2
104	52	1	3	2	1	2	1	1	1	2	1
105	46	1	1	1	1	1	2	2	2	3	2
106	55	1	1	1	2	3	1	1	1	2	2
107	43	1	2	2	1	1	2	2	1	3	2
108	52	1	1	1	1	2	2	2	2	2	2
109	45	1	2	2	2	2	2	2	2	3	2
110	52	2	1	2	2	2	1	1	2	1	1

表 4.3　最適尺度化されたカテゴリースコア

Type1

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	-0.5673079	-0.5917914	-0.9254558	-0.3292346	-0.8241982	-0.9696602	-0.5327743	-0.8869627	-0.6672534	-1.0231223
2	-0.5873348	-0.5452766	-0.5644464	1.1231075	-0.7612506	-0.6520364	-0.0540313	-0.6835263	-0.8617743	-0.6093932
3	1.1546426	-0.3550446	0.1470265	-0.7938729	1.2848478	0.4110621	1.4205811	0.2788497	0.1974356	0.4631943
4		1.4921125	1.3428757		0.3006009	1.2106344	-0.8337755	1.2916392	1.3315920	1.1693211

Type2

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	-0.5520169	-0.6165673	-1.0627928	-0.2232932	-0.8503634	-0.9780857	-1.1744627	-0.9088254	-1.1741107	-1.2089955
2	-0.6023184	-0.5515667	-0.6026956	1.0927710	-0.8803839	-0.6356417	-0.4858100	-0.6807804	-0.4849876	-0.4129017
3	1.1543353	-0.3196339	0.5874855	-0.8694778	0.9104462	0.3962242	0.7736854	0.3219651	0.7590250	0.6488272
4		1.4877678	1.0780029		0.8203012	1.2175033	0.8865873	1.2676407	0.9000734	0.9730701

（a）質的データのままで PLS に適用し予測する

（b）Type1 の方法で分析し予測する

（c）Type2 の方法で分析し予測する

なお，新しいデータを適用する際，Q1～Q10 のカテゴリー番号を数量化する必要があるが，これには，100 人のデータで最適尺度化されたカテゴリースコア（表 4.3）に置き換えて予測を行う。

４．２．結　果

(a)質的変数のまま，(b)Type1，(c)Type2 の結果を表 4.4 に示す。score との差（右 3 列）を見ると，各児童において，(a)，(b)，(c)の 3 つの差のうち，(a)が最小となったのは 1 人，(b)が最小となったのは 2 人，(c)が最小となったのは 7 人であり，数量化の効果が見られる。また，(a)，(b)，(c)それぞれの差の二乗和（最下列）に注目すると，(a)より(b)，(c)の方が小さく，質的変数のままで予測を行うよりも，数量化してから，予測をす

表 4.4　分析結果

児童番号	score	予測結果			scoreとの差		
		(a)	(b)	(c)	(a)	(b)	(c)
101	58	47.40042	54.66702	55.71089	10.59958	3.33298	2.28911
102	35	52.59543	50.8996	50.38544	-17.59543	-15.89960	-15.38544
103	38	50.98628	50.45163	49.70157	-12.98628	-12.45163	-11.70157
104	52	49.08289	50.3378	51.91628	2.91711	1.66220	0.08372
105	46	50.16606	49.62298	46.55941	-4.16606	-3.62298	-0.55941
106	55	50.20145	50.29768	48.90605	4.79855	4.70232	6.09395
107	43	45.23465	47.20462	44.19467	-2.23465	-4.20462	-1.19467
108	52	53.38026	51.05355	51.18096	-1.38026	0.94645	0.81904
109	45	45.75614	44.44456	43.22948	-0.75614	0.55544	1.77052
110	52	54.06716	47.93806	54.49698	-2.06716	4.06194	-2.49698
scoreとの差の二乗和					651.2290665	492.3322989	427.8024174

る方が性能がよいことがわかる。(b)と(c)と比べると，(b)は，説明変数 **X** のみで数量化を行った **X***を用いて分析を行った結果であり，(c)は目的変数 **y** と説明変数 **X** の両方の情報を用いて数量化を行って得られた **X***を用いて分析を行った結果であるため，目的変数の情報も用いた方がより性能がよいと判断される。ただし，対象児童を変えると，質的変数のままで予測した場合の差が小さくなるケースもあり得ることから，今後，検討が必要と考える。

５．結　論

本稿では，部分的最小二乗回帰において，質的データが含まれたデータを分析できる手法を提案した。具体的には，最適尺度法のアルゴリズム（PRINCIPALS）を PLS-R に統合し，その数量化のパターンとして，説明変数の行列に対して数量化を行ってから PLS-R を適用する方法と目的変数と説明変数の両方を合わせた行列を用いて数量化を行ってから PLS-R を適用する方法の 2 つを提案した。これにより，数量化してから PLS-R を実行できること，数値例より，数量化を行ってから PLS-R を適用させる方が予測がよくなることが示唆された。

今後の課題としては，個体の数よりも項目数が多く，かつ，質的データが混在する場合の扱いができるようにすること，シミュレーションによる細かい性能の評価を行うことなどがあげられる。

参考文献

Esposito Vinzi, V.and Russolillo, G. (2013). Partial least squares algorithms and methods, ***Wiley Interdiscip Rev*: *Comoatatinal Stat***, **5**(1), 1-19.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Young, F.W., Takane, Y. and de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. ***Psychometrika,* 43**，279–281.

Wold, H. (1966). Non linear estimation by iterative least squares procedure, Research paper in Statistics: Festschift for J. Neyman. F. David.

森裕一・黒田正博・足立浩平（2017），最小二乗法・交互最小二乗法，共立出版.

文部科学省（2015），パブリックユースデータ（擬似データ），文部科学省 (https://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/sonota/1404609.htm).