SVDを用いた行列におけるデータ誤差の導出法について

01507350　東京都立工業高等専門学校　*保福 一郎　HOFUKU Ichiro
東京理科大学　大島 邦夫　OSHIMA Kunio

1　はじめに

任意の行列に対するQR分解，SVD(Singular Value Decomposition)等は，様々な目的のために適用されている．例えば情報検索分野でのこれらの行列の分解は，1つの単語に対する誤った情報を考慮にいれたデータベースを作成する場合等で適用されている[2]．その手法の主な概略は，与えられたデータベースを行列で表現し，その中の誤情報（または余情報）を削除させるためまずデータベースの役割を果たす行列を分解する．そしてその分解された行列間での操作により情報を損失させ，データベースの内容を変更させるというものである．本発表では，SVDを用いてデータとして与えられた行列を分解し，情報損失とその影響について言及する．与えられる行列は，著者らが今までに開発してきたランキング決定の際に適用される既約行列（評価行列）とする[3]．この行列を基に生成した解析結果と，情報損失として修正した行列との解析結果とを比較・検討するのである．

2　準　備

2.1　評価行列

本章では，与えられる評価行列の生成法及びその解析結果の解説を簡単に与える．

1つの集合 $\psi_A = \{a(1), a(2), \cdots, a(n)\}$ において行列 M_A を以下の規則にしたがって生成する．

$$\begin{cases} a[i,i] = 0.5, & (i = 1, 2, \cdots, n) \\ a[i,j] = a(j) \text{ に対する } a(i) \text{ の優位率}, & \\ \qquad (1 \leq i,\ j \leq n \ \ i \neq j) & \end{cases} \tag{2.1}$$

例として ψ_A をある教科の試験を受験した学生3名の集合

$$\psi_A = \{a(1), a(2), a(3)\}$$

とし，試験内容は設問数が3つ (X_1, X_2, X_3) であり，各学生の獲得得点数が表1であったとする．ここで集合 ψ_A の各要素の優位率の関係からその集合に対応する評価行列 M_A を求める．ここで学生 $a(i)$ が設問 X_l, $(l = 1, 2, 3)$ に対し獲得した点数を $s(a_i, x_l)$ とする．行列 M_A の成分 $a[i,j]$,（$1 \leq i, j \leq 3$）は $a(j)$ に対する $a(i)$ の優位率を表しているため，本適用例においては，

表 1: 学生と各設問との関係

	X_1	X_2	X_3	合計点
$a(1)$	3	1	4	8
$a(2)$	5	7	3	15
$a(3)$	4	7	2	13
平均点	4	5	3	

$$a[i,j] = \frac{1}{3}\sum_{l=1}^{3} \frac{s(a_i, x_l)}{s(a_i, x_l) + s(a_j, x_l)}, \quad (1 \leq i,\ j \leq 3,\ \ i \neq j)$$

とする．この手法により生成された評価行列 M_A は，

$$M_A = \begin{pmatrix} 0.5 & 0.357143 & 0.406746 \\ 0.642857 & 0.5 & 0.551852 \\ 0.357143 & 0.448148 & 0.5 \end{pmatrix}$$

となる．

行列 M_A は既約行列であることからPerron-Frobeniusの定理を適用することにより，絶対値最大の正の固有値 $\lambda(A)$ に対応した正の固有ベクトル $\boldsymbol{r}(A)$ が求められる．この手法により生成されたベクトルをランキングベクトルと呼ぶ．ランキングベクトルは以下の3つの特性(a)～(c)をもつ．

(a) ランキングベクトルは正ベクトルである．
(b) ランキングベクトルの各成分の大小は，その成分に値する力を表す．
(c) 勝率の高い要素に対し，高い優位率をもつ要素の方が，勝率の低い要素に対し，高い優位率をもつ要素より，ランキングベクトルの数値が高くなる．

2.2 SVD

本節では，SVD に関する簡単な解説を与える．

Singular Value Decomposition

行列 A を，$rank(A)=r$ の $m\times n$ 行列とする．一般性を失うことなく $m\geq n$ と仮定することができる．このとき，A の SVD は

$$A=U\Sigma V^T \tag{2.2}$$

で表され，$U^TU=I_m$，$V^TV=I_n$ であり $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_n)$，$\sigma_i>0,(1\leq i\leq r)$，$\sigma_j=0,(j\geq r+1)$ である．

ここで A の Frobenius Norm を $||A||_F$ で表すと，

$$\begin{aligned}||A||_F &= ||U\Sigma V^T||_F=||\Sigma V^T||_F\\ &= ||\Sigma||_F=\sqrt{\sum_{j=1}^{r}\sigma_j^2}\end{aligned}$$

となる．したがって A の k-階数近似を $A(k)$ とすれば，行列 A と $A(k)$ との差は，

$$\begin{aligned}||A-A(k)||_F &= \min_{\mathrm{rank}(X)\leq k}||A-X||_F \quad (2.3)\\ &= \sqrt{\sigma_{k+1}^2+\cdots+\sigma_r^2}\end{aligned}$$

となる．

3 解 析

本章では，第2章で生成した M_A の SVD を計算し，式 (2.3) についての M_A の k-階数近似 $M_A(k)$ を求める．

式 (2.2) の分解による行列 M_A の U, Σ, V^T は，

$$U=\begin{pmatrix}-0.510788 & -0.413787 & -0.753575\\ -0.683530 & ,-0.336172 & ,0.647901\\ -0.521424 & 0.846031 & -0.111123\end{pmatrix}$$

$$\Sigma=\begin{pmatrix}1.43727 & 0 & 0\\ 0 & 0.15296 & 0\\ 0 & 0 & 0.0067595\end{pmatrix}$$

$$V^T=\begin{pmatrix}-0.61299 & -0.52729 & -0.58839\\ -0.79008 & 0.41371 & 0.45236\\ 0.0048967 & 0.74217 & -0.67020\end{pmatrix}$$

となる．ここで，行列 M_A は正則であるため，k-階数近似は，$M_A(1),M_A(2),M_A(3)$ となる．これらの行列のランキングベクトルを $\boldsymbol{r}_A(k)$ とおくと，

表 2: ランキングベクトルの成分変化

	標本分散	情報エントロピー
$M_A(3)$	0.0088139	1.57233
$M_A(2)$	0.00872748	1.57246
$M_A(1)$	0.00937187	1.57157

$$\boldsymbol{r}_A(3)=(0.504667,0.679438,0.532612)^T$$
$$\boldsymbol{r}_A(2)=(0.505237,0.678997,0.532634)^T$$
$$\boldsymbol{r}_A(1)=(0.510788,0.683530,0.521424)^T$$

となり，その時の行列 M_A の情報損失量 $l(k)$ を

$$l(k)=\frac{||M_A-M_A(k)||_F}{||M_A||_F}$$

とすると，

$$l(3)=0,\quad l(2)=0.00468,\quad l(1)=0.106$$

となる．次に行列 $M_A(k)$ の情報損失量に対するランキングベクトルの成分の影響の度合いを調べる．成分の変化の度合いを調べる手法としては，各成分の標本分散及び $\boldsymbol{r}_A(k)$ を l_1-norm で正規化し，完全事象系として情報エントロピーを適用した2つの場合を考える．表2はその結果を表す．表2を参照すると，行列 M_A の情報損失によるランキングベクトルの成分同士の変化の度合いが少ないことが解る．

参考文献

[1] 保福一郎, 大島邦夫, 多群間の成績を考慮した混合ランキングの生成法, 2001 年度日本OR 学会　秋期研究発表会アブストラクト集, pp.241-242, 2001.

[2] Michael W. Berry and Zlatko Drmac, *Matrices, Vector Spaces, and Information Retrieval*, SIAM Review, Vol.41, No.2, pp.335-362

[3] 大島邦夫, 保福一郎, ランキングベクトルとウェイトを適用した試験結果におけるランキング法について, 日本応用数理学会論文誌, Vol. 6 (1996), No.1, pp.133-146.