D-045

アンケートデータ欠損補完を目的とした確率的テンソル主成分分析の利用に関する検討

A Study on Use of Probabilistic Principal Component Analysis for Imputation of Missing Values in Questionnaire Data

福田 智広 † 吉川 大弘 † 古橋 武 †
Tomohiro Fukuta Tomohiro Yoshikawa Takeshi Furuhashi

1 はじめに

近年，企業が顧客の需要や製品に対する評価を把握する方法の 1 つに，アンケートを用いた市場調査がある．得られたアンケートデータに，主成分分析などの多変量解析手法を用いて解析することで，販売戦略に役立てることができる．また，広く用いられているアンケート調査手法の一つに評定尺度法がある．評定尺度法では，複数の評価対象と複数の質問項目が用意され，回答者は各対象について，各質問項目に複数段階の評点を付けることで印象を表現する．この評定尺度法により得られたアンケートデータは，3 階のテンソルで表現できる．

しかし一方，特に紙面で行うアンケートでは，見落としなどにより，回答が未記入となる部分が存在する場合があり，そのような欠損を含むアンケートデータに対しては多変量解析手法が適用できないため，何らかの形で欠損を補完する必要がある．そこで本稿では，3 階のテンソル構造のアンケートデータに対して，確率的主成分分析をテンソル空間に拡張した，確率的テンソル主成分分析を用いた欠損補完手法を提案する．実際のアンケートデータに対して提案手法を適用し，その性能評価を行う．

2 提案手法

確率的主成分分析 (Probabilistic Principal Component Analysis : PPCA) は，主成分分析に確率的モデルを適用したものであり，データの潜在変数を用いることで，欠損を補完できる [1]．本稿では，この手法を 3 階のテンソルに拡張し，データの欠損を補完する手法を提案する．なお，以下で単に “テンソル” と表現したときは，すべて 3 階のテンソルを指す．

2.1 Tucker 分解

代表的なテンソル分解手法に，Tucker 分解 [2] がある．Tucker 分解により，テンソル構造のデータ $\underline{X} \in \mathbb{R}^{n\times m\times l}$ が与えられたとき，

$$\underline{X} = \underline{G} \times_1 U_1 \times_2 U_2 \times_3 U_3 \tag{1}$$

と分解できる．ここで，$\underline{G} \in \mathbb{R}^{k\times k\times k}$ はコアテンソルと呼ばれ，$U_1 \in \mathbb{R}^{n\times k}$，$U_2 \in \mathbb{R}^{m\times k}$，$U_3 \in \mathbb{R}^{l\times k}$ は射影行列と呼ばれる．また，$\times_n$ は n 番目のモードに対するテンソル積で，n-モード積と呼ばれる．Tucker 分解では，$U_n U_n^T = I$ という条件のもとで

$$\min_{\underline{G},U_1,U_2,U_3} \|\underline{X} - \underline{G} \times_1 U_1 \times_2 U_2 \times_3 U_3\| \tag{2}$$

を計算するものである．

2.2 提案手法

テンソル構造のデータ $\underline{X}$ において，評定尺度法によるアンケートデータでは，n が質問項目数，m が対象項目数，l が回答者数に対応している．そこで本稿では，1 番目のモードを質問モード，2 番目のモードを対象モード，3 番目のモードを回答者モードと呼ぶこととする．

また $\underline{X}$ のモード展開において，質問モード展開では質問項目 ×(対象項目 × 回答者) 行列 $X_{質}$ ができ，対象モード展開では対象項目 ×(回答者 × 質問項目) 行列 $X_{対}$，回答者モード展開では回答者 ×(質問項目 × 対象項目) 行列 $X_{回}$ ができる．提案手法では，質問および対象モード展開行列に PPCA を適用し，欠損データの補完を行う．

提案手法の手順を以下に述べる．

手順 1: モード展開した行列 X_n において，欠損部分がない行列 X_{nObs}(観測部分と呼ぶ) と欠損部分がある行列 X_{nMiss} に分ける．

$$X_n = [X_{nObs}, X_{nMiss}] \tag{3}$$

手順 2: 欠損部分がない行列 X_{nObs} に高階特異値分解 ($Higher\ Order\ Singular\ Value\ Decomposition$: $HOSVD$) を適用し，U_{nObs} を求める．

†名古屋大学

手順 3: 欠損部分がない行列から求めた U_{nObs} を用いて，コアテンソル $\underline{G}$ を式 (4) により計算する．ここで，回答者モードの射影行列は，回答者モードの欠損部分を各回答者の平均値で補完した行列を用いて求める．

$$\underline{G} = \underline{X_{Obs}} \times_1 U_{\text{質}\,Obs}^T \times_2 U_{\text{対}\,Obs}^T \times_3 U_{\text{回}}^T \quad (4)$$

手順 4: 手順 2 で求めた U_{nObs} を用いて，各モードにおける潜在変数 z を求める．

$$z = (U_{nObs}^T U_{nObs})^{-1} U_{nObs}^T X_{nMiss} \quad (5)$$

手順 5: 手順 4 で求めた z を用いて，補完値を求める．

$$X_{nMiss} = U_{nObs} z \quad (6)$$

手順 6: 手順 4，5 で補完した各モード展開行列を評点 (1〜5) に規格化したのち，$HOSVD$ を適用し，U_n を求める．

手順 7: 手順 2 で求めた U_3，$\underline{G}$ および手順 6 で求めた U_1，U_2 によりテンソル再構築を行うことで，欠損部分が補完されたテンソル $\underline{X_{imp}}$ が求まる．

$$\underline{X_{imp}} = \underline{G} \times_1 U_{\text{質}} \times_2 U_{\text{対}} \times_3 U_{\text{回}} \quad (7)$$

3 実験

3.1 実験方法

5 段階評定尺度法による実際のアンケートデータに対して，欠損部分以外の平均値を用いて補完する従来手法 [3]，CF(協調フィルタリング) 法 [4]，および提案手法を適用し，欠損補完に対する精度の比較を行った．ここでは欠損箇所はランダムに作り，欠損割合はデータ全体の 5%〜45% の 5% 刻みとした．評価指標には，真値を正しく補完できたかを示す正答率と，評点の傾向を捉えることができたかを示す評点傾向を用いた．

3.2 結果と考察

各手法による補完精度を図 1 に示す．提案手法は欠損率が低い場合において，他の手法よりも精度よく補完できていることがわかる．しかし，欠損率が 25% 以上において，最も精度が悪くなってしまっている．これは，欠損率が高くなるにしたがって，使用できる観測部分が減り，射影行列をうまく推定できなかったためであると考えられる．一方で，従来手法と CF 法の精度が欠損率に関わらず一定である理由として，これら 2 つの手法の補完値は，評点の平均値付近に偏る傾向があり，また用いたデータが，平均値付近の評点が多いという特徴があったためであると考えられる．

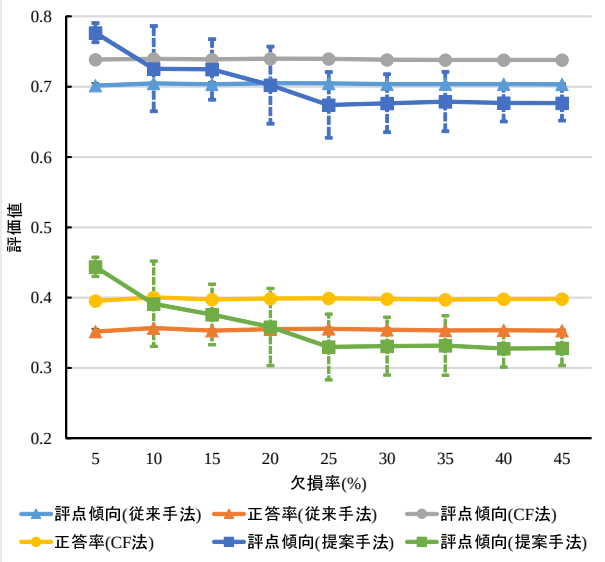

図 1: 補完精度

4 おわりに

本稿では，確率的主成分分析に基づく，3 階のテンソル構造のアンケートデータの欠損補完手法を提案した．実際のアンケートに適用し，低い欠損率においては，提案手法が従来手法よりも精度が高いことを示した．今後の課題として，欠損率が高い場合における有効な補完手法，および回答者間の類似性を考慮した欠損補完方法に対する検討などが挙げられる．

参考文献

[1] L Qu, J Hu, L Li, Y Zhang : PPCA-based missing data imputation for traffic flow volume: a systematical approach, IEEE Trans, Intelligent Transportation Systems, vol.10, pp.512-522, 2009.

[2] T.G.Kolda : Multilinear operators for higher-order decompositions. Technical Report SAND2006-2081, Sandia National Laboratories, Albuquerque, NM and Livermore 2006.

[3] I Myrtveit, E Stensrud, UH Olsson : Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, IEEE Trans, Software Engineering, vol.27, pp.999-1013, 2001.

[4] 神嶌 敏弘：推薦システムのアルゴリズム，人工知能学会誌，vol.22-23, 2007-2008.