3F3-1

確率的テンソル主成分分析を用いたアンケートデータの欠損補完に関する検討

A Study on Imputation of Questionnaire Data Using Probabilistic Principal Component Analysis

福田智広 Tomohiro Fukuta　吉川大弘 Tomohiro Yoshikawa　古橋武 Takeshi Furuhashi

名古屋大学
Nagoya University

Questionnaire is often carried out in order to design a marketing strategy by analyzing acquired data. However, there are often some missing values in a questionnaire. The missing data affects the analysis, because the multivariate analysis methods can be applied to only complete data. Thus, it is important to impute these missing values. The most common method in the imputation methods for these missing values is the mean imputation. However, it does not consider the feature of data. Another common method is the collaborative filtering which considers the feature of data, while it is affected by the characteristics of whole data. In order to get the characteristics of detailed data, this paper focuses on Probabilistic Principal Component Analysis (PPCA). This method is extended to three-order tensor data. This paper applies this method to actual questionnaire data and shows the accuracy comparing with the conventional methods.

1. はじめに

近年，企業が市場調査を通して，自社製品やサービスに対する顧客の需要や評価を把握することは，マーケティングにおいて重要である．販売戦略を立てるための市場調査の方法の 1 つにアンケート調査がある [柳澤 07]．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 [Osgood 57] 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．また，この方法で得られたアンケートデータは，図 1 のような 3 階のテンソルで表現できる．

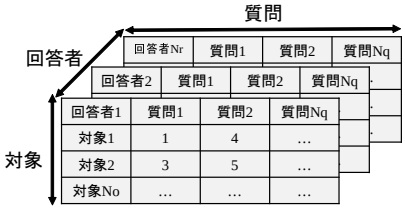

図 1: 評定尺度法によるアンケートデータ

しかし，データの中には，未記入などによって欠損部分が存在する場合がある．一般にアンケート解析に用いられる多変量解析手法では，完全データを想定しているため，欠損部分があるデータをそのまま利用できない．一方で，欠損を持つデータを除いて解析を行うことは，得られた情報の損失につながる．そのため，解析をする上で，何らかの形で欠損を補完する必要がある．この欠損補完では一般的に，平均値で補完する手法 [Myrtveit 01] が用いられる．しかしこの方法では，データ全体の特徴を用いるのみで，質問間，対象間および回答者間の特徴を考慮した補完をすることはできない．また，データ間の相関に基づく補完方法として，協調フィルタリングを用いた方法 [神嶌 07] があるが，やはりデータ全体の傾向に偏りやすいという問題がある．

そこで本稿では，確率的主成分分析 (Probabilistic Principal Component Analysis： PPCA)[Tipping 99] を用いた欠損補完を行う．PPCA は，データ間の特徴を潜在変数として考慮することができる．本稿では，PPCA をテンソルへ拡張することで，質問間，対象間および回答者間の特徴を考慮し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，従来手法および協調フィルタリングと比較して欠損補完の精度が高いことを示す．

連絡先: 福田 智広，名古屋大学，
〒 464-8603 名古屋市千種区不老町，
TEL: 052-789-2793，
fukuta@cmplx.cse.nagoya-u.ac.jp

2. 欠損補完手法

2.1 平均値補完

欠損箇所に対して，質問項目の平均値または回答者の平均値を挿入することで欠損値を補完する．以下では，質問項目の平均値を用いて補完する手法を従来法 (質)，回答者の平均値を用いて補完する手法を従来法 (回) とする．

2.2 協調フィルタリング

協調フィルタリング (Collaborative Filtering) とは，データの相関を利用した欠損補完手法である．本稿では，質問間の類似度を計算して補完値を算出する．以下に具体的な方法を示す．

回答者 × 質問項目のデータ行列を x とする．このとき回答者 a，質問項目 b の評点は x_{ab} と表せる．ここで回答者 n，質問項目 m が欠損しており，補完することを考える．まず，質問項目 m と m 以外の質問項目 a の類似度 $p_{a,m}$ を式 (1) で求める．

$$p_{a,m} = \frac{\sum_{k \in Y_{am}} (x_{ka} - \overline{x}'_a)(x_{km} - \overline{x}'_m)}{\sqrt{\sum_{k \in Y_{am}} (x_{ka} - \overline{x}'_a)^2} \sqrt{\sum_{k \in Y_{am}} (x_{mk} - \overline{x}'_m)^2}} \quad (1)$$

ここで，Y_{am} は二つの質問項目に共通に回答した回答者集合である．また $\overline{x}'_a = \sum_{k \in Y_{am}} x_{ka} / |Y_{am}|$ である．ただし，質問項目 m と質問項目 a に共通で回答した回答者が一人以下なら

ば，$p_{a,m} = 0$ とする．回答者 n，質問項目 m の補完値 $\hat{x}_{nm}$ は，式 (1) の類似度で重み付けした各質問項目の回答者 n への評点の平均で補完する．質問項目 m に評点をつけた回答者を R_m と表すと，補完値は式 (2) で求まる．

$$\hat{x}_{nm} = \overline{x}_m + \frac{\sum_{j \in R_m} p_{j,m}(x_{j,m} - \overline{x}'_j)}{\sum_{j \in R_m} |p_{j,m}|} \tag{2}$$

2.3 確率的主成分分析

確率的主成分分析 (PPCA) とは，主成分分析に確率的モデルを適応したものである．通常の主成分分析と比べて，データの欠損値を確率的に扱うことができ，工夫によりその補完ができるという利点がある．PPCA のモデル式を式 (3) に表す．

$$x = Wz \tag{3}$$

x は $D \times N$ のデータ行列 (評点)，z は $q \times N$ の潜在変数であり，ガウス分布に従う．ここで $q < D$ である．W は $D \times q$ の負荷量行列であり，最尤推定で求まる．

この PPCA を用いた欠損補完手法を説明する [Qu 09]．まず，データ x を欠損部分がない $D \times N_o$ 行列 x_{obs}(観測部) と欠損部分がある $D \times N_m$ 行列 x_{miss}(欠損部) に分ける．ここで，$N_o < N$，$N_m = N - N_o$ である．x_{obs} を用いて，観測部の負荷量行列 W_{obs} を最尤推定で求め，その後潜在変数 z を式 (4) により求める．

$$z = (W_{obs}^T W_{obs})^{-1} W_{obs}^T x_{miss} \tag{4}$$

ここで W_{obs} は $D \times q$ 行列，z は $q \times N_m$ 行列となる．観測部の潜在変数 z を用いて，欠損部の補完値 x_{imp} を式 (5) で計算する．

$$x_{imp} = W_{obs} z \tag{5}$$

2.4 確率的テンソル主成分分析

本稿では，2.3 で示した PPCA を 3 次元に拡張した確率的テンソル主成分分析を提案する．以降にその具体的方法について説明する．まずテンソル mode 展開は，3 階のテンソルを図 2 のように行列に展開して表現するものである．質問 mode 展開は，質問項目 ×(対象項目 × 回答者) 行列 $X_{質}$，対象 mode 展開は，対象項目 ×(回答者 × 質問項目) 行列 $X_{対}$，回答者 mode 展開は，回答者 ×(質問項目 × 対象項目) 行列 $X_{回}$ でそれぞれ表される．PPCA をテンソルへ拡張したモデルを式 (6) に示す [Timmerman 00]．

$$\underline{X} = \underline{Z} \times_{質} U_{質}{}^T \times_{対} U_{対}{}^T \times_{回} U_{回}{}^T \tag{6}$$

ここで，$\underline{X}$ はデータテンソル，$\underline{Z}$ はコアテンソル，U_n は n モードにおける射影行列である．$U_{質}$ は質問項目の特徴を，$U_{対}$ は評価対象の特徴を，$U_{回}$ は回答者の特徴を表す．また，$U_{質}$ の列 i ベクトルは質問の第 i 基底ベクトル，$U_{回}$ の列 j ベクトルは対象の第 j 基底ベクトルと呼び，列番号が小さいほど，データ $\underline{X}$ の特徴をより表している．基底ベクトルの大きさおよび符号が類似している項目は，類似した特徴を示す．$\times_n$ は n モード積を表す．U_n は X_n に対して式 (7) に示す高階特異値分解 (High Order Singular Value Decomposition: HOSVD) を行うことで計算される．

$$X_n = U_n \Sigma_n V_n \tag{7}$$

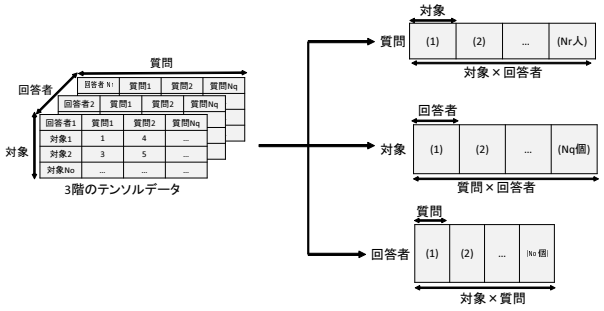

図 2: mode 展開

2.5 提案手法

2.5.1 質問と対象の特徴を考慮したデータ補完

2.4 で示したテンソルの mode 展開を用いて，質問と対象の特徴を考慮したデータ補完を行う手法を提案する．図 1 のようなアンケートデータを，質問 mode 展開すると質問項目 ×(対象項目 × 回答者) 行列 $X_{質}$ ができ，対象 mode 展開すると対象項目 ×(回答者 × 質問項目) 行列 $X_{対}$ ができる．この 2 つの mode で展開を行った行列を用いて，欠損を補完する手順について以下に述べる．

手順 1: mode 展開した行列 X_n において，欠損部分がない行列 X_{obs} と欠損部分がある行列 X_{miss} に分ける．

$$X_n = [X_{obs}, X_{miss}] \tag{8}$$

手順 2: 欠損部分がない行列 X_{obs} に HOSVD を適用し，U_{obs} を求める．

手順 3: 手順 2 で求めた U_{obs} を用いて，潜在変数 z を算出する．

$$z = (U_{obs}^T U_{obs})^{-1} U_{obs}^T X_{unobs} \tag{9}$$

手順 4: 手順 3 で求めた z を用いて，補完値を求める．

$$X_{miss} = W_{obs} z \tag{10}$$

手順 5: 手順 4 で求めた補完値を評点 (1〜5) に規格化を行う．具体的には，補完値の最大値 Imp_{max} と最小値 Imp_{min} を求め，式 (11) により，0.0〜5.0 に規格化を行う．

$$\tilde{I} = aI + b \tag{11}$$

ここで，$\tilde{I}$ は規格化後の補完値，I は規格化前の補完値を示し，a および b は，それぞれ式 (12)，(13) で求まる．

$$a = \frac{5}{Imp_{max} - Imp_{min}} \tag{12}$$

$$b = -\frac{5 \times Imp_{min}}{Imp_{max} - Imp_{min}} \tag{13}$$

手順 6: 手順 1〜5 によって，質問 mode および対象 mode で求めた補完値の平均を，切り上げにより整数化し，欠損部分に補完する．

$X_{質}$ は質問に着目した行列であるため，これを用いることで各質問の特徴を考慮することができる．また $X_{対}$ は評価対象に着目した行列であるため，各対象の特徴を考慮することができる．以下ではこの手法を提案手法 1 と呼ぶ．

2.5.2 回答者間の類似性を考慮したデータ補完

ここでは，回答者間の類似性に着目して欠損を補完する手法の手順について述べる．

手順 1: 質問 mode$X_{質}$ の 質問 × 対象 の行列 1 つが回答者の評点行列を示す．この評点行列を欠損のない回答者群 X_{obs} と欠損のある回答者群 X_{miss} に分ける．

手順 2: X_{miss} の中で，欠損箇所が最も少ない回答者 Res_{miss} と最も類似した回答者を X_{obs} から選ぶ．このとき式 (14) に示す RMSE (Root Mean Square Error) を用いて最も評点の差が小さい回答者 Res_{obs} を選ぶ．

$$RMSE = \sqrt{\frac{1}{n}\sum_{o,q}(i_{oq} - j_{oq})^2} \tag{14}$$

ここで i_{oq} は Res_{obs} の対象 o，質問 q における評点，j_{oq} は Res_{miss} の対象 o，質問 q における評点である．また，n は Res_{miss} の評点が付いている項目数である．

手順 3: 手順 2 で選んだ Res_{obs} の負荷量行列 W を用いて，潜在変数 z を求める．ここで x_{res} は，Res_{obs} の評点行列 (対象 × 質問) である．

$$z = (W^T W)^{-1} W^T x_{res} \tag{15}$$

手順 4: 手順 3 で求めた潜在変数 z を用いて補完値を算出する．

$$x_{miss} = Wz \tag{16}$$

手順 5: 補完した回答者 Res_{miss} を欠損のない回答者群 X_{obs} に加え，手順 2〜5 が X_{miss} 群のすべての回答者に適用されるまで繰り返す．

回答者の潜在変数を利用することで，その回答者の評点傾向を捉えて欠損を補完することができる．以下ではこの手法を提案手法 2 と呼ぶ．

3. 実験

実際のアンケートデータに対して，従来法 (質)，(回) と CF(協調フィルタリング) 法および提案手法 1，2 を適用し，欠損補完した際の精度の比較を行う．

3.1 アンケートデータ

実験に用いたアンケートデータについて説明する．1014 名の回答者に対して，次世代型サービスに関する Web アンケートを行った．6 個の次世代型サービスに対する説明文がそれぞれ評価対象である．回答者は各対象には 10 個の質問項目，合計で 60 個の質問に回答した．回答は 1〜5 の 5 段階評点尺度法を用いて行った．

3.2 実験方法

3.1 で説明したアンケートデータに対して，従来法 (質)，(回) と CF 法および提案手法 1，2 を用いてそれぞれ欠損補完を行った．全体の 1 割の回答者にあたる，100 人の評点には欠損箇所がないとし，残り 9 割の回答者の評点に欠損箇所を作成した．ここで，欠損箇所はランダムに作り，欠損割合はデータ全体の 5%，10%，20%，40% とした．評価指標として，真値を正しく補完できたかを示す正答率と，真値に近い値を補完できたかを示す $RMSE$ を用いて各手法を比較した．欠損箇所作成から欠損補完までを 1 試行とし，これを 10 試行行い，正答率の平均値と $RMSE$ の平均値を求めた．

$$\text{正答率} = \frac{N_{true}}{N_{miss}} \tag{17}$$

$$RMSE = \sqrt{\frac{1}{N_{miss}}\sum(T_{true} - T_{imp})^2} \tag{18}$$

ここで，N_{true} は補完値が真値と一致した数，N_{miss} は欠損数を示す．また，T_{true} は元データの評点，T_{imp} は補完した評点であり，$RMSE$ の値が小さいほど，真値に近い値を補完できていることを示す．

3.3 結果と考察

各欠損率における正答率および $RMSE$ を表 3.3(a)-(d) に示す．表 3.3(a)-(d) に示すように，全欠損率で正答率は提案手法 2 が最も高く，他の手法よりも真値を正確に補完できていることがわかる．また，$RMSE$ は従来法 (質) が最も小さく，真値に近い値を多く補完していることがわかる．一方で，提案手法 1 では正答率，$RMSE$ ともに，他の手法と比べて大きく下回った．

表 1：正答率と RMSE

(a) 欠損率 5%

	正答率	$RMSE$
従来法 (質)	47.9%	0.879
従来法 (回)	41.0%	0.999
CF 法	38.4%	0.997
提案手法 1	34.3%	1.14
提案手法 2	51.2%	0.904

(b) 欠損率 10%

	正答率	$RMSE$
従来法 (質)	48.4%	0.878
従来法 (回)	41.6%	1.00
CF 法	37.5%	1.01
提案手法 1	33.4%	1.16
提案手法 2	50.9%	0.913

(c) 欠損率 20%

	正答率	$RMSE$
従来法 (質)	47.6%	0.877
従来法 (回)	41.2%	0.999
CF 法	35.3%	1.10
提案手法 1	32.5%	1.14
提案手法 2	50.7%	0.918

表 1：正答率と RMSE
(d) 欠損率 40%

	正答率	$RMSE$
従来法 (質)	47.5%	0.877
従来法 (回)	41.1%	1.00
CF 法	29.6%	1.35
提案手法 1	34.1%	1.11
提案手法 2	49.6%	0.938

各手法における欠損率 5% のときの補完値の分布を図 3 に示す．縦軸はデータ数，横軸は評点を示す．欠損数は，評点 3 が一番多く，ついで 2,4,1,5 の順となっている．図から，従来法 (質) では，補完した評点がすべて 2〜4 となっていることがわかる．その結果，真値との誤差である RMSE の値は小さくなったと考えられる．一方，CF 法と提案手法 2 については，評点が 1 や 5 となるものも補完できている．また，従来法 (回) と提案手法 1 では，補完値が評点 3 に集中していることがわかる．

図 3 について，各手法における各評点に対する正答率を図 4 に示す．縦軸は正答率，横軸は評点を示す．提案手法 2 では，多くの評点において正答率が一番高く，その結果表 3.3 において他の手法よりも全体の正答率が高くなったと考えられる．また従来法 (質) は，評点 2〜4 においては提案手法 2 と同程度の正答率を示しているが，評点 1 と 5 では正答率が 0 となっていた．一方で，提案手法 2 と同様に評点 1 や 5 となるものも補完していた CF 法では，特に評点 4 と 5 の正答率が低いことがわかる．

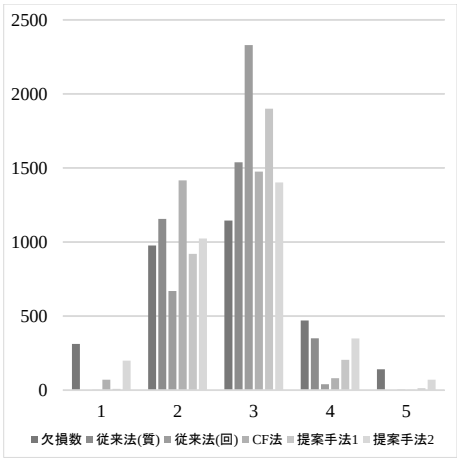

図 3: 補完値分布 (欠損率 5%)

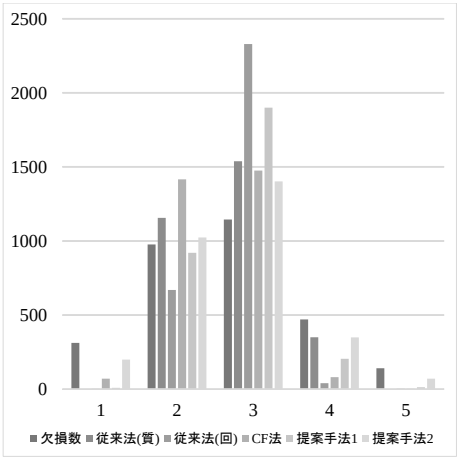

図 4: 各評点の正答率 (欠損率 5%)

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 Web アンケートに適用し，提案手法 2(回答者間の類似性を考慮) では従来法よりも真値を補完できる割合が高いことを示した．今後の課題として，回答者間の類似性について，評点の付け方を考慮した方法に対する検討などが挙げられる．

参考文献

[柳澤 07] 柳澤 秀吉, 村上 存, 福島 清暁：製品意匠の感-性評価における多様性分析：携帯電話のデザインへの適用 (OS12-2 感性・感情の設計), 設計工学・システム部門講演会講演論文集, pp.48-51, 2007.

[Osgood 57] Osgood C, Suck G, Tannenbaum P : The Measurement of Meaning, University of Illinois Press(1957).

[Myrtveit 01] 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.

[神嶌 07] 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.

[Tipping 99] E.Tipping, M.Bishop : Mixtures of Probabilistic Principal Component Analyzers, Neural computation, vol.11, pp.443-482, 1999.

[Qu 09] 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.

[Timmerman 00] ME Timmerman, HAL Kiers : Three-mode principal component analysis: Choosing the numbers of components and sensitivity to local optima, British Journal of Mathematical, vol.53, pp.1-16, 2000.