JSDSS 日本ソーシャルデータサイエンス論文誌
第 7 巻　第 1 号（2023 年 3 月

確率的潜在意味解析におけるパラメータ推定時に利用可能な初期値設定法の提案

寺澤 眞之介[†1]　大竹 恒平[†2]　生田目 崇[†1]

概要：大量データからのルール抽出や因果関係の発見を目指した機械学習手法が様々な分野で浸透しつつある．機械学習手法の多くでは，適当な初期値から設定されたペナルティを削減するように繰り返し計算することによってパラメータを得る．パラメータが多くても計算機の性能で解を得られるという長所がある反面，設定する初期値によって得られる解が異なるという問題もある．本論文では，マーケティングのセグメンテーションなどでよく使われる確率的潜在意味解析を対象に，パラメータ推定時に利用可能な初期値の設定方法について提案する．この手法は初期値を明示的に設定することで一意な解を得られるだけでなく，反復回数の削減による計算効率の向上も目指したものであり，実データを用いた提案手法の検証を行い，その有効性を確認する．

キーワード：確率的潜在意味解析，セグメンテーション，コレスポンデンス分析，計算効率

1. はじめに

近年，情報通信技術の進展に伴い，様々な業種・領域において多様な大規模データが蓄積されつつある．これらのデータに対する機械学習手法の適応は，とりわけマーケティング実務におけるデータ分析を大きく変えた．大量のデータから意味のあるルールや因果関係をデータドリブンに求めようとする以外にも，複雑なモデルを仮定した高度な分析が行えるようになった．

従来からの多変量解析は変数間の相関関係をもとに因果関係や構造評価をするものであった．したがって，行列からパラメータの求解を行うことが一般的であった．これに対して，多くの機械学習手法に共通する大きな特徴としては，反復計算による解法が用いられることが挙げられる．計算機環境の進化がこうした手法を可能にしたわけであるが，反面で，反復解法ならではの欠点もある．一つは計算時間が長くかかることであり，もう一つは解の安定性がないことである．

こうした問題に対して，本論文ではマーケットセグメンテーションなどで広く使われるようになった確率的意味解析(probabilistic Latent Semantic Analysis; pLSA)を取り上げ，初期値依存の問題と計算回数削減を目指した一つの方法を提案する．そして，実データを用いた検証を通して，本論文で提案する方法の有効性を確かめる．

2. 確率的潜在意味解析

確率的潜在意味解析は潜在クラスモデルの一種であり事象が生起する代表的なパターンを抽出するトピックモデルの一つである[1]．潜在クラスモデルは，観測されていない少数の潜在的な変数を用いて，観測された変数の関係を説明するモデルであり[2]，pLSA は各観測対象や変数が潜在クラスに所属確率を扱ったモデルである．

pLSA は，行の要素x_iと列の要素y_jの背後に共通する潜在的な意味クラスz_kがあると仮定し，次元圧縮を行うことで，行列の行要素と列要素が同時共起することが多いようなグループを抽出する手法である．pLSA が利用される領域としては文章や音声や，小売店や EC におけるクラスタリングやマーケットバスケット分析などが挙げられる．

pLSA の共起確率を(1)式に，そのグラフィカルモデルを図 1 に示す．

$$P(x_i, y_j) = \sum_k P(x_i|z_k)P(y_j|z_k)P(z_k) \tag{1}$$

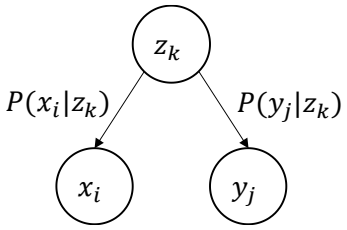

図 1 pLSA のグラフィカルモデル

(1)式とx, yの共起回数$N(x_i, y_j)$を用いて，データ全体の対数尤度は次のようにして与えられる．

†1 中央大学 (連絡先：nama@kc.chuo-uac.jp)
†2 東海大学
投稿日：2022 年 12 月 15 日
採録日：2023 年 3 月 18 日

$$
\begin{aligned}
L &= \sum_z N(x_i, y_j) \log P(x_i, y_j) \\
&= \sum_i \sum_j N(x_i, y_j) \log \sum_k P(x_i|z_k) P(y_j|z_k) P(z_k) \\
&= \sum_i \sum_j N(x_i, y_j) \times \\
&\qquad \log \sum_k P(z_k|x_i, y_j) \frac{P(x_i|z_k) P(y_j|z_k) P(z_k)}{P(z_k|x_i, y_j)}
\end{aligned} \tag{2}
$$

(2)式において，Lを最大にするような$P(x_i|z_k), P(y_j|z_k), P(z_k)$を求めることになるが，この時，イェンセンの不等式を用いてこのLの下限を与え，その下限を最大化することで，パラメータ推定をする．ただし，潜在クラスモデルで求めるパラメータは多いため，一般には EM アルゴリズム（Expectation-Maximization アルゴリズム）を用い，反復による解の改善が閾値以下になったところで収束と判定する．この時，pLSA の E ステップと M ステップはそれぞれ以下の式で計算される．この式が示すように，代数的な計算のみで更新が行われるところが特徴である．

＜E ステップ＞

$$
\begin{aligned}
P(z_k|x_i, y_j) &= \frac{P(x_i, y_j, z_k)}{P(x_i, y_j)} \\
&= \frac{P(x_i|z_k) P(y_j|z_k) P(z_k)}{\sum_k P(x_i|z_k) P(y_j|z_k) P(z_k)}
\end{aligned} \tag{3}
$$

＜M ステップ＞

$$
P(x_i|z_k) = \frac{\sum_j N(x_i, y_j) P(z_k|x_i, y_j)}{\sum_i \sum_j N(x_i, y_j) P(z_k|x_i, y_j)} \tag{4}
$$

$$
P(y|z) = \frac{\sum_i N(x_i, y_j) P(z_k|x_i, y_j)}{\sum_i \sum_j N(x_i, y_j) P(z_k|x_i, y_j)} \tag{5}
$$

$$
P(z) = \frac{\sum_i \sum_j N(x_i, y_j) P(z_k|x_i, y_j)}{\sum_i \sum_j \sum_k N(x_i, y_j) P(z_k|x_i, y_j)} \tag{6}
$$

3. 提案手法

3.1 提案の方針

pLSA では(3)式から(6)式に示した，EM アルゴリズム[3]により解を求める．EM アルゴリズムは更新するごとに尤度が向上することは知られているが，初期値に依存してその最終解は異なり[4,5,6]，また著者のこれまでの経験上，全く異なる解が得られても，尤度は大きく変わらないという知見を得ている．pLSA で用いられるデータは行をトランザクションとしたときに列の各変数の反応もしくは発生件数をまとめた 2 次元の共起行列である．そこで，コレスポンデンス分析[7]を先に行い，その結果をもとに初期値を与えることによって，より近い反応をする行もしくは列が当初から同じ潜在クラスに入りやすいようにする．

ただし，コレポンデンス分析は固有値問題に帰結されるため，その行スコアと列スコアの範囲は実数全体となる[8]．そこで，初期値を与える際には，ニューラルネットワークのアクティベーション関数の一種である，softmax 関数によって基準化している．

類似する方法としては，文献[9]や[10]によって示されているが，pLSA における行要素と列要素の対称関係もしくは，確率的要素を考慮した方法ではない．本論文における方法は，初期値設定において pLSA の特徴をできる限り保持するように与えているところに特徴があるといえる．

3.2 初期値の作成方法

初期値作成の流れを以下に示す．

1. 共起行列に対してコレスポンデンス分析を実施し，正準相関係数，行スコア，列スコアを算出する．この際の次元数を，pLSA の潜在クラス数に合わせる．
2. 各要素の行スコア，列スコアに対して，正準相関係数を掛け合わせ，各要素の座標を算出する．
3. 行の要素，列の要素ごとに座標に対して非階層型クラスタ分析の一つである k-means 法[11]によってクラスタリングをする．この時のクラスタ数も pLSA の潜在クラス数と揃える．
4. 各クラスタの重心座標を算出し，各要素と各クラスタの重心との内積を算出する．
5. 算出した内積に対して，softmax 関数を用いて各クラスタにおける各要素の合計が 1 になるようにする．
6. 算出した値の行方向を$P(x|z)$の初期値，列方向を$P(y|z)$の初期値とする．

3.3 クラスタ番号の突き合わせと性能評価

初期値の違いによる pLSA の安定性を評価するためには，複数回実験を実施し，得られた結果間において，似たクラスタのクラスタ番号を特定する必要がある．そこで，1 回目の結果のクラスタ番号を固定し，2 回目以降のクラスタ番号の割当てを，最適化問題を利用して行う．このとき，n 回目の結果において，$P(x, y|z) = P(x|z)P(y|z)$ が成り立つ．1 回目の結果のクラスタ番号を固定しているため，1 回目の結果の潜在クラスiと，$n = l$回目の結果の潜在クラスjの近接度sim_lを(7)式のように定義する．

$$
\begin{aligned}
sim_l(i, j) = \sum_x \sum_y & P(x, y|z = i, n = 1) \circ \\
& P(x, y|z = j, N - l)
\end{aligned} \tag{7}
$$

l回目ですべてのi, jに対して，$m \times m$の行列に並べたものを 1 回目の結果とl 回目の近接度SIM_lとする．さらに，割当

問題の変数として，バイナリ変数$t_{(i,j)}$ を用いて(8)式のような行列Vを作成する．ここで，$t_{(i,j)}=1$ となるとき，1 回目の結果の潜在クラスiと，l回目の結果の潜在クラスjが同じクラスとなる．なお，行列AとBのアダマール積を$A\circ B$を記している．

$$V=\begin{pmatrix}1 & \cdots & 0 & t_{(1,1)} & \cdots & t_{(1,m)}\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & \cdots & 1 & t_{(m,1)} & \cdots & t_{(m,m)}\end{pmatrix} \tag{8}$$

Vのk行目をV_kとすると割当問題を以下に示す．目的関数の(9)式は，1 回目の pLSA の結果と，l回目の pLSA の結果の近接度の最大化である．(10)式はV_kの合計が 2 になるように，(11) 式はVの$m+1$列目から $2m$ 列目までの各列の合計が 1 になるような制約条件である．すなわち，1 回目の潜在クラスiに相当するl回目の潜在クラスjが一つずつ，かつ，l回目の各潜在クラスjが一つずつ割り当てるようにしている．

$$\max \quad VV=\sum_{k=1}^{m}(V_k^{\mathsf{T}}\times V_k)\circ SIM_l \tag{9}$$

$$\text{s.t.} \quad \sum_{i=1}^{m}t_{(i,j)}=2, \qquad j=1,\dots,m \tag{10}$$

$$\sum_{j=1}^{m}t_{(i,j)}=1, \qquad i=1,\dots,m \tag{11}$$

$$t_{(i,j)}\in\{0,1\}, \qquad i,j=1,\dots,m \tag{12}$$

これを複数の得られた結果毎に行う．

そして，pLSA の性能を以下の 3 つの観点から評価する．

1. EM アルゴリズム収束時の対数尤度
2. EM アルゴリズムが収束するまでの反復回数
3. $P(x|z),P(y|z)$ の確率が上位の要素の共通度

1 では，pLSA の解の妥当性を評価する．pLSA の解は対数尤度が最大になるように最尤推定されている．そのため，収束時の対数尤度の大きさで解の妥当性を評価する．

2 の観点では，pLSA は EM アルゴリズムを利用して最尤推定を行っているため，解が収束するまで反復を繰り返す．そこで反復回数を測定して，これを収束速度とみなす．

3 の共通度として Jaccard 係数[12]を利用する．Jaccard 係数とは 2 つの集合に含まれる要素の和集合に対する共通の要素の割合であり，集合の類似度を表す．すなわち，Jaccard 係数が高いほうが pLSA の解が安定していると言える．

4. 評価実験

4.1 データの概要

本節では，前節で提案した初期値設定方法と，従来の乱数で発生させた初期値で pLSA を実行した際の性能を評価し，比較を行う．

評価実験として，あるスーパーマーケットの ID-POS データを利用する．対象データの概要を以下に示す．

- 分析対象期間 2015 年 1 月 1 日～2015 年 12 月 31 日
- 会員数 134,058
- 商品カテゴリ数 26 （例：“野菜”，“果物”）
- 年間売上件数 62,975,264 件

4.2 評価実験の流れ

評価実験の流れを以下に示す．

1 行を「会員番号」，列を「商品カテゴリ」，要素を「各会員が各カテゴリの商品を購入した点数」とした共起行列を作成する．
2 作成した共起行列を基に，提案手法の初期値を作成する．
3 初期値を以下の 2 通りで設定し，初期値ごとに pLSA を 10 回ずつ実行する．本研究の評価実験では，データの列数 (=26) を考慮し，潜在クラス数をともに 4 とした．
  3.1 実行する毎に初期値を全て乱数で発生させる
  3.2 提案手法で作成した初期値を利用して，$P(x|z),P(y|z)$ を固定し，$P(z)$の初期値を乱数で発生させる．
4 2 通りの初期値ごとに，10 回の実行結果について，クラスタ番号を統一する．
5 10 回の結果の性能を前節で述べた方法で評価する．
6 2 通りの初期値設定方法の性能を比較する．

以降で，2 通りの初期値から実行した pLSA の各結果のクラスタ番号を一致させ，pLSA の性能の評価を行い，比較する．

4.3 提案手法による初期値の作成

前節で述べた初期値作成方法に従い，データから作成された共起行列から，提案手法の初期値を作成する．作成した共起行列は，行を「会員番号」，列を「商品カテゴリ」としているため，$(134,058\times 26)$ の行列となる．

1 共起行列に対してコレスポンデンス分析を実施した結果として，表 1 に正準相関係数，表 2 に行の得点，

表 3 に列の得点を示す. この際の次元数は, 評価実験の次元数と合わせるため 4 次元とした.

表 1　正準相関係数

	正準相関係数
V1	0.397
V2	0.356
V3	0.289
V4	0.235

表 2　行スコア（一部）

	V1	V2	V3	V4
1	−0.212	0.819	0.695	0.171
2	−1.399	1.139	−0.532	0.334
3	−1.352	0.980	−0.753	0.752
4	−0.271	−0.396	0.530	0.479
5	0.563	2.026	0.095	0.351
6	0.180	1.122	−1.238	−1.682
7	0.039	−1.130	2.565	0.784
8	−1.108	0.752	−1.298	0.224
9	1.130	−2.709	0.106	2.017
10	0.003	−1.287	2.919	2.793

表 3　列スコア（一部）

	V1	V2	V3	V4
1	−0.902	0.793	−0.392	0.415
2	−0.062	0.178	−0.125	−0.929
3	−0.877	0.345	−0.465	0.643
4	0.110	−0.629	−0.397	−0.662
5	−0.893	0.949	−0.946	0.251
6	0.304	0.420	−1.414	−1.019
7	−0.719	0.688	−0.936	−0.149
8	−0.419	0.503	−0.230	−0.063
9	−0.561	0.520	−0.050	−0.008
10	−0.639	0.685	0.000	0.251

2　各要素の行の得点，列の得点に対して，正準相関係数を掛け合わせ，各要素の座標を算出した．さらに，行の要素, 列の要素ごとに座標に対して k-means 法でクラスタリングをし, 行方向の結果を表 4 に, 列方向の結果を表 5 に示す．この時のクラスタ数も pLSA の潜在クラス数と揃えるため，4 クラスに分割した.

表 4　座標とクラスタ番号（行方向，一部）

	V1	V2	V3	V4	クラスタ番号
1	−0.084	0.292	0.201	0.040	3
2	−0.555	0.406	−0.154	0.078	3
3	−0.537	0.349	−0.218	0.177	3
4	−0.108	−0.141	0.153	0.112	3
5	0.224	0.722	0.027	0.082	3
6	0.071	0.399	−0.358	−0.395	3
7	0.015	−0.402	0.742	0.184	1
8	−0.440	0.268	−0.375	0.053	3
9	0.448	−0.965	0.031	0.473	4
10	0.001	−0.458	0.844	0.656	1

表 5 座標とクラスタ番号（列方向，一部）

	V1	V2	V3	V4	クラスタ番号
1	−0.358	0.282	−0.113	0.097	2
2	−0.025	0.063	−0.036	−0.218	2
3	−0.348	0.123	−0.134	0.151	2
4	0.044	−0.224	−0.115	−0.155	4
5	−0.354	0.338	−0.273	0.059	2
6	0.121	0.150	−0.409	−0.239	2
7	−0.285	0.245	−0.271	−0.035	2
8	−0.166	0.179	−0.066	−0.015	2
9	−0.223	0.185	−0.014	−0.002	2
10	−0.253	0.244	0.000	0.059	2

3　各クラスタの重心座標を算出し, 各要素と各クラスタの重心との内積を算出した. 行方向の結果を表 6 に, 列方向の結果を表 7 に示す.

表 6 各要素と各クラスタの重心のベクトルの内積（行方向，一部）

	1	2	3	4
1	0.036	−0.501	0.051	−0.199
2	−0.347	−0.585	0.180	−0.278
3	−0.354	−0.366	0.174	−0.252
4	0.043	0.069	0.008	−0.042
5	0.069	−0.605	0.039	−0.235
6	−0.200	−0.473	0.026	0.047
7	0.395	0.015	−0.061	−0.090
8	−0.388	−0.204	0.141	−0.126
9	0.337	1.758	−0.202	0.389
10	0.486	0.427	−0.051	−0.193

表 7 各要素と各クラスタの重心のベクトルの内積
（列方向，一部）

	1	2	3	4
1	0.036	−0.501	0.051	−0.199
2	−0.347	−0.585	0.180	−0.278
3	−0.354	−0.366	0.174	−0.252
4	0.043	0.069	0.008	−0.042
5	0.069	−0.605	0.039	−0.235
6	−0.200	−0.473	0.026	0.047
7	0.395	0.015	−0.061	−0.090
8	−0.388	−0.204	0.141	−0.126
9	0.337	1.758	−0.202	0.389
10	0.486	0.427	−0.051	−0.193

4　算出した内積に対して，softmax 関数を用いて各クラスタにおける各要素の合計が 1 になるようにした．算出した値の行方向を$P(x|z)$ の初期値（表 8），列方向を$P(y|z)$の初期値（表 9）とする．なお，表中における E は指数表記であり，10 のべき乗を表す．

表 8 $P(x|z)$の初期値

	1	2	3	4
1	7.161E−06	6.455E−07	7.931E−06	5.688E−06
2	4.883E−06	5.935E−07	9.021E−06	5.253E−06
3	4.849E−06	7.386E−07	8.964E−06	5.390E−06
4	7.207E−06	1.142E−06	7.593E−06	6.655E−06
5	7.401E−06	5.818E−07	7.834E−06	5.482E−06
6	5.654E−06	6.636E−07	7.734E−06	7.268E−06
7	1.026E−05	1.081E−06	7.085E−06	6.341E−06
8	4.686E−06	8.686E−07	8.677E−06	6.116E−06
9	9.672E−06	6.179E−06	6.154E−06	1.023E−05
10	1.123E−05	1.633E−06	7.159E−06	5.721E−06

表 9 $P(y|z)$の初期値

	1	2	3	4
1	0.007188	0.04283	0.02576	0.03432
2	0.009334	0.03860	0.02820	0.03933
3	0.006092	0.04174	0.03108	0.03515
4	0.007505	0.03663	0.04078	0.04136
5	0.007617	0.04379	0.02676	0.03366
6	0.013010	0.03930	0.03577	0.03810
7	0.007555	0.04259	0.02810	0.03507
8	0.008683	0.04047	0.02781	0.03656
9	0.007977	0.04076	0.02619	0.03639
10	0.008208	0.04132	0.02531	0.03555

4.4 クラスタ番号の突き合わせ

残る$P(z)$については適当な初期値推定方法がないため，ランダムに発生させた．そして pLSA を 10 回ずつ実行した結果をもとに，3.3 節に示した方法によりクラスタ番号を突き合わせる．最適化問題の解を表 10，表 11 に示す．なお，表 10 が乱数を初期値とした場合であり，表 11 は本論文の方法で初期値設定をした場合である．

これら表では，クラスタ番号iに相当する，$n = l$ 回目の結果のクラスタ番号jが 1 となる．すなわち，表 10 の$n = 2$ 回目でクラスタ番号$j = 3$は，1回目の結果のクラスタ番号 $i = 4$ と同じクラスとしてクラスタ番号が突き合わされることなる．

表 10　クラスタ番号の突き合わせ（乱数）

		i						i			
n	j	1	2	3	4	n	j	1	2	3	4
1	1	1	0	0	0	6	1	0	0	1	0
	2	0	1	0	0		2	1	0	0	0
	3	0	0	1	0		3	0	0	0	1
	4	0	0	0	1		4	0	1	0	0
2	1	1	0	0	0	7	1	0	0	0	1
	2	0	1	0	0		2	0	0	1	0
	3	0	0	0	1		3	1	0	0	0
	4	0	0	1	0		4	0	1	0	0
3	1	0	1	0	0	8	1	0	1	0	0
	2	1	0	0	0		2	1	0	0	0
	3	0	0	0	1		3	0	0	0	1
	4	0	0	1	0		4	0	0	1	0
4	1	0	1	0	0	9	1	0	1	0	0
	2	1	0	0	0		2	1	0	0	0
	3	0	0	0	1		3	0	0	0	1
	4	0	0	1	0		4	0	0	1	0
5	1	0	0	1	0	10	1	0	0	1	0
	2	1	0	0	0		2	0	1	0	0
	3	0	0	0	1		3	0	0	0	1
	4	0	1	0	0		4	1	0	0	0

表 11　クラスタ番号の突き合わせ（初期値）

		i						i			
n	j	1	2	3	4	n	j	1	2	3	4
1	1	1	0	0	0	6	1	0	1	0	0
	2	0	1	0	0		2	0	0	1	0
	3	0	0	1	0		3	0	0	0	1
	4	0	0	0	1		4	1	0	0	0
2	1	0	0	1	0	7	1	0	1	0	0
	2	1	0	0	0		2	0	0	0	1
	3	0	1	0	0		3	0	0	1	0
	4	0	0	0	1		4	1	0	0	0
3	1	0	1	0	0	8	1	0	1	0	0
	2	0	0	0	1		2	0	0	0	1
	3	0	0	1	0		3	0	0	1	0
	4	1	0	0	0		4	1	0	0	0
4	1	0	1	0	0	9	1	0	1	0	0
	2	0	0	0	1		2	0	0	0	1
	3	0	0	1	0		3	0	0	1	0
	4	1	0	0	0		4	1	0	0	0
5	1	0	1	0	0	10	1	0	1	0	0
	2	0	0	0	1		2	0	0	1	0
	3	0	0	1	0		3	0	0	0	1
	4	1	0	0	0		4	1	0	0	0

4.5 有効性の検討

ここでは, 3.3 節で述べた 3 つの評価指標に従って, pLSA の性能を評価する．まず，初期値の違いによる収束時の対数尤度を表 12 に示す.

表 12　2 通りの初期値における対数尤度の比較

回数	乱数	初期値
1	−8.640.E+08	−8.634.E+08
2	−8.635.E+08	−8.634.E+08
3	−8.642.E+08	−8.634.E+08
4	−8.641.E+08	−8.634.E+08
5	−8.634.E+08	−8.634.E+08
6	−8.634.E+08	−8.634.E+08
7	−8.634.E+08	−8.634.E+08
8	−8.634.E+08	−8.634.E+08
9	−8.641.E+08	−8.634.E+08
10	−8.634.E+08	−8.634.E+08
平均	−8.637.E+08	−8.634.E+08

表 12 の結果から，2 通りの初期値で実行した pLSA の対数尤度について, 10 回の平均値が提案手法の初期値を用いた場合若干精度が高いという結果になったが，ほとんど等しいことが分かる．このことから，提案手法で pLSA を実行しても，従来の手法と同等の妥当性がある解を得られていると言える．したがって，提案手法で求めた pLSA の解も，妥当な解として解釈に至ることが可能である.

次に，表 13 に各回の収束までの反復回数をまとめる.

表 13　2 通りの初期値における反復回数の比較

回数	乱数	初期値
1	414	163
2	974	687
3	327	222
4	254	216
5	410	167
6	424	312
7	220	273
8	295	191
9	347	160
10	479	164
平均	414.4	255.5

表 13 から，EM アルゴリズムの反復回数について，乱数で初期値を発生させた際と，提案手法を比較すると，提案手法の方が乱数での初期値設定時より，4 割程度削減できていることが分かる．すなわち，提案手法のように，初期値の段階で似た傾向を持つ要素をまとめておくことにより，EM アルゴリズムの反復回数が削減できることが分かった．さらに，提案手法の方が，従来の手法よりも反復回数という点で優れていると考えることができる.

最後に，クラスタ番号を突き合せた pLSA の結果について，2 通りの初期値における$P(x|z)$, $P(y|z)$ の各クラスタの Jaccard 係数と，初期値を乱数で発生させた際の Jaccard 係数から作成した初期値を利用した際の Jaccard 係数を引いた値を表 14 に示す.

本研究では 10 回の結果に対して Jaccard 係数を用いるため，分子を 1〜10 回目のすべての組合せ 45 通りの共通要素数の合計，分母を 1〜10 回目のすべての組合せ 45 通りで出現する要素を対象とする．これを各クラスタについて行い，それらの平均を Jaccard 係数とする．また，上位の要素数を$P(x|z)$, $P(y|z)$ それぞれについて，100 個と 5 個とする.

表 14 $P(x|z), P(y|z)$の各クラスタの Jaccard 係数

	乱数(A)	初期値(B)	差分$(B-A)$
$P(x\|z)_1$	0.2155	0.7814	0.5660
$P(x\|z)_2$	0.2111	0.3548	0.1437
$P(x\|z)_3$	0.1302	0.5369	0.4067
$P(x\|z)_4$	0.1662	0.5935	0.4273
$P(x\|z)$平均	0.1807	0.5666	0.3859
$P(y\|z)_1$	0.5095	0.6881	0.1786
$P(y\|z)_2$	0.3198	0.5149	0.1951
$P(y\|z)_3$	0.3107	0.5852	0.2745
$P(y\|z)_4$	0.2944	0.7206	0.4263
$P(y\|z)$平均	0.3586	0.6272	0.2686

表 14 より，一番右の列の提案手法と従来の手法の差分が，すべて正の値をとっていることから，提案手法は Jaccard 係数が高いと分かる．すなわち，提案手法の方が確率上位の要素の変動が少ないと言えることから，pLSA の解が比較的安定していることが分かる．特に，xは顧客を示しており行数は列数に比べて各段に多い．乱数で初期値を与えた場合，かなり低い Jaccard 係数も見受けられるが，提案手法では，上位を見たときに平均で 50%が共通した顧客となっており，解の安定性がこの点からも読み取れる．

5. まとめと今後の課題

本論文では，pLSA の初期値依存性の問題に対して，初期値設定方法の提案を目的とした．この目的を果たすために，pLSA が次元圧縮した結果から近い要素をまとめる手法であることに着目し，初期値設定の段階で関係が強いと考えられる要素をまとめることで，安定した解を反復回数が少なく得られると考えた．そのために，コレスポンデンス分析, k-means 法, softmax 関数を用いて初期値を作成する手法を提案した．

さらに，提案した手法で作成した初期値と，従来の乱数で発生させた初期値で pLSA の性能を評価・比較を行い，提案手法の有用性について検討する評価実験を行った．

本研究では，pLSA の初期値依存性に対する，改善のための初期値設定方法を提案し，その有用性を示してきた．しかし，初期値作成の段階で，k-means 法を用いているため，k-means 法の初期値依存性の問題は依然として残る．そのため，近い要素をまとめる手法について改めて検討する必要がある．また，本研究では，潜在クラスの大きさ$P(z)$に関しては乱数で発生させていた．そのため，pLSA の結果が一意に決まることなく，多少のばらつきを持っている．これにより，提案手法を用いて分析を行った際にも，$P(z)$の初期値により解釈に多少の影響があると考えられる．そこで，$P(x|z), P(y|z)$の初期値を基に$P(z)$の初期値も作成する手段を検討する必要がある．また，本研究の数値実験においては，潜在クラス数を 4 と固定したが潜在クラス数を変えたときの挙動についての考慮する必要があると考える．

謝辞　本研究は中央大学特定課題研究助成の成果の一部である．

参考文献

[1] Hoffmann, T. “Probabilistic Latent Semantic Analysis,” *Proceedings of Uncertainity in Artificial Intelligence, UAI'99*, 8 pages (1999)
[2] J. A. Hagenaars and A. L. McCutcheon (eds), *Applied Latent Class Analysis. Cambridge*: Cambridge University Press (2002)
[3] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley (2007)
[4] アナリティクデザインラボ，PLSA（確率的潜在意味解析），http://www.analyticsdlab.co.jp/column/plsa.html（2022 年 12 月 8 日最終閲覧）
[5] 岩﨑幸子, “クラスタリングの安定化と相互類似関係を考慮したアソシエーション分析～データマイニング手法の困った問題への対処案～,” 数理システムユーザーコンファレンス 2018 (2018)
[6] 廣瀬英雄,「推薦システム」，pp. 66-69，共立出版 (2023)
[7] A. Farahat and F. Chen, “Improving Probabilistic Latent Semantic Analysis with Principal Component Analysis,” *Proceedings of 11st Conference of the European Chapter of the Association for Computational Linguistics*, pp.105-112 (2006)
[8] 内山俊郎, “情報理論的クラスタリングを用いた確率的潜在意味解析の性能向上”, 電子情報通信学会論文誌 D, Vol.J100-D No.3, pp.419-426 (2017)
[9] Sten-Erik Clausen/藤本一男(訳・解説),「対応分析入門 －原理から応用までー」，オーム社 (2015)
[10] 中村永友,「R で学ぶデータサイエンス 2　多次元データ解析法」，共立出版 (2009)
[11] 神鳥敏弘, “データマイニング分野のクラスタリング手法(1): －クラスタリングを使ってみよう！－,” 人工知能学会誌, 18 巻, 1 号, pp. 59-65 (2003)
[12] 大竹恒平，南場浩平，岡誠，生田目崇, “Twitter を活用した初対面時の対話の活性化に関する研究,” 日本ソーシャルデータサイエンス論文誌, 6 巻, 1 号（2022）

JSDSS 日本ソーシャルデータサイエンス論文誌
第 7 巻　第 1 号（2023 年 3 月

Proposal of Initial Values Set for Estimating Parameters of probabilistic Latent Semantic Analysis

Shinnosuke TERASAWA[†1]　Kohei OTAKE[†2]　Takashi NAMATAME[†3]

Abstract: Machine learning methods aimed at extracting rules from large amounts of data and discovering causal relationships are spreading in various fields. Many machine learning methods obtain parameters by iteratively reducing a loss function value from a reasonable initial value. Although there is an advantage that a solution can be obtained with the performance of a computer even if there are many parameters, there is also a problem that the obtained solution differs depending on the initial values to be set. In this paper, we propose a method of setting initial values for probabilistic Latent Semantic Analysis (pLSA), which is often used in marketing segmentation. This method not only obtains a unique solution by explicitly setting the initial value, but also aims to improve computational efficiency by reducing the number of iterations. We verify the proposed method using real data and confirm its effectiveness.

Keywords: probabilistic Latent Semantic Analysis, Segmentation, Correspondence Analysis, Computational Efficiency

†1 Chuo University (Correspondence Author: nama@kc.chuo-u.ac.jp)
†2 Tokai University