情報処理学会論文誌　データベース　Vol.13　No.4　1–12 (Oct. 2020)

主成分距離行列シルエットクラスタリングによる潜在因子ラベル付けモデル

大槻 明[1,a)]

受付日 2020年3月6日, 採録日 2020年7月10日

概要：従来の因子分析モデルでは，抽出因子数の正当性を確認するところまでは統計的に行われているが，抽出された潜在因子のラベル付け（潜在因子の解釈や意味付け）は，分析者の主観で行われているケースが多い．しかし，分析者の主観が入る時点で潜在因子の解釈に対する科学的根拠が損なわれてしまう可能性がある．そこで本研究では，潜在因子のラベル付けに科学的根拠を持たせるための補助的なアプローチとして，主成分得点からユークリッド距離を求めて距離行列を作成し，この距離行列を対象にシルエット分析を用いたクラスタリングを行うことで潜在因子にラベル付けを行うクラスタリングモデルを考案した．本クラスタリングモデルは，k-means のクラスタ数 k を自動推定できるよう拡張したモデルであるため，因子分析だけに限らず距離行列を対象とした様々なクラスタリングにも応用可能なモデルである．

キーワード：因子分析，主成分分析，シルエット分析，距離行列，k-means クラスタリング

Model for Labeling to Latent Factor by Silhouette Clustering Using Principal Component Distance Matrix

Akira Otsuki[1,a)]

Received: March 6, 2020, Accepted: July 10, 2020

Abstract: Although the conventional factor analysis model performed statistically up to the point where the number of extracted factors was verified, but labeling of extracted latent factors (Interpretation and meaning of latent factors) are often performed under the subjectivity of analysts. However, the scientific basis for the interpretation of latent factors may be compromised when analyst subjectivity involves. Therefore, in this study, as an auxiliary approach to provide a scientific basis for labeling latent factors, a distance matrix is created by calculating the Euclidean distance from the principal component scores, and clustering with silhouette analysis is performed on this distance matrix to devise a model that labels latent factors. Because this clustering model is extending k-means with automatic estimation of the number of cluster, can deal with various cases using distance matrix not apply only to factor analysis model.

Keywords: factor analysis, principal component, silhouette analysis, distance matrix, k-means clustering

1. はじめに

従来の因子分析モデルでは，抽出因子数の正当性を確認するところまでは統計的に行われているが，抽出された潜在因子のラベル付け（潜在因子の解釈や意味付け）は，分析者の主観で行われているケースが多い．しかし，分析者主観が入る時点で潜在因子の解釈に対する科学的根拠が損なわれてしまう可能性がある．

そこで本研究では，潜在因子のラベル付けに科学的根拠を持たせるための補助的なアプローチとして，主成分得点からユークリッド距離を求めて距離行列を作成し，この距離行列を対象に，シルエット分析を用いたクラスタリングを行うことで，潜在因子にラベル付けを行うクラスタリングモデルについて考案した．なお，本クラスタリングモデルは，k-means のクラスタ数 k を自動推定できるよう拡張

[1] 日本大学
Nihon University, Chiyoda, Tokyo 101–8360, Japan
[a)] otsuki.akira@nihon-u.ac.jp

したモデルであるため，因子分析だけに限らず距離行列を対象とした様々なクラスタリングにも応用可能なモデルである．

因子分析との比較検証の結果，従来の因子分析では，潜在因子数は推定できても，特に，高負荷量の因子を対象とした潜在因子のラベル付けを行うことが難しかった．対して本提案モデルでは，シルエット分析を用いて最適なクラスタ数 K に分割したうえで潜在因子数を抽出するモデルであるため，従来手法よりも科学的根拠を持たせる形でラベル付けを行うことが可能であることを示した．

2. 従来の因子分析における潜在因子の解釈

因子分析の結果解釈に関する論述は，古くは 1968 年に Armstrong ら [1] が論文「On the interpretation of factor analysis」のなかで行っており，この論文では，研究者は因子分析の解釈に例を用いて説明するのみであり，この例には因子分析の解釈にかかる信頼性や妥当性が省略されていると主張されている．最近でも，Pohlmann [2] が 1992 年～2002 年に The Journal of Educational Research で公開された論文を調査し，教育研究学分野における因子分析の解釈にかかる問題について論じている．具体的には，分析者は因子分析の結果を評価するために十分な情報を提供していないことが問題であると指摘している．つまり，分析者の恣意的な解釈により，因子分析の結果が解釈されていることを示唆している．

日本に目を向けると，柳井 [3] が，1998 年から 1999 年までに教育心理学研究・心理学研究で発表された論文を調査した結果について論じているが，主に因子分析法の利用をめぐる諸問題（因子回転の問題）についてしか論じられていない．つまり，日本においては潜在因子の解釈に関する議論はこれまでさかんには行われてこなかったと考えられる．

潜在因子数の推定に関する手法としては，最尤法を使用する場合には，多変量正規分布から乖離している変数を削除する，といった多変量正規分布を条件とすることがよく知られている [4]．また，堀 [5] のまとめによると，スクリーテストや平行分析などもよく用いられている．さらに，Lee ら [6] は歪度と尖度から分析対象の変数を選択している．

以上をまとめると，本論文で取り上げた先行研究では，潜在因子を抽出する手法については数多く研究されてきたが，抽出された潜在因子に対するラベル付け（解釈や意味付け）に深く言及している研究は少なかった．つまり，抽出された潜在因子を元に，分析者の主観でラベル付けが行われるケースが多いと考えられる．具体例を示すと，中川ら [7] の論文では，因子負荷量 0.3 以上の因子を潜在因子として抽出としており，松岡ら [8] の論文では，因子負荷量．500 以上の因子を潜在因子として抽出している．三保ら [9] の論文では，「因子パターンの値が各因子で ±0.40 以上であった 39 項目を 4 因子の解釈の対象項目とした」とされている．このように，分析者の主観が入る時点で潜在因子の解釈に対する科学的根拠が損なわれてしまう可能性があると考えられる．

ゆえに本研究では，潜在因子のラベル付けに科学的根拠を持たせるための補助的なアプローチとして，主成分得点からユークリッド距離を求めて距離行列を作成し，この距離行列を対象に，シルエット分析を用いたクラスタリングを行うことで，潜在因子にラベル付けを行うモデルについて提案する．

3. 分析対象データ

厚生労働省の平成 25 年若年者雇用実態調査の概況 [10] で公開されている表 22 のデータを用いることとした．この理由は，同表は，最終学校卒業後初めて勤務した会社をやめた主な理由について，男女，年齢および最終学歴などの別ごとにまとめられたものであり，ここから，従来手法と本モデルで，初めて勤務した会社をやめた主な理由の潜在因子を推定できるかについて比較検証するためである．実際に分析に用いたデータを付録 A に示す．付録 A における各項目（x1～x16）は，行に示されている，性別，年齢，最終学歴，在職中の雇用形態および勤続期間の属性ごとの「会社をやめた主な理由」をそれぞれ示している．

4. 従来の因子分析モデルによる潜在因子数の推定およびラベル付け

4.1 潜在因子数の推定

2 章で述べた先行研究の手法 [6], [7], [8], [9] を参考に因子分析を行う．まず，潜在因子数の決定であるが，中川ら [7] の論文では，「5) 抽出因子数決定のための基準」のなかで述べているように，「固有値が s1 よりも大きいもので，因子負荷量 0.3 以上の潜在因子」としている．松岡ら [8] の論文では，因子負荷量．500 以上の高い負荷量を解釈基準とした，とされている．三保ら [9] の論文では，「第 1 因子から第 7 因子にかけて 10.21，6.33，3.78，2.33，2.16，1.67，1.44 となった。3 因子から 5 因子の結果を比較して，解釈可能性から，因子数を 4 とし，主因子法で共通性を推定した。」としか書かれておらず，結局はこの論文の**図 1** で示されている 4 分野を設定しているだけであった．つまり，先行研究では，分析者の主観により潜在因子の抽出方法（閾値）がバラバラである．そこで，堀 [6] のまとめを参考に平行分析およびスクリーテストを用いることとした．平行分析の結果を図 1 に，スクリーテストの結果を**図 2** にそれぞれ示す．

図 1 の×が主成分解，つまり共通性の推定をしないときのスクリープロットを表している．そして，△が因子分析のスクリープロットであり，ここから因子数を決めることができる．具体的には，点線（2 種）が平行分析を表して

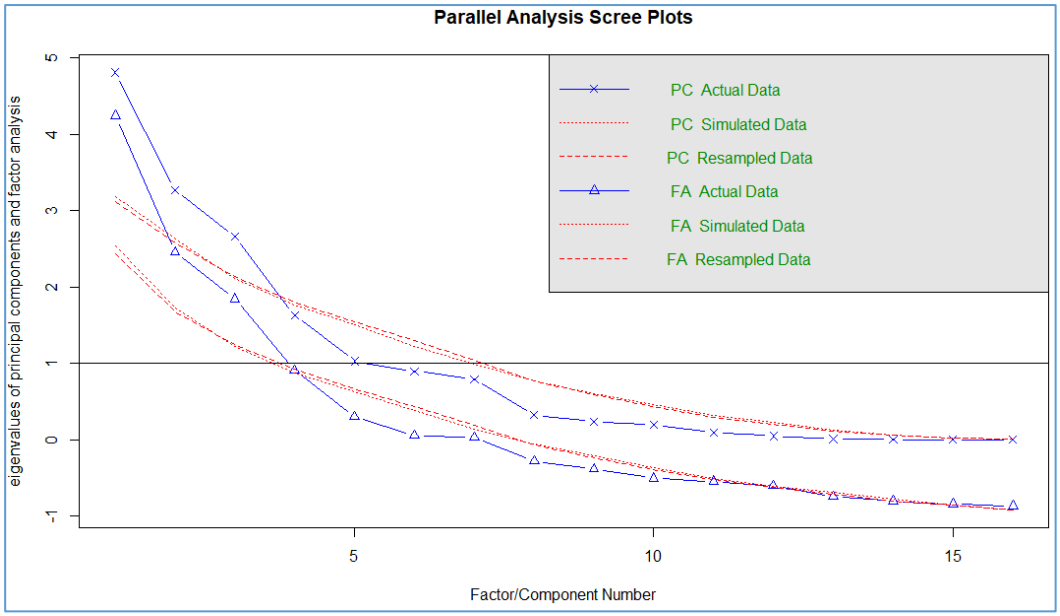

図 1　平行分析による因子数の決定
Fig. 1　Decision of number of factors.

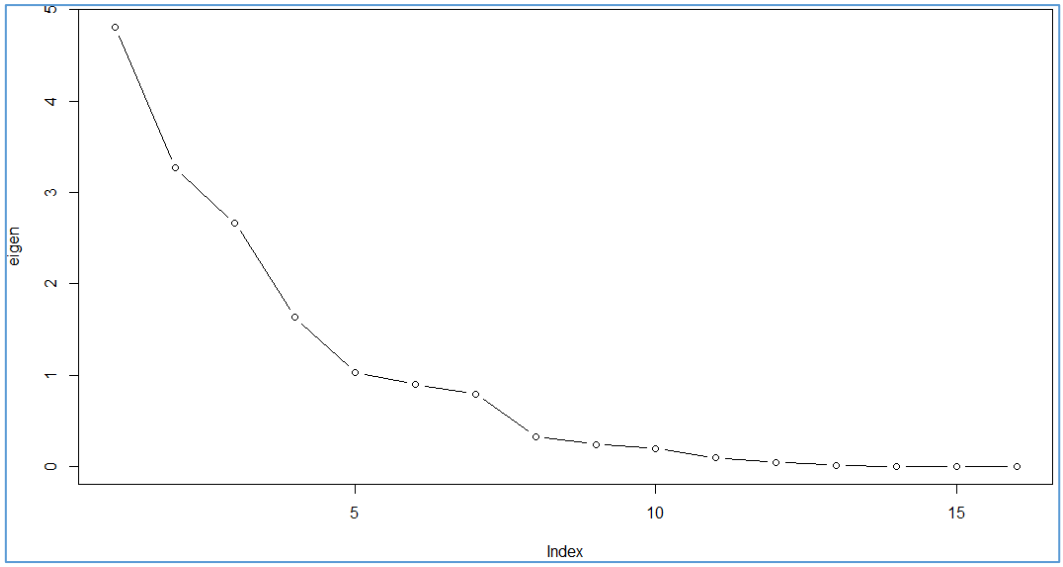

図 2　スクリーテストによる因子数の決定
Fig. 2　Decision of number of factors by scree test.

おり，これは同じデータセットの乱数表から作られる相関行列や，データセットの数値をリサンプリングして作られる相関行列のスクリープロットである．このラインより上に来る因子が有意と判断されるため，平行分析からは 3 因子と判断される．

また，スクリーテスト（図 2）は，相関行列の固有値を固有値順位に対してプロット（スクリープロット）し，これに最下位固有値から傾向線を引き，その傾向線から離れる固有値の順位が因子数となる．そして，三保らの手法を参考にすると，3 因子から 5 因子の結果を比較した場合，解釈可能性から因子数は 3 か 4 と仮定される．また，中川らの「固有値が 1 以上」の基準を参考にしても，因子数は 3 か 4 と仮定される．

4.2　潜在因子のラベル付け

前節から，因子数を 3 および 4 として因子分析を行った．なお，因子得点を求める方法には回帰法（トンプソン法）を，また回転法には promax を用いて因子分析を行った．この結果を**表 1**，**表 2**，**表 3**，**表 4**，**表 5** に示す．

前述のように，中川ら [7] の論文では，因子負荷量 0.3 以上の因子を潜在因子として抽出としており，松岡ら [8] の論文では，因子負荷量．500 以上の高い負荷量の因子を潜在因子として抽出している．三保ら [9] の論文では，「因子パターンの値が各因子で ±0.40 以上であった 39 項目を 4

項目	上位ラベル	項目と上位ラベルの関係性の根拠
x1 仕事が自分に合わない	労働環境	厚生労働省「改正労働基準法案の制定経緯に係る資料２」iでは，労働環境問題の一つとして長時間労働問題を捉えており，この長時間労働問題は業種・職種といった仕事内容に関係していることを示唆しているため，仕事内容は労働環境の一つであると考えられる．
x2 自分の技能・能力が活かせられなかった	自己都合	会社側はこの社員の能力が十分活かされる仕事を任せていたと主張する可能性もあり，また「自分の技能・能力は今の会社では活かされず，他の会社等でなら活かされる」という考え方自体が自己都合的な考え方であると考えられるため．
x3 責任のある仕事を任されたかった	労働環境	ここでいう責任は仕事に対するものであり，仕事内容に含まれると考えられるため，x1 同様の根拠により上位ラベルとして「労働環境」を付けた．
x4 ノルマや責任が重すぎた	労働環境	ここでいうノルマや責任は仕事に対するものであり，仕事内容に含まれると考えられるため，x1 同様の根拠により上位ラベルとして「労働環境」を付けた．
x5 会社に将来性がない	労働環境	会社自体が労働をする上で必要な環境の一つであると考えられるため．
x6 賃金の条件がよくなかった	労働環境	厚生労働省サイト「労働条件・職場環境に関するルール」iiから，賃金は労働環境の一つであると考えられるため．
x7 労働時間・休日・休暇の条件がよくなかった	労働環境	厚生労働省サイト「労働条件・職場環境に関するルール」ii から，労働時間等は労働環境の一つであると考えられるため．
x8 人間関係がよくなかった	労働環境	厚生労働省サイト「労働条件・職場環境に関するルール」ii では，職場のパワハラに関する言及もあるため，人間関係は労働環境の一つであると考えられるため．
x9 不安定な雇用状態が嫌だった	労働環境	不安定な雇用状態とは，雇用期間（労働条件）のことを指していると考えられ，労働条件は労働をする上で必要な環境の一つであると考えられるため．
x10 健康上の理由	自己都合	自身の健康上の理由で会社を辞めることは，自己都合による退職理由であると考えられるため．
x11 結婚、子育てのため	自己都合	自身の結婚，子育てのために会社を辞めることは，自己都合による退職理由であると考えられるため．
x12 介護、看護のため	自己都合	自身の介護，看護のために会社を辞めることは，自己都合による退職理由であると考えられるため．
x13 独立して事業を始めるため	自己都合	独立して事業を始めるために会社を辞めるということは，自己都合による退職理由であると考えられるため．
x14 家業をつぐ又は手伝うため	自己都合	家業をつぐ又は手伝うために会社を辞めるということは，自己都合による退職理由であると考えられるため．
x15 1 つの会社に長く勤務する気がなかったため	自己都合	１つの会社に長く勤務する気がないという退職理由は，自己都合による退職理由であると考えられるため．
x16 倒産、整理解雇又は希望退職に応じたため	会社都合	倒産、整理解雇又は希望退職は，明らかに会社側都合による退職理由であると考えられるため．

[i] 厚生労働省「改正労働基準法案の制定経緯に係る資料 2」，https://www.mhlw.go.jp/file/05-Shingikai-11201000-Roudoukijunkyoku-Soumuka/0000142961.pdf
[ii] 厚生労働省サイト「労働条件・職場環境に関するルール」，https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/koyou_roudou/roudouseisaku/chushoukigyou/joken_kankyou_rule.html

因子の解釈の対象項目とした」とされている．つまり，先行研究では分析者の主観によって潜在因子抽出の基準はバラバラである．そこで，中川ら [7] の論文では，多くの論文をサーベイした結果，因子負荷量 ±0.30 以上が妥当として潜在因子の解釈を試みていたため，本研究でもこの閾値を採用することとした．なお，たとえば表 1 の x7 のように，複数の因子が因子負荷量 ±0.30 となる場合は高い方の因子を採用した．さらに，相関分析では，相関係数が．70 以上で「強い相関がある」と定義されている [11]．このように高い負荷量で見るほど科学的根拠が増すと考えられるため，負荷量 $\geqq 0.7$ を閾値とした場合についても検証した．

まず，因子数 3 のケースで因子負荷量 ±0.30 の基準を用いた場合の潜在因子のラベル付けであるが（表 1〜表 2），Factor1（潜在因子）は「労働環境と自己都合の混在」とい

表 1　初めて勤務した会社を辞めた理由尺度の因子分析と尺度構成結果（因子数 3）
Table 1　Result of the factor analysis and scale composition about the reason leaving the company (Number of factors is 3).

	Factor1	Factor2	Factor3	Ave	StDev
x1	-0.41	0.15	0.93	0.22	0.55
x2	0.45	0.12		0.29	0.17
x3	0.75	-0.11	0.14	0.26	0.36
x4	-0.58	0.25	0.21	-0.04	0.38
x5	0.50	1.08	0.24	0.61	0.35
x6	0.33	0.30		0.32	0.02
x7	-0.34	0.60	0.16	0.14	0.39
x8	-0.85	-0.28		-0.56	0.28
x9	0.42	-0.56	0.29	0.05	0.43
x10	-0.84	0.15		-0.34	0.49
x11	-0.21	-0.10	-0.90	-0.40	0.35
x12	0.14	-0.15	0.20	0.06	0.15
x13	-0.79			-0.79	0.00
x14	1.01	0.61	-0.11	0.50	0.46
x15	0.62			0.62	0.00
x16	0.14	0.73	-0.26	0.20	0.41

表 2　潜在因子の解釈（因子数 3，因子負荷量 ≧ ±0.30）
Table 2　Interpretation of latent factors (Number of factors is 3 and Factor loading ≧ ±0.30).

Factor1	労働環境	x3 責任のある仕事を任されたかった
	労働環境	x4 ノルマや責任が重すぎた
	労働環境	x6 賃金の条件がよくなかった
	労働環境	x8 人間関係がよくなかった
	自己都合	x2 自分の技能・能力が活かせられなかった
	自己都合	x10 健康上の理由
	自己都合	x13 独立して事業を始めるため
	自己都合	x14 家業をつぐ又は手伝うため
	自己都合	x15 1 つの会社に長く勤務する気がなかったため
Factor2	労働環境	x5 会社に将来性がない
	労働環境	x7 労働時間・休日・休暇の条件がよくなかった
	労働環境	x9 不安定な雇用状態が嫌だった
	自己都合	x16 倒産、整理解雇又は希望退職に応じたため
Factor3	労働環境	x1 仕事が自分に合わない
	自己都合	x11 結婚、子育てのため

表 3　潜在因子の解釈（因子数 3，因子負荷量 ≧ ±0.70）
Table 3　Interpretation of latent factors (Number of factors is 3 and Factor loading ≧ ±0.70).

Factor1	労働環境	x3 責任のある仕事を任されたかった
	労働環境	x8 人間関係がよくなかった
	自己都合	x10 健康上の理由
	自己都合	x13 独立して事業を始めるため
	自己都合	x14 家業をつぐ又は手伝うため
Factor2	労働環境	x5 会社に将来性がない
	自己都合	x16 倒産、整理解雇又は希望退職に応じたため
Factor3	労働環境	x1 仕事が自分に合わない
	自己都合	x11 結婚、子育てのため

表 4　初めて勤務した会社を辞めた理由尺度の因子分析と尺度構成結果（因子数 4）
Table 4　Result of the factor analysis and scale composition about the reason leaving the company (Number of factors is 4).

	Factor1	Factor2	Factor3	Factor4	Ave	StDev
x1	-0.25	0.12	0.93	0.21	0.25	0.43
x2	0.89	-0.17	0.57		0.43	0.45
x3	0.96	-0.18	0.13		0.30	0.48
x4	0.10	0.97			0.54	0.43
x5	0.28	1.00	0.27		0.51	0.34
x6	0.15	0.26	0.29	0.23	0.23	0.05
x7	-0.24	0.53	0.14	0.38	0.21	0.29
x8	-0.69	-0.25	0.11		-0.28	0.33
x9	0.50	-0.54	0.25	-0.16	0.01	0.40
x10	-0.50	0.11	0.55		0.05	0.43
x11	-0.13	-0.12	-0.93	0.21	-0.24	0.42
x12	-0.19	0.29	-0.51		-0.14	0.33
x13	-0.77				-0.77	0.00
x14	0.69	0.57	-0.10	-0.31	0.21	0.42
x15	0.55	0.21			0.38	0.17
x16	0.69	-0.22			0.23	0.46

うラベルを，Factor2 は「労働環境」のラベルを付けることがそれぞれ可能であると考えられる．しかし，Factor3 は潜在因子が「労働環境（x1）」，「自己都合（x11）」の 1 つずつしかなく，2 つ以上の変数が揃わないとラベル付けをする根拠に乏しいと考えられるため，Factor3 はラベル付けを行うことが困難である．なお，表 2，表 3，表 5，表 8 の 2 列目に x1～x16 の上位カテゴリが示されているが，これらは，次表に示す考察を通じて著者により付けられた上位カテゴリラベルである．

次に，因子数 3 のケースで因子負荷量 ±0.70 の基準を用いた場合の潜在因子のラベル付けであるが（表 1，表 3），

表 5　潜在因子の解釈（因子数 4，因子負荷量 ≧ ±0.30）
Table 5　Interpretation of latent factors (Number of factors is 4 and Factor loading ≧ ±0.30).

Factor1	労働環境	x3 責任のある仕事を任されたかった
	労働環境	x8 人間関係がよくなかった
	自己都合	x2 自分の技能・能力が活かせられなかった
	自己都合	x10 健康上の理由
	自己都合	x13 独立して事業を始めるため
	自己都合	x14 家業をつぐ又は手伝うため
	自己都合	x15 1 つの会社に長く勤務する気がなかったため
	自己都合	x16 倒産、整理解雇又は希望退職に応じたため
Factor2	労働環境	x4 ノルマや責任が重すぎた
	労働環境	x5 会社に将来性がない
	労働環境	x7 労働時間・休日・休暇の条件がよくなかった
	労働環境	x9 不安定な雇用状態が嫌だった
Factor3	労働環境	x1 仕事が自分に合わない
	自己都合	x11 結婚、子育てのため
	自己都合	x12 介護、看護のため
Factor4		NA（該当する因子が無い）

Factor1 は，因子負荷量 ±0.30 以上のときと同様に「労働環境と自己都合の混在」というラベルを付けることが可能であると考えられる．しかし，Factor2 を同様の閾値でみた場合は，閾値を超える因子数は，「労働環境（x5）」と「自己都合（x16）」の 1 つずつしかない．同様に，Factor3 も閾値を超える因子数も，「労働環境（x1）」，「自己都合（x11）」の 1 つずつしかないため，これらも前述と同様にラベル付けを行う根拠に乏しい結果となった．

次に，因子数 4 のケースで因子負荷量 ±0.30 の基準を用いた場合の潜在因子のラベル付けであるが（表 4〜表 5），Factor4 に該当する因子が存在していない時点で，潜在因子数 4 は妥当ではないと考えられる．具体的に，Factor4 で因子負荷量 ≧ ±0.30 の変数は x7 と x14 しかないが，x7 は Factor2 の因子負荷量が最も高いため Factor2 に所属することになり，x14 は Factor1 の因子負荷量が最も高いため Factor1 に所属することになる，結果 Factor4 に所属する因子は存在しないことになる．ゆえに，因子数 4 のケースで因子負荷量 ≧ ±0.70 の基準を用いた場合の検証は行わないこととした．

5.　提案モデル

5.1　主成分分析および累積寄与率

本モデルで主成分分析を用いる理由は，n 個の観測データを（合成変数）に要約し，クラスタリングをするための距離行列を作成するためである．データを要約する手法には，主成分分析以外にも，対応分析 [12] や MDS（多次元尺度構成法）[13] などが存在するが，まず，対応分析はカテゴリカル（質的）データを対象とした手法であり，本研究が対象とするデータは量的データである．また，主成分分析はデータの散らばり具合の情報量として保持して合成変数を作成し，それを軸としてデータを要約するのに対し，MDS は対象間の距離を維持したまま小さな次元に要約する．つまり，主成分分析では合成変数が何かしらの意味を持つのに対し MDS での軸は特に意味は持たない．本研究では，潜在因子を抽出することを目的とするため，合成変数に何かしらの意味を持たせることが可能である主成分分析を採用することとした．

3 章の対象データ（付録 A）を元に主成分分析を行い，累積寄与率を調べて第 n 主成分得点までを使うかを調査する．**表 6** に累積寄与率を求めた結果を示しているが，今回は累積寄与率が 90%を超える第 5 主成分得点までを使用することとした．

表 6　主成分分析の結果
Table 6　Result of cumulative proportion.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	…	PC16
0.40	0.67	0.82	0.87	0.92	0.95	0.97	…	1.00

5.2　主成分得点からユークリッド距離を計算

潜在因子は，付録 A であれば説明変数（x1〜x16）から算出するため，付録 A の行列を転値したうえで，第 5 主成分までの得点を用いて各説明変数間のユークリッド距離（Euclidean Distance：ED）を求める．例として説明変数 x1，x2 間のユークリッド距離 ED(x1, x2) は次式により求められる．なお，$\mathrm{pc}_{1\sim5}$ は第 1〜第 5 主成分得点を表す．

$$\mathrm{ED(x1,x2)} = \sqrt{(\mathrm{x1}_{pc1}-\mathrm{x2}_{pc1})^2+(\mathrm{x1}_{pc2}-\mathrm{x2}_{pc2})^2+,\cdots,+(\mathrm{x1}_{pc5}-\mathrm{x2}_{pc5})^2} \quad (1)$$

表 7　主成分得点距離行列
Table 7　Principal component distance matrix.

	x1	x2	…	x16
x1	ED(x1,x1)	ED(x1,x2)	…	ED(x1,x16)
x2	ED(x2,x1)	ED(x2,x2)	…	ED(x2,x16)
…	…	…	…	…
x16	ED(x16,x1)	ED(x16,x2)	…	ED(x16,x16)

そして，式 (1) によって算出された x1～16 間のユークリッド距離を元に，**表 7** に示すような主成分得点距離行列を作成する．

5.3　主成分距離行列を対象としたクラスタリング

前節の主成分得点距離行列を対象に，シルエット分析 [14] を用いてクラスタ数 K を自動推定する．シルエット分析は，クラスタ内の面が適切に分割されているかどうかを検証するための手法であり，本研究ではシルエット（silhouette）値を用いて分析する．

シルエット分析以外にも，k-means のクラスタ数 k を自動推定する手法には，X-means [15] や G-means [16] が存在する．X-means も G-means も，クラスタ分割を停止する条件が異なるだけで少ないクラスタ数から k-means を行い，分割されたクラスタをさらに再分割を繰り返しながら最適なクラスタ数 k に分割するというアルゴリズム自体は同じである．ここで，X-means は BIC（ベイズ情報量基準）を停止基準に用いており，G-means はクラスタ分割後の各クラスタ内のサンプルに対して，ガウス分布に従うかどうかについて Anderson–Darling 検定を行っている．BIC の問題点は，ベイズファクターは 2 つのモデルの尤度の比，つまり，数値の大小は 2 つを比較してどちらが良いか，という評価しかできないため，クラスタ内凝集度や他クラスタとの乖離度を総合的に評価することが難しい．また，事前分布によってはパラメータが増えるなど計算量が膨大になる．G-means の問題点は，クラスタ内の凝集度については考慮されているが，他クラスタとの乖離度については考慮されていない．以上から，本研究では，クラスタ内凝集度と他クラスタとの乖離度が総合的に評価できるシルエット分析を用いることとした．

シルエット値は，他のクラスタの点と比べてその点が自身のクラスタ内の他の点にどれくらい相似しているかを示す尺度であり，i 番目の点のシルエット値 S_i は式 (2) のように求められる．

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{2}$$

ここで，a_i は i 番目の点（ノード）から i と同じクラスタの他の点までの平均距離，つまりクラスタ内凝集度を表し，b_i は i 番目の点から別のクラスタの点までの最小平均距離，つまり他クラスタとの乖離度をそれぞれ表す．

つまり，平均シルエット値が最も大きいときにクラスタ分割を行うと，クラスタ内凝集度が最も高く，また他のクラスタとの離散度が最も高い状態でクラスタリングを行うことができる．今回は，**図 3** に示すとおり，最も平均シルエット幅が大きい数値は 3 であったため，K = 3 としてクラスタリングを行った．この結果を**図 4** に示す．

図 4 の各クラスタのメンバをリスト化したものが**表 8** である．表 8 から Cluster1 は労働環境に問題を感じて退職したメンバが多いクラスタであり，Cluster3 は自己都合による退職したメンバが多いクラスタであることが明らかになった．さらに，Cluster2 は労働環境に加え自己都合による退職したメンバのクラスタであることが明らかとなった．以上から，「労働環境（Cluster1）」と「労働環境および自己都合の混在（Cluster2）」および「自己都合（Cluster3）」という 3 つの潜在因子が抽出できたと考えられる．

表 8　クラスタ内メンバリスト（K = 3）
Table 8　Member list in the cluster (K = 3).

Cluster1	労働環境	x1 仕事が自分に合わない
	労働環境	x6 賃金の条件がよくなかった
	労働環境	x7 労働時間・休日・休暇の条件がよくなかった
	労働環境	x8 人間関係がよくなかった
Cluster2	労働環境	x4 ノルマや責任が重すぎた
	労働環境	x5 会社に将来性がない
	労働環境	x9 不安定な雇用状態が嫌だった
	自己都合	x2 自分の技能・能力が活かせられなかった
	自己都合	x10 健康上の理由
	自己都合	x11 結婚、子育てのため
Cluster3	労働環境	x3 責任のある仕事を任されたかった
	自己都合	x12 介護、看護のため
	自己都合	x13 独立して事業を始めるため
	自己都合	x14 家業をつぐ又は手伝うため
	自己都合	x15 1 つの会社に長く勤務する気がなかったため
	自己都合	x16 倒産、整理解雇又は希望退職に応じたため

5.4　クラスタリングの結果（潜在因子の）解釈

表 9，**表 10**，**表 11** は，付録 A を前節のクラスタ 1～3 ごとに分割し，「①Cluster1～3 の平均」列，「①/①～③の合計」列を追加したものである．そして，次式の IoC_{ij}（Interpretation of Cluster）により，各クラスタの評価を行った．

$$IoC_{ij} = \frac{ave_{ij}}{sum(ave_{i=1\sim 3j})} \tag{3}$$

i はクラスタ番号 1～3 を表し，j は 1，2，...，17 の値（番号）を表す．つまり j は付録 A の行サンプル a～q に対応している．たとえば，$j = 1$ は a を，$j = 17$ は q をそれぞれ表す．なので，ave_{ij} は i 番目のクラスタの j 行目のサンプルの全変数（x1～x16）の平均スコアを表す．また，$sum(ave_{i=1\sim 3,j})$ は，j 行目のサンプルのクラスタ 1～3 の

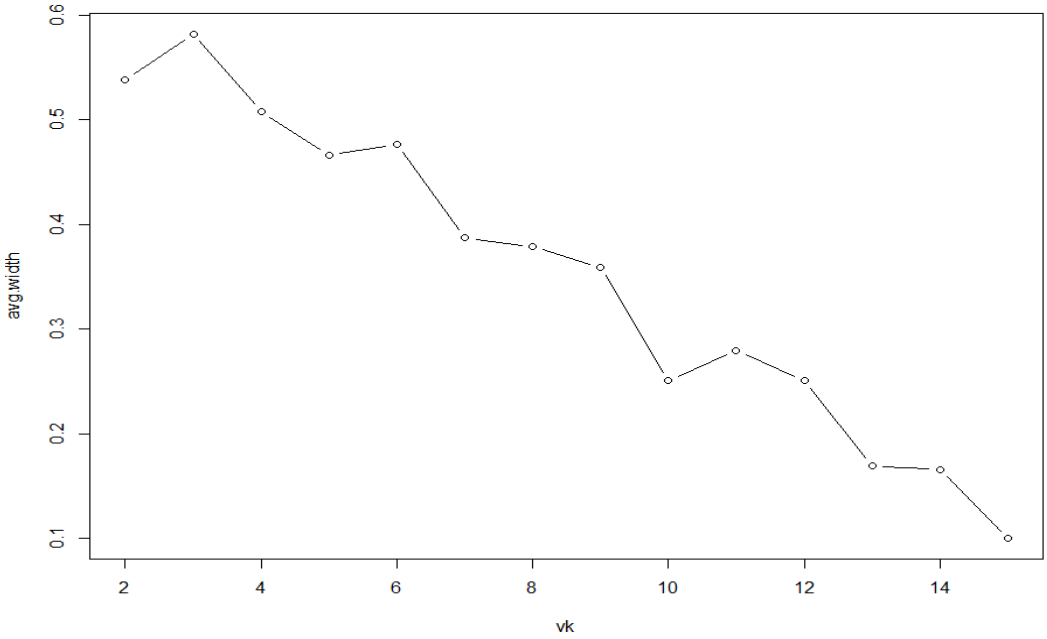

図 3　シルエット値の推移
Fig. 3　Change in the silhouette value.

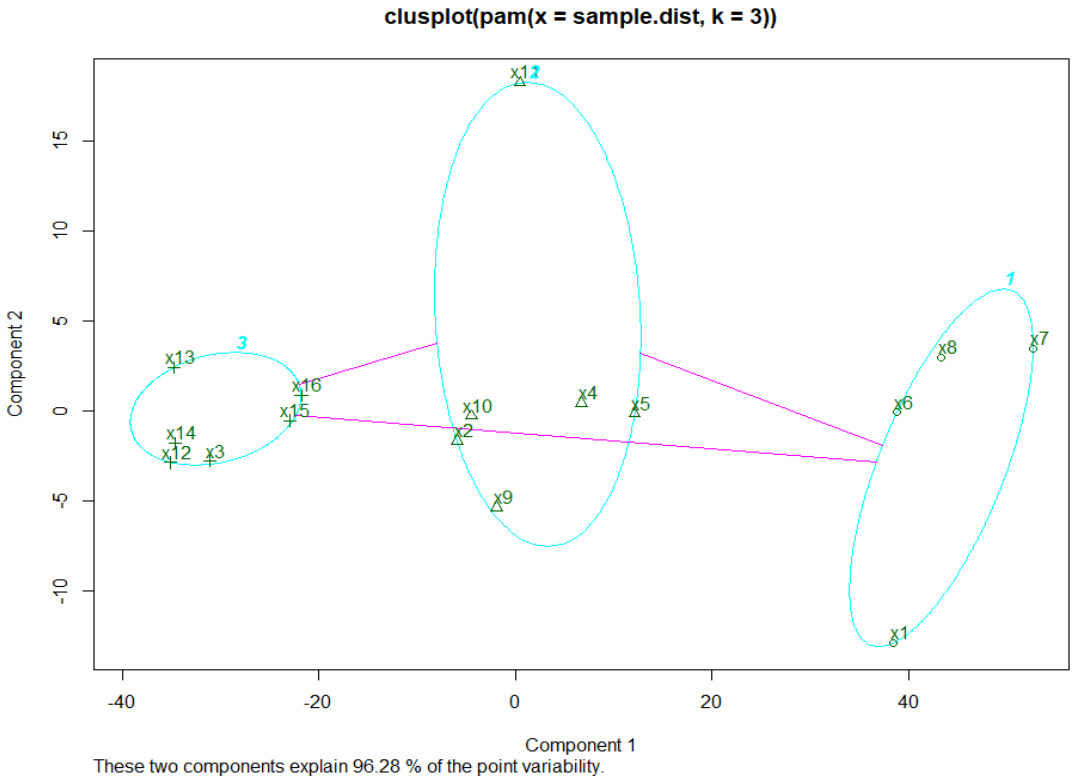

図 4　クラスタリングの結果（K = 3）
Fig. 4　Result of clustering (K = 3).

平均スコアの合計を表す．つまり，IoC_{ij} はクラスタ 1〜3 の合計に占める各クラスタの割合を，行サンプルの j 行の単位で算出する式である．なお，表 9〜表 11 では，ave_{ij} は「①Cluster1〜3 の平均」の列に，IoC_{ij} は「①/①〜③の合計」の列にそれぞれ表示されている．また，出典元 [10] によると，付録 A の値は選択割合であると述べられているため，IoC_{ij} の値が高いほど，a〜q の全被験者属性の選択率が高い列変数の集まりであると考えられる．ゆえに，クラスタ 1〜3 の重要度は，「クラスタ 1 > クラスタ 2 > クラスタ 3」という解釈ができ，さらに，全クラスタ（1〜3）の中で，クラスタ 1 の割合が過半数以上を占めていたため，a〜q の全被験者属性が退職した理由に最も起因している潜在要因

表 9　IoC_{ij} を用いたクラスタ 1 の分析結果
Table 9　Analysis result of cluster 1 using IoC_{ij}.

	x1	x6	x7	x8	①Cluster1の平均	①/①〜③の合計
a	22.0	22.1	22.7	15.7	20.63	0.65
b	16.1	14.7	21.8	22.8	18.85	0.61
c	23.3	20.1	23.8	22.7	22.48	0.70
d	18.7	18.3	22.6	22.1	20.43	0.63
e	17.4	17.3	21.8	16.7	18.30	0.60
f	19.0	17.9	20.3	19.6	19.20	0.63
g	23.4	19.6	22.3	26.7	23.00	0.66
h	14.3	21.5	20.8	24.0	20.15	0.63
i	19.7	16.4	25.3	15.5	19.23	0.60
j	14.5	16.8	11.6	14.1	14.25	0.57
k	18.9	20.1	24.7	18.2	20.48	0.64
l	18.6	16.0	19.8	21.0	18.85	0.61
m	27.3	20.0	21.8	24.2	23.33	0.66
n	21.1	20.5	27.3	19.8	22.18	0.65
o	19.9	23.3	25.5	19.2	21.98	0.61
p	12.3	18.2	24.7	15.6	17.70	0.57
q	7.7	18.8	19.8	20.0	16.58	0.54

表 10　IoC_{ij} を用いたクラスタ 2 の分析結果
Table 10　Analysis result of cluster 2 using IoC_{ij}.

	x2	x4	x5	x9	x10	x11	②Cluster2の平均	①/①〜③の合計
a	8.9	10	15.2	8.7	7.3	3.0	8.85	0.28
b	7.1	12	9.9	8.7	9.0	15.0	10.28	0.33
c	5.6	9.4	12.4	8.0	9.0	5.0	8.23	0.26
d	7.2	10.3	13.2	9.8	8.6	8.1	9.53	0.30
e	8.5	12.1	11.9	8.1	7.9	11.9	10.07	0.33
f	6.6	9.3	11.6	9.3	7.0	10.3	9.02	0.30
g	6.7	11.5	13.8	7.4	9.7	8.3	9.57	0.27
h	8.2	11.7	8.3	8.6	9.5	15.6	10.32	0.32
i	10.1	14.1	13.9	6.9	9.5	8.7	10.53	0.33
j	10.6	5.1	9.6	18.7	4.9	4.1	8.83	0.35
k	8.7	11.0	14.5	8.2	7.6	5.4	9.23	0.29
l	7.2	11.2	10.2	9.1	8.9	13.6	10.03	0.33
m	8.8	14.2	13.1	11.3	10.9	4.0	10.38	0.29
n	9.6	11.6	10.1	10.5	8.5	8.2	9.75	0.29
o	9.7	15.6	14.9	9.1	8.4	13.1	11.80	0.33
p	6.6	8.7	16.2	8.1	6.9	16.3	10.47	0.34
q	8.8	9.2	13.6	6.6	8.2	17.5	10.65	0.35

は「労働環境」であることが明らかとなった．

本節では，5.3 節で生成された 3 クラスタの重要度を IoC_{ij} で評価し，最も重要度の高いクラスタを抽出することによってクラスタリング結果の解釈を行った．IoC_{ij} は

表 11　IoC_{ij} を用いたクラスタ 3 の分析結果
Table 11　Analysis result of cluster 3 using IoC_{ij}.

	x3	x12	x13	x14	x15	x16	③Cluster 3 の平均	①/①〜③の合計
a	2.1	1.3	0.9	1.9	4.1	4.1	2.40	0.08
b	1.4	0.5	1.0	0.3	3.8	4.7	1.95	0.06
c	0.8	2.6	0.1	0.2	3.2	2.2	1.52	0.05
d	1.9	0.7	0.4	0.9	5.0	4.7	2.27	0.07
e	1.9	0.6	1.6	1.4	3.4	4.8	2.28	0.07
f	1.4	0.5	0.4	0.8	4.6	5.0	2.12	0.07
g	1.8	0.6	1.3	1.2	4.1	4.7	2.28	0.07
h	1.9	0.5	2.7	0.2	1.9	2.5	1.62	0.05
i	2.0	1.6	1.0	1.6	2.9	4.0	2.18	0.07
j	5.0	1.7	0.2	2.2	1.5	1.3	1.98	0.08
k	2.5	0.6	0.7	1.6	4.3	5.5	2.53	0.08
l	1.0	1.2	1.2	0.5	3.6	3.4	1.82	0.06
m	1.9	0.2	0.1	0.7	3.4	4.7	1.83	0.05
n	1.8	2.2	0.0	0.6	4.9	4.2	2.28	0.07
o	3.1	0.1	0.4	0.9	6.4	3.7	2.43	0.07
p	1.9	1.3	1.3	2.6	4.5	6.5	3.02	0.10
q	1.8	0.6	6.4	2.0	4.5	6.0	3.55	0.12

全クラスタの平均値に占める各クラスタの平均値の割合，つまりは平均値の分布を比較することでクラスタの重要度を評価するモデルであるが，このほかにも度数分布（ヒストグラム）を用いてクラスタの重要度を評価する方向性も考えられるため，最後にヒストグラムと IoC_{ij} との比較検証を通じて IoC_{ij} の有効性について検証する．

IoC_{ij} と同様に，5.3 節で生成された 3 クラスタの元データ（付録 A の回答割合）をヒストグラムで可視化（分析）した結果を**図 5**，**図 6**，**図 7** に示す．図 5〜図 7 のいずれも，横軸が元データの回答割合の度数区間を表し，縦軸が頻度を表している．図 5〜図 7 を比較すると，Cluster1 が他のクラスタに比べて回答割合の高いメンバが集まっていることが理解できる．しかし，IoC_{ij} では全クラスタ（1〜3）のなかで，クラスタ 1 の割合が全クラスタのなかで過半数以上を占めていたという状況まで明らかにできていたが，ヒストグラムではそこまでの状況を明らかにすることはできなかった．ゆえに，IoC_{ij} の方が本研究におけるクラスタの重要度を評価する手法としては有効であると考えられる．

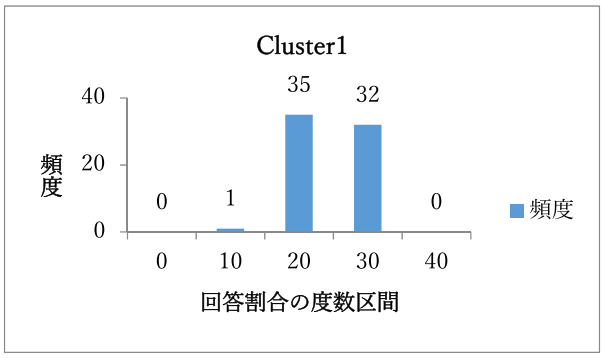

図 5　ヒストグラムによるクラスタ 1 の分析結果
Fig. 5　Analysis result of cluster 1 using histogram.

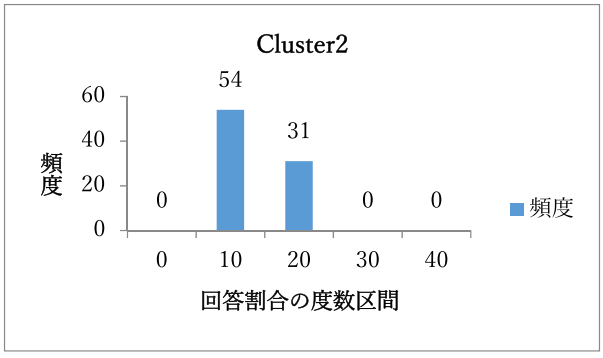

図 6　ヒストグラムによるクラスタ 2 の分析結果
Fig. 6　Analysis result of cluster 2 using histogram.

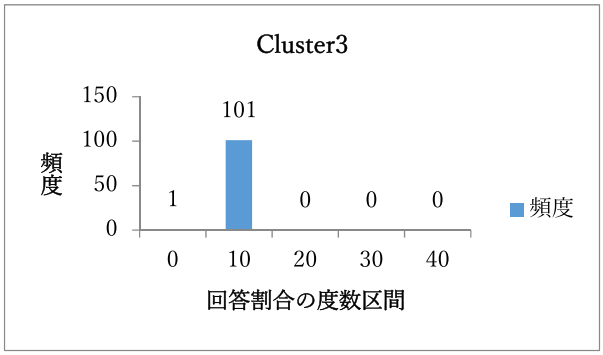

図 7　ヒストグラムによるクラスタ 3 の分析結果
Fig. 7　Analysis result of cluster 3 using histogram.

6.　おわりに

潜在因子の解釈に関する論述は，古くは 1968 年に J. Scott らの論文から行われているが，従来の因子分析では，因子回転や因子数の選定に関する議論は数多く行われてきたが，抽出された潜在因子の解釈は分析者の主観で行われてきたケースが多い．しかし，主観が入ってしまう時点で科学的根拠が大きく損なわれてしまう可能性がある．

ゆえに，本研究では，主成分得点からユークリッド距離を求めて主成分得点距離行列を作成し，シルエット（silhouette）分析を用いてこの距離行列から最適なクラスタ数 K を抽出することで，最適な潜在因子の集合知を得るアプローチについて提案した．

今後の課題としては，複数の意味が混合する潜在因子が特定できたときに，その解釈が可能となるアプローチについて研究していきたい．また，分析対象のデータの違いによって潜在因子のラベル候補となる用語が複数できてしまうケースも考えられるため，これらのなかからラベル候補となる用語を絞り込む方法についても研究していきたい．さらに，本研究で実証したデータは個票データではなく要約されたデータであったため，今後は個票データでも本モデルの有用性を確認していきたい．

参考文献

[1] Armstrong, J.S. and Soelberg, P.: On the interpretation of factor analysis (1968), available from ⟨https://repository.upenn.edu/marketing_papers/14⟩.
[2] Pohlmann, J.T.: Use and Interpretation of Factor Analysis in The Journal of Educational Research: 1992-2002, *The Journal of Educational Research*, Vol.98, No.1, pp.14–23 (2004).
[3] 柳井晴夫：因子分析法の利用をめぐる問題点を中心にして，*The Annual Report of Educational Psychology in Japan*，Vol.39, pp.96–108 (2000).
[4] Boomsma, A.: The robustness of LISREL against small sample sizes in factor analysis models, Jöreskog, K.G. and Wold, H. (Eds.), Systems under indirect observation: Causality, structure, prediction (part 1), pp.149–173 (1982).
[5] 堀　啓造：因子分析における因子数決定法，香川大学経済論叢，Vol.77, No.4, pp.35–37 (2005).
[6] Lee, S.-M., Terada, M., Shimizu, K. and Lee, M.-H.: Comparative Analysis of Work Values Across Four Nations, *Journal of Employment Counseling*, Vol.54, pp.132–144 (2017).
[7] 中川有加，西田みゆき，柳井晴夫：日本の看護学研究における因子分析法の利用，聖路加看護大学紀要，No.31, pp.8–16 (2005).
[8] 松岡　緑，西田真寿美，関　文恭：因子分析による看護学生の老人像に関する研究，九州大学医療技術短期大学部紀要，Vol.8, pp.35–43 (1981).
[9] 三保紀裕，清水和秋：大学進学理由と大学での学習観の測定—尺度の構成を中心として，キャリア教育研究，Vol.29, pp.43–55 (2011).
[10] 厚生労働省：平成 25 年若年者雇用実態調査の概況の表 22，入手先 ⟨https://www.mhlw.go.jp/toukei/list/dl/4-21c-jyakunenkoyou-h25_gaikyou.pdf⟩.
[11] 首都大学東京授業用コンテンツ：「相関分析」，入手先 ⟨https://infolit.uec.tmu.ac.jp/lit/contents/office2010/statistics/04/⟩（参照 2019-05-03）.
[12] Greenacre, M.: Correspondence Analysis in Practice, Chapman and Hall (2016).
[13] 高根芳雄：多次元尺度法，東京大学出版会 (1980).
[14] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, Vol.20, pp.53–65 (1987).
[15] Pelleg, D. and Moore, A.: *X-means: Extending K-means with Efficient Estimation of the Number of Clusters*, Carnegie Mellon University (2000).
[16] Hamerly, G. and Elkan, C.: Learning the k in k-means, available from ⟨https://papers.nips.cc/paper/2526-learning-the-k-in-k-means.pdf⟩ (accessed 2020-05-23).
[17] Zumel, N. and Mount, J.: *Practical Data Science with R*, Second Edition, ISBN 9781617295874 (2019).

付　　録

A.1　性・年齢階級・最終学歴・雇用形態・初めて勤務した会社での勤続期間階級，最終学校卒業後初めて勤務した会社をやめた主な理由別在学していない若年労働者割合 [10]

（「その他」や「不明」は潜在因子の意味解釈が困難なため除外した．またデータに-が含まれている項目や本分析に関係が無いと考えられる項目も除外した）．

（単位：%）

		x1 仕事が自分に合わない	x2 自分の技能・能力が活かせられなかった	x3 責任のある仕事を任されたかった	x4 ノルマや責任が重すぎた	x5 会社に将来性がない	x6 賃金の条件がよくなかった	x7 労働時間・休日・休暇の条件がよくなかった	x8 人間関係がよくなかった	x9 不安定な雇用状態が嫌だった	x10 健康上の理由	x11 結婚、子育てのため	x12 介護、看護のため	x13 独立して事業を始めるため	x14 家業をつぐ又は手伝うため	x15 1つの会社に長く勤務する気がなかったため	x16 倒産、整理解雇又は希望退職に応じたため
a	男	22.0	8.9	2.1	10.0	15.2	22.1	22.7	15.7	8.7	7.3	3.0	1.3	0.9	1.9	4.1	4.1
b	女	16.1	7.1	1.4	12.0	9.9	14.7	21.8	22.8	8.7	9.0	15	0.5	1.0	0.3	3.8	4.7
c	20 ～ 24 歳	23.3	5.6	0.8	9.4	12.4	20.1	23.8	22.7	8.0	9.0	5.0	2.6	0.1	0.2	3.2	2.2
d	25 ～ 29 歳	18.7	7.2	1.9	10.3	13.2	18.3	22.6	22.1	9.8	8.6	8.1	0.7	0.4	0.9	5.0	4.7
e	30 ～ 34 歳	17.4	8.5	1.9	12.1	11.9	17.3	21.8	16.7	8.1	7.9	11.9	0.6	1.6	1.4	3.4	4.8
f	高校卒	19.0	6.6	1.4	9.3	11.6	17.9	20.3	19.6	9.3	7.0	10.3	0.5	0.4	0.8	4.6	5.0
g	専修学校（専門課程）修了	23.4	6.7	1.8	11.5	13.8	19.6	22.3	26.7	7.4	9.7	8.3	0.6	1.3	1.2	4.1	4.7
h	高専・短大卒	14.3	8.2	1.9	11.7	8.3	21.5	20.8	24.0	8.6	9.5	15.6	0.5	2.7	0.2	1.9	2.5
i	大学卒	19.7	10.1	2.0	14.1	13.9	16.4	25.3	15.5	6.9	9.5	8.7	1.6	1.0	1.6	2.9	4.0
j	大学院修了	14.5	10.6	5.0	5.1.0	9.6	16.8	11.6	14.1	18.7	4.9	4.1	1.7	0.2	2.2	1.5	1.3
k	正社員	18.9	8.7	2.5	11.0	14.5	20.1	24.7	18.2	8.2	7.6	5.4	0.6	0.7	1.6	4.3	5.5
l	正社員以外	18.6	7.2	1.0	11.2	10.2	16.0	19.8	21.0	9.1	8.9	13.6	1.2	1.2	0.5	3.6	3.4
m	6 か月～ 1 年未満	27.3	8.8	1.9	14.2	13.1	20.0	21.8	24.2	11.3	10.9	4.0	0.2	0.1	0.7	3.4	4.7
n	1 年～ 2 年未満	21.1	9.6	1.8	11.6	10.1	20.5	27.3	19.8	10.5	8.5	8.2	2.2	0.0	0.6	4.9	4.2
o	2 年～ 3 年未満	19.9	9.7	3.1	15.6	14.9	23.3	25.5	19.2	9.1	8.4	13.1	0.1	0.4	0.9	6.4	3.7
p	3 年～ 5 年未満	12.3	6.6	1.9	8.7	16.2	18.2	24.7	15.6	8.1	6.9	16.3	1.3	1.3	2.6	4.5	6.5
q	5 年～ 10 年未満	7.7	8.8	1.8	9.2	13.6	18.8	19.8	20.0	6.6	8.2	17.5	0.6	6.4	2.0	4.5	6.0

大槻 明　（正会員）

2010 年慶應義塾大学大学院理工学研究科後期博士課程卒業．博士（工学）．東京大学特任研究員，お茶の水女子大学特任講師，東京工業大学特任准教授を経て，現在，日本大学教授．情報知識学会理事．関心分野は，グラフマイニング，ビッグデータサイエンス等．2012 年情報知識学会第 9 回論文賞，2012 年リバネス研究費エディテージ大賞，2018 年度社会情報学会関東支部研究発表会支部賞を受賞．

（担当編集委員　若林 啓）