《論　文》

探索的因子分析と主成分分析との使い分け

奥　　喜　正

1．はじめに

一般的な統計ユーザが頻繁に使用する多変量データ解析手法に，主成分分析と探索的因子分析（本稿では因子分析と記す）がある．双方の手法は，多くの変数がもつ情報を少数の新変数で要約することを目的にしており，解を求めるアルゴリズムなどでは相互に類似点も多い．現状では，心理学や経営学の分野では因子分析を使用する傾向が多く，一方，工学や計量経済学では主成分分析が好んで利用されているようにも窺える．そして，これら手法を利用した論文を数多拝読すると，それらの手法の相違点を顧みずに混乱して使用しているケースが見受けられる．統計ソフトウエアの一部では標準化された主成分得点だけ出力するものも見受けられ，このような事情により混乱がさらに増幅しているようにも思われる．本稿では，このような統計ユーザの利用状況を踏まえて，この2手法を出来るだけ正しく使い分けることができるように，ユーザにとっての使用上の留意点を明瞭にすることを目的として本研究を行う．そして，因子分析最終解の不安定性問題に関しては，シミュレーション分析を実行して因子得点に基づく製品マップの座標値変動を，主成分分析によるものと比較して検討を行う．本稿では，因子分析については探索的因子分析に限定し，各変数について標準化して分析を行う相関行列に基づく主成分分析（Principal Component Analysis：PCA）のみを考察の対象にする．

2－1．アルゴリズムの基本

因子分析で解を求める際に利用する最小二乗基準について最初に述べる．階数ｋの行列Ｘを階数Ｌの行列X_Aで最小二乗近似することを考える（ここで，Ｌ＜ｋとする）．

$$|| X - X_A ||^2 = \mathrm{trace}(X - X_A)(X - X_A)^T$$

これを最小にする最小二乗解 X_Aは，行列Ｘの固有値分解あるいはスペクトル分解の第Ｌ項までの和

$$X_A = \sum_{i=1}^{L} \lambda_i \mathbf{t}_i \mathbf{t}_i^T$$

によって与えられる．

つぎに，上記の最小二乗解を求めるために必要な**固有値分解**（スペクトル分解）について述べる．ｐ次実対称行列Ｘの固有値 λ_1，λ_2，$\cdots$，λ_p はすべて相異なる実数であるとして，対応する固有値ベクトルは $\mathbf{t}_1$，$\mathbf{t}_2$，$\cdots$，$\mathbf{t}_p$ とする．このとき，これらの列ベクトルを並べた行列Ｔ

$$T = [\mathbf{t}_1 \mathbf{t}_2 \cdots \mathbf{t}_p]$$

は直交行列になる．固有値を対角成分に並べた対角行列をｐ ｘ ｐ型行列 Ｄ＝diag（λ_1，$\cdots$，λ_p）にすると

$$T^T X T = D$$

とできる．実対称行列Ｘは直交行列Ｔを用いて対角行列Ｄに変形できる．また，この式は

$$X = TDT^T = \lambda_1 \mathbf{t}_1 \mathbf{t}_1^T + \lambda_2 \mathbf{t}_2 \mathbf{t}^T + \cdots \lambda_p \mathbf{t}_p \mathbf{t}_p^T$$

とも書けて，**固有値分解**あるいはスペクトル分解という．この式の右辺で，Ｌ番目（Ｌ＜ｐ）までの項の和によって，行列Ｘに関して最小二乗解をもたらす階数Ｌの行列が得られる．

２－２．アルゴリズムの概要

因子分析で，初期因子負荷行列Ａを求めるアルゴリズムはつぎの如くである．因子分析の基本モデルは

$$R^* = AA^T$$

である．ｐ次の相関行列 R^* は対角成分には，分析対象の標準化変数の分散である１ではなく，共通性の推定値が入る．すなわち，標準化された変数の分散１について

共通因子（共通性）＋独自因子（独自性）＝１

の如く仮定する．主な考察対象の分散である共通因子の「共通性」以外に，個々の変数がもつ「独自性」という**独自分散が存在する**ことを仮定する点が主成分分析とは異なっており，因子分析の一大特徴である．独自因子を説明する例として，物理学の評価点数は，曖昧な概念である理系的才能という「共通因子」だけでは説明できないであろう．物理学固有の才能，例えば，仮説を証明するために適切な実験計画や実験器具を考案し，

それらを実行する才能は，純粋数学とは異なる才能であり物理学に固有な才能・素養を独自因子と捉えるのが因子分析というモデルである．

因子分析ではｐ次の相関行列 R^* の固有値分解によって最小二乗解を得る．これは，ｋ＜ｐとして階数ｋの行列で，標本相関行列 R^* の近似を行うときは

$$R^* = T_1 D_1 T_1^{\mathrm{T}}$$

が成立する．ここで，$T = [\mathbf{t}_1 \mathbf{t}_2 \cdots \mathbf{t}_k]$，$D_1 = \mathrm{diag}(\lambda_1 \ \lambda_2 \cdots \lambda_k)$ である．よって，最小二乗解の初期因子負荷行列に関する階数ｋの行列は

$$A = T_1 D_1^{1/2} = [\sqrt{\lambda_1}\mathbf{t}_1 \ \sqrt{\lambda_2}\mathbf{t}_2 \cdots \sqrt{\lambda_k}\mathbf{t}_k]$$

のように得られる．

一方，主成分分析の解を求める手順は伝統的にはつぎのようになる．ｐ個の標準化された変数を要素にもつ変数ベクトル $\boldsymbol{x} = (x_1 \ x_2 \cdots \ x_p)$ において，それらの変数による合成変数

$$a_1 x_1 + a_2 x_2 + \cdots + a_p x_p = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$$

の分散，$\mathrm{Var}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\mathrm{Var}(\boldsymbol{x})\boldsymbol{a} = \boldsymbol{a}^{\mathrm{T}}R\boldsymbol{a}$ が最大になるように重み係数 $\boldsymbol{a}$ を求めることが目的である．ここで，標準化データ行列をスタートにする．この際，重み係数については $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} = 1$ という条件を課して

$$F \equiv \boldsymbol{a}^{\mathrm{T}}R\boldsymbol{a} - \lambda(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} - 1)$$

関数Fの最大化問題と捉え，ベクトルａで微分してゼロベクトルにおく．$2R\boldsymbol{a} - 2\lambda\boldsymbol{a} = \boldsymbol{0}$ より

$$R\boldsymbol{a} = \lambda\boldsymbol{a}$$

の関係を得る．この方程式を解いて，相関行列Ｒの固有値は相異なる実数であるとすると

$$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$$

のとき，対応する固有ベクトルを $\boldsymbol{a}_1$，$\boldsymbol{a}_2$，・・・，$\boldsymbol{a}_p$ とすると，第１主成分は $\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{x}$，第２主成分は $\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{x}$，・・・ のように主成分得点が得られる．このプロセスを繰り返すことは相関行列Ｒの固有値分解，すなわち，スペクトル分解によって相関行列Ｒを近似することに等しくなる．

２－３．探索的因子分析と主成分分析それぞれの分析目的

探索的因子分析には明瞭なる仮説がなくて，諸変数の背後にある共通因子を探求した

い場合に使用される．因子分析には，因子に関する仮説を検討する**確認的因子分析**があるが，それを使用する前段階に探索的因子分析を使用することも多い［1］［2］．本稿では，探索的因子分析に限定して因子分析を検討する．

探索的因子分析の使用目的は，多くの変数がもつ分散のデータ情報の**要約**と**圧縮**である［1］．**データ情報の要約**は，例えば，少数の共通因子あるいは階数が小さい**因子負荷行列**A_1により

$$A_1 = T_1 D_1^{1/2} = [\sqrt{\lambda}_1 \mathbf{t}_1 \sqrt{\lambda}_2 \mathbf{t}_2 \cdots \sqrt{\lambda}_k \mathbf{t}_k]$$

のように，**変数群の要約がｋ個**（ここで，**ｋ＜ｐ**）**の共通因子によって行われる**．また，**データ圧縮**は被験者の因子得点，**因子得点行列**Ｆによりなされて，それらは他のデータ解析メソッドのクラスター分析などに再利用され，被験者のグルーピングや消費者群のセグメント化に利用される．

例えば，本稿のシミュレーション分析で使用する各種アルコール飲料のイメージデータについて，因子分析を実行した場合の**データ圧縮例**を具体的に示す．因子分析の解析結果として得られたアルコール類の因子得点が表１である．それら因子得点を変数として，クラスター分析のWARD法で分析した分析結果がつぎのデンドログラムであり，350 x 10のデータ行列が下記の如く10 x 2の製品マップ用行列と樹系図に圧縮された訳である．この結果，諸アルコール類を，飲みやすい「ラガー，ドライビール」，高級でムード感が強い「ワイン，ウイスキー，ブランデー」，そして，飲みやすくもなく高級感の乏しいその他に銘柄の３分類ができた．

一方，主成分分析では，**少数の合成変数を作成すること**が主な目的で，第１主成分，第２主成分･･･のように求める．利用例として，多くの説明変数をもつ回帰分析で，多重共線性を回避するために，多くの説明変数群を少数の主成分得点群にまとめる，**主成分回帰**が典型的な例である［1］．

表１．諸アルコール類の因子得点

銘柄	因子得点１	因子得点２
ウイスキー	−0.20	0.73
ウオッカ	−0.93	−0.25
ジン	0.02	−0.02
テキーラ	−0.66	−0.37
ドライ	1.01	−0.52
ブランデー	−0.57	0.66
ラガー	0.80	−0.64
ワイン	0.78	0.84
焼酎	−0.08	−0.34
日本酒	−0.18	−0.09

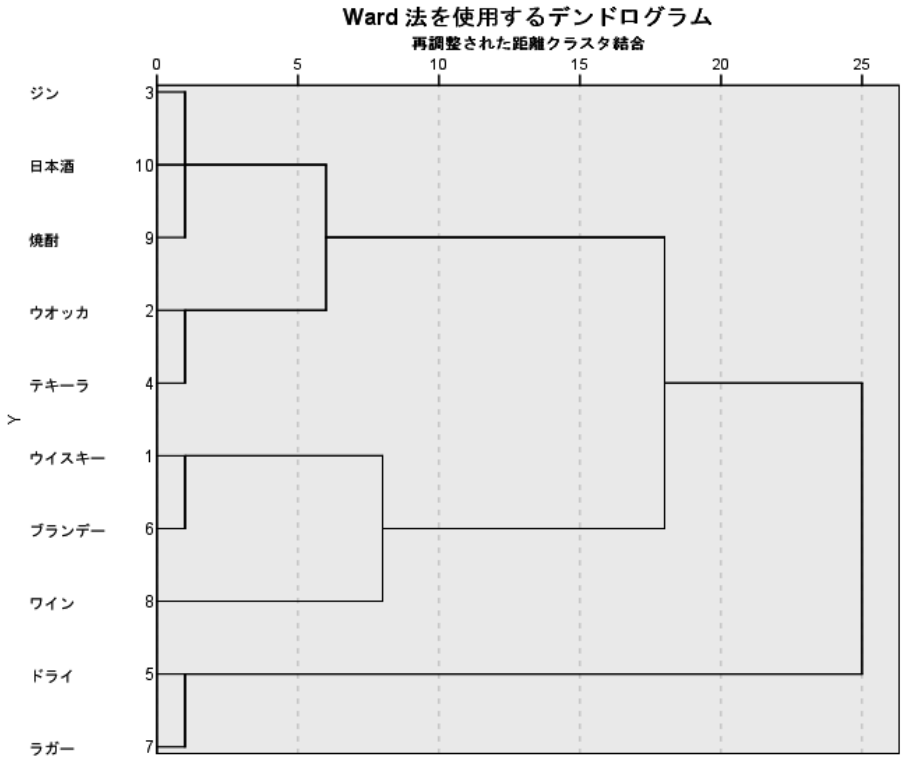

図１．因子得点に基づくクラスター分析の樹系図

入力データと主成分得点による主成分分析の４分類を以下の表２に示す．本稿では，この主成分分析の４分類を「足立の４分類」と名付ける．統計ソフトウエアの一部は主成分分析（PCA）の解である主成分得点について，表２の②④のみを出力するものもある．また，表２の④に相当する主成分分析は標準化データを入力して標準化された主成分得点を出力するものであるが，これは探索的因子分析に類似するものである．このケースの主成分分析が，因子分析と主成分分析との混乱利用を助長してきたとも窺える．特に，素データからの主成分分析と標準化データからの主成分分析では本質的に異なっており，相互に変換して解を求めることはできない［2］．すなわち，素データを分析して，標準化しない主成分得点を算出することは，主成分分析に特有のものである．

表２．　足立による主成分分析の４分類［2］．

		主成分得点出力		
		標準化（－）	標準化（+）	
入力データ	素データ	①	②	共分散行列のPCA
	標準化データ	③	④	相関行列のPCA

3．因子分析解の不安定性の検討

主成分分析に比べて，因子分析による解は不安定であるとか，一意に定まらないというような疑問点が投げかけられる．それは，因子分析ではモデルに独特な共通性の推定方法，初期解の回転方法の違いなどが存在し，多くのバリエーションが存在することが因子分析に対する疑問点の原因になっていよう．そこで，本稿では因子分析の直交解を求める方法について推定法は最小二乗法，回転はバリマックス直交回転に固定した．これら疑問点について，アルコール飲料アンケートデータを分析して得られる因子得点による製品マップで，分析データの微小変化によりもたらされる製品マップにおける諸対象の座標値変動を，主成分分析のそれと比較検討して考察を加える．つまり，因子負荷量と因子得点を算出して，**因子分析では因子得点を利用して製品マップを作製する**．同様の手順を主成分分析でも行い，**主成分分析では主成分得点によって製品マップを作製する**．製品マップは因子得点により作成されるが，因子得点行列Fは因子負荷行列をA，標準化データ行列をZとすると

$$F = ZA\ (A^{T}A)^{-1}$$

の如く得られる．よって，因子得点Fはデータ Z の微小変化により，相関行列Rも⊿R変化して，ゆえに因子負荷行列Aも影響を受けてΔA変動し，結果として製品マップの諸対象の座標値Fも変動する．

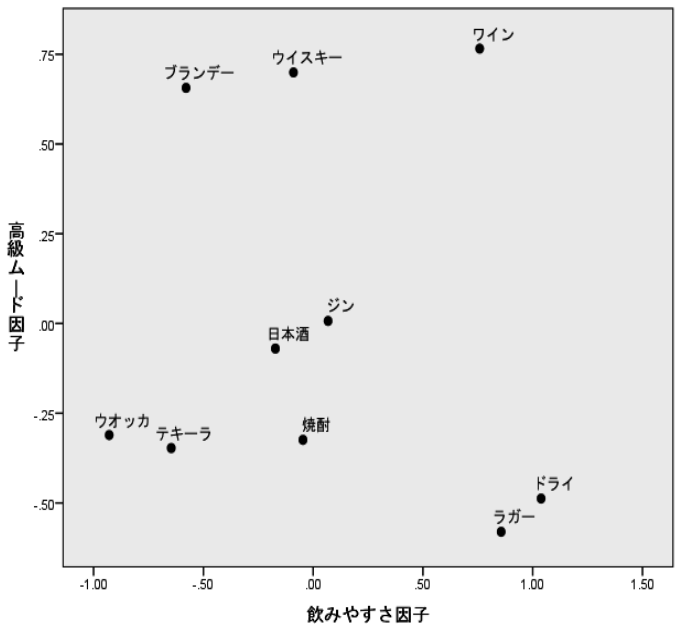

図２．因子得点によるアルコール類製品マップ（80%乱数５によるサブデータ）

表３．アルコール製品マップにおける諸アルコール飲料の二次元座標値の変動　80%

			因子分析	主成分分析	結果
	対象NO	対象名	標本分散	標本分散	標本分散比
第一因子	1	ウイスキー	0.00348	0.00442	1.271
	2	ウオッカ	0.00136	0.00166	1.221
	3	ジン	0.00584	0.00694	1.188
	4	テキーラ	0.00267	0.00338	1.264
	5	ドライ	0.00093	0.00124	1.335
	6	ブランデー	0.00249	0.00294	1.178
	7	ラガー	0.00407	0.00431	1.061
	8	ワイン	0.00390	0.00364	0.934
	9	焼酎	0.00325	0.00370	1.138
	10	日本酒	0.00377	0.00482	1.279
第二因子	1	ウイスキー	0.00308	0.00353	1.148
	2	ウオッカ	0.00477	0.00533	1.117
	3	ジン	0.00645	0.00813	1.259
	4	テキーラ	0.00457	0.00605	1.323
	5	ドライ	0.00341	0.00459	1.344
	6	ブランデー	0.00156	0.00232	1.487
	7	ラガー	0.00588	0.00762	1.295
	8	ワイン	0.00191	0.00247	1.293
	9	焼酎	0.00486	0.00591	1.218
	10	日本酒	0.00447	0.00600	1.342

表４．アルコール製品マップにおける諸アルコール飲料の二次元座標値の変動　60%

			因子分析	主成分分析	結果
	対象NO	対象名	標本分散	標本分散	標本分散比
第一因子	1	ウイスキー	0.0099	0.01248	1.264
	2	ウオッカ	0.0033	0.00461	1.407
	3	ジン	0.0139	0.01434	1.030
	4	テキーラ	0.0041	0.00530	1.285
	5	ドライ	0.0029	0.00314	1.096
	6	ブランデー	0.0056	0.00680	1.207
	7	ラガー	0.0133	0.01551	1.168
	8	ワイン	0.0126	0.01219	0.968
	9	焼酎	0.0117	0.01542	1.317
	10	日本酒	0.0152	0.01963	1.291
第二因子	1	ウイスキー	0.0065	0.00863	1.318
	2	ウオッカ	0.0116	0.01232	1.064
	3	ジン	0.0168	0.02114	1.261
	4	テキーラ	0.0075	0.00985	1.305
	5	ドライ	0.0081	0.01037	1.279
	6	ブランデー	0.0076	0.00935	1.233
	7	ラガー	0.0097	0.01271	1.314
	8	ワイン	0.0099	0.01226	1.237
	9	焼酎	0.0143	0.01657	1.162
	10	日本酒	0.0092	0.01229	1.339

本稿の分析データは，各種アルコール飲料のイメージを調査目的としたマーケティングリサーチデータで，被験者にはアルコールをよく嗜む男性35人を選び，10種類のアルコール類について，各種アルコールイメージを喚起させる７個の属性項目に答えてもらって得た 350 x10 型のイメージデータ行列Ｚである．

元データ行列から一様乱数（[0，5]）を利用して微小変化（約80%と約60%のデータを抽出した）させたテストデータをそれぞれ40個作成する．40個のテストデータを使用してシミュレーション分析を行い，データの微小変化によって，因子分析の因子得点による製品マップの，諸アルコール飲料二次元座標値がどの程度影響を受けて変動するかを，主成分分析による製品マップの諸座標値の変動状態と比較検討する．得られた40個の製品マップにおいて，各アルコール飲料の座標値変動をそれぞれの標本分散として捉え，双方のデータ解析手法において求めた．結果として，約80％の場合も60%のデータの場合も，主成分分析よりも因子分析による製品マップの２次元座標値の変動は小さくて（表３，４），因子分析で推定法と回転法を固定した場合では，データの微小変化による因子分析解の変動は主成分分析のそれよりも必ずしも大きくはなく，因子分析解が不安定であるとは本稿では見做せなかった．

４．考察

因子分析と主成分分析との間で使用上混乱が起き易いのは，変数を標準化する相関行列から分析を行う場合，いわゆる足立の４分類③，④の場合である．共分散行列から分析を行い，主成分得点も標準化しないケースでは，対応する因子分析のモデルは存在しないので混乱は起きない［2]．足立による４分類の④にあたる主成分分析は標準化変数データを入力して標準化された主成分得点を出力するものであるが，この主成分分析は因子分析と類似しているので，利用時に混乱の対象になり易い．

因子分析では標準化した変数の分散１について共通性とは別の，個々の変数が保持する独自分散を分析対象から排除して，因子負荷行列には共通性のみの分散情報を持ち込むという点で，④のケースの主成分分析とは異なるのである．他方，主成分分析では変数が持つ分散をすべて主成分，共通因子によって説明できると仮定するので，変数の保持する分散はすべて因子行列に持ち込まれる．つまり，相関行列の対角成分に共通性の推定値を挿入するのか，あるいは１を代入するかという点で両者は異なることになり，残りは固有値分解あるいはスペクトル分解を実行して相関行列の近似を実行して解を求めるので，両者の計算アルゴリズムの基本は類似的になる．

本稿では，因子分析で推定法と回転法を固定した場合では，データの微小変化による因子分析解の変動は主成分分析のそれよりも必ずしも大きくはなく，因子分析解の不安定性は主張できなかった．つまり，主成分分析よりも因子分析による解の２次元座標値

変動は小さかったが，その原因は主成分分析の主成分は独自分散や誤差分散をも含むものであり，因子分析における分析対象の分散である共通性よりも大きいので，結果として座標値変動も大きくなったとも推察できる．

ところで，共通性の値の大小により，因子分析では継続的な調査では属性変数の絞り込みが行える．他方，主成分分析では変数の分散が分析中すべて保持されるので，データ情報の圧縮が主な目的の場合には適切なメソッドであり，主成分回帰がその好例である．

表５．データ数Ｎがパラメータ数よりもあまり大きくないデータとその主成分分析結果

	立派な	役立つ	よい	大きい	力がある	強い	速い	騒がしい	誠実な	忙しい
僧侶	3.2	2.7	3.7	2.8	2.6	2.6	2.2	1.4	3.3	1.8
銀行員	3.4	3.5	3.4	2.5	2.2	2.6	3.2	2.1	4.1	4.2
漫画家	3	3.2	3.5	2.2	2.1	2.2	3.3	3.4	3.4	4.3
デザイナー	3.2	3.2	3.5	2.6	2.5	2.6	3.6	2.9	3.2	4
保母	4.2	4.6	4.5	3.1	3	3.2	2.8	3.3	4.5	4.9
大学教授	4	4	3.8	3.4	3.2	3.1	2.4	1.5	3.7	3
医師	4	4.8	3.9	3.5	3.8	3.7	3.2	2.1	3.7	4.5
警察官	3.7	4.6	4.1	3.4	4	4.1	4.3	3.4	4.2	4
新聞記者	3.6	4.3	3.7	2.9	3.5	3.6	4.7	4.2	3.9	5
船乗り	3.6	3.6	3.5	3.5	4.2	4.2	3.5	3.5	3.5	3.5
スポーツ選手	3.7	3.2	3.7	3.9	4.7	4.7	4.9	3.5	3.7	4.1
作家	3.4	3.7	3.5	3.1	2.7	2.4	2.3	1.8	3.3	3.3
俳優	3.2	3.2	3.6	2.9	2.2	2.5	3.3	3.3	2.8	4.3

説明された分散の合計

	初期の固有値			抽出後の負荷量平方和			回転後の負荷量平方和		
成分	合計	分散の %	累積 %	合計	分散の %	累積 %	合計	分散の %	累積 %
1	4.917	49.167	49.167	4.917	49.167	49.167	3.242	32.423	32.423
2	2.148	21.480	70.648	2.148	21.480	70.648	3.091	30.905	63.328
3	1.772	17.724	88.371	1.772	17.724	88.371	2.504	25.043	88.371
4	.443	4.430	92.801						
5	.386	3.861	96.662						
6	.158	1.583	98.245						
7	.108	1.077	99.322						
8	.032	.322	99.645						
9	.026	.260	99.905						
10	.010	.095	100.000						

因子抽出法: 主成分分析

回転後の成分行列[a]

	成分		
	1	2	3
役立つ	.877	.177	.169
よい	.858	.154	-.001
立派な	.841	.469	-.098
誠実な	.820	.130	.190
力がある	.214	.945	.208
大きい	.258	.926	-.096
強い	.237	.904	.327
騒がしい	.022	.129	.933
速い	-.075	.433	.858
忙しい	.429	-.162	.814

因子抽出法: 主成分分析
回転法: Kaiser の正規化を伴うﾊﾞﾘﾏｯｸｽ法
a. 5 回の反復で回転が収束しました。

さらに，データ数Nがパラメータ数pに比べてあまり大きくないときでも主成分分析は有効な解を得ることができて，表5がその例である．本稿では，各変数を標準化して分析する相関行列に基づく主成分分析のみを考察の対象に限定したが，それは諸変数の分散の大小という情報が重要である場合を除いて，分散共分散行列から主成分分析を実行すべき根拠は少ないからである．実際，相関行列からの場合と分散共分散行列からの場合の主成分分析による製品マップの座標値およびその変動には差はあまり認められない（表6，表7）．

表6．主成分分析による製品マップでの諸アルコール飲料の二次元座標値の変動．
―相関行列からの場合と共分散行列からの場合―

			相関行列	共分散行列
	対象NO	対象名	標本分散	標本分散
第一因子	1	ウイスキー	0.0044	0.0044
	2	ウオッカ	0.0017	0.0017
	3	ジン	0.0069	0.0074
	4	テキーラ	0.0034	0.0033
	5	ドライ	0.0012	0.0013
	6	ブランデー	0.0029	0.0031
	7	ラガー	0.0043	0.0045
	8	ワイン	0.0036	0.0042
	9	焼酎	0.0037	0.0037
	10	日本酒	0.0048	0.0047
第二因子	1	ウイスキー	0.0035	0.0036
	2	ウオッカ	0.0053	0.0052
	3	ジン	0.0081	0.0081
	4	テキーラ	0.0061	0.0060
	5	ドライ	0.0046	0.0052
	6	ブランデー	0.0023	0.0024
	7	ラガー	0.0076	0.0075
	8	ワイン	0.0025	0.0024
	9	焼酎	0.0059	0.0060
	10	日本酒	0.0060	0.0059

表7．相関行列と分散共分散行列からの主成分分析による製品マップの座標値

	相関行列からのPCA		共分散行列からのPCA	
	第一因子	第二因子	第一因子	第二因子
ウイスキー	−0.225	0.786	−0.215	0.779
ウオッカ	−1.053	−0.282	−1.053	−0.289
ジン	0.096	−0.005	0.061	−0.004
テキーラ	−0.719	−0.367	−0.747	−0.369
ドライ	1.095	−0.531	1.101	−0.519
ブランデー	−0.657	0.741	−0.629	0.727
ラガー	0.853	−0.673	0.874	−0.666
ワイン	0.925	0.868	0.906	0.869
焼酎	−0.125	−0.386	−0.122	−0.379
日本酒	−0.205	−0.129	−0.193	−0.131

そして，諸変数に関する先だった知識によって，独自因子に関与する分散が小さいことが暗示される場合には主成分分析の使用がやや適切で，逆に事前情報がない場合には因子分析の使用が好ましいように窺える．

参考文献

［1］Hair, Jr.J.F., Black,W.C., Babin, B.J. & Anderson, R.E. Multivariate Data Analysis: A Global Perspective, 7th ed. Pearson.（2010）

［2］足立浩平　多変量データ解析法．ナカニシヤ出版（2006）

［3］奥喜正，高橋裕　データ解析の実際．丸善プラネット（2013）

［4］Mulaik, S.A. Blurring the Distinction between Component Analysis & Common Factor Analysis. Multivariate Behavioral Research, **25**, 53-59.（1990）

［5］足立浩平　心理統計学と多変量データ解析　計算機統計学，Vol.14，pp.139-161.（2001）

［6］奥喜正　因子分析と主成分分析の関係，日本経営数学会第37回全国研究大会報告要旨集，38-41.（2015）

［7］Mulaik, S.A. Foundation of Factor Analysis, 2nd ed. Chapman & Hall.（2010）