

因子分析 (Factor Analysis)

因子分析 (Factor Analysis)

多変量解析の手法で、目的変数がなく説明変数が量的データの場合の手法は主成分分析と因子分析がある。どちらの手法も、多くの変数から少数の潜在変数を生成することを目的としている。類似している手法なので比較されることが多いが、主成分分析と因子分析では潜在変数の生成の仕方が異なる。

主成分分析での潜在変数はデータの分散を最大限に説明する新しい変数で、必ず直交している必要がある。データの分散を最大限に説明するため主成分分析の潜在変数の 1 つとして総合力が含まれる。それ以外の潜在変数は相反する概念のものとなる。(例えば、総合力と文系能力と理系能力など)

因子分析での潜在変数はデータの背後にある潜在的な共通因子(構造的な要因)を特定するための変数で、直交する必要はない。また潜在的な共通因子を表現するため、総合力は存在しない。潜在変数は 1 つの概念となる。(例えば、文系能力と理系能力など)

主成分分析は説明変数の個数が個体数より多くても実行できるが、因子分析は説明変数の個数より個体数が多くなければならない。

因子分析の流れ

観察変数 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ について、 $n > m$ 個の共通因子 ($\mathbf{f} = (f_1, f_2, \dots, f_m)$) で説明することを考える。それ以外にも各データは特有の要因として独立因子 ($\mathbf{e} = (e_1, e_2, \dots, e_n)$) を持つとする。共通因子と独立因子の線形結合によって観察変数を表すときの係数行列は因子負荷量と呼ばれる。因子負荷量は各因子と観察変数との相関の度合いを示す。因子負荷量の値から各因子がどのような潜在変数であるかを判断する。また因子負荷量の列ベクトル(各因子の負荷量)の 2 乗和をその因子が説明する共通性と呼ぶ。さらに \mathbf{f} は各観測データにおける共通因子の影響を示すスコアとして解釈される。

$$x_i = \sum_{j=1}^m c_{ij} f_j + d_i e_i$$

行列で表現すると

$$\mathbf{x} = \mathbf{C}\mathbf{f} + \mathbf{D}\mathbf{e}$$

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}, \mathbf{f} \in \mathbb{R}^{m \times 1}, \mathbf{e} \in \mathbb{R}^{n \times 1}.$$

関係式の右辺に未知数が多すぎるため、直接解くことはできない。よって共通性を仮定して因子負荷量を求める。また因子分析では因子負荷量が一意に決まるわけではない。よって回転方法を指定することで一意に定める。因子分析の流れは、[共通性の初期値を決める → 共通性を推定する → 求めた共通性をもとに因子負荷量を回転法の基準で一意に定める] のようになる。最初に定める共通性を共通性の初期値と呼ぶ。共通性の初期点としてはすべて 1 にする

方法、観測変数の相関行列の対角成分にする方法などがよく用いられる。因子負荷量を求める方法としては最尤法、最小2乗法、主因子法などが挙げられる。回転法はバリマックス回転、プロマックス回転などが挙げられる。

因子軸の回転方法

因子分析ではパラメータの制約が足りないため、共通性を推定するだけでは因子負荷量は一意に定まらない。よってデータの適合とは別に、解を識別させるための基準を外的に導入するために回転法を用いる。また回転させることで解釈性を向上させる目的もある。

回転基準を設定する方法には2種類ある。直交回転は、因子軸が直交することを仮定するモデル。因子軸が直交するとは、因子間の相関が0であることを意味する。よって直交回転を行うと因子間相関がすべて0となる。斜交回転は、因子軸が斜交することを仮定するモデル。因子軸が斜交するとは、因子間の相関があることをいみする。ただし、データによっては斜交回転を施しても相関がほとんどない場合もある。斜交回転のほうが柔軟なモデリングを行っているため単純構造になりやすい。単純構造とは、それぞれの項目について、一つの因子の負荷量は高く、それ以外の因子は0に近いような因子負荷行列のこと。

バリマックス回転

バリマックス回転はバリマックス基準を満たす回転方法。バリマックス基準とは、因子負荷行列(の2乗)の列の分散の和を最大にする基準。つまり、特定の因子についてある項目は負荷量の絶対値が高く、別の項目は小さいような因子負荷行列を求める。分散(バリエーション)を最大(マックス)にするためバリマックス基準と呼ばれる。バリマックス回転は直交回転である。

プロマックス回転

因子負荷行列をなるべくバリマックス回転に最小二乗基準で近くづくように回転させる方法。直交の制約を外せるので、プロマックス回転は斜交回転である。

chatGPT による試験対策問題

因子分析は、多変量解析手法の一つで、観測された多次元データを少数の潜在変数(因子)に集約する方法である。この方法により、観測変数間の共通性を特定し、データの構造を明らかにすることができる。

問 1: 基本用語の説明

因子分析において、以下の用語を説明しなさい。

1. 共通因子 (Common Factor)
2. 特殊因子 (Specific Factor)
3. 因子負荷量 (Factor Loading)
4. 共通性 (Communality)

因子分析では観測データ ($\mathbf{x} = (x_1, x_2, \dots, x_n)$) を観測データで共通する要因である共通因子 ($\mathbf{f} = (f_1, f_2, \dots, f_m)$) と各観測データ特有の要因である特殊因子 ($\mathbf{e} = (e_1, e_2, \dots, e_n)$) の線形結合で説明することを考える。この時の共通因子の係数行列 \mathbf{C} を因子負荷量と呼び、共通因子と観測データの相関を表す。因子負荷量の列ベクトルの二乗和を共通性と呼び、観測データの分散のうち共通因子で説明できる部分を表す。

$$\mathbf{x} = \mathbf{C}\mathbf{f} + \mathbf{D}\mathbf{e}$$

\mathbf{D} は特殊因子の重みを対角成分に並べた対角行列。

問 2: 因子分析のプロセス

因子分析を実施する際、以下のプロセスを簡潔に説明しなさい。

1. 因子の抽出方法（例：主成分分析や最尤法）
2. 因子の回転方法（例：バリマックス回転やプロマックス回転）
3. 因子数の決定方法（例：固有値基準やスクリープロット）

1. 因子の抽出方法

因子の抽出では、観測データの背後に存在する潜在変数（因子）を求めます。以下の方法が一般的です。

- **主成分分析**: 観測変数の分散を最大限説明する直交因子を抽出します。データの次元削減にも適用されます。
- **最尤法 (Maximum Likelihood Method)**: 観測データの確率モデルを基に因子負荷量を推定し、分散共分散構造を最適化します。この方法は仮定に基づく推定を行うため、統計的検定が可能です。

2. 因子の回転方法

抽出された因子負荷量を解釈しやすくするため、回転操作を行います。以下の方法がよく使用されます。

- **バリマックス回転 (Varimax Rotation)**: 因子負荷量の分散を最大化する直交回転法で、因子構造を単純化し解釈を容易にします。
- **プロマックス回転 (Promax Rotation)**: 因子間の相関を許容する斜交回転法で、より現実的な因子構造をモデル化します。

3. 因子数の決定方法

適切な因子数を決定することで、過剰適合や不足適合を防ぎます。以下の基準が使用されます。

- **固有値基準 (Eigenvalue Criterion)**: 固有値が 1 以上の因子を採用する基準。分散の説明力が十分にある因子を選択します。
- **スクリープロット (Scree Plot)**: 固有値をプロットし、カーブの屈折点（エルボーポイント）で因子数を決定します。

問 3: 因子負荷量の解釈

以下の観測データを基に因子分析を実施すると仮定する。以下の質問に答えなさい。

1. 因子負荷量が高い変数の特徴をどのように解釈するか。
2. 因子負荷量が低い変数が示す可能性のある問題点を挙げなさい。

1. 因子負荷量が高い変数の特徴をどのように解釈するか

因子負荷量が高い変数は、対応する共通因子によってその変数の分散が大部分説明されていることを意味します。これにより、以下のような解釈が可能です：

- 該当する観測変数は、対応する因子と強い関連性を持ち、その因子の特徴を代表する重要な指標である。
- 因子負荷量の値が高いほど、その変数は因子の定義や解釈において重要な役割を果たす。

例えば、教育分野の因子分析において、数学のテストスコアが「学術能力」という因子に対して高い負荷量を持つ場合、数学のスコアは「学術能力」を説明する重要な要素であると解釈されます。

2. 因子負荷量が低い変数が示す可能性のある問題点

因子負荷量が低い変数は、以下の問題点を示唆している可能性があります：

- 該当因子との関連性が弱い: - 変数が対応する因子とほとんど関連しておらず、その因子では十分に説明できないことを意味します。
- 誤差（特殊因子）の影響が大きい: - 変数の分散の大部分が特殊因子（観測データ特有の要因や誤差）によるものである可能性があります。

- 他の因子との関連性が強い: - 因子負荷量が低い変数は、他の因子に高い負荷量を持つ可能性があり、因子の解釈を複雑にする要因となることがあります。
- 測定の不適合: - 該当する変数が、調査や測定の目的に対して不適切である、または外れ値やデータの歪みによる影響を受けている可能性があります。

例えば、アンケート調査で因子負荷量が低い項目がある場合、それは調査内容と無関係な回答が多かったか、質問項目が曖昧であった可能性を示しています。

問 4: 因子分析の応用

因子分析の結果を用いて、以下の状況における応用例を説明しなさい。

1. 教育分野における学生の学習スタイルの分類
2. マーケティング分野における顧客セグメンテーション

1. 教育分野における学生の学習スタイルの分類

因子分析は、学生の学習スタイルやパフォーマンスを特徴付ける潜在要因を明らかにするために用いられます。例えば：

- アンケート調査を実施し、学生の学習に対する態度や習慣に関する複数の観測項目（例: 「自己管理能力」「協働学習の好み」「独学の頻度」など）を収集します。
- 因子分析を適用することで、「自己主導型学習」「協調型学習」「受動型学習」などの潜在的な学習スタイルを抽出できます。
- これにより、学生をそれぞれの学習スタイルに基づいて分類し、個別化された教育プログラムの設計や支援方針の策定に活用できます。
- 例えば、「自己主導型学習」の学生にはオンライン教材の提供が効果的であり、「協調型学習」の学生にはグループディスカッションの機会を増やすといった具体的な対策が考えられます。

2. マーケティング分野における顧客セグメンテーション

因子分析は、顧客の購買行動や価値観を説明する潜在因子を明らかにすることで、マーケティング戦略の立案に役立ちます。例えば：

- 顧客アンケートや購買データから、さまざまな観測項目（例: 「価格重視」「ブランド重視」「利便性の重要性」「環境意識」など）を収集します。
- 因子分析を実施することで、「価格重視型」「ブランド重視型」「環境意識型」などの顧客セグメントを抽出できます。
- これに基づき、顧客をセグメントごとに分類し、それぞれに適したマーケティング施策を展開します。
- 例えば、「価格重視型」の顧客には割引キャンペーンを中心とした施策を行い、「ブランド重視型」の顧客にはプレミアム製品を強調する広告戦略を採用する、といった具体的なアプローチが可能です。