

因子分析 Factor Analysis

樋口さぶろお <https://hig3.net>

龍谷大学大学院 先端理工学研究科 数理・情報科学コース

計算科学特論 L13(2026-07-13 Mon)

最終更新: Time-stamp: "2026-08-08 Sat 08:50 JST hig"

今日の目標

- 因子分析の考え方を説明できる



L12-Q1

TA Prob and Sol: 移動平均モデル MA(1) の例

1 次の移動平均モデル MA(1) モデルを考える (授業中に使った記号で).
 $\beta_0 = 0$ とする.

- ① 確率変数 X_2 を, ϵ_t, θ_1 で表そう.
- ② 母期待値 $E[X_2^2]$ を σ と θ_1 で表そう.
- ③ 母自己共分散 $\text{Cov}[X_2, X_3]$ を σ と θ_1 で表そう.
- ④ 母自己共分散 $\text{Cov}[X_2, X_4]$ を求めよう.

略解

- ① $X_2 = \epsilon_2 + \theta_1 \epsilon_1$.
- ② $E[(\epsilon_2 + \theta_1 \epsilon_1)^2] = \sigma^2(1 + \theta_1)^2$.
- ③ $E[(\epsilon_2 + \theta_1 \epsilon_1)(\epsilon_3 + \theta_1 \epsilon_2)] = \sigma^2 \theta_1$.
- ④ 0.

L12-Q2

TA Prob and Sol:AR(4) の状態空間モデル

4 次の自己回帰モデル AR(4) モデルを状態空間モデルとして考える (授業の記号で).

- ① 状態ベクトルを書こう.
- ② システムモデル, 観測モデルに現れる行列の, F, G, H を書こう.

略解

$$\textcircled{1} \quad \mathbf{X}_t = \begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{bmatrix}$$

② $\boldsymbol{x}_t = F\boldsymbol{x}_{t-1} + G\boldsymbol{v}_t, y_t = H\boldsymbol{x}_t + 0 \cdot w_t$ とすると

$$F = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, H = [1 \quad 0 \quad 0 \quad 0].$$

多変量データセット

データ行列 x_{ij} ($i = 1, \dots, n, j = \dots, p$)

- j : 10 種競技の種目番号, p : 競技数. 変数.
- i : 選手番号, n : エントリー選手数. 観測, データ点.
- x_{ij} : 選手 i の, 競技 j の得点.
- 横ベクトル $\mathbf{x}_i$: 選手 i の各種目の記録のリスト.

<https://www.rdocumentation.org/packages/factoextra/versions/2.0.0/topics/decathlon2> 得点でなく記録

データ点=確率変数の実現値: 選手 (の競技結果)

因子分析の動機

10 種競技のデータを、「瞬発力」「持久力」のような、各選手の持つ (p よりも少数の) m 個の共通因子で説明したい.

$$\mathbf{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix}.$$

因子分析の確率モデル

確率変数 これらには、 n 個の実現値がある．

p 次元縦ベクトル $\mathbf{X}$ 記録

p 次元縦ベクトル $\boldsymbol{\epsilon}$ ノイズ

m 次元縦ベクトル $\mathbf{f}$ 共通因子 潜在変数

求めるべき係数・定数

定数の p 次元縦ベクトル $\boldsymbol{\mu}_0$

係数（因子負荷行列） Λ : p 行 m 列の行列．

個々のデータに対する $\mathbf{f}$ の実現値

$$\mathbf{X} = \boldsymbol{\mu}_0 + \Lambda \mathbf{f} + \boldsymbol{\epsilon}.$$

仮定：（母期待値，母分散を仮定し，分布は仮定しないが，まずは正規分布を想像）「ぜんぶ「母平均値ゼロ」「独立だったらそうなる共分散ゼロ」」

$$E[\epsilon_j] = 0, \quad \text{Cov}[\epsilon_j, \epsilon_k] = \delta_{jk} \sigma_j^2,$$

$$E[f_\ell] = 0, \quad \text{Cov}[f_\ell, f_h] = \delta_{\ell h} \cdot 1,$$

$$\text{Cov}[\epsilon_{\cdot}, f_\ell] = 0.$$

ミッション

機械学習で言うと教師なし学習. 因子負荷行列を「いい感じに」定めたい.
クラスター分析を思い出そう「いい感じに」クラスターメンバーシップ
を定めたい

主成分分析との比較

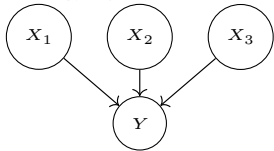
次元圧縮

- 主成分分析では、入力されたデータ行列に対し、主成分という低次元の量がアウトプットされる．変換のための主成分負荷量がある．
- 因子分析では、入力されたデータ行列に対し、共通因子という低次元の量がアウトプットされる．変換のための因子負荷量がある．

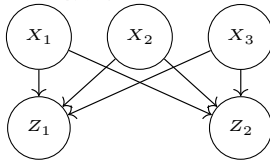
潜在変数の (不) 存在

- 主成分分析では、測定できていない変数はない
- 因子分析では、あるのに直接測定できていない**潜在変数** (latent variable) として共通因子がある
 - ▶ 心理学や教育学では、そのような変数の存在を仮定しがちなので、自然に、因子分析を使うことになる．

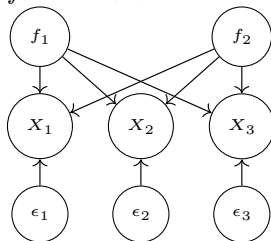
重回帰分析

 Y の教師データがある

主成分分析

 Z の教師データがない

因子分析

 f は潜在変数

L13-Q1

Quiz(共分散行列の変換)

$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}$ を m 個の確率変数のベクトルとする.

$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$ を, 適当な次元の定数の行列 R により $\mathbf{Y} = R\mathbf{X}$ で定める.

X_i の共分散 $\text{Cov}[X_i, X_j]$ を (i, j) 成分とする行列を分散-共分散行列 Σ_{XX} といい, $\text{Cov}[\mathbf{X}, \mathbf{X}]$ と略記する.

- Σ は, 行列 $(\mathbf{X} - \boldsymbol{\mu})^t(\mathbf{X} - \boldsymbol{\mu})$ の各成分の母期待値からなるものであることを示そう.
- $\text{Cov}[\mathbf{Y}, \mathbf{Y}] = R\text{Cov}[\mathbf{X}, \mathbf{X}]^t R$ を示そう.

因子負荷行列の推定

$$E[\mathbf{X}] = \boldsymbol{\mu}_0.$$

標本平均値 $\frac{1}{n} \sum_{i=1}^n x_{ij}$ によって, パラメタ $(\mu_0)_j$ を推定することがよく行われる.

$$\begin{aligned} \text{Cov}[\mathbf{X}, \mathbf{X}] &= \text{Cov}[\Lambda \mathbf{f}, \Lambda \mathbf{f}] + \text{Cov}[\boldsymbol{\epsilon}, \boldsymbol{\epsilon}] \\ &= \Lambda \text{Cov}[\mathbf{f}, \mathbf{f}] \Lambda^t + \text{diag}(\sigma_1^2, \dots, \sigma_p^2) = \Lambda \Lambda^t + \text{diag}(\sigma_1^2, \dots, \sigma_p^2) \end{aligned}$$

不偏標本共分散行列に近くなるように右辺を推定することがよく行われる (主因子法).

$\boldsymbol{\epsilon}, \mathbf{f}$ がすべて多次元正規分布と仮定し, すべてのパラメタを一度に最尤推定することもよく行われる.

ベイズ推定もできる

L13-Q2

Quiz(因子分析の共分散行列)

因子分析のモデル

$$\mathbf{X} = \boldsymbol{\mu}_0 + \Lambda \mathbf{f} + \boldsymbol{\epsilon}.$$

において,

$$\begin{aligned} \text{Cov}[\mathbf{X}, \mathbf{X}] &= \text{Cov}[\Lambda \mathbf{f}, \Lambda \mathbf{f}] + \text{Cov}[\boldsymbol{\epsilon}, \boldsymbol{\epsilon}] \\ &= \Lambda \text{Cov}[\mathbf{f}, \mathbf{f}] \Lambda^t + \text{diag}(\sigma_1^2, \dots, \sigma_p^2) = \Lambda^t \Lambda + \text{diag}(\sigma_1^2, \dots, \sigma_p^2) \end{aligned}$$

を導出しよう.

解の任意性

(誤差がなくても)「正しい推定」は一意ではない (= 識別可能性がない)
推定された Λ , $\mathbf{f}$ に対して, m 次の直交行列 P に対して, $P\Lambda$, ${}^tP\mathbf{f}$ も同じ資格を持つ.

さらに, $\text{Cov}[P\mathbf{f}, P\mathbf{f}] = E$ である.

直交行列 $\Leftrightarrow P {}^tP = E$

解の選択

因子負荷行列は、人間が解釈しやすいものを選ぶ
どっちがいい？

クラスター分析

変数	第 1 因子 総合学力	第 2 因子 理系的
X_1 数学	0.80	0.50
X_1 理科	0.90	0.49
X_2 国語	0.85	-0.51
X_4 社会	0.95	-0.49

変数	第 1 因子 理系力	第 2 因子 文系力
X_1 数学	0.80	0.03
X_1 理科	0.90	0.04
X_2 国語	0.01	0.85
X_4 社会	0.03	0.95

心理学では、後者が好まれる状況が多い。

解の回転

直交回転 $\Lambda' = P\Lambda$ は、 m 次元の因子空間で軸を取り直すことに相当．ある原理（例：因子負荷行列にゼロに近い成分が多い）にしたがって綺麗な軸を選ぶ方法が多数提案されている（○○回転）

斜交回転 P として、直交でない正則行列を許したものを斜交回転という

P が、各列ベクトルの長さが 1 の正則行列のときも、 $\text{Cov}[P\mathbf{f}, P\mathbf{f}] = E$ 以外は成立する

回転じゃないじゃん！ 回転＋シアじゃん！

制約が緩い分、原理（例：因子負荷行列にゼロに近い成分が多い）を徹底しやすい．ただし、 $\mathbf{f}$ の各成分は独立でなくなる（多変量正規分布なら、共分散行列が対角でなくなる）

因子分析の手順

因子分析の手順

- ① 共通因子の個数 p を決める.
- ② 因子負荷量 Λ と, 独自因子 ϵ_j の分散 $\sigma_1^2, \dots, \sigma_m^2$ を推定する.
- ③ 因子空間の回転を行って共通因子の解釈をする.
- ④ 共通因子 (因子得点) $f_1, \dots, f_n$ を求める.