

拡散過程に対する構造方程式モデリングの統計的推測

草野 彰吾 (大阪大学大学院 基礎工学研究科), 内田 雅之 (大阪大学大学院 基礎工学研究科)

構造方程式モデリングとは

構造方程式モデリング (Structural Equation Modeling, SEM) とは**直接観測できない変数である潜在変数間の関係を調べるための統計解析法**である。ここでは SEM の例として、株価データを用いた業界間の関係の検証を考える。研究者は「金融業界の動向 (金融系因子) が製造業界の動向 (製造業系因子) に影響を及ぼす」という仮説を検証したいとする (図 1)。

このような関係を検証したい

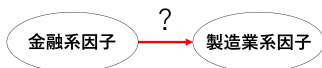


図 1 分析のモチベーション。

しかし、**業界の動向は直接観測することができない**。そこで、これらの因子は株価データを用いて、因子モデルによって観測されると考える (図 2)。

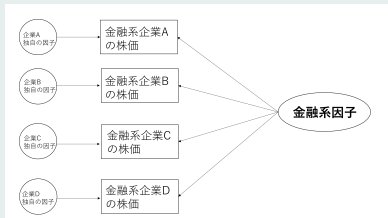


図 2 金融系因子の因子モデルのパス図。

このとき、金融系因子の因子モデルと製造業系因子の因子モデルを結合することで、金融系因子と製造業系因子の関係を株価データを用いて検証することができる (図 3)。

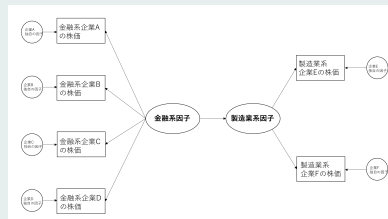


図 3 SEM のパス図。

このように、**SEM は潜在変数間の関係をデータによって検証できる**ことから、心理学をはじめ、経済学、医学、工学など様々な分野で使用されている。

研究の動機

近年、金融データを代表とした高頻度時系列データが容易に取得できるようになったことから、連続時間確率過程に対する統計的推測の研究が盛んに行われている。連続時間確率過程に対する多変量解析法として、これまで主成分分析 (Ait-Sahalia and Xiu (2019)) や因子分析 (Kusano and Uchida (2022)) などが研究されてきたが、連続時間確率過程に対する SEM はほとんど研究されてこなかった。そこで、本研究では高頻度データに基づく拡散過程に対する SEM を提案する。

拡散過程に対する SEM のモデル

確率過程 X_1 が因子モデル:

$$X_{1,t} = \Lambda_{x_1,m} \xi_{m,t} + \delta_{m,t},$$

- m はモデルに割り当てられた番号,
- $\{X_{1,t}\}_{t \geq 0}$ は p_1 次元観測可能確率過程,
- $\{\xi_{m,t}\}_{t \geq 0}$ は $k_{1,m} (\leq p_1)$ 次元潜在因子過程,
- $\{\delta_{m,t}\}_{t \geq 0}$ は p_1 次元潜在因子過程,
- $\Lambda_{v,m} \in \mathbb{R}^{p_1 \times k_{1,m}}$

に従うとする. また, 確率過程 $X_{n,t}$ が因子モデル:

$$X_{2,l} = \Lambda_{X, m} \eta_{m,l} + \varepsilon_{m,l},$$

- $\{X_{2,t}\}_{t \geq 0}$ は p_2 次元観測可能確率過程,
- $\{\eta_{m,t}\}_{t \geq 0}$ は $k_{2,m} (\leq p_2)$ 次元潜在因子過程,
- $\{\varepsilon_{m,t}\}_{t \geq 0}$ は p_2 次元潜在因子過程,
- $\Lambda_{x_{2,m}} \in \mathbb{R}^{p_2 \times k_{2,m}}$

に従うとする。さらに、潜在因子間の関係を

$$\eta_{m,t} = B_{0,m}\eta_{m,t} + \Gamma_m\xi_{m,t} + \zeta_{m,t},$$

- $\{\zeta_{m,t}\}_{t \geq 0}$ は $k_{2,m}$ 次元潜在因子過程,
- $B_{0,m} \in \mathbb{R}^{k_{2,m} \times k_{2,m}}$ は対角成分が 0 の負荷行列, $\Gamma_m \in \mathbb{R}^{k_{2,m} \times k_{1,m}}$,
- $\Psi_m = \mathbb{I}_{k_{1,m}} - B_{0,m}$

とする。ただし、 $I_{k_2 m}$ は $k_2 m$ 次元単位行列とする。ここで、 $\{\xi_{m,j}\}_{j \geq 0}$ は拡散過程

$$d\xi_{m,t} = B_{1,m}(\xi_{m,t})dt + S_{1,m}dW_{1,t}, \quad (t \in T), \quad \xi_{m,0} = c_{1,m}.$$

- $B_{1,m} : \mathbb{R}^{k_{1,m}} \rightarrow \mathbb{R}^{k_{1,m}}, S_{1,m} \in \mathbb{R}^{k_{1,m} \times r_1}, c_{1,m} \in \mathbb{R}^{k_{1,m}},$
- $\Sigma_{\xi\xi,m} = S_{1,m} S_{1,m}^\top,$
- W_1 は r_1 次元標準ブラウン運動

に従い, $\{\delta_{m,t}\}_{t \geq 0}$ は拡散過程

$$d\delta_{m,t} = B_{\gamma_m}(\delta_{m,t})dt + S_{\gamma_m}dW_{\gamma_t}, \quad (t \in T), \quad \delta_{m,0} = c_{\gamma_m}.$$

- $B_{2,m} : \mathbb{R}^{p_1} \rightarrow \mathbb{R}^{p_1}$, $S_{2,m} \in \mathbb{R}^{p_1 \times r_2}$, $c_{2,m} \in \mathbb{R}^{p_1}$,
- $\Sigma_{\delta\delta,m} = S_{2,m} S_{2,m}^\top$,
- W_2 は r_2 次元標準ブラウン運動

に従い, $\{\varepsilon_{m,t}\}_{t \geq 0}$ は拡散過程

$$d\varepsilon_{m,t} = B_{3,m}(\varepsilon_{m,t})dt + S_{3,m}dW_{3,t}, \quad (t \in T), \quad \varepsilon_{m,0} = c_{3,m}.$$

- $B_{3,m} : \mathbb{R}^{P_2} \rightarrow \mathbb{R}^{P_2}$, $S_{3,m} \in \mathbb{R}^{P_2 \times r_3}$, $c_{3,m} \in \mathbb{R}^{P_2}$,
- $\Sigma_{EE,m} = S_{3,m} S_{3,m}^T$,
- $W_{3,t}$ は r_3 次元標準ブラウン運動

に従い, $\{Z_m\}_{m \geq 0}$ は拡散過程

$$d\tilde{z}_{m,t} = B_{4,m}(\tilde{z}_{m,t})dt + S_{4,m}dW_{4,t}, \quad (t \in T), \quad \tilde{z}_{m,0} = c_{4,m}.$$

- $$\begin{aligned} B_{4,m} : \mathbb{R}^{k_{2,m}} &\rightarrow \mathbb{R}^{k_{2,m}}, \quad S_{4,m} \in \mathbb{R}^{k_{2,m} \times r_A}, \quad c_{4,m} \in \mathbb{R}^{k_{2,m}}, \\ \Sigma_{\zeta\zeta,m} &= S_{4,m} S_{4,m}^\top, \\ W_{4,\cdot} &\text{は } r_A \text{ 次元標準ブラウン運動} \end{aligned}$$

に従うと仮定する。ただし、 $W_{1,i}$, $W_{2,i}$, $W_{3,i}$, $W_{4,i}$ は独立であり、 $\mathbf{T}$ は行列の転置を表すとする。また、 $X_i = (X_{1,i}^T, X_{2,i}^T)^T$ とし、 $\{X_{ij}^T\}_{j=1}^n$ を離散観測データ、 h_n を刻み幅とし、 $i_n^* = ih_n$, $T = nh_n$ とする。以下、 $\{\xi_{m,i}\}_{i \geq 0}$, $\{\delta_{m,i}\}_{i \geq 0}$, $\{\varepsilon_{m,i}\}_{i \geq 0}$, $\{\zeta_{m,i}\}_{i \geq 0}$ がエルゴード的拡散過程 (Kessler (1997)) である場合を考える。

パラメータと分散構造

SEM では、分散構造に関するパラメータ

$$\Lambda_{X1, m}, \Lambda_{X2, m}, \Gamma_m, \Psi_m, \Sigma_{\xi\xi, m}, \Sigma_{\delta\delta, m}, \Sigma_{\xi\delta, m}, \Sigma_{\zeta\zeta, m} \quad (1)$$

を推定する。(1)のうち、未知で重複していない要素をまとめて未知パラメータ $\theta_m \in \Theta_m$ を定義する。ただし、 $\Theta_m \in \mathbb{R}^{q_m}$ はコンパクトな凸集合とする。また、パラメータの真値を $\theta_{m,0} \in \text{Int } \Theta_m$ とする。さらに、分散構造を

$$\Sigma_m(\theta_m) = \begin{pmatrix} \Sigma_{X_1 X_1, m}(\theta_m) & \Sigma_{X_1 X_2, m}(\theta_m) \\ \Sigma_{X_2 X_1, m}(\theta_m)^\top & \Sigma_{X_2 X_2, m}(\theta_m) \end{pmatrix}$$

と定義する。ただし、

$$\Sigma_{Y, Y, m}(\theta_m) = \Lambda_{Y, m} \Sigma_{\varepsilon \varepsilon, m} \Lambda_{Y, m}^T + \Sigma_{\delta \delta, m}.$$

$$\Sigma_{Y, Y, m}(\theta_m) = \Lambda_{Y, m} \Sigma_{\mathcal{E}, m} \Gamma_m^T \Psi_m^{-1} \Lambda_{Y, m}^T$$

$$\Sigma_{Y_m, Y_m}(\theta_m) = \Lambda_{Y_m} \Psi_m^{-1} (\Gamma_m \Sigma_{\xi\xi m} \Gamma_m^T + \Sigma_{\eta\eta m}) \Psi_m^{-1T} \Lambda_{Y_m}^T + \Sigma_{cc m}$$

である。拡散過程に対する SEM では、 $\Sigma_m(\theta_m)$ の推定に実現共分散

$$Q_{XX} = \frac{1}{T} \sum_{i=1}^n (X_{t_i^e} - X_{t_{i-1}^e})(X_{t_i^e} - X_{t_{i-1}^e})^\top$$

を用いる。

数値シミュレーション

真のモデル

確率過程 $X_{1,t}$ が因子モデル:

$$X_{1,t} = \Lambda_{x_1,m} \xi_{m,t} + \delta_{m,t},$$

- $\{X_{1,t}\}_{t \geq 0}$ は 4 次元観測可能確率過程,
- $\{\xi_{m,t}\}_{t \geq 0}$ は 2 次元潜在因子過程,
- $\{\delta_{m,t}\}_{t \geq 0}$ は 4 次元潜在因子過程,
- $\Lambda_{x_1,m} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (\Lambda_{x_1,m})_{21} & 0 \\ 0 & 1 & 0 \\ 0 & (\Lambda_{x_1,m})_{42} & 0 \end{pmatrix}^T \in \mathbb{R}^{4 \times 2}$

に従うとする。また、確率過程 $X_{2,t}$ が因子モデル:

$$X_{2,t} = \Lambda_{x_2,m} \eta_{m,t} + \varepsilon_{m,t},$$

- $\{X_{2,t}\}_{t \geq 0}$ は 2 次元観測可能確率過程,
- $\{\eta_{m,t}\}_{t \geq 0}$ は 1 次元潜在因子過程,
- $\{\varepsilon_{m,t}\}_{t \geq 0}$ は 2 次元潜在因子過程,
- $\Lambda_{x_2,m} = (1, (\Lambda_{x_2,m})_{21})^T \in \mathbb{R}^{2 \times 1}$

に従うとする。さらに、潜在因子間の関係を

$$\eta_{m,t} = \Gamma_m \xi_{m,t} + \zeta_{m,t},$$

- $\{\zeta_{m,t}\}_{t \geq 0}$ は 1 次元潜在因子過程,
- $\Gamma_m \in \mathbb{R}^{1 \times 2}$

とする。また、拡散過程 $\{\xi_{m,t}\}_{t \geq 0}$ は 2 次元 Ornstein-Uhlenbeck (OU) 過程

$$d\xi_{m,t} = -(A_{1,m}\xi_{m,t} - \mu_{1,m})dt + S_{1,m}dW_{1,t}, \quad (t \in [0, T]), \quad \xi_{m,0} = 0,$$

- $A_{1,m} \in \mathbb{R}^{2 \times 2}$, $\mu_{1,m} \in \mathbb{R}^2$, $W_{1,t}$ は 2 次元標準ブラウン運動,
- $S_{1,m} \in \mathbb{R}^{2 \times 2}$, $\Sigma_{\xi\xi,m} = S_{1,m}S_{1,m}^T$ は正定値行列

に従い、拡散過程 $\{\delta_{m,t}\}_{t \geq 0}$ は 4 次元 OU 過程

$$d\delta_{m,t} = -(A_{2,m}\delta_{m,t} - \mu_{2,m})dt + S_{2,m}dW_{2,t}, \quad (t \in [0, T]), \quad \delta_{m,0} = 0,$$

- $A_{2,m} \in \mathbb{R}^{4 \times 4}$, $\mu_{2,m} \in \mathbb{R}^4$, $W_{2,t}$ は 4 次元標準ブラウン運動,
- $S_{2,m} \in \mathbb{R}^{4 \times 4}$, $\Sigma_{\delta\delta,m} = S_{2,m}S_{2,m}^T$ は正定値対角行列

に従い、拡散過程 $\{\varepsilon_{m,t}\}_{t \geq 0}$ は 2 次元 OU 過程

$$d\varepsilon_{m,t} = -(A_{3,m}\varepsilon_{m,t} - \mu_{3,m})dt + S_{3,m}dW_{3,t}, \quad (t \in [0, T]), \quad \varepsilon_{m,0} = 0,$$

- $A_{3,m} \in \mathbb{R}^{2 \times 2}$, $\mu_{3,m} \in \mathbb{R}^2$, $W_{3,t}$ は 2 次元標準ブラウン運動,
- $S_{3,m} \in \mathbb{R}^{2 \times 2}$, $\Sigma_{\varepsilon\varepsilon,m} = S_{3,m}S_{3,m}^T$ は正定値対角行列

に従い、拡散過程 $\{\zeta_{m,t}\}_{t \geq 0}$ は 1 次元 OU 過程

$$d\zeta_{m,t} = -(A_{4,m}\zeta_{m,t} - \mu_{4,m})dt + S_{4,m}dW_{4,t}, \quad (t \in [0, T]), \quad \zeta_{m,0} = 0,$$

- $A_{4,m} \in \mathbb{R}$, $\mu_{4,m} \in \mathbb{R}$, $S_{4,m} > 0$, $W_{4,t}$ は 1 次元標準ブラウン運動

に従う。このとき、パラメータは

$$\theta_m = ((\Lambda_{x_1,m})_{21}, (\Lambda_{x_1,m})_{42}, (\Lambda_{x_2,m})_{21}, (\Gamma_m)_{11}, (\Gamma_m)_{12}, (\Sigma_{\xi\xi,m})_{11}, (\Sigma_{\xi\xi,m})_{12}, (\Sigma_{\xi\xi,m})_{22}, (\Sigma_{\delta\delta,m})_{11}, (\Sigma_{\delta\delta,m})_{22}, (\Sigma_{\delta\delta,m})_{33}, (\Sigma_{\delta\delta,m})_{44}, (\Sigma_{\varepsilon\varepsilon,m})_{11}, (\Sigma_{\varepsilon\varepsilon,m})_{22}, \Sigma_{\zeta\zeta,m})^T \in \Theta_m$$

となる。ただし、パラメータ空間は

$$\Theta_m = [-100, -0.1] \cup [0.1, 100]^5 \times [0.1, 100] \times [-100, -0.1] \cup [0.1, 100] \times [0.1, 100]^8$$

とする。また、真値を

$$\theta_{m,0} = (2, 3, 3, 1, 2, 2, 2, 4, 1, 4, 4, 1, 1, 9, 4)^T \in \Theta_m$$

とする。このモデルのパス図は図 4 のようになる。

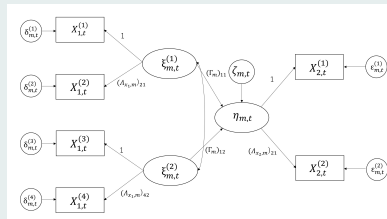


図 4 真のモデルのパス図。

シミュレーション設定

サンプリングを $(n, h_n, T) = (10^6, 10^{-4}, 10^3)$ とし、イテレーションを 10000 とし、最適化の初期値を真値 $\theta_{m,0}$ とした。また、適合度検定の有意水準は $\alpha = 0.05$ とする。

誤特定モデル ($m^* = M_1$)

因子数を $k_{1,M_1} = 1$, $k_{2,M_1} = 1$ とする。真のモデルの因子数は $k_{1,m} = 2$, $k_{2,m} = 1$ であったから、このモデルは因子数を誤特定していることに注意する。また、識別可能性のために、制約

- $\Lambda_{x_1, M_1} = (1, (\Lambda_{x_1, M_1})_{21}, (\Lambda_{x_1, M_1})_{31}, (\Lambda_{x_1, M_1})_{41})^\top \in \mathbb{R}^{4 \times 1}$,
- $\Lambda_{x_2, M_1} = (1, (\Lambda_{x_2, M_1})_{21})^\top \in \mathbb{R}^{2 \times 1}$,
- $\Sigma_{\delta\delta, M_1} \in \mathbb{R}^{4 \times 4}$ と $\Sigma_{\varepsilon\varepsilon, M_1} \in \mathbb{R}^{2 \times 2}$ は正定値対角行列

を置く。このとき、パラメータは

$$\theta_{M_1} = ((\Lambda_{x_1, M_1})_{21}, (\Lambda_{x_1, M_1})_{31}, (\Lambda_{x_1, M_1})_{41}, (\Lambda_{x_2, M_1})_{21}, \Gamma_{M_1}, \Sigma_{\varepsilon\varepsilon, M_1}, (\Sigma_{\delta\delta, M_1})_{11}, (\Sigma_{\delta\delta, M_1})_{22}, (\Sigma_{\delta\delta, M_1})_{33}, (\Sigma_{\delta\delta, M_1})_{44}, (\Sigma_{\varepsilon\varepsilon, M_1})_{11}, (\Sigma_{\varepsilon\varepsilon, M_1})_{22}, \Sigma_{\zeta\zeta, M_1})^\top \in \theta_{M_1}$$

となる。ただし、パラメータ空間は

$$\theta_{M_1} = \{[-100, -0.1] \cup [0.1, 100]\}^5 \times [0.1, 100]^8$$

である。また、分散構造は

$$\Sigma_{M_1}(\theta_{M_1}) = \begin{pmatrix} \Sigma_{X_1 X_1, M_1}(\theta_{M_1}) & \Sigma_{X_1 X_2, M_1}(\theta_{M_1}) \\ \Sigma_{X_1 X_2, M_1}(\theta_{M_1})^\top & \Sigma_{X_2 X_2, M_1}(\theta_{M_1}) \end{pmatrix}$$

である。ただし、

$$\Sigma_{X_1 X_1, M_1}(\theta_{M_1}) = \Lambda_{x_1, M_1} \Sigma_{\zeta\zeta, M_1} \Lambda_{x_1, M_1}^\top + \Sigma_{\delta\delta, M_1},$$

$$\Sigma_{X_1 X_2, M_1}(\theta_{M_1}) = \Lambda_{x_1, M_1} \Sigma_{\zeta\zeta, M_1} \Gamma_{M_1} \Lambda_{x_2, M_1}^\top,$$

$$\Sigma_{X_2 X_2, M_1}(\theta_{M_1}) = \Lambda_{x_2, M_1} (\Gamma_{M_1}^2 \Sigma_{\zeta\zeta, M_1} + \Sigma_{\zeta\zeta, M_1}) \Lambda_{x_2, M_1}^\top + \Sigma_{\varepsilon\varepsilon, M_1}$$

である。このモデルのパス図は図 5 のようになる。

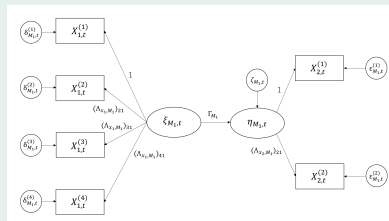


図 5 誤特定モデルのパス図。

図 6 より、モデルが特定された場合、つまり定理 1 の設定 ($m^* = m$) において、 $\hat{\theta}_{m^*,n}^{(1)} = \hat{\theta}_{m,n}^{(1)}$ の漸近正規性がわかる。ただし、赤線は理論分布である。また、図 7 より、定理 2 の設定 ($m^* = m$) において、帰無仮説の下で、検定統計量 $T_{m^*,n} = T_{m,n} = F_{m,n}(\hat{\theta}_{m,n})$ が自由度 6 のカイ二乗分布に分布収束していることがわかる。ただし、赤線は自由度 6 のカイ二乗分布である。さらに、誤特定モデルの場合 ($m^* = M_1$) に対する検定問題

$$\begin{cases} H_0 : \Sigma_m(\theta_m) = \Sigma_{M_1}(\theta_{M_1}), \\ H_1 : \Sigma_m(\theta_m) \neq \Sigma_{M_1}(\theta_{M_1}) \end{cases}$$

を考える。このとき、検定統計量は $T_{m^*,n} = T_{M_1,n} = F_{M_1,n}(\hat{\theta}_{M_1,n})$ であり、棄却域は

$$\{T_{M_1,n} > \chi_6^2(0.05) = 15.51\}$$

となる。数値シミュレーションの結果、10000 回中 10000 回帰無仮説が棄却された。この結果は対立仮説が正しい場合、帰無仮説を正しく棄却できることを示しており、**検定の一致性が成り立っている**ことがわかる。

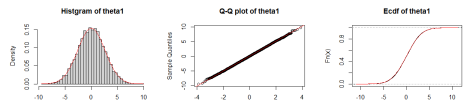


図 6 $\sqrt{n}(\hat{\theta}_{m,n}^{(1)} - \theta_{m,0}^{(1)})$ のヒストグラム (左)、Q-Q プロット (中央)、経験分布関数 (右)。

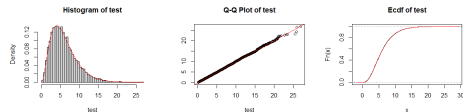


図 7 $T_{m,n}$ のヒストグラム (左)、Q-Q プロット (中央)、経験分布関数 (右)。

参考文献

- Alt Sahalia, Y. and Xiu, D. (2019). Principal component analysis of high-frequency data. *Journal of the American Statistical Association*, **114**(525), 287-303.
- Harville, D. A. (1998). *Matrix algebra from a statistician's perspective*. Taylor & Francis.
- Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, **24**(2), 211-229.
- Kuzano, S. and Uchida, M. (2022). Statistical inference in factor analysis for diffusion processes from discrete observations. *arXiv preprint arXiv:2202.01481*.