

強スパイク固有値モデルにおける高次元共分散行列の推定

筑波大学・数理物質科学 小西 啓介
 筑波大学・数理物質系 矢田 和善
 筑波大学・数理物質系 青嶋 誠

本講演では、高次元共分散行列とその逆行列の推定を考える。データを潜在空間とノイズ空間に分けて推定を行うことで、非スパースな高次元共分散行列の推定法を与える。

平均にゼロベクトル、共分散行列に p 次元正定値対称行列 Σ をもつ母集団を考える。 Σ の固有値分解を $\Sigma = \sum_{j=1}^p \lambda_j \mathbf{h}_j \mathbf{h}_j^T$, $\lambda_1 \geq \dots \geq \lambda_p (> 0)$ とし、次のスパイクモデルを仮定する。

$$\lambda_j = c_j p^{\alpha_j} \quad (j = 1, \dots, m) \quad \text{and} \quad \lambda_j = c_j \quad (j = m+1, \dots, p) \quad (1)$$

ただし、 $c_j (> 0)$, $\alpha_j \in (1/2, 1]$ はともに p に依存しない未知の実数、 m は p に依存しない未知の自然数とする。モデル (1) は Aoshima and Yata (2018, Sinica) で提唱された強スパイク固有値モデルの一つであることに注意する。この母集団から n 個のデータベクトル $\mathbf{x}_1, \dots, \mathbf{x}_n$ を無作為に抽出し、 $p \times n$ データ行列 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ を定義する。標本共分散行列 $\mathbf{S} = n^{-1} \mathbf{X}^T \mathbf{X}$ の固有値分解を $\mathbf{S} = \sum_{j=1}^p \hat{\lambda}_j \hat{\mathbf{h}}_j \hat{\mathbf{h}}_j^T$ で与える。ただし、 $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p (\geq 0)$ である。さらに、双対標本共分散行列 $\mathbf{S}_D = n^{-1} \mathbf{X}^T \mathbf{X}$ の固有値分解を $\mathbf{S}_D = \sum_{j=1}^n \hat{\lambda}_j \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j^T$ で与える。

高次元共分散行列の推測について、Bickel and Levina (2008, AS) によるスパース推測法が有名である。しかしながら、スパース推測法が適さない事例も多くある。Aoshima and Yata (2018) の研究により、ゲノム等の高次元データの構造は非スパースであることが確認され、稠密なデータ構造により、いくつかの固有値が強くスパイクすることが例示された。それゆえ、強スパイク固有値モデルのもと、高次元共分散行列の推測を行うことが重要となる。Fan et al. (2013, JRSS-B) では、データにモデル (1) を仮定し、従来型 PCA に基づく高次元共分散行列とその逆行列の推定法 (POET) を提案した。しかしながら、従来型の PCA は高次元のもとで不一致性を持つことが知られている。一方で、Yata and Aoshima (2012, JMA) は、高次元のもとで一致性を持つ新しい PCA として、ノイズ掃き出し法を提案した。Wang and Fan (2017, AS) は、ノイズ掃き出し法による固有値推定を用いて POET を改良し、新たな高次元共分散行列の推定法 (S-POET) を提案した。しかしながら、S-POET の高次元固有ベクトル推定には不一致性が残り、さらに、 Σ^{-1} の推測までその方法論を拡張できていなかった。

本講演では、ノイズ掃き出し法とスパース推測法を融合させ、新たな高次元共分散行列とその逆行列の推定法を提案する。ノイズ掃き出し法による固有値と固有ベクトルの推定量は以下で与えられる。

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\text{tr}(\mathbf{S}_D) - \sum_{i=1}^j \hat{\lambda}_i}{n - j} \quad \text{and} \quad \tilde{\mathbf{h}}_j = \frac{\mathbf{X} \hat{\mathbf{u}}_j}{(n \tilde{\lambda}_j)^{1/2}}, \quad j = 1, \dots, n-1$$

モデル (1) において、強スパイクする最初の m 次元の潜在空間をノイズ掃き出し法で推定し、残りのノイズ空間をスパース推測することで新たな推定法を構築する。 $T(\cdot)$ を Bickel and Levina (2008) で与えられたスパース化作用素とし、 $\tilde{\mathbf{w}}_i = (\mathbf{I}_p - \sum_{j=1}^m \tilde{\mathbf{h}}_j \tilde{\mathbf{h}}_j^T) \mathbf{x}_i$ なるデータ変換を考え、 $\tilde{\Sigma}_2 = \sum_{i=1}^n \tilde{\mathbf{w}}_i \tilde{\mathbf{w}}_i^T / n$ に対して $T(\tilde{\Sigma}_2)$ の固有値分解を $T(\tilde{\Sigma}_2) = \sum_{j=1}^p \hat{\lambda}_j \hat{\mathbf{h}}_j \hat{\mathbf{h}}_j^T$ で与える。ただし、 $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p (> 0)$ である。そのとき、 Σ^{-1} の推定量を以下のように与える。

$$\tilde{\Sigma}^{-1} = \sum_{j=1}^m \tilde{\lambda}_j^{-1} \tilde{\mathbf{h}}_j \tilde{\mathbf{h}}_j^T + \left(\mathbf{I}_p - \sum_{j=1}^m \tilde{\mathbf{h}}_j \tilde{\mathbf{h}}_j^T \right) \left(\sum_{j=1}^{p-m} \hat{\lambda}_j^{-1} \hat{\mathbf{h}}_j \hat{\mathbf{h}}_j^T \right) \left(\mathbf{I}_p - \sum_{j=1}^m \tilde{\mathbf{h}}_j \tilde{\mathbf{h}}_j^T \right)$$

当日は、 $\tilde{\Sigma}^{-1}$ の高次元一致性を示し、POET との理論的な比較を行う。さらに、本方法論による Σ の推定量も与え、数値実験と実データ解析も交えて、提案手法の性能を検証する。