Discussion: Posterior Asymptotics of High-Dimensional Spiked Covariance Model

By Kim S, Lee K, Park S and Lee Jaeyong

Luis Carvalho

Dept of Mathematics and Statistics
Boston University
Visiting Professor at Università Cattolica del Sacro Cuore, Italy

Objective Bayes Conference
Athens, June 11th, 2025

Recap

- **Setup**: given n observations $X_i \overset{\text{iid}}{\sim} N_p(0, \Sigma)$, what are the posterior convergence rates assuming that Σ has a good spectral gap?

Recap

- **Setup**: given n observations $X_i \overset{\text{iid}}{\sim} N_p(0, \Sigma)$, what are the posterior convergence rates assuming that Σ has a good spectral gap?
- Under a *generalized shrinkage inverse Wishart prior* and given some technical conditions, more importantly,
  - A high dimensional setting, $\mathbf{n/p \to 0}$
  - The **k** largest (spiked) eigenvalues of the true covariance matrix are sufficiently separated by a constant value
  - The non-spiked eigenvalues are bounded away from zero and above

  the authors provide posterior convergence rates borrowing from the asymptotics on the sample covariance by Wang and Fan (2017) and recent results based on the inverse Wishart prior (Lee *et al*, 2024)

Recap

- Berger, Sun, and Song (2020) proposed marrying the inverse Wishart prior with the **reference** prior (Yang and Berger, 1994): with $\Sigma = U \Lambda U^\top$, $U \in O(p)$, $\Lambda = \text{Diag}\{\lambda_i\}$ and $\mathcal{A} = \{(\lambda_1, \ldots, \lambda_p) : \lambda_1 > \cdots > \lambda_p > 0\}$,

$$\pi(\Sigma) \propto \frac{\text{etr}(-\Sigma^{-1} H/2)}{|\Sigma|^a [\prod_{i<j} (\lambda_i - \lambda_j)]^b} I(\mathcal{A})$$

BOSTON UNIVERSITY

Recap

- Berger, Sun, and Song (2020) proposed marrying the inverse Wishart prior with the **reference** prior (Yang and Berger, 1994): with $\Sigma = U\Lambda U^\top$, $U \in O(p)$, $\Lambda = \text{Diag}\{\lambda_i\}$ and $\mathcal{A} = \{(\lambda_1, \ldots, \lambda_p) : \lambda_1 > \cdots > \lambda_p > 0\}$,

$$\pi(\Sigma) \propto \frac{\text{etr}(-\Sigma^{-1}H/2)}{|\Sigma|^{a}[\prod_{i<j}(\lambda_i - \lambda_j)]^b} I(\mathcal{A})$$

- Thus,

$$\pi(U, \Lambda) \propto \frac{\text{etr}(-U\Lambda^{-1}U^\top H/2)}{|\Sigma|^{a}[\prod_{i<j}(\lambda_i - \lambda_j)]^{b-1}} I(\mathcal{A})$$

and, if $H = hI_p$ and $b = 1$,

$$\pi(U, \Lambda) \propto \prod_{i=1}^{p} \lambda_i^{-a} e^{-\frac{h}{2\lambda_i}} I(\mathcal{A})$$

Recap

- Here, the authors assume a **generalized** shrinkage prior: with $H = hI_p$, $b = 1$, and **unordered** eigenvalues

$$\pi(U, \Lambda) \propto \prod_{i=1}^{p} \lambda_i^{-a_i} e^{-\frac{h}{2\lambda_i}}$$

Recap

- Here, the authors assume a **generalized** shrinkage prior: with $H = hI_p$, $b = 1$, and **unordered** eigenvalues

$$\pi(U, \Lambda) \propto \prod_{i=1}^{p} \lambda_i^{-a_i} e^{-\frac{h}{2\lambda_i}}$$

- With the spectral decomposition of the sample covariance as $\sum_i X_i X_i^\top / n = QWQ^\top$ and setting $\Gamma = U^\top Q$ the posterior is

$$\pi(U, \Lambda | X) \propto \prod_{i=1}^{p} \lambda_i^{-a_i - n/2} \mathrm{etr}(-\Lambda^{-1} \Gamma^\top (hI_p + W) \Gamma / 2)$$

BOSTON UNIVERSITY

Recap

- Here, the authors assume a **generalized** shrinkage prior: with $H = hI_p$, $b = 1$, and **unordered** eigenvalues

$$\pi(U, \Lambda) \propto \prod_{i=1}^{p} \lambda_i^{-a_i} e^{-\frac{h}{2\lambda_i}}$$

- With the spectral decomposition of the sample covariance as $\sum_i X_i X_i^\top / n = QWQ^\top$ and setting $\Gamma = U^\top Q$ the posterior is

$$\pi(U, \Lambda | X) \propto \prod_{i=1}^{p} \lambda_i^{-a_i - n/2} \text{etr}(-\Lambda^{-1} \Gamma^\top (hI_p + W) \Gamma / 2)$$

- Gibbs sampling seems to be then straightforward: with c_i the i-th diagonal entry of $\Gamma^\top (hI_p + W)\Gamma$,
  - $\lambda_i | U, X \overset{\text{iid}}{\sim} \text{Inv-Gamma}(a_i + n/2 - 1, c_i/2)$
  - $\Gamma | \Lambda, X \sim \text{Bingham}(\Lambda^{-1}, hI_p + W)$

BOSTON UNIVERSITY

Questions

- Let's take a look at the induced prior on Σ: since for any permutation $\sigma \in \mathcal{P}_p$, with Λ the ordered eigenvalues of Σ,

$$\Sigma = U\Lambda U^\top = (UP_\sigma)(P_\sigma^\top \Lambda P_\sigma)(UP_\sigma)^\top := U_{\cdot,\sigma}\Lambda_\sigma U_{\cdot,\sigma}^\top$$

with $\varsigma(\Sigma) := \{\sigma \in \mathcal{P}_p : U_\sigma \Lambda_\sigma U_\sigma^\top = U\Lambda U^\top = \Sigma\}$,

$$\pi(\Sigma) = \pi(\varsigma(\Sigma)) \propto \sum_{\sigma \in \mathcal{P}_p} \prod_{i=1}^{p} \lambda_{\sigma(i)}^{-a_i} e^{-\frac{h}{2\lambda_i}} \prod_{i<j} (\lambda_i - \lambda_j)^{-1} I(\mathcal{A})$$

BOSTON UNIVERSITY

Questions

- Let's take a look at the induced prior on Σ: since for any permutation $\sigma \in \mathcal{P}_p$, with Λ the ordered eigenvalues of Σ,

$$\Sigma = U\Lambda U^\top = (UP_\sigma)(P_\sigma^\top \Lambda P_\sigma)(UP_\sigma)^\top := U_{\cdot,\sigma}\Lambda_\sigma U_{\cdot,\sigma}^\top$$

with $\varsigma(\Sigma) := \{\sigma \in \mathcal{P}_p : U_\sigma \Lambda_\sigma U_\sigma^\top = U\Lambda U^\top = \Sigma\}$,

$$\pi(\Sigma) = \pi(\varsigma(\Sigma)) \propto \sum_{\sigma \in \mathcal{P}_p} \prod_{i=1}^{p} \lambda_{\sigma(i)}{}^{-a_i} e^{-\frac{h}{2\lambda_i}} \prod_{i<j} (\lambda_i - \lambda_j)^{-1} I(\mathcal{A})$$

- So, more complex than IW, which might spell trouble in practical applications?

Questions

- Let's take a look at the induced prior on Σ: since for any permutation $\sigma \in \mathcal{P}_p$, with Λ the ordered eigenvalues of Σ,

$$\Sigma = U\Lambda U^\top = (UP_\sigma)(P_\sigma^\top \Lambda P_\sigma)(UP_\sigma)^\top := U_{\cdot,\sigma}\Lambda_\sigma U_{\cdot,\sigma}^\top$$

with $\varsigma(\Sigma) := \{\sigma \in \mathcal{P}_p : U_\sigma \Lambda_\sigma U_\sigma^\top = U\Lambda U^\top = \Sigma\}$,

$$\pi(\Sigma) = \pi(\varsigma(\Sigma)) \propto \sum_{\sigma \in \mathcal{P}_p} \prod_{i=1}^{p} \lambda_{\sigma(i)}{}^{-a_i} e^{-\frac{h}{2\lambda_i}} \prod_{i<j} (\lambda_i - \lambda_j)^{-1} I(\mathcal{A})$$

- So, more complex than IW, which might spell trouble in practical applications?
- The complexity that arises from identifying the spectral decomposition of Σ cannot be avoided: no free lunch? See, e.g. (Papastamoulis and Ntzoufras, 2022)

BOSTON UNIVERSITY

Questions

- There is another source of non-identifiability: when $p > n$, $W = \text{Diag}(n\hat{\lambda}_1, \ldots, n\hat{\lambda}_n, 0, \ldots, 0)$, so with $V = I_n \oplus V_2$ for $V_2 \in O(p-n)$,

  $$\Gamma^\top (hI_p + W)\Gamma = \Gamma^\top V^\top V(hI_p + W)V^\top V\Gamma := \tilde{\Gamma}^\top (hI_p + W)\tilde{\Gamma}$$

  that is, $\Gamma = [\Gamma_1 \quad \Gamma_2]$ is only identified up to its first n columns in Γ_1

BOSTON UNIVERSITY

Questions

- There is another source of non-identifiability: when $p > n$, $W = \mathrm{Diag}(n\hat{\lambda}_1, \ldots, n\hat{\lambda}_n, 0, \ldots, 0)$, so with $V = I_n \oplus V_2$ for $V_2 \in O(p-n)$,

$$\Gamma^\top (hI_p + W)\Gamma = \Gamma^\top V^\top V(hI_p + W)V^\top V\Gamma := \tilde{\Gamma}^\top (hI_p + W)\tilde{\Gamma}$$

  that is, $\Gamma = [\Gamma_1 \quad \Gamma_2]$ is only identified up to its first n columns in Γ_1

- So it seems that $\Gamma \notin O(p)$ but

$$\Gamma \in O(p)/(O(p-n) \times \mathcal{P}_p) = \mathcal{V}_{p,n}/\mathcal{P}_n$$

  almost the Grassmannian $\mathcal{G}_{n,p}$!

Questions

- In the main two lemmas that provide shrinkage rates of the posterior the authors define a set $D_{\epsilon,l}$,

$$D_{\epsilon,l} = \left\{ \Gamma \in O(p) \, : \, \min_{\sigma \in \mathcal{P}_k} \inf_{V_2 \in O(p-k)} \|(P_\sigma \oplus V_2) - \Gamma\|_F < \epsilon \right\}$$

  that seems to identify Γ; is that the main motivation?
- If so, wouldn't it be better to just identify Γ directly?
- This should also impact the normalizing constant of the prior, which should be important when selecting k to avoid multiplicity issues (Scott and Berger, 2010)

BOSTON UNIVERSITY

Questions

- The authors recommend that a_i be specified based on the data for faster convergence rates: with $\hat{\lambda}_i$ the eigenvalues of the sample covariance and $t \in [\hat{\lambda}_{k+1}, \hat{\lambda}_n]$,

$$a_i = \frac{nt}{2(\hat{\lambda}_i - t)} + 2, \quad i = 1, \dots, k$$

Is there a way to specify them in an **objective** way?

Questions

- The authors recommend that a_i be specified based on the data for faster convergence rates: with $\hat{\lambda}_i$ the eigenvalues of the sample covariance and $t \in [\hat{\lambda}_{k+1}, \hat{\lambda}_n]$,

$$a_i = \frac{nt}{2(\hat{\lambda}_i - t)} + 2, \quad i = 1, \ldots, k$$

  Is there a way to specify them in an **objective** way?
- There are clear connections to factor analysis [e.g. (Anderson, 1964), (Wang and Fan, 2017)], so interest might lie on **parsimonious** representations of Σ [e.g. (Frühwirth-Schnatter, Hosszejni, Lopes, 2025)], but:
  - Selecting k might incur on biases, so would it be possible to maintain the same posterior rates?
  - Is there a way to specify a_i as a function of p and n to induce lower k with high probability *a priori*?

BOSTON UNIVERSITY

Thank you

for a great paper and talk!