High dimensional covariance matrix Estimation

Debdeep Pati (FSU), Anirban Bhattacharya (TAMU), Natesh Pillai (Harvard), David Dunson (Duke)
debdeep@stat.fsu.edu, anirban@stat.tamu.edu, pillai@fas.harvard.edu, dunson@duke.edu

Introduction

Covariance matrix estimation in $p \gg n$ setting (genomic applications), very high dimensional model space, an unstructured $p \times p$ covariance matrix has $O(p^2)$ free parameters. **Key: parsimonious modeling**. **Solution: Factor models**, explain dependence through shared dependence on fewer latent factors:

$$y_i = \mu + \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma), \quad i = 1, \ldots, n$$

$\mu \in \mathbb{R}^p$ a vector of means, assumed $\mu = 0$, $\eta_i \in \mathbb{R}^k$ latent factors, Λ a $p \times k$ matrix of factor loadings with $k \ll p$, ϵ_i has diagonal covariance $\Sigma = \sigma^2 I_p$. Hence $\mathrm{Var}(y_i) := \Omega = \Lambda\Lambda' + \Sigma$. Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Liu, 2011 ...), Minimax optimal rates established in Cai, Zhang and Zhou (2010), Bayesian counterpart lacks a theoretical framework in terms of posterior convergence rates

Objectives

A prior $\Pi(\Lambda \otimes \sigma^2)$ induces a prior distribution $\Pi(\Omega)$, How does the posterior behave assuming data sampled from fixed truth? Castillo and van der Vaart (2012) point mass mixtures, computationally inefficient due to search of huge model space - calls for conts. shrinkage priors.
$\mathcal{C}_n$: cone of covariance matrices of size $p \times p$ satisfying sparsity constraints; see (A0) - (A4). We observe $y_i \sim N_p(0, \Omega_{0n})$, $\mathbf{y}^{(n)} = (y_1, \ldots, y_n)$ For $\|\cdot\|_2$ denoting the operator norm, find minimum sequence $\epsilon_n \to 0$ such that

$$\lim_{n\to\infty} \mathrm{E}_{\Sigma_{0n}} \Pi_n \big[\|\Omega_n - \Omega_{0n}\|_2 > M\epsilon_n \mid \mathbf{y}^{(n)} \big] = 0$$

Can we achieve optimal rate of convergence ϵ_n even when $p = e^{n^\alpha}$, $0 < \alpha < 1$?

New priors

We propose independent priors for columns of Λ. Point mass mixture priors are widely used but they have computational issues.

- (Ppm) $\lambda_{jh} \sim (1 - \pi_h)\delta_0 + \pi_h g(\cdot)$, $\pi_h \sim \mathrm{Beta}(1, \lambda p + 1)$. $g(\cdot)$ has Laplace like or heavier tails, $\sigma^2 \sim \mathrm{Ga}(a, b), k \sim \mathrm{Poiss}(\lambda)$

Propose new prior (Pcs) that are computationally amenable, but statistically as efficient as the point mass priors. Idea is to introduce a local scale τ_h and a bunch of global scales $(\gamma_{1h}, \ldots, \gamma_{ph})$ for h th column

- (Pcs) $\lambda_{jh} \sim \mathrm{Laplace}(\tau_h \gamma_{jh}), \tau_h \sim \mathrm{Inverse\text{-}Ga}(a, b), (\gamma_{1h}, \ldots, \gamma_{ph}) \sim \mathrm{Dirichlet}(\alpha/p, \ldots, \alpha/p), \sigma^2 \sim \mathrm{Ga}(a, b), \mathrm{Poiss}(\lambda)$

Intuition behind Pcs

- Both have high concentration around sparse vectors, **high** prior probability of **large subsets being close to** 0.
- Pcs allows dependence among $(\gamma_1, \ldots, \gamma_p)$ forcing a large subset of the local scales γ_j to be close to zero and thus behaving similar to Ppm

Assumptions

(A0) $\Omega_{0n} \in \mathcal{C}_n$ are of the form

$$\Omega_{0n} = \Lambda_{0n}\Lambda_{0n}^{\mathrm{T}} + \Sigma_{0n}\,, \quad \Lambda_{0n} \in \Theta_{\Lambda}^{(p, k_{0n})}, \quad \Sigma_{0n} = \sigma_{0n}^2\, \mathrm{I}_p\,,$$

There exist sequences of positive real numbers c_n, s_n with $c_n \lesssim s_n$, such that,

(A1) $\lim_{n\to\infty} c_n k_{0n}^{3/2} \sqrt{\frac{s_n \log p_n}{n}} \sqrt{\log n} = 0$; $k_{0n}^{3/2} \sqrt{\frac{s_n \log p_n}{n}} (\log n)^{3/2} = O(1)$.

(A2) Each column of Λ_{0n} belongs to $l_0[s_n; p_n]$.

(A3) $\left\| \frac{1}{c_n} \Lambda_{0n}^{\mathrm{T}} \Lambda_{0n} - \mathrm{I}_{k_{0n}} \right\|_2 = o(k_{0n} \sqrt{\log k_{0n}/n})$.

(A4) There exists a constant $\sigma_0^{(1)}$ such that $\sigma_0^{(1)} \leq \sigma_{0n}^2 \leq c_n$.

Main results

- With $(A0) - (A4)$, $\|\cdot\|_2$ and $s_n k_{0n} \gtrsim \log p_n$, both Ppm and Pcs lead to a convergence rate $\epsilon_n = c_n k_{0n}^{3/2} \sqrt{\frac{s_n \log p_n}{n}} \sqrt{\log n}$
- We obtain near minimax rate as with $(A0) - (A4), \|\cdot\|_2$, with $k_{0n} = O(1)$, the minimax rate is $c_n \sqrt{\frac{s_n \log p_n}{n}}$

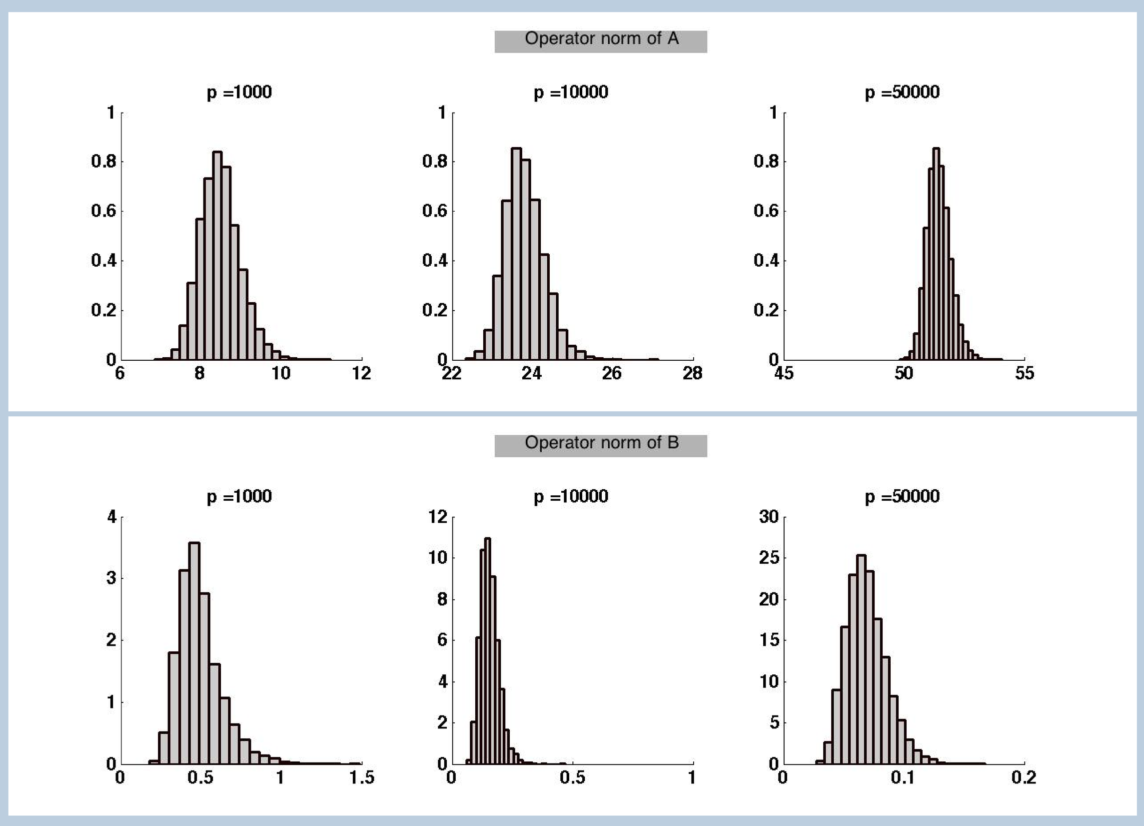
Insights into the assumptions

- random matrix theory - "tall" and "skinny" matrices properly normalized act as approximate isometry
- In light of (A3), a plausible mechanism generating the truth $\lambda_{0jh} \sim (1 - \pi)\delta_0 + \pi \mathrm{N}(0, 1)$ with $\pi = s/p$
- With prob. $1 - e^{-C'k}$ for constants $C', C > 0$

$$\left\| \frac{1}{p} \Lambda_0^{\mathrm{T}} \Lambda_0 - \pi \mathrm{I}_k \right\|_2 \leq C \frac{\sqrt{k}}{\sqrt{p}} \|\pi \mathrm{I}_k\|_2$$

- Hence, we expect with large probability,

$$\left\| \frac{1}{s} \Lambda_0^{\mathrm{T}} \Lambda_0 - \mathrm{I}_k \right\|_2 = o(k_{0n} \sqrt{\log k_{0n}/n}) \quad (A3)$$



$s = \pi \times p$. $A_{p\times k}$ i.i.d. $(1 - \pi)\delta_0 + \pi N(0, 1)$. $B = (1/s)A^{\mathrm{T}}A - \mathrm{I}_k$

Conclusion

- Consistent estimation even if $p = O(e^{n^\alpha})$ for $\alpha \in (0, 1)$
- Prior concentration, prior probability of subset size very important
- Developed new shrinkage priors Pcs which achieve this
- Computation is very fast using Pcs, hence potentially useful