Bayesian structure learning in graphical models

Seung Bong Jung

Seoul National University

February 17, 2021

Topic

- Banerjee and Ghoshal, "Bayesian structure learning in graphical models", Journal of Multivariate Analysis, 2015, 136, 147-162

Outline

Outline

Notations

Notation 1.1

Let V *be the set of indices of* p *vertices of* p*-dimensional random vector* $\mathsf{X} \sim \mathsf{N}_p(0, \Omega^{-1})$. *If induced undirect graph* G *has edges* $\mathsf{E} \subset \{(i,j) \in \mathsf{V} \times \mathsf{V} : i < j\}$, *denote* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$.

Notation 1.2

- $\mathcal{M}$: *the linear space of symmetric matrices of order* p.
- $\mathcal{M}^+$: *the cone of positive definite matrices of order* p.
- $\mathcal{P}_G$: *the cone of positive definite matrices of* p *with zero entries corresponding to each missing edge in* E.

Notations

Notation 1.3

Suppose $\mathrm{A} = ((a_{ij}))$ *is a* $p \times p$ *matrix and* $x \in \mathbb{R}^p$.

- *If* $1 \leq r < \infty$, $||\mathrm{A}||_r = (\sum_{i,j=1}^{p} |a_{ij}|^r)^{\frac{1}{r}}$, $||x||_r = (\sum_{i=1}^{p} |x_i|^r)^{\frac{1}{r}}$
- $||\mathrm{A}||_\infty = \max_{i,j} |a_{ij}|$, $||x||_\infty = \max_i |x_i|$.
- *If* $\mathrm{A} \in \mathcal{M}$, $\mathrm{eig}_i(\mathrm{A})$ *denotes* i*th biggest eigenvalue of* A.

Notation 1.4

Suppose $r \times s$ *matrix* $\mathrm{A} = ((a_{ij}))$ *is given.*

- $||\mathrm{A}||_{(r,s)} = \sup(||\mathrm{Ax}||_s : ||\mathrm{x}||_r = 1, \mathrm{x} \in \mathbb{R}^p)$

Remark 1.1

Assume $\mathrm{A} \in \mathcal{M}$. *Clearly,* $||\mathrm{A}||_2 = \sqrt{\mathrm{tr}(\mathrm{A}^t\mathrm{A})}$, $||\mathrm{A}||_{(2,2)} = \max_i |\mathrm{eig}_i(\mathrm{A})|$

Notations

Notation 1.5

Suppose r_n, s_n are two numerical sequences.

- *If $\frac{r_n}{s_n}$ is bounded as $n \to \infty$, we say $r_n = O(s_n)$ or equivalently $r_n \lesssim s_n$. In case $r_n \lesssim s_n$ and $s_n \lesssim r_n$ both hold, denote this by $r_n \asymp s_n$*
- *If $\frac{r_n}{s_n} \to 0$ as $n \to \infty$. Then we denote this by $r_n = o(s_n)$ or equivalently $r_n \ll s_n$.*
- *If $\frac{r_n}{s_n} \to 1$ as $n \to \infty$, we say $r_n \sim s_n$.*

Notation 1.6

Suppose random sequence X_n is given. We say $X_n = O_p(\delta_n)$, if $^{\exists}M > 0$ such that $P(|X_n| > M\delta_n) \to 0$ as $n \to \infty$. Similarly, $X_n = o_p(\delta_n)$, if $^{\forall}\epsilon > 0, P(|X_n| > \epsilon\delta_n) \to 0$ as $n \to \infty$.

Notations

Notation 1.7

- $\#\mathsf{T}$ *denotes the cardinality of a set* T. *Define the symmetric matrix* $\mathsf{E}_{(i,j)} = ((\mathbb{1}_{\{(i,j),(j,i)\}}(l,m)))$, *where* $\mathbb{1}$ *is an indicator function.*
- *For a subset* A *of a metric space* (S,d), *the minimum number of* d*-balls of size* ϵ *in* S *needed to cover* A *is denoted by* $\mathsf{N}(\epsilon,\mathsf{A},\mathsf{d})$

Notation 1.8

- *The Hellinger distance between two probability densities* q_1 *and* q_2 *is defined as* $\mathsf{h}(\mathsf{q}_1,\mathsf{q}_2) = ||\sqrt{\mathsf{q}_1} - \sqrt{\mathsf{q}_2}||_2 = (\int(\sqrt{\mathsf{q}_1(\mathsf{x})} - \sqrt{\mathsf{q}_2(\mathsf{x})})^2 \mathsf{dx})^{\frac{1}{2}}$

Preliminaries

Theorem 1.1

Let $A, B \in \mathcal{M}$. *Then followings hold :*

$$||A||_\infty \leq ||A||_{(2,2)} \leq ||A||_2 \leq p||A||_\infty$$
$$||AB||_2 \leq \min\{||A||_{(2,2)}||B||_2, ||B||_{(2,2)}||A||_2\}$$

(Proof) Use spectral decomposition to A, B.

Theorem 1.2

Suppose $X \sim N_p(0, \Omega^{-1})$, *where* $\Omega = ((\omega_{ij}))$. *Then* $\omega_{ij} = \omega_{ji} = 0$ *implies conditional independence of* X_i *and* X_j *given* $(X_r : r \neq i, j)$ *with* $i \neq j$.

(Proof) w.l.o.g. assume $i = 1, j = 2$. Let $\Sigma = \Omega^{-1}$. Consider Σ as following :

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Preliminaries

where Σ_{11} is 2×2 matrix, Σ_{22} is $(p-2) \times (p-2)$ matrix, and Σ_{12} is $2 \times (p-2)$ matrix, $\Sigma_{21} = \Sigma_{12}^{t}$.

Let $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Omega_{11}^{-1}$. Then, we have $X_1, X_2 | X_{r:r \neq 1,2} \sim N_2(\mathbf{0}, \Sigma_{11.2}^{-1})$. Hence, $Cov(X_1, X_2 | X_{r:r \neq 1,2}) = \omega_{12} = \omega_{21}$.

Under normality, X_1, X_2 given $(X_r : r \neq 1, 2)$ are independent if and only if their conditional covariance given other covariates is 0. Thus, we get the desired result.

Outline

Prior settings

Graphical lasso

Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N_p(0, \Sigma)$, where $\Omega = \Sigma^{-1}$. Then we clearly see that $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^t$ is MLE of Σ. Then the graphical lasso solves the following optimization problem with penalty parameter $\lambda \geq 0$ and constraint $\Omega \in \mathcal{M}^+$:

$$\text{minimize} \quad \underbrace{-\log\det(\Omega) + \text{tr}(\hat{\Sigma}\Omega)}_{\text{negative data log-likelihood}} \quad \underbrace{+\frac{\lambda}{n}||\Omega||_1}_{\text{Lasso type penalty}} \tag{1}$$

Prior settings

Remark 2.1

Let $\mathcal{U}(\epsilon_0, s) = \{\Omega \in \mathcal{M}^+ : \#\{(i,j) : 1 \le i < j \le p, \omega_{ij} \neq 0\} \le s, 0 < \epsilon_0 \le \text{eig}_1(\Omega) \le \text{eig}_p(\Omega) \le \epsilon_0^{-1} < \infty\}$ *Denote the minimizer of (1) by* Ω^* *with true precision matrix* Ω_0. *Assuming* $\Omega_0 \in \mathcal{U}(\epsilon_0, s)$, *Rothman et al.(2008) has shown* $||\Omega^* - \Omega_0||_2 = O_p(\sqrt{\frac{(p+s)\log p}{n}})$.

Hence $||\Omega^* - \Omega_0||_2 \xrightarrow{P} 0$ *as* $n^{-1}(p+s)\log p \to 0$.

Theorem 2.1

Suppose Ω^* *is the solution of (1) with true precision matrix* $\Omega_0 \in \mathcal{U}(\epsilon_0, s)$. *Then* $||\Omega^*||_{(2,2)} = O_p(1)$ *and* $||\Omega^{*-1}||_{(2,2)} = O_p(1)$

Prior settings

(proof)

- For the first equation, by triangle inequality and result of theorem 1.1, $||\Omega^*||_{(2,2)} \leq ||\Omega^* - \Omega_0||_{(2,2)} + ||\Omega_0||_2$.
- By remark 2.1 and result of theorem 1.1, we see that $||\Omega^* - \Omega_0||_{(2,2)} = o_p(1)$. Since $||\Omega_0||_{(2,2)}$ is constant, we have $||\Omega^*||_{(2,2)} = O_p(1)$.
- Similarly, by triangle inequality and Cauchy-Shwartz inequality, one can see

$$\begin{aligned} ||\Omega^{*-1}||_{(2,2)} &\leq ||\Omega_0^{-1}||_{(2,2)} + ||\Omega^{*-1} - \Omega_0^{-1}||_{(2,2)} \\ &\leq ||\Omega_0^{-1}||_{(2,2} + ||\Omega_0^{-1}||_{(2,2)}||\Omega^{*-1} - \Omega_0^{-1}||_{(2,2)}||\Omega^{*-1}||_{(2,2)} \end{aligned}$$

,which leads to $||\Omega^{*-1}||_{(2,2)} \leq \frac{||\Omega_0^{-1}||_{(2,2)}}{1-||\Omega_0^{-1}||_{(2,2)}||\Omega^{*-1}-\Omega_0^{-1}||_{(2,2)}}$.

- Again, for $||\Omega^* - \Omega_0||_{(2,2)} = o_p(1)$ and $||{\Omega_0}^{-1}||_{(2,2)}$ is constant, we have the second equation.

Prior settings

Wang's method(2012)

- Put Exponential prior $\lambda \exp(-\lambda x)$ on diagonal elements and Laplace prior $\frac{\lambda}{2} \exp(-\lambda|x|)$ on off-diagonal elements.
- But by absolute continuity of priors, the probability of event $\{\omega_{ij} = 0\}$ is 0 and so is the posterior probability of the event.
- Hard to introduce sparsity to the model.

Banerjee & Ghosal's method(2015)

- Put point-mass prior on the events $\{\omega_{ij} = 0\}$.
- Able to make posterior inference about the sparse structure of the model.

Prior settings

- Let $\Gamma = (\gamma_{ij} : 1 \leq i < j \leq p)$ be a $\binom{p}{2}$-dimensional vector of edge-inclusion vector, where $\gamma_{ij} = \mathbb{1}\{(i,j) \in \mathrm{E}\}$.
- $\#\Gamma$: the number of non-zero elements in Γ.
- Put the same priors on precision matrix suggested by Wang.
- Define $\mathcal{M}_0^+$ to be subset of $\mathcal{M}^+$, consisting of positive definite matrices of order p, whose eigenvalues are bounded by two fixed positive numbers.
- Now assume true precision matrix belongs to $\mathcal{M}_0^+$, unless otherwise stated in this section.
- Suppose $\gamma_{ij} \overset{i.i.d}{\sim} \mathrm{Ber}(q)$, where q is fixed. To introduce sparsity to the precision matrix, put restriction on $\#\Gamma$, namely the size of model.

Prior settings

Prior 1

- For fixed positive number $\bar{\gamma}$, we have

$$p(\Gamma) \propto q^{\#\Gamma}(1-q)^{\binom{p}{2}-\#\Gamma}\mathbb{1}(\#\Gamma \leq \bar{\gamma})$$

Prior 2

- Choose the prior distribution on $\bar{R}$ such that $p(\bar{R} > a_1 m) \leq e^{-a_2 m \log m}$ for some $a_1, a_2 > 0$ and $^{\forall} m \in \mathbb{N}$.
- Then, we have another prior $p(\Gamma|\bar{R}) \propto q^{\#\Gamma}(1-q)^{\binom{p}{2}-\#\Gamma}\mathbb{1}(\#\Gamma \leq \bar{R})$ leading to $p(\Gamma) \propto q^{\#\Gamma}(1-q)^{\binom{p}{2}-\#\Gamma}P(\bar{R} \geq \#\Gamma)$.

Prior settings

Posterior

- From Prior 1,2, we have following posterior :

$$
\begin{aligned}
p(\Omega, \Gamma | X^{(n)}) &\propto p(X^{(n)} | \Omega, \Gamma) p(\Omega | \Gamma) p(\Gamma) \\
&= (2\pi)^{-np/2} [\det(\Omega)]^{n/2} \exp(-\frac{n}{2} \mathrm{tr}(\hat{\Sigma}\Omega)) \\
&\times \prod_{\gamma_{ij}=1} \{\frac{\lambda}{2} \exp(-\lambda|\omega_{ij}|)\} \prod_{i=1}^{p} \{\frac{\lambda}{2} \exp(-\frac{\lambda}{2}\omega_{ii})\} \\
&\times p(\Gamma) \mathbb{1}_{\mathcal{M}_0^+}(\Omega)
\end{aligned}
$$

Prior settings

Posterior

- Consequently, $p(\Omega, \Gamma|X^{(n)}) \propto C_\Gamma Q(\Omega, \Gamma|X^{(n)})$, where

$$C_\Gamma = (2\pi)^{-np/2} q^{\#\Gamma} (1-q)^{\binom{p}{2}-\#\Gamma} (\lambda/2)^{p+\#\Gamma} \beta(\Gamma)$$

$$\beta(\Gamma) = \begin{cases} p(\bar{R} \geq \#\Gamma), & \text{for Prior 2} \\ \mathbb{1}(\#\Gamma \leq \bar{r}), & \text{for Prior 1} \end{cases}$$

$$Q(\Omega, \Gamma|X^{(n)}) = [\det(\Omega)]^{n/2} \exp(-\frac{n}{2} tr(\hat{\Sigma}\Omega)) \prod_{\gamma_{ij}=1} \exp(-\lambda|\omega_{ij}|)$$

$$\times \prod_{i=1}^{p} \exp(-\frac{\lambda}{2}\omega_{ii}) \mathbb{1}_{\mathcal{M}_0^+}(\Omega)$$

- The posterior is very intractable due to positive constraint on Ω.One possible method is RJMCMC, but the result may be unreliable due to too large number of models to search over. $(2^{\binom{p}{2}})$.

Posterior concentration

Claim 2.1

Suppose $A, B \in \mathcal{M}$. *Then,* $tr([AB]^2) \leq tr(A^2B^2)$

(proof) Use the equation $2(AB)^2 - 2A^2B^2 = (AB - BA)^2$ and the fact that $AB - BA$ is skew-hermitian if $A, B \in \mathcal{M}$.

Lemma 2.1

If p_{Ω_k} *is the density of* $N_p(0, \Omega_k^{-1})$, $k = 1, 2$, *then for all* $\Omega_k \in \mathcal{M}_0^+$, $k = 1, 2$, *and* $d_i, i = 1, \ldots, p$, *eigenvalues of* $A = \Omega_1^{-1/2}\Omega_2\Omega_1^{-1/2}$, *we have followings for sufficiently small* $\delta > 0$ *and some constant* $c_0 > 0$.

(i) $c_0^{-1}||\Omega_1 - \Omega_2||_2^2 \leq \sum_{i=1}^{p} |d_i - 1|^2 \leq c_0||\Omega_1 - \Omega_2||_2^2$

(ii) $h(p_{\Omega_1}, p_{\Omega_2}) < \delta$ *implies* $\max_i |d_i - 1| < 1$

and $||\Omega_1 - \Omega_2||_2^2 \leq c_0 h^2(p_{\Omega_1}, p_{\Omega_2})$

(iii) $h^2(p_{\Omega_1}, p_{\Omega_2}) \leq c_0||\Omega_1 - \Omega||_2^2$.

Posterior concentration

(Proof of (i))

- Since $\Omega_1 \in \mathcal{M}_{\prime}^{+}$, both $||\Omega_1||_2$ and $||\Omega_1^{-1}||_2$ are bounded by some constant.
- $$||\Omega_1 - \Omega||_2^2 = ||\Omega_1^{\frac{1}{2}}(\mathsf{I_p} - \mathsf{A})\Omega_1^{\frac{1}{2}}||_2^2 \leq ||\Omega_1||_{(2,2)}^2||(\mathsf{I_p} - \mathsf{A}||_2^2$$
$$= ||\Omega_1||_{(2,2)}^2 \mathsf{tr}(\mathsf{I_p} - \mathsf{A})^2 = ||\Omega_1||_{(2,2)}^2 \sum_{i=1}^{p}(d_i - 1)^2$$
- $$\sum_{\mathsf{i}=1}^{\mathsf{p}}(\mathsf{d_i} - 1)^2 = ||\mathsf{I_p} - \mathsf{A}||_2^2 = ||\Omega_1^{-\frac{1}{2}}(\Omega_1 - \Omega_2)\Omega_1^{-\frac{1}{2}}||_2^2$$
$$\leq ||\Omega_1^{-1}||_{(2,2)}^2||\Omega_1 - \Omega_2||_2^2$$

Posterior concentration

(Proof of (ii) & (iii))

- By direct calculations,
$$\frac{1}{2}h^2(p_{\Omega_1}, p_{\Omega_2}) = 1 - \{\prod_{i=1}^{p} \frac{1}{2}(d_i^{\frac{1}{2}} + d_i^{-\frac{1}{2}})\}^{-\frac{1}{2}} \cdots (*)$$
- Assuming $h(p_{\Omega_1}, p_{\Omega_2}) < \delta$ and rearranging terms, we have
$$\prod_{i=1}^{p} \frac{1}{2}(d_i^{\frac{1}{2}} + d_i^{-\frac{1}{2}}) \leq (1 - \frac{\delta^2}{2})^{-2} = 1 + \eta$$
- From GM-AM inequality, $\max_i \frac{1}{2}(d_i^{\frac{1}{2}} + d_i^{-\frac{1}{2}}) \leq 1 + \eta$.
- With sufficiently small $\delta > 0$, η can be made sufficiently small, so that solving the inequality $\frac{1}{2}(d_i^{\frac{1}{2}} + d_i^{-\frac{1}{2}})$, we have $|d_i - 1| < 1, ,^{\forall} i$.
- Since $|d_i - 1| < 1$, by a Taylor's expansion, for some constant $c_1, c_2 > 0$,

Posterior concentration

$$1 + c_1 \sum_{i=1}^{p} (d_i - 1)^2 \leq 2^{-p} \prod_{i=1}^{p} (d_i^{\frac{1}{2}} + d_i^{-\frac{1}{2}}) \approx (1 - h^2)^{-2}$$

$$\leq 1 + c_2 \sum_{i=1}^{p} (d_i - 1)^2$$

- Since h is made sufficiently small, using Taylor's expansion, $(1 - h^2)^{-2} \sim 1 + h^2$. By the lower estimate in the above inequality, $c_1 \sum_{i=1}^{p} (d_i - 1)^2 \leq h^2$. With the result of (i), this establishes (ii).
- To prove (iii), since Hellinger distance is bounded, we may assume $||\Omega_1 - \Omega_2||_2$ is sufficiently small. Again by (i), $\max_i |d_i - 1| < 1$ and so with the upper estimate in the above inequality and the result of (i), we obtain (iii).

Posterior concentration

Claim 2.2

Suppose p_0, p_1 *are two probability densities. Then* $K(p_0, p_1) \geq h^2(p_0, p_1)$

(note) $K(p_0, p_1) = \int p_0 \log(\frac{p_0}{p_1})$, which is KL divergence.
(proof) If $x < 1$, we have $-\log(1-x) \geq x$. So,

$$K(p_0, p_1) = -2 \int p_0 \log(\sqrt{\frac{p_1}{p_0}}) \geq 2 \int p_0 (1 - \sqrt{\frac{p_1}{p_0}}) = 2 - 2 \int \sqrt{p_0 p_1}$$
$$= h^2(p_0, p_1)$$

Posterior concentration

Claim 2.3

Let p_{Ω_i} *be density of* $N_p(0, \Omega_i^{-1})$, $i = 0, 1$ *and* $d_i, i = 1, 2, \ldots, p$ *be eigenvalues of* $\Omega_0^{-\frac{1}{2}} \Omega_1 \Omega_0^{-\frac{1}{2}}$. *Then followings hold:*

$$K(p_{\Omega_0}, p_{\Omega_1}) = -\frac{1}{2} \sum_{i=1}^{p} (\log d_i - 1 + d_i)$$

$$V(p_{\Omega_0}, p_{\Omega_1}) = \frac{1}{4} (\sum_{i=1}^{p} (\log d_i - 1 + d_i))^2 + \frac{1}{2} \sum_{i=1}^{p} (1 - d_i)^2$$

where $V(p_0, p_1) = \int p_0 \log^2(\frac{p_0}{p_1})$

(proof) Use the fact $E(Z^t A Z) = tr(A\Sigma), Var(Z^t A Z) = 2tr(A\Sigma A\Sigma)$, where $Z \sim N_p(0, \Sigma)$ and $A \in \mathcal{M}$. The result is direct from some computations.

Posterior concentration

Claim 2.4

Let $p \in \mathbb{N}$ and choose $\bar{r} \in \mathbb{N}$ such that $\bar{r} < \binom{p}{2}/2$. If $j \leq \bar{r}$,

$$\binom{\binom{p}{2}}{j} \leq \binom{p + \binom{p}{2}}{\bar{r}}$$

Theorem 2.2

Let $X^{(n)} = (X_1, \ldots, X_n) \sim N_p(0, \Omega_0^{-1})$, where $\Omega_0 \in \mathcal{U}(\epsilon_0, s)$ for some $0 < \epsilon_0 < \infty$ and $0 \leq s \leq p(p-1)/2$. Also assume that priors are either Prior1 or Prior2 with $q < 1/2$ and the range of eigenvalues of matrices in $\mathcal{M}_0^+$ is sufficiently broad to contain $[\epsilon_0, \epsilon_0^{-1}]$. Then the posterior distribution of Ω satisfies

$$E_0[P\{||\Omega - \Omega_0||_2 > M\epsilon_n | X^{(n)}\}] \to 0,$$

for $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$ and a sufficiently large $M > 0$.

Posterior concentration

(Proof)

- Let $B(p_{\Omega_0}, \epsilon_n) = \{p_\Omega : K(p_{\Omega_0}, p_\Omega) \leq \epsilon_n^2, V(p_{\Omega_0}, p_\Omega) \leq \epsilon_n^2\}$. Denote the prior of Ω by Π.
- Also let $\mathcal{P}_n$ be the set of all densities p_Ω such that the graph induced by Ω has maximum number of edges $\bar{r} < \binom{p}{2}/2$ and each entry of Ω is at most L in absoulte value, where $\bar{r}$ and L depend only on n and to be determined.
- The proof is by verifying each condition in the remark 2.2, which is the result of theorem 2.1. in "Convergence Rates of Posterior Distributions"

Posterior concentration

Remark 2.2

Suppose that for a sequence ϵ_n *with* $\epsilon_n \to 0$ *and* $n\epsilon_n^2 \to \infty$, *a constant* $C > 0$ *and sets* $\mathcal{P}_n \subset \mathcal{P}$, *where* $\mathcal{P}$ *is a space of densities, we have*

- *(Condition 1)* $\log N(\epsilon_n, \mathcal{P}_n, d) \leq n\epsilon_n^2$
- *(Condition 2)* $\Pi(\mathcal{P}_n^c) \leq \exp(-n\epsilon_n^2(C+4))$
- *(Condition 3)* $\Pi(B(p_{\Omega_0}, \epsilon_n)) \geq \exp(-n\epsilon_n^2 C)$

Then for sufficiently large M, $E_0[\Pi(d(p, p_0) \geq M\epsilon_n)|X^{(n)}] \to 0$.

- We shall verify (Condition1) first. We set metric d as Frobenius distance. Then $N(\epsilon_n, \mathcal{P}_n, d) = \sum_{j=1}^{\bar{r}} (\frac{L}{\epsilon_n})^j \binom{\binom{p}{2}}{j}$. Taking sufficiently large n and fixed constants $b_1, b_2 > 0$,

Posterior concentration

$$\log\{\sum_{i=1}^{\bar{r}}(\frac{\mathrm{L}}{\epsilon_n})^j\binom{\binom{p}{2}}{j}\} \leq \log\{\bar{r}(\frac{\mathrm{L}}{\epsilon_n})^{\bar{r}}\binom{p+\binom{p}{2}}{\bar{r}}\} \tag{2}$$
$$\lesssim \log\bar{r} + \bar{r}\log\mathrm{L} + \bar{r}\log\epsilon_n^{-1} + \bar{r}\log p$$

- Choose $\bar{\mathrm{r}} \sim \frac{\mathrm{b}_1\mathrm{n}\epsilon_\mathrm{n}^2}{\log\mathrm{n}}$ and $\mathrm{L} \sim \frac{\mathrm{b}_2\mathrm{n}\epsilon_\mathrm{n}^2}{\log\mathrm{n}}$. Then we have

$$\log\bar{r} + \bar{r}\log\mathrm{L} + \bar{r}\log\epsilon_n^{-1} + \bar{r}\log p \approx n\epsilon_n^2$$

- Next, we verify (Condition2).

Posterior concentration

- Suppose p_Ω does not belong to $\mathcal{P}_n$. By definition of $\mathcal{P}_n$, there exists an entry of Ω that exceeds L in magnitude or the number of edges in the graph induced by Ω exceeds $\bar{r}$. The probability of latter event is $P(\bar{R} > \bar{r}) \leq \exp(-a_2' b_1 n\epsilon_n^2)$. Note b_1 is chosen to be sufficiently large. To bound the probability of the formal event, if we naively say that prior on all entries of Ω is given by Exponential prior, the probability of each entry exceeds L is $\int_{\mathrm{L}}^{\infty} \lambda \exp(-\lambda\omega_{ij})\mathrm{d}\omega_{ij} = \exp(-\lambda \mathrm{L})$. For there are total $p + \binom{p}{2}$ distinct entries in Ω, assuming independence on entries for convenience, we have the probability of former event $(p + \binom{p}{2}) \exp(-\lambda \mathrm{L}) \lesssim \binom{p}{2} \exp(-\mathrm{L})$. Hence to sum up, there exists a constant $\mathrm{C} > 0$ such that

$$\Pi(\mathcal{P}_n^c) \leq \exp(-n\epsilon_n^2(\mathrm{C}+4))$$

Posterior concentration

- Assume $p_\Omega \in \mathrm{B}(p_{\Omega_0}, \epsilon_n)$. By claim 2.2., $\mathrm{K}(p_{\Omega_0}, p_\Omega) \geq \mathrm{h}^2(p_{\Omega_0}, p_\Omega)$.
- By Taylor's expansion to $\mathrm{K}(\mathrm{p}_{\Omega_0}, \mathrm{p}_\Omega)$, provided that it is valid, and using the result of theorem 1.1 and lemma 2.1,

$$\mathrm{K}(p_{\Omega_0}, p_\Omega) \sim \frac{1}{4}\sum_{i=1}^{p}(1-d_i)^2 \leq c^{-2}\mathrm{p}^2||\Omega-\Omega_0||_\infty^2$$

- Thus, if $c^{-2}\mathrm{p}^2||\Omega-\Omega_0||_\infty^2 \leq \epsilon_n^2$, then the above approximation is valid and so $\mathcal{P} = \{p_\Omega : ||\Omega-\Omega_0||_\infty \leq c\epsilon_n/p\} \subset \mathrm{B}(p_{\Omega_0}, \epsilon_n)$.
- Hence it suffices to get a lower estimate of the prior probability of $\mathcal{P}$.
- Because of constraint $\mathcal{M}_0^+$, the components of Ω are not independently distributed. However, a small neighborhood of $\Omega_0 \in \mathcal{U}(\epsilon_0, s)$ lies within $\mathcal{M}_0^+$, which was the assumption. Hence, the constraint can only increase prior concentration of $\mathrm{B}(p_{\Omega_0}, \epsilon_n)$, as the prior probability of $\mathcal{P}$ tends to increase. So in order to obtain lower bound for the prior probability of $\mathcal{P}$, we may assume independence on components of Ω.

Posterior concentration

- So we have the estimate

$$\Pi(||\Omega_0 - \Omega||_\infty \leq c\epsilon_n/p) \gtrsim (\frac{c\epsilon_n}{p})^{p+s} \tag{3}$$

- Because $\log((\frac{c\epsilon_n}{p})^{p+s}) \approx n\epsilon_n^2$, we have (Condition 3).

Posterior concentration

Remark 2.3

Similar to frequentist graphical lasso, we have $||\Omega - \Omega^*||_2 = O(\epsilon_n)$ *with posterior probability tending to one in probability. This gives,*

$$\frac{\int_{||\Omega-\Omega^*||_2 \leq \epsilon_n} f(\Omega) \prod_{(i,j)\in \bar{E}} d\omega_{ij}}{\int_{\Omega \in \mathcal{M}_0^+} f(\Omega) \prod_{(i,j)\in \bar{E}} d\omega_{ij}} \to 1$$

, for a bounded and positive measurable function $f(\cdot)$ *of* Ω

Outline

Approximation of Posterior

- Let $\Omega_\Gamma = ((\omega_{\Gamma,ij}))$ be the precision matrix in the model Γ
- Reformulate $Q(\Omega, \Gamma | X^{(n)})$ as

$$Q(\Omega, \Gamma | X^{(n)}) = \exp(-\frac{n}{2} h_\Gamma(\Omega_\Gamma) \mathbb{1}_{\mathcal{M}_0^+}(\Omega_\Gamma)$$

where

$$h_\Gamma(\Omega_\Gamma) = -\log\det(\Omega_\Gamma) + \mathrm{tr}(\hat{\Sigma}\Omega_\Gamma) + \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} |\omega_{\Gamma,ij}| + \frac{\lambda}{n} \sum_{i=1}^{p} \omega_{\Gamma,ii}.$$

- Let $\bar{E}_\Gamma = \{(i,j) \in V \times V : i = j, \text{ or } \gamma_{ij} = 1, i \neq j\}$
- By integrating $p(\Gamma, \Omega | X^{(n)})$ with respect to $\omega_{\Gamma,ij}$,

$$p(\Gamma | X^{(n)}) \propto C_\Gamma \int_{\Omega_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{2} h_\Gamma(\Omega_\Gamma)) \prod_{(i,j) \in \bar{E}_\Gamma} d\omega_{\Gamma,ij}$$

Approximation of Posterior

- Suppose $h_\Gamma(\Omega_\Gamma)$ is uniquely minimized at $\Omega_\Gamma{}^*$, which is graphical lasso. Define $\Delta_\Gamma = \Omega_\Gamma - \Omega_\Gamma^* = ((u_{\Gamma,ij}))$
- $p(\Gamma|X^{(n)}) \propto C_\Gamma \int_{\Omega_\Gamma^*+\Delta_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{2} h_\Gamma(\Delta_\Gamma + \Omega_\Gamma^*)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij}$
- $g_\Gamma(\Delta_\Gamma) = -\log\det(\Delta_\Gamma + \Omega_\Gamma^*) + \mathrm{tr}(\hat{\Sigma}\Delta_\Gamma) + \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} (|u_{\Gamma,ij} + \omega^*_{\Gamma,ij}| - |\omega^*_{\Gamma,ij}|) + \frac{\lambda}{n} \sum_{i=1}^{p} u_{\Gamma,ii}$
- $h_\Gamma(\Delta_\Gamma + \Omega_\Gamma{}^*) = g_\Gamma(\Delta_\Gamma) + h_\Gamma(\Omega_\Gamma^*) + \log\det(\Omega_\Gamma{}^*)$
-

$$p(\Gamma|X^{(n)}) \propto C_\Gamma \exp(-\frac{n}{2} h_\Gamma(\Omega_\Gamma{}^*))[\det(\Omega_\Gamma{}^*)]^{-\frac{n}{2}} \times \int_{\Omega_\Gamma^*+\Delta_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{2} g_\Gamma(\Delta_\Gamma)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij}$$

Approximation of Posterior

- Recall that Laplace approximation for twice differentiable function h on $\mathbb{R}^d$ with unique minimizer $\hat{\mathrm{x}}$ and $M > 0$,

$$\int e^{-Mh(\mathrm{x})}\mathrm{dx} \approx e^{-Mh(\hat{\mathrm{x}})}(2\pi)^{\frac{d}{2}} M^{-\frac{d}{2}} [\det(D^2 h(\mathrm{x}))|_{\mathrm{x}=\hat{\mathrm{x}}}]^{-\frac{1}{2}} \tag{4}$$

- Define the matrix $\mathrm{H_B} = ((h_{\mathrm{B}}\{(i,j),(l,m)\}))$, where $h_{\mathrm{B}}\{(i,j),(l,m)\} = \mathrm{tr}(\mathrm{B}^{-1}\mathrm{E}_{(i,j)}\mathrm{B}^{-1}\mathrm{E}_{(l,m)})$.
- The Hessian matrix of g_Γ is the $\#\bar{\mathrm{E}}_\Gamma \times \#\bar{\mathrm{E}}_\Gamma$ matrix $\mathrm{H}_{\Delta_\Gamma + \Omega^*_\Gamma}$, whose $\{(i,j),(l,m)\}$th entry for $(\mathrm{i},\mathrm{j}),(\mathrm{l},\mathrm{m}) \in \bar{\mathrm{E}}_\Gamma$ is given by

$$\frac{\partial^2 \mathrm{g}_\Gamma(\Delta_\Gamma)}{\partial u_{\Gamma,ij}\partial u_{\Gamma,lm}} = \mathrm{tr}((\Delta_\Gamma + \Omega^*_\Gamma)^{-1}\mathrm{E}_{(i,j)}(\Delta_\Gamma + \Omega^*_\Gamma)^{-1}\mathrm{E}_{(l,m)})$$

- Since Ω^*_Γ is the unique minimizer of h_Γ, g_Γ is minimized only at 0.

Approximation of Posterior

- Assuming that $\omega^*_{\Gamma,ij}$ is not 0, by (2),

$$p^*(\Gamma|X^{(n)}) \propto C_\Gamma \exp(-\frac{n}{2} h_\Gamma(\Omega^*_\Gamma))(\frac{4\pi}{n})^{\frac{\#\bar{E}}{2}} [\det(H_{\Omega^*_\Gamma})]^{-\frac{1}{2}} \tag{5}$$

- If $\omega^*_{\Gamma,ij}$ is 0 for at least one (i,j), the approximation is not valid because g_Γ is not differentiable. We may call the model making Laplace approximation as "regular model".
- By showing that we can ignore non-regular model, we can consider only regular models, which leads to significantly reduced number of models to search over.

Approximation of Posterior

- In fact, the validity of approximation also depends on $[\det(H_{\Omega_\Gamma^*})]^{-\frac{1}{2}}$. But this term is always bounded away from 0, regardless of regularity, since the minimum eigenvalue of the Hessian $H^*_{\Omega_\Gamma}$ is bounded away from 0.
- The idea to find posterior mode of $p^*(\Gamma|X^{(n)})$ is bad. The posterior heavily decays, even if we consider only regular model. Also, there are too many models to search over.

Outline

Ignorability of non-regular models

- For convenience, suppose that the vector Γ has t 1s on its first components and the rest of them are 0. The model Γ is non-regular when the corresponding graphical lasso to the model Γ has at least one 0 entry on the edge which is presented in the model Γ. So among those t 1s, say the last r of them have corresponding graphical solution equal to 0. In this context, we obtain submodel Γ' of Γ with first $(t - r)$ 1s and rest 0s

Ignorability of non-regular models

Lemma 3.1

For the corresponding regular submodel Γ' of Γ, the graphical lasso solutions corresponding to models Γ and Γ' are identical.

(proof)

- Our convex optimization problem is to minimize

$$f(\Omega_\Gamma) = -\log\det(\Omega_\Gamma) + \mathrm{tr}(\hat{\Sigma}\Omega_\Gamma) + \frac{2\lambda}{n}\sum_{\gamma_{ij}=1}|\omega_{\Gamma,ij}| + \frac{\lambda}{n}\sum_{i=1}^{p}\omega_{\Gamma,ii}$$

- Define a function $a : \mathcal{M} \to \mathcal{M}$ by $a(X) = 2X - \mathrm{diag}(X)$. Then, it is clear that a is injection and linear map. Hence $a(X) = 0$ only when $X = 0$.

Ignorability of non-regular models

- By KKT condition,

$$\frac{\partial f(\Omega_\Gamma)}{\partial \Omega_\Gamma} = -a(\Omega_\Gamma^{-1} - \hat{\Sigma} - \frac{\lambda}{n} G) = 0 \Leftrightarrow \Omega_\Gamma^{-1} - \hat{\Sigma} - \frac{\lambda}{n} G = 0$$

  where $G = ((g_{ij}))$ is a matrix with if $i = j$, $g_{ij} = 1$ and otherwise $g_{ij} = \frac{\omega_{\Gamma,ij}}{|\omega_{\Gamma,ij}|}$ for $\omega_{\Gamma,ij} \neq 0$ and $|g_{ij}| \leq 1$ for $\omega_{\Gamma,ij} = 0$.

- Since Ω_Γ^* is graphical lasso, by (30), $\Omega_\Gamma^{*-1} - \hat{\Sigma} - \frac{\lambda}{n} G = 0$. Let (i, j) be a pair such that $\gamma_{ij} = 1$ but $\omega_{\Gamma,ij}^* = 0$. By the definition of regular submodel Γ', $\Omega_{\Gamma'}^*$ also satisfies $\Omega_\Gamma^{-1} - \hat{\Sigma} - \frac{\lambda}{n} G = 0$ because $\omega_{\Gamma,ij}^* = 0$ for any $\gamma_{ij}' = 0$. $\Omega_{\Gamma'}^* = \Omega_\Gamma^*$ follows from the uniqueness of solution of the given optimization problem.

Ignorability of non-regular models

Theorem 3.1

Consider the prior on Γ *either Prior1 or Prior2 with* $q < \frac{1}{2}$. *The posterior probability of a non-regular model* Γ *is always less than that of the corresponding regular submodel* Γ'.

(proof)

- It is clear that $\{u_{\Gamma,ij} : ||\Delta_\Gamma||_2 \leq \epsilon_n\} \subset \{u_{\Gamma',ij} : ||\Delta_{\Gamma'}||_2 \leq \epsilon_n\} \cdots (*)$.
- Recall that
  $p(\Gamma|X^{(n)}) \propto C_\Gamma \int_{\Omega_\Gamma^* + \Delta_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{2} h_\Gamma(\Delta_\Gamma + \Omega_\Gamma^*)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij}$ and
  $h_\Gamma(\Delta_\Gamma + \Omega_\Gamma{}^*) = g_\Gamma(\Delta_\Gamma) + h_\Gamma(\Omega_\Gamma^*) + \log\det(\Omega_\Gamma{}^*)$

Ignorability of non-regular models

- As a consequence of Lemma 3.1, the term $h_\Gamma(\Omega_\Gamma^*) + \log\det(\Omega_\Gamma^*)$ coincide, because $\Omega_\Gamma^* = \Omega_{\Gamma'}^*$.
- Thus $$\frac{p(\Gamma|X^{(n)})}{p(\Gamma'|X^{(n)})} = \frac{C_\Gamma \int_{\Delta_\Gamma + \Omega_\Gamma^* \in \mathcal{M}_0^+} \exp(-\frac{n}{2} g_\Gamma(\Delta_\Gamma)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij}}{C_{\Gamma'} \int_{\Delta_{\Gamma'} + \Omega_{\Gamma'}^* \in \mathcal{M}_0^+} \exp(-\frac{n}{2} g_{\Gamma'}(\Delta_{\Gamma'})) \prod_{(i,j)\in \bar{E}_{\Gamma'}} du_{\Gamma',ij}}$$
- Recalling the remark 2.2,

$$\begin{aligned} \frac{p(\Gamma|X^{(n)})}{p(\Gamma'|X^{(n)})} &\approx \frac{C_\Gamma \int_{||\Delta_\Gamma||_2 \le \epsilon_n} \exp(-\frac{n}{2} g_\Gamma(\Delta_\Gamma)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij}}{C_{\Gamma'} \int_{||\Delta_{\Gamma'}||_2 \le \epsilon_n} \exp(-\frac{n}{2} g_{\Gamma'}(\Delta_{\Gamma'})) \prod_{(i,j)\in \bar{E}_{\Gamma'}} du_{\Gamma',ij}} \\ &\le \frac{C_\Gamma \int_{||\Delta_\Gamma||_2 \le \epsilon_n} \exp(-\frac{n}{2} g_\Gamma(\Delta_\Gamma)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij}}{C_{\Gamma'} \int_{||\Delta_\Gamma||_2 \le \epsilon_n} \exp(-\frac{n}{2} g_{\Gamma'}(\Delta_{\Gamma'})) \prod_{(i,j)\in \bar{E}_{\Gamma'}} du_{\Gamma',ij}} \end{aligned} \tag{6}$$

- The term $-\log\det(\Delta_\Gamma + \Omega_\Gamma^*) + \mathrm{tr}(\hat{\Sigma}\Delta_\Gamma)$ in g_Γ is larger than that in $g_{\Gamma'}$.

Ignorability of non-regular models

- Hence, combining this fact with the inequality (6),

$$\frac{p(\Gamma|X^{(n)})}{p(\Gamma'|X^{(n)})} \leq \frac{C_\Gamma}{C_{\Gamma'}} \int_{||\Delta_\Gamma||_2 \leq \epsilon_n} \exp(-\frac{n}{2}\frac{2\lambda}{n} \sum_{\gamma_{ij}=1,\gamma'_{ij}=0} |u_{\Gamma,ij}|) \prod_{(i,j)\in \bar{E}_\Gamma \cap \bar{E}^c_{\Gamma'}} du_{\Gamma,ij}$$

$$\leq \frac{C_\Gamma}{C_{\Gamma'}} \int \exp(-\lambda \sum_{\gamma_{ij}=1,\gamma'_{ij}=0} |u_{\Gamma,ij}|) \prod_{(i,j)\in \bar{E}_\Gamma \cap \bar{E}_{\Gamma'}} du_{\Gamma,ij}$$

$$= \frac{C_\Gamma}{C_{\Gamma'}} (\frac{2}{\lambda})^{\#\Gamma - \#\Gamma'} = (\frac{q}{1-q})^r \frac{\beta(\Gamma)}{\beta(\Gamma')} \leq (\frac{q}{1-q})^r$$

- The last inequality follows from the fact if the Prior2 is used, then $P(\bar{R} \geq \#\Gamma) \leq P(\bar{R} \geq \#\Gamma')$ since $\#\Gamma > \#\Gamma'$. Similarly, one can see $\frac{\beta(\Gamma)}{\beta(\Gamma')} \leq 1$ for Prior 1.
- For $q < \frac{1}{2}$, $\frac{q}{1-q} < 1$. So we conclude that posterior of Γ' is larger that of Γ. This leads us to concentrate only on regular model with $q < \frac{1}{2}$.

Outline

Error in Laplace approximation

Lemma 3.2

For any regular model Γ, with probability tending to one, the remainder term of function $h_\Gamma(\Omega_\Gamma)$, defined in the beginning of section (3), is bounded by $(p + \#\Gamma)||\Delta_\Gamma||_2^2(C_1||\Delta_\Gamma||_2 + C_2||\Delta_\Gamma||_2^2)/2$ for some constants $C_1, C_2 > 0$.

- By Taylor's expansion,

$$h_\Gamma(\Omega_\Gamma) = h_\Gamma(\Omega_\Gamma{}^*) + \frac{1}{2}\mathrm{vec}(\Delta_\Gamma)^t H_{\Omega_\Gamma^*}\mathrm{vec}(\Delta_\Gamma) + R_n \tag{7}$$

- Using the integral form of the remainder,

$$h_\Gamma(\Omega_\Gamma) = h_\Gamma(\Omega_\Gamma{}^*) + \mathrm{vec}(\Delta_\Gamma)^t[\int_0^1 (1-v)H_{\Omega_\Gamma^* + v\Delta_\Gamma}dv]\mathrm{vec}(\Delta_\Gamma) \tag{8}$$

Error in Laplace approximation

- Subtracting (8) from (7)

$$R_n = \text{vec}(\Delta_\Gamma)^t [\int_0^1 (1-v)(H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}) dv] \text{vec}(\Delta_\Gamma) \tag{9}$$

- Applying Cauchy-Shwartz inequality and theorem 1.1 to (9),

$$\begin{aligned} |R_n| &\leq ||\Delta_\Gamma||_2^2 || \int_0^1 (1-v)(H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}) dv ||_{(2,2)} \\ &\leq ||\Delta_\Gamma||_2^2 \int_0^1 (1-v) ||H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}||_{(2,2)} dv \\ &\leq \frac{1}{2} ||\Delta_\Gamma||_2^2 \max_{0 \leq v \leq 1} ||H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}||_{(2,2)} \\ &\leq \frac{1}{2} ||\Delta_\Gamma||_2^2 (p + \#\Gamma) \max_{0 \leq v \leq 1} ||H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}||_\infty \end{aligned} \tag{10}$$

Error in Laplace approximation

- It can be shown that each entry of $H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*} \leq C_1||\Delta_\Gamma||_2 + C_2||\Delta_\Gamma||_2^2$ for all $0 \leq v \leq 1$ and some constant $C_1, C_2 > 0$.
- Therefore, $||H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}||_\infty \leq C_1||\Delta_\Gamma||_2 + C_2||\Delta_\Gamma||_2^2$ for all $0 \leq v \leq 1$ and so $\max_{0 \leq v \leq 1} ||H_{\Omega_\Gamma^* + v\Delta_\Gamma} - H_{\Omega_\Gamma^*}||_\infty \leq C_1||\Delta_\Gamma||_2 + C_2||\Delta_\Gamma||_2^2$
- From inequality (10), we get the desired result

Error in Laplace approximation

Theorem 3.2

The error in the Laplace approximation tends to 0 in probability if $(p+\#\Gamma)^2\epsilon_n = n^{-\frac{1}{2}}(p+\#\Gamma)^{\frac{5}{2}}(\log p)^{\frac{1}{2}} \to 0$ *in probability, hence asymptotically negligible, where* $\epsilon_n = n^{-\frac{1}{2}}(p+\#\Gamma)^{\frac{1}{2}}(\log p)^{\frac{1}{2}}$, *which is posterior convergence rate established in theorem 2.2.*

(proof)

- With Taylor's expansion,

$$p(\Gamma|X^{(n)}) \propto C_\Gamma \int_{\Omega_\Gamma^*+\Delta_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{2}(h_\Gamma(\Omega_\Gamma^*) + \frac{1}{2}\mathrm{vec}(\Delta_\Gamma)^t H_{\Omega_\Gamma^*}\mathrm{vec}(\Delta_\Gamma) + R_n)) \prod_{(i,j)\in \bar{E}_\Gamma} du_{\Gamma,ij} \tag{11}$$

- For notational simplicity, denote $\prod_{(i,j)\in\bar{E}} du_{\Gamma,ij}$ by $d\Delta_\Gamma$.

Error in Laplace approximation

- By the remark 2.2,

$$\frac{\int_{||\Delta_\Gamma||_2 \leq \epsilon_n} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t H_{\Omega_\Gamma^*}\mathrm{vec}(\Delta_\Gamma) - \frac{n}{2}R_n)\mathrm{d}u_{\Gamma,ij}}{\int_{\Omega_\Gamma^*+\Delta_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t H_{\Omega_\Gamma^*}\mathrm{vec}(\Delta_\Gamma) - \frac{n}{2}R_n)\mathrm{d}u_{\Gamma,ij}} \to 1 \quad (12)$$

- With sufficiently large n, for $||\Delta_\Gamma||_2 \leq \epsilon_n$, by the result of lemma 3.2, $|R_n| \leq \frac{1}{2}C(p+\#\Gamma)||\Delta_\Gamma||_2^2\epsilon_n$ for some constant $C > 0$
- Thus, the upper and lower bounds of the integral $\int_{||\Delta_\Gamma||_2 \leq \epsilon_n} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t H_{\Omega_\Gamma^*}\mathrm{vec}(\Delta_\Gamma) - \frac{n}{2}R_n)\mathrm{d}u_{\Gamma,ij}$ are given by

$$\int_{||\Delta_\Gamma||_2 \leq \epsilon_n} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t(H_{\Omega_\Gamma^*} \mp C(p+\#\Gamma)\epsilon_n I)\mathrm{vec}(\Delta_\Gamma))\mathrm{d}u_{\Gamma,ij}$$

- If $(p+\#\Gamma) \to 0$, since the minimum eigenvalue of $H_{\Omega_\Gamma^*}$ is bounded away from 0,

$$\int_{||\Delta_\Gamma||_2 > \epsilon_n} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t(H_{\Omega_\Gamma^*} \mp C(p+\#\Gamma)\epsilon_n I)\mathrm{vec}(\Delta_\Gamma))\mathrm{d}u_{\Gamma,ij} \to 0.$$

Error in Laplace approximation

- Thus the bounds can be simplified into

$$\int_{\Delta_\Gamma + \Omega_\Gamma^* \in \mathcal{M}_0^+} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t(\mathrm{H}_{\Omega_\Gamma^*} \mp \mathrm{C}(\mathrm{p}+\#\Gamma)\epsilon_\mathrm{n}\mathrm{I})\mathrm{vec}(\Delta_\Gamma))\mathrm{d}u_{\Gamma,ij}$$

- Using the above bound, the ratio of the actual integral to the approximate integral has upper bound and lower bounds given by

$$\frac{\int_{\Delta_\Gamma + \Omega_\Gamma^* \in \mathcal{M}_0^+} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t(\mathrm{H}_{\Omega_\Gamma^*} \mp \mathrm{C}(\mathrm{p}+\#\Gamma)\epsilon_\mathrm{n}\mathrm{I})\mathrm{vec}(\Delta_\Gamma))\mathrm{d}u_{\Gamma,ij}}{\int_{\Omega_\Gamma^* + \Delta_\Gamma \in \mathcal{M}_0^+} \exp(-\frac{n}{4}\mathrm{vec}(\Delta_\Gamma)^t\mathrm{H}_{\Omega_\Gamma^*}\mathrm{vec}(\Delta_\Gamma))\mathrm{d}u_{\Gamma,ij}}$$
$$= [\det(\mathrm{I} \mp \mathrm{C}(p+\#\Gamma)\epsilon_n(\mathrm{H}_{\Omega_\Gamma}{}^*)^{-1}]^{-\frac{1}{2}} \tag{13}$$

- The above expression must lie between $[1 \mp \mathrm{C}(p+\#\Gamma)\epsilon_n\{\mathrm{eig}_1(\mathrm{H}_{\Omega_\Gamma}{}^*)\}^{-1}]^{-\frac{p+\#\Gamma}{2}}$, as the dimension of $\mathrm{H}_{\Omega_\Gamma}{}^*$ is $p+\#\Gamma$. Again the minimum eigenvalue of $\mathrm{H}_{\Omega_\Gamma}{}^*$ is bounded away from 0. So if $(p+\#\Gamma)^2\epsilon_n \to 0$, the above bound on the ratio goes to 1.

Outline

Simulation

- Simulation with models with sparse precision matrix
- The mean of models used is 0. The models with precision matrix $\Omega = ((\omega_{ij}))$ or covariance matrix $\Sigma = ((\sigma_{ij}))$ used in the simulation are followings:
  AR(1): $\sigma_{ij} = 0.7^{|i-j|}$
  AR(2): $\omega_{ii} = 1, \omega_{i-1,i} = \omega_{i,i-1} - 0.5, \omega_{i,i-2} = \omega_{i-2,i} = 0.25$
  Star: $\omega_{ii} = 1, \omega_{1,i} = \omega_{i,1} = 0.1$ and $\omega_{ij} = 0$, otherwise.
  Circle: $\omega_{ii} = 2, \omega_{i-1,i} = \omega_{i,i-1} = 1, \omega_{1,p} = \omega_{p,1} = 0.9$

Simulation

- We use specificity(SP), sensitivity(SE), and Matthews Correlation Coefficient(MCC) to measure the performance of the algorithm which will be described soon. Here, specificity and sensitivity are defined in the context of edge.

$$\begin{aligned} \mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \quad \mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{aligned} \tag{14}$$

- where TP,TN,FP, and FN respectively denote the true positives(edges included which are present in the true model), true negatives(edges excluded which are absent in the true model), false positives(edges included which are absent in the true model), and false negatives(edges excluded which are present in the true model)

Simulation

- Note that for estimated $\hat{\Omega}$, we say there is no edge between node i and j if $|\hat{\omega}_{ij}| \leq 0.1^3$.
- To compare Bayesian method with frequentist method, we also do simulation for frequentist method with the same models.
- Corresponding to each model, generate samples of size $n = 100, 200$ and dimension $p = 30, 50, 100$.
- The penalty parameter λ for the graphical lasso algorithm is chosen such that $\frac{\lambda}{n} = 0.5$ for model AR(1),AR(2),and Circle, and $\frac{\lambda}{n} = 0.2$ for Star. Set $q = 0.4$.
- Run 50 replications of each model and calculate SP, SE, and MCC in each replication. Then, average each measure over the replications and calculate standard deviation of each measure also.
- Provide ROC curves corresponding to the various models with values of the penalty parameter ranging between 0 1, plotting Sensitivity against False Positive Rate in case $n = 100$ and $p = 30, 50$

Simulation

- For Bayesian method, we set the restriction of model size $\bar{r}$ by that of frequentist method with the same penalty parameter.
- The idea to find posterior mode is not good. For example, in AR(1) model with $n = 100, p = 30$, we have $\bar{r} = 29$. We have to search total $\sum_{k=0}^{29} \binom{\binom{p}{2}}{k}$ models. Considering $\binom{\binom{p}{2}}{k} \geq 10^{45}$ when $k = 29$, this is impossible.
- Also, the posterior tends to decay.
- Hence we need to resort to other method. One possibility was RJMCMC, which turned out to be inappropriate for simulation.

Simulation

- We may use Metropolis-Hastings algorithm to obtain MCMC samples of model indicator Γ. In each replication of model, we choose final model by Median Probability Model(denoted by "MPP"). We may suggest symmetric proposal distribution. (500 MCMC sample with 100 burn-in)
- Suppose there are two p-dimension models $G = (V, E)$ and $G' = (V, E')$ with model indicators Γ, Γ' respectively. Considering Γ as vector, we denote $\Gamma(i, j) = 1$ if there is edge between node i and j and $\Gamma(i, j) = 0$ otherwise. Let $\pi(\Gamma|\Gamma')$ symmetric proposal distribution, which samples Γ' uniformly from the graphs that differ from Γ in one position. To be more specific, we set $\Gamma'(i, j) = 1$ for one pair (i,j) uniformly sampled from $V \times V \setminus E$, i.e., $\Gamma(i, j) = 0$ and $\Gamma'(i, j) = 0$ for one pair (i,j) uniformly sampled from E, i.e., $\Gamma(i, j) = 1$. So we have equal probability to set $\Gamma'(i, j) = 1$ or $\Gamma'(i, j) = 0$. Set $\Gamma'(-i, -j) = \Gamma(i, j)$, i.e. set Γ' as the same as Γ except for pair (i, j). Using this symmetric proposal distribution, we have following algorithm for MCMC sampling of Γ.

Simulation

Algorithm 1 Metropolis-Hastings algorithm for sampling $\mathbf{\Gamma}$

1: Initial model indicator : $\mathbf{\Gamma}^{(0)}$, Given data : $\mathbf{X^{(n)}}$
2: **for** i=0,1,...,k **do**
3: $\quad \mathbf{\Gamma}^{\text{temp}} \sim \pi(\mathbf{\Gamma}|\mathbf{\Gamma}^{(i)})$
4: $\quad$ Check regularity of $\mathbf{\Gamma}^{\text{temp}}$
5: $\quad$ If $\mathbf{\Gamma}^{\text{temp}}$ is regular, $\mathbf{\Gamma}^{\text{cand}} = \mathbf{\Gamma}^{\text{temp}}$. Else, repeat $3 \sim 4$.
6: $\quad \alpha_i = \min\{1, \frac{\mathbf{p^*(\Gamma^{cand}|X^{(n)})}}{\mathbf{p^*(\Gamma^{(i)}|X^{(n)})}}\}$
7: $\quad U_i \sim U(0,1)$
8: $\quad$ If $U_i \leq \alpha_i$, $\mathbf{\Gamma}^{(i+1)} = \mathbf{\Gamma}^{\text{cand}}$. Else, $\mathbf{\Gamma}^{(i+1)} = \mathbf{\Gamma}^{(i)}$.
9: **end for**

Simulation

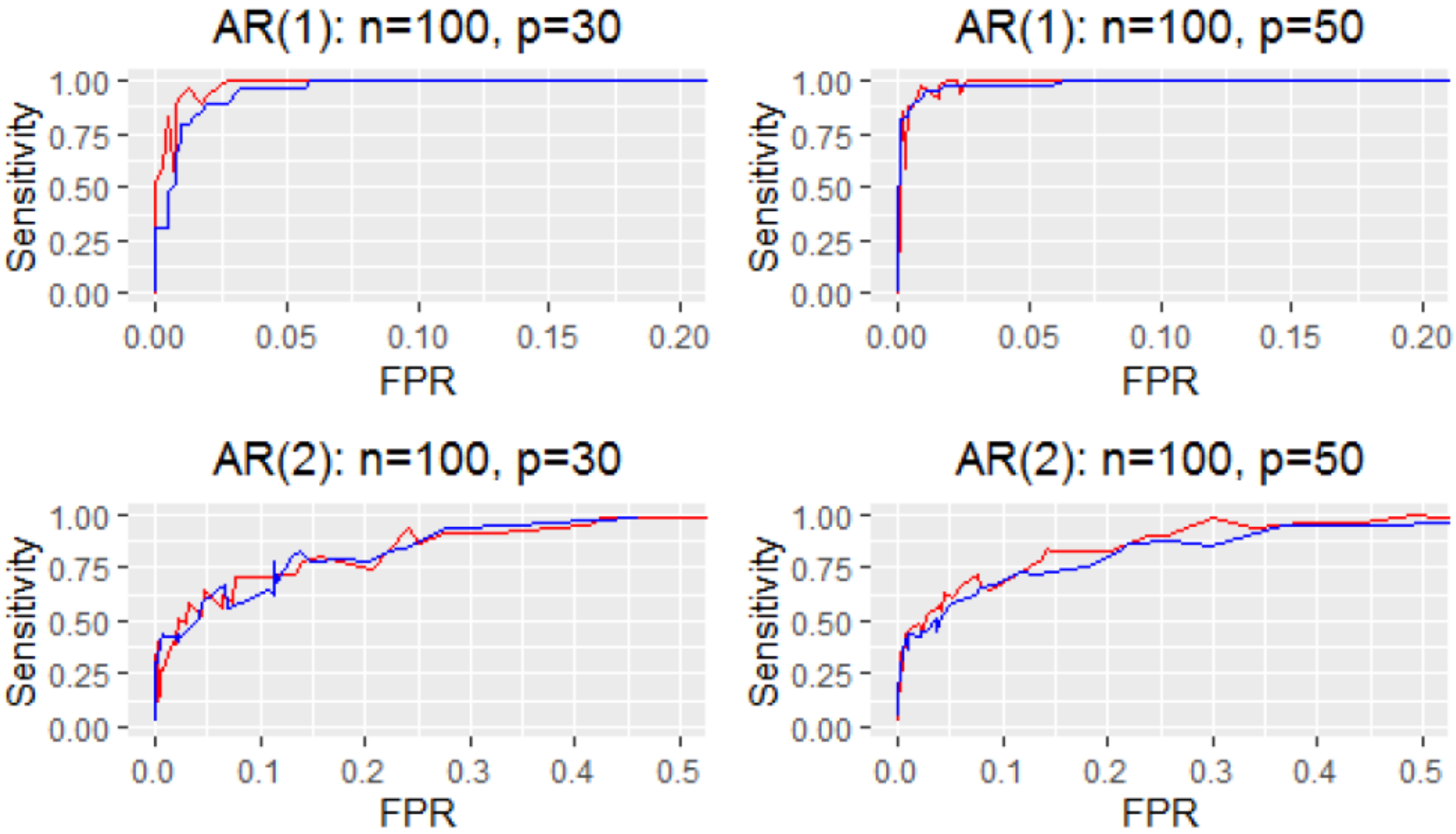

Figure 1: ROC curves for AR(1) and AR(2) structures. Red line : GL, Blue line: MPP

Simulation

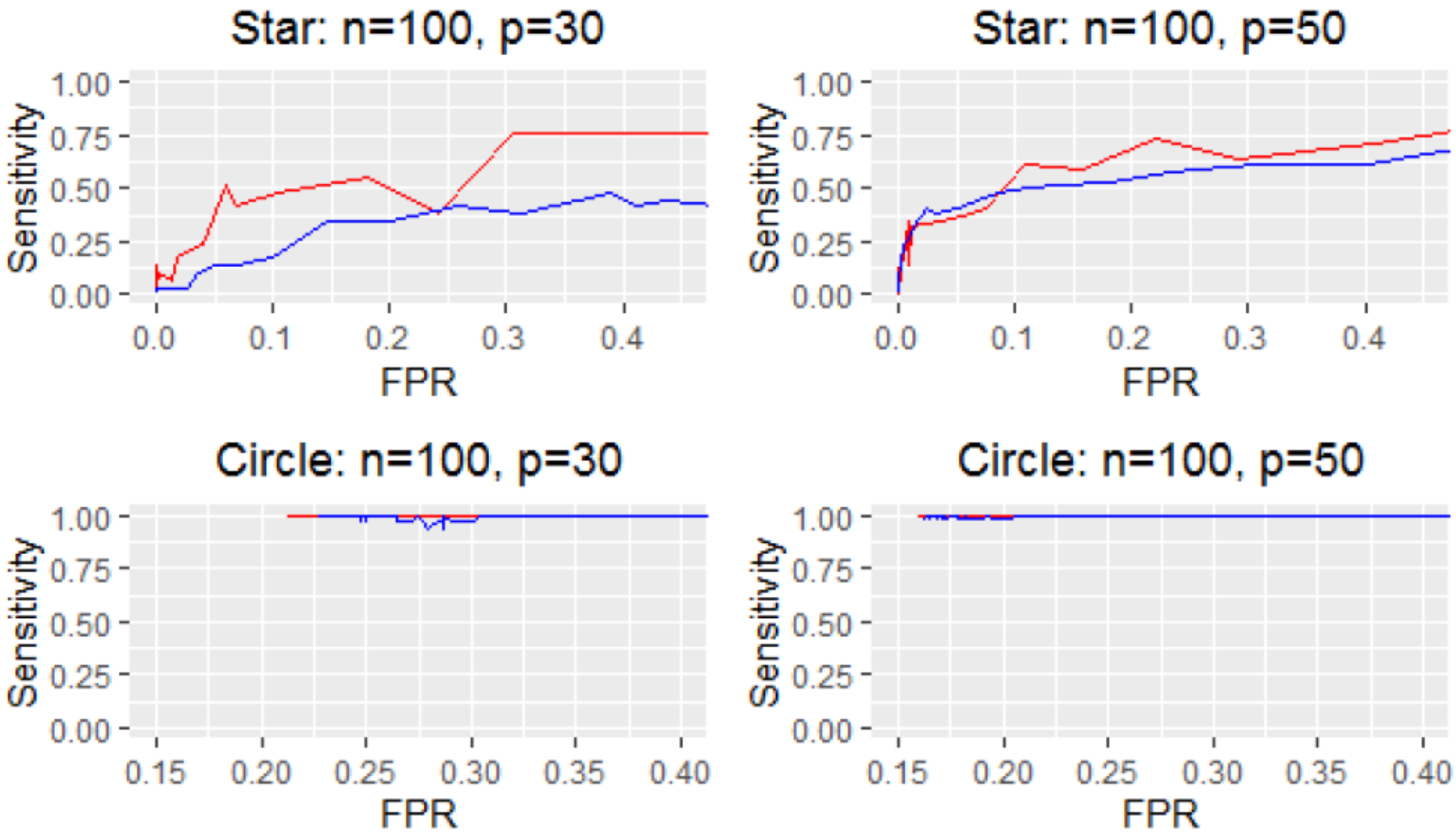

Figure 2: ROC curves for Star and Circle. Red line : GL, Blue line: MPP

Simulation

Model	p	n=100						n=200					
		MPP			GL			MPP			GL		
		SP	SE	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE	MCC
AR(1)	30	0.977 (0.009)	0.919 (0.054)	0.813 (0.048)	0.977 (0.011)	0.970 (0.045)	0.844 (0.055)	0.980 (0.007)	0.972 (0.038)	0.861 (0.041)	0.982 (0.007)	0.992 (0.012)	0.886 (0.035)
	50	0.986 (0.004)	0.893 (0.044)	0.802 (0.038)	0.985 (0.005)	0.969 (0.038)	0.839 (0.041)	0.990 (0.003)	0.962 (0.032)	0.869 (0.032)	0.990 (0.004)	0.992 (0.011)	0.888 (0.035)
	100	0.994 (0.001)	0.872 (0.038)	0.804 (0.030)	0.993 (0.002)	0.957 (0.029)	0.837 (0.029)	0.995 (0.001)	0.959 (0.027)	0.878 (0.028)	0.995 (0.001)	0.994 (0.010)	0.888 (0.022)
AR(2)	30	0.983 (0.007)	0.434 (0.050)	0.546 (0.046)	0.978 (0.010)	0.475 (0.044)	0.562 (0.038)	0.988 (0.007)	0.485 (0.031)	0.613 (0.029)	0.988 (0.006)	0.488 (0.029)	0.614 (0.029)
	50	0.987 (0.004)	0.409 (0.034)	0.518 (0.032)	0.983 (0.005)	0.483 (0.032)	0.559 (0.026)	0.993 (0.003)	0.464 (0.024)	0.609 (0.026)	0.993 (0.003)	0.485 (0.023)	0.621 (0.027)
	100	0.991 (0.002)	0.400 (0.023)	0.499 (0.022)	0.989 (0.002)	0.475 (0.026)	0.536 (0.022)	0.997 (0.001)	0.442 (0.019)	0.607 (0.017)	0.996 (0.001)	0.486 (0.017)	0.630 (0.013)
Star	30	0.965 (0.006)	0.229 (0.077)	0.225 (0.080)	0.954 (0.013)	0.297 (0.104)	0.256 (0.098)	0.990 (0.006)	0.167 (0.093)	0.263 (0.122)	0.994 (0.004)	0.237 (0.084)	0.384 (0.109)
	50	0.965 (0.007)	0.319 (0.074)	0.264 (0.061)	0.948 (0.009)	0.479 (0.104)	0.331 (0.067)	0.994 (0.002)	0.387 (0.078)	0.516 (0.074)	0.994 (0.003)	0.498 (0.088)	0.606 (0.061)
	100	0.940 (0.005)	0.992 (0.010)	0.484 (0.018)	0.940 (0.004)	1.000 (0.000)	0.489 (0.013)	0.988 (0.002)	1.000 (0.000)	0.793 (0.019)	0.988 (0.002)	1.000 (0.000)	0.787 (0.023)
Circle	30	0.784 (0.021)	1.000 (0.000)	0.448 (0.022)	0.746 (0.015)	1.000 (0.000)	0.411 (0.013)	0.754 (0.013)	1.000 (0.000)	0.418 (0.012)	0.724 (0.018)	1.000 (0.000)	0.392 (0.015)
	50	0.828 (0.010)	0.976 (0.019)	0.395 (0.016)	0.830 (0.009)	1.000 (0.000)	0.408 (0.011)	0.836 (0.007)	0.996 (0.009)	0.413 (0.010)	0.834 (0.006)	1.000 (0.000)	0.412 (0.008)
	100	0.892 (0.003)	0.984 (0.012)	0.371 (0.008)	0.891 (0.004)	1.000 (0.000)	0.376 (0.006)	0.905 (0.003)	0.997 (0.006)	0.402 (0.007)	0.904 (0.003)	1.000 (0.000)	0.400 (0.006)

Simulation

- Bayesian method performs slightly better than the frequentist method in terms of specificity, but suffers in sensitivity.
- This is because we chose q $< \frac{1}{2}$ to focus on regular models, allowing less edges to the model.
- The performance of Bayesian model tends to get lower as the dimension grows.
- Sensitivity was not good in AR(2) and Star models for both methods.
- In ROC curves, as the penalty parameter λ tends to get larger, sensitivity grows at the cost of higher false positive rate.

Illustration with real data

- Check tendency of performance of Bayesian method with real data
- Protein data from Rpacakge sparsebn consisiting of $n = 7466$ observations of $p = 11$ continuous variables corresponding to different proteins in human immune system cells.
- Penalty parameter λ is chosen such that $\frac{\lambda}{\mathrm{n}} = 0.4$ and $\mathrm{q} = 0.4$.
- For Bayesian method, we obtain 10000 MCMC samples with 2000 burn-in.

Illustration with real data

Method	SP	SE	MCC
MPP	0.636	0.364	0.000
GL	0.606	0.455	0.0602

Table 2: Simulation result for protein data

True Network

Frequentist Network

Bayesian Network

Illustration with real data

- The model estimated by Bayesian method shows less edges, hence suffers in sensitivity.
- But frequentist method performs is not good in specificity compared to Bayesian method and more edges are connected to Jnk, Akt in the model estimated by frequentist method than true network. Along with table 2, one can see what we wanted to verify with real data.

Discussion

- Bayesian method has mainly two drawbacks. Efficiency and Sensitivity.
- To explain poor efficiency, we need to check regularity condition and sample Γ again until we have regular model.
- Also, in estimating precision matrix Ω_Γ corresponding to regular model Γ, we use frequentist method. The algorithm of it requires very long time as the dimension grows especially in sparse models.
- However, the method is brilliant in that it introduced the sparsity to the model. One may hope this method to be improved in efficiency and sensitivity. Also, one can consider the problem of choosing penalty parameter λ.

References

- Ghosal, Ghosh and Van Der Vaart, "Convergence Rates of Posterior Distributions", Annals of Statistics, 2000, 28(2), 500-531
- Wang, "Bayesian Graphical Lasso Models and Efficient Posterior Computation", Bayesian Analysis, 2012, 7(4), 867-886
- A.J.Rothman et al, "Sparse permutation invariant covariance estimation", Electronic Journal of Statistics, 2008, 2, 494-515
- Liu and Martin, "An empirical G-Wishart prior for sparse high-dimensional Gaussian graphical models", arXiv, 2019, 1912.03807, 1-36

The End