Bayesian estimation of sparse precision matrices

Subhashis Ghoshal,
North Carolina State University

CANSSI–SAMSI Workshop:
Geometric Topological and Graphical Model Methods in Statistics Fields Institute, Toronto, Canada, May 22-23, 2014

Based on collaborations with Sayantan Banerjee

Estimation of Large Precision Matrix

- Consider multivariate Gaussian data $X \sim N_p(0, \Sigma)$.
- Let $\Omega = \Sigma^{-1}$ be the precision matrix.
- If the p variables are represented as the vertices of a graph G, then the absence of an edge between any two vertices j and j', which means conditional independence given others, is equivalent to $\omega_{jj'} = 0$.
- A graph can be used to control sparsity in Ω.

Graphical Lasso for Sparse Precision Matrix

- Developed in various papers — Meinshausen and Bühlman (2006), Yuan and Lin (2007), Banerjee et al. (2008), Friedman, Hastie and Tibshirani (2008).
- Maximize $\log\det\Omega - \mathrm{tr}(S\Omega) - \lambda\|\Omega\|_1$ subject to p.d. Ω, where S is the sample covariance matrix $n^{-1}\sum_{i=1}^{n} X_i X_i'$.
- Computation is doable in $O(p^3)$ steps by R package Glasso. Faster algorithms are possible assuming some special structure.

Convergence rate of Graphical Lasso

- Convergence rate studied by Rothman et al. (2008). If $\lambda \asymp \sqrt{(\log p)/n}$, convergence rate in Frobenius (aka Euclidean) norm is $\sqrt{((p+s)\log p)/n}$, where s is the number of non-zero off-diagonal entries.

Bayesian Graphical Lasso

- Wang (2012): Put independent exponential prior on diagonal entries, Laplace on off-diagonals, subject to positive definiteness restriction.
  - $\Omega \sim \mathcal{P}$: $\omega_{ii} \overset{\text{iid}}{\sim} \lambda e^{-\lambda \omega_{ii}}$, $\omega_{ii} > 0$, $\omega_{ij} \overset{\text{ind}}{\sim} \frac{\lambda}{2} e^{-\lambda |\omega_{ij}|}$, $i \neq j$,
  - $\Omega \sim \mathcal{P} | \Omega \in \mathcal{M}^+$, the set of positive definite matrices.
- Posterior mode is graphical Lasso.
- Full posterior easily computable by MCMC.
- Not a real sparse prior. Posterior sits on non-sparse matrices, and hence cannot converge near the truth in high dimension.

(Really sparse) Bayesian Graphical Lasso

- Real sparsity can be introduced by an extra point mass at zero for off-diagonal entries.
- Γ: $\gamma_{i,j} = \mathbb{1}((i,j) \in E)$, $p(\Omega|\Gamma) \propto \prod_{\gamma_{ij}=1} \{\exp(-\lambda|\omega_{ij}|)\} \prod_{i=1}^{p} \left\{\exp\left(-\frac{\lambda}{2}\omega_{ii}\right)\right\} \mathbb{1}_{\mathcal{M}^+}(\Omega)$.
- $p(\Gamma) \propto q^{\#E}(1-q)^{\binom{p}{2}-\#E} | \#\Gamma \leq R$,
- Maximum model size R has Poisson-like tail.

Posterior distribution

$$
\begin{aligned}
p(\Omega, \Gamma | X) &\propto p(X|\Omega, \Gamma) p(\Omega|\Gamma) p(\Gamma) \\
&= \{\det(\Omega)\}^{n/2} \exp\left\{-\frac{n}{2}\mathrm{tr}(\hat{\Sigma}\Omega)\right\} \\
&\quad \times \prod_{\gamma_{ij}=1} \{\exp(-\lambda|\omega_{ij}|)\} \prod_{i=1}^{p} \left\{\exp\left(-\frac{\lambda}{2}\omega_{ii}\right)\right\} \\
&\quad \times q^{\#E}(1-q)^{\binom{p}{2}-\#E}.
\end{aligned}
$$

Model posterior probabilities

$$
p(\Gamma|X) \propto \int_{\Omega \in \mathcal{M}^+} \exp(\frac{n}{2} h(\Omega)) \prod_{(i,j)\in\mathcal{V}_\Gamma} d\omega_{ij},
$$

where

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

Issues

- Computation becomes a challenge. Traditional MCMC/RJMCMC are too slow.
- What can we say about posterior convergence rates?

Posterior Convergence Rate

Theorem

Let $X_1, \ldots, X_n \overset{iid}{\sim} N_p(0, \Omega^{-1})$ *and the true precision matrix* $\Omega_0 \in \mathcal{U}(s, \epsilon_0) = \{\Omega : \#\{(i,j) : \omega_{ij} \neq 0, i \neq j\} \leq s,\ 0 < \epsilon_0 \leq \min \mathrm{eig}_j(\Omega) \leq \max \mathrm{eig}_j(\Omega) \leq \epsilon_0^{-1} < \infty\}$. *Then for some* $M > 0$, *the posterior probability* $\mathrm{P}(\|\Omega - \Omega_0\|_2 > M\epsilon_n \mid X) \to 0$, *for* $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ *and* $\|\cdot\|_2$ *stands for the Frobenius (Euclidean) norm.*

Steps in Proving Posterior Convergence Rate

- Frobenius distance is comparable with the Hellinger distance between $N_p(0, \Omega)$ and $N_p(0, \Omega')$, the square root of their Kullback-Leibler (KL) divergence and the Euclidean norm for eigenvalues $d_1, \ldots, d_p$ of $\Omega_0^{-1/2} \Omega \Omega_0^{-1/2}$ centered by 1: $\sum_{j=1}^{k} |d_j - 1|^2$. If either Hellinger or Frobenius is small, all d_js are uniformly close to 1, allowing Taylor's expansion.
- Use general theory of posterior convergence rate [G, Ghosh and van der Vaart (2000)] by bounding Hellinger entropy of a "sieve" by $n\epsilon_n^2$ with at least $1 - e^{-bn\epsilon_n^2}$ prior probability, and assuring that the prior probability of the ϵ_n-size KL neighborhood of the true density is at least $e^{-n\epsilon_n^2}$.
- In view of the equivalence of distances, need bounding entropy and obtain prior concentration in terms of Frobenius norm.

Steps (contd.)

- Define sieve $\mathcal{P}_n$ so maximum number of edges $\bar{r} < \binom{p}{2}/2$ and each entry at most L, where $\bar{r}$ and L are to be determined.
- Metric entropy $\leq \log[\bar{r}\left(\frac{L}{\epsilon_n}\right)^{\bar{r}}\binom{\binom{p}{2}}{\bar{r}}]$ need to make $\leq n\epsilon_n^2$.
- Choose $L \in [bn\epsilon_n^2, bn\epsilon_n^2 + 1]$ to ensure that $\binom{p}{2}\exp(-L) \leq \exp(-b'n\epsilon_n^2)$. Note tail probability $\leq e^{-bn\epsilon_n^2}$.
- Requirement on $\bar{r}$ becomes $\log \bar{r} + \bar{r}\log p + \bar{r}\log(\frac{1}{\epsilon_n}) + \bar{r}\log(n\epsilon_n^2) \asymp n\epsilon_n^2$.
- $\mathrm{P}(\mathcal{P}_n^c) \leq \mathrm{P}(\bar{R} > \bar{r}) + \exp(-b_3 n\epsilon_n^2)$ holds if $\bar{r}$ is like $n\epsilon_n^2/\log n$ under the Poisson tail condition.
- Bounding the KL divergences by $\sum_{j=1}^{p}|d_j - 1|^2$, suffices to lower bound $\mathrm{P}\{\max|d_j - 1| < \epsilon_n/p\} = \mathrm{P}\{\|\Omega - \Omega_0\|_\infty < \epsilon_n/p\} \geq (c'\epsilon_n/p)^{p+s}$ using "independence".
- $(p+s)(\log p + \log(1/\epsilon_n)) \asymp n\epsilon_n^2$, giving convergence rate $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$.

Approximate Posterior Model Probabilities

- We use Laplace approximation — in each submodel, expand log posterior density around posterior mode and evaluate normal integrals analytically.
- Posterior mode is graphical lasso restricted to the submodel.
- To find the Laplace approximation, need to calculate the Hessian.

Hessian

$U = \Omega - \Omega^*$, where Ω^* is the graphical lasso solution in the submodel.

$$p\{\Gamma|X\} \propto \exp\{n\, h(\Omega^*)/2\}\, \{\det(\Omega^*)\}^{-n/2} \int_{U+\Omega^* \in \mathcal{M}^+} \exp\{n\, g(U)/2\},$$

where $g(U)$ is

$$\log\det(U+\Omega^*) - \mathrm{tr}(\widehat{\Sigma} U) - \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} (|u_{ij} + \omega^*_{ij}| - |\omega^*_{ij}|) - \frac{\lambda}{n} \sum_{i=1}^{p} u_{ii}.$$

Hessian of $g(U)$ is the $\#\mathcal{V}_\Gamma \times \#\mathcal{V}_\Gamma$ matrix $H_{U+\Omega^*}$, with $\{(i,j),(l,m)\}$th entry

$$-\mathrm{tr}\left\{(U+\Omega^*)^{-1} E_{(i,j)} (U+\Omega^*)^{-1} E_{(l,m)}\right\},$$

$E_{(i,j)}$ is a binary matrix with 1 only at (i,j)th and (j,i)th location.

Laplace approximation

$$
\begin{aligned}
p^*\{\Gamma|X\} &\propto C_\Gamma \exp\{n\, h(\Omega^*)/2\}\{\det(\Omega^*)\}^{-n/2} \exp\{n\, g(0)/2\} \\
&\quad \times (2\pi)^{\#\mathcal{V}_\Gamma/2}(n/2)^{-\#\mathcal{V}_\Gamma/2}[\det\{-\frac{\partial g(U)}{\partial U \partial U^T}|_0\}]^{-1/2} \\
&= C_\Gamma \exp\{n\, h(\Omega^*)/2\}(2\pi)^{\#\mathcal{V}_\Gamma/2}(n/2)^{-\#\mathcal{V}_\Gamma/2}\{\det(H_{\Omega^*})\}^{-1/2}.
\end{aligned}
$$

Approximation is meaningful (i.e. differentiability hold) only if all the graphical lasso estimates of the off-diagonal elements corresponding to the graph generated by Γ are non-zero — coined as "regular models". Other models are non-regular models.

Ignorability of non-regular models

For a given nonregular submodel Γ, define its regular counterpart to be the model Γ by removing the edges having graphical lasso solution zero. Then as defined above, the graphical lasso solution corresponding to the two models are identical.

Theorem

If $q < 1/2$, then the posterior probability of a non-regular model Γ is always less than that of its regular submodel Γ'.

Accuracy of Laplace approximation

Theorem

The error in Laplace approximation of the posterior probability of a graphical model structure is asymptotically small if $(p+s)^2\epsilon_n \to 0$, *where* ϵ_n *is the posterior convergence rate, that is, the error in the Laplace approximation tends to zero if* $n^{-1/2}(p+s)^{5/2}(\log p)^{1/2} \to 0$.

Proof uses the bound — if sparsity is s, then with probability tending to 1, the remainder term in the expansion of $h(\Omega)$ around Ω^*, is bounded by $(p+s)\|\Omega-\Omega^*\|_2^2(C_1\|\Omega-\Omega^*\|_2 + C_2\|\Omega-\Omega^*\|_2^2)/2$.

How to test the method?

So far tested on

- AR(1) model, $\sigma_{ij} = 0.7^{|i-j|}$.
- AR(2) model, $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.5$, $\omega_{i,i-2} = \omega_{i-2,i} = 0.25$.

Compute the so called median probability model $\{j : \mathrm{P}(X_j \text{ included in model—data}) \geq 1/2\}$, based only on regular models that are with Hamming distance 1 (there are $O(p)$ such models instead of 2^p). We monitor specificity, sensitivity and Matthews Correlation Coefficient.

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

How is the method working?

Model	p	$n = 100$			$n = 200$		
		SP	SE	MCC	SP	SE	MCC
AR(1)	30	0.977	0.941	0.831	0.986	0.996	0.907
		(0.003)	(0.019)	(0.015)	(0.002)	(0.003)	(0.014)
	50	0.987	0.953	0.841	0.991	0.992	0.903
		(0.002)	(0.013)	(0.010)	(0.001)	(0.004)	(0.008)
	100	0.977	0.875	0.724	0.961	0.867	0.739
		(0.001)	(0.026)	(0.019)	(0.006)	(0.034)	(0.028)
	500	0.909	0.585	0.310	0.953	0.761	0.541
		(0.004)	(0.026)	(0.012)	(0.006)	(0.019)	(0.019)
AR(2)	30	0.975	0.470	0.546	0.987	0.495	0.617
		(0.003)	(0.014)	(0.013)	(0.002)	(0.008)	(0.008)
	50	0.983	0.462	0.541	0.993	0.489	0.629
		(0.001)	(0.013)	(0.011)	(0.001)	(0.005)	(0.007)
	100	0.943	0.460	0.383	0.938	0.453	0.438
		(0.003)	(0.015)	(0.010)	(0.008)	(0.005)	(0.007)
	500	0.781	0.383	0.104	0.831	0.434	0.183
		(0.005)	(0.077)	(0.010)	(0.004)	(0.014)	(0.007)

How is the method working? (contd.)

Model	p	$n = 100$			$n = 200$		
		SP	SE	MCC	SP	SE	MCC
Block	30	1.000	0.423	0.831	1.000	0.429	0.524
		(0.000)	(0.047)	(0.037)	(0.000)	(0.060)	(0.049)
	50	1.000	0.381	0.481	1.000	0.402	0.502
		(0.000)	(0.044)	(0.040)	(0.000)	(0.036)	(0.030)
	100	1.000	0.330	0.445	1.000	0.349	0.460
		(0.000)	(0.021)	(0.017)	(0.000)	(0.027)	(0.021)
Star	30	0.947	0.289	0.228	0.995	0.210	0.378
		(0.004)	(0.038)	(0.036)	(0.001)	(0.032)	(0.041)
	50	0.945	0.492	0.332	0.993	0.475	0.585
		(0.003)	(0.034)	(0.025)	(0.000)	(0.034)	(0.024)
	100	0.990	1.000	0.827	0.988	1.000	0.792
		(0.000)	(0.000)	(0.004)	(0.000)	(0.000)	(0.008)
Circle	30	0.733	1.000	0.399	0.719	1.000	0.388
		(0.004)	(0.000)	(0.003)	(0.005)	(0.000)	(0.004)
	50	0.831	1.000	0.409	0.833	1.000	0.411
		(0.003)	(0.000)	(0.003)	(0.002)	(0.000)	(0.003)
	100	0.891	1.000	0.378	0.903	1.000	0.399
		(0.001)	(0.000)	(0.002)	(0.008)	(0.000)	(0.002)

In the real world

We analyze closing prices of 452 stocks from the S&P 500 index during January 1, 2003 to January 1, 2008. The stocks are categorized into 10 Global Industry Classification Standard (GICS) — "Health Care", "Materials", "Industrials", "Consumer Staples", "Consumer Discretionary", "Utilities", "Information Technology", "Financials", "Energy", "Telecommunication Services".
Denoting Y_{tj} as the closing stock price for the jth stock on day t, we construct the 1257×452 data matrix S with entries
$s_{tj} = \log(Y_{(t+1)j}/Y_{tj})$, $t = 1, \ldots, 1257$, $j = 1, \ldots, 452$,
standardized to have mean zero and standard deviation one. We find the median probability model. The following color coded graph show interrelationships.

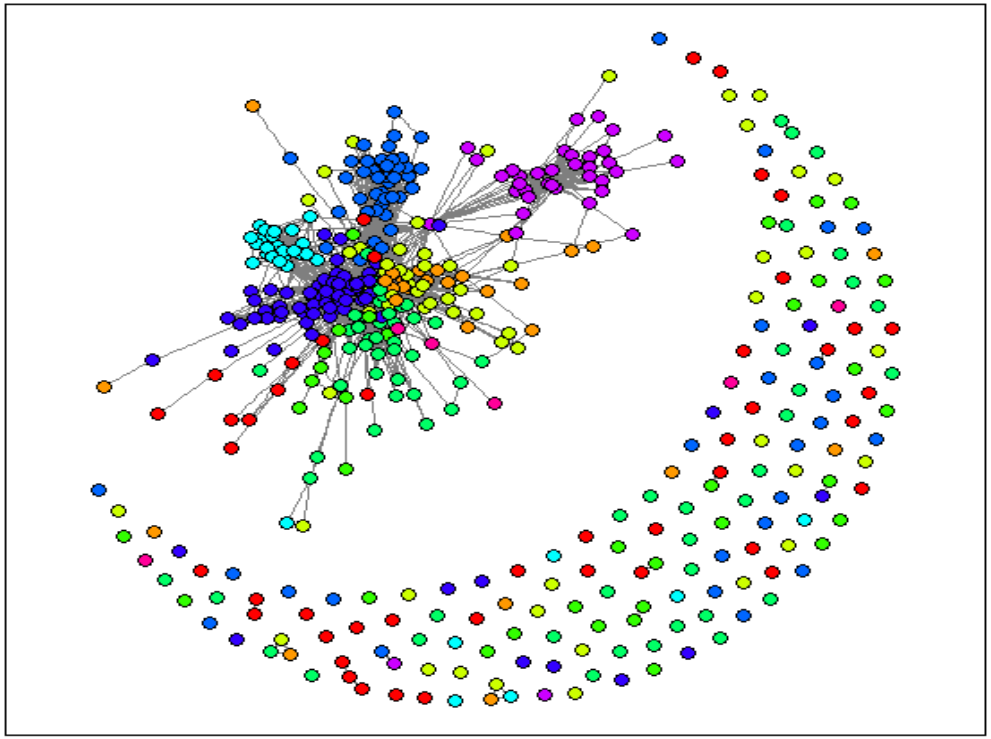

Figure: Graphical structure of the median probability model selected by the Bayesian graphical structure learning method.

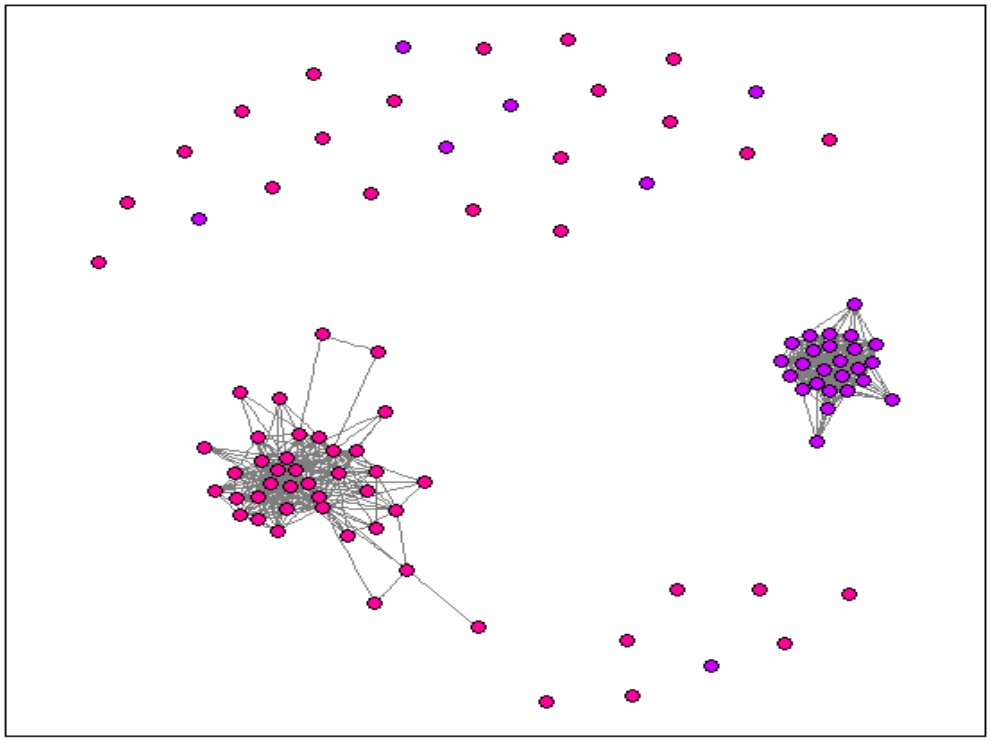

Figure: Graphical structure corresponding to the subgraph corresponding to the sectors "Utilities" [red] and "Information Technology" [violet].

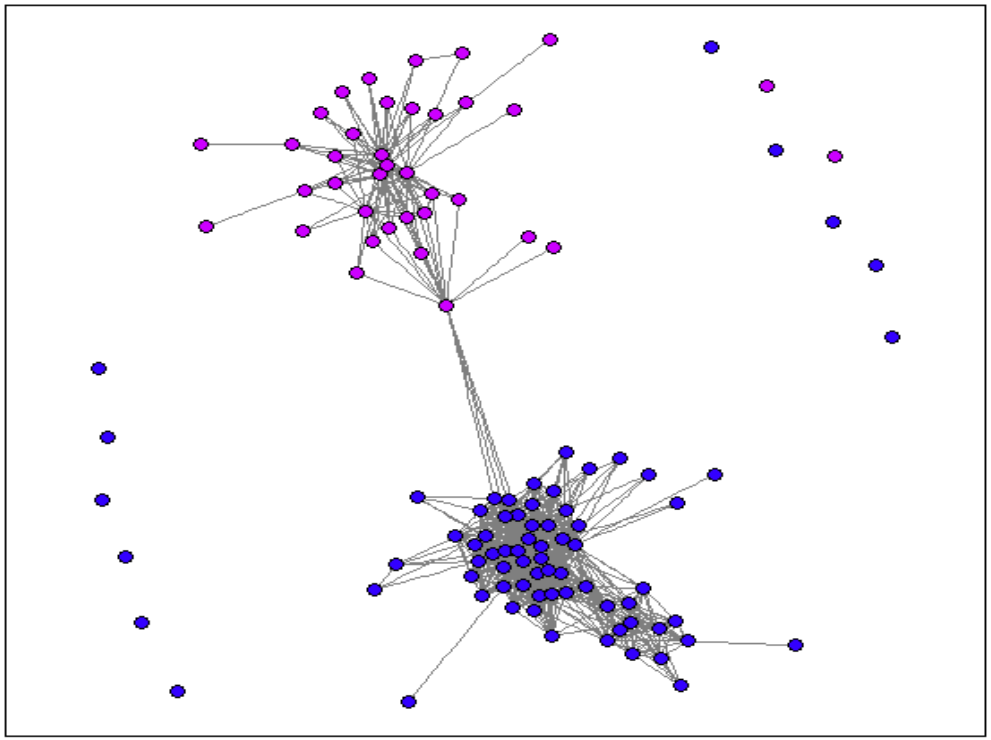

Figure: Graphical structure corresponding to the subgraph corresponding to the sectors "Financials" [blue] and "Energy" [violet].

Thank you