Bayesian inference for Gaussian graphical models beyond decomposable graphs

Kshitij Khare

Department of Statistics
University of Florida

Workshop on "High-Dimensional Covariance Matrices, Networks and Concentration Inequalities"

Outline

- Review of Bayesian Gaussian graphical models
- Generalized Bartlett graphs/Generalized $\mathcal{G}$-Wishart distributions
- Examples

Motivation

- Availability of high-dimensional data or "big data" from various applications
- Number of variables (p) much larger than (or sometimes comparable to) the sample size (n)
- Examples:
  - Biology: gene expression data
  - Environmental science: climate data on spatial grid
  - Finance: returns on thousands of stocks

Goal: Understanding relationships between variables

- Common goal in many applications: Understand complex network of relationships between variables
- Covariance matrix: a fundamental quantity to help understand multivariate relationships
- Even if estimating the covariance matrix is not the end goal, it is a crucial first step before further analysis

Challenges in high-dimensional estimation

- Covariance matrix (often denoted by Σ) has $O(p^2)$ unknown paramters
- If $p = 1000$, we need to estimate roughly 1 million parameters
- If sample size n is much smaller (or even same order) than p, this is not viable
- The sample covariance matrix (classical estimator) can perform very poorly in high-dimensional situations (not even invertible when $n < p$)

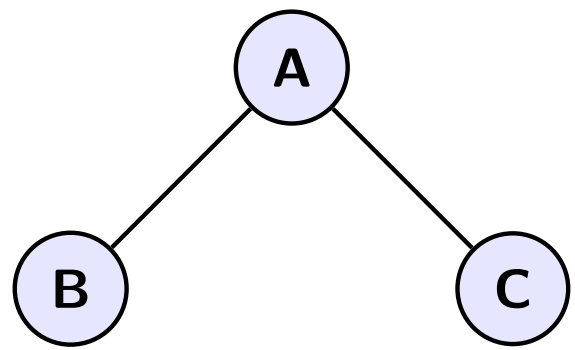
Is there a way out?

- **Reliably estimate small number of parameters** in Σ or $\Omega = \Sigma^{-1}$
- Set insignificant parameters to zero
- Gives rise to sparse estimates of Σ or Ω
- Sparsity pattern can be represented by graphs/networks

Gaussian Graphical Models: Sparsity in Ω

- Assume Ω (inverse covariance matrix) is sparse: corresponds to assuming conditional independences
- Sparsity pattern in Ω can be represented by an undirected graph $G = (V, E)$
- Build a graph from sparse Ω

$$\Omega = \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 2 & 0 \\ 0.3 & 0 & 1.2 \end{pmatrix} \end{array} \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array}$$



Are these models useful? Appropriate?

- **Many physical networks are assumed to be sparse**
- Complex networks (internet, citation networks, social networks) tend to be sparse [Newman, 2003]
- Genetic networks are sparse [Gardner et al, 2003, Jeong et al, 2001]

What do we want to learn from the data?

- **Data**: $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ i.i.d. $\mathcal{N}_p\left(\mathbf{0}, \Omega^{-1}\right)$

- **Model selection**: Estimating the underlying sparsity pattern/underlying graph $G = (V, E)$

- **Estimation**: Given a graph G, finding an estimate $\hat{\Omega} \in \mathbb{P}_G$ where

$$\mathbb{P}_G = \{\Omega : \Omega \text{ positive definite }, \Omega_{ij} = 0 \; \forall \; (i,j) \notin E\}$$

Bayesian estimation: $\mathcal{G}$-Wishart priors

- Popular class of priors on $\mathbb{P}_G$ (Dawid and Lauritzen (1993, AOS), Roverato (2000, Biometrika))
- Density with parameters $\delta > 0$ (**single shape parameter**) and U (positive definite, scale parameter) given by

$$\pi_{GW,\delta,U}(\Omega) \propto |\Omega|^{\frac{\delta}{2}} \exp\left(-\frac{1}{2} tr(\Omega U)\right).$$

- Conjugate prior: If prior is $\mathcal{G}$-Wishart, posterior is also $\mathcal{G}$-Wishart
- Closed form computations from posterior not possible in general
- Various approaches developed to sample from posterior

Bayesian estimation: Letac-Massam priors

- Introduced in Letac and Massam (2007, AOS)
- Generalization of $\mathcal{G}$-Wishart distributions with **multiple shape parameters** for differential shrinkage
- Conjugate prior: If prior is Letac-Massam, posterior is also Letac-Massam
- **Defined when the underlying graph G is decomposable**
- Closed form computations feasible

Decomposable graphs

A graph is decomposable if it does not contain a cycle of size greater than or equal to 4 as an induced subgraph

Vertices	Percentage
2	100
3	100
4	83
5	71
6	52
7	32
8	15
9	4.5
10	0.9

Figure: Percentages of decomposable graphs among connected non-isomorphic graphs with at most 10 vertices.

Our goal

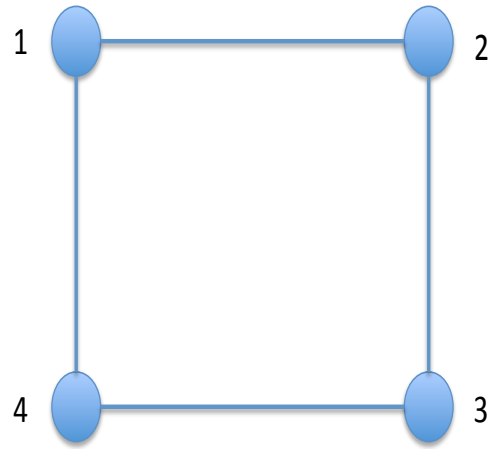
Develop a class of prior distributions on $\mathbb{P}_G$ which

- has multiple shape parameters
- contains $\mathcal{G}$-Wishart distributions as special case
- tractable posterior computation for a much larger class of graphs than decomposable graphs

Transformation to Cholesky space

- Let $\Omega = LDL^T$, where L is lower triangular with diagonal entries 1 and D is diagonal with positive diagonal entries
- Entries of L and D have a parallel interpretation as appropriate conditional regression coefficients and conditional variances
- Remember $\Omega \in \mathbb{P}_G$
- $L_I = ((L_{ij}))_{i>j,(i,j)\in E}$ - functionally independent entries of L

Transformation to Cholesky space: Example

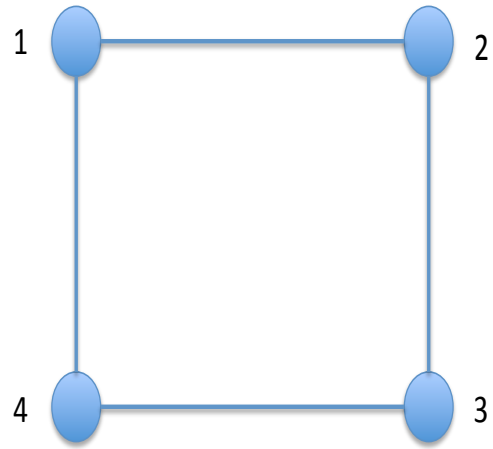


$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{21} & 0 & \Omega_{41} \\ \Omega_{21} & \Omega_{22} & \Omega_{32} & 0 \\ 0 & \Omega_{32} & \Omega_{33} & \Omega_{43} \\ \Omega_{41} & 0 & \Omega_{43} & \Omega_{44} \end{bmatrix} \qquad L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \mathbf{L_{21}} & 1 & 0 & 0 \\ L_{31} & \mathbf{L_{32}} & 1 & 0 \\ \mathbf{L_{41}} & L_{42} & \mathbf{L_{43}} & 1 \end{bmatrix}$$

Transformation to Cholesky space: Observations

- Entries of $\Omega \in \mathbb{P}_G$ not algebraically "free" of each other
- They are constrained by the fact that Ω is positive definite
- Entries in L_I and D can take arbitrary real and non-negative real values respectively
- Fact: For each $(i, j) \notin E$, L_{ij} is a (multivariate) polynomial in entries of L_I, D, D^{-1}
- Transformation $\Omega \to (L_I, D)$ bijection with Jacobian $\prod_{j=1}^{p} D_j^{\nu_j}$, where $\nu_j := |\{i : i > j, (i, j) \in E_\sigma\}|$

Transformation to Cholesky space: Example



$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \mathbf{L_{21}} & 1 & 0 & 0 \\ L_{31} & \mathbf{L_{32}} & 1 & 0 \\ \mathbf{L_{41}} & L_{42} & \mathbf{L_{43}} & 1 \end{bmatrix} \qquad L_{31} = 0,\ L_{42} = -\frac{L_{41} L_{21} D_{11}}{D_{22}}$$

Generalized $\mathcal{G}$-Wishart distributions

- (K., Rajaratnam, Saha (2018, JRSSB)) **Generalized G-Wishart density** with parameters $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_p) \in \mathbb{R}_+^p$ and $U \in \mathbb{M}_p^+$ is given by

$$\pi_{GGW,\boldsymbol{\delta},U}(\Omega) \propto \left(\prod_{i=1}^{p} D_{ii}(\Omega)^{\frac{\delta_i}{2}} \right) \exp\left(-\frac{1}{2} tr(\Omega U) \right).$$

- **G-Wishart density arises as a special case** of the generalized G-Wishart (by considering all the δ_i's to be equal)

- Family of generalized G-Wishart distributions defined above is a **conjugate family of prior distributions** for Gaussian graphical models.

Generalized Bartlett (GB) graphs

- $G' = (V, E')$ is called a cover of $G = (V, E)$ if $E \subseteq E'$
- (K., Rajaratnam and Saha (2018, JRSSB)) An undirected graph G is defined to be **Generalized Bartlett** if and only if $G = (V, E)$ has a decomposable cover $\widetilde{G} = (V, \widetilde{E})$ such that every triangle in $\widetilde{E}$ contains an edge from E
- That is for any $u, v, w \in V$ such that $(u, v), (v, w), (u, w) \in \widetilde{E}$, at least one of $(u, v), (v, w), (u, w)$ belongs to E

Examples of GB graphs

- All decomposable graphs
- Any cycle
- Any $2 \times k$ or $3 \times k$ lattice

Examples of GB graphs

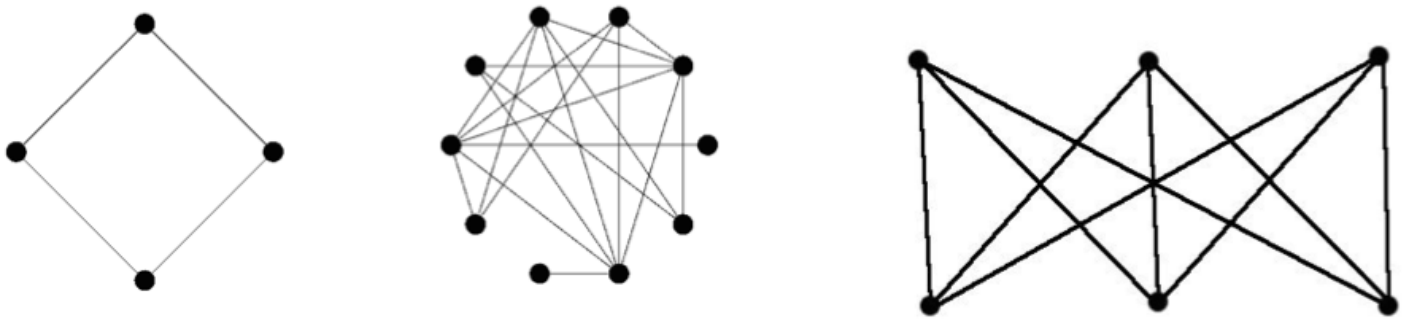


Figure: (left and middle) Examples of Generalized Bartlett graphs. (right) The bipartite graph $K_{3,3}$: the only 6-vertex connected graph which is not Generalized Bartlett.

Generalized Bartlett graphs: size comparison with decomposable graphs

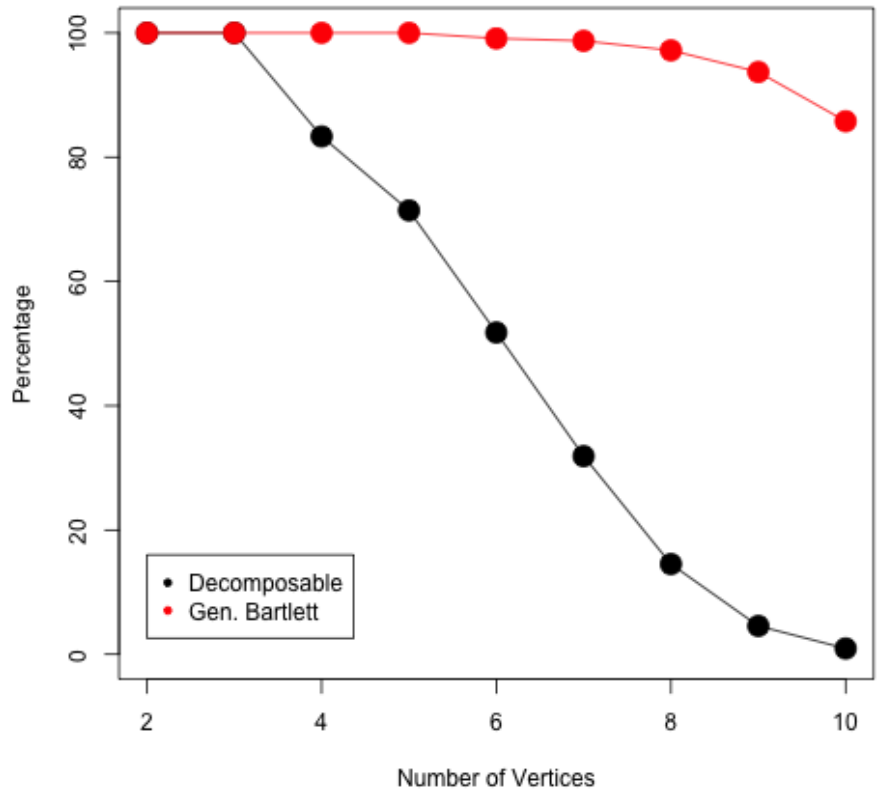


Figure: Plot comparing the percentage of Generalized Bartlett graphs with decomposable graphs

Properties of GB graphs

- Any subgraph of a GB graph is a GB graph
- If all prime components of a graph are GB, then the graph is GB
- Other expansion properties ...

Generalized $\mathcal{G}$-Wishart distributions and Generalized Bartlett graphs

If $\Omega \sim \pi_{GGW,\delta,U}$, then the induced density for (L_I, D) is proportional to

$$\left(\prod_{i=1}^{p} D_{ii}^{\frac{\delta_i}{2}+\nu_i}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} U_{ij} \sum_{k=1}^{\min(i,j)} L_{ik}L_{jk}D_{kk}\right)$$

Theorem (K., Rajaratnam and Saha (2018))

Let G be a Generalized Bartlett Graph. If $\Omega = LDL^T \sim \pi_{GGW,\delta,U}$ then

- **the conditional posterior density of any independent entry** L_{ij}, given all other entries of L_I and $\widetilde{D}$, **is univariate normal**,
- **the conditional posterior density of** $\widetilde{D}_k = \frac{D_k}{D_{k-1}}$ given L_I and other entries of $\{\widetilde{D}_{k'}\}_{k' \neq k}$ is either **a Generalized Inverse Gaussian or Gamma**.

Recall our goal

Develop a class of prior distributions on $\mathbb{P}_G$ which

- has multiple shape parameters
- contains $\mathcal{G}$-Wishart distributions as special case
- tractable posterior computation for a much larger class of graphs than decomposable graphs

Comparison with Accept-Reject

- Accept-Reject algorithm to sample from $\mathcal{G}$-Wishart for a general graph G developed in Carvalho and Wang (2010)
- Can be easily extended to Generalized $\mathcal{G}$-Wishart
- Accept-Reject algorithms can suffer from very low acceptance probability issues as the dimension increases

Comparison with Accept-Reject

- Carvalho and Wang's algorithm does well in small dimensional settings: example in their paper is with $p = 7$, size of largest prime component 4
- But problem of low acceptance probabilities emerges if we increase the size of the largest prime component
- If G is a 12-cycle, the accept-reject algorithm needs 5 hours per iteration on average
- On the other hand, 10000 iterations of the Gibbs sampler can be performed in 4 minutes, and lead to accurate estimates of posterior quantities

Comparison with Metropolis-Hastings

- Metropolis-Hastings algorithm to sample from $\mathcal{G}$-Wishart for a general graph G developed in Mitsakakis et al. (2011)
- Can be easily extended to sample from Generalized $\mathcal{G}$-Wishart with parameters δ and U
- Let $U^{-1} = T^t T$ be the Cholesky decomposition of U^{-1}.
- If $\frac{t_{ij}}{t_{jj}}$ is large for every $i > j, (i,j) \in E$, then the acceptance probabilities in the Metropolis-Hastings algorithm can be too small

Comparison with Metropolis-Hastings

- Let G be a 5×3 grid
- Choose U such that $t_{ij}/t_{jj} = \lambda$ for every $i > j, (i,j) \in E$
- Fact: There exists a matrix $g_{\delta}(\Omega)$ such that

$$E_{U,\delta}[g(\Omega)_{ij}] = U_{ij}$$

  for every $i > j, (i,j) \in E$

- Can use the maximum difference between the MCMC estimate and theoretical expectation of $g(\Omega)_{ij}$ as a measure of accuracy of a given MCMC algorithm

Comparison with Metropolis-Hastings

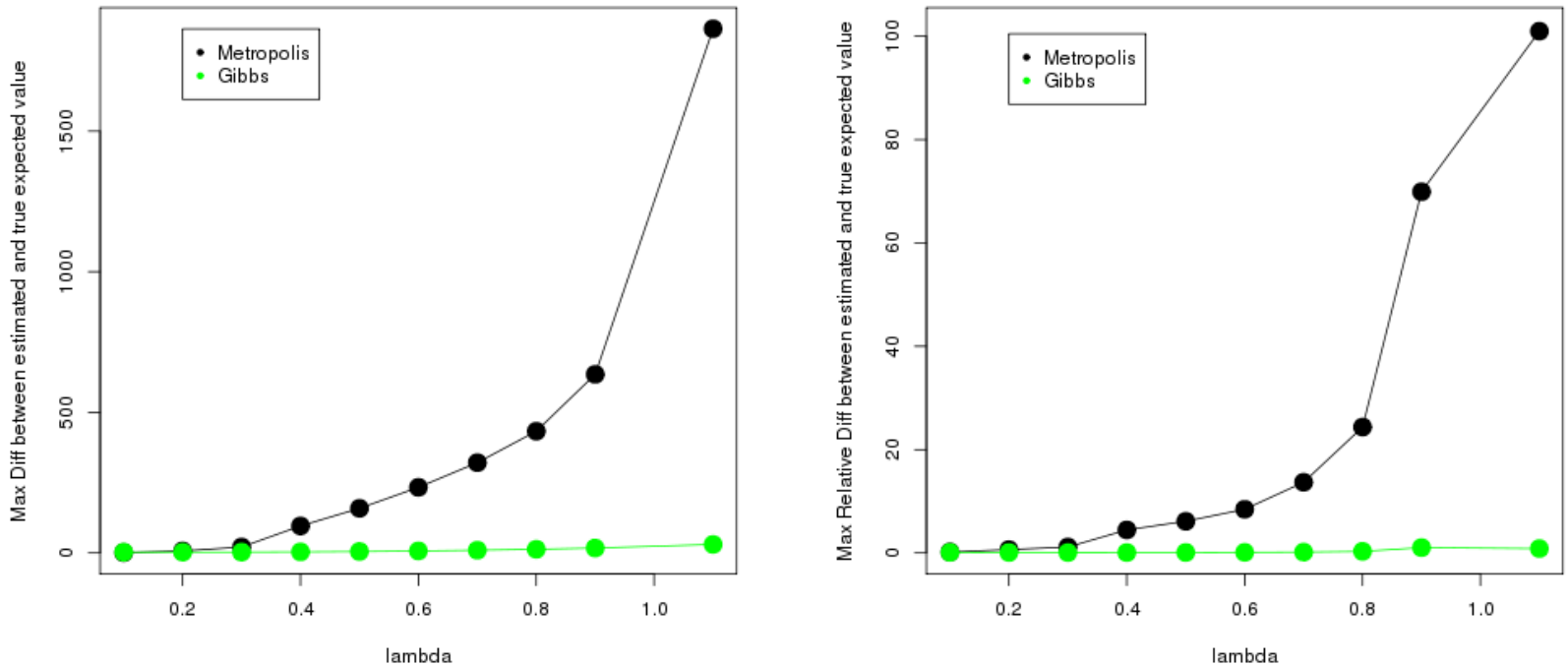


Figure: Plots comparing the maximum entry wise difference and the maximum entry wise relative difference between the estimated and true expected values for the Metropolis-Hastings algorithm versus the Gibbs sampling algorithm.

Model selection example: Goal

- Our Bayesian framework can also be used for efficient model selection in conjunction with penalized methods such as *Glasso*
- *Glasso* (or any other penalized ℓ_1 approach) has a penalty parameter ρ which controls the level of sparsity in the resulting graph and needs to be chosen
- On method to select ρ is provided in Banerjee et al. (2007), can also use cross-validation
- The idea is to consider a grid of ρ (penalty parameter) values, and use Bayesian methods to choose the "best" value

Model selection example: Setup

- For given p, a "true" sparse graph $G = (V, E)$ with p vertices is chosen by taking a WOR sample of size $\frac{p*(p-1)}{2} * 0.01$ from the total number of possible edges.
- Generate a "true" precision matrix $\Omega_0 \in \mathbb{P}_G$, and n i.i.d samples from $N(0, \Omega_0^{-1})$
- Let S denote sample covariance matrix
- A grid of 50 values of ρ ranging from 1 to 0.01 is considered
- Let $G_1, G_2, \cdots, G_{50}$ represent the corresponding graphs

Model selection example: Deviance Information Criterion

- Use the Deviance Information Criterion (DIC) as a measure of how well a given graph/model fits the data
- $DIC = 2\bar{D} - D(\bar{\Omega})$, where $D(\Omega) = n * (tr(\Omega S) - \log(|\Omega|))$
- $\bar{D}$ is the posterior expectation of $D(\Omega)$, and $\bar{\Omega}$ is the posterior expectation of Ω
- For every graph $G_1, G_2, \cdots, G_{50}$, use Generalized $\mathcal{G}$-Wishart prior with $U = \frac{n}{p}\text{tr}(S)I_p$ and $\delta_j = (U_{jj} + nS_{jj})/S_{jj}$ for $1 \leqslant j \leqslant p$
- Posterior expectations can be computed using Gibbs sampling
- Choose graph with best DIC score

Model selection example: Results

p	n	Specificity		Sensitivity	
		Glasso-Ban	gen. G-Wishart	Glasso-Ban	gen. G-Wishart
50	100	1	0.9830	0.5833	1
100	100	1	0.9714	0	0.8663
200	100	1	0.9316	0.0007538	0.7781
500	200	0.9999	0.9166	0.0041	0.5570
1000	300	0.9899	0.9214	0.0023	0.2772

Table: Model selection comparison of *Glasso* (with penalty parameter chosen by Banerjee et al. (2007)) and generalized G-Wishart based Bayesian approach

Questions?