

Inference in Gaussian covariance graph models

Bala Rajaratnam
Stanford University

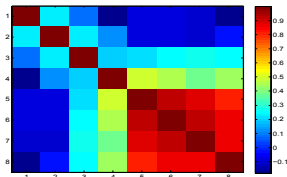
O-Bayes09, International Workshop on Objective Bayes Methodology

(Joint work with Kshitij Khare, Stanford University)

Discovering complex multivariate dependencies: An example

Gene expression data (yeast) for 8 genes involved in galactose utilization
 (Gasch et al. (2000))

$$R = \begin{bmatrix} 1.00 & 0.24 & 0.08 & -0.18 & -0.07 & -0.08 & -0.10 & -0.18 \\ 0.24 & 1.00 & 0.23 & 0.12 & -0.08 & -0.07 & -0.10 & -0.03 \\ 0.08 & 0.23 & 1.00 & 0.20 & 0.21 & 0.26 & 0.28 & 0.26 \\ -0.18 & 0.12 & 0.20 & 1.00 & 0.50 & 0.46 & 0.39 & 0.44 \\ -0.07 & -0.08 & 0.21 & 0.50 & 1.00 & 0.91 & 0.88 & 0.81 \\ -0.08 & -0.07 & 0.26 & 0.46 & 0.91 & 1.00 & 0.92 & 0.87 \\ -0.10 & -0.10 & 0.28 & 0.39 & 0.88 & 0.92 & 1.00 & 0.87 \\ -0.18 & -0.03 & 0.26 & 0.44 & 0.81 & 0.87 & 0.87 & 1.00 \end{bmatrix}$$



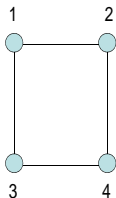
- In many real-life applications p large, n relatively small, $\binom{p}{2}$ correlations
- Task: Estimate covariance matrix $\Sigma \in \mathbb{P}^+$
- $\hat{\Sigma}_{MLE} = S$ is often a poor estimator unless $n \gg p$
- Goal: Make most of relatively small sample size
- Tools: Sparsity, Regularization, Graphs.

Literature review

- Extensive literature
- **Bayesian approach:** Daniels and Kass (1999, 2001); Daniels and Pourahmadi (2002); Consonni and Veronese (2003); Sun and Berger (2007); Letac and Massam (2007); Rajaratnam et al. (2008) and many others
- Tools: Priors on the eigenvalues, eigenvectors, correlation coefficients, conditional covariances, elements of the Cholesky decomposition, Givens angles
- **Frequentist approach:** Ledoit and Wolf (2004); Huang et al. (2006); Meinshausen and Bühlmann (2006); Bickel and Levina (2007, 2008); Dahl et al. (2007); Yuan and Lin (2007); Friedman, Hastie and Tibshirani (2007); Chaudhuri et al (2007) and many others
- Tools: Minimization of a loss function, L1 penalty, sparsity, regularization, minimization of a loss function, shrinkage, iterative algorithms

Sparse estimation through graphical models

- Tools to discover structure in high-dimensional data
- Undirected Graph $G = (V, E)$, V set of p vertices, E set of edges



$$V = \{1,2,3,4\}$$
$$E = \{(1,2), (1,3), (2,4), (3,4)\}$$

- \mathbb{P}^+ set of positive definite matrices; P_G set of positive definite matrices with zero restrictions according to G

$$P_G := \{\Sigma : \Sigma \in \mathbb{P}^+, \Sigma_{ij} = 0 \text{ if } (i, j) \notin E\}$$

Concentration graph models

- Induce sparsity or zeros in the precision or concentration matrix (inverse covariance matrix) - introduced by Dempster(1972)
- In Gaussian case, this leads to conditional independencies, i.e., if $\mathbf{X} = (X_1, X_2, \dots, X_p)' \sim \mathcal{N}(0, \Sigma)$, then

$$X_i \perp X_j \mid \mathbf{X}_{rest} \Leftrightarrow (\Sigma^{-1})_{ij} = 0$$

- Given a graph G , the Gaussian concentration graph model corresponding to G is the set of probability distributions

$$\mathcal{N}_G^* := \{\mathcal{N}(0, \Sigma) : \Sigma^{-1} \in P_G\}$$

- Graph G encodes pairwise conditional independencies
- Gives rise to regular exponential family models
- Frequentist inference: Edwards (2000), Lauritzen (1996), Whittaker (1991)
- Bayesian inference: Dawid and Lauritzen (1993), Letac and Massam (2007), Rajaratnam et al. (2008)

Covariance graph models

- Induce sparsity or zeros in the covariance matrix - introduced by Cox and Wermuth (1996)
- In Gaussian case, this leads to marginal independencies, i.e., if $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then

$$X_i \perp X_j \Leftrightarrow \Sigma_{ij} = 0$$

- Given a graph G , the Gaussian covariance graph model corresponding to this graph is the set of probability distributions

$$\mathcal{N}_G := \{\mathcal{N}(\mathbf{0}, \Sigma) : \Sigma \in P_G\}$$

- Graph G encodes marginal independencies
- Not a regular exponential family model

A simple Gaussian model

- Consider $\mathbf{X} \sim \mathcal{N}_4(0, \Sigma_G)$ $\mathbf{X} = (X_1, X_2, X_3, X_4)'$
- Independencies: $X_1 \perp X_3, X_2 \perp X_4, X_3 \perp X_4 \Rightarrow \sigma_{13} = \sigma_{24} = \sigma_{34} = 0$.
- Symbolically,

$$\Sigma_G = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & 0 \\ 0 & \sigma_{23} & \sigma_{33} & 0 \\ \sigma_{14} & 0 & 0 & \sigma_{44} \end{bmatrix}$$

- Graphical representation: $\overset{4}{\bullet} - \overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet}$

Maximum likelihood estimation under restricted model

- Given data $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ i.i.d. $\mathcal{N}(0, \Sigma_G)$, how to evaluate $\hat{\Sigma}_{MLE} \in P_G$ which maximizes the log-likelihood

$$l(\Sigma) = \text{const} - \frac{n}{2} \log |\Sigma| - \frac{n}{2} \text{tr}(\Sigma^{-1} S)$$

- More difficult problem; Curved exponential family; Multiple modes
- Sufficient condition for existence of MLE: $n > p$
- Necessary condition for certain Gaussian covariance graph models
- Kauermann (1996), Wermuth et al. (2006), Chaudhuri et al (2007)
- For full model, i.e., no zero restrictions on Σ , $\text{MLE} = S = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i \mathbf{Y}_i^T$
- Natural exponential family; Canonical parameter $\Omega = \Sigma^{-1}$
- Bayesian inference in full model: Traditional choice of conjugate prior Inverse Wishart distribution (Diaconis-Ylvisaker prior)

Bayesian estimation under restricted model \mathcal{N}_G

- Given graph G , consider the restricted model

$$\mathcal{N}_G = \{\mathcal{N}(0, \Sigma) : \Sigma \in P_G\}$$

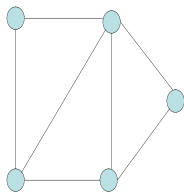
- The model \mathcal{N}_G gives rise to a curved exponential family (Efron, 1975, 1978; Amari, 1982)
- The canonical parameter $\Omega = \Sigma^{-1}$ lies in a complicated lower-dimensional space
- Can no longer use standard DY approach for regular exponential families
- Want a class of priors defined on the space of covariance matrices with fixed zeros.
- Need a new methodology

Desirable properties from class of priors

- Priors should facilitate inference in high dimensions
- Possible to sample from posterior distribution using a simple mechanism
- Possible to compute posterior mean/mode in closed form
- Flexibility (many shape parameters)
- Objective Bayesian inference (O-Bayes!), more robust priors
- Posterior linearity
- Too ambitious in the curved setting ???

Tools: Decomposable graphs

- A graph G is **decomposable** if it does not contain a chordless cycle of length ≥ 4
- Useful subclass of graphs



(a) Decomposable Graph



(b) Non-decomposable Graph

Why decomposable graphs?

If $G = (V, E)$ is decomposable and $\Sigma \in P_G$, then \exists an ordering of the vertices such that if $\Sigma = LDL^T$ is the modified Choleski decomposition, then

$$L_{ij} = 0, \forall i > j, (i, j) \notin E$$

- Zeroes in Σ reflected in L
- Existence of such an ordering characterizes decomposable graphs
- Any ordering will not work
- Focus attention on G decomposable, and vertices (variables) are ordered so that zeroes in Σ are reflected in L
- Define

$$\mathcal{L}_G = \{L : L_{ij} = 0 \forall i < j, \text{ or } (i, j) \notin E, \text{ and } L_{ii} = 1, \forall 1 \leq i, j \leq m\}$$

Wishart distributions for covariance graph models

- Define θ_G (the modified Choleski space) by

$$\theta_G := \{(L, D) : L \in \mathcal{L}_G, D \text{ diagonal with } D_{ii} > 0 \forall 1 \leq i \leq m\}$$

- Consider the class of distributions on θ_G with density
 (w.r.t. $\prod_{i>j, (i,j) \in E} dL_{ij} \prod_{i=1}^m dD_{ii}$)

$$\pi_{U, \alpha}(L, D) \propto e^{-\frac{(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^p \alpha_i \log D_{ii})}{2}} \quad \forall (L, D) \in \theta_G$$

- New class of Wishart distributions
- Scale parameter U (positive definite)
- Shape parameter $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)$ (with non-negative entries)

Properties of our class: Conjugacy

- Suppose we observe data $(\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n)$ under $\mathcal{N}_G = \{\mathcal{N}(\mathbf{0}, \Sigma) : \Sigma \in P_G\}$

- If prior is $\pi_{U,\alpha}$, posterior distribution of (L, D) is given by

$$\pi_{U,\alpha}(L, D \mid \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n) \propto e^{-\frac{\text{tr}((LDL^T)^{-1}(nS+U)) + \sum_{i=1}^p (n+\alpha_i) \log D_{ii}}{2}} \quad \forall (L, D) \in \theta_G$$

Hence the posterior distribution belongs to the same family i.e.

$$\pi_{U,\alpha}(\cdot \mid \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n) = \pi_{nS+U, (n+\alpha_1, n+\alpha_2, \dots, n+\alpha_p)}(\cdot)$$

- Our class of distributions form a **conjugate** family of priors for Gaussian covariance graph models

Properties of our class: Normalizing constant

- Density specified only up to normalizing constant
- When is normalizing constant finite?
- Define $\mathcal{N}^{\prec}(i) := \{j : (i, j) \in E, i > j\}$
- $\mathcal{N}^{\prec}(i)$ is the number of neighbors less than i
- Sufficient condition:

Proposition

$$\int_{\theta_G} e^{-\frac{\text{tr}((LDL^T)^{-1}U) + \sum_{j=1}^p \alpha_j \log D_{jj}}{2}} \prod_{(i,j) \in E, i > j} dL_{ij} \prod_{i=1}^p dD_{ii} < \infty$$

if

$$\alpha_i > |\mathcal{N}^{\prec}(i)| + 2 \quad \forall i = 1, 2, \dots, p.$$

Sampling from the posterior distribution

- Goal: To sample from posterior distribution
- Tool: Block Gibbs sampler
- Partition (L, D) into blocks so that the conditional distribution of each block given the others are standard and easy to sample from
- Define

$$L_{\cdot v}^G := (L_{uv})_{u > v, (u,v) \in E}, \quad v = 1, 2, \dots, p-1$$

- (L, D) can be partitioned as $(L_{\cdot 1}^G, L_{\cdot 2}^G, L_{\cdot 3}^G, \dots, L_{\cdot p-1}^G, D)$

Block Gibbs sampler

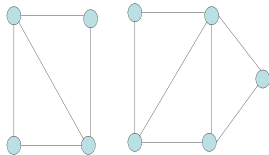
- Conditional distribution of $L_{\cdot v}^G$ given others is multivariate normal, $v = 1, 2, \dots, p - 1$
- Conditional distribution of D_{ii} given others is Inverse-Gamma
- Can be used to sample from $\pi_{U, \alpha}$ using the block Gibbs sampling algorithm
- Can explore posterior easily even in high-dimensions
- Can rigorously prove convergence of the block Gibbs sampler using sufficient conditions in Athreya et al. (1996)

Tools: Homogeneous graphs

- A graph G is **homogeneous** if it is decomposable and does not contain the 4-path, $\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet} - \overset{4}{\bullet}$, as an induced subgraph
- There exists an ordering of the vertices such that if $\Sigma \in \mathbb{P}^+$ and $\Sigma = LDL^T$ is the modified Choleski decomposition, then

$$\Sigma \in P_G \Leftrightarrow L \in \mathcal{L}_G \Leftrightarrow L^{-1} \in \mathcal{L}_G$$

- We demonstrate that \exists a constructive way to obtain such an ordering
- In essence, zeroes in $\Sigma \in P_G$ are reflected in $L \in \mathcal{L}_G$, $L^{-1} \in \mathcal{L}_G$



(a) Homogeneous
Graph

(b) Non-homogeneous
Graph

Homogeneous graphs: Covariance Hyper Markov and other properties

- Hyper Markov properties - concept introduced by Dawid and Lauritzen (1993) - conditional independence properties at the level of the priors

Theorem

If $(L, D) \in \pi_{U, \alpha}$ and $\Sigma = LDL^T$, then

$$D_{ii} = \Sigma_{i|pa(i)} \perp \Sigma_{pr(i)} \quad \forall 1 \leq i \leq p$$

- Establishes “strong directed covariance hyper Markov property” with respect to the given ordering of vertices for our class of priors
- Enables evaluation of $\mathbf{E}[\Sigma^{-1}]$, $\mathbf{E}[\Sigma]$, Laplace transform under $\pi_{U, \alpha}$ in closed form
- Facilitates decision theoretic estimation in covariance graph models
- Our priors satisfy a generalized notion of posterior linearity

Objective Bayes alternative: Reference priors

- Hyperparameter choice is not always clear
- Reference prior framework based on Berger and Bernardo (1992)
- Reference prior for the parameter $(L, D) \in \theta_G$ has density

$$\pi_R(L, D) \propto \frac{1}{|D|},$$

An improper density from our class with $U = 0$ and $\alpha = (2, 2, \dots, 2)$

- Posterior density has shape parameter $\alpha = (2 + n, 2 + n, \dots, 2 + n)$
- Posterior density using reference prior is proper if

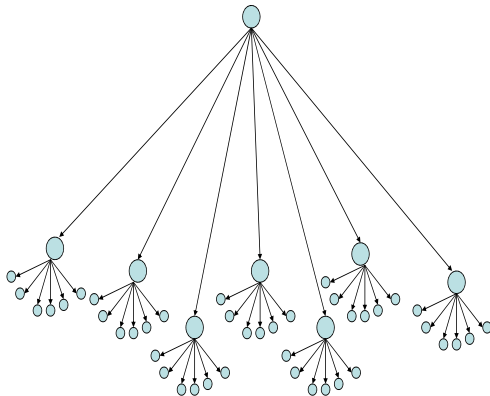
$$n > \max_{1 \leq i \leq p} |\mathcal{N}^{\setminus}(i)|,$$

i.e., sparse !!

- Challenge/Difficulty for G homogeneous: above does not hold if $n < p$

An example

- G homogeneous with 50 vertices
- Hasse diagram



An example

- Simulate $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ from $\mathcal{N}_{50}(\mathbf{0}, \Sigma)$ with known Σ
- Use prior $U = 0, \alpha$ with $\alpha_i = 2|\mathcal{N}^{\prec}(i)| + 5$, $i = 1, 2, \dots, 50$
- Can compute posterior mean $\Sigma_{mean} := \mathbf{E}[\Sigma | \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n]$ in closed form
- Use Gibbs sampling procedure
- Use burn-in of 1000 iterations, and calculate mean estimate by averaging over next 1000 iterations (Run using R, time taken 139.77 seconds). Relative error of estimate

$$\frac{\|\hat{\Sigma} - \Sigma_{mean}\|_2}{\|\Sigma_{mean}\|_2} \leq 0.018$$

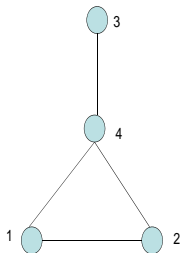
Choice of hyperparameters

- When substantial prior information available: subjective Bayesian
- When partial prior information available:
 - $U_{tr} = \frac{\text{tr}(S)}{p} I_p$
 - $\alpha^1(c)$ with $\alpha_i^1 = c, \forall 1 \leq i \leq p$
 - $\alpha^2(c)$ with $\alpha_i^2 = 2|\mathcal{N}^{\leftarrow}(i)| + c, \forall 1 \leq i \leq p$ (related to posterior linearity)
 - Empirical Bayes procedure: Let the data decide. Available when G homogeneous
- When no prior information is available:
 - Objective Bayesian reference prior framework of Berger and Bernardo
 - In our model, this corresponds to an improper prior with $U = 0, \alpha = \alpha^1(2)$

Simulation example 1

- Graph G with 4 vertices: First a small dimensional example

- $\Sigma_G = \begin{bmatrix} 1 & 0.4 & 0 & 0.5 \\ 0.4 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0.7 \\ 0.5 & 0.5 & 0.7 & 2 \end{bmatrix}$



- 10000 simulations; For each simulation, generate n i.i.d. samples from $\mathcal{N}(0, \Sigma_G)$, and construct estimators
- Goal: Explore flexibility of our class of priors
- Mean squared error = $\mathbf{E} \left[(\hat{\Sigma} - \Sigma)^2 \right]$, (other losses are also investigated)

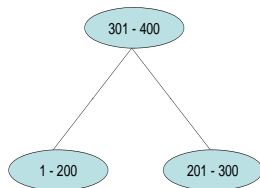
Simulation example 1

Sample size	5	7	10	20	25	30	40
MLE	50.103	5.425	3.065	1.416	1.102	0.909	0.686
$(U_{tr}, \alpha^1(5))$	13.296	6.553	3.740	1.569	1.2	0.972	0.712
$(U_{tr}, \alpha^2(5))$	4.231	3.256	2.434	1.292	1.028	0.860	0.668
$(U_{tr}, \alpha^1(8))$	3.312	2.646	2.05	1.164	0.946	0.801	0.652
$(U_{tr}, \alpha^2(8))$	3.137	2.568	2.024	1.160	0.944	0.803	0.619
$(U_{tr}, \alpha^1(11))$	3.220	2.624	2.05	1.165	0.950	0.808	0.626
$(U_{tr}, \alpha^2(11))$	3.644	2.968	2.309	1.282	1.036	0.876	0.671
Risk redc. vs. MLE	94 %	53 %	34 %	18 %	14 %	12 %	10 %

Table: Table of mean squared error for various estimators

Simulation Example 2

- Graph G with 400 vertices
- Σ known



Explore flexibility: Shrink variables $\{301, 302, \dots, 400\}$ more than others;
 $\alpha_i^*(\mathbf{c}) = |\mathcal{N}^{\prec}(i)| + \mathbf{c}$, $\forall 1 \leq i \leq 300$, $\alpha_i^*(\mathbf{c}) = 2|\mathcal{N}^{\prec}(i)| + \mathbf{c}$, $\forall 301 \leq i \leq 400$

Simulation Example 2

Sample size	300	410	850	1600
MLE	-	1619	517	272
$(U_{tr}, \alpha^*(10))$	1353	1010	499	268
$(I, \alpha^*(10))$	1589	1149	511	270
$(U_{tr}, \alpha^*(14))$	1336	979	475	262
$(I, \alpha^*(14))$	1436	1032	482	264
$(U_{tr}, \alpha^*(18))$	1764	1228	536	281
$(I, \alpha^*(18))$	1767	1218	536	282
Risk redc. vs. MLE	-	39 %	8 %	4 %

Table: Table of mean squared error for various estimators

Conclusion

Our class of priors provides a rich theoretical framework for Bayesian inference in covariance graph models

- Closed form Bayes estimation - useful in very high dimensions
- Flexibility (ρ shape parameters)
- Conditional distributions are easy to sample from; Block Gibbs sampling procedure can be used to sample from the posterior
- If G homogeneous,
 - Normalizing constant available in closed form, facilitates model selection
 - Covariance Hyper Markov properties
 - $\mathbf{E}[\Sigma^{-1}]$, $\mathbf{E}[\Sigma]$, Laplace transform under $\pi_{U,\alpha}$ available in closed form, facilitates closed form inference in decision theoretic framework
 - Generalized notion of posterior linearity for curved exponential families

Conclusion

- Bayesian framework allows for inference in the case when $n < p$, not possible in general in the frequentist framework
- Examples: Better estimation - Our priors can lead to substantial risk reduction
- Objective Bayesian Inference, other objective priors ??
- Choosing the graph: model selection (Covgnet algorithm)
- Choice of hyperparameters