Biological network inference using Bayesian graphical models

Christine B. Peterson

May 23, 2016
Introduction

Goal

- Learn biological networks from high-throughput data

Possible data sources

- Gene expression levels
- Protein or metabolite abundances
- Other continuous measurements

Statistical approach

- Gaussian graphical modeling in Bayesian framework
Challenge: understand protein interactions in cancer

**Acute myeloid leukemia (AML)**

- Rapid progression
- Five-year survival rates vary from 15–70% by subtype

**Key changes** in behavior of cancer cells

- Avoid normal process of cell death
- Uncontrolled cell division

Reflect alterations to **protein signaling networks**
AML data set

Protein quantifications

- Measured using reverse phase protein array (RPPA)
- 18 proteins involved in cell death and cell-cycle regulation

Subject population

- Newly diagnosed AML patients
- Sample sizes 17 – 68 per group for 4 subtypes

Goal: infer protein network for each subtype
Inference of multiple networks

Goal

- Infer networks for multiple sample groups

Existing approaches

- Separate inference for each subgroup
- Pooled inference
- Methods penalizing differences across groups

Proposed method

- Learn network using graphical model
- Use Bayesian priors to
  - Favor common edges in related networks
  - Learn which networks are related
Undirected graphical models

- Graph structure summarizes conditional independence
- Node = variable
- Missing edge = variables independent given all others

Fundamental assumption

- True conditional dependence structure sparse

Benefits

- Allows visualization of relationships
- Reduces overfitting
- Focuses on direct relationships
Gaussian graphical model

Likelihood

Multivariate normal constrained by graph $G$

$$p(x_i|\Omega) = \mathcal{N}(0, \Omega^{-1}), \quad i = 1, \ldots, n$$

- $x_i =$ vector of observed data for sample $i$
- Precision matrix $\Omega = \Sigma^{-1}$
- Entry $\omega_{ij} = 0$ if no edge between $i$ and $j$ in $G$
Gaussian graphical model

Goal of inference

- Infer sparse $\Omega$
- Pattern of zeros in $\Omega$ corresponds to graph
Graphical model inference

**Graphical lasso** [Friedman et al., 2008]

- Impose $L_1$ penalty on entries of $\Omega$
- Maximize penalized log-likelihood

$$\log \det(\Omega) - \text{trace}(S\Omega) - \rho \|\Omega\|_1$$

$S =$ sample covariance

$\|\Omega\|_1 =$ sum of absolute values of elements of $\Omega$

**Adaptive graphical lasso**

- Include a tailored weight for each $\omega_{ij}$ in penalty term
- Avoid overpenalization of large $\omega_{ij}$s
Achieve sparsity through choice of prior

**Advantages**

- Informative priors
- Error control
- Integration into hierarchical model

**Disadvantages**

- Computational tractability
Analogue to graphical lasso

Background: Bayesian lasso [Park and Casella 2008]
- Double exponential priors on regression coefficients
- MAP under Bayesian lasso = lasso estimate

Extension to graphical models [Wang 2012]
- Double exponential priors on off-diagonal entries of $\Omega$
Bayesian graphical lasso

Prior formulation

\[ p(\Omega | \lambda) \propto \prod_{i<j} \left[ \frac{\lambda}{2} \exp \left\{ -\lambda |\omega_{ij}| \right\} \right] \prod_{i=1}^{p} \left[ \frac{-\lambda}{2} \exp \left\{ \frac{\lambda}{2} \omega_{ii} \right\} \right] I[\Omega \in M^+] \]

- Restricted to space of symmetric positive definite matrices
- Prior for off-diagonal entries \( \omega_{ij} \) spiked around 0
- Sharpness of spike determined by shrinkage parameter \( \lambda \)
- Adaptive version: different \( \lambda_{ij} \)'s for different entries in \( \Omega \)
Conjugate prior

Background: Wishart distribution
- Multivariate generalization of $\chi^2$/gamma distributions
- Conjugate for precision matrix in multivariate normal

G-Wishart distribution [Roverato 2002]
- Restrict domain of Wishart given constraints of graph structure
Prior formulation

\[ p(\Omega \mid G) \propto |\Omega|^{(b-2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(D\Omega) \right\} I[\Omega \in M^+(G)] \]

Advantages

- Bayesian graphical lasso shrinks entries in precision matrix
- \( G \)-Wishart restricted to matrices with 0s corresponding to \( G \)
  i.e. \( \omega_{ij} = 0 \) whenever edge \((i, j)\) not included in \( G \)
- More difficult computationally, but clearer interpretation
Multiple network inference

Graphical model for each sample group over common set of variables

Group 1

Group 2

Group 3
Proposed model

\[
\text{posterior} \propto \text{likelihood} \times \text{prior}
\]
Proposed model

posterior $\propto$ likelihood $\times$ prior

$$\propto \prod_{k=1}^{K} p(X_k | \Omega_k) \times \cdots$$

**Gaussian graphical model** specific to group $k$

$$x_{k,i} \sim \mathcal{N}(0, \Omega_k^{-1}), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K$$

- $x_{k,i} = \text{observed data for sample } i \text{ in group } k$
Proposed model

posterior $\propto$ likelihood $\times$ prior

$$\propto \prod_{k=1}^{K} p(X_k|\Omega_k) \times \cdots$$

Components of prior

- Prior encouraging sparsity of $\Omega_k$s
- Prior favoring shared graph structure across related sample groups
- Spike-and-slab prior on parameters measuring network relatedness
- Edge-specific informative prior
Proposed model

posterior \propto \text{likelihood} \times \text{prior}

\propto \prod_{k=1}^{K} \left[ p(X_k|\Omega_k)p(\Omega_k|G_k) \right] \times \cdots

**Goal:** Sparse version of precision matrices

**Solution:** $G$-Wishart prior on $\Omega_k$
How to link inference of graphs?

Supergraph relating groups

- Represented by $K \times K$ matrix $\Theta$
- Nonzero entries indicate network relatedness
- Magnitude of entries measures pairwise graph similarity
Prior linking graph structures across groups

posterior $\propto$ likelihood $\times$ prior

$$
\propto \prod_{k=1}^{K} \left[ p(\mathbf{X}_k|\Omega_k)p(\Omega_k|G_k) \right] \prod_{i<j} \left[ p(\mathbf{g}_{ij}|\nu_{ij}, \Theta) \right] \times \cdots
$$

Goal: Encourage selection of common edges in related subgroups

Solution: Markov random field prior

$$
p(\mathbf{g}_{ij}|\nu_{ij}, \Theta) \propto \exp(\nu_{ij}\mathbf{1}^t\mathbf{g}_{ij} + \mathbf{g}_{ij}'\Theta\mathbf{g}_{ij})
$$

- $\mathbf{g}_{ij} =$ binary indicators of inclusion of edge $(i, j)$ across $K$ groups
- Larger values of $\theta_{km}$ encourage common edges between $G_k$ and $G_m$
Prior on edge-specific parameters

\[ \text{posterior} \propto \text{likelihood} \times \text{prior} \]

\[
\propto \prod_{k=1}^{K} \left[ p(X_k | \Omega_k) p(\Omega_k | G_k) \right] \prod_{i<j} \left[ p(g_{ij} | \nu_{ij}, \Theta) p(\nu_{ij}) \right] \times \cdots
\]

**Goal:** Encourage graph sparsity + incorporate reference information

**Solution:**
- Common prior with small mean encourages overall sparsity
- Informative prior specific to \( \nu_{ij} \) incorporates reference information on particular edge
Selection prior on network similarity

**Goal:** Sparse version of $\Theta$

**Solution:** Spike-and-slab mixture prior

$$p(\theta_{km}|\gamma_{km}) = (1 - \gamma_{km}) \cdot \delta_0 + \gamma_{km} \cdot \text{Gamma}(\alpha, \beta)$$

- $\gamma_{km} =$ indicator of network relatedness
- For non-related graphs, Dirac delta “spike” forces $\theta_{km}$ to be 0
- For related graphs, $\theta_{km}$ comes from Gamma “slab”
Selection prior on network similarity

\[ \text{posterior} \propto \text{likelihood} \times \text{prior} \]

\[
\propto \prod_{k=1}^{K} \left[ p(X_k|\Omega_k)p(\Omega_k|G_k) \right] \prod_{i<j} \left[ p(g_{ij}|\nu_{ij}, \Theta)p(\nu_{ij}) \right] p(\Theta|\gamma) \times \cdots
\]

**Goal:** Sparse version of \( \Theta \)

**Solution:** Spike-and-slab mixture prior

\[
p(\theta_{km}|\gamma_{km}) = (1 - \gamma_{km}) \cdot \delta_0 + \gamma_{km} \cdot \text{Gamma}(\alpha, \beta)
\]

- \( \gamma_{km} = \) indicator of network relatedness
- For non-related graphs, Dirac delta “spike” forces \( \theta_{km} \) to be 0
- For related graphs, \( \theta_{km} \) comes from Gamma “slab”
Selection prior on network similarity

posterior $\propto$ likelihood $\times$ prior

$$\propto \prod_{k=1}^{K} \left[ p(X_k|\Omega_k)p(\Omega_k|G_k) \right] \prod_{i<j} \left[ p(g_{ij}|\nu_{ij}, \Theta)p(\nu_{ij}) \right] p(\Theta|\gamma) \times \cdots$$

**Goal:** Sparse version of $\Theta$

**Solution:** Spike-and-slab mixture prior

$$p(\theta_{km}|\gamma_{km}) = (1 - \gamma_{km}) \cdot \delta_0 + \gamma_{km} \cdot \text{Gamma}(\alpha, \beta)$$

- $\gamma_{km} = \text{indicator of network relatedness}$
- For non-related graphs, Dirac delta “spike” forces $\theta_{km}$ to be 0
- For related graphs, $\theta_{km}$ comes from Gamma “slab”
Selection prior on network similarity

posterior $\propto$ likelihood $\times$ prior

$$\propto \prod_{k=1}^{K} \left[ p(X_k | \Omega_k) p(\Omega_k | G_k) \right] \prod_{i<j} \left[ p(g_{ij} | \nu_{ij}, \Theta)p(\nu_{ij}) \right] p(\Theta | \gamma)p(\gamma)$$

**Goal:** Sparse version of $\Theta$

**Solution:** Spike-and-slab mixture prior

$$p(\theta_{km} | \gamma_{km}) = (1 - \gamma_{km}) \cdot \delta_0 + \gamma_{km} \cdot \text{Gamma}(\alpha, \beta)$$

- $\gamma_{km}$ = indicator of network relatedness
- For non-related graphs, Dirac delta “spike” forces $\theta_{km}$ to be 0
- For related graphs, $\theta_{km}$ comes from Gamma “slab”
- Independent Bernoulli priors on latent indicators
Posterior inference

\textbf{posterior} \propto \text{likelihood} \times \text{prior}

\begin{equation*}
\propto \prod_{k=1}^{K} \left[ p(\mathbf{X}_k | \Omega_k)p(\Omega_k | G_k) \right] \prod_{i<j} \left[ p(g_{ij} | \nu_{ij}, \Theta)p(\nu_{ij}) \right] p(\Theta | \gamma)p(\gamma)
\end{equation*}

\textbf{Goal}: Obtain posterior sample of parameters

\textbf{MCMC}: At each iteration, sample parameters from full conditionals
posterior $\propto$ likelihood $\times$ prior

$$
\propto \prod_{k=1}^{K} \left[ p(X_k | \Omega_k) p(\Omega_k | G_k) \right] \prod_{i<j} \left[ p(g_{ij} | \nu_{ij}, \Theta) p(\nu_{ij}) \right] p(\Theta | \gamma) p(\gamma)
$$

**Goal:** Obtain posterior sample of parameters

**MCMC:** At each iteration, sample parameters from full conditionals

1. Sample graph and precision matrix for each group

   **Challenge:** Prior normalizing constant intractable

   **Solution:** Double Metropolis-Hastings algorithm
Posterior inference

\[ \text{posterior} \propto \text{likelihood} \times \text{prior} \]

\[ \propto \prod_{k=1}^{K} \left[ p(X_k | \Omega_k)p(\Omega_k | G_k) \right] \prod_{i<j} p(g_{ij}|\nu_{ij}, \Theta)p(\nu_{ij}) \left[ p(\Theta | \gamma)p(\gamma) \right] \]

**Goal:** Obtain posterior sample of parameters

**MCMC:** At each iteration, sample parameters from full conditionals

2. Sample network relatedness parameters

  **Challenge:** Changing dimension of parameter space
  **Solution:** Between-model and within-model moves
Posterior inference

**posterior** $\propto$ likelihood $\times$ prior

$$\propto \prod_{k=1}^{K} \left[ p(X_k|\Omega_k)p(\Omega_k|G_k) \right] \prod_{i<j} \left[ p(g_{ij}|v_{ij}, \Theta)p(v_{ij}) \right] p(\Theta|\gamma)p(\gamma)$$

**Goal**: Obtain posterior sample of parameters

**MCMC**: At each iteration, sample parameters from full conditionals

3. Sample edge-specific parameters
   Standard Metropolis-Hastings step
Posterior inference

**Edge selection**
- Threshold marginal posterior probabilities of inclusion (PPI)

**Choice of thresholds**
- Median probability model: PPI > 0.5
- Control Bayesian false discovery rate (FDR)

$$FDR = \frac{\sum (1 - PPI) \times I(\text{discovery})}{\# \text{ discoveries}}$$
Performance on learning related graph structures

Simulation setup

- Three true precision matrices with $p = 20$
- Network for first group is an AR(2) model with 37 edges
- Add/remove 5 edges to get second graph, repeat to get third
- Generate 50 samples from multivariate normal for each group

Methods compared

- Fused and group graphical lasso [Danaher et al. 2014]
- Separate Bayesian estimation using G-Wishart prior
- Joint Bayesian method
Simulation results

Accuracy of graph structure learning measured using

- True positive rate (TPR)
- False positive rate (FPR)
- Matthews correlation coefficient (MCC)

<table>
<thead>
<tr>
<th>Method</th>
<th>TPR</th>
<th>FPR</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fused graphical lasso</td>
<td>0.93</td>
<td>0.52</td>
<td>0.33</td>
</tr>
<tr>
<td>Group graphical lasso</td>
<td>0.93</td>
<td>0.55</td>
<td>0.31</td>
</tr>
<tr>
<td>Separate Bayesian</td>
<td>0.52</td>
<td>0.010</td>
<td>0.65</td>
</tr>
<tr>
<td>Joint Bayesian</td>
<td><strong>0.58</strong></td>
<td><strong>0.008</strong></td>
<td><strong>0.70</strong></td>
</tr>
</tbody>
</table>

- Bayesian methods have good specificity, but lower sensitivity
- Joint method improves sensitivity over separate estimation
Performance across model sizes

ROC curve

Best performance from proposed joint Bayesian method
Inferred protein networks for four AML subtypes

- Networks for M0 and M1 have least shared structure
- Red edges = common to all four groups

Subtype M0

Subtype M1
Comments on multiple network modeling

**Network features of interest**
- Hub nodes, differential connections
- Degree of structural difference across subgroups

**Conclusions** on joint Bayesian model
- Increases sensitivity over current Bayesian approaches
- Improves overall performance of graph structure learning

[Peterson, Stingo, and Vannucci. *JASA*, 2015]
Additional graphical modeling work

Integrating reference information into network inference

- **Goal**: Infer cellular metabolic network under neuroinflammation
- **Challenge**: Limited sample size
- **Solution**: Construct Bayesian prior to favor known connections

[Peterson et al. *Statistics and Its Interface*, 2013]

Joint variable and graph selection

- **Goal**: Identify pathway-linked proteins relevant to cancer survival
- **Challenge**: Coordinated weak effects, network structure unknown
- **Solution**: Simultaneously learn network and relevant predictors

[Peterson, Stingo, and Vannucci. *Statistics in Medicine*, 2016]
Future directions and conclusions

Future directions

• Improve scaling
• Link edge values across groups rather than binary structure
• Extend to heavy-tailed/non-normal data
• Additional areas of application

Conclusions on Bayesian graphical modeling approaches

• Flexible framework for capturing structure of biological problems
• Allows control of error rates for edge selection