Bayesian Robust Learning in Chain Graph Models for Integrative Pharmacogenomics

Anindya Bhadra
www.stat.purdue.edu/~bhadra

Purdue University
Multi-platform genomic data can be naturally modeled using chain graphs.

The biological hierarchy (CNA $\rightarrow$ mRNA $\rightarrow$ proteins $\rightarrow$ drug responses) gives the directed edges between platforms.

Undirected edges determine the conditional independence structure within each platform in the Gaussian case.

Goal: To develop an inference procedure for chain graph models robust to an assumption of normality.

Joint work with Moumita Chakraborty and Min Jin Ha (MD Anderson) and Veera Baladandayuthapani (Michigan). Supported by NSF Grant DMS-2014371.
Figure: (a) Chain graph structure for CNA, mRNA, RPPA and drug layers, (b) Empirical density plot of MAPK1 CNA levels. The $H$-score defined in the text as a measure of non-normality is equal to 0.988 for MAPK1 CNA. (c) Normal q-q plot of data corresponding to MAPK1 CNA levels (d) $H$-scores across multi-platform genomic data and 20 drugs.
Non-normality can appear in multiple layers. Indeed, in our experience, this is the norm rather than the exception.

Inference of conditional independence structure assuming a GGM is erroneous.

Common approaches for single layer graphs:

- The nonparanormal (Liu et al., 2009, JMLR).
- Bayesian copula-based approaches (Pitt et al., 2006, Biometrika).
- Robust and alternative multivariate $t$ (Finegold and Drton, 2011 AoAs; 2014 BA).
- Bhadra et al. (2018, Biometrics).
A limitation of the nonparanormal/copula-based methods

- The nonparanormal and copula-based approaches assume that the data can be “transformed to normality.”

- Specifically, if \((Y_1, \ldots, Y_p)\) follows a nonparanormal, then there exist monotone \(f_1, \ldots, f_p\) such that \((f_1(Y_1), \ldots, f_p(Y_p))\) follows a multivariate Gaussian.

- A critical assumption for the identifiability of the nonparanormal is that the means and variances are preserved before and after transformation, i.e, \(E(Y_i) = E[f_i(Y_i)]\) and \(V(Y_i) = V[f_i(Y_i)]\) (Eq. (3), Liu et al., 2009, JMLR).

- We want to handle cases where these moments may not even exist (examples: horseshoe or \(t\) distributed marginals with low df).
The models of Finegold and Drton (2011 AoAS; 2014 BA)

- Basic model:
  \[(Y_1/d_1, \ldots, Y_p/d_p) \sim \mathcal{N}(0, \Sigma^{-1})\]

- Finegold and Drton (2011): \(d_1 = \cdots = d_p \sim \text{InvGamma}(\tau/2, \tau/2)\)

- Finegold and Drton (2014): \(d_i \overset{\text{ind}}{\sim} \text{InvGamma}(\tau/2, \tau/2)\).

The first case gives the usual multivariate \(t\) (after marginalizing out the shared latent variable), the second model was termed the “alternative” multivariate \(t\).

- Zeros in \(\Sigma^{-1}\) determine the conditional uncorrelatedness (resp., conditional independence in the Gaussian case).
The model of Bhadra et al. (2018, Biometrics)

- Unclear why a $t$ distributed marginal is appropriate for all margins as in Finegold and Drton.

- Bhadra et al. (2018) allow $d_i$ to be almost arbitrary non-negative random variables that can model both polynomially and exponentially decaying tails.

- The trouble is in interpreting zeros in $\Sigma^{-1}$. It signifies neither conditional independence (the Gaussian case) nor conditional uncorrelatedness (the $t$ case).

- The main result of Bhadra et al. is that

\[ \{\Sigma^{-1}\}_{i,j} = 0 \iff P(Y_i < 0 \mid Y_{-\{i,j\}}) = P(Y_i < 0 \mid Y_{-i}) \]

- Zero patterns in $\Sigma^{-1}$ determines the sign independence pattern.
Models for multi-layer data: the Gaussian chain graph case

- One way is to specify via layer-wise node-conditional regressions:

\[
X(l) | X[1:l-1] \sim N_{|T_l|}(\beta_l X[1:l-1], J_l^{-1}), \quad l = 2, \ldots, L,
\]

\[
X(1) \sim N_{q_1}(0, J_1^{-1}).
\]

- Non-zero entries in \( \beta_l \) and \( J_l \) encode directed and undirected edges respectively.

- \((u - v) \in E \) when the \((v, u)\)th entry in \( J_l \) equals zero for nodes \( u \) and \( v \) in the same layer \( l \).

- Similarly, \((u \rightarrow v) \in E \) when the \((v, u)\)th entry of \( \beta_l \) is zero, for \( \mathcal{L}(u) < \mathcal{L}(v) \) and \( \mathcal{L}(v) = l \).

- Examples: Ha et al. (2021, JASA), Lin et al. (2016, JMLR) and many others.
We apply the sign independence framework of Bhadra et al. (2018) to the chain graph model of Ha et al. (2021):

\[ D_l X(l) = B_l D_{[1:l-1]} X_{[1:l-1]} + \varepsilon_l, \quad \varepsilon_l \sim N_{|T_l|}(0, K_l^{-1}), \quad 2 \leq l \leq L, \]

\[ \varepsilon_1 = D_1 X_{(1)}, \quad \varepsilon_1 \sim N_{q_1}(0, K_1^{-1}), \]

where \( D_l \) is diagonal matrix of scale variables for the nodes in layer \( l \).
What are we able to infer?

Theorem 1

(i) (At least one node is non-normal). Conditional sign-independence follows from $B$ and $K$ as:

(a) ($u$ and $v$ in the same layer). Suppose $\mathcal{L}(u) = \mathcal{L}(v)$ and $\rho = k_{uv} = k_{vu}$. Then $\rho = 0$ if and only if $X_u \perp^s X_v | Z_u$, where $Z_u = X_{[1:\mathcal{L}(u)]} \setminus \{X_u, X_v\}$.

(b) ($u$ and $v$ in different layers). Suppose $\mathcal{L}(u) < \mathcal{L}(v)$ and $\rho = B_{vu}$. Then $\rho = 0$ if and only if $X_u \perp^s X_v | Z_d$, where $Z_d = X_{[1:\mathcal{L}(v)-1]} \setminus X_u$.

(ii) (Between normal nodes). $\rho = 0$ if and only if $X_u \perp X_v | Z_u$ for $\mathcal{L}(u) = \mathcal{L}(v)$ and $X_u \perp X_v | Z_d$ for $\mathcal{L}(u) < \mathcal{L}(v)$. 
The model for $D$

- A key benefit of the normal scale mixture framework as in Bhadra et al. (2018) is that it is possible to *reverse engineer* the mixing variables $d_i$ from a knowledge of the marginal tails of $y_i$.

- The main tool is a result of Barndorff-Nielsen et al. (1982) that says if $(y_i \mid d_i) \sim N(0, d_i)$ and marginally

  - *(Polynomial tails).* If $f(y_i) \propto |y_i|^{2\lambda_i-1}$, as $|y_i| \to \infty$, then $p(d_i) \propto d_i^{\lambda_i-1}$, as $d_i \to \infty$.

  - *(Exponential tails).* If $f(y_i) \propto |y_i|^{2\lambda_i-1} \exp\{- (2\psi_i)^{1/2} |y_i|\}$, as $|y_i| \to \infty$, then $p(d_i) \propto d_i^{\lambda_i-1} \exp(-\psi_i d_i)$, as $d_i \to \infty$. 
The model for $D$

- Our prior for $d_i$ in this paper is a mixture:

$$d_i \mid \pi_i \sim \omega_i p_i + (1 - \omega_i) \delta_1,$$

$$\omega_i \sim \text{Bernoulli}(\pi_i),$$

$$\pi_i \sim \text{Beta}(a_i, b_i).$$

- $\pi_i$ is the probability that node $i$ will be non-normal, its prior hyperparameters $a_i$ and $b_i$ are selected via $p$-values of KS test for normality for node $i$.

- But we also want to leave a non-zero probability for a node being normal. Hence the Dirac mass at 1.

- Conditional on a node being normal, the Barndorff-Nielsen result from the previous slide is used to select the hyperparameters for $p_i$. 
Inference under a mixture prior on $D$

- We are still able to infer conditional sign independence between two nodes where at least one is non normal.

- Similarly, we are able to infer conditional independence between two normal nodes.

- Except, now these conclusions are true with some probability determined by $\pi_i$. 
The priors on $B_l$ and $\mathcal{K}_l$

- Recall that the RCGM model is written as a sequence of partial regressions.
- $B_l$ is a matrix of regression coefficients connecting layer $l-1$ and $l$ and $\mathcal{K}_l$ is the precision matrix among the nodes in layer $l$.
- We use spike-and-slab priors for all, except for the diagonal terms in $\mathcal{K}_l$, which are assigned gamma priors.
- Inference proceeds via MCMC in the usual manner, details are in the supplement to the paper.
Numerical experiments

- We compare three methods:
  - RCGM
  - BANS (Ha et al., 2021, JASA): Performs Bayesian estimation in Gaussian chain graphs.
  - LBBM (Lin et al. 2016, JMLR): Performs $\ell_1$ penalized estimation in Gaussian chain graphs (lasso for $B_i$, glasso for $K_i$, proceed via ADMM).

- Caveat: Although RCGM and BANS take a Bayesian approach, they use the node-conditional/pseudo likelihoods rather than full likelihood for estimation (currently computationally very expensive).

- Metric for comparison: the performance in sign recovery, calculated via Hamming loss.
Numerical experiments

Figure: ROC curves for the simulation setting \((q, L, n, p_E) = (50, 4, 200, 0.08)\) across high, medium and low levels of non-normality \(\pi\), where \(q\), \(L\) and \(p_E\) denote the dimension of graph, number of layers and sparsity respectively. Panels (a) and (b) correspond to scaling by Exponential(mean = 2.5) and Inv-Gamma(shape = 3, rate = 6) respectively.
Pharmacogenomics in lung cancer

**Figure:** Sankey diagram showing connectivity between the 4 platforms across 10 pathways. Each box in the left three columns is a pathway-molecular platform combination, and widths of the lines between them are proportional to the number of directed edges connecting them. Gray lines denote edges between pathway-platform blocks and drugs.
Pharmacogenomics in lung cancer

**Figure:** The estimated multilayered network for DNA Damage Response pathway. Blue and red edges indicate positive and negative dependencies, while CD and CSD stand for conditionally dependent and conditionally sign-dependent edges respectively. The width of the edges is proportional to the posterior inclusion probabilities.
Main references

- Chakraborty, M., Baladandayuthapani, V., Bhadra, A. and Ha, M. J. (2022+). Bayesian Robust Learning in Chain Graph Models for Integrative Pharmacogenomics. (submitted). [arXiv:2111.11529]

- Bhadra, A., Rao, A. and Baladandayuthapani, V. (2018). Inferring network structure in non-normal and mixed discrete-continuous genomic data. *Biometrics, 74*, 185–195.

- Ha, M. J., Stingo, F. C. and Baladandayuthapani, V. (2021). Bayesian structure learning in multilayered genomic networks. *Journal of the American Statistical Association, 116*, 605–618.

- Lin, J., Basu, S., Banerjee, M. and Michailidis, G. (2016). Penalized maximum likelihood estimation of multi-layered Gaussian graphical models. *Journal of Machine Learning Research, 17(146)*, 1–51.