Sparse estimation of a two-level network for predicting breast cancer patients' treatment responses via quadratic discriminant analysis

Liang Shan

Division of Preventive Medicine

Department of Medicine

The University of Alabama at Birmingham

Feb 28, 2020

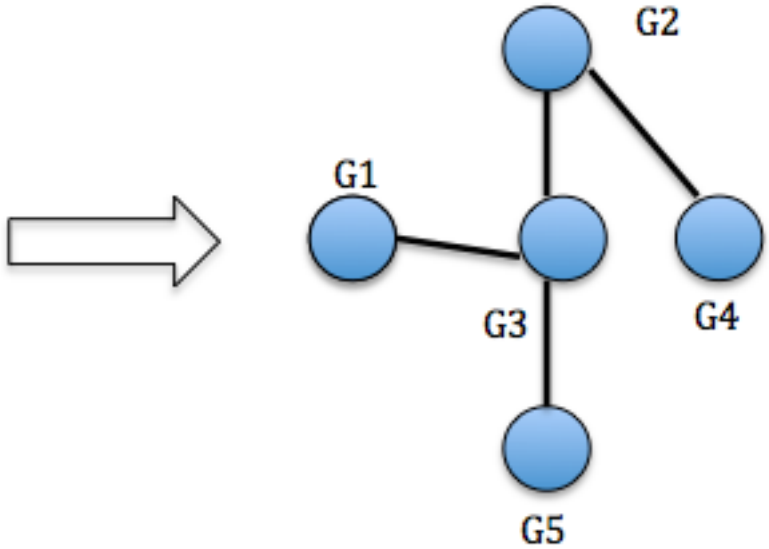
Outline

Background of Gaussian Graphical Model

- A graph with p nodes and edges between the p nodes could be represented by $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the nodes set and E is a subset of the set $V \times V$.
- Under multivariate Gaussian distribution, the elements of inverse covariance matrix (or precision matrix) Ω indicate conditional dependency between pairs of variables.
- Specifically, if $\omega_{i,j} = 0$, variables i and j are conditionally independent, otherwise, they are dependent given all other variables.
- That is to say, $\omega_{i,j} = 0 \Leftrightarrow (i,j)$ and $(j,i) \notin E$ and $\omega_{i,j} \neq 0 \Leftrightarrow (i,j)$ and $(j,i) \in E$
- Therefore, $\Omega \longrightarrow E \longrightarrow$ Graph (i.e. Network).

Gaussian Graphical Model and Gene Network

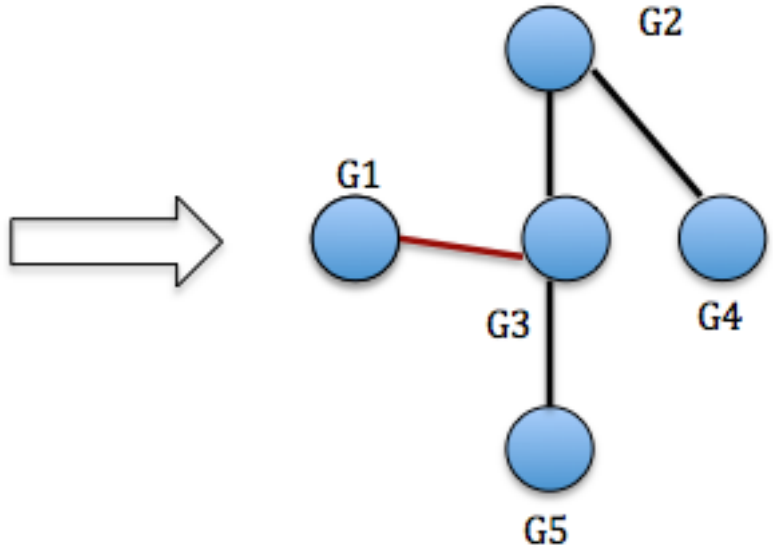
$$
\begin{array}{c} \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} G_1 & G_2 & G_3 & G_4 & G_5 \end{array} \\
\begin{pmatrix}
\omega_{11} & 0 & \omega_{13} & 0 & 0 \\
0 & \omega_{22} & \omega_{23} & \omega_{24} & 0 \\
\omega_{13} & \omega_{23} & \omega_{33} & 0 & \omega_{35} \\
0 & \omega_{24} & 0 & \omega_{44} & 0 \\
0 & 0 & \omega_{35} & 0 & \omega_{55}
\end{pmatrix}
\end{array}
$$



- $\omega_{ij} = 0$ means G_i and G_j are conditionally independent $\Leftrightarrow$ G_i and G_j are disconnected in the gene network
- $\omega_{ij} \neq 0$ means G_i and G_j are conditionally dependent $\Leftrightarrow$ G_i and G_j are connected in the gene network

Gaussian Graphical Model and Gene Network

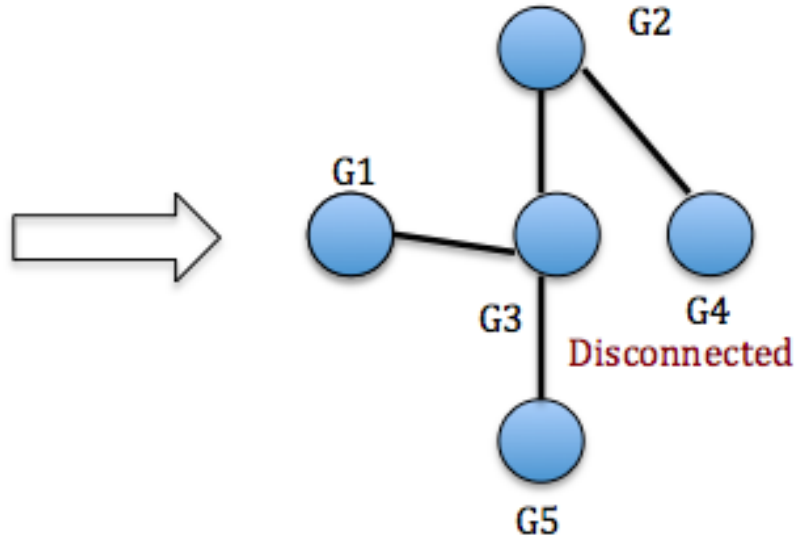
$$
\begin{array}{c} \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} G_1 & G_2 & G_3 & G_4 & G_5 \end{array} \\
\begin{pmatrix}
\omega_{11} & 0 & \omega_{13} & 0 & 0 \\
0 & \omega_{22} & \omega_{23} & \omega_{24} & 0 \\
\omega_{13} & \omega_{23} & \omega_{33} & 0 & \omega_{35} \\
0 & \omega_{24} & 0 & \omega_{44} & 0 \\
0 & 0 & \omega_{35} & 0 & \omega_{55}
\end{pmatrix}
\end{array}
$$



$\omega_{13} \neq 0 \Leftrightarrow G_1$ and G_3 are connected in the gene network

Gaussian Graphical Model and Gene Network

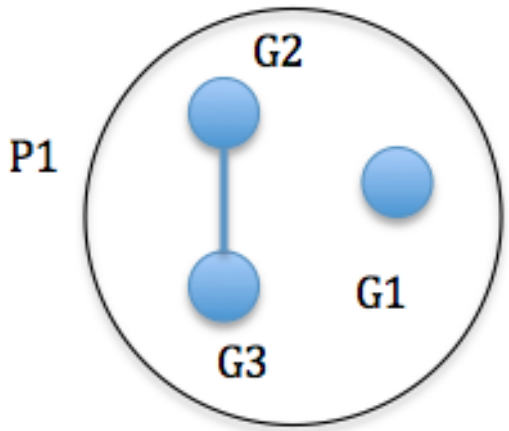
$$
\begin{array}{c} \\ G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} G_1 & G_2 & G_3 & G_4 & G_5 \end{array} \\
\begin{pmatrix}
\omega_{11} & 0 & \omega_{13} & 0 & 0 \\
0 & \omega_{22} & \omega_{23} & \omega_{24} & 0 \\
\omega_{13} & \omega_{23} & \omega_{33} & 0 & \omega_{35} \\
0 & \omega_{24} & 0 & \omega_{44} & 0 \\
0 & 0 & \omega_{35} & 0 & \omega_{55}
\end{pmatrix}
\end{array}
$$



$\omega_{45} = 0 \Leftrightarrow G_4$ and G_5 are disconnected in the gene network

Why two-level network analysis?

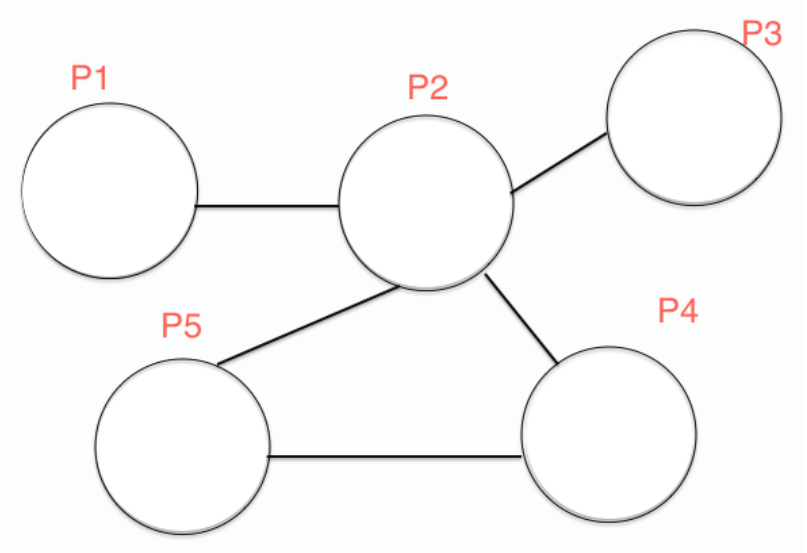
- Definition: the two-level network analysis is a network based analysis in terms of two layers, higher-level variables (e.g., pathways) and lower-level variables (e.g., genes).
- People like to know how lower level variables interact within a higher-level variable and how higher-level variables interact.



An Example of the Gene Network within a Pathway

Why two-level network analysis?

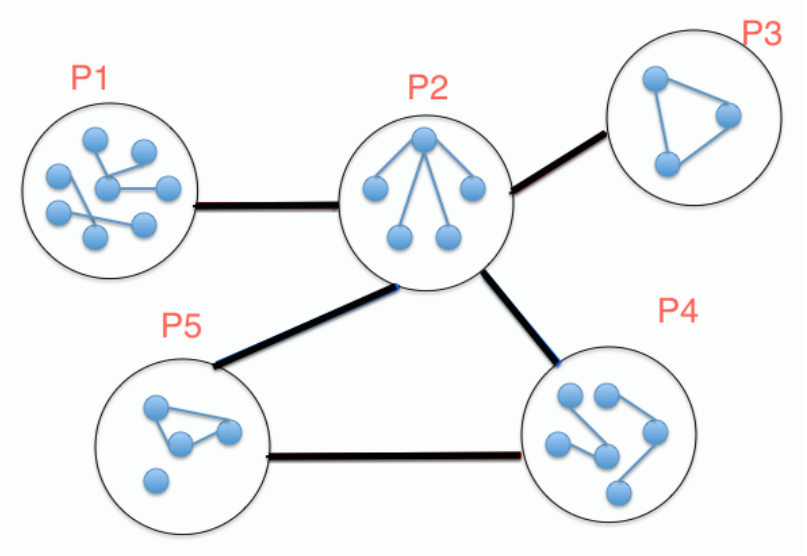
- Definition: the two-level network analysis is a network based analysis in terms of two layers, higher-level variables (e.g., pathways) and lower-level variables (e.g., genes).
- People like to know how lower level variables interact within a higher-level variable and how higher-level variables interact.



An Example of the Pathway Network

Why two-level network analysis?

- Definition: the two-level network analysis is a network based analysis in terms of two layers, higher-level variables (e.g., pathways) and lower-level variables (e.g., genes).
- People like to know how lower level variables interact within a higher-level variable and how higher-level variables interact.



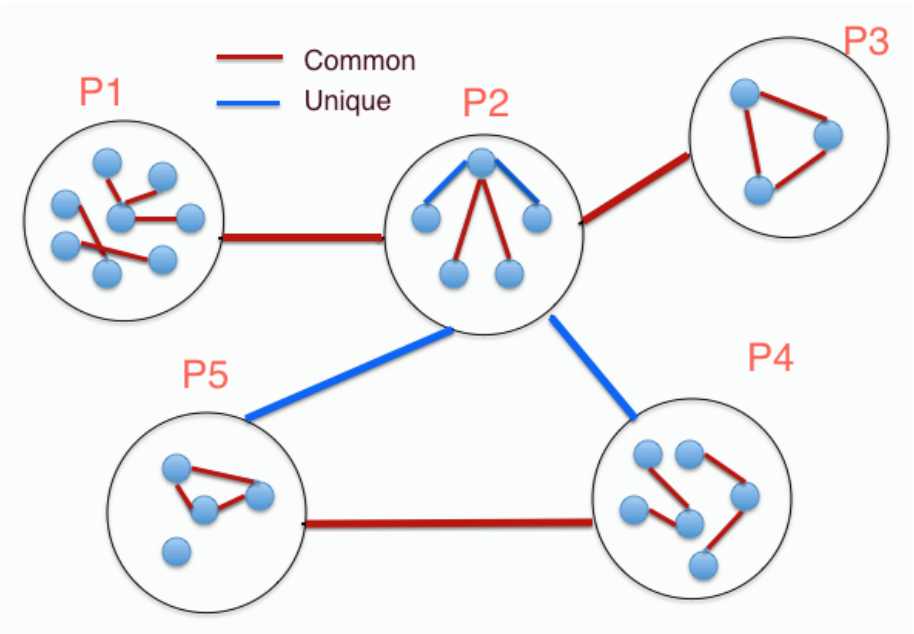
An Example of the Two-level Network

Why joint estimation?

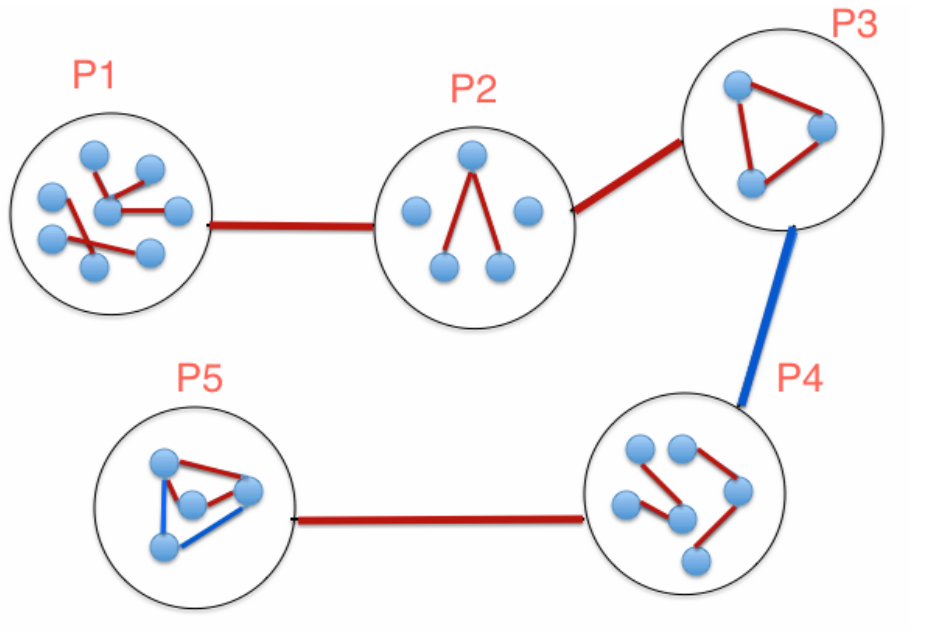
- Biological systems are highly dynamic; we usually face observations collected from different states. For instance, normal or diseased.
- If we estimate the two-level network separately, common structures are ignored. If we estimate them together, differences are masked.
- Many works have been proposed to jointly estimate multiple graphical models (Guo *et al.*, 2011; Danaher *et al.*, 2014; Shan and Kim, 2018; Shan *et al.*, 2019).
- There are also papers that focus on inferring the differential network directly (Yuan *et al.*, 2017).

Joint Estimation of the Two-level Network

- Common two-level network exists among heterogeneous classes.
- Yet differences exist among heterogeneous classes in terms of the defined two-level network.



(a) White Breast Cancer Patients



(b) Non-White Breast Cancer Patients

The Two-level Gaussian Graphical Model

Settings:

- The heterogeneous dataset has p variables and M classes ($M \geqslant 2$).
- The p variables are in K pre-specified groups, denoted by $P_1, \ldots, P_K$.
- There are p_k variables within the kth group P_k.
- The m^{th} class contains n_m observations $(\underline{x}_1^m, \ldots, \underline{x}_{n_m}^m)$, where $\underline{x}_i^m = (x_{i,1}^m, \ldots, x_{i,p}^m)$ $i = 1, \ldots, n_m$.
- We have M $n_m \times p$ matrices, where

$$X_{n_m \times p}^m = \begin{bmatrix} x_{1,1}^m & \cdots & x_{1,p}^m \\ \vdots & \ddots & \vdots \\ x_{n_m,1}^m & \cdots & x_{n_m,p}^m \end{bmatrix} = \begin{bmatrix} \underline{x}_1^m \\ \vdots \\ \underline{x}_{n_m}^m \end{bmatrix} \quad m = 1, \ldots, M$$

The Two-level Gaussian Graphical Model

Assumptions:

1. Within each class m, $\underline{x}_1^m, \ldots, \underline{x}_{n_m}^m \in \mathbb{R}^p$ are i.i.d MN $(\underline{0}, (\Omega^{(m)})^{-1})$, where

$$\Omega^{(m)} = \begin{bmatrix} \omega_{1,1}^{(m)} & \cdots & \omega_{1,p}^{(m)} \\ \vdots & \ddots & \vdots \\ \omega_{p,1}^{(m)} & \cdots & \omega_{p,p}^{(m)} \end{bmatrix}$$

   is symmetric positive definite.

2. Observations from different classes are independent of each other.

The Two-level Gaussian Graphical Model

- $\Omega_{kk}^{(m)}$ tells us information on the lower-level network.
- $\Omega_{kk'}^{(m)}$ tells us information on the higher-level network.

$$\Omega^{(m)} = \begin{bmatrix} \Omega_{11}^{(m)} & \Omega_{12}^{(m)} & \cdots & \Omega_{1K}^{(m)} \\ \Omega_{21}^{(m)} & \Omega_{22}^{(m)} & \cdots & \Omega_{2K}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{K1}^{(m)} & \Omega_{K2}^{(m)} & \cdots & \Omega_{KK}^{(m)} \end{bmatrix}$$

The Two-level Gaussian Graphical Model

Let the conditional dependency among variables in the kth and $k^{'}$th groups in class m be written as a p_k by $p_{k'}$ sub-block precision matrix $\Omega^{(m)}_{kk'}$, which is

$$\Omega^{(m)}_{kk'} = \begin{bmatrix} \omega^{kk'(m)}_{1,1} & \omega^{kk'(m)}_{1,2} & \cdots & \omega^{kk'(m)}_{1,p_{k'}} \\ \omega^{kk'(m)}_{2,1} & \omega^{kk'(m)}_{2,2} & \cdots & \omega^{kk'(m)}_{2,p_{k'}} \\ \vdots & \vdots & \ddots & \vdots \\ \omega^{kk'(m)}_{p_k,1} & \omega^{kk'(m)}_{p_k,2} & \cdots & \omega^{kk'(m)}_{p_k,p_{k'}} \end{bmatrix}$$

For instance, $\omega^{kk'(m)}_{1,2}$ indicates the conditional dependency between gene 1 in pathway k and gene 2 in pathway k' for class m.

The Two-level Gaussian Graphical Model

Reparameterization within a class:

- We reparameterize the sub-block precision matrix $\Omega_{kk'}^{(m)}$ by introducing a higher level factor $\theta_{kk'}^{(m)}$ for the kth and k'th groups in class m.
- The partial correlation between ith variable in the kth group and jth variable in the k'th group in class m can be written as $\omega_{i,j}^{kk'(m)} = \theta_{kk'}^{(m)} \gamma_{i,j}^{kk'(m)}$, where $\theta_{kk'}^{(m)} \geqslant 0$, $1 \leqslant k, k' \leqslant K$.
- The constraints are $\theta_{kk'}^{(m)} = \theta_{k'k}^{(m)}$, $\gamma_{i,j}^{kk'(m)} = \gamma_{j,i}^{k'k(m)}$, and $\theta_{kk}^{(m)} = 1$ (i.e. $\omega_{i,j}^{kk(m)} = \gamma_{i,j}^{kk(m)}$).

The Two-level Gaussian Graphical Model

Therefore, the precision matrix for class m is written as

$$\Omega^{(m)} = \begin{bmatrix} \Omega_{11}^{(m)} & \Omega_{12}^{(m)} & \cdots & \Omega_{1K}^{(m)} \\ \Omega_{21}^{(m)} & \Omega_{22}^{(m)} & \cdots & \Omega_{2K}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{K1}^{(m)} & \Omega_{K2}^{(m)} & \cdots & \Omega_{KK}^{(m)} \end{bmatrix}$$

$$= \begin{bmatrix} \theta_{11}^{(m)}\Gamma_{11}^{(m)} & \theta_{12}^{(m)}\Gamma_{12}^{(m)} & \cdots & \theta_{1K}^{(m)}\Gamma_{1K}^{(m)} \\ \theta_{21}^{(m)}\Gamma_{21}^{(m)} & \theta_{22}^{(m)}\Gamma_{22}^{(m)} & \cdots & \theta_{2K}^{(m)}\Gamma_{2K}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{K1}^{(m)}\Gamma_{K1}^{(m)} & \theta_{K2}^{(m)}\Gamma_{K2}^{(m)} & \cdots & \theta_{KK}^{(m)}\Gamma_{KK}^{(m)} \end{bmatrix}$$

The Joint Estimation Method

Reparameterization across classes:

- We assume the two-level networks across classes share some common structure, which is composed of the higher-level network and the lower-level network.
- That is to say, $\theta_{kk'}^{(m)} = \alpha_{kk'}\beta_{kk'}^{(m)}$ $(1 \leqslant k \neq k' \leqslant K,\ 1 \leqslant m \leqslant M)$ and $\gamma_{i,j}^{kk(m)} = \iota_{i,j}^{kk}\rho_{i,j}^{kk(m)}$ $(1 \leqslant k \leqslant K,\ 1 \leqslant i,j \leqslant p_k,\ 1 \leqslant m \leqslant M)$.
- The constraints for the first decomposition are $\alpha_{kk'} \geqslant 0$, $\alpha_{kk'} = \alpha_{k'k}$ and $\beta_{kk'}^{(m)} = \beta_{k'k}^{(m)}$ $(1 \leqslant k \neq k' \leqslant K,\ 1 \leqslant m \leqslant M)$,
- The constraints for the second are $\iota_{i,j}^{kk} \geqslant 0$, $\iota_{i,j}^{kk} = \iota_{j,i}^{kk}$ and $\rho_{i,j}^{kk(m)} = \rho_{j,i}^{kk(m)}$ $(1 \leqslant k \leqslant K,\ 1 \leqslant i \neq j \leqslant p_k,\ 1 \leqslant m \leqslant M)$ and $\iota_{i,i}^{kk} = 1$ (i.e. $\gamma_{i,i}^{kk(m)} = \rho_{i,i}^{kk(m)}$).

The Joint Estimation Method

By adding information of common structures into the model, we have $\Omega^{(m)}_{kk'} = \theta^{(m)}_{kk'}\Gamma^{(m)}_{kk'} = \alpha_{kk'}\beta^{(m)}_{kk'}\Gamma^{(m)}_{kk'}$ so that

$$\Omega^{(m)} = \begin{bmatrix} \theta^{(m)}_{11}\Gamma^{(m)}_{11} & \theta^{(m)}_{12}\Gamma^{(m)}_{12} & \cdots & \theta^{(m)}_{1K}\Gamma^{(m)}_{1K} \\ \theta^{(m)}_{21}\Gamma^{(m)}_{21} & \theta^{(m)}_{22}\Gamma^{(m)}_{22} & \cdots & \theta^{(m)}_{2K}\Gamma^{(m)}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta^{(m)}_{K1}\Gamma^{(m)}_{K1} & \theta^{(m)}_{K2}\Gamma^{(m)}_{K2} & \cdots & \theta^{(m)}_{KK}\Gamma^{(m)}_{KK} \end{bmatrix}$$

$$= \begin{bmatrix} \Gamma^{(m)}_{11} & \alpha_{12}\beta^{(m)}_{12}\Gamma^{(m)}_{12} & \cdots & \alpha_{1K}\beta^{(m)}_{1K}\Gamma^{(m)}_{1K} \\ \alpha_{21}\beta^{(m)}_{21}\Gamma^{(m)}_{21} & \Gamma^{(m)}_{22} & \cdots & \alpha_{2K}\beta^{(m)}_{2K}\Gamma^{(m)}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{K1}\beta^{(m)}_{K1}\Gamma^{(m)}_{K1} & \alpha_{K2}\beta^{(m)}_{K2}\Gamma^{(m)}_{K2} & \cdots & \Gamma^{(m)}_{KK} \end{bmatrix}$$

The Joint Estimation Method

Finally, our problem could be decomposed into M individual optimization problems at the $(t+1)$th iteration:

$$\min_{\{\Omega^{(m)}_{kk'},\Gamma^{(m)}_{kk}\}} \quad n_m[\mathrm{trace}(S^{(m)}\Omega^{(m)}) - \log|\Omega^{(m)}|] + \lambda_1 \sum_{k\neq k'} \frac{\sum_{\substack{1\leqslant i\leqslant p_k \\ 1\leqslant j\leqslant p_{k'}}} |\omega^{kk'(m)}_{i,j}|}{(\sum_{m=1}^{M} \sum_{\substack{1\leqslant i\leqslant p_k \\ 1\leqslant j\leqslant p_{k'}}} |\omega^{kk'(m)^{(t)}}_{i,j}|)^{1/2}} + \lambda_2 \sum_{k=1}^{K} \sum_{1\leqslant i\neq j\leqslant p_k} \frac{|\gamma^{kk(m)}_{i,j}|}{(\sum_{m=1}^{M} |\gamma^{kk(m)^{(t)}}_{i,j}|)^{1/2}} \quad (1)$$

The solution to (1) could be efficiently solved using the weighted Glasso algorithm by Friedman *et al.* (2008).

The Joint Estimation Method

Tuning parameters selection:
The tuning parameters λ_1 and λ_2 in (1) control the sparsity of the estimator. For now, we select them using the Bayesian Information Criterion (BIC), defined as:

$$\mathrm{BIC}(\lambda_1,\lambda_2) = \sum_{m=1}^{M}\{n_m[\mathrm{trace}(S^{(m)}\hat{\Omega}^{(m)}_{\lambda_1,\lambda_2}) - \log|\hat{\Omega}^{(m)}_{\lambda_1,\lambda_2}|] + df_m \log(n_m)\},$$

- $df_m = \#\{(i,j) : i < j, \hat{\omega}^{(m)}_{i,j} \neq 0\}$
- $\hat{\Omega}^{(m)}_{\lambda_1,\lambda_2}$ is the $\hat{\Omega}^{(m)}$ when we impose tuning parameters λ_1 and λ_2

We denote the joint estimation method for the two-level Gaussian graphical models as JMGGM.

Data

- The University of Texas M.D. Anderson Cancer Center
- Total number of samples: $n = 278$;
  - number of classes: $M = 2$;
  - sample sizes: $n_1 = 56$ (pCR) vs. $n_2 = 222$ (RD) or $n_1 = 114$ (ER-) vs. $n_2 = 164$ (ER+);
- Total number of genes: $p = 22283$;
  - number of pathways: $K = 1320$;
  - number of genes in pathways: $p_k = 4$ to 778;

Human Breast Cancer Gene Expression Data

Sample ID	Class	$Pathway_1$			...	$Pathway_{1320}$		
		$Gene_1$	...	$Gene_{54}$	...	$Gene_1$	...	$Gene_{16}$
1	pCR	13.18	...	11.48	...	10.24	...	9.77
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
114	pCR	12.87	...	9.36	...	9.43	...	9.19
115	RD	13.72	...	7.90	...	10.73	...	9.38
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
278	RD	13.18	...	8.47	...	10.20	...	9.74

Endpoints

- Endpoint D: response to preoperative chemotherapy
  - pathological complete response (pCR) or residual invasive cancer (RD)
  - pCR means no residual invasive cancer in the breast or lymph nodes
  - Research have shown that pCR is a strong indicator of long-term cancer free survival for breast cancer patients (Hess *et al.*, 2006; Popovici *et al.*, 2010)
  - # of pCR vs RD is very unbalanced (56:222)
- Endpoint E: the clinical estrogen-receptor status as established by immunohistochemistry.
  - ER- or ER+
  - ER– is one of the histologic characteristics that indicate more chemotherapy-sensitive cancer (Hess *et al.*, 2006)
  - # of ER- vs ER+ is relatively balanced (114:164)
- pCR prediction is a moderately difficult problem, compared to ER-status (Fan *et al.*, 2009; Popovici *et al.*, 2010; Cai *et al.*, 2012)

Why to choose QDA-based analysis?

- We assume different inverse covariance structure across heterogeneous classes, so quadratic discriminant analysis (QDA) is selected over LDA.
- However, conventional QDA has been shown to have shortcomings in high-dimensional settings.
- Only a few research have applied sparse estimates of covariance/precision matrices to QDA in high-dimensional settings.
  1. Sun and Zhao (2015) applied a sparse version of QDA for classification problems.
  2. Pavlenko *et al.* (2012) and Le and Hastie (2014) applied the sparse estimate of block-diagonal precision matrices to QDA, which assume no pathway level network.

Application Steps

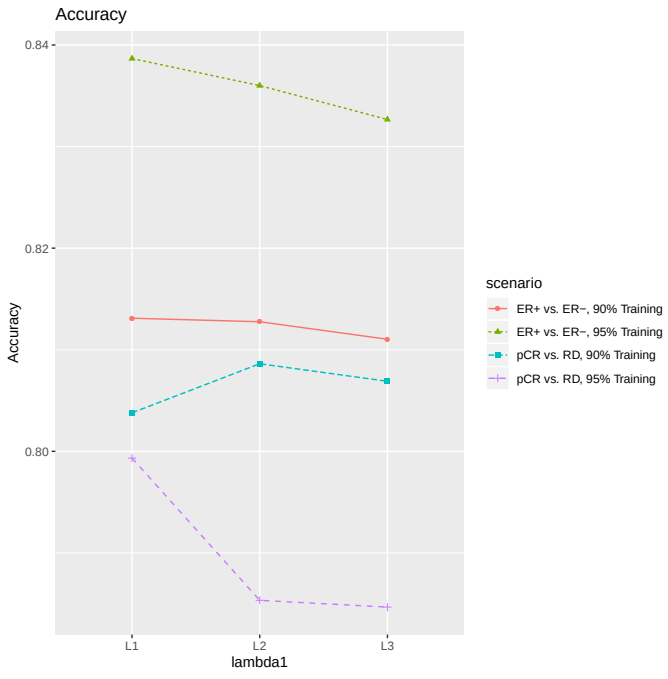
1. Divide the data into a training set (95% **or** 90%) and a testing set, stratified by response to preoperative chemotherapy **or** ER status.
2. Use the training set to estimate the two-level precision matrix under different λ_1 levels (L1,L2,**or** L3), and plug in the estimation to the QDA score formula to classify the testing set observations.
3. Calculate classification measurements, including ACC, MCC, TPR, TNR, PPV, and NPV.
4. Repeat the process 100 times and get the mean of the 100 classfication measurement sets.

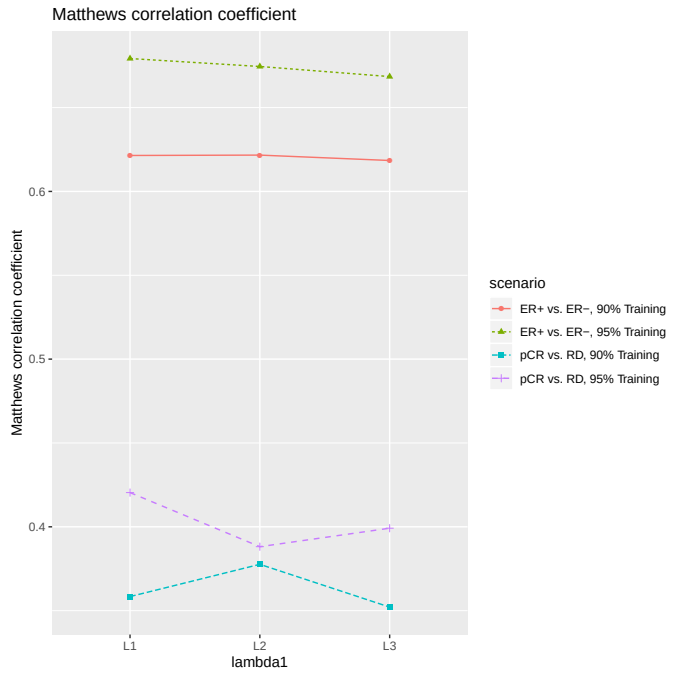
Revisit the Joint Estimation Method

$$\min_{\{\Omega_{kk'}^{(m)},\Gamma_{kk}^{(m)}\}} \quad n_m[\mathrm{trace}(S^{(m)}\Omega^{(m)}) - \log|\Omega^{(m)}|] + \lambda_1 \sum_{k\neq k'} \frac{\sum_{\substack{1\leqslant i\leqslant p_k \\ 1\leqslant j\leqslant p_{k'}}} |\omega_{i,j}^{kk'(m)}|}{(\sum_{m=1}^{M} \sum_{\substack{1\leqslant i\leqslant p_k \\ 1\leqslant j\leqslant p_{k'}}} |\omega_{i,j}^{kk'(m)^{(t)}}|)^{1/2}} + \lambda_2 \sum_{k=1}^{K} \sum_{1\leqslant i\neq j\leqslant p_k} \frac{|\gamma_{i,j}^{kk(m)}|}{(\sum_{m=1}^{M} |\gamma_{i,j}^{kk(m)^{(t)}}|)^{1/2}} \quad (2)$$

Note: λ_1 majorly controls the sparsity level of the higher-level network, and choice of the optimal λ_1 is related to sample size, so we tried different levels of λ_1 (L1<L2<L3).

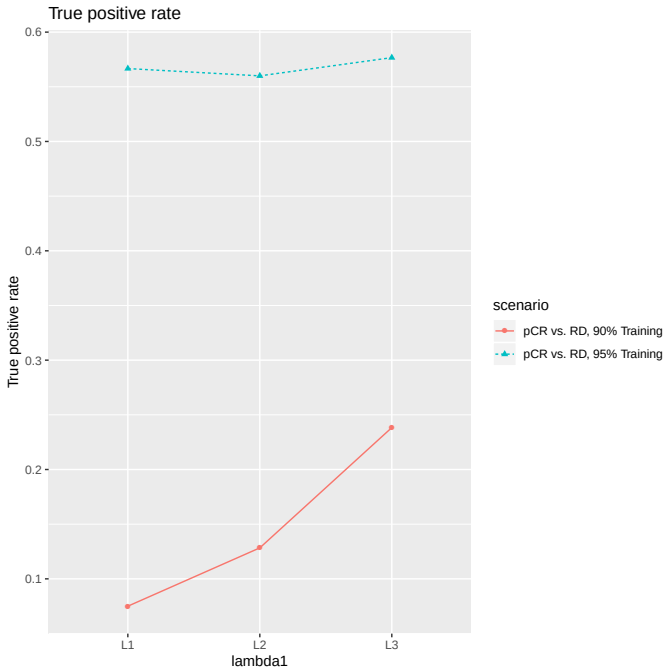
Results_Accuracy and Matthews Correlation Coefficient

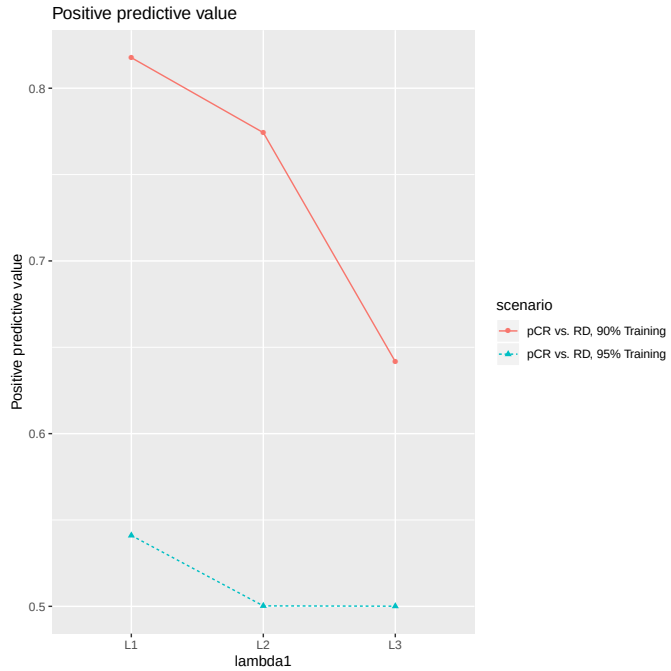




- Prediction performance for ER status, in terms of ACC and MCC, is generally better than that for response to preoperative chemotherapy.
- MCC is a balanced measure for unbalanced data.

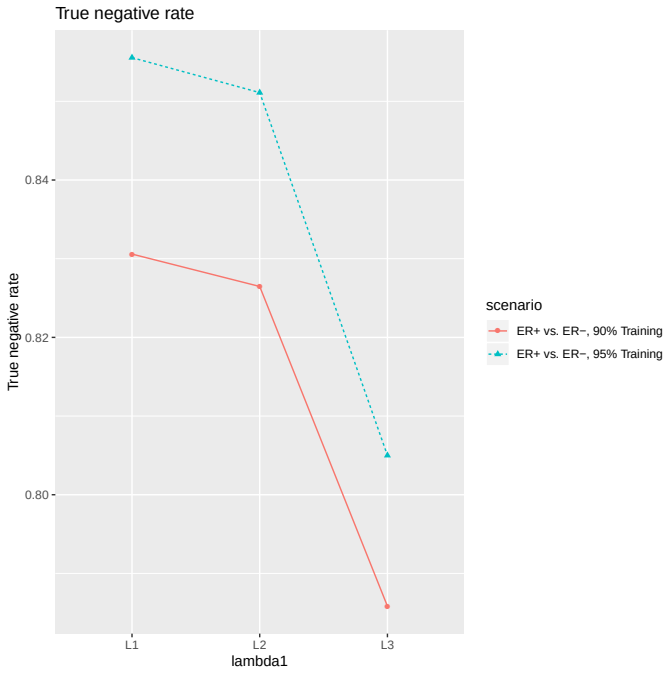
Results_pCR

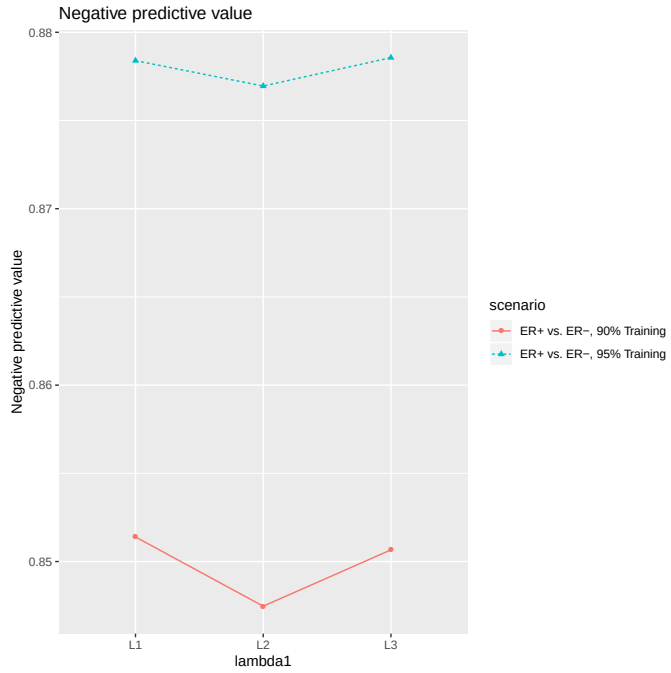




- As λ_1 increase, sparsity level increases in terms of the two-level network, better for explanation.
- TPR and PPV behave oppositely with different training set percentages.
- pCR is the minority class for the very unbalanced data.

Results_ER-





- As λ_1 increase, sparsity level increases a little in terms of the two-level network.
- TNR and NPV behave in the same direction with different training set percentages, more training data is better.
- ER- is the minority class for the relative balanced data.

Future Research

- Unbalanced data: method and QDA
- compare with other methods that have been developed for high-dimensional QDA
- Apply to TCGA dataset

Acknowledgements

- The University of Alabama at Birmingham
  - Sejong Bae, PhD
  - Dongquan Chen, PhD MSHI
- Virginia Tech
  - Inyoung Kim, PhD
- This study was in part supported by P30 CA13148

Questions?

THANKS!!