ORF350 (Spring 2017) Analysis of Big Data **Assignment: 6**

Probabilistic Graphical Model

Lecturer: Han Liu *Email: hanliu@princeton.edu*

Due Date: This assignment is due on 5PM Thursday May 11 2017.

Submission: Please submit your hard copy at the ORF 350 dropbox in the Sherrerd Hall student lounge. Electronic submissions will not be accepted unless there is an extreme and compelling case. Late submission should refer to the the Late Day policy section in the Class Rules.

In the lectures, we have learned two very power methods for high dimensional regression and classification: the Lasso (or ℓ_1-penalized least square regression) and sparse logistic regression (or ℓ_1-penalized logistic regression). In this problem set, we will teach you how these two methods are related to a very powerful data analysis tool–"Probabilistic graphical models"!

Given d random variables $X_1, \ldots, X_d$, an undirected graphical model is a statistical model (thus, by definition it is a set of probability distributions) associated with an undirected graph (Each node of the graph corresponds to one of these variables). As will be explained in the later part of this question, this graph specifies the conditional dependence structures between these random variables. By applying the Lasso and logistic regression, we can exploit purely observational data to recover and visualize the conditional dependence structures of random variables.

Before digging into the formal definition and theory of the graphical model, let's first get a taste of how the graphical model can be used as a powerful tool for data analysis.

Question 1. Nationwide GDP Growth Correlation (30 points)

Big Data! Big World!

Background:
In this modern era of globalization, we live in a world where the economic growth of any nation may depend on the prosperity of many other nations–with whom the nation share political ties, geographical dependence, or trade partnerships. These economic relationships are often complex and difficulty to summarize, but can be very useful information for global investors and policy makers. One powerful approach is to visualize these relationships using a graph that connects countries whose economic growths are correlated. We will do just that!
In particular, we use Gross Domestic Product (GDP) as the measure for economic growth of a country. The GDP is the monetary values of all the goods and services produced in a country during a year, and it is commonly used to access a country's economic growth and health. In this problem, we will use this GDP data to estimate a graphical model that represents the economic dependence between different countries.

Load `gdp.Rdata`, you will get a numerical matrix `gdp` where each row has GDP annual growth rates of certain country from 2001 to 2014. Our goal is to conduct a graph analytics of the GDP growth among countries using the graphical model tool. Install and load the packages `glmnet` and `igraph`. You can also find a function called `graphplot` in `q4.r` for visualizing the estimated graph.

Question 1.1 (10 points): Note that the matrix `gdp` has missing values. Directly delete countries that have no GDP rates reported. Then, for a given country, replace `NA`'s with the mean value of the reported growth rates across the years.

(Hint: The function `is.na` will be useful.)

Question 1.2 (20 points):

Given a Country A, we use the Lasso regression to find out the countries whose GDP growth rates are significantly associated (in the linear regression sense) with Country A. In particular, do the following analysis:

1. Create a $d \times d$ matrix of zeros, $\boldsymbol{M}$, where d is the number of different countries in the dataset. This will be your adjacency matrix.

   In an adjacency matrix, $\boldsymbol{M}_{jk} = 1$, if there is an edge between the j-th and k-th nodes of the graph. $\boldsymbol{M}_{jk} = 0$, otherwise. (The diagonal entries do not matter.)

   In this problem, the j-th and k-th nodes represent the j-th and k-th countries, respectively.

   You will populate this matrix in the next few steps. (Observe that $\boldsymbol{M}$ must be symmetric since we are working with undirected graphs.)

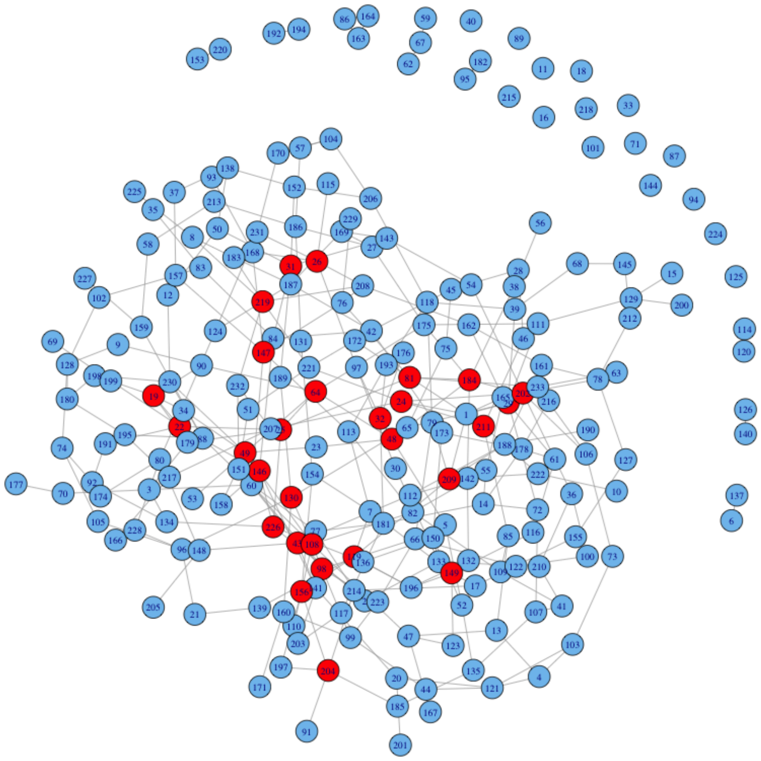
2. For each country, apply Lasso so that we can find the neighborhood of every node and thus recover the whole graph. In particular, call the function `glmnet` in `R` to regress the GDP growth rate of the country (call it Country A) over the GDP growth rates of all other countries with the tuning parameter $\lambda = 1$.

3. Country A's GDP growth is associated with Country B's growth if Lasso chooses a non-zero coefficient for the variable associated with Country B. Use this fact to fill in the correct entries in $\boldsymbol{M}$.

   (Note 1: It is possible that we select country B as the neighborhood of country A but do not select country A as the neighborhood as country B. In this case, we adopt the "**AND**" rule to address the contradiction here, meaning that country A and B are connected in the graph if and only if both A and B are in the neighborhood of each other.)

   (Note 2: As a result, the countries selected by Lasso in this approach are considered the neighborhood of country A in the GDP growth interaction graph. We will provide theoretical justifications later in Part III.)

4. Visualize the estimated graph by the function `graphplot` provided in `q4.r` (which takes an adjacency matrix as input). This function will print the graph in a file called `gdp_nodewise.png` with nodes colored red if they have degrees greater than or equal to five and colored blue if they have degrees less than five. Your graph should be similar to the following one.



5. List the country names that correspond to the red nodes in the graph.

Question 2. Preliminary Theories (25 points)

In this part, we rigorously establish the theory of the graphical model. First, we provide a brief lesson on **independence** and **conditional independence** and their application to graphical models.

Time to Learn!

Independence and Conditional Independence:
Suppose we have d random variables $\{X_j\}_{j=1}^d$, and for an index set A we let $\boldsymbol{X}_A = \{X_j : j \in A\}$. For three index sets A, B and C such that $A \cap B = \emptyset$, $B \cap C = \emptyset$ and $A \cap C = \emptyset$, we say $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are globally independent if

$$p_{\boldsymbol{X}_A,\boldsymbol{X}_B}(\mathbf{x}_A, \mathbf{x}_B) = p_{\boldsymbol{X}_A}(\mathbf{x}_A) p_{\boldsymbol{X}_B}(\mathbf{x}_B)$$

for arbitrary realizations $\mathbf{x}_A, \mathbf{x}_B \in \mathbb{R}^d$. And we say $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are conditionally independent given $\boldsymbol{X}_C$ if

$$p_{\boldsymbol{X}_A,\boldsymbol{X}_B \mid \boldsymbol{X}_C}(\mathbf{x}_A, \mathbf{x}_B \mid \mathbf{x}_c) = p_{\boldsymbol{X}_A \mid \boldsymbol{X}_C}(\mathbf{x}_A \mid \mathbf{x}_C) p_{\boldsymbol{X}_B \mid \boldsymbol{X}_C}(\mathbf{x}_B \mid \mathbf{x}_C)$$

for arbitrary realizations $\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C \in \mathbb{R}^d$, where $p(\cdot)$ could be a probability density or mass function depending on whether the random variable is continuous or discrete.

Application to Graphical Models:
The key idea of the graphical model is to represent conditional independence relationship by a graph. More specifically, we define a graph $G = (V, E)$. Here $V = \{1, \ldots, d\}$ is the node set corresponding the random variables $X_1, \ldots, X_d$. $E \subset V \times V$ is the edge set. For each pair of nodes $j, k \in V$, an edge (j, k) is not in the edge set E if and only if the corresponding variables X_j and X_k are conditionally independent given the rest of the variables $\boldsymbol{X}_{\backslash j}$ (where $\backslash j := \{\ell : \ell \neq j, k\}$). More rigorously we say

$$\forall j, k \in V, (j, k) \notin E \quad \text{if and only if} \quad X_j \perp X_k \mid X_{\backslash\{j,k\}}. \tag{0.1}$$

Therefore, this graph actually visualizes the the conditional dependence relationship between any pair of variables.

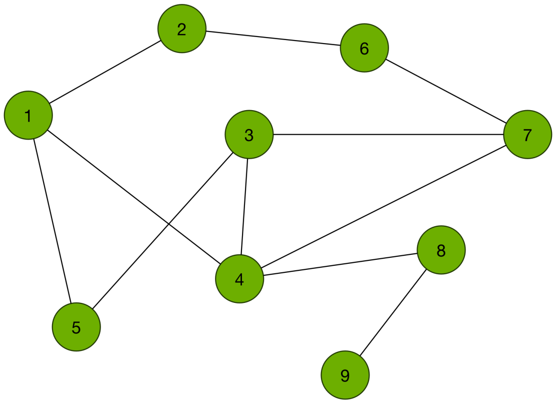
Separation:
In the graph theory, there is an important concept called "separation": Let $A, B, C \subset V$ such that every two of them have no intersection. We say a node set C separates A and B if any path connecting the nodes in A and B must contain at least one node in C (In another word, the removal of C renders A and B disconnected).
Even though the graph G is defined only based on pairwise conditional independence (i.e., each time, each edge (j, k) only corresponds to the conditional independence relationship between two variables X_j and X_k). However, an amazing result shows that it actually encodes a more powerful relationship! More specifically, under very weak assumptions (which always holds in our class), for the graph G we develop in (0.1), it holds that if C separates A and B, then $\boldsymbol{X}_A \perp \boldsymbol{X}_B \mid \boldsymbol{X}_C$. So here we establish the correspondence between separation and conditional independence.

Now, we apply this knowledge in some exercises.

Question 2.1 (10 points): Construct concrete examples to show that global independence cannot imply conditional independence and conditional independence cannot imply global independence either.

(Hint 1: For the example of global independence not leading to conditional independence, consider the independent coin tosses. Construct some event conditional on which the coin tosses are not independent any more. For the example of conditional independence not leading to global independence, consider two responses in the linear regression model that share the same design covariate.)

Question 2.2 (15 points): Consider the graphical model as illustrated below. Denote the random variable corresponding to node i by X_i.



Answer the following questions:

- Does the node set $\{3\}$ separate node set $\{5\}$ and node set $\{7\}$?
- Does the node set $\{8\}$ separate node set $\{7\}$ and node set $\{9\}$?
- Is it true that $X_1 \perp X_8 \mid \boldsymbol{X}_{\setminus\{1,8\}}$?
- Is it true that $X_1 \perp X_8 \mid X_4$?
- Is it true that $X_1 \perp X_3 \mid X_5$?

Question 3. The Gaussian Graphical Model (25 points)

Having introduced the general concept of undirected graphical models, we now focus on a specialized case: the Gaussian graphical model. We aim to provide theoretical justification on why we can use node-wise Lasso in Part I to estimate the graph.

Time to Learn!
Gaussian Graphical Models
Suppose $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{X}$ is a d-dimensional random vector and $\boldsymbol{\Sigma}$ is a d-by-d covariance matrix. It turns out that the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ encodes the information of the conditional dependence structure of $\boldsymbol{X}$, as indicated by the following theorem.

Theorem 0.1. *Suppose $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{X}$ is a d-dimensional random vector and $\boldsymbol{\Sigma}$ is a d-by-d matrix. Then for any different $j, k \in \{1, ..., d\}$, we have*

$$X_j \perp X_k \mid X_{\setminus\{j,k\}} \Longleftrightarrow \boldsymbol{\Theta}_{jk} = 0$$

where $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, $\perp$ means independence, and $\mid X_{\setminus\{j,k\}}$ means conditional on all the variables expect X_j and X_k. Therefore estimating the edge set E is equivalent to estimating the support of $\boldsymbol{\Theta}$.

To prove the the theorem above, we use two steps that correspond to two small questions below. Without loss of generality, we can always assume that $j = 1$ and $k = 2$. We will use the following result directly without proof. For those who are interested in deriving the result, please see the blog posts. For any $A \subseteq \{1, ..., d\}$, we have

$$\boldsymbol{X}_A \,|\, \boldsymbol{X}_{A^c} \sim N(\boldsymbol{\Sigma}_{AA^c}\boldsymbol{\Sigma}_{A^cA^c}^{-1}\boldsymbol{X}_{A^c}, \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AA^c}\boldsymbol{\Sigma}_{A^cA^c}^{-1}\boldsymbol{\Sigma}_{A^cA}), \tag{0.2}$$

where A^c is the complement set of A.

Question 3.1 (10 points):

Suppose $A = \{1, 2\}$. We partition the covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ in the following pattern:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AA^c} \\ \boldsymbol{\Sigma}_{A^cA} & \boldsymbol{\Sigma}_{A^cA^c} \end{bmatrix}, and \ \boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{AA} & \boldsymbol{\Theta}_{AA^c} \\ \boldsymbol{\Theta}_{A^cA} & \boldsymbol{\Theta}_{A^cA^c} \end{bmatrix}.$$

Show that $\boldsymbol{\Theta}_{AA}^{-1} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AA^c}\boldsymbol{\Sigma}_{A^cA^c}^{-1}\boldsymbol{\Sigma}_{A^cA}$. Combined with the results (0.2), we know that $\boldsymbol{X}_A \,|\, \boldsymbol{X}_{A^c} \sim N(\boldsymbol{\Sigma}_{AA^c}\boldsymbol{\Sigma}_{A^cA^c}^{-1}\boldsymbol{X}_{A^c}, \boldsymbol{\Theta}_{AA}^{-1})$.

(Hint: Start from the fact that $\boldsymbol{\Theta\Sigma} = \mathbf{I}$)

Question 3.2 (10 points): Use the result derived from Question 3.1 to reach the final conclusion.

(Hint: Two Gaussian variables are independent to each other if and only if their covariance is zero.)

Now let's discuss how to recover the support of $\boldsymbol{\Theta}$ under the Gaussian graphical model, or equivalently speaking, the edge set E in the graph G. In Part I, we run nodewise Lasso to find significant association between covariates. This is actually called the method of *neighborhood pursuit*, which means that we estimate $\boldsymbol{\Theta}$ by investigating the neighborhood of each node in the graph. Here we justify this method theoretically under the Gaussian graphical model. For X_j, define its neighbor $\mathcal{N}(X_j) := \{k : \Theta_{jk} \neq 0\}$. Without loss of generality, we look at the neighborhood of X_1, and we denote the other variables by $\boldsymbol{X}_{\backslash 1}$ for notational convenience. By (0.2), we can express X_1 as

$$X_1 = \underbrace{\boldsymbol{\Sigma}_{1,\backslash 1}\boldsymbol{\Sigma}_{\backslash 1,\backslash 1}^{-1}}_{\boldsymbol{\beta}^T}\boldsymbol{X}_{\backslash 1} + \epsilon, \tag{0.3}$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\backslash 1,\backslash 1}^{-1}\boldsymbol{\Sigma}_{\backslash 1,1}$ and $\mathrm{Var}(\epsilon) = \Sigma_{11} - \boldsymbol{\Sigma}_{1,\backslash 1}\boldsymbol{\Sigma}_{\backslash 1,\backslash 1}^{-1}\boldsymbol{\Sigma}_{\backslash 1,1}$.

Question 3.3 (5 points): Show that $\epsilon \perp \boldsymbol{X}_{\backslash 1}$ and $\boldsymbol{\beta} = -\Theta_{11}^{-1}\boldsymbol{\Theta}_{\backslash 1,1}$, which implies that the support of $\boldsymbol{\beta}$ is the same as $\boldsymbol{\Theta}_{\backslash 1,1}$. In this way, recovering the support of $\boldsymbol{\beta}$ is equivalent to recovering the support of $\boldsymbol{\Theta}_{\backslash 1,1}$, i.e., the neighborhood of X_1 on the graph. This question justifies the method of neighborhood pursuit.

Question 4. The Ising model (20 points)

In this part, we introduce another kind of probabilistic graphical model called the Ising graphical model, which is very important in statistical physics. Under the Ising model, $X_1, ..., X_d \in \{-1, 1\}$ and for $\{\beta_j\}_{j=1}^d \in \mathbb{R}$ and $\{\beta_{jk}\}_{j\neq k} \in \mathbb{R}$ such that $\beta_{jk} = \beta_{kj}$, we have

$$P(X_1 = x_1, X_2 = x_2, ..., X_d = x_d) = \frac{1}{Z}\exp(\sum_{j=1}^{d}\beta_j x_j + \sum_{j<k}\beta_{jk}x_j x_k), \tag{0.4}$$

where $Z = \sum_{x_1,x_2,...,x_d\in\{-1,1\}}\exp(\sum_{j=1}^d \beta_j x_j + \sum_{j<k}\beta_{jk}x_j x_k)$. Similarly, we will show that under this Ising model, $\{\beta_{jk}\}_{j,k=1}^d$ capture the conditional dependence structure of $\{X_j\}_{j=1}^d$ and that the neighborhood pursuit method is still effective in recovering the support of $\{\beta_{jk}\}_{j,k=1}^d$.

Question 4.1 (10 points):

Analogous to Theorem 0.1, show that $X_j \perp X_k \mid X_{\setminus\{j,k\}} \iff \beta_{jk} = 0$ for different j and k.

(Hint: There will be many exponential quantities in your equations. Denoting them by short letters like A_1 and A_2 will very much simplify the equations you need to verify.)

Question 4.2 (10 points):

Show that

$$P(X_j = 1 | \boldsymbol{X}_{\setminus j} = \mathbf{x}_{\setminus j}) = \frac{1}{1 + \exp(-2(\beta_j + \sum_{k \neq j} \beta_{jk} x_k))},$$

which is a logistic regression model if we treat X_j as the class label and $\boldsymbol{X}_{\setminus j}$ as the covariates.

(Therefore from the perspective of the neighborhood pursuit, we can run logistic regression of X_j over $\boldsymbol{X}_{\setminus j}$ to estimate $\{\beta_{jk}\}_{k=1}^{d}$. Then by the result from Question 3.8, we can recover the correspondent graphical model.)