Project

Problem 1. *The Population Version of the PC Algorithm*
The PC algorithm can be utilized to learn the causal graph between a set of random variables $V = \{X_1, \cdots, X_p\}$ up to Markov equivalence classes. Here, we will implement the population version of the PC algorithm with a given significance level.

- Implement the following parts of the PC algorithm:

Algorithm 1

Input: Vertex set V, observational data D, significance level α.
Output: Estimated skeleton $\hat{G}$ including v-structures.
(1) Construct the complete undirected graph on the vertex set V.
(2) Perform conditional independence tests at a given significance level α and delete edges based on the tests.
(3) Orient v-structures.

The observational data D is a $n \times p$ matrix where n is the number of samples. The ith column of matrix D corresponds to observational data of variable X_i.

Let $\hat{\rho}_{X_i,X_j|S}$ be the sample partial correlation of two variable X_i, X_j given the variables in the set $S \subseteq V \backslash \{X_i, X_j\}$. In part (2), assume that the joint distribution of random variables in V is a multivariate normal distribution. In order to test whether $\hat{\rho}_{X_i,X_j|S}$ is zero or not, you can compare $|\hat{\rho}_{X_i,X_j|S}|$ with the threshold $\Phi(1-\alpha/2)^{-1}/\sqrt{n-|S|-3}$ where $\Phi(.)$ denotes the cdf of $\mathcal{N}(0,1)$.

- Plot the graph $\hat{G}$ for the observational data.
- Try to orient as many undirected edges as possible by applying Meek rules.
- How many DAGs are in the corresponding equivalence class?
- The time complexity of an algorithm is commonly estimated by counting the number of elementary operations performed by the algorithm. Considering conditional independence tests as the elementary operations of the PC algorithm, what is the time complexity of this algorithm? (Let $d_{\max}$ be the maximum degree of the graph).

Problem 2. *Analyzing the Stock Market data*
The goal of this task is to learn the causal structure among different technology companies by observing their stock price over a period of time. The list of these companies appear in Table 1.

The prices of these companies were sampled every 2 minutes for seven market days (03/03/2008 - 03/10/2008), i.e., $T = 7$. We can assume that the underlying dynamic which describes the relationships between the log-prices of these companies is jointly Gaussian. This is due to the Black-Scholes model of the market. Therefore, directed information values can be estimated using the following equation,

Name	code
Apple Inc.	APPL (1)
Cisco Systems	CSCO (2)
Dell Inc.	DEL (3)
EMC Corporation	EMC (4)
Google Inc.	GOG (5)
Hewlett-Packard	HP (6)
Intel	INT (7)
Microsoft	MSFT (8)
Oracle	ORC (9)
International Business Machines	IBM (10)
Texas Instruments	TXN (11)
Xerox	XRX (12)

Table 1: List of Companies in the analysis

$$I(x \to y||\underline{z}) = \frac{1}{2}\sum_{t=1}^{T} \log \frac{|\Sigma_{y_1^t \underline{z}_1^{t-1}}||\Sigma_{x_1^{t-1} y_1^{t-1} \underline{z}_1^{t-1}}|}{|\Sigma_{y_1^{t-1} \underline{z}_1^{t-1}}||\Sigma_{x_1^{t-1} y_1^t \underline{z}_1^{t-1}}|}, \tag{1}$$

where $\Sigma_{y_1^t \underline{z}_1^{t-1}}$ is the covariance matrix of $(y(1), ..., y(t), \underline{z}(1), ..., \underline{z}(t-1))$. The log-price of the technology companies are given in a form of a Matlab file. In this file, you will find an array M, where $M(i, j, k)$ denotes the j-th sample of the log-price of k-th company at day i. Learn the directed information graph among these companies and plot it.