Homework # 3

Due to 19.05.2020
Please submit before the TA session on 19.05

Problem 1. Consider a linear model $x = Bx + e$. The LiNGAM algorithm after the ICA step, obtains the following matrix,

$$W = \begin{pmatrix} 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0.3 & -0.3 \\ -0.2 & 0.3 & -0.5 & 0.1 & 0 \\ -0.8 & 0.4 & 0 & 0 & 0 \\ -0.6 & 0 & 0.6 & 0 & 0 \end{pmatrix}.$$

Find matrix B using the remaining steps of the LiNGAM algorithm.
Hint: $W = PD(I - B)$. Find the correct permutation matrix P and the diagonal matrix D.

Problem 2. The covariance matrix between the observed nodes of a jointly Gaussian network is given below. Using the algorithm discussed in the class, find the complete network (latent nodes and the edges between latent and observed).
Hint: The complete network is a tree and has 11 nodes.

$$\Sigma \approx \begin{pmatrix} 0.1913 & 0.0073 & -0.0008 & -0.0034 & -0.0018 & 0.0001 & 0.0000 \\ 0.0073 & 0.3320 & -0.0022 & -0.0093 & -0.0813 & 0.0003 & 0.0001 \\ -0.0008 & -0.0022 & 0.3839 & 0.2703 & 0.0005 & -0.0481 & -0.0193 \\ -0.0034 & -0.0093 & 0.2703 & 1.5064 & 0.0023 & -0.0339 & -0.0136 \\ -0.0018 & -0.0813 & 0.0005 & 0.0023 & 0.2240 & -0.0001 & 0.0000 \\ 0.0001 & 0.0003 & -0.0481 & -0.0339 & -0.0001 & 0.2954 & 0.0365 \\ 0.0000 & 0.0001 & -0.0193 & -0.0136 & 0.0000 & 0.0365 & 0.2187 \end{pmatrix}$$

Please use the file uploaded on the course web that contains the actual covariance matrix.

Problem 3. You are given the realizations of a linear autoregressive model of six time series, i.e.,

$$\vec{X}_t = \mathbf{A}\vec{X}_{t-1} + \vec{W}_t,$$

where $\vec{X}_t = (X_{1,t}, ..., X_{6,t})^T$ and $\vec{W}_t$ is a vector of i.i.d Gaussian random variables with mean zero. The file "P1" contains a multidimensional arrays XX of dimension $6 \times 300 \times 20$, where $XX(p, n, t)$ represents the sample of $X_{p,t}$ at n-th trial.

I) Find the causal structure among these 6 time series by estimating the directed information quantities.
Hint: In case that all processes are jointly Gaussian, we have

$$I(X \to Y||Z) = \frac{1}{2}\sum_{t=1}^{T} \log \frac{|\Sigma_{Y_1^t Z_1^{t-1}}||\Sigma_{X_1^{t-1} Y_1^{t-1} Z_1^{t-1}}|}{|\Sigma_{Y_1^{t-1} Z_1^{t-1}}||\Sigma_{X_1^{t-1} Y_1^t Z_1^{t-1}}|},$$

where $|\Sigma_{Y_1^t Z_1^{t-1}}|$ denotes the determinant of the covariance matrix of $(Y_1, ..., Y_t, Z_1, ..., Z_{t-1})$. To decide whether your estimated DI is zero or positive use 0.6 as the threshold.

II) Learn the coefficient matrix $\mathbf{A}$ using linear regression, i.e., by solving

$$\min_{\mathbf{B}\in\mathbb{R}^{6\times 6}} ||\mathbf{Y} - \mathbf{B}^T\mathbf{X}||^2,$$

where $\mathbf{Y} = [\vec{X}_T, ..., \vec{X}_2]$, and $\mathbf{X} = [\vec{X}_{T-1}, ..., \vec{X}_1]$.
Compare the adjacency matrix of the causal structure in part I and the support of the coefficient matrix $\mathbf{A}$.

Problem 4. Implementation of Chow-Liu algorithm. The second file, "P2", contains i.i.d. samples of a multivariate normal distribution. Implement the Chow-Liu algorithm discussed in the class to find the best tree approximation for this dataset.
Hint: Since the variables are jointly Gaussian, you can estimate the mutual information quantities by estimating the covariance matrices.