Homework 1

Xiaocheng Zhou (1155184323)

March 1, 2023

1. From path diagram to equations

   (a) Write down the explicit form of the model.

   measurement equation:

$$\begin{bmatrix} \text{PVD} \\ \text{IHD} \\ \text{CVA} \\ \text{BMI} \\ \text{WHR} \\ \text{SBP} \\ \text{DBP} \\ \text{InTG} \\ \text{LDL-C} \\ \text{PON11} \\ \text{PON12} \\ \text{FBG} \\ \text{SELE2} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \lambda_{2,1} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{3,1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_{5,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_{7,3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{9,4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{11,5} & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{12,6} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \eta \\ \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \end{bmatrix}$$

   structural equation:

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_3 + \gamma_4\xi_4 + \gamma_5\xi_5 + \gamma_6\xi_2\xi_4 + \gamma_7\xi_3\xi_5 + \gamma_8\xi_4\xi_5 + \delta$$

   or

$$\eta = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \\ \xi_2\xi_4 \\ \xi_3\xi_5 \\ \xi_4\xi_5 \end{bmatrix} + \delta$$

   (b) Interpret the model in terms of the variables provided in the rectangles and ellipses.

   First, in the measurement equation, based on some medical knowledge (provided by this question), *PVD*, *IHD*, and *CVA* are measured to form the latent variable **cardiovascular heart disease** (η), also the outcome for this model; *WHR* and BMI are measured to obtain the observed variables for forming the latent variable **body shape** (ξ_1); *SBP* and *DBP* are measured for achieving the latent

variable **blood pressure** (ξ_2); *LDL-C* and *InTG* are measured for geting the latent variable **lipid control** (ξ_3); *PON11* and *PON12* are measured for forming the latent variable **gene-inflammatory** (ξ_4); and *FBG* and *SELE2* are measured for achieving the latent variable **gene-lipid control** (ξ_5). Second, in the structural equation, the *cardiovascular heart disease* outcome is linearly affected by all rest latent variables we discussed above and some interactive terms of them, including **body shape**, **blood pressure**, **lipid control**, **gene-inflammatory**, **gene-lipid control**, the interactive term of **blood pressure and gene-inflammatory**, interactive term of **lipid control and gene-lipid control**, and interactive term of **gene-inflammatory and gene-lipid control**.

(c) Is the model identified? Why?
Yes, the SEM model is identified, since the factor loading matrix has a non-overlapping structure and the values of $\lambda_{1,1}$, $\lambda_{4,2}$, $\lambda_{6,3}$, $\lambda_{8,4}$, $\lambda_{10,5}$, and $\lambda_{13,6}$ are fixed as 1, then the only non-singular matrix **M** that can cause unidentifiablity is **I**.

(d) List and interpret the fixed and unknown parameters in the model.
I would like to take several values in loading matrix and coefficients in structural equation for examples. The coefficient of InTG to lipid control is fixed to 1, which presents that the InTG has the same increasing rate with the designed latent variable lipid control, and every unit change in lipid control will change 0.377 unit ($\lambda_{9,4}$) in LDL-C. The interaction of gene-lipid control and lipid control has a relatively large effect ($\gamma_7 = 1.223$) on the cardiovascular heart disease, and all the other latent variables will affect the PVD, IHD, and CVA indirectly through the disease of cardiovascular heart.

(e) Discuss all possible submodels of model (a).
i. We can consider any random combination or higher powers of $\xi_{\{1:5\}}$, as explanatory variables in the SEM;
ii. We can use delete one or more observed variables that related to the latent variables;
iii. We can delete the latent variable and only use one of the observed variable to reflect the latent trait.

(f) Discuss the differences between SEMs and conventional regression models. The regression models are designed for finding the relationship between variables that depict the effect that variables cast on outcome, where all variables are observed, while SEMs are latent model that contain variables that cannot be seen, which are able to find the relationship among latent and observed variables. Moreover, SEMs possess the capability to model latent constructs, including direct and interactive effects among the latent variable and observed variables. It also avoids the multicollinearity problem by viewing some observed variables are generated by latent factors, however, the conventional regression model may collapse when encountering such problems.

2. From equations to path diagram

(a) Write the model in a matrix form and draw the corresponding path diagram.
measurement equation:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} \mu_1 & a_1 \\ \mu_2 & a_2 \\ \mu_3 & a_3 \\ \mu_4 & a_4 \\ \mu_5 & a_5 \\ \mu_6 & a_6 \\ \mu_7 & a_7 \\ \mu_8 & a_8 \\ \mu_9 & a_9 \\ \mu_{10} & a_{10} \end{bmatrix} \begin{bmatrix} 1 \\ c \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{2,1} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{4,2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{6,3} & 0 \\ 0 & 0 & \lambda_{7,3} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_{9,4} \\ 0 & 0 & 0 & \lambda_{10,4} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix}$$

structural equation

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \gamma_4 & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 & 0 & 0 \\ \gamma_5 & \gamma_6 & 0 & \gamma_7 & \gamma_8 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_1^2 \\ \xi_1\xi_2 \\ \xi_2^2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$$

The corresponding path diagram is shown in Figure 1 below.

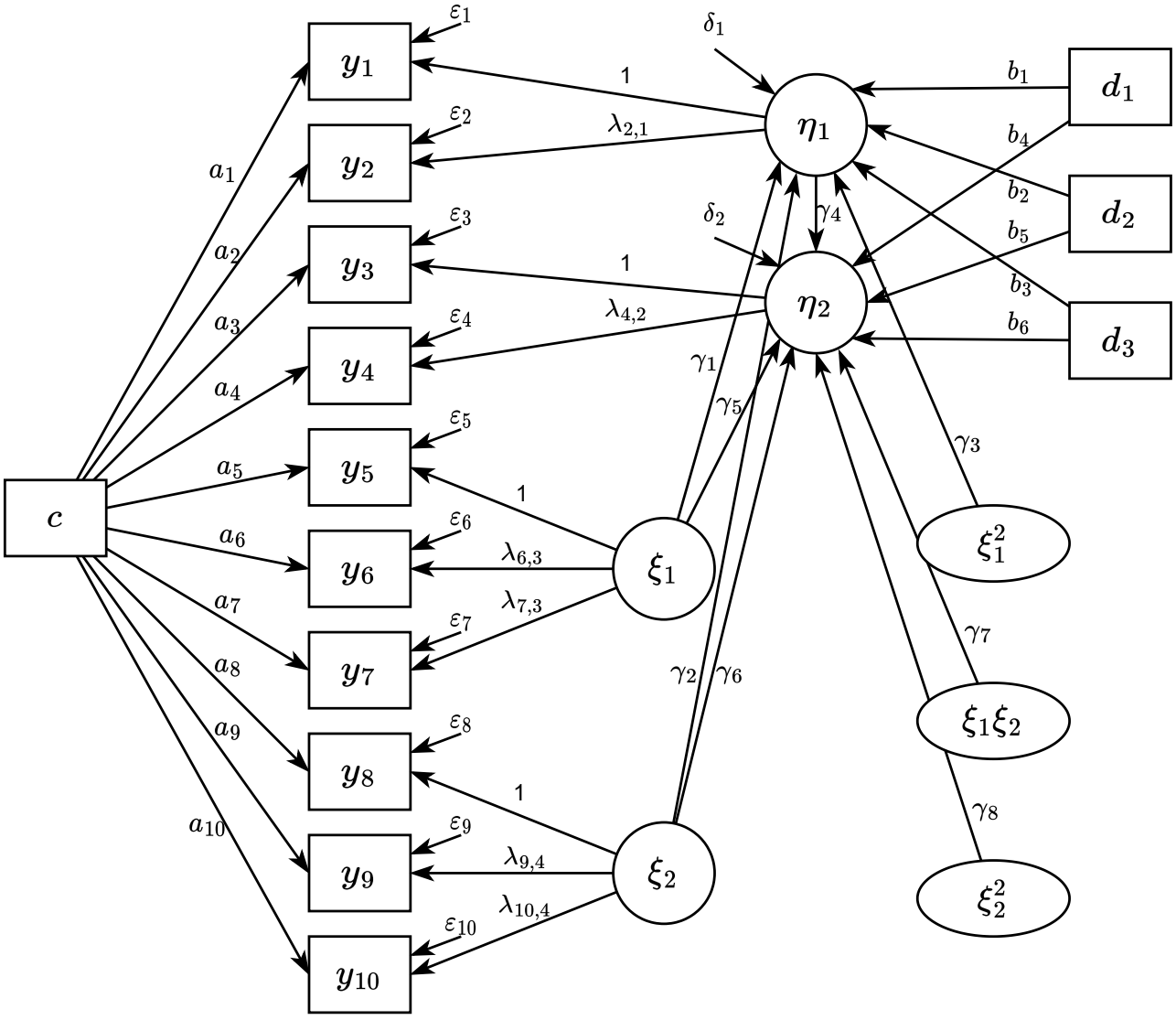

Figure 1: Path diagram of the equations in Q2

(b) In a Bayesian analysis, what prior distributions do we usually assign to the model parameters? Why?
We should choose non-informative priors when we don't have much information about the distribution of a data set. However, if we have information from other data sets or experts, we can choose an "informative prior" that encodes the your knowledge by designing the hyper-parameters, which helps us develop a more manageable MCMC algorithm for statistical inference. No matter if we have prior knowledge about this data set, we should pick a conjugate prior for target distribution, because it will facilitates our computation for posterior distribution.

(c) Derive the full conditional distributions in detail for the unknown parameters. To Be Continue...