***Note:***

1. Due 11:59PM, December 21, 2016.
2. Electronic submission to your instructor's email.
3. You are VERY MUCH encouraged to form teams to discuss proofs and program algorithms. If so, please acknowledge your teammate(s)' contributions at the beginning of your submitted homework. You must independently write your homework based on your own understanding.
4. Choose any programming language you like, R, Python, Matlab, C/C++, Julia, etc.

## *Examples and Implementations*

*[Bayesian approach to Latent Class Models: Definition, Simulation, Estimation and The Choice of Number of Classes] This Problem is a simulation study of latent class models, which is a widely useful and effective class of models for studying multivariate discrete data. The latent class models have a long history and wide applications in disease diagnosis, psychology, psychiatrics, pattern recognition, data compression, etc. You will be asked to simulate data from latent class models given parameters, and then hide the true parameters and fit the latent class models.*

To specify a latent class model with $M_0$ classes, we define $\boldsymbol{y}_i$, to be a vector of length $K$ indicating individual $i$'s binary response to $K$ items, $\eta_i \in \{1, \dots, M_0\}$ to be individual $i$'s unobserved latent class, and $\pi_j = P(\eta_i = j)$ to be the probability that individual $i$ is in class $j$ for $j = 1, \dots, M_0$. Here we assume there are $N$ subjects.

For example, in the studies investigating major depressive disorder, investigators obtain information on the symptoms through NIMH Diagnostic Interview Schedule. The data $\boldsymbol{y}_i$ is a vector representing the presence or absence of $K$ symptoms of depression for individual $i$, $\eta_i$ is individual $i's$ true but unknown depression class, and $\pi_j$ is the proportion of individuals in the population of which our sample is representative in depression class $j$.

Given $\eta_i$, elements $y_{ik}$ of $\boldsymbol{y}_i$ are assumed to be mutually independent so that the distribution of $\boldsymbol{y}_i$ is

$$f(\boldsymbol{y}_i; \boldsymbol{\pi}, \boldsymbol{p}) = \sum_{j=1}^{M_0} \pi_j \prod_{k=1}^{K} p_{jk}^{y_{ik}} \left(1 - p_{jk}\right)^{1-y_{ik}},$$

where $p_{jk} = P(y_{ik} = 1 \mid \eta_i = j)$ is the probability that individual $i$, who is in class $j$, will have a positive response to item $k$.

1) Draw the directed acyclic graph (DAG), $G$, with nodes $\{y_{ik}\}, \{p_{jk}\}, \{\pi_j\}, \{\eta_i\}$, so that the joint distribution with density $f(\boldsymbol{y}_i; \boldsymbol{\pi}, \boldsymbol{p}, \eta_i)$ is Markov to $G$. (*Note:* if we condition on an individual's latent class $\eta_i$, her binary response vector $\boldsymbol{y}_i$ is independent of $\boldsymbol{\pi}$. Also, use

minimal number of edges.)

2) In the DAG you drew, for a directed arrow from $\eta_i$ to $y_{ik}$, write the mathematical condition on $f(\boldsymbol{y}_i; \boldsymbol{\pi}, \boldsymbol{p}, \eta_i)$ that will make it disappear. State its interpretation.

3) Simulate a dataset, $D^*$, with $N = 300$ subjects, $M_0 = 3$ classes, $K = 5$ symptoms, with
$$\{p_{jk}\} = \begin{bmatrix} 0.1 & 0.9 & 0.1 & 0.15 & 0.1 \\ 0.4 & 0.4 & 0.45 & 0.5 & 0.4 \\ 0.95 & 0.1 & 0.9 & 0.9 & 0.9 \end{bmatrix},$$
and $\boldsymbol{\pi} = (0.5, 0.3, 0.2)'$. Calculate and tabulate the frequency of each *K*-dimension binary patterns ($2^K$ in total) and the *observed* pairwise log odds ratios $\hat{\psi}_{k,k'}^{obs,N} = \log \frac{\widehat{P_N}(y_{ik}=1, y_{ik'}=1)\widehat{P_N}(y_{ik}=0, y_{ik'}=0)}{\widehat{P_N}(y_{ik}=0, y_{ik'}=1)\widehat{P_N}(y_{ik}=1, y_{ik'}=0)}$ for all pairs of $(k, k')$ if 0/0 does not occur. (*Note*: fix a seed if you'll need me to reproduce your results.)

4) For ease of estimation, we reparametrize the model with $\{g_{jk} = \log\left(\frac{p_{jk}}{1-p_{jk}}\right)\}_{j=1,k=1}^{M^{fit},K}$, and $\{a_j = \log(\pi_j/\pi_{M^{fit}})\}_{j=1}^{M^{fit}-1}$, where $M^{fit}$ is the number of classes you specify when fitting the model that could be $M_0$ or not. Show the likelihood $f(\boldsymbol{Y} \mid \boldsymbol{a}, \boldsymbol{g})$, where $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_1^N, \boldsymbol{a} = \{a_j\}, \boldsymbol{g} = \{g_{jk}\}$.

5) Assuming a Bayesian model, we need to specify prior distributions for the parameters in our latent class model. For a model with $M^{fit}$ classes, let priors $g_{jk} \sim N(0, variance = 9/4)$, and $a_j \sim N(0, 9/4)$. Write out the full-conditional distributions (densities if continuous) for: $f(g_{jk} \mid \{g_{-j,-k}\}, \boldsymbol{\eta}, \boldsymbol{Y})$, $f(a_j \mid \{a_{-j}\}, \boldsymbol{\eta})$, and $f(\eta_i \mid \boldsymbol{a}, \boldsymbol{g}, \boldsymbol{Y})$ up to proportionality constants.

6) Fit a Bayesian latent class model with three classes ($M^{fit} = M_0 = 3$), using your simulated data, and the priors specified in 5). Obtain the sequence of values for each parameter that are drawn from the posterior, $\left\{p_{jk}^{(t)}\right\}_{t=t_0}^{t_1}$, $\left\{\pi_j^{(t)}\right\}_{t=t_0}^{t_1}$, $\left\{\eta_i^{(t)}\right\}_{t=t_0}^{t_1}$, $j = 1, \ldots M^{fit}, k = 1, \ldots, K, i = 1, \ldots, N$, where $t_0$ and $t_1$ are the indices of the start and end of your sampling chain, respectively. (*Note*: you may use JAGS, WinBUGS and call them from R. **You must submit your code as well.**)

7) Visualize/Plot your estimated posterior distributions: $f(p_{jk} \mid \boldsymbol{Y}, M^{fit} = 3)$, $f(\pi_j \mid \boldsymbol{Y}, M^{fit} = 3)$, $P(\eta_i = j \mid \boldsymbol{Y}, M^{fit} = 3), j = 1, \ldots, M^{fit}, k = 1, \ldots, K, i = 1, \ldots, N$. (*Hint*: compare the estimated posteriors with the true parameter values that were used to simulate the data $D^*$. For the posteriors of the individual class indicators $\{\eta_i\}$, just randomly choose 4 individuals.)

8) At each iteration from the kept sampling chain, $t = t_0, \dots, t_1$, simulate one data sets $D^{(t)}$ with 300 subjects following the latent class model with parameters, $\left\{p_{jk}^{(t)}\right\}_{j=1,k=1}^{M^{fit},K}, \boldsymbol{\pi}^{(t)}, \left\{\eta_i^{(t)}\right\}_{i=1}^{N}$; Compute the all the finite-sample-based pairwise log odds ratios from $D^{(t)}$ and denote it by $\{\hat{\psi}_{k,k'}^{(t),N}\}$. Compare the set of values $\{\hat{\psi}_{k,k'}^{(t),N}\}$ to $\hat{\psi}_{k,k'}^{obs,N}$, for each pair $(k, k')$. What do you see? (Note: you may choose a few interesting pairs $(k, k')$ to demonstrate what you find.)

9) Repeat 5) to 8) for $M^{fit} = 2, 4$. Summarize your results. (Note: you may choose a few interesting pairs $(k, k')$ you used in 8) to demonstrate what you find.)

10) Summarize your experience with this simulation study of latent class model, e.g., what's the statistical mechanism that gives rise to the dependence among symptoms (can refer to the DAG), or do we have evidence in the data about the true number of classes, etc.