PLSC 40502: Problem Set 3

February 19, 2025

This problem set is due at **11:59 pm on Wednesday, March 5th**.

Please upload your solutions as a .pdf file saved as "Yourlastname_Yourfirstinitial_pset3.pdf"). In addition, an electronic copy of your .Rmd file (saved as "Yourlastname_Yourfirstinitial_pset3.Rmd") must be submitted to the course website at the same time. We should be able to run your code without error messages. In addition to your solutions, please submit an annotated version of this `.rmd` file saved as "Yourlastname_Yourfirstinitial_pset3_feedback.rmd", noting the problems where you needed to consult the solutions and why along with any remaining questions or concerns about the material. In order to receive credit, homework submissions must be substantially started and all work must be shown.

Problem 1

In this problem, we will be working with data from the 2023 Cooperative Election Study (CES) to understand patterns in issue positions and preferences among voters. The code below loads in a subset of the questions asked to 24,500 respondents in the "Common Content" questionnaire and applies an initial conversion of the columns with the "labelled" type from a numeric code to a factor variable.

```
ces <- read_dta("data/CES23_Common_subset.dta") %>% mutate(across(where(is.labelled), as_factor))
```

Below are the demographic columns. You will not be using these directly in the model, but they may be useful for diagnostics and exploration in the latter parts of the problem.

- `gender4` - Four category gender variable (Male, Female, Non-binary, Other)
- `birthyr` - Birth year
- `educ` - Educational attainment
- `race` - Race
- `hispanic` - Whether respondent is Hispanic
- `pid3` - Three-category party ID
- `pid7` - Seven-category party ID
- `ideo5` - Five category ideology

For the mixture model, you'll be using the 17 following responses related to policy attitudes. Each of these is a "support/oppose" response to a particular policy policy proposal. Those questions are listed below

- `CC23_321a` - Ban assault rifles
- `CC23_321c` - Require criminal background checks on all gun sales
- `CC23_321d` - Increase the number of police on the street by 10 percent, even if it means fewer funds for other public services.
- `CC23_323a` - Grant legal status to all illegal immigrants who have held jobs and paid taxes for at least 3 years, and not been convicted of any felony crimes
- `CC23_323b` - Increase the number of border patrols on the US-Mexican border
- `CC23_323c` - Build a wall between the U.S. and Mexico
- `CC23_324a` - Always allow a woman the right to obtain an abortion as a matter of choice

- `CC23_324d` - Expand access to abortion, including making it more affordable, broadening the types of providers who can offer care, and protecting access to abortion clinics.
- `CC23_326a` - Give the Environmental Protection Agency power to regulate carbon dioxide emissions
- `CC23_326b` - Require that each state use a minimum amount of renewable fuels (wind, solar, and hydroelectric) in the generation of electricity even if electricity prices increase.
- `CC23_326d` - Increase fossil fuel production in the U.S.
- `CC23_328a` - Relax local zoning laws in your state to allow for construction of more apartments and condos.
- `CC23_328b` - Expand federal tax incentives to encourage developers to build homes for people who make less than half of the average income in your area
- `CC23_328c` - Require able-bodied adults under 64 years of age who do not have dependents to have a job in order to receive Medicaid.
- `CC23_328d` - Repeal the Affordable Care Act
- `CC23_328e` - Expand Medicaid to cover individuals making less than 25,000 and families making less than 40,000 a year.
- `CC23_328f` - Forgive up to 20,000 of student loan debt for each person

Note that this is a subset of the full set of questions asked in the CES. Here, we've tried to remove questions that are logically dependent on each other (e.g. always allowing abortion is mutually exclusive with making abortion illegal in all circumstances) to make the subsequent modeling assumptions a bit more plausible

Part A

To begin, convert the responses into binary indicators that take on a value of 1 if the respondent supports the policy and 0 if they do not support the policy. Code other responses (e.g. "skipped") as missing data (`NA`). Subset the sample down to those respondents who answered Support or Oppose to each of the 17 policy questions (no missing responses) - you will be using this dataset for the remainder of the problem. How many observations do you have in this dataset?

Part B

Assume the following data-generating process. We observe N respondents each giving binary responses to J questions (here, $J = 17$). Y_i denotes the J-length vector of responses for unit i with components Y_{ij}. Assume that each respondent has a latent cluster indicator z_i drawn from a categorical distribution with K categories and parameter π which lies on the simplex. $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \in (0, 1) \forall k \in \{1, 2, \ldots, K\}$

$$z_i \sim \text{Categorical}(\pi)$$

Conditional on z_i, each of the j responses for unit i is independently distributed Bernoulli with cluster-and-question-specific parameter $\beta_{k,j} \in (0, 1)$

$$Y_{ij} | z_i = k \sim \text{Bernoulli}(\beta_{k,j})$$

Our targets of inference are the $K \times J$ matrix of cluster-and-question-specific response probabilities β and the baseline K-length cluster prevalence parameter π

Write the log-likelihood $\ell(\beta, \pi; Y)$.

Part C

Write the "complete data" log-likelihood $\ell(\beta, \pi | Y, Z)$. Simplify as much as possible.

Part D - The "E-step"

Find the conditional distribution $p(Z_i|Y_i, \beta, \pi)$.

Write a function in R that takes as input the observed Y, a given number of clusters K, and values of the parameters β and π and returns an $N \times K$ matrix of "responsibility parameters" $\gamma_{i,k}$. That is, the conditional probability that each observation belongs to each cluster given Y and parameters β and π.

Part E - The "M-step"

Write down the Q-function $Q(\beta, \pi|\beta^{(t)}, \pi^{(t)})$. That is, find the expectation of the complete data log-likelihood from Part C, taking the expectation over the latent variables Z with respect to the conditional distribution from Part D evaluated at the current iteration's value of the parameters.

Hint: The current values of the parameters $\beta^{(t)}$ and $\pi^{(t)}$ enter into the Q-function only through the "responsibility parameters" from the previous iteration $\gamma_i^{(t)}$, so you can write the function just in terms of $\gamma_i^{(t)}$ and the parameters to be optimized (β and π).

Find the values of β and π that maximize the Q-function (you should get a straightforward expression for each in closed form). Implement a function in R that takes a matrix of responsibility parameters $\gamma^{(t)}$ and returns these values $\beta^{(t+1)}$ and $\pi^{(t+1)}$

Part F

Write a function to implement the EM algorithm in R to obtain maximum likelihood estimates of the cluster means $\hat{\beta}$ and cluster prevalence $\hat{\pi}$. Your function should take as input a data matrix Y that consists of N rows (observations) and J columns (questions). It should run the EM algorithm until convergence. Assess convergence by comparing the change in the per-observation log-likelihood from iteration-to-iteration. You should use a convergence tolerance of 1×10^{-6} (that is, if the absolute per-observation log-likelihood does not change by more than this amount, your algorithm has converged). Return the MLE estimates of β and π, the matrix of responsibility parameters γ for each observation computed at the MLEs of β and π, and the value of the maximized log-likelihood.

Implement the mixture model using two clusters: $K = 2$. Set the random seed to 60637. Run the algorithm 10 times using random initializations and select the solution that results in the highest log-likelihood. Provide a substantive interpretation of the clusters that you obtain. Make a histogram of the estimated responsibility parameter for cluster 1 across respondents in your data. Comment on what you find.

Part G

Next, implement the mixture model with five clusters $K = 5$. Again set the random seed to 60637 and use 10 random initializations. Provide a substantive interpretation of each of the clusters. Feel free to use the provided covariates to help you characterize each of these latent types).

Calculate the prevalence of each of the five clusters accounting for sample selection by using the provided CES weights (`commonweight`) (take the weighted means of the responsibility parameters across observations in the sample). Compare these to the unweighted prevalences. What types of voters are over-represented in the sample; what types are under-represented?

Part H

Now you'll compare your cluster model to a standard 2-parameter probit IRT model. We assume that the probability that respondent i answers "Support" to question j is

$$Pr(Y_{ij} = 1) = g^{-1}(\beta_j \theta_i - \alpha_j)$$

where g^{-1} is the inverse normal CDF function, α_j and β_j are the "difficulty" and "discrimination" parameters for question j and θ_i is the "ideal point" of respondent i.

The `emIRT` library allows for fast estimation of item response theory models using the EM algorithm. Install and load this library.

```
library(emIRT)
```

Use the `binIRT()` function to estimate the two-parameter probit model on the survey responses. Read the documentation for this function carefully and code the data matrix appropriately. You'll also need to make an initial matrix of starting values and priors using the `getStarts()` and `makePriors()` functions. See the example for `binIRT` for more information. Set the seed to 60637 prior to generating the (random) starting values.

Plot a histogram of the posterior means of the ideal points θ for respondents in the sample. Interpret the latent dimension θ. On which questions does this parameter provide high discrimination between responses? On which questions does this parameter provide low discrimination between responses?

Part J

Using your responsibility parameter estimates from Part G, assign each respondent to the cluster which is "most responsible" for that observation (that is, "hard cluster" your respondents). For each of these clusters, plot a histogram of the ideal point estimates from part H. Comment on your findings. What have you discovered?