Modern Clustering Analysis

Question Posed by Amy Herring to Arijit Dey

Readings

General Knowledge of (Bayesian) Clustering Methods

A wide variety of (Bayesian approaches) for clustering are available, and in the course of the independent study, the student will explore the literature and identify methods of interest. A few approaches to help get started include those described in the papers linked below.

- Introduction to Bayesian clustering
- Bayes mixture models
- Dombowsky clustering paper
- Bayesian pyramids

Preliminary Exam Independent Study

Note: the take-home portion of the prelim exam should be provided via an R markdown or Quarto file in a GitHub repository for purposes of reproducibility. The repository should include the source data and R code for processing, analysis, and preparation of the independent study report.

Goals/learning objectives include the following.

1. Gain valuable experience with biomedical data wrangling skills, including data cleaning, merging, transforming, and validation of data in standard flat file format. Explore potential impact of features such as missing values and distributions of variables on analysis results.
2. Implement reproducible workflow for data input, transformation, and analysis and document decisions and trade-offs made during data cleaning and preparation for downstream clarity.
3. Strengthen applied data analysis skills through open-ended problem-solving in a "real world" (i.e., messy) data setting.
4. Practice identifying advantages and drawbacks of standard and cutting-edge methods for clustering of data at fixed time points and longitudinally.

The National Health and Nutrition Examination Study I: Epidemiologic Follow-Up Study (NHEFS) is a nationally-representative study of US adults involving longitudinal data on behaviors and health conditions. NHEFS is a longitudinal study following adults aged 25 to 74 years enrolled in the first National Health and Nutrition Examination Survey (NHANES I). NHEFS involves a series of follow-up surveys. The first wave of follow-up data collection, the 1982-1984 NHEFS (ICPSR 8900), included all participants who were between 25 and 74 years of age at their NHANES I exam. The second wave, the 1986 NHEFS (ICPSR 9466), included the NHEFS cohort who were 55-74 years at their baseline examination and not known to be deceased at the time of the first wave. The third wave, the 1987 NHEFS, was conducted for the entire non-deceased NHEFS cohort. Thus data are available at baseline (NHANES I) and three subsequent follow-up occations (NHEFS waves). The Inter-university Consortium for Political and Social Research (ICPSR) has made these data available in flat ASCII files for public use.

The primary questions of interest are what patient clusters exist at each time point, and whether patient clusters vary over the four survey times.

1. Data processing, preparation, and exploratory analysis:
    a. Download the NHANES I data.
    b. Download the NHEFS data from Michigan's IPCSR. Separate files are available for the 1982-1984 data collection wave, the 1986 wave, and the 1987 wave.
    c. Before analyzing the data, we will need to subset each file and then combine across data files. We will only extract selected variables from each file (using for example read_fwf()), and records can be linked using the sample sequence number (SEQNO). Note that some questions are only asked of participants based on their participation at a prior wave, so check documentation carefully to be sure you have the most complete data (there will be some missing data regardless). First, identify relevant variables in each of the 4 waves (NHANES I and the three NHEFS waves) to measure the following.
        - Basic demographics: sex, age, marital status, and vital status (measurements after a patient's death, supplied by a proxy reporter, should not be included, so at each wave we wish to limit to live patients, e.g. using TRACSTAT from the 1986 wave)
        - Basic biometrics: weight and height
        - Health behaviors: current smoking and current alcohol use
        - General health (e.g., GENERAL from 1986, HEALTH from 1986 measuring past 12 month health)
        - History/occurrence of the following medical conditions (note the documentation – once a patient reports a condition, say in 1982, then the subsequent surveys just ask about additional events, so you'll use variables across surveys to determine occurrence of these conditions). From these you just want to know at each wave whether they've had each event, and if so, whether they've had additional subsequent events.
            – arthritis
            – heart attack
            – coronary bypass surgery
            – small stroke (e.g., TIA)
            – stroke (e.g., CVA)
        - Whether or not participant is currently taking medication for diabetes; whether or not they're currently taking medication for high blood pressure
    d. Present exploratory data analysis for these variables at each time. Note the fraction of data points that are missing at each time for each variable. Provide a "Table 1" that describes the data. Discuss any potential challenges for subsequent analysis based on this exploratory data analysis, including (but not limited to) whether the same type of information is available for each variable at each time point.
    e. What number (and fraction) of participants provided data at all four occasions? What variables from baseline (above) are predictive of missing data in one of the NHEFS waves? What implications might missing data have on subsequent analysis?
2. Clustering Analysis and Interpretation
    a. Cluster the baseline data (NHANES I) using multiple methods, including k-means clustering, a standard Bayesian mixture model (e.g., as implemented in Dombowsky et al.), the CLAMR approach proposed by Dombowsky et al., and the Bayesian pyramids approach of Gu et al. Discuss any challenges in implementing each method to the data, and provide evidence of how similar

or different the clustering solutions are across methods. Which features are most influential in determining cluster membership? How do the features vary across clusters?

b. Explore the relationship between baseline clusters and subsequent mortality. Can certain clusters be characterized as higher risk than others? If so, which ones? Do the clustering approaches vary in their ability to predict death in later waves?

c. Now repeat the exercise in part 6 separately for each of the NHEFS follow-up waves. Discuss how similar/different the clusters are across waves. (Hint: it may be helpful to subset to a common group of participants/times to make comparisons more interpretable.)

d. Enumerate several challenges one could foresee in describing how clustering changes over time and address the extent to which these may be accommodated using the existing literature.

e. To what extent do you believe meaningful clusters exist in these data? Provide evidence from the data to support your response.