STAT 207, Spring 2022
Takehome Quiz # 1
Released April 27. Due May 1 on Canvas

Instructions:

- Remember to use the template provided on the Canvas web page.
- Start your paper with an abstract that contains a short description of the problem and the main findings. Then, the first part of the body of the paper will correspond to an introduction with a description of the problem and an exploratory data analysis. The methods and the analysis will follow. Please organize and present the materials in the best possible way. Be informative but concise. There is no page limit to the paper, but ideally you should aim for no more than 10 pages (including any figures and tables).
- The paper will finish with concluding remarks and references (if any). Tables and figures, if any, need to be part of the text. Do not append them to the end of the paper. Please append the codes at the end of the report or as a separate file. Please make them tidy and include minimal comments. Your score of the quiz does not take into consideration of the codes, but I may read them to better understand the report.
- The purpose of this 'quiz' is not to identify groundbreaking findings, but to provide a solid analysis of a real dataset and identify potential areas of advantages/disadvantages/caveats of the modeling approach. The focus of the grading will be on both the correctness of the implementation, and communication and presentation of the data, results, and findings.

The file **SimMovieRating.csv** contains simulated movie rating data from 20 users on 30 movies. The ratings are in the range of 1 to 10. The file **SimMovieCovariates.csv** contains three binary covariates describing each movie: whether the movie is an action movie, whether the movie is a romance movie, and whether Keanu Reeves is a cast member of the movie. In this take-home quiz, you will analyze this simulated movie rating data using a Normal-Inverse-Wishart model.

1. Using quantitative and graphical tools to summarize the main features of this dataset. Describe any patterns in the ratings that you see from the exploratory data analysis.
2. Consider the ratings matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times p}$ for n raters and p movies. Since the raw ratings $\tilde{X}$ are on the scale from 1 to 10, let us pre-process the data matrix so that it has 0 mean. Denote the de-meaned rating matrix as $\boldsymbol{X}$.
3. Consider the model

$$X_{ij} \sim_{ind} N(\boldsymbol{u}_i^T \boldsymbol{v}_i, \sigma^2), \quad i = 1, ..., n, j = 1, ..., p$$

where $\boldsymbol{u}_i \in \mathbb{R}^K$ is a K dimensional latent user feature vector, and $\boldsymbol{v}_i \in \mathbb{R}^K$ is a K dimensional latent movie feature vector. That is, we consider the expected ratings can be decomposed into user and movie effects.

We put independent Gaussian priors on $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ so that

$$\boldsymbol{u}_i|\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u \sim_{ind} N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u), \quad i = 1, ..., n$$

$$\boldsymbol{v}_i|\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v \sim_{ind} N(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v), \quad j = 1, ..., p$$

For the hyperpriors, we let

$$\boldsymbol{\mu}_u|\boldsymbol{\Sigma}_u \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_u/\kappa_0), \quad \boldsymbol{\Sigma}_u \sim \text{InvWishart}(\nu_0, \boldsymbol{\Lambda}_0)$$

$$\boldsymbol{\mu}_v|\boldsymbol{\Sigma}_v \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_v/\kappa_0), \quad \boldsymbol{\Sigma}_v \sim \text{InvWishart}(\nu_0, \boldsymbol{\Lambda}_0)$$

Finally, we let $p(\sigma^2) \propto 1/\sigma^2$. Write out the steps of a Gibbs sampler to analyze the data.

4. Let $\boldsymbol{\mu}_0 = 0, \nu_0 = p + 1, \boldsymbol{\Lambda}_0 = \boldsymbol{I}, \kappa_0 = 1$. Implement the sampler for $K = 1$ and visualize the posterior means of the $p \times 1$ matrix $\boldsymbol{V}$.

5. Comparing the matrix $\boldsymbol{V}$ with the movie covariates, can you provide potential interpretation of the latent movie feature matrix $\boldsymbol{V}$?

6. Suppose now we have one new user. Without knowing any previous movie watching records of this user, we consider $\boldsymbol{u}_{n+1}$ follow the same distribution as $(\boldsymbol{u}_1, ..., \boldsymbol{u}_n)$ independently. Describe how you can draw samples from the posterior predictive distribution of $\tilde{X}_{i,1}$ and $\tilde{X}_{i,15}$, i.e., the predicted rating of movie 1 and 15. Visualize this bivariate distribution using the posterior samples you obtained. *For the question, please report your results on the original data scale instead of the de-meaned scale.*

7. Repeat the question 4 to 6 with $K = 2$. Compare your model results to the previous results with $K = 1$. Describe any notable differences (and you do not need to restrict yourself to the aspects asked in the previous questions).

8. Provide a critical review of your findings. This can be a discussion of any difference (or lack of difference) between different models/exploratory analysis; any limitations you see in the models; potentially better modeling approaches that you do not have time to explore; etc. You may discuss in the broader context of predicting user ratings beyond this particular dataset.