STAT 656: Bayesian Data Analysis
Fall 2023
Homework 2

Note: For each question, no credit will be given unless work is shown.

Synthetic data (100 points)

The file `hw2_synthetic.csv` is a dataset of count-valued measurements $\mathbf{y} = \{y_1, \ldots, y_n\}$, with $y_i \in 0, 1, \ldots$. Each output y_i has an associated $x_i = (x_{i,1}, x_{i,2}) \in \mathbb{R}^2$, and write $\mathbf{x} = \{x_1, \ldots, x_n\}$'s as $\mathbf{x}$. We model y_i as

$$y_i \mid \beta \sim \text{Poisson}(e^{f(x_i,\beta)}).$$

Here, the exponential is to ensure the Poisson rate is always positive, and the function $f(x_i, \beta) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}^2 + \beta_4 x_{i,2}^2 + \beta_5 x_{i,1} x_{i,2}$

1. (25 points) With the provided data, perform a Bayesian analysis on the parameters of the model above to decide which terms in the expression for $f(x)$ you think are important. State clearly what your prior over β is, and how you arrived at your conclusion, including any useful figures (especially of the posterior distribution). You can use Stan.

2. (25 points) Having decided which terms in f are important, keep only those and discard the rest, resulting in a possibly simpler model. Now perform a Bayesian analysis over the parameters of this model. Note that you are using the data twice, once to select the model and next to fit the it, but we will not worry about that. Compare the posteriors for both models.

3. (25 points) Perform posterior predictive checks for both models, being sure to explain what you are doing. Which model do you think fits the data better?

4. (25 points) Use the results from the second model to create a contour plot showing the log average Poisson intensity as a function of x. In other words, plot $\log \mathbb{E}_{p(\beta|\mathbf{x},\mathbf{y})}[\exp(f(x, \beta)]$ as a function of the two components of x (you can restrict the component ranges from -10 to $+10$).

Applied problem (100 points)

Scientists at the Notlem lab are researching the conversion of stem cells into pancreatic β-cells to treat patients with type 1 diabetes. They recently developed two new chemical modulators for improved conversion, and wish to identify their concentration settings that maximize conversion.

You are employed as a consultant for the lab. In an initial meeting with the Notlem scientists, you were informed that they intend to conduct their study in the following manner:

1. Place stem cells in each well of a 24-well rectangular plate.

2. Assign pairs of concentrations for the two chemical modulators to each well in the manner you specify (you get to specify 24 concentration pairs per plate).

3. Incubate the entire plate, and measure the expression level of a specific protein for each well (giving you 24 measurements). The expression-level resides in $\mathbb{R}$, with larger values signifying better conversion.

The scientists also state that

- their resource constraints limit the number of plates to 3. Since each plate has 24 wells, you can get at most 72 measurements.
- their time constraints limit the number of experimental runs to 2. Thus you can tell the scientists 24 pairs of concentrations followed by 48, or you can tell them 48 pairs followed by 24. In either case, you can wait for their measurements for the first set before choosing the concentrations for the second.
- you should only consider concentrations between 0 and 80 for the first chemical modulator, and between 0 and 30 for the second chemical modulator, because anything outside this region is not thought to improve conversion, and that
- polynomial models of conversion as a function of chemical modulators' concentrations (e.g. $f(x, \beta)$ like the previous question, but possibly with cubic or even higher terms) have been shown to enjoy some measure of success for this type of problem in previous literature. (Unlike the earlier question, here the measurements are real valued, so you don't need the exp and Poisson parts, just Gaussian noise.)

After you select a design, you can upload it to `Brightspace` as a CSV file. Any design strategy you decide to adopt must satisfy the scientists' resource constraints. Again, please be sure to note that an entire plate will be incubated at one point in time. Thus, your only possible design strategies are to upload

- two files at separate times, the first containing the initial 24 runs, and the second, the final 48 runs, or
- two files at separate times, the first containing the initial 48 runs, and the second, the final 24 runs.

The deadline to submit the first design is 1159pm Sept 22, and the deadline for the second is 1159pm Sept 29, else your request will be ignored and you will only have partial information

See `sample_design.csv` for a sample design CSV file with 24 design points. Please follow this format exactly (if you generate it with R, use the `write.table` command with the `col.names` and `row.names` options set to `FALSE`). The scientists will e-mail you the results in an augmented CSV file. No credit will be given unless you provide all your reasoning, computations, and discussion of results in a concise and coherent manner.

Your task is to (you will do this twice, after each response from the scientists):

- tell the scientists the specific pairs of concentrations to run in the study, explaining your thinking
- analyze the resulting (cumulative) data to build a Bayesian regression model of stem cell conversion as a function of the two chemical modulators' concentrations,
- use the model to create a contour plot of the posterior predictive mean of conversion as a function of concentration settings,
- use the model to calculate the posterior predictive distribution of the concentration settings that yield maximum conversion, and finally
- construct the posterior predictive distribution of the conversion corresponding to a specific point prediction of the concentration that yields maximum conversion.

Feedback (optional)

Brief comments on each of these points would be greatly appreciated.

(a) Do the instructors present material at an adequate pace during lecture (too slow/too fast)?

(b) What general material would you like the instructors to spend more time on?

(c) Which topics/ideas/concepts in lecture were not well-explained? Brief comments are appreciated.

(d) Any further comments/questions/feedback?