ELL880: Major Test

May 2, 2023

Maximum Marks: 30

Instructions:

- **Please clearly indicate the question number, and part number if applicable, at the start of each response.**
- **Please read all questions carefully.**
- **Please ensure that your responses are to-the-point and that you write only what is asked for on the answer script you submit.**
- **While the exam is open-notes, all your answers must be written entirely in your own words, without any copying from anywhere.**
- **Please try to be clear and careful with all mathematical notation, so that there is no ambiguity in the expressions/formulae you write down. Try to stick to the kind of notation used in class as far as possible.**

1. Consider the `Kline` data set on diversity of tool types in different societies discussed in class. For each society, the data set contained its population P, contact level C (low/high), and count of tool types T. In the dynamic innovation/loss model, we had assumed an innovation rate that was independent of the current number of tool types the society has. But it might also be the case that having more tools itself increases the innovation rate; let us now work with this assumption.

   (a) Write down a reasonable differential equation model for the rate of change of tool types as a function of time, incorporating the above assumption in a suitable fashion (and also retaining population and contact level as predictors). If you are making any additional assumptions, please state them clearly and try to justify them. Clearly indicate the meaning/interpretation of all parameters in your model. **[4]**

   (b) Solve the above equation for the steady state value of the number of tool types. **[1]**

   (c) Now write out a full statistical model (including priors) for fitting the values in the Kline data set to the equilibrium model obtained in part (b). Try to choose suitable priors for all parameters. **[2]**

2. Consider the trolley problem data set discussed in class. Suppose we model the response R (on a scale of 1 to 7) as a function of the binary action A, intention I, and contact C features as in class; but now we also want to add in the *story type* S as a predictor of the response, with partial pooling (multilevel modelling). There are 12 different story types in the data set. We need not consider any other predictors here.

   (a) Draw the DAG showing the assumed relationships between all the variables being considered here. **[1]**

   (b) Write out a full statistical model corresponding to the DAG, suitably incorporating the story type with partial pooling. Clearly indicate the meaning/interpretation of all parameters in your model, and specify reasonable priors on all of them. In particular, indicate which parameter will reflect/control the degree of pooling of the story types. **[4]**

   (c) Now suppose we also sought to include gender of the participant as an additional predictor in the above model. Would it be reasonable to infer that any observed effect of gender on response in such a model could be interpreted as a causal effect? Explain why or why not. **[2]**

3. Consider a data set of the outcomes of applications for research grants, which records the following variables: gender of the applicant G, academic discipline in which the grant was submitted D, and outcome of the grant F (funded or not).

(a) Draw a DAG to represent the plausible relationships between these variables. **[1]**

(b) As per the backdoor criterion, which variables should be stratified by to estimate the *total* causal effect of gender on obtaining grant funding? **[1]**

(c) Write down a *binomial* GLM (i.e., modelling *counts* of different outcome types in different scenarios) for estimating this total causal effect, similar to what we did for the Berkeley admissions data. Ensure that the meaning/interpretation of all variables and parameters in your model is clearly indicated, and priors are specified on all parameters. **[3]**

(d) Now modify the above model in order to estimate the *direct* causal effect of gender on funding outcome, justifying your model with reference to the DAG drawn in part (a). **[2]**

(e) Suppose there is a general perception that some academic discipline (let it be D=1) discriminates against women in awarding grant funding. Explain how the anticipation of such discrimination might lead to it being masked in the inference of the direct causal effect as in part (d). Draw a corresponding DAG to illustrate this situation and show what kind of confounding may occur. **[3]**

4. Suppose we wish to do Gaussian Process regression for time-series data, where each data point is some continuous outcome variable V at a given time point T. Write down a full statistical model for this purpose, using a quadratic kernel function. Ensure that the meaning/interpretation of all parameters and variables in your model is clear, and that all priors are specified. **[6]**