# ST440/540 – Exam 2 - Due April 17

THIS IS AN EXAM - DO NOT DISCUSS THE PROBLEM WITH ANYONE (INCLUDING OTHER STUDENTS OR THE TA)! If you have questions, please email me.

The data you will analyze are from the paper (Supplemental Materials 2) Cretaceous climates: Mapping paleo-Koppen climatic zones using a Bayesian statistical analysis of lithologic, paleontologic, and geochemical proxies. The data are on the course website. This paper uses various proxies to estimate the climate millions of years ago. We will analyze only mean annual temperature (MAT). Information about MAT is inferred from fossil records. For example, if a 100M year old alligator bone is found in Mexico, and alligators only live in areas with MAT in a certain range, then we can conclude the MAT was within this range in Mexico 100M years ago.

All of the observations in the dataset have two locations, current (i.e., the lat/long where the proxy was found) and paleo (i.e., where the current lat/lon was hundreds of millions of years ago); we will use only paleo latitude for this analysis and it is denoted $X_i$ for sample $i$. Depending on the type of proxy (animal, plant fossil, etc) there are two types of observations: quantitative ("Temperature, C") and interval("Min Temp" and "Max Temp"). For **quantitative** observations a best guess at the MAT (degrees Celcius) is provided and denoted $Y_i$. If observation $i$ is an **interval observation** then all we know is that MAT was in the interval $(l_i, u_i)$.

1. **Quantitative data only**: First use only quantitative data. Your objective is to estimate MAT as a function of paleo latitiude in each of nine Cretaceous time slices (Berriasian/Valanginian, ..., Maastrichtian). Let $f_i \in \{1, ..., 9\}$ denote the time slice for observation $i$. For an observation from time slice $k$ (thus $f_i = k$) let $\text{E}(Y_i) = g_k(X_i)$ and the objective is to estimate $g_1, ..., g_9$. For simplicity, assume $g_k(X)$ is a quadratic function of $X$ for all $k$. You should use all the data in one model fit but allow for different parameters by time slice.

   (a) Write a Bayesian statistical model for these data including prior distributions. Argue that this model is a reasonable choice for this analysis.

   (b) Use model selection methods to compare 2-3 models for the regression coefficients. Use the optimal model for all subsequent analyses (including MCMC convergence).

   (c) Verify that the MCMC algorithm has converged.

   (d) Check that the model fits the data well. If you see evidence of lack of fit, suggest (but do not implement) changes to the model that might address these deficancies.

   (e) Plot (with uncertianty) the estimated curves $g_1(X)$, ..., $g_9(X)$ as a function of $X$.

   (f) Does MAT change across the time slices? Justify your answer.

2. **Mixed data**: In this analysis, you will use both the quantitative and interval observations.

   (a) Write a Bayesian statistical model for these data including prior distributions. Argue that this model is a reasonable choice for this analysis.

   (b) Using the model for the regression coefficients selected in Question 1, fit the model and plot the estimated curves $g_1(X)$,...,$g_9(X)$ (with uncertainty).

   (c) Does adding the interval data improve your analysis?

Your paper should be written as a professional document with full sentences, clearly labeled figures and tables and few spelling/grammar errors. Organize your report with subsections corresponding to the questions above, i.e., 1a, 1b, ..., 2c. Summarize your analysis in a PDF document that is **no more than two pages long** (12 font, single space, standard margins). Append your code to the end of this document and submit a single document. **In-class students should turn in the exam in class on Monday, April 17. Online students should submit the exam on moodle.**

HAVE FUN!