ST740 – Data Analysis Project – Due 10/25

THIS IS AN EXAM, YOU MUST WORK INDEPENDENTLY! Do not discuss the exam with anyone other than the professor, including other students or the TA.

For this assignment, you will analyze climate reconstruction data (posted on the course website) from Croke et al (2021)[1] using statistical methods similar to Cahill et al (2023+)[2]. This analysis reconstructs past annual precipitation in Brisbane, AU using observed precipitation (mm), called the Rainfall Index (RFI), and six climate proxy variables. The proxy variables are from various sources such as ice cores and tree rings and have been standardized to have mean zero and variance one. The time period of interest is 1612-2017, and all seven data sources have extensive missing data during this period.

Let Y_t be the RFI in year t+1611 (so Y_1 corresponds to year 1612). The autoregressive time-series model is $Y_1 \sim \text{Normal}(\mu_0, \sigma_0^2)$ and

$$Y_t | Y_{t-1} \sim \text{Normal}(\mu_0 + \phi_0(Y_{t-1} - \mu_0), (1 - \phi_0^2)\sigma_0^2),$$

where μ_0 is the mean, σ_0^2 is the variance and ϕ_0 controls the temporal dependence. The form of the mean and variance are such that the marginal (over the Y_j for $j \neq t$) mean and variances are $\text{E}(Y_t) = \mu$ and $\text{Var}(Y_t) = \sigma_0^2$ for all t.

To impute these missing observations we use several proxy variables. Denote proxy variable $j \in \{1, ..., 6\}$ in year $t + 1611$ as X_{tj}. These variables are related to RFI using a linear relationship

$$\text{E}(X_{jt}) = \mu_{jt} = \alpha_j + \beta_j Y_t$$

for intercept α_j and slope β_j. To account for temporal dependence, we use another autoregressive model with $X_{j1} \sim \text{Normal}(\mu_{j1}, \sigma_j^2)$ and

$$X_{jt} | X_{jt-1} \sim \text{Normal}\left(\mu_{tj} + \phi_j(X_{jt-1} - \mu_{jt}), (1 - \phi_j^2)\sigma_j^2\right).$$

The Bayesian model is completed with uninformative priors $\mu_0, \alpha_j, \beta_j \sim \text{Normal}(0, 100^2)$, $\phi_j \sim \text{Uniform}(0, 1)$ and $\tau_j^2, \sigma_j^2 \sim \text{InvGamma}(0.1, 0.1)$.

1. Assuming the proxy data are complete, i.e., X_{jt} is observed for all j and t, and $\phi_0 = \phi_1 = ... = \phi_6 = 0$, derive the distribution of Y_t given $X_{1t}, ..., X_{6t}$ are discuss how the proxy data are informative about the RFI.

2. Derive the full (given all parameters and other RFI and proxy observations) conditional distributions of Y_t, X_{jt} and β_j.

3. Construct an MCMC algorthm (either step-by-step in `R` or in `JAGS` or similar program) to sample from the joint posterior. Verify that your MCMC chain(s) have converged and provide an adequate approximation to the posterior.

4. Conduct Bayesian tests, one for each proxy, that the proxy variables are associated with RFI.

[1] Jacky Croke, John V Rtkovsky, Kate Hughes, Micheline Campbell, Sahar Amirnezhad-Mozhdehi, Andrew Parnell, Niamh Cahill, and Ramona Dalla Pozza. A palaeoclimate proxy database for water security planning in Queensland Australia. Scientific Data, 8(1):292, 2021.

[2] Niamh Cahill, Jacky Croke, Micheline Campbell, Kate Hughes, John Vitkovsky, Jack Eaton Kilgallen, and Andrew Parnell. A Bayesian Hierarchical Time Series Model for Reconstructing Hydroclimate from Multiple Proxies. Environmetrics, 2023. In press.

5. Plot the observed and imputed values (which uncertainty) of Y_t from 1850-1950 and comment on the imputation.

6. Divide the time period into nine intervals: 1612-1649, 1650-1699, ... 1950-1999, 2000-2017. For interval $k \in \{1, ..., 9\}$, let $\bar{Y}_k = \sum_{t \in P_k} Y_t / n_k$, where P_k are the n_k years in period k. Plot the posterior distribution of $\bar{Y}_k$ for all k and test that $\bar{Y}_k < \bar{Y}_9$, separately for each period $k = \{1, ..., 8\}$. What do you conclude about the climate in Brisbane?

7. Idenfity 2-3 limitations of your analysis and discuss how these could be addressed in future research.

Your write-up should be written in manuscript form with complete sentences and paragraphs and carefully labelled figures and tables. Do not submit a markdown document. Include sufficient detail of your model and results so that another student in the class could reproduce your results. It should be a single PDF document. The main text is limited to 5 pages (double-spaced) including tables and figures but excluding derivations and code, which should be given as an appendix that does not count towards the 5 page limit. A large part of your grade will be on presentation (clear writting, clean figures and tables, etc) so take time to polish your work.

HAVE FUN!