MA 578 — Bayesian Statistics

Project Ideas

(Due: Tuesday, 12/10/19)

Here are some ideas for your final project. The general goal is to perform a full Bayesian analysis, including prior specification, posterior characterization, either via sampling and/or analytical derivations, posterior predictive and/or model checks, and conclusions. In some cases the goal can be more specific, such as reporting posterior intervals for parameters of interest or performing model selection, but it will invariantly require you to carefully choose hyper-priors, usually via cross-validation or empirical Bayes criteria. The outcome of the project is a short research article with: introduction, contextualizing the problem and dataset; model, with prior and likelihood; posterior derivation and/or sampling with summaries as results; and conclusions and/or discussion.

1. (Spatial linear model) Recall that one of the main motivations behind using a hierarchical linear model (HLM) is to achieve a richer covariance structure on the response $\mathbf{y}$. In this project you'll explore a specific type of structure where the correlation patterns are related to the spatial distance between locations where the data were observed. You have n observations $\mathbf{y} = [y(s_i)]_{i=1}^n$ measured at *sites* with locations s_i, and assume that

$$\mathbf{y} \,|\, \beta, \sigma^2, \tau^2 \sim N\big(X\beta, \sigma^2 I_n + \tau^2 H(\phi)\big). \tag{1}$$

   Matrix H specifies spatial correlation using a *range* parameter ϕ:

$$H(\phi)_{ij} = \exp\left\{\frac{-||s_i - s_j||_2}{\phi}\right\},$$

   where $||s_i - s_j||_2$ is the (Euclidean) distance between sites i and j. Assume a non-informative prior for β and σ^2, but try and elicit informative priors for τ^2 and ϕ. In particular, assume that $\phi \sim \mathsf{Inv\text{-}Gamma}(\alpha, \alpha)$.

   You can use the `meuse` dataset in package `sp`[1]. You'll be spatially regressing[2] log of lead concentration as a function of distance to river Meuse in meters and soil type. Here are some tasks you can tackle in this project:

   (a) Setup a hierarchical linear model version for (1), so it is simpler to obtain the posterior on β, σ^2, τ^2 and ϕ.

   (b) Design a Metropolis-within-Gibbs sampler for the parameters, with a random walk MH step for ϕ. When deriving the step for β, recall that if, for example, $\mathbf{z} \sim N(X\beta, \sigma^2 V)$, then $C^{-\top} z \sim N(C^{-\top} X\beta, \sigma^2 I_n)$ with C the Cholesky factor of V.

   (c) You have to be careful when selecting the priors on τ^2 and ϕ. You can try to fix the hyper-parameters on τ^2 to make the prior less informative, and then try to calibrate α using cross-validation.

[1] In R, `library(sp); data(meuse)`. Check the help file on `meuse` for more information about the dataset.

[2] Variances σ^2 and τ^2 are nicknamed the *nugget* and (partial) *sill*, respectively.

2. (Missing data) Consider a balanced dataset $\mathbf{y}$ that is stratified into J groups with the observations in the j-th group $\mathbf{y}_j$ having a p-variate normal distribution. Your main goal is to infer the group means θ_j and covariances Σ_j, and to this end you adopt the following hierarchical model:

$$\mathbf{y}_j \,|\, \theta_j, \Sigma_j \overset{\text{ind}}{\sim} N(\theta_i, \Sigma_j), \qquad \theta_j \,|\, \mu, T \overset{\text{ind}}{\sim} N(\mu, T), \qquad \Sigma_j \overset{\text{ind}}{\sim} \textsf{Inv-Wishart}(\nu, \Lambda^{-1}),$$

for $j = 1, \ldots, J$, and with μ and T having the usual semi-conjugate distributions. Moreover, you want to assess how your estimates for θ_j and Σ_j change as you observe missing data in $\mathbf{y}$.

You can use the classical "iris" dataset in this project[3]. Describe the posterior for the full dataset and then simulate missing data by sampling 10% of the positions in the dataset at random and re-analyzing the dataset. You can vary the proportion of missing data for a more thorough study. A few ideas to enrich your project:

   (a) You can adopt a non-informative hyper-prior for μ and T or specify their parameters using empirical Bayes or cross-validation. We have discussed a Gibbs sampler for a simpler hierarchical model in class. Each conditional posterior step should be similar to what we have for the simpler model, but you need to derive a new step to sample from Σ_j. Recall that if $W \sim \textsf{Wishart}(\nu, S)$ then $W^{-1} \sim \textsf{Inv-Wishart}(\nu, S^{-1})$, and check R's `rWishart` function.

   (b) For each pair of dimensions in the dataset (e.g., for the iris dataset, the dimensions are petal length and width and sepal length and width and so $p = 4$) plot posterior contour curves for θ_j and plot the data as points, colored by group. Compare the plots as you resample your estimates under different levels of missing data; plot the missing points differently, say, using hollow points. Also, describe the posterior distribution of the missing data and compare it to the actual values you withheld as missing.

   (c) Sample replicated data and check normality by comparing Mahalanobis distances $\sum_{j=1}^{J} (\mathbf{y}_j - \theta_j)^\top \Sigma_j^{-1} (\mathbf{y}_j - \theta_j)$. Note that these statistics should follow a χ^2_{pJ} distribution if the data are actually normal, so we can use them directly or their quantiles under the χ^2 distribution.

3. (Censored regression) Suppose that you want to model a linear regression with responses z_i, $i = 1, \ldots, n$, and predictors $\mathbf{x}_i$,

$$z_i \,|\, \beta, \sigma^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_i^\top \beta, \sigma^2),$$

but you actually only know that $l_i \leq z_i < u_i$, that is, $\mathbf{z}$ is *censored*. Censoring thresholds l_i and u_i do not vary for each observation but are usually predetermined, say, during survey design, to belong to a number of choices. These choices are specified by censoring class y_i, so $l_{y_i} \leq z_i < u_{y_i}$. The parameters of interest here are still β and σ^2, as in a regular regression, but now we have the missing z_i as nuisance parameters.

[3] In R, `data(iris)`.

For this project, you can use the `Affairs` dataset from package `AER`. In this classic dataset[4], censored response `affairs` counts the number of extramarital affairs per year and it can have six values, as seen below; all other variables are predictors.

y_k	l_k	u_k
0	0	1
1	1	2
2	2	3
3	3	4
7	4	12
12	12	∞

(a) Design a Gibbs sampler to estimate the posterior distribution of β and σ^2. Note that when sampling z_i you have a *truncated* normal distribution; check package `truncnorm` and function `rtruncnorm`.

(b) Since we have counts as the response, a more natural model would be $z_i \,|\, \beta \overset{\text{ind}}{\sim} \mathsf{Poisson}(\mu_i(\beta))$ with $\log \mu_i(\beta) = \mathbf{x}_i^\top \beta$ (canonical link). Note that $z_i = y_i$, uncensored, if $y_i \in \{0, 1, 2, 3\}$ since z_i is now integer. Design a new Gibbs sampler for this truncated log linear regression. For the conditional step $\beta \,|\, \mathbf{z}, \mathbf{y}$ you can use a Metropolis-Hastings step with the second order Taylor expansion random walk proposal we discussed in class.

(c) For both censored regressions, normal[5] and Poisson, summarize the posterior distribution of coefficients and conduct posterior predictive checks. In particular, comment on estimated values and which predictors are more relevant, and discuss which model best explains the data.

4. (Bivariate Poisson distribution) In many applications we observe pairs of counts that are correlated. For example, this is often seen in team sports, where competition drives the correlation: if one team scores, the other team is more motivated to strike back. This need to model correlation motivates the definition of a bivariate Poisson or *Holgate* distribution: if $\widetilde{X} \sim \mathsf{Po}(\alpha - \delta)$, $\widetilde{Y} \sim \mathsf{Po}(\beta - \delta)$, and $U \sim \mathsf{Po}(\delta)$ are independent, then $X = \widetilde{X} + U$ and $Y = \widetilde{Y} + U$ follow a Holgate distribution with parameters α, β, and δ, denoted by $X, Y \sim \mathsf{Holgate}(\alpha, \beta, \delta)$. In this case, X and Y are marginally Poisson distributed with means α and β respectively, but $\mathrm{Cov}(X, Y) = \delta$. It can be shown that, with $\psi = \delta/[(\alpha - \delta)(\beta - \delta)]$,

$$\mathbb{P}(X, Y) = (\alpha - \delta)^X (\beta - \delta)^Y e^{-\alpha - \beta + \delta} \sum_{u=0}^{\min\{X,Y\}} \frac{\psi^u}{(X-u)!(Y-u)!u!}.$$

Suppose then that you observe pairs $x_i, y_i \overset{\text{iid}}{\sim} \mathsf{Holgate}(\alpha, \beta, \delta)$ for $i = 1, \ldots, n$. For instance, Holgate[6] discusses an interesting application of the bivariate Poisson to study accident

[4]The original paper can be found at `https://fairmodel.econ.yale.edu/rayfair/pdf/1978A200.PDF`

[5]Also known as *tobit* regression.

[6]Holgate, P., (1964) "Estimation for the bivariate Poisson distribution", Biometrika 51 (1), 241–245.

proneness. In `accidents.csv` you'll find a dataset with number of accidents sustained by shunters in two consecutive periods. To perform a Bayesian analysis, set a non-informative prior for α, β, and δ.

(a) Design a Gibbs sampler to estimate the posterior $\mathbb{P}(\alpha, \beta, \delta \,|\, \mathbf{x}, \mathbf{y})$. While we have the mass function for each observed pair in closed form above, it is easier to include nuisance parameters $u_i \,|\, \delta \overset{\text{iid}}{\sim} \mathsf{Po}(\delta)$ and note that, conditional on u_i, $x_i - u_i \,|\, \alpha, \delta \overset{\text{iid}}{\sim} \mathsf{Po}(\alpha - \delta)$ and $y_i - u_i \,|\, \beta, \delta \overset{\text{iid}}{\sim} \mathsf{Po}(\beta - \delta)$ are independent.

(b) Based on the accidents dataset, Holgate reports maximum likelihood estimates of $\widehat{\alpha} = 1.270$, $\widehat{\beta} = 0.975$, and $\widehat{\delta} = 0.257$. Design an expectation-maximization (EM) method to estimate the posterior mode. It should be similar to the Gibbs sampler above, but, at the t-th iteration, estimating

$$u_i^{(t)} = \mathbb{E}[u_i \,|\, x_i, y_i, \alpha = \alpha^{(t)}, \beta = \beta^{(t)}, \delta = \delta^{(t)}]$$

by expectation and then optimizing for α, β, and δ given these mean estimates for u_i. After finding the posterior mode, compute a Laplace approximation of the posterior on α, β, and δ.

(c) Summarize the posterior distribution for α, β, and δ from your Gibbs sampler and compare it to the Laplace approximation, to Holgate's reported estimates, and to a naive approach with independent pairs $x_i \,|\, \alpha \overset{\text{iid}}{\sim} \mathsf{Po}(\alpha)$ and $y_i \,|\, \beta \overset{\text{iid}}{\sim} \mathsf{Po}(\beta)$. Perform posterior predictive checks for these models and report which is more adequate to describe the accidents data.