STA2311 (Fall 2023) - Homework 2

Due December 19, 2023

Instructions: This homework is to be completed using R Markdown. Each question is worth 100/3 marks and includes both theoretical and coding aspects. For each question, you must first derive the relevant mathematical formulas and typeset them in LaTeX using notation consistent with that from lecture, carefully justifying all of your steps (using complete sentences) and making references to relevant material from lecture or elsewhere (with appropriate citations). You must then implement your work in `R` in order to numerically solve the specific problem asked in the question, thoroughly commenting your code and formatting it with appropriate indentations and whitespace. Aside from base packages, you must *not* use any `R` packages except for the Tidyverse package (which is optional).

Formatting and Submission: In R Markdown, your output must be in .pdf format, rather than HTML or Microsoft Word (ugh). While the numerical inputs and outputs of your code should appear within the main text of your paper (i.e., the main text should include sentences along the lines of "We initialize the algorithm at $\alpha^{(0)} = 0.5$" and "The algorithm converges after 39493 iterations, yielding the final estimate $\hat{\alpha} = 0.043$"), your code should appear in an appendix at the end of your paper; see here for instructions. You must submit a hard copy of your .pdf to Rob, and you must submit the .Rmd file which generates your paper using the virtual assignment dropbox on Quercus. Someone running your .Rmd on another machine should be able to reproduce your document *exactly*, so remember to set seeds whenever they are appropriate.

Collaboration: While you may discuss the homework with your peers, the work you submit should be entirely your own.

1. Consider a cubic spline regression model: for $i = 1, \ldots, n$, given a covariate $x_i \in \mathbb{R}$, we observe

$$Y_i = \eta(x_i) + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and

$$\eta(x) = \sum_{j=0}^{3} \alpha_j x^j + \sum_{k=1}^{K} \psi_k (x - \gamma_k)_+^3$$

for some fixed $K \in \mathbb{N}$, where $(a)_+ := \max(0, a)$. Here the α_j and ψ_k are $\mathbb{R}$-valued coefficients, and the γ_k are $\mathbb{R}$-valued parameters called *knots*. We divide the range spanned by the observed values of x into K intervals of equal length, $I_1, \ldots, I_K$, and assume that each interval I_k contains at most one knot. We impose the following priors on the parameters in the model:

$$\begin{aligned}
\sigma^2 &\sim \text{InvGamma}(0.1, 0.1) \\
\alpha_j &\sim \mathcal{N}(0, 10), \quad 0 \leq j \leq 3 \\
\psi_k &\sim \mathcal{N}(0, 10), \quad 0 \leq j \leq K \\
\gamma_j &\sim \text{Unif}(I_j), \quad 0 \leq j \leq K.
\end{aligned}$$

Using the data in `spline.txt`, design and implement an MCMC sampler to sample from the joint posterior. Compare and contrast the results for $K \in \{2, \ldots, 5\}$.[1]

[1] Alternatively, if you are feeling adventurous, you can make K a "parameter" itself and design your own *reversible jump MCMC* algorithm to sample from the joint posterior that includes K.

2. Consider a probit regression model: for $i = 1, \ldots, n$, given a vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^p$, we observe

$$Y_i \sim \text{Bernoulli}(\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})),$$

for some vector of coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. Consider augmented data $\tilde{\boldsymbol{Y}}_{\text{aug}} = \{(Y_i, \phi_i)\}_{i=1}^n$, where $\phi_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, 1)$ is a latent variable and we only observe $Y_i = \mathbb{1}_{\phi_i > 0}$. We impose a non-informative prior on $\boldsymbol{\beta}$: $p(\boldsymbol{\beta}) \propto 1$. Some data is provided in `probitDA.txt`.

   (a) Implement a Gibbs sampler (*not* a Metropolis-within-Gibbs sampler) to sample from the posterior distribution of $\boldsymbol{\beta} \mid \boldsymbol{Y}$ by sampling from the conditional distributions of $\boldsymbol{\beta}$ and each ϕ_i using the augmented data $\tilde{\boldsymbol{Y}}_{\text{aug}}$.

   (b) Consider "parameter-expanded" augmented data $\boldsymbol{Y}_{\text{aug}} = \{(Y_i, \xi_i)\}_{i=1}^n$, where $\xi_i = \sigma\phi_i$ and σ^2 is given the improper prior $p(\sigma^2) \propto \sigma^{-2}$. Implement a Gibbs sampler (*not* a Metropolis-within-Gibbs sampler) to sample from the posterior distribution of $\boldsymbol{\beta} \mid \boldsymbol{Y}$ by sampling from the conditional distributions of $(\boldsymbol{\beta}, \sigma^2)$ and each ξ_i. To sample $(\boldsymbol{\beta}, \sigma^2)$, first sample from $\sigma^2 \mid \boldsymbol{Y}_{\text{aug}}$ and then sample from $\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{Y}_{\text{aug}}$.

   (c) Compare the performance of the two samplers using visual and statistical diagnostics.

3. Consider a simplified version of sparse linear regression: for $i = 1, \ldots, n$, we observe

$$\boldsymbol{Y}_i \overset{iid}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2 I)$$

for some vector of coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ that is assumed to be sparse. Our aim is to reduce the dimensionality of the model by removing the elements of $\boldsymbol{\beta}$ which we believe are 0. We place a prior on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ as follows:

$$\begin{aligned} \beta_i \mid \lambda_i, \tau &\sim \mathcal{N}(0, \lambda_i^2 \tau^2) \\ \lambda_i &\sim \text{Cauchy}^+(0, 1), \end{aligned}$$

where $\text{Cauchy}^+(\mu, \sigma)$ is the half-Cauchy distribution with location parameter μ and scale parameter σ. We also impose the priors $\tau \sim \text{Cauchy}^+(0, 1)$ and $p(\sigma) \propto 1/\sigma$. Using the data provided in `horse.txt` (with $p = 20$ and $n = 100$), implement an MCMC sampler to sample from the posterior distribution of $\boldsymbol{\beta} \mid (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. Based on the results of your sampler, decide which covariates should remain in the model.