Assignment 1: Gibbs samplers for conjugate Bayesian analysis

Class
Instructor: Luiz Max de Carvalho

June 6, 2023

Hand-in date: 07/06/23.

General guidance

- State and prove all non-trivial mathematical results necessary to substantiate your arguments;
- Do not forget to add appropriate scholarly references *at the end* of the document;
- Mathematical expressions also receive punctuation;
- Please hand in a single PDF file as your final main document.

  Code appendices are welcome, *in addition* to the main PDF document.

Background

Linear regression is a work-horse of modern Statistics. Its wide applicability, robustness to violations of assumptions and ease of interpretation are major assets. In this assignment we will consider Consider some data $\boldsymbol{y} \in \mathbb{R}^n$, along with a $n \times P$ matrix of covariates $\boldsymbol{X}$. Now, consider the following model for $\boldsymbol{y}$:

$$\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{A} \sim \text{Normal}\left(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{A}\right),$$

where $\boldsymbol{\beta} \in \mathbb{R}^P$ is a vector of coefficients and $\boldsymbol{A}$ is a $n \times n$ positive-definite **covariance matrix**. Note that this implies that $E_{\boldsymbol{\beta}}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with

$$\boldsymbol{\epsilon} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{A}).$$

Here, $\theta = (\boldsymbol{\beta}, \boldsymbol{A})$ are the unknown quantities in the model. We will assign the following joint prior structure to θ:

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{Normal}\left(\boldsymbol{m}_0, \boldsymbol{\Sigma}_B\right), \\ \boldsymbol{A} &\sim \text{Inverse-Wishart}(\boldsymbol{V}_0, a_0), \end{aligned}$$

where $\boldsymbol{\Sigma}_B$ is a $P \times P$ covariance matrix,$\boldsymbol{V}_0$ is a $n \times n$ scale matrix and $a_0 > n-1$ is a scalar encoding the prior degrees of freedom.

Hint: <u>If</u> you are confused about how to specify $\boldsymbol{V}_0$, a good starting point might be to assume conditional independence [1] <u>*a priori*</u>, i.e.,

$$\boldsymbol{V}_0 = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix},$$

where $\boldsymbol{w} = \{w_1, \dots w_n\}$ and σ^2 are hyperparameters that need to be specified.

Questions

1. Consider the posterior distribution

$$p\left(\boldsymbol{\beta}, \boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{y}\right) \propto f\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{A}, \boldsymbol{X}\right) \pi\left(\boldsymbol{\beta}, \boldsymbol{A}\right).$$

   Deduce the full conditional distribution of

   (a) Each component of $\boldsymbol{\beta}$, β_j;

   (b) The vector of coefficients $\boldsymbol{\beta}$;

   (c) The covariance matrix $\boldsymbol{A}$;

2. From the previous calculations, show how to sample from $p\left(\boldsymbol{\beta}, \boldsymbol{A} \mid \boldsymbol{X}, \boldsymbol{y}\right)$ using[2]

   (a) A Gibbs sampler that updates each β_j individually;

[1] Also known as assuming the errors are conditionally not autocorrelated.

[2] Remember to prove that the algorithm you are proposing is actually a Gibbs sampler!

(b) A Gibbs sampler that updates $\boldsymbol{\beta}$ as whole.

3. Discuss the theoretical characteristics of the samplers in the previous item: are they ergodic? Are they geometrically ergodic? Can you give a rate?

4. Evaluate both samplers empirically; which one would you recommend? Do your findings agree with the theory? Which diagnostics/measures of efficiency did you choose and why?

   Hint: You are going to need some data for this. If you do not have a favourite data set, consider generating a synthetic one.

5. Using the best algorithm you could construct, show how to produce samples from the posterior predictive

$$\tilde{p}(\tilde{y} \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}, \theta) p(\theta \mid \boldsymbol{y}, \boldsymbol{X}) \, d\theta.$$

   Hint: Look at this density like the expectation that it is.