Bayesian linear regression - Mathematical part

Estevão Batista do Prado

October 22, 2018

1 Introduction

Before introducing Bayesian linear regression, in this Section we present the classical linear regression model and parameter estimation via maximum likelihood in order to show the different aspects between the classical and Bayesian approaches.

1.1 Linear regression

Let $\{Y_i\}_{i=1}^n$ be independent and identically distributed random variables, $\{y_i\}_{i=1}^n$ their observed values and $\{x_{ij}\}_{j=1}^d$ the j-th explanatory variable for i. Consider $\mathbf{y} = (y_1, ..., y_n)^\top$ an $n \times 1$ column vector and $\mathbf{X} = (x_{i1}, ..., x_{id})_{i=1}^n$ an $n \times d$ design matrix containing all variables that may be associated to the response variable $\mathbf{y}$. The classical linear regression model may be written

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

$\boldsymbol{\beta} = (\beta_0, ..., \beta_{d-1})^\top$ is the $d \times 1$ vector of parameters, $\mathbf{I}_n$ an $n \times n$ identity matrix , and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$ the $n \times 1$ vector of errors. In addition, the likelihood function of $\mathbf{y}$ given $\boldsymbol{\beta}$ and σ^2 is defined as follows

$$\prod_{i=1}^n f(Y_i = y_i|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big\}.$$

In order to make inference, the $\boldsymbol{\beta}$ estimates are obtained by maximizing the likelihood function or its log. For mathematical convenience, once the parameter values that maximize both functions are the same, it is commonly utilized the log-likelihood, which is given by

$$\mathrm{ln} f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\mathrm{ln}(2\pi) - \frac{n}{2}\mathrm{ln}(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1}$$

Considering that $\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} = (\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta})^\top$ is a scalar, differentiating equation (1) with respect to $\boldsymbol{\beta}$ and solving for $\boldsymbol{\beta}$, we have

$$\begin{aligned}\frac{\partial\ \mathrm{ln} f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}} &= \frac{1}{2\sigma^2}2\mathbf{X}^\top\mathbf{y} - 2\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}\\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},\end{aligned}$$

the maximum likelihood (ML) estimator. Similarly, the ML estimator for σ^2 is given by

$$\begin{aligned}
\frac{\partial \ \ln f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\hat{\sigma}^2} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2(\hat{\sigma}^2)^2} = 0, \\
&= \hat{\sigma}^2 n + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0, \\
\hat{\sigma}^2 &= \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n}.
\end{aligned}$$

Also, there are some statistical properties such as consistency that are verified for $\boldsymbol{\beta}$. For instance, $\hat{\boldsymbol{\beta}}$ is an asymptotically consistent estimator for $\boldsymbol{\beta}$ if $P(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}| > a) \to 0$ as n $\to \infty$ for every $a > 0$. For finite samples, it is consistent when $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. That is,

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}], \\
&= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})], \\
&= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}], \\
&= \boldsymbol{\beta},
\end{aligned}$$

since $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{I}$ and $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. Furthermore, the variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}] &= \mathrm{Var}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \ \mathrm{Var}[\mathbf{y}]((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}$$

In Bayesian linear regression, the parameters $\boldsymbol{\beta}$ and σ^2 are estimated in a different way and usually through stochastic simulation methods. For instance, in the classical context the asymptotic distribution for $\hat{\boldsymbol{\beta}}$ will always be $\mathrm{N}_d(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$, since the Central Limit Theorem holds this result [1, page 244]. On the other hand, in the Bayesian context the distribution of $\hat{\boldsymbol{\beta}}$ will not necessarily be Normal, since it depends on the choice of the prior distribution.

1.2 Bayesian Linear regression

In this Section we present the Bayesian linear regression models. Also, we introduce the Bayesian perspective, prior distributions and estimation methods.

1.2.1 Basics concepts of Bayesian inference

Under the Bayesian perspective, we aim, both through data and subjective information, to draw conclusions about a certain unknown quantity of interest by using probabilistic models. Under the classical point of view, the inference is also made utilizing probabilistic models, but only the information from the data is considered and any other extra information is not incorporated into the decision process.

The inclusion of the subjective information in the inference process is the main point in Bayesian analysis. That is made by inserting a prior distribution that describes all the available knowledge that one can have about the quantity of interest, and its choice may influence the final results depending on how much data is available. That is, the more data, the less the impact of the prior information on the final conclusions [2].

There are many types of prior distributions, such as non-informative, conjugate, improper etc. The non-informative ones are based on the sampling distribution and their idea is to have a default prior distribution when there is no information about the problem at hand. For instance, the Jeffreys prior is non-informative and it is proportional to the Fisher Information, which is the expected value of the second derivative of the log-likelihood function with respect to the parameter of interest [3]. Although the Jeffreys prior is called non-informative, the Fisher Information quantify the variability of the parameter based on the available data. That is, the higher the value of the Fisher Information, the more concave is the log-likelihood, thus evidencing that the data helps to estimate the quantity of interest.

The conjugate priors are a class of distributions that present the same parametric form of the likelihood function and their choice is frequently related to mathematical and computational convenience [3]. As a consequence of conjugacy, the posterior distribution may be obtained analytically and posterior samples are generated straightforwardly. On the other hand, improper priors are distributions that, in their parametric space, do not integrate to 1. For instance, in some cases Jeffreys priors are improper, but the posterior distribution is proper; see Section 3.2 of [2].

Consider $\boldsymbol{\theta} = (\theta_0, ..., \theta_{d-1})^\top$ an unknown vector of parameters that we are interested in estimating and $\mathbf{y} = (y_1, ..., y_n)^\top$ an $n \times 1$ column vector assumed to be a realization of the random variable whose distribution is $p(y_i|\boldsymbol{\theta})$. The likelihood function of the y_i is given by

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}). \tag{2}$$

All the information from the observations y_i about $\boldsymbol{\theta}$ is included in (2). The difficulty in estimating $\boldsymbol{\theta}$ becomes an optimization problem of maximizing the likelihood function (or its logarithm). In constrast, under the Bayesian methodology the estimate of $\boldsymbol{\theta}$ is given by the joint posterior distribution, which is defined by the Bayes' theorem,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathcal{L}(\theta|\mathbf{y})p(\boldsymbol{\theta})}{\int_\Theta \mathcal{L}(\theta|\mathbf{y})p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{3}$$

where Θ represents the parametric space of $\boldsymbol{\theta}$ and $p(\boldsymbol{\theta})$ the prior distribution. Equation (3) can also be written as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \mathcal{L}(\theta|\mathbf{y})p(\boldsymbol{\theta}), \tag{4}$$

since $\int_\Theta \mathcal{L}(\theta|\mathbf{y})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal distribution of $\mathbf{y}$ and does not depend on $\boldsymbol{\theta}$. The posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ provides all the information that one can have about $\boldsymbol{\theta}$. For instance, it is possible to evaluate $p(\boldsymbol{\theta}|\mathbf{y})$ and its mean, median, variance and some other quantities such as quantiles in order to have point and interval estimates. Besides, the posterior distribution frequently has no closed form, thus depending on computational methods to be obtained.

When the posterior distribution is available, one can be interested about the predictive posterior distribution, which is utilized to predict unobserved values of the response outcome, $\tilde{\mathbf{y}}$, and the marginal distribution of $\mathbf{y}$. To obtain these two distributions, the constant that was not considered in (4) is necessary.

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{y}) &= \int_{\Theta} p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \qquad \tilde{\mathbf{y}} \sim p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y}), \\ p(\mathbf{y}) &= \int_{\Theta} \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned}$$

The advantages of the Bayesian inference when compared to the classical one are that all the information available is considered and its probabilistic interpretation about the estimates is straightforward [4]. The prior knowledge is inserted through the prior distribution and combined with the data, represented by the likelihood function, all inference is carried out based on the posterior distribution. In parallel, the interpretation of the interval estimates does not involve assumptions about replications of the experiment. That is, in the Bayesian context given the interval estimates, $\boldsymbol{\theta}$ belongs to it with $(1-\alpha)\%$ probability.

1.2.2 Linear model: conjugate priors

Here we consider a Bayesian linear model in the form

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_d),$$

where $\sigma^2 > 0$, $\mathbf{I}_d$ an identity matrix, $\boldsymbol{\beta} = (\beta_0, ..., \beta_{d-1})^\top$ a $d \times 1$ vector, $\mathbf{X}$ an $n \times d$ design matrix and we assume that ϵ_i's are independent. The likelihood function is also

$$f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \Big\}$$

Under the Bayesian point of view, the inference process involves data and prior information. Thus, we assume a $\mathbf{N}_d(\mathbf{m}, \sigma^2\mathbf{V})$, which is a conjugate prior distribution for $\boldsymbol{\beta}|\sigma^2$ as follows

$$f(\boldsymbol{\beta}|\sigma^2, \mathbf{m}, \mathbf{V}) = (2\pi\sigma^2)^{-d/2} |\mathbf{V}|^{-d/2} \exp\Big\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \mathbf{m})^\top \mathbf{V}^{-1} (\boldsymbol{\beta} - \mathbf{m}) \Big\}.$$

For σ^2, we also set a conjugate prior distribution given by an Inverse Gamma denoted by $\mathrm{IG}(a, b)$ in the form of

$$f(\sigma^2|a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{a-1} \exp\Big\{ -\frac{b}{\sigma^2} \Big\},$$

where $a > 0$ and $b > 0$. Since we have the likelihood function and the proper priors, we can then find the posterior distribution to make inference on the parameters $\boldsymbol{\beta}$ and σ^2. Using the Bayes' theorem, we have

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) &= \frac{f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}|\sigma^2, \mathbf{m}, \mathbf{V}) f(\sigma^2|a, b)}{f_{\mathrm{y}}(\mathbf{y})}, \qquad (5) \\ &\propto f_{\mathrm{y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}|\sigma^2, \mathbf{m}, \mathbf{V}) f(\sigma^2|a, b), \\ &\propto (\sigma^2)^{-\frac{n}{2} - \frac{d}{2} + a - 1} \exp\Big\{ -\frac{A}{2\sigma^2} \Big\}, \end{aligned}$$

where

$$
\begin{aligned}
A &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\mathbf{m})^\top\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m}) + 2b,\\
&= \mathbf{y}^\top\mathbf{y} - \mathbf{y}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{V}^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^\top\mathbf{V}^{-1}\boldsymbol{\beta} + \mathbf{m}^\top\mathbf{V}^{-1}\mathbf{m} + 2b,\\
&= \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X}+\mathbf{V}^{-1})\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}) + (\mathbf{m}^\top\mathbf{V}^{-1}\mathbf{m}+2b+\mathbf{y}^\top\mathbf{y}) - (\mathbf{y}^\top\mathbf{X}+\mathbf{m}^\top\mathbf{V}^{-1})\boldsymbol{\beta}.
\end{aligned}
$$

For convenience, let $\boldsymbol{\Lambda} = (\mathbf{X}^\top\mathbf{X}+\mathbf{V}^{-1})^{-1}$ a $d \times d$ matrix and $\boldsymbol{\mu} = (\mathbf{X}^\top\mathbf{X}+\mathbf{V}^{-1})^{-1}(\mathbf{X}^\top\mathbf{y}+\mathbf{V}^{-1}\mathbf{m})$ a $d \times 1$ vector. Hence,

$$
\begin{aligned}
A &= \boldsymbol{\beta}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta} + \mathbf{m}^\top\mathbf{V}^{-1}\mathbf{m} + 2b + \mathbf{y}^\top\mathbf{y},\\
&= (\boldsymbol{\beta}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}) - \boldsymbol{\mu}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu} + \mathbf{m}^\top\mathbf{V}^{-1}\mathbf{m} + 2b + \mathbf{y}^\top\mathbf{y}.
\end{aligned}
$$

Finally, the joint posterior distribution for $\boldsymbol{\beta}$ and σ^2 is given by

$$
\begin{aligned}
f(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X}) \propto & f_{\mathrm{y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)f(\boldsymbol{\beta}|\sigma^2,\mathbf{m},\mathbf{V})f(\sigma^2|a,b),\\
&\propto(\sigma^2)^{-\frac{d}{2}}\exp\Big\{-\frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}{2\sigma^2}\Big\}\\
&\quad\times(\sigma^2)^{-\frac{n}{2}+a-1}\exp\Big\{-\frac{\mathbf{m}^\top\mathbf{V}^{-1}\mathbf{m}-\boldsymbol{\mu}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}+2b+\mathbf{y}^\top\mathbf{y}}{2\sigma^2}\Big\}.
\end{aligned}
\tag{6}
$$

Therefore, the equation (6) shows that the posterior distribution $f(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X})$ is proportional to the multiplication of kernels of the $\mathbf{N}_n(\boldsymbol{\mu},\sigma^2\boldsymbol{\Lambda})$ and IG ($a^* = -\frac{n}{2}+a$, $b^* = b + \frac{\mathbf{m}^\top\mathbf{V}^{-1}\mathbf{m}-\boldsymbol{\mu}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu}+\mathbf{y}^\top\mathbf{y}}{2}$). The manipulations presented above are partially available in [2], [4] and [16]. Additionally, [4] shows the normalizing constant and the marginal distribution of $\mathbf{y}$, which are necessary for equation (5).

1.2.3 Linear model: non-conjugate prior

In order to illustrate a Bayesian linear model with non-conjugate prior where Metropolis-Hastings may be helpful, consider $\mathbf{y} \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$ and assume that σ^2 is known. Also, consider a Laplace distribution as prior for each β_i such as

$$
f(\beta_i|m_i,v_i) = (2v_i)^{-d/2}\exp\Big\{-\sum_{i=1}^{d}\frac{|\beta_i-m_i|}{v_i}\Big\},
$$

where $\beta_i \in \mathbb{R}$, $m_i \in \mathbb{R}$ and $v_i > 0$. In this case, the posterior distribution is given by

$$
\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) &\propto f_{\mathrm{y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})\prod_{i=1}^{d}f(\beta_i|m_i,v_i),\\
&\propto \exp\Big\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \sum_{i=1}^{d}\frac{|\beta_i-m_i|}{v_i}\Big\}.
\end{aligned}
$$

The equation above has no closed form of a known p.d.f or p.m.f and MCMC methods can be utilized in order to obtain samples from the posterior distributions of β_i. Although the Gibbs Sampling is an MCMC method, in this case it cannot be used since it is not possible to sample directly from $f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2)$. In contrast, Metropolis-Hasting can be applied.

1.2.4 MCMC methods

In order to estimate the unknown quantities in the Bayesian models, Markov Chain Monte Carlo (MCMC) methods are useful tools to sample from those posterior distributions that have no closed form. In this section we briefly introduce two MCMC algorithms commonly used: Gibbs Sampling [5, 6] and Metropolis-Hastings [7, 8].

In the statistical context, Monte Carlo integration is convenient when it is not possible to analytically compute a finite integral such as

$$\int g(\theta)p(\theta)d\theta, \tag{7}$$

where $p(\cdot)$ is a p.d.f or p.m.f and $g(\cdot)$ a integrable function. In short, i.i.d samples are generated from $p(\cdot)$ and evaluated at $g(\cdot)$ and then averaged. For a sufficiently large number of samples, the Strong Law of Large Numbers holds this average converges almost surely to (7) [9]. When it is not possible to directly sample from $p(\cdot)$, Gibbs Sampling and Metropolis-Hastings are alternatives to generate samples $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(K)}$ from $p(\cdot)$ no longer independent (now with a Markovian dependence structure) that evaluated at $t(\cdot)$ and averaged over the samples, also converge almost surely to (7).

The Gibbs Sampling is a stochastic simulation algorithm via Markov Chain utilized when the joint posterior distribution has no closed form but all its conditional distributions do. For instance, consider that $p(\boldsymbol{\theta}|\mathbf{y})$ is the joint posterior distribution and suppose that its conditional distributions may be written as

$$p(\theta_k|\theta_0, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_{d-1}, \mathbf{y}), \text{ where } k = 0, ..., d-1.$$

The idea is to successively sample from each conditional distribution of θ_k in order to obtain samples from the joint posterior distribution. The Gibbs Sampling is described as following

Algorithm Gibbs Sampling

1. Initialize $t = 1$ and define initial values $\theta_0^{(0)}, \theta_1^{(0)}, ..., \theta_{d-1}^{(0)}$ for the vector $\boldsymbol{\theta} = (\theta_0, \theta_2, ..., \theta_{d-1})^\top$.

2. Sample

$$\begin{aligned}
\theta_0^{(t)} &\sim p(\theta_0|\theta_1^{(t-1)}, \theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_{d-1}^{(t-1)}, \mathbf{y});\\
\theta_1^{(t)} &\sim p(\theta_1|\theta_2^{(t)}, \theta_3^{(t-1)}, \theta_4^{(t-1)}, ..., \theta_{d-1}^{(t-1)}, \mathbf{y});\\
&\vdots\\
\theta_{d-1}^{(t)} &\sim p(\theta_{d-1}|\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_{d-2}^{(t)}, \mathbf{y});
\end{aligned}$$

3. Take $t = t + 1$ and return to the step 2 until the desired sample has been obtained for each θ_k.

Similarly to the Gibbs Sampling, Metropolis-Hastings is also a stochastic algorithm that generates samples with Markovian dependence structure from a certain distribution (or kernel). Further, Metropolis-Hastings is more flexible than Gibbs Sampling, since it can be utilized to generate samples from distributions that have or not closed form. When it is possible to generate directly from the conditional or joint distribution, it is appropriate to use Gibbs Sampling, once Metropolis-Hastings has an acceptance and rejection step. But when at least one of the conditional distributions have no closed form, Metropolis might be an alternative. In some cases Metropolis and Gibbs Sampling are combined because some conditionals have closed form and others do not. This hybrid algorithm is called Metropolis-within-Gibbs and was proposed by [10] and [11].

Again, suppose we desire to obtain a sample from the joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. The Metropolis-Hastings is based on a proposal distribution $q(\boldsymbol{\theta}^{(t-1)})$, which generates candidate values $\boldsymbol{\theta}^*$ that are accepted as values from $p(\boldsymbol{\theta}|\mathbf{y})$ with certain probability. In its first version [7], the $q(\boldsymbol{\theta}^{(t-1)})$ is only Normal, where the mean has a Markovian structure and the variance is constant. A more general framework was proposed by [8] in which the proposal distribution can be other than Normal, and since then many adaptive Metropolis algorithms have been introduced, which basically differ by the specification of the covariance matrix of the proposal distribution [12, 13, 14, 15]. Below, the Metropolis algorithm proposed by [8] is described.

Algorithm Metropolis-Hastings

```
1. Initialize $t = 1$ and define the initial values $\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_d^{(0)}$ for the vector $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_{d-1})$;
2. Sample $\boldsymbol{\theta}^*$ from the proposal distribution $q(\boldsymbol{\theta}^{(t-1)})$;
    (a) Compute
        $\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})q(\boldsymbol{\theta}^{(t-1)})}\right\}$,    (8)
    (b) Compute $u \sim U[0,1]$. If $u < \alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^*)$, then $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$, otherwise, $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$;
3. Take $t = t + 1$ and return to the step 2 until the desired posterior sample has been obtained.
```

The choice of the proposal distribution must be based on the support of $\boldsymbol{\theta}$. For example, if $\boldsymbol{\theta} \in \mathbb{R}^d$ it is appropriate to choose a proposal that is also supported on $\mathbb{R}^d$, otherwise the results generated by Metropolis might be invalid.

One important characteristic of the MCMC algorithms is that they generate chains that need to converge. That is, for each component of $\boldsymbol{\theta}$ or $\boldsymbol{\beta}$, a chain of values with Markovian dependence structure is generated and one must verify its convergence to the joint posterior distribution. Due to it, we usually set a warm-up period, which is called burn-in, from which the samples start being considered and all inference is made utilizing only these samples; see [9] for theoretical convergence results of the MCMC methods.

References

[1] W. Feller. An Introduction to Probability Theory and Its applications. 1967; John Wiley, Ed.3, vol. 1.

[2] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis. 2014; CRC, Ed.3, Boca Raton.

[3] C. Robert. The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation. 2007; Springer, Ed. 2, New York.

[4] A. O'Hagan. Kendall's Advanced Theory of Statistics: Bayesian Inference. 1994; Arnold, Ed. 2B, London.

[5] S. Geman, D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions Pattern Analysis and Machine Intelligence. 1984; 6:721–741.

[6] A. Gelfand, A. Smith. Sampling based approaches to calculating marginal densities. Journal of American Statistical Association. 1990; 85:398–409 Springer, New York.

[7] N. Metropolis, A. Rosenbluth, M. Teller, E. Teller. Equations of state calculations by fast computing machines. Journal of Chemistry and Physics. 1953;1087:1091–21.

[8] W. Hastings. Monte Carlo sampling using Markov chains and their applications. Biometrika. 1970; 57: 97-109.

[9] C. Robert, G. Casella. Monte Carlo Statistical Methods. 2004; Springer, New York.

[10] P. Muller. A generic approach to posterior integration and Gibbs sampling. Technical report, Purdue University, West Lafayette, Indiana.

[11] P. Muller. Alternatives to the Gibbs Sampling scheme. Technical report, Institute of Statistics and Decision Science, Duke University, Durham, North Carolina.

[12] D. Gamerman. Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing. 1997;57:68–7.

[13] H. Haario, E. Saksman, J. Tamminen. An adaptive Metropolis algorithm. Bernoulli. 2001;223:243–7.

[14] M. Vihola. Robust adaptive Metropolis algorithm with coerced acceptance rate. Statistics and Computing. 2012;997:1008–843.

[15] I. S. Mbalawata, S. Sarkka, M. Vihola, H. Haario. Adaptive Metropolis algorithm using variational Bayesian adaptive Kalman Filter. Computational Statistics and Data Analysis. 2015;101:115–83.

[16] D. Gamerman, H. F. Lopes. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. 2006; Chapman and Hall/CRC, Ed. 2, London.