MCMC Algorithms

Bayesian Psychometric Models, Lecture 8

Deriving Posterior Distributions from Conjugate Priors

A benefit of conjugate priors is that posterior distributions can be derived analytically, which makes algorithms much more efficient. This document describes how to do so for Bayesian Linear Models. Following this, we will inspect the Gibbs_Sampling.R file to see how Gibbs Sampling works (with our derived posterior distributions). We will then inspect Metropolis-Hastings_Sampling.R to see how the Metropolis-Hastings algorithm works.

Linear Models

The classic linear model is one where, for a respondent r ($r = 1, \dots, N$), a dependent variable (or outcome), y_r, is predicted via a set of independent variables (or predictors), x_{rv} where $v = 1, \dots, V$ and, sometimes, their interactive product, by means of a set of regression coefficients β_v:

$$Y_r = \beta_0 + \beta_1 x_{r1} + \beta_2 x_{r2} + \cdots + \beta_V x_{rV} + e_r = \boldsymbol{x}_r \boldsymbol{\beta} + e_r$$

Here, e_r is the residual or error term for respondent r, with $e_r \sim N(0, \sigma_e^2)$.

For our example data, the linear model is thus:

$$Y_r = \beta_0 + \beta_{Height} x_{Height,r} + \beta_{Group2} x_{Group2,r} + \beta_{Group3} x_{Group3,r} + \beta_{Height*Group2} x_{Height,r} x_{Group2,r} + \beta_{Height*Group3} x_{Height,r} x_{Gr}$$

The right-hand side is the vector notation of the regression equation for a single respondent, with $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_V]^T$ a vector of regression coefficients with size $(V+1) \times 1$ and $\boldsymbol{x}_r = [1, x_{r1}, \dots, x_{rV}]$ a vector of a constant of 1 (for multiplying the intercept β_0) and V predictors, for a size of $1 \times (V+1)$.

When put together across all respondents, we get a common matrix format of the linear model that we will use to build the Bayesian Linear Model under Gibbs Sampling and Metropolis-Hastings:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

Here:

- $\boldsymbol{Y}$ is size $N \times 1$
- $\boldsymbol{X}$ is size $N \times (V+1)$, with the first column being full of 1s
- $\boldsymbol{\beta}$ is size $(V+1) \times 1$
- $\boldsymbol{e}$ is size $N \times 1$ and $\boldsymbol{e} \sim MVN\left(\boldsymbol{0}, \boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{I}_{(N \times N)}\right)$, where:
    - $\boldsymbol{I}_{(N \times N)}$ is an $(N \times N)$-sized identity matrix

Bayesian Estimation of Model for Example Data

Before we dive into how the algorithms work, we will first examine how to estimate a Bayesian version of our model. To do so we need to specify:

1. Likelihood function (from the model)
2. Prior distributions for all model parameters

Likelihood Function

Using the general matrix form of the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, express the likelihood function:

Here, note the determinant of $|\Sigma_e| = \left|\sigma_e^2\boldsymbol{I}\right| = \left|\sigma_e^2\right| |\boldsymbol{I}| = \left(\sigma_e^2\right)^N = \sigma_e^{2N}$

$$f(\boldsymbol{Y}|\boldsymbol{X}, \beta, \sigma_e^2) = \frac{1}{\sqrt{(2\pi)^V \left|\sigma_e^2\boldsymbol{I}\right|}} \exp\left(\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\sigma_e^2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right) = (2\pi)^{-\frac{V}{2}}\left(\sigma^2\right)^{-\frac{N}{2}} \exp\Bigg(-\frac{1}{2\sigma_e^2}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T (\boldsymbol{Y} - \boldsymbol{X}$$

Often, the normalizing constants are removed, which leaves:

$$f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma_e^2) \propto \left(\sigma_e^2\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma_e^2}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right)$$

We will come back to this distribution later for now, please note that because $\boldsymbol{\Sigma}_e = \sigma_e^2\boldsymbol{I}_{(N\times N)}$, an equivalent expression is possible using a series of independent univariate normal distributions:

$$f(\boldsymbol{Y}|\boldsymbol{X}, \beta, \sigma_e^2) = \prod_{r=1}^{N} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(\frac{-(Y_r - \boldsymbol{X}_r\boldsymbol{\beta})^2}{2\sigma_e^2}\right)$$

Alternatively, $\left(Y_r|\boldsymbol{X}_r, \boldsymbol{\beta}, \sigma_e^2\right) \sim N\left(\boldsymbol{X}_r\boldsymbol{\beta}, \sigma_e^2\right)$.

Prior Distributions

The choice of prior distributions is commonly very difficult. As conjugate priors make algorithms more efficient, we will begin there.

Choosing conjugate prior distributions in Bayesian linear models has one complicating factor: The likelihood for $\boldsymbol{\beta}$ is conditional on σ_e^2. So, let's start there. In particular:

$$f\left(\boldsymbol{\beta}, \sigma_e^2\right) = f\left(\boldsymbol{\beta} \mid \sigma_e^2\right) f\left(\sigma_e^2\right)$$

Further, we can start to see how a conjugate prior may look by altering the likelihood function slightly by noting that:

$$\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right) = \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$$

Where $\hat{\boldsymbol{\beta}}$ is the result of the solution to the "normal equations" consisting only of $\boldsymbol{Y}$ and $\boldsymbol{X}$:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

This leads to a rewritten likelihood function of:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma_e^2) \propto \left(\sigma_e^2\right)^{-\frac{\nu}{2}} \exp\left(-\frac{\nu\sigma_0^2}{2\sigma_e^2}\right) \left(\sigma_e^2\right)^{-\frac{N-\nu}{2}} \exp\left(-\frac{1}{2\sigma_e^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)$$

Where $\nu = N - V$.

Then, if you squint really hard, you can see that the conjugate prior for the residual variance is an inverse-gamma distribution:

$$f(\sigma_e^2) \propto \left(\sigma_e^2\right)^{-\frac{\nu_0}{2}} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma_e^2}\right)$$

The inverse-gamma distribution is often notated with parameters α and β. Here, $\alpha = \frac{\nu_0}{2}$ and $\beta = \frac{1}{2}\nu_0\sigma_0^2$ where ν_0 and σ_0^2 are the prior values of the denominator degrees of freedom and estimated residual variance.

For $\boldsymbol{\beta}$, the conditional prior is a multivariate normal distribution with prior mean $\boldsymbol{\mu}_0$ and prior covariance matrix Σ_0.

$$f\left(\boldsymbol{\beta} \mid \sigma_e^2\right) \propto \left(\sigma_e^2\right)^{-\frac{V}{2}} \exp\left(-\frac{1}{2\sigma_e^2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)^T \Sigma_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_0\right)\right)$$

Posterior Distribution with Conjugate Priors

The posterior distribution becomes a mess, but it can be summarized by looking first at σ_e^2 and the at $\boldsymbol{\beta}$. The entire derivation won't be provided, however, the joint posterior distribution $f\left(\boldsymbol{\beta},\sigma_e^2 \mid \boldsymbol{Y},\boldsymbol{X}\right)$ is a product of inverse-gamma and normal distributions.

For each derivation outline, we must take the product of the prior and data likelihoods, factor them, then figure out the form of the posterior.

Fun Fact About Multivariate Normal Distributions

It can be shown that a random vector $\boldsymbol{z}$ follows a multivariate normal distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{V}$ if and only if

$$f\left(\boldsymbol{z} \mid \boldsymbol{m},\boldsymbol{V}\right) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{z}^T\boldsymbol{V}^{-1}\boldsymbol{z} - 2\boldsymbol{z}^T\boldsymbol{V}^{-1}\boldsymbol{m}\right)\right)$$

Outline of Derivation of Posterior for β

The prior distribution for $\boldsymbol{\beta}$ is:

$$f\left(\boldsymbol{\beta} \mid \sigma_e^2\right) = \left|2\pi\boldsymbol{\Sigma}_0\right|^{-1} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right)$$

The data likelihood for $\boldsymbol{\beta}$ is:

$$f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma_e^2) = \left(2\pi\sigma_e^2\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma_e^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right)$$

As $\boldsymbol{\beta}$ only exists in the exponents, we can work with those exclusively. For the prior distribution:

$$\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0$$

For the data:

$$\frac{1}{\sigma_e^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)=\frac{\boldsymbol{Y}^T\boldsymbol{Y}}{\sigma_e^2}-\frac{2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}+\frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}}{\sigma_e^2}$$

The product (which is a sum in the exponent):

$$\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+\boldsymbol{\beta}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right)+\left(\frac{\boldsymbol{Y}^T\boldsymbol{Y}}{\sigma_e^2}-\frac{2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}+\frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}}{\sigma_e^2}\right)$$

Removing the terms that do not depend on $\boldsymbol{\beta}$ and rearranging terms leaves:

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}+\frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}}{\sigma_e^2}-2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+-\frac{2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}$$

Then, combining like terms for the first two parts and the second two parts leaves:

$$\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_0^{-1}+\frac{\boldsymbol{X}^T\boldsymbol{X}}{\sigma_e^2}\right)\boldsymbol{\beta}-2\boldsymbol{\beta}^T\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+\frac{\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}\right)$$

This fits the MVN pattern shown above where $\boldsymbol{\beta}$ follows a multivariate normal distribution with inverse variance:

$$\boldsymbol{\Sigma}_n^{-1}=\left(\boldsymbol{\Sigma}_0^{-1}+\frac{\boldsymbol{X}^T\boldsymbol{X}}{\sigma_e^2}\right),$$

and variance inverse times mean:

$$\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta}_n=\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+\frac{\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}\right)$$

Therefore the mean vector is:

$$\boldsymbol{\beta}_n=\boldsymbol{\Sigma}_n^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+\frac{\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}\right)=\left(\boldsymbol{\Sigma}_0^{-1}+\frac{\boldsymbol{X}^T\boldsymbol{X}}{\sigma_e^2}\right)\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0+\frac{\boldsymbol{X}^T\boldsymbol{Y}}{\sigma_e^2}\right)$$

The posterior distribution for $\boldsymbol{\beta}$ conditional on σ_e^2 is normal, with mean $\bar{\boldsymbol{\beta}}$ and variance $\boldsymbol{\Sigma}_\beta$:

$$\bar{\boldsymbol{\beta}}=\left(\Sigma_0^{-1}+\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\Sigma_0^{-1}\boldsymbol{\beta}_0+\boldsymbol{X}^T\boldsymbol{Y}\right)$$

Derivation of Posterior for σ_e^2

The product of prior times data likelihood for σ_e^2 is:

$$f(\sigma_e^2|\boldsymbol{\beta},\boldsymbol{X},\boldsymbol{Y})\propto\left[\left(\sigma_e^2\right)^{-\frac{\nu_0}{2}+1}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma_e^2}\right)\right]\times\left[\left(\sigma_e^2\right)^{-\frac{N}{2}}\exp\left(-\frac{1}{2\sigma_e^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right)\right]$$

$$= \left(\sigma_e^2\right)^{-\frac{\nu_0 - N}{2}} \exp\left(-\frac{\nu_0\sigma_0^2 + (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta})}{2\sigma_e^2}\right)$$

This is the form of an inverse gamma distribution, so the posterior distribution for σ_e^2 is inverse gamma, with:

$$\alpha_n = \frac{\nu_0 + N}{2}$$
$$\beta_n = \frac{\nu_0\sigma_0^2 + (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta})}{2}$$

Side Note: When You Read "Assuming Known σ_e^2"

Although "knowing σ_e^2" sounds implausible (if I *knew* σ_e^2, I would likely know $\boldsymbol{\beta}$, and probably would have lots of other impressive abilities…), we start here this because in an MCMC algorithm, "known" can also mean "conditional on this specific value of σ_e^2".

Side Note: Issues with Non-Invariance of Priors

The priors selected today, while conjugate, lead to a posterior distribution that is non-invariant under transformations of the parameters. That is, if we took $\gamma = \log\left(\sigma_e^2\right)$, then it can be shown that the posterior would be different. If this is an issue, a Jefferies or Reference prior could be used, which leaves the posterior invariant under transformation of parameters. The Jeffries prior for our case is:

$$f(\boldsymbol{\beta}, \sigma_e^2) \propto \frac{1}{\sigma_e^2},$$

which is uniform across all values of parameters.

Gibbs Sampling Using Complete Conditional Distributions

See Gibbs_Sampling.R file.

Metropolis-Hastings for σ_e^2

See Metropolis-Hastings_Sampling.R file.