Multivariate Linear Regression with Bayesian Inference

Joaquín Martínez-Minaya

October 7, 2024

1 Multivariate Linear Regression Model

We define the multivariate linear regression model as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$

where:

- $\boldsymbol{Y}$ is an $n \times m$ matrix of response variables, where n is the number of observations and m is the number of response variables.
- $\boldsymbol{X}$ is an $n \times k$ matrix of covariates, where k is the number of covariates (including the intercept).
- $\boldsymbol{B}$ is a $k \times m$ matrix of coefficients.
- $\boldsymbol{E}$ is an $n \times m$ matrix of errors, assumed to follow a multivariate normal distribution: $\boldsymbol{E} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

2 Prior Distributions

We use conjugate prior distributions for the coefficients matrix $\boldsymbol{B}$ and the covariance matrix $\boldsymbol{\Sigma}$.

2.1 Prior for the Coefficient Matrix $\boldsymbol{B}$

The prior distribution for the coefficient matrix $\boldsymbol{B}$ given the covariance matrix $\boldsymbol{\Sigma}$ is defined as a matrix normal distribution:

$$\boldsymbol{B} \mid \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{B}_0, \boldsymbol{\Sigma} \otimes \boldsymbol{A}^{-1})$$

where:

- $\boldsymbol{B}_0$ is a $k \times m$ matrix of prior means, typically set to zero.
- $\boldsymbol{A}$ is a $k \times k$ precision matrix for the covariates. By default, it is set as an identity matrix, $\boldsymbol{A} = \boldsymbol{I}_k$.
- $\boldsymbol{\Sigma}$ is the covariance matrix of the response variables.

2.2 Prior for the Covariance Matrix $\boldsymbol{\Sigma}$

The prior distribution for the covariance matrix $\boldsymbol{\Sigma}$ follows an inverse-Wishart distribution:

$$\boldsymbol{\Sigma} \sim \mathrm{IW}(\nu_0, \boldsymbol{V}_0)$$

where:

- ν_0 is the degrees of freedom parameter, set as $\nu_0 = m + 2$.
- $\boldsymbol{V}_0$ is a $m \times m$ scale matrix, set as an identity matrix $\boldsymbol{V}_0 = \boldsymbol{I}_m$.

3 Posterior Distributions

Given the prior distributions and observed data, the posterior distributions for the parameters are derived as follows:

3.1 Posterior of $\boldsymbol{B} \mid \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X}$

The conditional posterior distribution of the coefficient matrix $\boldsymbol{B}$ given the covariance matrix $\boldsymbol{\Sigma}$ is a matrix normal distribution:

$$\boldsymbol{B} \mid \boldsymbol{\Sigma}, \boldsymbol{Y}, \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{B}_n, \boldsymbol{\Sigma} \otimes (\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{A})^{-1})$$

where:

$$\boldsymbol{B}_n = (\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{A})^{-1}(\boldsymbol{X}^T \boldsymbol{Y} + \boldsymbol{A}\boldsymbol{B}_0)$$

3.2 Marginal Posterior of $\boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{X}$

The marginal posterior distribution of the covariance matrix $\boldsymbol{\Sigma}$ follows an inverse-Wishart distribution:

$$\boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{X} \sim \mathrm{IW}(\nu_0 + n, \boldsymbol{V}_0 + \boldsymbol{S})$$

where:

$$\boldsymbol{S} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_n)^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}_n) + (\boldsymbol{B}_n - \boldsymbol{B}_0)^T \boldsymbol{A}(\boldsymbol{B}_n - \boldsymbol{B}_0)$$

3.3 Marginal Posterior of $\boldsymbol{B} \mid \boldsymbol{Y}, \boldsymbol{X}$

The marginal posterior distribution of the coefficient matrix $\boldsymbol{B}$ is a matrix t-distribution:

$$\boldsymbol{B} \mid \boldsymbol{Y}, \boldsymbol{X} \sim \mathcal{T}(\boldsymbol{B}_n, \boldsymbol{\Sigma}_n \otimes (\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{A})^{-1}, \nu_n)$$

where:

- $\boldsymbol{B}_n$ is the posterior mean of $\boldsymbol{B}$.
- $\boldsymbol{\Sigma}_n = (\boldsymbol{V}_0 + \boldsymbol{S})/(\nu_0 + n - m + 1)$ is the scale matrix for $\boldsymbol{\Sigma}$.
- $\nu_n = \nu_0 + n - m + 1$ is the degrees of freedom for the t-distribution.

4 Posterior Predictive Distribution

The posterior predictive distribution for a new observation $\boldsymbol{x}^*$ is given by:

$$\boldsymbol{Y}^* \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{x}^* \sim \mathcal{T}(\boldsymbol{x}^* \boldsymbol{B}_n, \boldsymbol{\Sigma}_n(1 + \boldsymbol{x}^{*T}(\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{x}^*), \nu_n)$$

This distribution captures both the uncertainty in the estimation of $\boldsymbol{B}$ and the variability in the new data point $\boldsymbol{Y}^*$.