Bayesian Linear Models and GLMs

Søren Lund Pedersen

1 Introduction

In this article, I explore the relationship between Bayesian Linear Models and Generalized Linear Models (GLMs). The Bayesian approach builds on Bayes' theorem, allowing us to incorporate prior beliefs into our models. This framework can be extended to both linear models and GLMs, providing a unified approach for various types of data.

2 Linear Models

Linear models express a linear relationship between a dependent variable Y and independent variables, represented by coefficients $\boldsymbol{\beta}$. In the frequentist approach, $\boldsymbol{\beta}$ are fixed but unknown parameters estimated using maximum likelihood estimation (MLE). In contrast, Bayesian statistics treats $\boldsymbol{\beta}$ as random variables with associated probability distributions, capturing uncertainty in their values.

Bayesian inference often benefits from conjugate priors, especially when the likelihood belongs to the exponential family. For linear models (e.g., ordinary least squares regression), assuming normally distributed data and coefficients simplifies analysis. A common choice for the prior is the Gaussian prior, which leads to Ridge regression, or the double exponential (Laplace) prior, which leads to LASSO regression.

2.1 Conjugate Priors: Ridge and LASSO

Gaussian Prior (Ridge Regression):

$$P(\boldsymbol{\beta}) \propto \exp\left(-\frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2\right)$$

Using this prior, the MAP estimate minimizes:

$$\hat{\boldsymbol{\beta}}_{MAP} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N}(Y_i - \boldsymbol{\beta}^T X_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right].$$

This corresponds to Ridge regression, where the penalty term $\lambda \sum \beta_j^2$ shrinks coefficients toward zero, reducing model complexity.

Double Exponential Prior (LASSO Regression):

$$P(\boldsymbol{\beta}) \propto \exp\left(-\lambda\|\boldsymbol{\beta}\|_1\right)$$

With this prior, the MAP estimate becomes:

$$\hat{\boldsymbol{\beta}}_{MAP} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (Y_i - \boldsymbol{\beta}^T X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

This formulation leads to LASSO regression, which encourages sparsity in the coefficients by imposing an ℓ_1-norm penalty.

2.2 Relating to OLS by Minimizing Negative Log-Posterior

The connection between Bayesian MAP estimation and ordinary least squares (OLS) can be seen through the minimization of the negative log-posterior. The MAP estimate is given by:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \left[-\log P(Y \mid \beta) - \log P(\beta) \right].$$

For a linear regression model with Gaussian errors, the likelihood is:

$$P(Y \mid \beta) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \beta^T X_i)^2 \right).$$

Taking the negative log of the likelihood yields:

$$-\log P(Y \mid \beta) \propto \sum_{i=1}^{N} (Y_i - \beta^T X_i)^2.$$

If we assume a noninformative (uniform) prior $P(\beta) \propto 1$, then $-\log P(\beta)$ is constant and can be ignored. Thus:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \beta^T X_i)^2.$$

This shows that, without a prior, MAP estimation recovers the OLS solution.

3 Bayesian GLMs

We extend the Bayesian framework to non-Gaussian models using GLMs. After selecting the appropriate probability distribution for the data and choosing a conjugate prior, we apply Bayes' theorem:

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \;=\; \frac{p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X})\, p(\boldsymbol{\beta})}{p(\mathbf{y} \mid \mathbf{X})},$$

where:

- $p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X})$ is the likelihood from the chosen exponential-family model,
- $p(\boldsymbol{\beta})$ is the prior on the coefficients,
- $p(\mathbf{y} \mid \mathbf{X})$ is the normalizing constant.

Due to the complexity of computing the normalizing constant, approximation techniques like Markov Chain Monte Carlo (MCMC) are often used.

3.1 Probit Regression and Latent Variable Interpretation

Probit regression is a type of GLM used for modeling binary outcomes, similar to logistic regression but using the normal CDF as the link function.

Consider a binary response $Y \in \{0, 1\}$ and predictors X. In probit regression, we model:

$$P(Y = 1 \mid X, \boldsymbol{\beta}) = \Phi(X^T \boldsymbol{\beta}),$$

where Φ is the CDF of the standard normal distribution.

A latent variable interpretation provides additional intuition:

- Assume there exists an unobserved continuous variable $Z = X^T \boldsymbol{\beta} + \epsilon$ where $\epsilon \sim N(0, 1)$.
- The observed binary outcome Y is determined by Z:

$$Y = \begin{cases} 1 & \text{if } Z > 0, \\ 0 & \text{if } Z \leq 0. \end{cases}$$

The probability that $Y = 1$ is then:

$$P(Y = 1 \mid X, \boldsymbol{\beta}) = P(Z > 0 \mid X, \boldsymbol{\beta}) = P(\epsilon > -X^T \boldsymbol{\beta}) = \Phi(X^T \boldsymbol{\beta}),$$

which matches the probit model.

The latent variable formulation provides a clear interpretation: the predictors X influence an underlying continuous propensity Z for the event $Y = 1$. When this latent propensity exceeds a threshold (zero), the event occurs. Bayesian probit regression then involves placing a prior on $\boldsymbol{\beta}$ and updating this with observed binary data to obtain a posterior distribution over $\boldsymbol{\beta}$. MCMC or other approximate methods are typically used to perform this inference due to the lack of closed-form solutions.