

# Introduction to Bayesian Linear Regression, Posterior Inference, and MCMC

Peter Sørensen

# Overview

- **Classical Linear Regression**
  - Model, inference, and limitations
- **Bayesian Linear Regression**
  - Motivation, priors, and posteriors
  - Conditional posteriors and inference
- **Computation and Applications**
  - MCMC and Gibbs sampling
  - Diagnostics and R implementation

# Introduction

- **Bayesian Linear Regression (BLR)** extends classical regression by incorporating **prior information** and producing **posterior distributions** over model parameters.
- **Advantages:**
  - Handles **high-dimensional** and **small-sample** problems.
  - Provides **full uncertainty quantification**.
  - Enables **regularization** and integration of **prior biological knowledge**.

# Applications in Genomics

- **Bayesian Linear Regression (BLR)** is widely applied in quantitative genetics and genomics.
- **Common use cases:**
  - Genome-Wide Association Studies (**GWAS**) and **fine-mapping** of causal variants.
  - **Genetic prediction** and **heritability estimation**.
  - **Pathway** and **gene-set enrichment** analyses.
  - Integrative **multi-omics** modeling (genome, transcriptome, epigenome).

# Classical Linear Regression

## Model

$$y = X\beta + e, \quad e \sim \mathcal{N}(0, \sigma^2 I_n)$$

- $y$ : outcomes
- $X$ : design matrix
- $\beta$ : coefficients
- $e$ : are the residuals
- $\sigma^2$ : residual variance

# Estimation

Regression effects:

$$\hat{\beta} = (X^{\top} X)^{-1} X^{\top} y$$

Residual variance:

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_i (y_i - x_i^{\top} \hat{\beta})^2$$

Inference via standard errors and  $t$ -tests, confidence intervals, and prediction intervals.

# Limitations

- No explicit control over **effect size distribution**
- Sensitive when **collinearity** is high
- **Not identifiable** when  $p > n$
- Uncertainty largely **asymptotic** unless normality assumptions hold

# Why Bayesian Linear Regression?

- Combines **likelihood** and **prior** to form the **posterior**.
- Priors express beliefs about **effect sizes**:
  - Normal  $\rightarrow$  many small effects
  - Spike-and-slab  $\rightarrow$  sparse effects
- Acts as a **regularizer**:
  - Shrinks small/noisy effects toward 0
  - Preserves large, important effects
- **Stable when**  $p > n$  due to prior information.
- Provides **full posterior distributions** for  $\beta$  and  $\sigma^2$ .



# Overview: Bayesian Linear Regression

- Combines data and prior knowledge using **Bayes' rule**.
- Uses **conjugate priors** to yield closed-form full conditionals.
- Employs **Gibbs sampling** to approximate the posterior distribution.
- Estimates **parameters, uncertainty, and predictions** from posterior draws.

# Bayesian Linear Regression with Gaussian Priors

Bayesian linear regression starts with the same model structure as classical linear regression.

$$y = X\beta + e, \quad e \sim \mathcal{N}(0, \sigma^2 I_n)$$

- $y$ :  $n \times 1$  vector of observed outcomes
- $X$ :  $n \times p$  design matrix of predictors
- $\beta$ :  $p \times 1$  vector of unknown coefficients
- $e$ : Gaussian noise with mean 0 and variance  $\sigma^2$

# Likelihood in Bayesian Linear Regression

Because the residuals are Gaussian, it follows that the marginal distribution of  $y$  is:

$$e \sim \mathcal{N}(0, \sigma^2 I_n)$$

The marginal distribution of  $y$  is:

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

This defines the **likelihood** the probability of the observed data given parameters  $\beta$  and  $\sigma^2$ :

$$p(y \mid X, \beta, \sigma^2) = \mathcal{N}(X\beta, \sigma^2 I_n)$$

# Introducing Priors

In Bayesian linear regression, we specify **prior distributions** that express our beliefs about parameters before seeing the data.

A common **conjugate prior** for the regression coefficients is:

$$\beta \mid \sigma_b^2 \sim \mathcal{N}(0, \sigma_b^2 I_p)$$

This reflects the belief that most effect sizes are small and centered near zero — consistent with the **polygenic assumption** in genetics.

## Role of the Prior Variance $\sigma_b^2$

The parameter  $\sigma_b^2$  acts as a **shrinkage (regularization) parameter**:

- Small  $\sigma_b^2 \rightarrow$  stronger shrinkage toward zero.
- Large  $\sigma_b^2 \rightarrow$  weaker shrinkage, allowing larger effects.

It controls the **strength of regularization** and is often treated as an **unknown hyperparameter** estimated from the data.

## Priors on Variance Components

We also place priors on the variance components to complete the hierarchical model.

$$\sigma_b^2 \mid S_b, v_b \sim S_b \chi^{-2}(v_b), \quad \sigma^2 \mid S, v \sim S \chi^{-2}(v)$$

Here:

- $S_b$  and  $v_b$  are user-defined hyperparameters that control the prior distribution on the **variance of regression coefficients**.
- $S$  and  $v$  are hyperparameters for the **residual variance**  $\sigma^2$ .

# Conjugate Priors and Regularization

Conjugate priors keep posteriors in the same family (e.g., scaled inverse-chi-squared), allowing **closed-form Gibbs updates**.

They also serve as **regularizers**:

- The prior on  $\beta$  shrinks small or noisy effects toward zero.
- Priors on variance components prevent overfitting, especially when  $p > n$ .

Thus, conjugate priors make Bayesian linear regression **efficient** and **stable**.

# Posterior Distribution

In Bayesian analysis, we combine the **likelihood** and **priors** using Bayes' rule to obtain the **joint posterior**:

$$p(\beta, \sigma_b^2, \sigma^2 \mid y) \propto p(y \mid \beta, \sigma^2) p(\beta \mid \sigma_b^2) p(\sigma_b^2) p(\sigma^2)$$

This posterior captures all **updated knowledge** about the unknown parameters after observing the data.

It forms the basis for computing **posterior means**, **credible intervals**, and **predictions**.



# Conjugacy and Gibbs Sampling

With **conjugate priors**, each parameter's **full conditional distribution** has a closed-form solution.

This makes **Gibbs sampling** a natural and efficient inference method.

- Parameters are updated one at a time, each from its conditional posterior.
- The resulting Markov chain explores the **joint posterior** of  $(\beta, \sigma_b^2, \sigma^2)$ .

Gibbs sampling thus provides an easy way to approximate the full posterior in Bayesian linear regression.

## Full Conditional for $\beta$

Given  $\sigma^2$ ,  $\sigma_b^2$ , and the data  $y$ , the regression coefficients have a **multivariate normal** conditional posterior:

$$\beta \mid \sigma^2, \sigma_b^2, y \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$$

where

$$\Sigma_\beta = \left( \frac{X^\top X}{\sigma^2} + \frac{I}{\sigma_b^2} \right)^{-1}, \quad \mu_\beta = \Sigma_\beta \frac{X^\top y}{\sigma^2}$$

This distribution represents our **updated belief** about  $\beta$  after observing the data, while holding  $\sigma_b^2$  and  $\sigma^2$  fixed.

## Comparison to Classical OLS

In classical regression, the OLS estimator is

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y, \quad y \sim \mathcal{N}(X\beta, \sigma^2 I)$$

The estimate of  $\beta$  is **independent of**  $\sigma^2$ ,  
since  $\sigma^2$  only scales the likelihood, not its maximum.

In Bayesian regression,  $\sigma^2$  appears explicitly in the posterior:

$$\Sigma_\beta = \left( \frac{X^\top X}{\sigma^2} + \frac{I}{\sigma_b^2} \right)^{-1}, \quad \mu_\beta = \Sigma_\beta \frac{X^\top y}{\sigma^2}$$

The term  $\frac{I}{\sigma_b^2}$  introduces **shrinkage**, regularizing estimates and stabilizing inference especially when  $p > n$  or predictors are highly correlated.

Thus, the Bayesian posterior mean is a **regularized, uncertainty-aware generalization** of OLS.

## Full Conditional for $\beta_j$

Instead of sampling  $\beta$  jointly, we can update each coefficient  $\beta_j$  **one at a time**, holding all others fixed efficient for large  $p$  or spike-and-slab models.

Let  $X_j$  be the  $j$ th column of  $X$  and define the **partial residual**:

$$r_j = y - X_{-j}\beta_{-j}$$

Then the conditional posterior for  $\beta_j$  is univariate normal:

$$\beta_j \mid D \sim \mathcal{N}\left(\frac{X_j^\top r_j}{X_j^\top X_j + \sigma^2/\sigma_b^2}, \frac{\sigma^2}{X_j^\top X_j + \sigma^2/\sigma_b^2}\right)$$

This corresponds to a **regularized least-squares update**. Residual updates **avoid matrix inversion**, scale to high dimensions, and extend naturally to **sparse (spike-and-slab)** models.

## Full Conditional for $\sigma_b^2$

The conditional distribution of the **prior variance**  $\sigma_b^2$ , given  $\beta$  and the hyperparameters, is a **scaled inverse-chi-squared**:

$$\sigma_b^2 \mid \beta \sim \tilde{S}_b \chi^{-2}(\tilde{v}_b)$$

where

$$\tilde{v}_b = v_b + p, \quad \tilde{S}_b = \frac{\beta^\top \beta + v_b S_b}{\tilde{v}_b}$$

At each Gibbs iteration,  $\sigma_b^2$  is sampled directly given  $\beta$ . This update reflects our revised belief about the **variability of effect sizes** after observing the current posterior draw of  $\beta$ .

## Full Conditional for $\sigma^2$

The conditional distribution of the **residual variance**  $\sigma^2$ , given  $\beta$  and the data, is also **scaled inverse-chi-squared**:

$$\sigma^2 \mid \beta, y \sim \tilde{S} \chi^{-2}(\tilde{v})$$

where

$$\tilde{v} = v + n, \quad \tilde{S} = \frac{(y - X\beta)^\top (y - X\beta) + vS}{\tilde{v}}$$

At each Gibbs iteration,  $\sigma^2$  is sampled directly given  $\beta$ . This captures our updated belief about the **residual variability** after accounting for the current linear predictor  $X\beta$ .

# Gibbs Sampling: Motivation

Bayesian inference often involves **complex posteriors** that lack closed-form solutions. To approximate these, we use **Markov Chain Monte Carlo (MCMC)** methods.

MCMC builds a **Markov chain** whose stationary distribution is the target posterior. Once the chain has **converged**, its samples can be used to estimate:

- Posterior means, variances, and credible intervals
- Predictive distributions
- Other functions of interest

Among MCMC algorithms, the **Gibbs sampler** is especially useful when all **full conditional distributions** are available in **closed form**.

# Gibbs Sampling: The Algorithm

For Bayesian linear regression with conjugate priors, the joint posterior is:

$$p(\beta, \sigma_b^2, \sigma^2 \mid y) \propto p(y \mid \beta, \sigma^2) p(\beta \mid \sigma_b^2) p(\sigma_b^2) p(\sigma^2)$$

We iteratively draw from the following **full conditionals**:

1. Sample  $\beta \mid \sigma_b^2, \sigma^2, y$
2. Sample  $\sigma_b^2 \mid \beta$
3. Sample  $\sigma^2 \mid \beta, y$

Each step updates one parameter given the latest values of the others. Repeating this sequence yields samples from the **joint posterior**  $p(\beta, \sigma_b^2, \sigma^2 \mid y)$ .

Because each conditional is **standard** (Normal or scaled inverse- $\chi^2$ ), Gibbs sampling is both **efficient** and **easy to implement**.



# Posterior Summaries

After running the Gibbs sampler, we obtain posterior draws  $\{\theta^{(t)}\}_{t=1}^T$  for parameters such as  $\beta_j$ ,  $\sigma^2$ , or  $\sigma_b^2$ .

We summarize the posterior distribution via:

- **Posterior mean**

$$\mathbb{E}[\theta \mid y] \approx \frac{1}{T} \sum_{t=1}^T \theta^{(t)}$$

- **Posterior median:** the median of  $\theta^{(t)}$
- **Credible interval (95%)**

$$[\theta]_{0.025}, [\theta]_{0.975}$$

These summaries describe the most probable values of  $\theta$  and their uncertainty after combining data and prior beliefs.

# Estimating Uncertainty

Bayesian inference provides **full posterior distributions**, not just point estimates. Uncertainty is quantified directly from the posterior samples:

- **Posterior standard deviation**

$$\text{SD}(\theta \mid y) \approx \sqrt{\frac{1}{T-1} \sum_{t=1}^T (\theta^{(t)} - \bar{\theta})^2}$$

The **width** of the credible interval reflects this uncertainty. Parameters with broader posteriors are estimated with less precision, and the degree of uncertainty depends on both the data and the prior.

# Posterior Prediction

Given a new observation  $x_{\text{new}}$ , we can predict using posterior draws:

1. Compute predicted means for each sample:

$$\hat{y}_{\text{new}}^{(t)} = x_{\text{new}}^{\top} \beta^{(t)}$$

2. Add residual uncertainty:

$$y_{\text{new}}^{(t)} \sim \mathcal{N}(x_{\text{new}}^{\top} \beta^{(t)}, \sigma^2(t))$$

The resulting samples  $\{y_{\text{new}}^{(t)}\}$  form a **posterior predictive distribution**, from which we can derive **predictive intervals** and evaluate **predictive accuracy**.

# Model Checking and Hypothesis Testing

Posterior samples enable rich **model diagnostics** and **hypothesis testing**:

- **Posterior probability of an event**

$$\Pr(\beta_j \neq 0 \mid y) \approx \frac{1}{T} \sum_{t=1}^T \mathbf{1}(\beta_j^{(t)} \neq 0)$$

- **Posterior predictive checks**

Simulate new datasets using posterior draws and compare them to the observed data to assess model fit.

- **Model comparison**

Bayes factors and marginal likelihoods can be approximated to formally test or compare competing models.

These tools extend Bayesian inference beyond estimation to **model validation**, **uncertainty quantification**, and **decision-making**.

# Convergence Diagnostics

Before interpreting MCMC results, we must check that the Gibbs sampler has **converged** to the target posterior distribution.

Convergence diagnostics assess whether the Markov chain has reached its **stationary distribution** and is producing valid samples.

Two basic strategies are:

- **Burn-in** – Discard early iterations (e.g., first 1000) to remove dependence on starting values.
- **Thinning** – Keep every  $k$ -th sample to reduce autocorrelation.

These steps improve sample quality and ensure reliable posterior summaries.

# Trace Plots

A simple yet powerful diagnostic is the **trace plot**, showing sampled parameter values  $\theta^{(t)}$  over iterations  $t$ .

- A **converged chain** fluctuates around a stable mean — no trend or drift.
- Multiple chains from different starting points should **overlap** and **mix well**.

Trace plots help detect: - Lack of stationarity (upward/downward trends) - Poor mixing or multimodality - Burn-in issues

Visual inspection is often the **first step** in assessing convergence.

# Autocorrelation

Samples from a Gibbs sampler are **correlated**, especially for tightly coupled parameters.

The **autocorrelation function (ACF)** quantifies dependence across lags  $k$ :

$$\hat{\rho}_k = \frac{\sum_{t=1}^{T-k} (\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^T (\theta^{(t)} - \bar{\theta})^2}$$

- High  $\hat{\rho}_k \rightarrow$  slow mixing and fewer effective samples
- Low  $\hat{\rho}_k \rightarrow$  better mixing and faster convergence

Reducing autocorrelation may require **more iterations**, **reparameterization**, or **thinning** the chain.

# Effective Sample Size (ESS)

Autocorrelation reduces the number of *independent* samples obtained.

The **effective sample size (ESS)** adjusts for this:

$$\text{ESS}(\theta) = \frac{T}{1 + 2 \sum_{k=1}^K \hat{\rho}_k}$$

- Small ESS  $\rightarrow$  chain is highly correlated, less informative
- Rule of thumb: ESS  $> 100$  per parameter for stable inference

ESS provides a quantitative measure of **sampling efficiency** and helps determine whether more iterations are needed.



## Gelman–Rubin Diagnostic ( $\hat{R}$ )

When running multiple chains, the **Gelman–Rubin statistic** compares between-chain and within-chain variability.

For  $m$  chains with  $T$  iterations each:

$$W = \frac{1}{m} \sum_{i=1}^m s_i^2, \quad B = \frac{T}{m-1} \sum_{i=1}^m (\bar{\theta}_i - \bar{\bar{\theta}})^2$$

The potential scale reduction factor:

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}}, \quad \hat{V} = \frac{T-1}{T} W + \frac{1}{T} B$$

- $\hat{R} \approx 1 \rightarrow$  convergence achieved
- $\hat{R} > 1.1 \rightarrow$  chains have **not converged**

## Geweke Diagnostic

The **Geweke test** checks whether early and late portions of a single chain have the same mean, indicating **stationarity**.

$$Z = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{\text{Var}(\bar{\theta}_A) + \text{Var}(\bar{\theta}_B)}}$$

Typically:

- Segment A = first 10% of the chain
- Segment B = last 50% of the chain

Under convergence,  $Z \sim \mathcal{N}(0, 1)$ .

- $|Z| \leq 2 \rightarrow$  chain likely stationary
- $|Z| > 2 \rightarrow$  potential non-convergence

These diagnostics ensure that posterior summaries reflect the **true target distribution**.

# Spike-and-Slab Bayesian Linear Regression

As in classical BLR, the outcome is modeled as:

$$y = Xb + e, \quad e \sim \mathcal{N}(0, \sigma^2 I_n)$$

where  $y$  is the  $n \times 1$  response,  $X$  the design matrix,  $b$  the regression coefficients, and  $\sigma^2$  the residual variance.

This defines the **likelihood**:

$$y \mid b, \sigma^2 \sim \mathcal{N}(Xb, \sigma^2 I_n)$$

The goal is to estimate  $b$  and identify which predictors truly contribute to  $y$ .

# Motivation for the Spike-and-Slab Prior

In standard Bayesian linear regression:

$$\beta_j \sim \mathcal{N}(0, \sigma_b^2)$$

This **Gaussian (shrinkage) prior** assumes all predictors have small effects, but it does **not allow exact zeros** — limiting variable selection.

The **spike-and-slab prior** addresses this by mixing two components:

- A **spike** at zero  $\rightarrow$  excluded predictors
- A **slab** (wide normal)  $\rightarrow$  active predictors

This yields **sparse**, interpretable models that select relevant variables.

# The Spike-and-Slab Mixture Prior

Each regression effect is drawn from a two-component mixture:

$$p(b_i \mid \sigma_b^2, \pi) = \pi \mathcal{N}(0, \sigma_b^2) + (1 - \pi) \delta_0$$

where:

- $\pi$  = prior probability that  $b_i$  is non-zero
- $\delta_0$  = point mass at zero

Thus, with probability  $\pi$  a predictor is active (slab), and with probability  $1 - \pi$  it is excluded (spike).

# Advantages of Spike-and-Slab Priors

This hierarchical mixture prior provides several benefits:

- **Sparsity** — allows exact zeros for irrelevant predictors
- **Interpretability** — binary indicators give posterior inclusion probabilities (PIPs)
- **Adaptivity** — the inclusion probability  $\pi$  is learned from the data
- **Balance** — captures both strong signals (detection) and small effects (prediction)

Hence, spike-and-slab models combine **variable selection** with **Bayesian uncertainty quantification**.

# Hierarchical Representation

We express each effect as:

$$b_i = \alpha_i \delta_i$$

where:

$$\alpha_i \mid \sigma_b^2 \sim \mathcal{N}(0, \sigma_b^2), \quad \delta_i \mid \pi \sim \text{Bernoulli}(\pi)$$

- $\alpha_i$ : effect size when predictor is included
- $\delta_i$ : binary inclusion indicator (0 or 1)

Marginalizing over  $\delta_i$  yields the spike-and-slab mixture prior above.

## Prior for the Inclusion Probability $\pi$

The overall sparsity level is controlled by  $\pi$ , assigned a **Beta prior**:

$$\pi \sim \text{Beta}(\alpha, \beta)$$

- Small  $\alpha$ , large  $\beta \rightarrow$  favor sparser models
- $\alpha = \beta = 1 \rightarrow$  uniform prior
- Larger  $\alpha \rightarrow$  denser models

This prior lets the **data determine the degree of sparsity**.



# Priors for Variance Components

Variance parameters use **scaled inverse-chi-squared** priors:

$$\sigma_b^2 \sim S_b \chi^{-2}(v_b), \quad \sigma^2 \sim S \chi^{-2}(v)$$

These are conjugate, providing closed-form conditional updates. Hyperparameters  $(S_b, v_b)$  and  $(S, v)$  encode prior beliefs about effect size variability and residual noise.

# Joint Posterior Structure

Combining the likelihood and priors, the joint posterior is:

$$p(\mu, \alpha, \delta, \pi, \sigma_b^2, \sigma^2 \mid y) \propto p(y \mid \mu, \alpha, \delta, \sigma^2) p(\alpha \mid \sigma_b^2) p(\delta \mid \pi) p(\pi) p(\sigma_b^2) p(\sigma^2)$$

This captures our **updated beliefs** about effects, inclusion indicators, and variance components.

# Gibbs Sampling for Spike-and-Slab BLR

Inference proceeds via **Gibbs sampling**, cycling through these conditional updates:

1.  $\alpha \mid D$
2.  $\delta \mid D$
3.  $\pi \mid D$
4.  $\sigma_b^2 \mid D$
5.  $\sigma^2 \mid D$

Here,  $D$  denotes the data and all other current parameter values. Each conditional follows a **standard distribution** (Normal, Bernoulli, Beta, scaled- $\chi^{-2}$ ). Iterating these updates generates samples from the joint posterior.

# Posterior Inclusion Probabilities

The **posterior inclusion probability (PIP)** measures how likely each predictor is truly associated with  $y$ :

$$\widehat{\Pr}(\delta_i = 1 \mid y) = \frac{1}{T} \sum_{t=1}^T \delta_i^{(t)}$$

- High PIP  $\rightarrow$  predictor is likely important
- Low PIP  $\rightarrow$  predictor likely irrelevant

PIPs summarize **variable relevance** and drive **Bayesian feature selection**.

# Summary of Bayesian Linear Regression

**Bayesian Linear Regression** combines **likelihood** and **prior** to form the **posterior**, enabling principled modeling, regularization, and uncertainty quantification.

- Inference via **MCMC** (often **Gibbs sampling**) with posterior draws for **means**, **credible intervals**, and **predictions**.
- **Spike-and-slab priors** enable **sparsity** and **variable selection**, assigning **exact zeros** to irrelevant predictors and identifying key variables via **posterior inclusion probabilities (PIPs)**.
- **Conjugate and mixture priors** yield **efficient** and **robust inference**, even when  $p > n$ .
- With **proper convergence checks**, Bayesian models provide **stable and reliable inference** across a wide range of data settings.

# Applications in Genomics

We have now seen the basic framework of **Bayesian Linear Regression (BLR)**

and will illustrate how it provides a **unified approach** for analyzing genetic and genomic data.

- **Genome-Wide Association Studies (GWAS)** and **fine-mapping** of causal variants.
- **Genetic prediction** and **heritability estimation**.
- **Pathway** and **gene-set enrichment** analyses.

These examples show how BLR connects **statistical modeling** with **biological interpretation** in quantitative genetics.