Introduction to Bayesian Linear Regression, Posterior Inference, and MCMC

Peter Sørensen

Table of contents

1 Introduction

Bayesian linear regression (BLR) extends the classical linear regression framework by incorporating prior information into the model and producing full posterior distributions over parameters, rather than single-point estimates. This approach offers several key advantages, particularly in the context of modern data analysis challenges such as high dimensionality, small sample sizes, and the need for uncertainty quantification.

In genomics and other biological applications, BLR is widely used for tasks such as mapping genetic variants, predicting genetic predisposition (e.g., polygenic risk scores), estimating genetic parameters like heritability, and performing gene set enrichment or pathway analyses. These applications benefit from BLR's ability to unify inference and prediction within a probabilistic framework.

The BLR model builds on the familiar linear regression formulation, where the observed outcome is modeled as a linear function of predictors plus Gaussian noise. However, unlike classical inference—which relies on least squares or maximum likelihood estimation and provides only point estimates and asymptotic intervals—Bayesian inference yields full posterior distributions over the unknown coefficients and variance. This allows for richer uncertainty quantification and more robust inference.

Several motivations drive the use of Bayesian methods in linear regression. First, BLR naturally quantifies uncertainty through posterior distributions, allowing the analyst to compute credible intervals, posterior probabilities, and predictive distributions. Second, prior distributions act as regularizers, helping to stabilize estimation in noisy or underdetermined settings, such as when the number of predictors p exceeds the number of observations n. Gaussian priors encourage shrinkage toward zero, while more structured priors (such as spike-and-slab) enable sparse or grouped solutions. Third, BLR makes it straightforward to incorporate external knowledge—such as biological relevance or prior experimental results—into the modeling process.

These notes begin by reviewing the classical linear regression model and its limitations. We then introduce the Bayesian linear regression model, outline the inference workflow, and show how to derive the full conditional posterior distributions for the model parameters using conjugate priors. Finally, we describe how posterior inference is performed using Gibbs sampling and conclude with practical considerations for implementation, diagnostics, and applications in R.

2 Classical Linear Regression

Classical linear regression is one of the most widely used statistical modeling tools. It provides a simple yet powerful framework for modeling the relationship between a response variable and a set of predictor variables. The goal is to estimate how changes in the predictors X affect the outcome y, assuming a linear relationship and normally distributed errors.

2.0.1 Model Specification

We start by specifying the standard linear model:

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

Here:

- y is the $n \times 1$ vector of observed outcomes,
- X is the $n \times p$ design matrix of predictors (which may include an intercept),
- β is the $p \times 1$ vector of unknown regression coefficients,
- σ^2 is the residual (error) variance,
- and ϵ is the vector of i.i.d. normal errors with mean zero and variance σ^2.

Because the errors are Gaussian, the distribution of y is:

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

This defines the **likelihood**—the probability model for the observed data given the parameters.

2.0.2 Parameter Estimation via OLS and MLE

In classical regression, parameters are estimated using **Ordinary Least Squares (OLS)**, which minimizes the residual sum of squares. Under the assumption of normally distributed errors, these estimators also correspond to the **Maximum Likelihood Estimators (MLE)**.

The OLS estimate for the regression coefficients is:

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

This estimator is only valid when $X^\top X$ is invertible, which requires that the predictors are linearly independent and that $n \geq p$.

The residual variance is estimated using:

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

which provides a measure of the average squared distance between the observed and fitted values.

2.0.3 Inference on Regression Coefficients

Once the model parameters are estimated, we can assess uncertainty and perform hypothesis tests.

The estimated **variance-covariance matrix** of $\hat{\beta}$ is:

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2 (X^\top X)^{-1}$$

From this, the **standard error** for each estimated coefficient $\hat{\beta}_j$ is:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \left[(X^\top X)^{-1}\right]_{jj}}$$

To test whether a coefficient is significantly different from zero, we use the ***t*-statistic**:

$$t_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$$

Under the null hypothesis $H_0 : \beta_j = 0$, and assuming Gaussian errors, this statistic follows a t-distribution with $n - p$ degrees of freedom.

A $100(1-\alpha)\%$ **confidence interval** for β_j is given by:

$$\hat{\beta}_j \pm t_{n-p,\,\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_j)$$

These tools allow us to quantify uncertainty in our parameter estimates and test hypotheses about individual effects.

2.0.4 Prediction and Uncertainty in Predictions

For a new observation x_{new} (a $p \times 1$ vector), the predicted mean response is:

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

The variance of this predicted response reflects both the uncertainty in the coefficient estimates and the inherent noise in the data:

$$\widehat{\mathrm{Var}}(\hat{y}_{\text{new}}) = \hat{\sigma}^2 \left(1 + x_{\text{new}}^\top (X^\top X)^{-1} x_{\text{new}}\right)$$

This leads to **prediction intervals** that are wider than confidence intervals for β because they account for additional variability in future observations.

2.0.5 Limitations of Classical Linear Regression

While classical linear regression is simple and interpretable, it has several well-known limitations:

- It does not allow for the incorporation of **prior knowledge** about parameters.
- There is **no explicit control** over the distribution of effect sizes.
- The model becomes **non-identifiable** when the number of predictors exceeds the number of observations ($p > n$), since $X^\top X$ is not invertible.
- Estimates can be unstable or highly variable in the presence of **multicollinearity** (highly correlated predictors).
- Uncertainty quantification relies on **asymptotic results** or the assumption of Gaussian errors, which may not always hold in practice.

These limitations motivate the development and application of **Bayesian linear regression**, which extends the classical framework by incorporating prior distributions and producing full posterior distributions for all unknown parameters.

3 Bayesian Linear Regression with Gaussian priors

Bayesian linear regression starts with the same model structure as classical linear regression. We assume the outcome vector y is generated from a linear function of predictors X with additive Gaussian noise. Specifically, the model is written as:

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

Here, y is an $n \times 1$ vector of observed outcomes, X is the $n \times p$ design matrix of predictors, β is the $p \times 1$ vector of unknown regression coefficients, and e is a vector of random errors assumed to be independent and identically distributed (i.i.d.) Gaussian noise with mean zero and constant variance σ^2. Because the residuals are Gaussian, it follows that the marginal distribution of y is:

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

This defines the **likelihood**—the probability model for the observed data, conditional on the unknown parameters β and σ^2.

3.0.1 Prior Distributions

To perform Bayesian inference, we must specify **prior distributions** that encode our beliefs about the parameters before observing the data.

A commonly used **conjugate prior** for the regression coefficients β is a multivariate normal distribution centered at zero:

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

This prior reflects a belief that most effect sizes are small and centered near zero, consistent with the **polygenic assumption** in genetics. The parameter σ_b^2 is the **prior variance** and acts as a **shrinkage parameter**:

- When σ_b^2 is small, the prior strongly favors values of β near zero, resulting in more shrinkage of estimates.
- When σ_b^2 is large, the prior becomes more diffuse, allowing for larger effect sizes and less shrinkage.

Thus, σ_b^2 controls the **prior belief about the magnitude of effect sizes**, and is often treated as an unknown hyperparameter to be estimated from the data (e.g., via hierarchical modeling or Gibbs sampling).

In addition, it is common to place **scaled inverse-chi-squared distributions** on the two variance parameters, σ_b^2 and σ^2:

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

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

Here:

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

These priors are **conjugate** to the Gaussian likelihood and the normal prior on β, which means they lead to posterior distributions in the same family (i.e., scaled inverse-chi-squared or Inverse-Gamma). This conjugacy simplifies derivations and enables closed-form **Gibbs sampling** steps.

These priors not only express prior knowledge or assumptions but also act as **regularizers**. In particular:

- The prior on β shrinks small/noisy effect estimates toward zero.
- The priors on variance parameters prevent overfitting and stabilize inference, especially in **high-dimensional** scenarios where $p > n$.

This makes conjugate priors a practical and computationally efficient choice in Bayesian linear regression models.

3.0.2 Posterior Distribution

The core of Bayesian analysis is to combine the likelihood with the prior distributions using **Bayes' rule**, which yields 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 encapsulates all updated knowledge about the unknown parameters after observing the data. It is the key quantity of interest in Bayesian inference and serves as the basis for computing summaries such as posterior means, credible intervals, or predictions.

Because we are using **conjugate priors**, the full conditional distributions of the parameters have **closed-form solutions**, which makes **Gibbs sampling** a natural and efficient inference strategy. In Gibbs sampling, we alternately sample from the conditional distributions of each parameter given the others.

3.0.2.1 Full Conditional for β

The full conditional distribution of the regression coefficients β, given σ^2 and the observed data y, is a multivariate normal distribution. We can write this in a compact and interpretable form using the **conditional posterior mean** μ_β and **conditional posterior covariance matrix** Σ_β:

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

where:

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

This conditional distribution reflects our updated belief about the regression coefficients after observing y, while conditioning on fixed values of σ_b^2 and σ^2.

3.0.2.2 Full Conditional for β_j

In practice, rather than sampling the entire vector β jointly, we can update each coefficient β_j **one at a time**, holding all others fixed. This is often more efficient for large p, and is particularly useful in Gibbs sampling frameworks like spike-and-slab models.

Let X_j be the jth column of the design matrix, and define the **partial residual**:

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

where X_{-j} is the matrix with the jth column removed, and β_{-j} is the vector of all coefficients except β_j.

This coordinate-wise update strategy is justified because the full conditional distribution of β is multivariate normal. When all other coefficients are held fixed, the conditional distribution of a single coefficient β_j given the data and remaining parameters is univariate normal. This is a standard result from the theory of the multivariate normal distribution, which implies that any subset of variables also follows a (conditional) normal distribution.

The full conditional for β_j is:

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

This update can be derived directly from the Gaussian likelihood and Gaussian prior on β_j, and corresponds to a regularized least-squares update. By cycling through all $j = 1, \ldots, p$ using the current residuals, we efficiently obtain samples from the full conditional distribution of β.

Residual updates are also attractive because they **avoid matrix inversion**, scale well to high dimensions, and naturally extend to models with sparsity indicators (e.g., spike-and-slab).

3.0.2.3 Comparison to Classical OLS

Recall that in classical linear regression, the **ordinary least squares (OLS)** estimator of β is:

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

This estimator is obtained by maximizing the likelihood under the assumption of Gaussian errors:

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

While the likelihood depends on σ^2, it **cancels out** when estimating β via OLS or MLE, because it only affects the **scale** of the likelihood, not the location of the maximum. As a result, the estimate of β is **independent of** σ^2.

In contrast, the Bayesian formulation includes prior information, and σ^2 appears explicitly in the posterior:

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

This introduces **dependence on** σ^2, meaning that uncertainty about the data also affects the precision of our belief about β.

Moreover, the additional term $\frac{I}{\sigma_b^2}$ in the posterior precision matrix encodes prior information about effect sizes. This shrinks the estimates toward zero and helps **regularize** the inference, particularly when $p > n$ or when predictors are highly correlated.

Thus, the Bayesian posterior mean:

$$\mu_\beta = \Sigma_\beta \cdot \frac{X^\top y}{\sigma^2}$$

can be seen as a **regularized, uncertainty-aware generalization** of the OLS estimate.

3.0.2.4 Full Conditional for σ_b^2

The full conditional distribution of the prior variance σ_b^2, given the current values of β and the hyperparameters, is a **scaled inverse-chi-squared distribution**:

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

where:

- $\tilde{v}_b = v_b + p$ is the updated degrees of freedom, with p the number of regression coefficients,
- $\tilde{S}_b = \dfrac{\beta^\top \beta + v_b S_b}{\tilde{v}_b}$ is the updated scale parameter.

This form is convenient for **Gibbs sampling**: at each iteration, a new value of σ_b^2 can be sampled directly, given the current value of β. It reflects our updated belief about the variability of the regression coefficients after observing the current posterior draw of β.

3.0.2.5 Full Conditional for σ^2

The full conditional distribution of the residual variance σ^2, given the current values of β and the data, is a **scaled inverse-chi-squared distribution**:

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

where:

- $\tilde{v} = v + n$ is the updated degrees of freedom, with n the number of observations,
- $\tilde{S} = \dfrac{(y - X\beta)^\top (y - X\beta) + vS}{\tilde{v}}$ is the updated scale parameter.

This form is convenient for **Gibbs sampling**: at each iteration, a new value of σ^2 can be sampled directly, given the current values of β. It reflects our updated belief about the residual variability in the data after accounting for the current linear predictor $X\beta$.

4 Gibbs Sampling

Bayesian inference often requires sampling from complex **posterior distributions** that cannot be computed analytically. In such cases, we rely on **Markov Chain Monte Carlo (MCMC)** methods to approximate the posterior using a sequence of dependent samples.

MCMC algorithms construct a **Markov chain** whose stationary distribution is the target posterior. Once the chain has **converged**, the sampled values can be used to estimate posterior expectations, make predictions, and conduct inference.

One of the simplest and most widely used MCMC algorithms is the **Gibbs sampler**. Gibbs sampling is especially convenient when all **full conditional distributions** of the model parameters are available in closed form.

In the Bayesian linear regression model with conjugate priors, the joint posterior distribution 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 can implement a Gibbs sampler by iteratively drawing 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 conditional on the latest values of the others. Repeating this sequence over many iterations yields samples from the **joint posterior** $p(\beta, \sigma_b^2, \sigma^2 \mid y)$.

Because each conditional distribution is standard (normal or scaled inverse-chi-squared), sampling is straightforward and efficient. Once the Gibbs sampler has **converged**, these posterior draws form the basis for inference.

4.1 Posterior Summaries and Inference from Gibbs Samples

After running the Gibbs sampler and obtaining T posterior draws of all parameters, we can use these samples to compute a wide range of quantities relevant to Bayesian inference. These include:

- **Posterior means and medians** as point estimates of parameters
- **Credible intervals** to quantify uncertainty
- **Posterior standard deviations** as measures of variability
- **Posterior probabilities** of hypotheses, such as $\Pr(\beta_j > 0 \mid y)$
- **Posterior predictive distributions** for new observations
- **Model diagnostics** such as convergence checks or residual analysis

These quantities allow us to summarize uncertainty, generate predictions, and make probabilistic statements about model parameters and data.

4.1.0.1 Posterior Summaries

Once we have a collection of posterior draws for a parameter θ (e.g., β_j, σ^2, or σ_b^2), we can summarize the posterior distribution using:

- **Posterior mean**:
$$\mathbb{E}[\theta \mid y] \approx \frac{1}{T} \sum_{t=1}^{T} \theta^{(t)}$$
- **Posterior median**: The median value of the sampled $\theta^{(t)}$.
- **Credible intervals**: For example, a 95% credible interval for θ can be obtained as the 2.5% and 97.5% quantiles of the posterior samples:
$$[\theta]_{0.025}, [\theta]_{0.975}$$

These summaries provide insight into the likely values of the parameter after accounting for uncertainty in both the data and prior beliefs.

4.1.0.2 Estimating Uncertainty

Bayesian inference provides **full posterior distributions**, not just point estimates. This allows us to directly quantify the uncertainty of parameters:

- **Posterior standard deviation**:
$$\mathrm{SD}(\theta \mid y) \approx \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} \left(\theta^{(t)} - \bar{\theta}\right)^2}$$
- This uncertainty is reflected in the **width** of the credible intervals and can vary across different parameters or under different priors.

4.1.0.3 Prediction

Given a new observation x_{new}, we can generate **posterior predictive distributions** using the sampled parameter values:

1. For each draw t, compute:
$$\hat{y}_{\text{new}}^{(t)} = x_{\text{new}}^{\top} \beta^{(t)}$$
2. Optionally, add residual noise from the corresponding draw of $\sigma^{2(t)}$:
$$y_{\text{new}}^{(t)} \sim \mathcal{N}\left(x_{\text{new}}^{\top} \beta^{(t)},\ \sigma^{2(t)}\right)$$
3. Use these $y_{\text{new}}^{(t)}$ samples to construct predictive intervals or evaluate predictive performance.

4.1.0.4 Model Checking and Hypothesis Testing

The posterior draws can also be used for **model diagnostics** or **hypothesis testing**:

- **Posterior probability of an event**, such as a non-zero effect:

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

- **Posterior predictive checks**: Simulate new datasets from the model using posterior draws and compare them to the observed data. Discrepancies may indicate model misfit.
- **Bayes factors** and **marginal likelihoods** can be computed or approximated for formal hypothesis testing or model comparison, though these often require specialized methods beyond standard Gibbs output.

These procedures allow us to move beyond point estimates and engage in a full Bayesian analysis that accounts for uncertainty in parameter estimation, prediction, and decision-making.

5 Convergence Diagnostics for Gibbs Sampling

Before interpreting results from a Gibbs sampler, it is crucial to assess whether the sampler has **converged** to the target posterior distribution. Convergence diagnostics help determine if the Markov Chain has reached its stationary distribution and is providing valid samples.

5.0.0.1 Burn-in and Thinning

- **Burn-in**: Discard initial samples (e.g., first 1000 iterations) to allow the chain to reach stationarity.
- **Thinning**: Keep every k-th sample to reduce autocorrelation. This helps with storage but does not improve convergence.

5.0.0.2 Trace Plots

A simple but effective tool is the **trace plot**: plotting sampled values of a parameter (e.g., $\beta_j^{(t)}$) against iteration number t:

- A converged chain should resemble a **stationary process** with no apparent trend.
- Multiple chains started from different initial values should **mix well** and overlap.

5.0.0.3 Autocorrelation

Gibbs samples are often correlated. We assess this using the **autocorrelation function (ACF)**:

- For lag k, the sample autocorrelation of parameter θ is:

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

- High autocorrelation suggests **slow mixing**, requiring longer chains or thinning.

5.0.0.4 Effective Sample Size

The **effective sample size** (ESS) adjusts for autocorrelation and reflects the number of independent samples:

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

- A small ESS means the chain is highly autocorrelated and less informative.
- As a rule of thumb, aim for ESS > 100 per parameter.

5.0.0.5 Gelman–Rubin Diagnostic (R)

When running **multiple chains**, the Gelman–Rubin statistic $\hat{R}$ compares between-chain and within-chain variance:

1. Let m be the number of chains and T the number of iterations per chain.
2. For each parameter θ, compute:
   - The within-chain variance:

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2$$

   - The between-chain variance:

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

3. The **potential scale reduction factor** is:

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

- A value $\hat{R} \approx 1$ indicates convergence.
- Values $\hat{R} > 1.1$ suggest that the chain has **not converged**.

5.0.1 Geweke Diagnostic

The **Geweke diagnostic** tests for stationarity by comparing the means of two segments of a single chain:

- Typically, the **first 10%** and the **last 50%** of the chain are used.
- For a parameter θ, the test statistic is:

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

  where:
  - $\bar{\theta}_A$ is the mean of the early window,
  - $\bar{\theta}_B$ is the mean of the late window.
- Under the null hypothesis of stationarity, Z approximately follows a standard normal distribution.

Values of Z far from zero (e.g., $|Z| > 2$) suggest that the chain has **not converged**, as early and late samples differ systematically.

Monitoring these diagnostics ensures that posterior summaries and predictions are based on reliable samples from the true posterior distribution.

6 Bayesian Linear Regression with Spike-and-Slab Priors

Similar to the Bayesian linear regression model with Gaussian priors, we begin by specifying the **likelihood** for the observed data. The response vector y is assumed to follow a Gaussian distribution, conditional on the regression parameters:

$$y \mid \mathbf{b}, \sigma^2 \sim \mathcal{N}\left(X\mathbf{b},\ \sigma^2 I_n\right)$$

where y is an $n \times 1$ vector of observed outcomes, X is an $n \times m$ design matrix of predictor variables, $\mathbf{b}$ is an $m \times 1$ vector of regression coefficients, and σ^2 is the residual variance. This defines the data-generating process: given the regression coefficients and residual variance, the outcomes are normally distributed around the linear predictor $X\mathbf{b}$.

In standard Bayesian linear regression (BLR), each regression coefficient β_j is typically assigned a Gaussian prior:

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

This reflects the belief that all predictors may contribute to the outcome, with effect sizes centered around zero and uncertainty governed by the prior variance σ_b^2. Such **shrinkage priors** perform well in settings where many small effects are expected. However, they do **not permit exact zeros**, limiting their utility for **variable selection** or enforcing **sparsity**.

To address this, we adopt a hierarchical model structure using **spike-and-slab priors**, a type of **two-component mixture model**. Conditional on the regression coefficients, the outcomes y follow a Gaussian distribution as above. At the second level of the hierarchy, however, each regression effect is assumed to arise from one of two components:

- A **slab**: a diffuse Gaussian distribution representing non-zero effects.
- A **spike**: a point mass at zero representing exactly zero effects.

This formulation allows us to directly model sparsity. Specifically, each coefficient b_i is assumed to follow the mixture prior:

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

where δ_0 denotes a point mass at zero, and π is the prior probability that b_i is non-zero.

Compared to standard Gaussian priors, **spike-and-slab priors** allow for exact zeros in regression coefficients. This enables **automatic variable selection** within a fully Bayesian framework, combining interpretability with uncertainty quantification.

The resulting **two-component mixture prior** offers several key advantages:

- **Sparsity**: Supports exact zeros in the coefficient vector, allowing the model to exclude irrelevant predictors.
- **Interpretability**: Posterior samples of the binary inclusion variables δ_i yield **posterior inclusion probabilities (PIPs)**, which help identify important predictors.

- **Adaptivity**: By placing a **Beta prior** on the sparsity parameter π, the model can learn the degree of sparsity directly from the data.
- **Prediction–detection trade-off**: The mixture structure balances the inclusion of small, potentially weak effects (for prediction) with the identification of stronger signals (for detection).

In summary, the spike-and-slab prior extends Bayesian linear regression to high-dimensional settings by enabling principled variable selection and adaptive regularization. The following sections derive the full conditional distributions used for inference via Gibbs sampling.

6.0.1 Prior Distributions

6.0.1.1 Spike-and-Slab Prior for Regression Effects

To explicitly model sparsity, we use a **spike-and-slab prior**, which introduces a hierarchical structure. Each regression coefficient b_i is expressed as:

$$b_i = \alpha_i \cdot \delta_i, \quad i = 1, \dots, m$$

Here, δ_i is a binary **inclusion indicator**, and α_i is the effect size when the predictor is active. We place the following priors:

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

That is, each δ_i is independently drawn from a Bernoulli distribution with success probability π, which represents the **a priori probability** that predictor i is relevant.

Marginalizing over δ_i, the prior for b_i becomes a two-component mixture:

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

where δ_0 denotes a point mass at zero. This expresses that with probability π, b_i is drawn from a Gaussian ("slab"), and with probability $1 - \pi$, it is exactly zero ("spike").

The parameter π controls the overall sparsity of the model. Importantly, the prior inclusion probability π is distinct from the **posterior inclusion probability** $\Pr(\delta_i = 1 \mid y)$, which is inferred from the data. A simple Monte Carlo estimate of this posterior probability is the average value of δ_i across samples from a Gibbs sampler.

6.0.1.2 Prior for the Inclusion Probability π

Rather than fixing π in advance, we often treat it as a random variable and assign it a **Beta prior**:

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

This prior is defined on the interval $[0, 1]$ and allows the data to inform the level of sparsity. The choice of (α, β) reflects prior beliefs:

- Small α and large β favor sparse models (most effects are zero).
- $\alpha = \beta = 1$ gives a uniform prior.
- Larger α relative to β favors denser models.

Because of conjugacy with the Bernoulli prior on δ_i, the posterior update of π is straightforward in Gibbs sampling.

6.0.1.3 Priors for Variance Parameters

As in the Gaussian BLR model, we assign **scaled inverse-chi-squared priors** to the variance components:

- For the prior variance of the effect sizes:

$$\sigma_b^2 \sim S_b \cdot \chi^{-2}(v_b)$$

- For the residual variance:

$$\sigma^2 \sim S \cdot \chi^{-2}(v)$$

These conjugate priors allow for closed-form updates in Gibbs sampling. The hyperparameters (S_b, v_b) and (S, v) encode prior beliefs about the variability of the coefficients and the residuals, and can be tuned to reflect prior knowledge or set to weakly informative values when such knowledge is limited.

6.0.2 Posterior Distribution

In the spike-and-slab Bayesian linear regression model, we introduce hierarchical priors for sparsity and variance components. The **joint prior distribution** over all model parameters factorizes as:

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

with the components defined as follows:

- $p(\alpha \mid \sigma_b^2)$: Normal priors on the latent effect sizes,
- $p(\delta \mid \pi)$: Bernoulli priors on binary inclusion indicators,

- $p(\pi)$: Beta prior on the inclusion probability,
- $p(\sigma_b^2)$ and $p(\sigma^2)$: Scaled inverse-chi-squared priors for the variance components.

These priors encode our initial beliefs about sparsity and effect magnitudes before seeing the data.

Combining the prior structure with the likelihood using **Bayes' rule**, we obtain the **joint posterior distribution** of all unknown parameters given the data y:

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

This expression defines the complete probabilistic model and captures our **updated beliefs** about the intercept μ, the regression effects α, the sparsity indicators δ, the prior inclusion probability π, and both variance parameters after observing the data.

Since this posterior is analytically intractable, inference proceeds via **Gibbs sampling**, where each parameter block is updated iteratively from its full conditional distribution.

6.0.3 Gibbs Sampling

In this hierarchical Bayesian model with spike-and-slab priors, all **full conditional posterior distributions** are of known standard form. This allows us to use **Gibbs sampling**, where each parameter is sampled from its conditional distribution given the data and all other current parameter values.

At each iteration of the Gibbs sampler, we cycle through the following updates:

$$[\mu \mid D], \quad [\alpha \mid D], \quad [\delta \mid D], \quad [\pi \mid D], \quad [\sigma_b^2 \mid D], \quad [\sigma^2 \mid D]$$

where D denotes the observed data and all other current parameter values. The remainder of this section outlines the key updates for each block.

6.0.3.1 Updating Effect Sizes α_i

Each latent effect α_i has a conditional posterior distribution that depends on whether the corresponding inclusion indicator δ_i is 0 or 1.

If $\delta_i = 0$, the effect $b_i = \alpha_i \cdot \delta_i = 0$ is excluded from the model, and the likelihood does **not** depend on α_i. In this case, α_i is not identifiable from the data, and its posterior is proportional to its prior:

$$p(\alpha_i \mid D, \delta_i = 0) \propto \mathcal{N}(0, \sigma_b^2)$$

Because α_i has no effect on the likelihood when $\delta_i = 0$, practical implementations typically set $b_i = 0$.

If $\delta_i = 1$, the effect contributes to the likelihood. Define the **partial residual** that excludes the contribution from predictor i:

$$r_i = y - 1\mu - X_{-i}b_{-i}$$

Then, the full conditional for α_i is Gaussian with mean and variance:

$$\alpha_i \mid D \sim \mathcal{N}\left(\frac{X_i^\top r_i}{X_i^\top X_i + \sigma^2/\sigma_b^2},\ \frac{\sigma^2}{X_i^\top X_i + \sigma^2/\sigma_b^2}\right)$$

This update corresponds to a **shrinkage estimator** of α_i that balances fit to the data (via $X_i^\top r_i$) with regularization (via σ_b^2).

6.0.3.2 Updating Inclusion Indicators δ_i

Each indicator $\delta_i \in \{0, 1\}$ determines whether the ith predictor is included in the model. Its full conditional is a **Bernoulli distribution** with success probability based on comparing the model fit with and without the predictor.

Let:

- RSS_0: residual sum of squares with $\delta_i = 0$,
- RSS_1: residual sum of squares with $\delta_i = 1$.

Then:

$$\Pr(\delta_i = 1 \mid D) = \frac{\exp\left(-\frac{1}{2\sigma^2}\text{RSS}_1\right)\pi}{\exp\left(-\frac{1}{2\sigma^2}\text{RSS}_0\right)(1-\pi) + \exp\left(-\frac{1}{2\sigma^2}\text{RSS}_1\right)\pi}$$

To sample δ_i, compute this probability and draw from the Bernoulli distribution.

6.0.3.2.1 Numerically Stable Version (Using Log-Odds)

To avoid numerical underflow when RSS values are large, compute the **log-odds**:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \frac{1}{2\sigma^2}(\text{RSS}_0 - \text{RSS}_1) - \log\left(\frac{1-\pi}{\pi}\right)$$

Then recover the probability θ_i using the **inverse-logit** (logistic) function:

$$\theta_i = \frac{\exp(K_i)}{1 + \exp(K_i)}$$

This provides a stable way to compute the probability of inclusion, especially when likelihood differences are large.

6.0.3.3 Updating π

With prior $\pi \sim \text{Beta}(\eta, \beta)$ and $\delta_i \sim \text{Bernoulli}(\pi)$, the conditional posterior is:

$$\pi \mid D \sim \text{Beta}\left(\sum_{i=1}^{m} \delta_i + \eta,\ m - \sum_{i=1}^{m} \delta_i + \beta\right)$$

6.0.3.4 Updating σ_b^2

The full conditional distribution of the prior variance σ_b^2, given the current values of the effect sizes α_i and inclusion indicators δ_i, is a **scaled inverse-chi-squared distribution**:

$$\sigma_b^2 \mid \alpha, \delta \sim \tilde{S}_b \cdot \chi^{-2}(\tilde{v}_b)$$

where:

- $p = \sum_{i=1}^{m} \delta_i$ is the number of included (non-zero) effects,
- $\tilde{v}_b = v_b + p$ is the updated degrees of freedom,
- $\tilde{S}_b = \dfrac{\sum_{i=1}^{m} \delta_i \alpha_i^2 + v_b S_b}{\tilde{v}_b}$ is the updated scale parameter.

This update accounts only for those coefficients currently included in the model ($\delta_i = 1$), reflecting the prior belief that excluded effects are exactly zero and thus do not contribute to the variance estimate.

This form allows direct sampling of σ_b^2 at each Gibbs iteration and reflects the updated uncertainty about the size of the non-zero regression effects.

6.0.3.5 Updating σ^2

The full conditional distribution of the residual variance σ^2, given the current values of β and the data, is a **scaled inverse-chi-squared distribution**:

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

where:

- $\tilde{v} = v + n$ is the updated degrees of freedom, with n the number of observations,
- $\tilde{S} = \dfrac{(y - X\beta)^\top (y - X\beta) + vS}{\tilde{v}}$ is the updated scale parameter.

This form is convenient for **Gibbs sampling**: at each iteration, a new value of σ^2 can be sampled directly, given the current values of β. It reflects our updated belief about the residual variability in the data after accounting for the current linear predictor $X\beta$.

This completes one iteration of the Gibbs sampler. Each step updates parameters from their full conditional distributions, enabling efficient posterior inference under the spike-and-slab prior.

6.0.3.6 Posterior inference

Each step in the Gibbs sampler involves only standard distributions (Gaussian, Bernoulli, Beta, scaled-inverse-chi-squared), allowing efficient and scalable posterior inference. Iterating these updates produces samples from the joint posterior, which can be used to estimate marginal posterior summaries such as:

- Posterior means or medians of effects,
- Posterior inclusion probabilities,
- Credible intervals for regression coefficients,
- Model sparsity levels.

6.0.4 Posterior Inclusion Probability

While π defines a global prior probability of inclusion, the **posterior inclusion probability** $\Pr(\delta_i = 1 \mid y)$ is computed **separately for each marker** after observing the data.

A Monte Carlo estimator of this probability is:

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

where $\delta_i^{(t)}$ is the sampled value of δ_i in iteration t of the Gibbs sampler.

This posterior quantity reflects our **updated belief** about whether each marker is truly associated with the trait, and is a key quantity used in Bayesian fine-mapping.