Introduction to Bayesian Linear Regression Models used in Gene Set Analyses

Peter Sørensen

Bayesian Linear Regression Models

Bayesian Linear Regression Models provide a flexible statistical framework for modeling complex biological and healthcare data. They support key applications such as:

- **Genome-wide association studies (GWAS)** and **fine-mapping** of causal variants
- **Polygenic risk scoring (PRS)** for predicting complex traits and disease risk
- **Gene and pathway enrichment analyses** to test biological hypotheses
- **Integrative multi-omics modeling** across the genome, transcriptome, epigenome, and proteome
- Applications to **registry-based healthcare data**, enabling population-level **risk prediction** and **disease modeling**

The Bayesian Linear Regression Model

The **Bayesian Linear Regression (BLR)** model builds:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

- $\mathbf{Y}$ represents the **observed outcomes** or **association measures** corresponding to the features in $\mathbf{X}$.
- $\mathbf{X}$ represents **molecular or genomic predictors** (e.g., genotypes, gene scores, annotations, pathway indicators).
- β — **effect sizes** quantifying how features in $\mathbf{X}$ explain variation in $\mathbf{Y}$
- ε — **residual noise** capturing unexplained variation

In the **Bayesian** formulation, each β_j is assigned a **prior distribution** reflecting beliefs about **effect size magnitude or sparsity** and determine how **information is shared** across features or biological layers.

Why Bayesian Linear Regression Models?

Regression effects can be estimated in many ways, but we focus on a **Bayesian hierarchical framework** because it:

- **Combines data and prior knowledge** to improve inference
- Provides a **natural way to regularize** and handle noisy or high-dimensional data
- Enables **flexible modeling** of diverse effect patterns:
  - Many small vs. few large effects
  - Structured effects (e.g., by pathway, gene set, or omic layer)
- Returns **uncertainty estimates** for all parameters $\rightarrow$ improving interpretability and model comparison

Through their **hierarchical structure**, BLR models naturally **integrate multiple biological layers** — linking genomic, transcriptomic, and other molecular data

Hyperpriors in Bayesian Modeling

The BLR model extends the linear framework by introducing **hierarchical priors** on parameters. Suppose we model regression coefficients b_j with a normal prior:

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

Here, σ_b^2 (the variance of the prior) controls how large the effects b_j are expected to be.

Instead of fixing σ_b^2 we treat it as unknown and assign it its own prior — the **hyperprior**:

$$\sigma_b^2 \sim \text{Inv-}\chi^2(\nu, S^2)$$

In practice, this parameter is **learned from the data** during estimation rather than fixed in advance.

Hierarchical Structure

The three levels in the model:

Level	Description	Example
1	Describes how data are generated given parameters	$y \sim \mathcal{N}(Xb, \sigma^2 I)$
2	Describes our beliefs about the parameters before seeing data	$b_i \sim \mathcal{N}(0, \sigma_b^2)$
3	Describes uncertainty about the prior's parameters	$\sigma_b^2 \sim \text{Inv-}\chi^2(\nu, S^2)$

This hierarchical structure allows the model to learn how strongly to shrink effect estimates from the data, while accounting for uncertainty in prior parameters and automatically regularizing effect sizes.

Simple and robust, but may not capture diverse effect-size distributions.

Adapting to Complex Biological Architectures

Complex traits arise from **heterogeneous effect-size distributions** — some features have large effects, many have small, and others are likely null. To capture this diversity, the BLR framework can be extended in two ways:

- **Data-driven grouping** of molecular features:
  The model *learns effect-size classes* from the data using a **mixture of variances** $\{\tau_k^2\}$ with probabilities $\{\pi_k\}$.
- **Biologically informed grouping** of molecular features:
  The model uses *prior biological knowledge* to assign features to groups *a priori*, each with its own variance τ_g^2 capturing within-group variability.

Both approaches enable the model to **adapt to complex genetic and molecular architectures** and **share information** across related features or omic layers

Multiple-Component BLR

A more flexible formulation assumes that effects come from a **mixture of normal distributions**:

$$\beta_j \mid d_j = k \sim \mathcal{N}(0, \tau_k^2), \qquad P(d_j = k) = \pi_k$$

- d_j is a **latent indicator variable** assigning effect j to component k.
- Each component k has its own variance τ_k^2, controlling expected effect size.
- π_k represents the **probability of membership** in each component.

This structure allows the model to **capture both large and small effects**, including potential nulls.
In practice, both π_k and τ_k^2 are **learned from the data**, making the model **data-driven and adaptive**, though not necessarily **biologically informed**.

Hierarchical (Group-Structured) BLR

In this **biologically informed extension**, we assign features to predefined **groups** (e.g., genes, pathways, or protein complexes) *a priori* and model group-specific variances:

$$\beta_j \sim \mathcal{N}(0, \tau^2_{g(j)}), \qquad \tau^2_g \sim p(\tau^2_g)$$

- Each feature j belongs to a group $g(j)$ defined before analysis.
- Each group g has its own variance τ^2_g, controlling effect size variability within that group.

This approach allows the model to **share information among related features** and **test enrichment across biological groups**. In practice, the group-level variances τ^2_g are **learned from the data**, enabling the model to **adapt shrinkage across biological structures**.

Indicator Variables and Posterior Inclusion Probabilities

In Bayesian variable selection, each feature j is assigned an **indicator variable**:

$$\delta_j = \begin{cases} 1, & \text{if feature } j \text{ has a non-zero effect} \\ 0, & \text{if feature } j \text{ has no effect.} \end{cases}$$

The model can be written as:

$$\beta_j = \alpha_j \, \delta_j, \qquad \alpha_j \sim \mathcal{N}(0, \tau^2), \quad \delta_j \sim \text{Bernoulli}(\pi)$$

where π is the prior inclusion probability.

After inference, we estimate $\text{PIP}_j = P(\delta_j = 1 \mid \text{data})$ — the **posterior inclusion probability** for feature j.

Understanding the Outputs of Bayesian Linear Regression (BLR)

The BLR model produces **posterior summaries** that describe the evidence, magnitude, and uncertainty of feature effects.

Key outputs:

1. **Posterior means of** β (effect sizes)
2. **Posterior inclusion probabilities (PIPs)**
3. **Variance component estimates** (σ^2, τ^2, ...)

Each plays a distinct and complementary role in interpretation.

Posterior Means of Effect Sizes (β)

- Represent the **estimated effect** of each feature (gene, SNP, or set)
- Computed as posterior averages:

$$\hat{\beta}_j = \mathbb{E}[\beta_j \mid \mathbf{Y}, \mathbf{X}]$$

- Interpretation:
  - Magnitude $\rightarrow$ **direction and strength** of association
  - Sign $\rightarrow$ **positive or negative effect**
  - Shrinkage $\rightarrow$ smaller estimates for weak or uncertain effects
- These are directly analogous to regression coefficients, but account for prior information and uncertainty.

Posterior Inclusion Probabilities (PIPs)

- Represent the **probability** that feature j has a nonzero effect:

$$\text{PIP}_j = \Pr(\beta_j \neq 0 \mid \mathbf{Y}, \mathbf{X})$$

- Quantifies **evidence of inclusion** in the model:
  - High PIP $\rightarrow$ strong support for inclusion
  - Low PIP $\rightarrow$ likely irrelevant or redundant feature
- Interpretation:
  - PIPs act as **Bayesian significance measures**
  - Useful for **ranking**, **fine-mapping**, and **feature prioritization**

In pathway analyses (e.g., Bayesian MAGMA), PIPs correspond to **gene- or set-level importance scores**.

Variance Component Estimates (τ^2, σ^2)

Variance components describe **uncertainty** and **heterogeneity** in effect sizes:

$$\beta_j \sim \mathcal{N}(0, \tau^2) \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

τ^2 is the variance component for the effect sizes
- Small τ^2: most effects are close to zero
- Large τ^2: more large-effect features expected

σ^2 is the variance component for the residuals
- Captures unexplained variation after accounting for $\mathbf{X}$

Together, these components describe the **genetic architecture**, enabling **heritability estimation**, **uncertainty quantification**, and **enrichment analysis** across sets or traits.

Interpreting the Outputs Together

Quantity	Interpretation	Typical use
$\hat{\beta}_j$	Direction and magnitude of effect	Effect estimation, prediction
PIP_j	Probability feature is truly associated	Fine-mapping, feature ranking
τ^2, σ^2	Variance in effect sizes and residuals	Genetic architecture, model fit

These summaries are **synergistic**:

- β tells *how much*
- PIP tells *how confident*
- Variance components tell *how complex*

Summary of BLR Model Structures

Model Type	Prior Structure	Biological Interpretation
Single-component BLR	One global variance τ^2	All features (across layers) share the same level of shrinkage — equal contribution assumption
Multiple-component BLR	Mixture of variances $\{\tau_k^2\}$	Features belong to different effect-size classes (e.g., large, small, null); grouping learned from data
Hierarchical (Biologically informed) BLR	Group-specific **mixtures of variances** $\{\tau_{gk}^2\}$	Features grouped *a priori* (e.g., by genes, pathways, or omic layers); within each group, effects can vary in size and sparsity

These models form a hierarchy of increasing flexibility and biological realism —

Motivation for Multivariate BLR

Many traits and molecular layers are **correlated** — they share genetic architecture and biological pathways.

To model these dependencies, we extend BLR to the **multivariate** setting:

- **Jointly models multiple traits or omic layers**
  $\rightarrow$ captures shared genetic or molecular effects
- **Borrows strength across correlated traits**
  $\rightarrow$ improves fine-mapping resolution and prediction accuracy
- **Estimates cross-trait effect patterns**
  $\rightarrow$ helps identify pleiotropic genes and shared biological pathways

The Multivariate Bayesian Linear Regression (MV-BLR) Model

In the **multivariate BLR** model, we model **multiple correlated outcomes** jointly:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

- $\mathbf{Y}$: $(n \times T)$ matrix of outcomes
  (e.g., association measures for T traits or omic layers)
- $\mathbf{X}$: $(n \times p)$ feature matrix
- $\mathbf{B}$: $(p \times T)$ matrix of effect sizes
- $\mathbf{E}$: $(n \times T)$ residual matrix

Each row of $\mathbf{Y}$ corresponds to an observation or gene, and each column to a trait, phenotype, or molecular layer.

Multivariate Error and Effect Priors

We extend the univariate priors to the multivariate setting:

$$\mathbf{e}_{i\cdot} \sim \mathcal{N}_T(\mathbf{0}, \Sigma_e)$$
$$\mathbf{b}_j \sim \mathcal{N}_T(\mathbf{0}, \Sigma_b)$$

- Σ_e: residual covariance among traits
- Σ_b: covariance of effect sizes across traits
- When Σ_e and Σ_b are diagonal, the model reduces to T independent univariate BLR models.

Allows **information sharing across correlated traits or omic layers** and can be used to identify **pleiotropic effects** and **cross-trait genetic architectures**.

Multivariate BLR (Structured MV-BLR)

The hierarchical structure can be extended to model **multiple traits** while preserving **biological grouping** of features:

$$\mathbf{b}_j \sim \mathcal{N}_T\big(\mathbf{0}, \Sigma_{b,g(j)}\big), \qquad \Sigma_{b,g} \sim p(\Sigma_{b,g})$$

- Each biological group g has its own **trait-level covariance matrix** $\Sigma_{b,g}$
- $\Sigma_{b,g}$ captures **correlations and scale of effects** across traits within that group

Enables information sharing both **within biological sets** and **across correlated traits**.

Indicator Variables and PIPs (Multivariate BLR)

In the **multivariate BLR**, each feature j may affect multiple outcomes (traits).
We extend the indicator variable to capture **cross-trait activity patterns**:

$$\delta_j = \begin{bmatrix} \delta_{j1} \\ \delta_{j2} \\ \vdots \\ \delta_{jT} \end{bmatrix}, \qquad \delta_{jt} = \begin{cases} 1, & \text{if feature } j \text{ affects trait } t \\ 0, & \text{otherwise.} \end{cases}$$

After inference, we estimate $\text{PIP}_{jt} = P(\delta_{jt} = 1 \mid \text{data})$ — the **posterior inclusion probability** that feature j affects trait t.

Multivariate BLR Outputs

In the multivariate setting, we generalize each posterior quantity:

Parameter	Interpretation
$\mathbf{B} = [\beta_{jt}]$	Effect matrix across traits (j: feature, t: trait)
PIP_j	Probability that feature j affects 1 trait
Σ_b	Covariance of effects across traits
Σ_e	Residual covariance among traits

These allow us to identify:

- **Shared genetic effects** (pleiotropy)
- **Trait-specific vs. shared signals**
- **Cross-trait enrichment** of biological sets

Overview of BLR Models used in Gene Set Analyses

Model Type	Feature Integration	Grouping Basis	Prior Structure	What It Captures
Single-component BLR	Combines all biological features in one model	None	One global variance (τ^2)	All features contribute equally; uniform shrinkage
Multiple-component BLR	Integrates all layers but allows heterogeneous contributions	Learned from data	Mixture of variances ($\{\tau_k^2\}$)	Large, small, and null effect classes
Hierarchical BLR	Groups features by biological structure (e.g., genes, pathways)	Defined *a priori*	Group-specific **mixture of variances** ($\{\tau_{gk}^2\}$)	Within-group heterogeneity; enrichment and

Learning at Different Levels

Model Level	Key Parameters Learned	What They Represent	How They Are Learned	What We Learn Biologically
Effect sizes	β	Strength and direction of association for each feature	Posterior mean/median given priors and data	Which features drive the outcome
Indicator variables	δ_j (single trait), δ_j (multi-trait)	Whether feature j is active (and for which traits)	Estimated as posterior inclusion probabilities (PIPs)	Which features are relevant, and whether effects are shared or trait-specific
Variance	τ^2, $\{\tau_k^2\}$, $\{\tau_{gk}^2\}$	Magnitude of expected	Inferred hierarchically	How strongly different

Estimating Variance Components During Model Fitting

In Bayesian Linear Regression, the **variance components** are *not fixed* — they are **estimated jointly** with the effect sizes β.

- The model defines hierarchical priors:

$$\beta_j \sim \mathcal{N}(0, \tau^2) \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  where τ^2 and σ^2 are **unknown parameters**.

- These variances are updated iteratively during MCMC or EM optimization:
  - τ^2 reflects the inferred spread of true effects
  - σ^2 reflects residual noise or unexplained variability

Inference alternates between sampling (or updating) β and re-estimating the variance components given the data and current effects.

Hierarchical Structure: Full Bayesian Estimation

Each variance component has its own prior, enabling uncertainty propagation:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$
$$\beta_j \sim \mathcal{N}(0, \tau^2) \qquad \tau^2 \sim \text{Inv-}\chi^2(\nu_\tau, S_\tau^2)$$
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_\sigma, S_\sigma^2)$$

- The model integrates over the **uncertainty in τ^2 and σ^2**, not just point estimates
- Posterior samples of these parameters describe:
  - The **degree of polygenicity** (via τ^2)
  - The **signal-to-noise ratio** (via τ^2/σ^2)

This full hierarchical treatment improves stability and interpretability compared to fixing variance components a priori.

In Grouped or Multi-Component BLR Models

Variance components can be **set-specific** or **mixture-specific**, and are all estimated from the data:

$$\beta_j \sim \mathcal{N}(0, \tau^2_{g(j)}), \qquad \tau^2_g \sim p(\tau^2_g)$$

or

$$\beta_j \sim \sum_{k=1}^{K} \pi_k \, \mathcal{N}(0, \tau^2_k), \qquad \tau^2_k \sim p(\tau^2_k)$$

- Each τ^2_g (or τ^2_k) is **estimated adaptively**
- Groups or components with strong evidence receive **larger** τ^2 (less shrinkage, more signal)
- Noisy or irrelevant groups shrink toward **smaller** τ^2

Variance components act as *adaptive shrinkage parameters*, controlling model complexity based on the data.

In the Multivariate BLR Model

Variance components generalize to covariance matrices:

$$\mathbf{b}_j \sim \mathcal{N}_T(\mathbf{0}, \Sigma_b), \qquad \mathbf{E}_{i\cdot} \sim \mathcal{N}_T(\mathbf{0}, \Sigma_e)$$

- Σ_b: covariance of effects across traits
  $\rightarrow$ estimated from shared signal among outcomes
- Σ_e: residual covariance among traits
  $\rightarrow$ estimated from correlated noise or shared environment

Estimating Σ_b and Σ_e enables discovery of **pleiotropy** and **cross-trait genetic structure**.

Why Variance Component Estimation Matters

Parameter	Role	Interpretation
τ^2	Effect-size variance	How much true genetic signal exists
σ^2	Residual variance	How much variation remains unexplained
τ_g^2 / τ_k^2	Group or component variance	Which sets/components are enriched
Σ_b	Cross-trait covariance	Pleiotropy or shared mechanisms

These variance components are not tuning parameters — they are **learned quantities** that describe the underlying biology. Their estimation is central to the interpretability of BLR.

Summary: Variance Components as Model-Driven Insights

- Estimated **jointly** with effect sizes and inclusion probabilities
- Control the **degree of shrinkage and complexity** in the model
- Reveal **biological structure** (e.g., gene-set enrichments, trait correlations)
- Provide **model-based evidence** for:
  - Polygenicity
  - Pleiotropy
  - Biological pathway relevance

In short, variance components are *learned descriptors of architecture*, not just technical parameters and are a cornerstone of the BLR framework.

References

Sørensen P, Rohde PD. *A Versatile Data Repository for GWAS Summary Statistics-Based Downstream Genomic Analysis of Human Complex Traits.*
medRxiv (2025). https://doi.org/10.1101/2025.10.01.25337099

Sørensen IF, Sørensen P. *Privacy-Preserving Multivariate Bayesian Regression Models for Overcoming Data Sharing Barriers in Health and Genomics.*
medRxiv (2025). https://doi.org/10.1101/2025.07.30.25332448

Hjelholt AJ, Gholipourshahraki T, Bai Z, Shrestha M, Kjølby M, Sørensen P, Rohde P. *Leveraging Genetic Correlations to Prioritize Drug Groups for Repurposing in Type 2 Diabetes.*
medRxiv (2025). https://doi.org/10.1101/2025.06.13.25329590

Gholipourshahraki T, Bai Z, Shrestha M, Hjelholt A, Rohde P, Fuglsang MK, Sørensen P. *Evaluation of Bayesian Linear Regression Models for Gene Set Prioritization in Complex Diseases.*
PLOS Genetics 20(11): e1011463 (2025).
https://doi.org/10.1371/journal.pgen.1011463