Models for Genomic Prediction

Aaron Lorenz

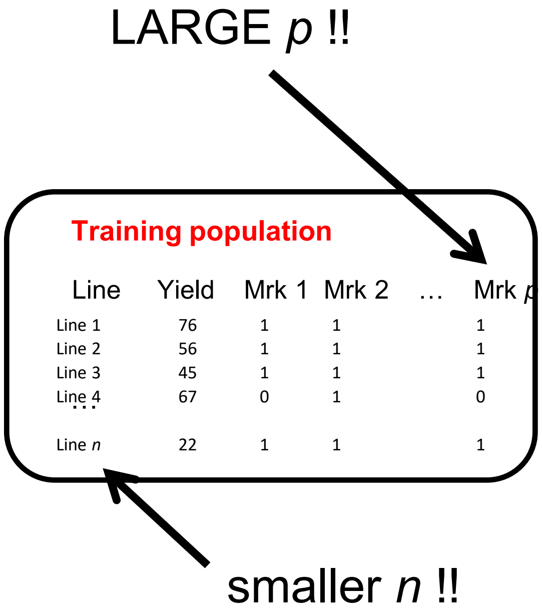
Data Bootcamp for Genomic Prediction in Plant Breeding

Ghent, Belgium

July 5-7, 2023

Genomic prediction models

Bayesian LASSO
G-BLUP
Principal component regression
Elastic net
RR-BLUP
BayesCpi
RKHS
Random forests
BayesD
Partial least squares
BayesC
LASSO
BayesA
BayesB
Dirichlet process regression
Support vector machine regression
Neural networks



LARGE *p* !!

Training population

Line	Yield	Mrk 1	Mrk 2	…	Mrk *p*
Line 1	76	1	1		1
Line 2	56	1	1		1
Line 3	45	1	1		1
Line 4	67	0	1		0
…					
Line *n*	22	1	1		1

smaller *n* !!

Getting around the ***large p, small n problem***

- Dimension reduction techniques
  - Singular value decomposition (essentially principal component analysis)
  - Partial least squares
  - Stepwise model selection strategies

- Ridge regression

- Random effects modeling

- Hierarchical modeling
  - Bayesian models

Baseline model

$$y_i = \mu + \sum_k \beta_k x_{ik} + e_i$$

$$\beta_k \sim ?$$

--More predictors than variables.

--Solution: fit predictors as random effects.

-- Constrain possible effects.

-- What distribution is β being sampled from?

Ridge regression BLUP (RR-BLUP): Convenient way of estimating genome-wide marker effects

RR-BLUP estimators

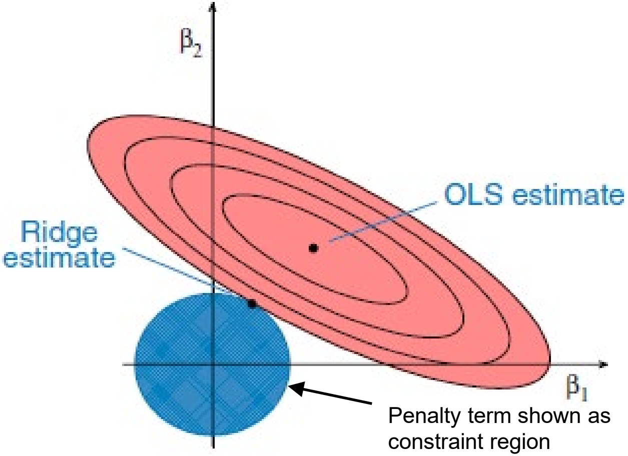
$$\hat{\beta} = (X^T X + \boxed{\lambda I})^{-1} X^T y$$

Where $\lambda = \sigma_e^2/\sigma_\beta^2$. Addition of λI term reduces collinearity and prevents $X^T X$ from becoming singular.

$$\beta_j \sim N(0, \sigma_\beta^2)$$

- Originally, ridge regression used grid search to find optimal λ.

- When $\lambda = \sigma_e^2/\sigma_\beta^2$ is used, $\hat{\beta}$ can be shown to be the BLUP of β, and it has become known as ridge regression BLUP (RR-BLUP).



Penalty term shown as constraint region

Bayesian ridge regression

$$y_i = \mu + \sum_k \beta_k x_{ik} + e_i$$

Prior variance, hyperparameter

$$\overbrace{p\left(\mu, \boldsymbol{\beta}, \sigma^2 \middle| \boldsymbol{y}, \sigma_\beta^2\right)}^{\text{Posterior density of the model}}$$

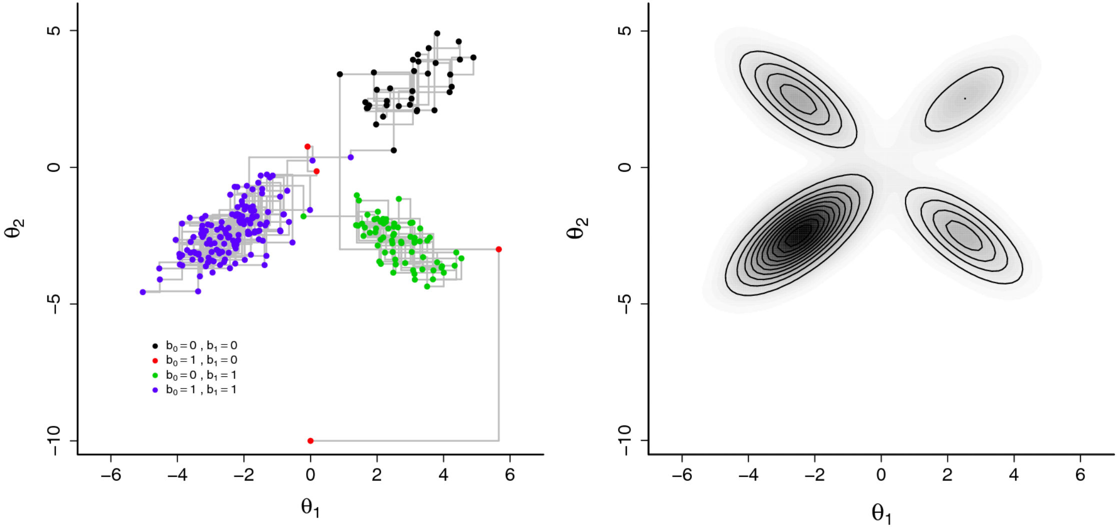
$$\propto p(\boldsymbol{y}|\mu, \boldsymbol{\beta}, \sigma^2)p(\mu, \boldsymbol{\beta}, \sigma^2|\sigma_\beta^2)$$

$$\propto \prod_{j=1}^{p} p\left(\beta_j \middle| \sigma_\beta^2\right)p(\sigma^2)$$

$$\beta_j \sim N(0, \sigma_\beta^2) \qquad \sigma_\beta^2 \sim \chi^{-2}$$

Use of Gibbs sampler to approximate parameters of the posterior distribution

A MCMC sampler that can explore the density of a complex multivariate distribution using only the conditional probabilities.

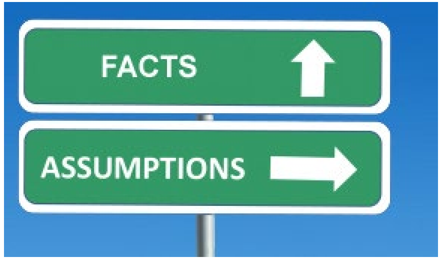


(a) First 250 Gibbs iterations

(b) Posterior distribution

Rodrigues et al. (2020)

Genomic prediction models galore!



- RR-BLUP ignores biological reality in two important ways (Bernardo, 2020):
    1. It assumes each marker effect is from the same normal distribution
    2. Epistasis is absent

- In order to circumvent these assumptions, many other genomic prediction models have been developed as well as adapted from other disciplines.
    - Bayesian models: BayesA, BayesB, BayesC, BayesCπ, BayesD, Bayesian LASSO
    - Elastic net
    - Reproducing kernel Hilbert spaces (RKHS)
    - Machine learning models such as support vector machine (SVM), random forest, and neural networks

Other prior distributions of marker effects

$$y_i = \mu + \sum_k \beta_j x_{ij} + e_i$$

$Bayesian\ ridge\ regression$

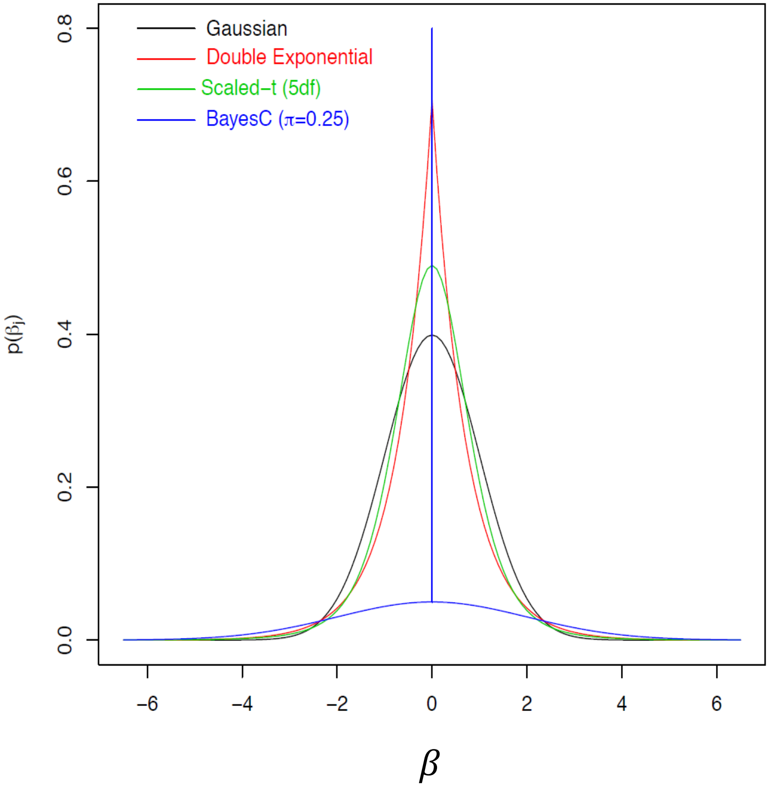
$\beta_j \sim N(0, \sigma_\beta^2)$

$LASSO\ (Double\ exponential)$

$\beta_j \sim DE(\lambda)$

$BayesC$

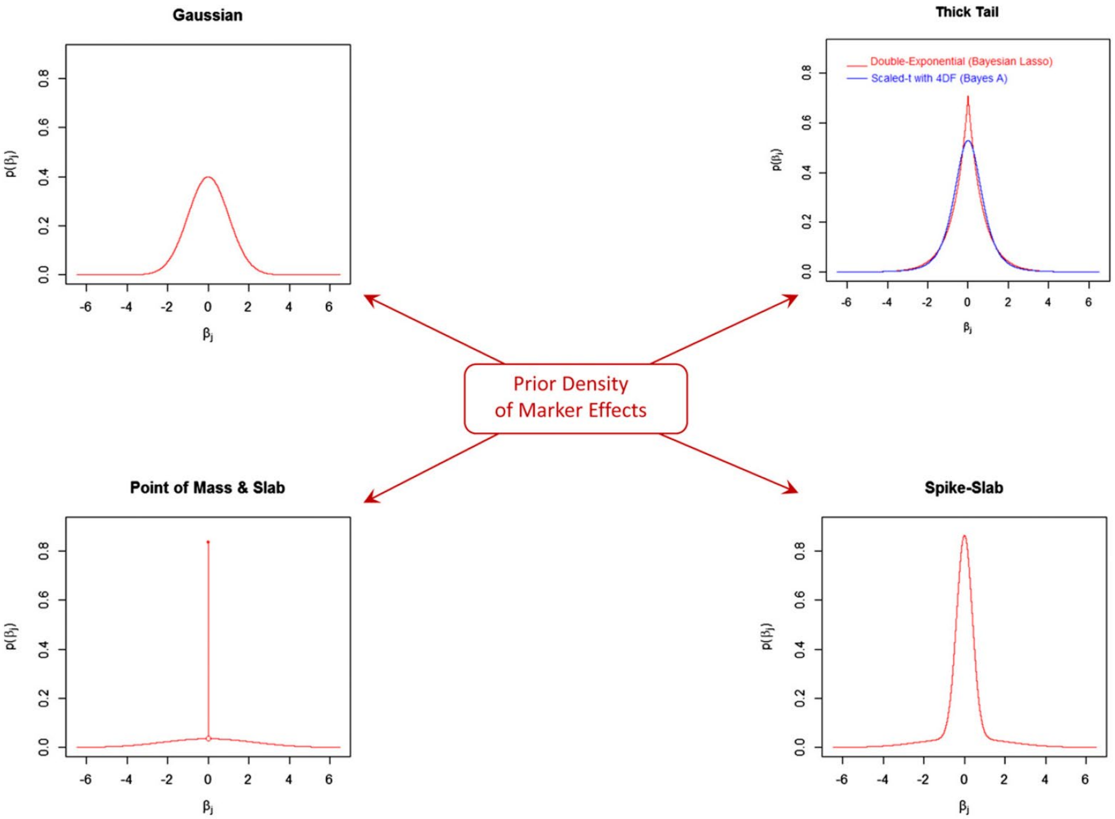
$$\beta_k = \begin{cases} 0 & \text{with prob } \pi \\ \sim \mathrm{N}(0,\sigma_\beta^2) & \text{with prob } (1-\pi) \end{cases}$$



Perez and de los Campos (2015)

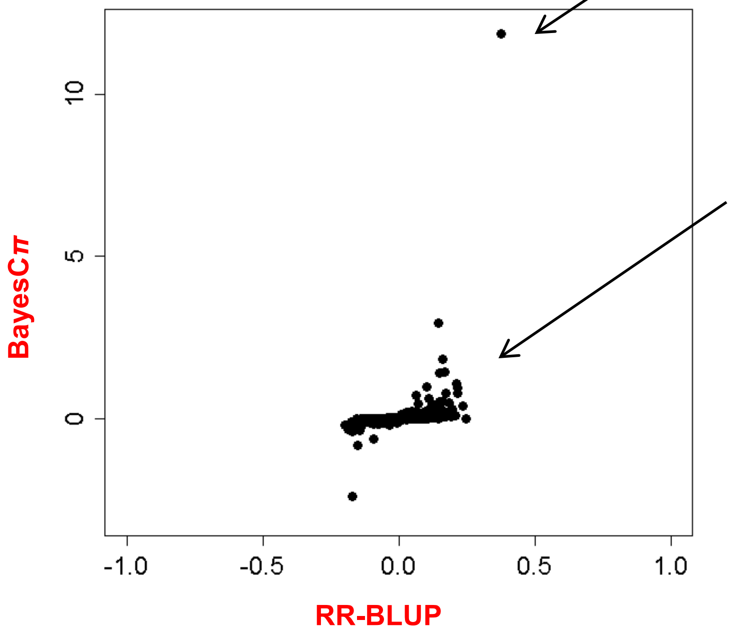
de Los Campos et al. (2013)

Priors



Prior Density of Marker Effects

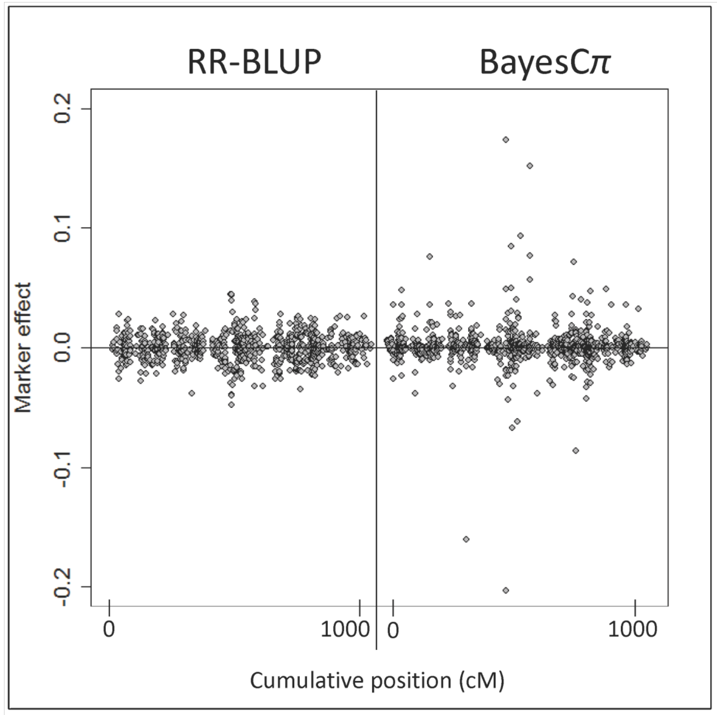
Marker effect estimates



Large-effect QTL simulated

Many small-effect QTL simulated

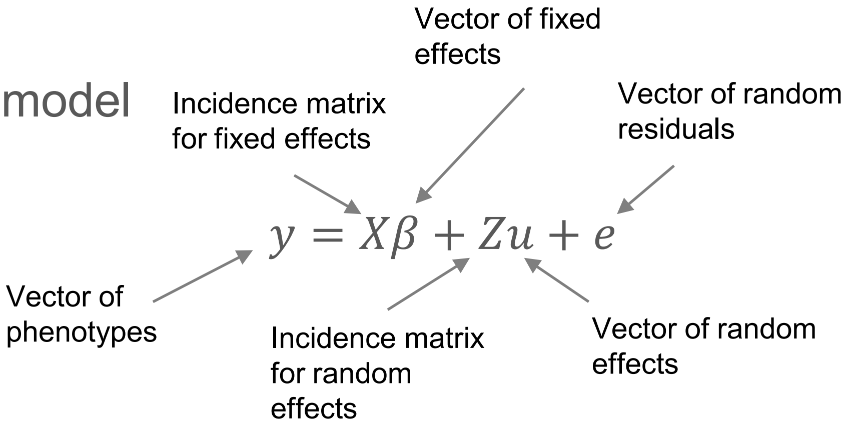
Comparing marker effects between models



Genomic best linear unbiased prediction (G-BLUP)

- Similar to traditional BLUP with pedigrees in a mixed model
- Pedigree relationship matrix is substituted with genomic relationship matrix
- Use genomic relationships in mixed-linear model to predict breeding value of relatives

- General mixed model



Vector of fixed effects

Incidence matrix for fixed effects

Vector of random residuals

$$y = X\beta + Zu + e$$

Vector of phenotypes

Incidence matrix for random effects

Vector of random effects

Incidence matrix for fixed effects

Vector of fixed effects

Vector of random residuals

$$y = X\beta + Zu + e$$

Vector of phenotypes

Incidence matrix for random effects

Vector of random effects

Random effects are assumed to be drawn from some underlying probability distribution and thus can be assigned a covariance structure.

Here, it is normally assumed that $u \sim MVN(0, G)$ where G describes the covariances among random effects.

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\hat{\mathbf{u}} = \boxed{\mathbf{G}}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Information between relatives being shared through G matrix

Genomic best linear unbiased prediction (G-BLUP)

$$y_i = u_i + \varepsilon_i$$

$$\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$$

$$\mathbf{G} = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nn} \end{bmatrix}$$

Ideal

G matrix calculated using causal polymorphisms

$$\mathbf{G}_C = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nn} \end{bmatrix}$$

Estimate

G matrix estimated using markers

$$\mathbf{G}_M = \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} & \cdots & \hat{G}_{1n} \\ \hat{G}_{21} & \hat{G}_{22} & \cdots & \hat{G}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{G}_{n1} & \hat{G}_{n2} & \cdots & \hat{G}_{nn} \end{bmatrix}$$

Hill and Weir (2011)

$$E(G_{ij}) = A_{ij}$$

$$E(G_{ij})\downarrow \;=\; Var(G_{ij})\;\uparrow$$

Realized relationships expected to vary *less* across genome

$$A = \begin{array}{c|ccccc}
 & \text{Ind1} & \text{Ind2} & \text{Ind3} & \text{Ind4} & \text{Ind5} \\
\hline
\text{Ind1} & & & & & \\
\text{Ind2} & 0.67 & & & & \\
\text{Ind3} & 0.50 & 0.33 & & & \\
\text{Ind4} & 0.33 & 0.23 & 0.35 & & \\
\text{Ind5} & 0.08 & 0.17 & 0.15 & 0.20 &
\end{array}$$

Realized relationships expected to vary *more* across genome

Calculation of **G**

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)}$$

Where Z is a centered marker matrix, Z = M – P

M = {-1, 0, 1} numerically codes the homozygote, heterozygote, and other homozygote

Each column of P contains the mean value of the corresponding column of M

p_i is the allele frequency of the second allele at marker locus I

This part, the denominator, scales the genomic relationship matrix so it is analogous to the numerator relationship matrix.

Equivalency between RR-BLUP and G-BLUP

$$\mathbf{y} = \mu + \sum_k \mathbf{x}_k \beta_k + \mathbf{e} \qquad \beta_k \sim N(0, \sigma_\beta^2)$$

$$\mathbf{u} = \sum_k \mathbf{x}_k \beta_k = \mathbf{X\boldsymbol{\beta}}$$

From MVN distribution properties:

$$\mathrm{var}(\mathbf{u}) = \mathbf{XX}^T \sigma_\beta^2 = \mathbf{G}\sigma_u^2$$

$$\mathbf{G} \propto \mathbf{XX}^T$$

Only valid with the normal prior!

Reproducing kernel Hilbert spaces

- Constitute regression functions that are linear combinations of a basis function provided by a reproducing kernel (RK).
- The RK function maps pairs of points from an input space to a feature space.
- The structure of the RK can be flexible, being linear or non-linear.
- Structure of the model is that of the standard Animal Model

$$y_i = \mu + u_i + \varepsilon_i$$

$$\mathbf{u} \sim N(0, \boldsymbol{K}\sigma_u^2)$$

Additive relationship matrix replaced with reproducing kernel

$$K(x_i, x_j) = exp\left\{-h \times \frac{\sum(x_{ik} - x_{ij})\text{^}2}{p}\right.$$

Summary

- A wide variety of models exist for genomic prediction designed to get around the “large p, small n” problem of estimating marker effects for complex traits

- Many models are very similar, but categories of models exist that attempt to model different assumptions about the genetic architecture of traits

- Under the assumptions of normality, it can be shown the RR-BLUP and G-BLUP are equivalent

- RKHS models using the same framework as G-BLUP, but the non-linear kernel allows for the potential to model non-additive effects

- In general, for complex traits in plant breeding scenarios, little differences between models have been found.