Linear Mixed Models

Evangelina López de Maturana & Oscar González-Recio

What you need to know

Reasoning	Mixed Model Equations	Solving algorithms	Variance component estimation	Bayesian framework
Why we need mixed models in GWP	How to create the MME	How to solve the equations system	How to estimate variances of random effects	Going bayesian

Mathematical representation of biological processes

$$P=G+E$$

$$y_i=g_i+E_i$$

macroenvironment

Cohort, diet, farm, year, age, location, parity, sex,

microenvironment

Unknown or difficult to measure effects (residual)

$$y_i=EnvironmentalEffects + g_i + e_i$$

$$y_i=X_ib_i +Z_ig_i+e_i$$

Mathematical representation of biological processes

$$P = G + E$$

$$y_i = g_i + E_i$$

Additive   Dominance   Epistasis

Epistasis: Generates additive variance

Hill et al (2008) https://doi.org/10.1371/journal.pgen.1000008

$$y_i = X_i b_i + Z_{ui} u_i + Z_{di} d_i + Z_{ei} u\#u_i + Z_{ei} u\#d_i + \ldots + e_i$$

$$y_i = X_i b_i + Z_{ui} u_i + e'_i$$

Mathematical representation of biological processes

$$P = G + E$$

$$y_i = g_i + E_i$$

Additive

Dominance

Epistasis

Generates additive variance

Hill et al (2008) https://doi.org/10.1371/journal.pgen.1000008

$$y_i = X_i b_i + Z_{ui} u_i + Z_{di} d_i + Z_{ei} u\#u_i + Z_{ei} u\#d_i + \ldots + e_i$$

$$y_i = X_i b_i + Z_{ui} u_i + e'_i$$

Linear mixed models

$$y_i = X_i b_i + Z_{ui} u_i + e'_i$$

$$\mathbf{e} \sim N(0, \mathbf{I} \otimes \sigma_e^2)$$

$$\mathbf{u} \sim N(0, \mathbf{A} \otimes \sigma_a^2)$$

"Fixed" terms

"Random" terms (assume some known distribution to the effect)

Linear mixed models: Mixed Model Equation (EMM)

$y_i = X_i b_i + Z_{ui} u_i + e'_i$

$$
\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{114} \\ Y_{115} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{124} \\ Y_{125} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{214} \\ Y_{215} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{224} \\ Y_{225} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{314} \\ Y_{315} \\ Y_{321} \\ Y_{322} \\ Y_{323} \\ Y_{324} \\ Y_{325} \end{pmatrix}
=
\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}
\begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}
+
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}
\begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix}
+
\begin{pmatrix} e_{111} \\ e_{112} \\ e_{113} \\ e_{114} \\ e_{115} \\ e_{121} \\ e_{122} \\ e_{123} \\ e_{124} \\ e_{125} \\ e_{211} \\ e_{212} \\ e_{213} \\ e_{214} \\ e_{215} \\ e_{221} \\ e_{222} \\ e_{223} \\ e_{224} \\ e_{225} \\ e_{311} \\ e_{312} \\ e_{313} \\ e_{314} \\ e_{315} \\ e_{321} \\ e_{322} \\ e_{323} \\ e_{324} \\ e_{325} \end{pmatrix}
$$

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + \mathbf{e}$$

Example of fixed effects (mean, and an effect with 2 levels)

Example of random effect with 3 levels (ignoring the covariance structure)

Linear mixed models: Mixed Model Equation (MME)

$$y_i = X_i b_i + Z_{ui} u_i + e'_i$$

When X and Z are too large, inversion of matrices is very computationally demanding, there are linear combinations, and solving the system becomes cumbersome.
Henderson proposed MME

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \alpha \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'Y} \\ \mathbf{Z'Y} \end{bmatrix}$$

Linear mixed models: Mixed Model Equation (EMM)

Increased complexity

$$\mathbf{y}_i = X_i\mathbf{b}_i + U_i\mathbf{u}_i + W_i\mathbf{p}_i + Z_{ui}\mathbf{u}_i + \mathbf{e}'_i$$

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}U & X'R^{-1}W & X'R^{-1}Z \\ U'R^{-1}X & U'R^{-1}U+\Theta_1 & U'R^{-1}W & U'R^{-1}Z \\ W'R^{-1}X & W'R^{-1}U & W'R^{-1}W+\Theta_2 & W'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}U & Z'R^{-1}W & Z'R^{-1}Z+\Theta_S \end{pmatrix}^{-1} \times \begin{pmatrix} X'R^{-1}Y \\ U'R^{-1}Y \\ W'R^{-1}Y \\ Z'R^{-1}Y \end{pmatrix}$$

Linear mixed models: Mixed Model Equation (EMM)

Marker regression

Decomposing the polygenic effect into the sum of SNP effects

$$y_i = X_i b_i + Z_{ui} u_i + e'_i$$

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{X}\beta_f + snp_1\beta_1 + snp_2\beta_2 + \ldots + snp_p\beta_p + \mathbf{e}$$

$$\mathbf{e} \sim N(0, \sigma_e^2)$$

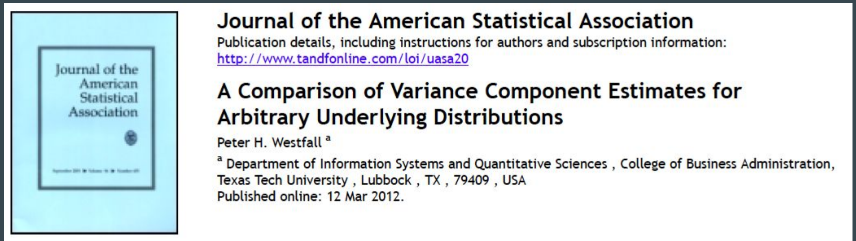
$$\mathbf{b}_i \sim N(0, \sigma_i^2)$$

$$\sigma_i^2 \sim \chi^{-2}_{(\upsilon,S)}$$

Variance component estimation
Methods

Estimate residual, genetic, permanent, marker, … variances

- ANOVA
- Maximum Likelihood (ML)
- Restricted Maximum Likelihood (REML)
- Minimum Norm Quadratic Unbiased Estimation (MINQUE I, II, III)
- Minimum Variance Quadratic Unbiased Estimation (MIVQUE)



Solve MME

Algorithms

- Gauss-seidel
- Choleski decomposition
- Preconditioned conjugate gradients (PCG)
- Gauss-seidel with residual updates

J. Dairy Sci. 91:360–366
doi:10.3168/jds.2007-0403

Technical Note: Computing Strategies in Genome-Wide Selection

A. Legarra*[1] and I. Misztal†
*Institut National de la Recherche Agronomique, UR631 Station d'Amélioration Génétique des Animaux, BP 52627, 32326 Castanet-Tolosan, France
†Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771

Solve MME

Algorithms

2,000 records
11,000 SNPs

Table 1. Computing times for different methods of solving the mixed-model equations in a case of genome-wide genetic evaluation[1]

Item	Cholesky decomposition[2]	LHS-GS	Matrix-free GS	Matrix-free GSRU	Matrix-free PCG
Convergence = 10^{-10}					
Time to set up	17 min	17 min	16 s	16 s	20 s
Time for solving	119 min	71 min	97 h	46 s	10 s
Number of iterations	1	164	161	164	20
Convergence = 10^{-14}					
Time to set up	17 min	17 min	[3]	16 s	20 s
Time for solving	119 min	170 min	[3]	74 s	12 s
Number of iterations	1	272	[3]	272	23

[1]LHS = left-hand side; GS = Gauss-Seidel; GSRU = Gauss-Seidel with residual update; PCG = preconditioned conjugated gradients.

[2]This is an exact method and the convergence measure is meaningless.

[3]Not tried.

Solve MME

Gauss Seidel with Residual Update

y corrected for all effects except the j^{th} effect, is equal to the current vector of residuals plus the current estimate of the jth effect.
Then, we can compute this j^{th} effect.

$$\mathbf{y} - \mathbf{X}_{1:j-1,:}\hat{\mathbf{a}}^{l+1}_{1:j-1} - \mathbf{X}_{j+1:n,:}\hat{\mathbf{a}}^{l}_{j+1:n} = \mathbf{e}^{l+1,j} + \mathbf{x}_j\hat{a}^l_j.$$

Then

$$\hat{a}_j^{l+1} = \frac{\mathbf{x}_j'\mathbf{e}^{l+1,j} + \mathbf{x}_j'\mathbf{x}_j\hat{a}_j^l}{\mathbf{x}_j'\mathbf{x}_j + \lambda}.$$

We update the residuals with the new j^{th} effect solution.

$$\mathbf{e}^{l+1,j+1} = \mathbf{e}^{l+1,j} - \mathbf{x}_j(\hat{a}_j^{l+1} - \hat{a}_j^l).$$

Solution from previous iteration

Updated solution in the new iteration

$\mathbf{x}_j'\mathbf{x}_j$ are constant, so they can be precomputed, updating vector products and residuals at each iteration, speeding up the algorithm

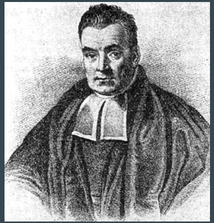
Bayesian framework

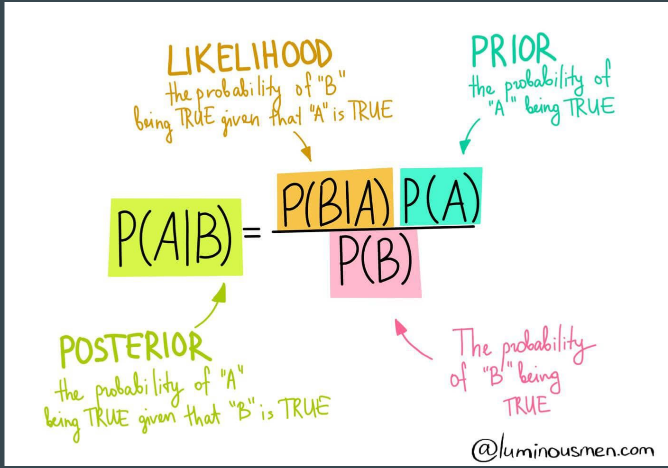
$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\beta_f + snp_1\beta_1 + snp_2\beta_2 + \ldots + snp_p\beta_p + \mathbf{e}$$

Assume a model for the data

$$p(\mathbf{y}|\boldsymbol{\theta}) = N(\mathbf{X\beta} + \cdots + \mathbf{Zu}, \sigma_e^2)$$

Bayes theorem



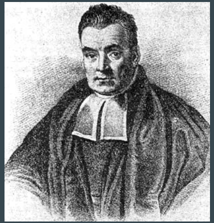


Bayesian framework

Assume a model for the data

$$p(\mathbf{y}|\boldsymbol{\theta}) = N(\mathbf{X}\boldsymbol{\beta} + \cdots + \mathbf{Z}\mathbf{u}, \sigma_e^2)$$

Bayes theorem



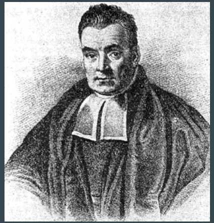
$$p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})$$

Bayesian framework

Assume a model for the data

$$p(\mathbf{y}|\boldsymbol{\theta}) = N(\mathbf{X}\boldsymbol{\beta} + \cdots + \mathbf{Z}\mathbf{u}, \sigma_e^2)$$

Bayes theorem



$$p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})$$

...proportionality

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

y=data

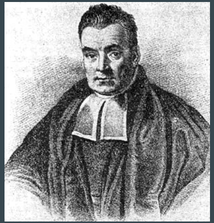
$\boldsymbol{\theta}$= unknown parameters, coefficients, variances, ...

Bayesian framework

Assume a model for the data

$$p(\mathbf{y}|\boldsymbol{\theta}) = N(\mathbf{X}\boldsymbol{\beta} + \cdots + \mathbf{Z}\mathbf{u}, \sigma_e^2)$$

Bayes theorem



$$p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})$$

...proportionality

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Choose priors

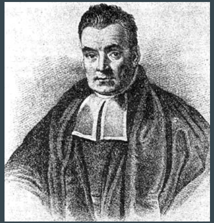
$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2)p(\boldsymbol{\beta}|\sigma_b^2)p(\mathbf{u}|\sigma_u^2)p(\sigma_b^2)p(\sigma_u^2)p(\sigma_e^2)$$

Bayesian framework

Assume a model for the data

$$p(\mathbf{y}|\boldsymbol{\theta}) = N(\mathbf{X}\boldsymbol{\beta} + \cdots + \mathbf{Z}\mathbf{u}, \sigma_e^2)$$

Bayes theorem



$$p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})$$

...proportionality

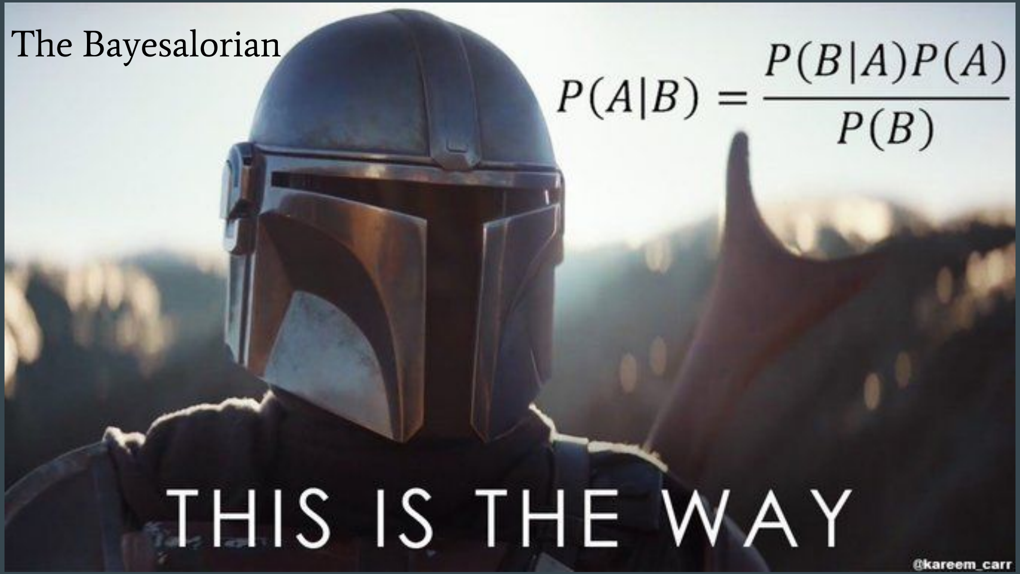
$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Choose priors

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2)p(\boldsymbol{\beta}|\sigma_b^2)p(\mathbf{u}|\sigma_u^2)p(\sigma_b^2)p(\sigma_u^2)p(\sigma_e^2)$$

Make inferences using McMC algorithms (Gibbs sampling, acceptance rejection, Metropolis-Hasting)

Bayesian framework



EXAMPLE

Create MME and solve the system using residual updates

ID	trait	Diet	SNP1	SNP2
1	20	1	1	2
2	25	1	1	1
3	30	2	0	2
4	35	2	0	1
5	20	3	2	1
6	30	3	2	0

HOMEWORK

Solve the mixed linear model

trait=mu+age+Diet+SNP1+SNP2+e

Using Gauss-Seidel with residual updates, with residual variance = 40 and SNP variance = 3.

ID	trait	Age	Diet	SNP1	SNP2
1	93	25	1	0	0
2	90	30	1	1	0
3	115	35	2	1	2
4	110	20	2	2	2
5	87	22	3	2	1
6	70	29	1	0	1
7	100	31	3	1	2

HOMEWORK

Solve the mixed linear model

trait=mu+age+Diet+Cohort+e

Using Gauss-Seidel with residual updates, with residual variance = 40 and cohort variance = 30.

ID	trait	Age	Diet	Cohort
1	93	25	1	1
2	90	30	1	2
3	115	35	2	1
4	110	20	2	2
5	87	22	3	3
6	70	29	1	3
7	100	31	3	3