Multitrait models in BGLR

Paulino Pérez-Rodríguez [1]
Gustavo de los Campos [2]

[1] Colegio de Posgraduados [2] Michigan State University

June, 2020.

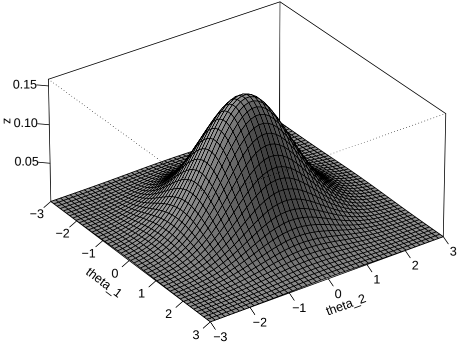
Contents

1. Important distributions
2. Multivariate Multiple Linear Regression
3. Multivariate Linear Mixed Model (Bayesian context)
4. Multitrait BGLR
5. Examples

Multivariate Normal distribution

The MN distribution is a generalization of the univariate normal distribution to higher dimensions. If $\boldsymbol{\theta} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the variance covariance matrix. The density function of $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\mu})\right\}$$



Matrix normal distribution or matrix Gaussian distribution

The MG distribution is a generalization of the MN distribution. The random matrix $\mathbf{\Theta}_{n\times q}$ follows a MG distribution $MG(\mathbf{M}, \mathbf{V}, \mathbf{W})$ and the density function is given by:

$$p(\mathbf{\Theta}|\mathbf{M}, \mathbf{V}, \mathbf{W}) = \frac{\exp\{-0.5tr[\mathbf{W}^{-1}(\mathbf{\Theta} - \mathbf{M})'\mathbf{V}^{-1}(\mathbf{\Theta} - \mathbf{M})]\}}{(2\pi)^{nq/2}|\mathbf{W}|^{n/2}|\mathbf{V}|^{q/2}},$$

where $\mathbf{M}$ is a location parameter, a $n \times q$ matrix, $\mathbf{V}$ is a $n \times n$ matrix, a scale parameter for rows and $\mathbf{W}$ is a $q \times q$, a scale parameter for columns.

It can be shown that $\mathbf{\Theta}_{n\times t} \sim MG(\mathbf{M}, \mathbf{V}, \mathbf{W})$, if and only if $vec(\mathbf{\Theta}_{n\times q}) \sim MN(vec(\mathbf{M}), \mathbf{W} \otimes \mathbf{V})$.

Inverse Wishart distribution

$\boldsymbol{\Theta} \sim IW_q(\boldsymbol{S}, df)$, where $\boldsymbol{S}_{q \times q}$ is a scale matrix, which should be positive definite, df corresponds to the degrees of freedom, $df > q - 1$. The density function is proportional to:

$$p(\boldsymbol{\Theta}|\boldsymbol{S}, df) \propto |\boldsymbol{S}|^{df/2}|\boldsymbol{\Theta}|^{-(df+q+1)/2} \exp\{-0.5 tr(\boldsymbol{S}\boldsymbol{\Theta}^{-1})\}$$

The expectation is defined if $df > q + 1$.

Multivariate Multiple Linear Regression

Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{iq})'$, $i = 1, ..., n$ (individuals) and $k = 1, ..., q$ (traits), and let $\boldsymbol{X}$ represent a matrix of covariates of dimension $n \times p$ and $\boldsymbol{B} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q)$ a matrix of regression coefficients. Without loss of generality, assuming that the the phenotypes have been centered by trait, the MMLR model can be written as:

$$\begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_{11} & \cdots & \beta_{1q} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pq} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1q} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nq} \end{pmatrix}$$

or in compact matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$

here for example $\boldsymbol{E} \sim MG(\boldsymbol{0}, \boldsymbol{I}, \boldsymbol{R}_0)$, with $\boldsymbol{R}_0$ matrix of dimensions $q \times q$.

Multivariate Linear Mixed Model (Bayesian context)

Without loss of generality consider:

$$\boldsymbol{Y} = \mathbf{1}\mu' + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U} + \boldsymbol{E},$$

with $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{E}$ as defined before, $\boldsymbol{\mu} = (\mu_1, ..., \mu_q)'$ a vector of intercepts for traits, and

$$\boldsymbol{U} = \begin{pmatrix} u_{11} & \cdots & u_{1t} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nt} \end{pmatrix}$$

is a matrix of random effects.

Continue...

With:

1. $\boldsymbol{E} \sim MG(\mathbf{0}, \boldsymbol{I}, \boldsymbol{R}_0)$, $\boldsymbol{R}_0$ (unknown) covariance matrix for residuals of dim $q \times q$.
2. $p(\boldsymbol{\mu}) \propto 1$.
3. $\boldsymbol{B} \sim MG(\mathbf{0}, \boldsymbol{I}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ (unknown) covariance matrix for regression coefficients.
4. $\boldsymbol{U} \sim MG(\mathbf{0}, \boldsymbol{K}, \boldsymbol{G})$, $\boldsymbol{K}$ (known) covariance matrix to model relationship between individuals and $\boldsymbol{G}$ (unknown) covariance matrix to model within subject relations.

We assign Inverse Wishart distributions to $\boldsymbol{R}_0$, $\boldsymbol{\Omega}$ and $\boldsymbol{G}$.

Continue...

Posterior distribution has no closed form, but all full conditionals necessary to implement a Gibbs sampler (Geman and Geman, 1984) have closed form. We need to sample from:

1. $\boldsymbol{\mu}|else$.
2. $\boldsymbol{B}|else$.
3. $\boldsymbol{U}|else$.
4. $\boldsymbol{R}_0|else$.
5. $\boldsymbol{\Omega}|else$.
6. $\boldsymbol{G}|else$.

Continue...

We consider several modeling strategies for $\boldsymbol{R}_0$, $\boldsymbol{\Omega}$ and $\boldsymbol{G}$. Without loss of generality consider the case of $\boldsymbol{\Omega}$.

Structure	**Prior**
	Hyper-parameters
Unstructured	$\boldsymbol{\Omega} \sim IW(\boldsymbol{\Omega} \mid \boldsymbol{S}_0, df_0)$;
	$\boldsymbol{S}_0$: prior scale matrix ($q \times q$) df_0: prior degree of freedom
Diagonal	$\boldsymbol{\Omega} = \mathrm{diag}(\omega_{11}, \ldots, \omega_{\mathrm{tt}})$, $p(\boldsymbol{\omega}) = \prod_{j=1}^{q} \chi^{-2}\left(\omega_{jj} \mid S_{0jj}, df_{0j}\right)$
	$\boldsymbol{S}_0$: prior scale vector ($q \times 1$) df_0: prior *df* vector ($q \times 1$)
Factor analytic	$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$; $\boldsymbol{\Lambda}$:factor loadings, $\boldsymbol{\Psi} = \mathrm{diag}(\psi_{\mathrm{jj}})$ $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = MVN(vec(\boldsymbol{\Lambda}) \mid \boldsymbol{0}, \sigma^2 \mathbf{I}) \times \prod_{j=1}^{q} \chi^{-2}\left(\psi_{jj} \mid S_{0jj}, df_{0j}\right)$
	$\boldsymbol{S}_0$: prior scale vector ($q \times 1$) df_0: prior *df* vector ($q \times 1$); σ^2: prior variance.
Recursive	$\boldsymbol{\Omega} = (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Psi} ((\mathbf{I} - \mathbf{B})^{-1})'$, $\boldsymbol{B}_{q \times q}$ regression coefficients; $\boldsymbol{\Psi} = \mathrm{diag}(\psi_{\mathrm{jj}})$. $p(\mathbf{B}, \boldsymbol{\Psi}) = MVN(vec(\mathbf{B}) \mid \boldsymbol{0}, \sigma^2 \mathbf{I}) \times \prod_{j=1}^{q} \chi^{-2}\left(\psi_{jj} \mid S_{0jj}, df_{0j}\right)$
	$\boldsymbol{S}_0$: prior scale vector ($q \times 1$) df_0: prior *df* vector ($q \times 1$); σ^2: prior variance.

Continue...

We can also assign different prior distribution to components of $\boldsymbol{B}$, for example mixture of normals (e.g. Cheng et al., 2018) or flat priors.

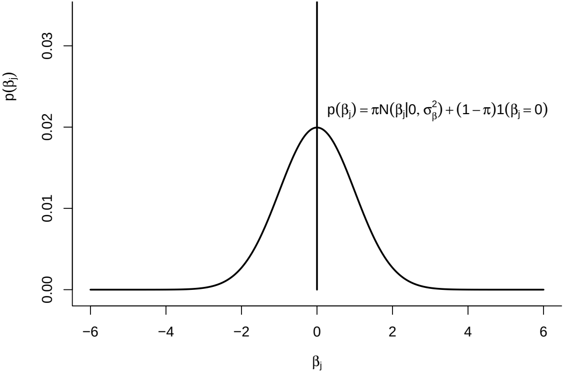


Figure 1: Spike-Slab.

Multitrait BGLR

BGLR now includes the function Multitrait that fits a MLMM. The function is a re-implementation/extension of the function MTM included in the MTM package (http://quantgen.github.io/MTM/vignette.html).

Key features:

1. Heavy computational routines implemented in the C programming language.
2. Gaussian and Gaussian mixtures priors for regression coefficients.
3. Different covariance structures for $\boldsymbol{R}_0$, $\boldsymbol{\Omega}$ and $\boldsymbol{G}$ (UNstructured, DIAGonal, Factor Analytic, RECursive).
4. Arbitrary number of random effects.

Visit https://github.com/gdlc/BGLR-R.

Continue...

The Multitrait function fits the model:

$$\boldsymbol{Y} = \mathbf{1}\mu' + \boldsymbol{X}_1\boldsymbol{B}_1 + \boldsymbol{X}_2\boldsymbol{B}_2 + \cdots + \boldsymbol{U}_1 + \boldsymbol{U}_2 + \cdots + \boldsymbol{E}$$

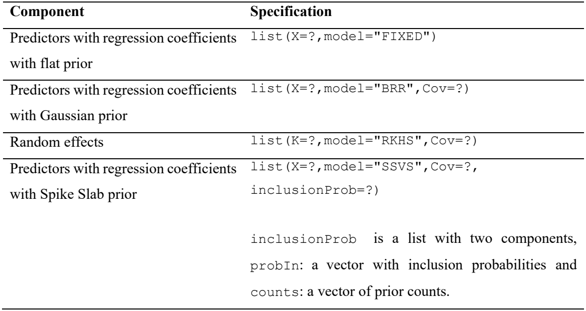
where all the terms have been defined before. The function implements a Gibbs sampler with scalar update. Missing values are imputed at each iteration of Gibbs algorithm.

We will discuss only the more basic models, but we present the information required to fit more complex models, the sample code is accesible in the github website.

Model specification

The user interface is similar to the BGLR function and is list based.

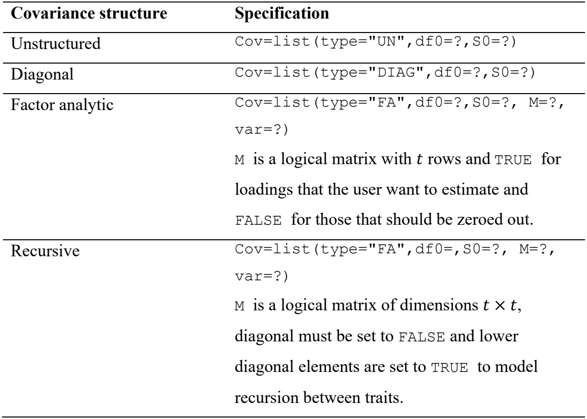
Table 2. Specifying components of the linear predictor.



Component	**Specification**
Predictors with regression coefficients with flat prior	`list(X=?,model="FIXED")`
Predictors with regression coefficients with Gaussian prior	`list(X=?,model="BRR",Cov=?)`
Random effects	`list(K=?,model="RKHS",Cov=?)`
Predictors with regression coefficients with Spike Slab prior	`list(X=?,model="SSVS",Cov=?, inclusionProb=?)` `inclusionProb` is a list with two components, `probIn`: a vector with inclusion probabilities and `counts`: a vector of prior counts.

Continue...

Table 3. Specifying covariance structures in R code.



Covariance structure	**Specification**
Unstructured	`Cov=list(type="UN",df0=?,S0=?)`
Diagonal	`Cov=list(type="DIAG",df0=?,S0=?)`
Factor analytic	`Cov=list(type="FA",df0=?,S0=?, M=?, var=?)` `M` is a logical matrix with t rows and `TRUE` for loadings that the user want to estimate and `FALSE` for those that should be zeroed out.
Recursive	`Cov=list(type="FA",df0=,S0=?, M=?, var=?)` `M` is a logical matrix of dimensions $t \times t$, diagonal must be set to `FALSE` and lower diagonal elements are set to `TRUE` to model recursion between traits.

Example 1: Fitting RR

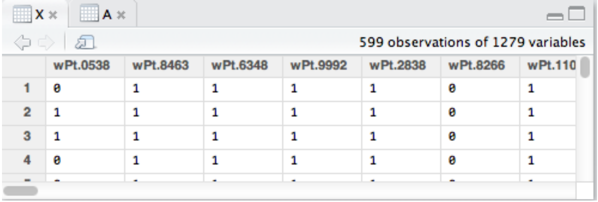
Data for $n = 599$ wheat lines evaluated in 4 environments, wheat improvement program, CIMMyT. The dataset includes $p = 1279$ molecular markers ($x_{ij}, i = 1, ..., n,\ j = 1, ..., p$) (coded as 0,1). The pedigree information is also available.

Lets load the dataset in R,

1. Load R.
2. Install BGLR package (if not yet installed).
3. Load the package.
4. Load the data.

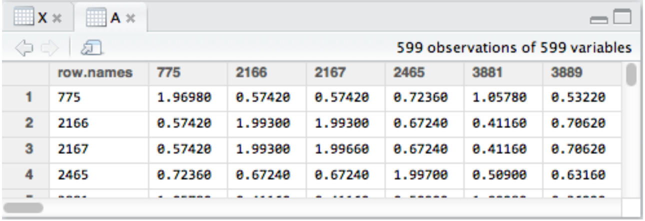
Continue...

You can explore the MM matrix, pedigree matrix within R,



599 observations of 1279 variables

	wPt.0538	wPt.8463	wPt.6348	wPt.9992	wPt.2838	wPt.8266	wPt.110
1	0	1	1	1	1	0	1
2	1	1	1	1	1	0	1
3	1	1	1	1	1	0	1
4	0	1	1	1	1	0	1



599 observations of 599 variables

	row.names	775	2166	2167	2465	3881	3889
1	775	1.96980	0.57420	0.57420	0.72360	1.05780	0.53220
2	2166	0.57420	1.99300	1.99300	0.67240	0.41160	0.70620
3	2167	0.57420	1.99300	1.99660	0.67240	0.41160	0.70620
4	2465	0.72360	0.67240	0.67240	1.99700	0.50900	0.63160

We fit the model:

$$\boldsymbol{Y} = \mathbf{1}\mu' + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

$\boldsymbol{B} \sim MG(\mathbf{0}, \boldsymbol{I}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega}$ (unknown) covariance matrix for effects. We assign an inverse Whishart distribution to $\boldsymbol{R}_0$ and $\boldsymbol{\Omega}$ (default settings in the software).

```
#Loading software
library(BGLR)

#Loading data
data(wheat)
y<-wheat.Y
X<-wheat.X
X<-scale(X)/sqrt(ncol(X))

#Linear predictor
ETA<-list(list(X=X,model="BRR"))

#Model fitting
set.seed(123)
fm<-Multitrait(y=y,ETA=ETA,nIter=10000,burnIn=5000)
```

Continue...

```
#Residual covariance matrix
fm$resCov

#Genetic covariance matrix
fm$ETA[[1]]$Cov

#Marker effects
fm$ETA[[1]]$beta

#Edited output
>fm$resCov
...
$type
[1] "UN"

$R
            1          2          4          5
1  0.53315070 0.07797869 -0.1055548 0.05279538
2  0.07797869 0.57551110  0.3034339 0.16497247
4 -0.10555484 0.30343389  0.6196055 0.11723637
5  0.05279538 0.16497247  0.1172364 0.59673421
```

Continue...

```
#Edited output
> fm$ETA[[1]]$Cov
...
$type
[1] "UN"

$Omega
           1          2          4          5
1  0.5632461 -0.1146953 -0.1213392 -0.2037660
2 -0.1146953  0.5058328  0.4081491  0.2661113
4 -0.1213392  0.4081491  0.4966503  0.2942788
5 -0.2037660  0.2661113  0.2942788  0.4959512
...
```

Equivalent code to specify covariance structures:

```
ETA<-list(list(X=X,model="BRR",Cov=list(type="UN")))
residual<-list(type="UN")

#Model fitting
set.seed(123)
fm<-Multitrait(y=y,ETA=ETA,resCov=residual,nIter=10000,burnIn=5000)
```

Example 2: Random effect model with unstructured covariance matrices

We fit the model:

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\mu}' + \boldsymbol{U} + \boldsymbol{E},$$

$\boldsymbol{U} \sim MG(\mathbf{0}, \boldsymbol{K}, \boldsymbol{G})$, $\boldsymbol{K}$ (known) covariance matrix to model relationship between individuals and $\boldsymbol{G}$ (unknown) covariance matrix to model within subject relations. We assign an inverse Whishart distribution to $\boldsymbol{R}_0$ and $\boldsymbol{G}$ (default settings in the software).

```
library(BGLR)
data(wheat)
K<-wheat.A
y<-wheat.Y

ETA<-list(list(K=K,model="RKHS"))

#Fit model
set.seed(123)
fm<-Multitrait(y=y,ETA=ETA,nIter=10000,burnIn=5000)
```

Continue...

```
#Residual covariance matrix
fm$resCov

#Genetic covariance matrix
fm$ETA[[1]]$Cov

#Random effects
fm$ETA[[1]]$u

#Edited output
>fm$resCov
...
$type
[1] "UN"
...
$R
             1         2          4           5
1  0.567446646 0.0343702 -0.1180475 0.004740618
2  0.034370198 0.5645890  0.2450678 0.151970705
4 -0.118047470 0.2450678  0.5043002 0.103934039
5  0.004740618 0.1519707  0.1039340 0.531484892
```

Continue...

```
#Edited output
>fm$ETA[[1]]$Cov
...
$type
[1] "UN"
...
$G
            1           2           4           5
1  0.30324330 -0.04113035 -0.06129195 -0.08841091
2 -0.04113035  0.28808179  0.26584470  0.15342103
4 -0.06129195  0.26584470  0.34676292  0.17047220
5 -0.08841091  0.15342103  0.17047220  0.30945972
```

Questions?

References

Geman, S. and Geman, D. (1984).
Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.
IEEE Transactions on pattern analysis and machine intelligence, 6, 6, 721–741.

Sorensen, D. and Gianola, D. (2007).
Likelihood, Bayesian, and MCMC methods in quantitative genetics.
Springer Science & Business Media.