Inferring Biadditive Models within the Bayesian Paradigm

Julie Josse[1]

Jean-Baptiste Denis[2]

Rapport technique 2012-4, 50 pp.

Unité Mathématiques et Informatique Appliquées
INRA
Domaine de Vilvert
78352 Jouy-en-Josas cedex
France

[1]Julie.Josse@Agrocampus-Ouest.Fr

[2]Jean-Baptiste.Denis@Jouy.Inra.Fr

© 2012 INRA

A first version of this technical report has been prepared for a talk given at the extraordinary MIAJ seminar of April, 4.

(25 avril 2012)

Abstract

So called *biadditive models* (most commonly known as AMMI models for Additive Main effect and Multiplicative Interaction) are frequently used to interpret the main traits of two ways data, for instance for the interpretation of *genotype by environment interactions.* Linked with PCA technics, they provide efficient empirical descriptions of matrix structures.

The use of Bayesian approaches in statistical analysis in increasing for many statistical models due to the new computer capacities and the existence of specialized algorithms to draw into posterior distributions.

Some work was already presented to deal with biadditive models in a Bayesian way. Here, we consider the point, proposing a new solution directly on the overparameterized model which allows one the use of standard softwares, for instance BUGS implementations.

We first give a detailed presentation of our proposal and then apply it to a real data set coming from the litterature, focusing on the interpretation.

In the appendix, the proposal to deal with overparameterized models is developped for any type of models.

Key Words biadditive - Bayesian - AMMI - overparameterization - GEI - BUGS

Résumé

Les modèles biadditifs (souvent appelés modèles AMMI pour Additive Main effect and Multiplicative Interaction) sont fréquemment employés pour l'interprétation de tableaux de données à deux entrées, par exemple pour l'interprétation des interaction génotype-milieu. Proches des analyses de données factorielles, ils représentent un outil efficace pour une description parcimonieuse de structure matricielles.

D'autre part, l'utilisation d'approches bayésienne en statistique se généralise. C'est la conséquence de la puissance accrue des calculateurs et de la disponibilité d'algorithmes spécialisés pour tirer des échantillons dans les distributions a posteriori.

Plusieurs travaux ont déjà été réalisés pour traiter les modèles biadditifs de manière bayésienne. Dans ce rapport, nous proposons une nouvelle approche qui prend directement en compte une définition surparamétrée de ces modèles. Son principal avantage est de permettre l'utilisation d'algorithmes génériques, comme ceux proposés dans les logiciels de la famille BUGS.

Un exemple de données de la litérature est traité de manière détaillée après une présentation formalisée de la proposition.

Dans l'annexe, la prise en compte de la surparamétrisation d'un modèle dans une approche bayésienne est traitée de manière complètement générale.

Mots clef biadditif - bayésien - AMMI - surparaméterisation - IGM - BUGS

Contents

1 Introduction **7**

2 Reminders **7**

2.1 About Bayesian statistics . . . 7

2.1.1 The Bayesian paradigm . . . 7

2.1.2 Advantages . . . 7

2.1.3 Drawbacks . . . 8

2.2 The biadditive model . . . 9

2.2.1 Definition . . . 9

2.2.2 Constraints . . . 9

2.2.3 Least squares estimation . . . 10

2.3 Literature proposals . . . 10

3 Defining a prior **11**

3.1 Using prior in an overparameterized framework . . . 12

3.2 Standard complete proposal . . . 13

3.3 Prior marginal distribution of the data . . . 14

4 Accessing to the posterior distribution **15**

4.1 Algebraically . . . 15

4.2 Graphical presentation: the DAG . . . 15

4.3 Simulation study . . . 16

5 Using posterior simulations: a worked example **21**

5.1 Comparing prior versus posterior . . . 21

5.2 Looking for a sensible number of multiplicative terms . . . 23

5.3 Interpreting the parameters . . . 23

5.4 Interpreting the genotype responses across the environments . . . 25

5.5 Missing values . . . 29

5.6 Exceeding a threshold . . . 30

5.7 Random versus fixed effects? . . . 30

5.8 Comparison with Perez et al. proposal . . . 32

6 Conclusion **32**

A Miscellaneous formulae 36

A.1 Conditional expectation and variance 36

A.2 First two moments of E_{ijk} 36

A.3 First two moments of a triple product 36

B von Mises-Fisher distribution 37

B.1 Definition 37

B.2 Interpretation of the von Mises-Fisher distribution 37

C Prior/posterior with overparameterized models 38

C.1 Linear tiny case 38

C.1.1 Definition 38

C.1.2 Reparameterization 39

C.1.3 Posterior in the new parameterization 40

C.1.4 Conclusion 40

C.2 Additive case 40

C.2.1 Definition 40

C.2.2 First reparameterization 40

C.2.3 Second reparameterization 42

C.3 General formulation 42

C.3.1 Definition of overparametrization 42

C.3.2 Dealing with overparameterization 43

C.3.3 Conclusion 44

C.4 Biadditive case 44

D Gibbs sampling algorithms used in the literature proposals 44

D.1 Brief reminder about Gibbs sampling 44

D.2 Hoff's proposal 45

D.3 Perez et al.'s proposal 46

E BUGS coding of the model 46

1 Introduction

Biadditive models [Denis.1992, Denis.1994, Denis.1996, Denis.1998] are frequently used to interpret the interaction between two factors. For instance in plant breeding, it is usual to study the interaction between varieties and environments for different purposes such as the selection of new varieties. Biadditive models present several advantages compared with the usual analysis of variance models with interaction terms: (i) simplification of the interaction scheme, (ii) better estimation because all results are used for the estimation of a unique combination[1], (iii) the possibility to have an idea of the interaction and estimate the error variance when replicates are absent, even with missing values.

In some recent papers [Perez.2011, Crossa.2011], it can be observed that the Bayesian approach was proposed for biadditive models. Most of the good properties of the Bayesian approaches are the consequences of a clear separation between the modelling on parameters (definition of priors) and the statistical analysis (getting information from the data set). In this report, we will have a look at that possibility, get some experience on it and propose a new treatment of such models with the Bayesian point of view.

2 Reminders

2.1 About Bayesian statistics

2.1.1 The Bayesian paradigm

Let $[Y \mid \theta]$ be the likelihood distribution[2] of a data set Y defined through a vectorial parameter θ. In Bayesian statistics, θ is not considered as having a fixed and unknown value but as being a random variable, the distribution of which characterizes the degree of certainty we have on the values it can take. So to complete the description model, the marginal distribution of θ, $[\theta]$, must be provided. This is the *a priori* distribution, or prior for shortness.

Performing a Bayesian statistical analysis is no more that applying twice the so-called Bayes theorem to get $[\theta \mid Y]$ denominated *a posteriori* distribution :

$$[\theta \mid Y] = \frac{[Y \mid \theta]\,[\theta]}{[Y]} \tag{1}$$

which is possible since the numerator is the joint distribution allowing the computation of $[Y]$, the marginal distribution of Y.

2.1.2 Advantages

More freedom about used probability distributions Thanks to stochastic algorithms (most known are MCMC), a lot of distributions are at hand. Among the points to underline when using a BUGS facility (see the Jags manual [Plummer.2011]), one can check that

[1] When a saturated interactive modelling is used the estimation of the expectation of one variety-environment combination relies only on the available observations for this combination.

[2] Square brackets stand for density distribution of a random variable; vertical bar for the conditioning operator. Then $[Y \mid \theta]$ reads as the density distribution of Y conditioned by θ.

- Truncated distributions can be incorporated in a standard way,
- Censored data can be used all the same,
- A large spectrum of discrete or continuous, scalar or multivariate distributions are available,
- Standard transformations are implemented.

No fear of non identifiability When no information is available from the data, we are left at the prior level of information (only for proper priors). For the same reason missing values are no more a problem. In fact, the viewpoint is shifted: we are not analyzing a data set, we are (i) defining a model on some phenomenum (establishing a prior distribution) and (ii) extracting the information relative to this model from available dataset(s) (getting the posterior distribution). For that reason, Bayesian approaches are very convenient for statistical meta-analyses.

A possibility to introduce knowledge When defining the prior distribution, one can incorporate, with much more flexibility than when defining the likelihood of a data set, the already knowledge about the phenomenon under study (expert knowledge, historical data, etc.). This is why sometimes, Bayesian approaches are linked to learning processes: the posterior is naturally the prior of new statistical analyses.

Of course, a sequential incorporation of data via such a prior/posterior scheme is equivalent to a unique global statistical inference.

Consistent inference for any transformation If a correct Bayesian inference is made on some set of parameters (say θ), a transportation of the probability distribution gives us a consistent inference to any function of it. No longer useful to decide if we are interested in the standard deviation (σ), the variance (σ^2) or the precision ($\frac{1}{\sigma^2}$)! Credibility intervals will correspond exactly.

Of course this is applicable to vectorial situations.

Ease of interpretation The output (posterior distribution) is of the same nature than the input (prior distribution) so if one is able to propose a prior, he must be able to well use the posterior. Moreover, compared to the frequentist approach, the notion of credible intervals can be seen as easier to interpret than the notion of confidence interval. Indeed, in the latter, the randomness is on the data (we consider all possible data, if we were able to do many experiments) whereas in the credible interval the randomness is on the parameter side.

2.1.3 Drawbacks

A prior has to be defined This is the main criticism again the Bayesian statistical practice. According to the statistician defining the prior, results will be different, then not objective. This is true, but can be viewed as an advantage because good statisticians will get better results which is fair? Also bad statisticians will be wrong as well when defining the Likelihood? Also, defining the prior imply a useful consideration of the model in use, which is often a profitable spent time.

By the way, it is worth recalling that the joint prior distribution on all model parameters has to be defined and not only the series of scalar marginals. This can be a quite tricky issue[3], sometimes solved with a reparameterization.

Difficult to question the model In classical statistics, there are several ways to question a model (via the study of residuals, for example) or at least to select a model within a family of models. This is not so easy in the Bayesian approach where a difficult point is the definition of fair priors for different models. Fair would be that the priors be equivalent at the level of every used observation but this is rarely the case because the number and natures of the parameters[4], is compulsorily different from a model to another.

The correct answer would be to imagine a hierarchy of the considered models, giving each a prior probability... this is not, for the moment, easily tractable.

2.2 The biadditive model

2.2.1 Definition

This class of models is known for many years since there is was described in 1923 in one of the Fisher's papers [Fisher.1923]. The terminology was proposed by Denis and Gower [Denis.1992] and we will stick to it, considering only the $B(m, a, b, \pi_Q)$[5] variety defined as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq} + E_{ijk} \tag{2}$$

where Y_{ijk} is the kth observation for the combination of genotype i and environment j. The number of observations for each combination, K_{ij}, being variable (but known) including zero which means missing value. Most analyses are carried out with $K_{ij} = 1$. The number of genotype is denoted with I, and the number of environments J. Here $Q \in \{0, ..., \min(I-1, J-1)\}$; when $Q = 0$, we are dealing with the additive model; when $Q = \min(I-1, J-1)$, it is the saturated interactive model (no restriction on the interactive term). E_{ijk} is the error term that includes the design effect, here we will assume that these errors are independent, centred, with same variance (σ_E) and normal distribution.

2.2.2 Constraints

This is one of the difficult point about biadditive models, overparameterization is present and there is a need for additional constraints to determine unique values for the parameters. We would like constraints which lead to an easy interpretation of the parameters, inducing simple statistical derivations and most often keeping the symetry between the levels of the two factors.

[3] There are two way: (i) use a multivariate distribution $[\theta_1, \theta_2, \theta_3]$ or (ii) go through conditional definition $[\theta_1][\theta_2 \mid \theta_1][\theta_3 \mid \theta_1, \theta_2]$.

[4] The number of parameters is also difficult to define in a Bayesian framework. However, criteria that mimics the AIC have been proposed such as the DIC to select models. One can also talked about the Bayes' Factor but which is hard to compute.

[5] The presence of m indicates the constant parameter μ, this of a the presence of the row main effect, b for the column main effect and Q precises the number of involved multiplicative terms in the interaction term. When there are no interaction the term π vanishes and when $Q = 1$, only π is left.

Let us define the effect term by

$$\begin{aligned}\mu_{ij} &= E\left(Y_{ij} \mid \mu, \alpha, \beta, \lambda, \gamma, \delta\right) \\ &= \mu + \alpha_i + \beta_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}. \qquad (3)\end{aligned}$$

It is well known [van_Eeuwijk.1996] that the parametric dimension of μ_{ij} is $(1+Q)\left((I+J)-(1+Q)\right)$ meanwhile the number of parameters in (3) is $(1+Q)(1+I+J)$ so that $(1+Q)(2+Q)$ additional constraints have to be introduced. Here we will use the standard ones:

$$\begin{aligned}\mathbf{1}'_I\alpha = \mathbf{1}'_J\beta &= 0 \\ \mathbf{1}'_I\gamma_q = \mathbf{1}'_J\delta_q &= 0 \text{ for } q \in \{1, ..., Q\} \qquad (4) \\ \gamma'\gamma = \delta'\delta &= \mathbf{I}_Q \\ \lambda_1 \leq \lambda_2 \leq \ ... &\leq \lambda_Q.\end{aligned}$$

where γ_q and δ_q are the qth columns of matrices γ and δ of respective sizes $(I \times Q)$ and $(J \times Q)$.

In fact this set of constraints is not sufficient since they don't fix the orientations of the (γ_q, δ_q) which can be simultaneously inverted. Most often, this indetermination is not considered. Here we will use the non standard specification of orientation borrowed from [Viele.2000] that the λs be not restricted and that γ_{1q}s and δ_{1q}s be always positive. Even if it looks like breaking the symetry between the levels of the factors out, giving a different role to the first levels, we will see later on that this is not the case.

2.2.3 Least squares estimation

When the number of replicates is constant ($K_{ij} = K > 0$), with the proposed constraints (4), the least squares estimates of the biadditive model are easy to obtain. Let $\overline{Y}$ be the $I \times J$ matrix of the mean by cell $\overline{Y_{ij}} = \frac{1}{K}\sum_k Y_{ijk}$.

$$\begin{aligned}\mu &= \left(\mathbf{1}_I\left(\mathbf{1}'_I\mathbf{1}_I\right)^{-1}\mathbf{1}'_I\right)\overline{Y}\left(\mathbf{1}_J\left(\mathbf{1}'_J\mathbf{1}_J\right)^{-1}\mathbf{1}'_J\right) \\ \alpha &= \overline{Y}\left(\mathbf{1}_J\left(\mathbf{1}'_J\mathbf{1}_J\right)^{-1}\mathbf{1}'_J\right) \\ \beta &= \overline{Y}'\left(\mathbf{1}_I\left(\mathbf{1}'_I\mathbf{1}_I\right)^{-1}\mathbf{1}'_I\right)\end{aligned}$$

and the parameters of the interaction terms are given by the singular vectors and singular values for the biggest Q singular values of the singular value decomposition of the row and column centered matrix

$$\left(\mathbf{I}_I - \mathbf{1}_I\left(\mathbf{1}'_I\mathbf{1}_I\right)^{-1}\mathbf{1}'_I\right)\overline{Y}\left(\mathbf{I}_J - \mathbf{1}_J\left(\mathbf{1}'_J\mathbf{1}_J\right)^{-1}\mathbf{1}'_J\right).$$

2.3 Literature proposals

Recently some authors [Viele.2000, Crossa.2011, Perez.2011] have used biadditive models within a Bayesian framework. Their motivation is that the Bayesian approach may offer solutions to issues that are often difficult to handle such as: unbalanced data, unequal cell size, heteroscedastic data, the difficult choice of the number of relevant dimensions for the interaction terms, etc. Moreover, a Bayesian approach also offers distributions of any quantity of interest (to be compare with point estimates and their confidence intervals of the maximum likelihood approach) which may permit to construct credible areas in the biplot for example. Finally, it allows one to take into account in the analysis previous information such as historical data which may be very interesting in the analysis of genotype-environment data. We can remark that compared to the surge of interest in using Bayesian approaches to study models of very distincs types, only few proposals have been made for the biadditive models. This can be explained by the difficulty to work in a overparametrization framework (as we will see later). More precisely, the problem can be quite easily tackle for the linear terms of the model but the major difficulty is to put priors on the interaction terms taking into account the constraints.

Viele and Srinivasan (2000) [Viele.2000] seems to be the first ones to propose a Bayesian treatment of such models. They put uniform priors on the first column of the matrices γ and δ. Since each vector has zero sum and unit length, it corresponds to put uniform prior on a I (respectively J) dimensional unit sphere. Then, they work sequentially and take as conditional prior of each (γ_q)(respectively (δ_q)) given the previous dimensions a uniform distribution on the correct subspace. Of course, due to the constraints (the vectors must be normalized and orthogonal to the previous ones), it is not easy to define the supports and to sample from uniform distributions with the correct supports. They proposed a method to do so and then used it in a specific Gibbs sampler. On a real data set [Crossa.2011], their approach seems quite promising and offers confidence regions helping in the interpretation of the results .

Spherical uniform distribution is a special case of von Mises-Fisher [VMF] distributions (see §B). Such distributions exist also for matrices. More precisely, the set of orthonormal matrices is called the Stiefel manifold [Chikuse.2002]. The von Mises-Fisher distributions are distributions over this manifold. [Hoff.2012, Hoff.2009, Smidl.2007] used these distributions in a Bayesian treatment of models based on singular value decompositions of particular matrices such as models for Principal Components Analysis. From a computational point of view, these models are closed to the linear-bilinear ones, the main difference being that the linear part is not included. More precisely, [Hoff.2012, Hoff.2009, Smidl.2007] proposed to use as prior the uniform distributions for matrices γ and δ, indeed a special case of von Mises-Fisher distributions. Using such priors allow them to ensure the orthonormality constraints at the posterior level since the posterior distributions are also VMF distributions[6]. More details about the method and the associated algorithm is given in appendix §D.

Very recently, in the framework of the analysis of genotype-environement data, [Perez.2011] proposed also to use as prior distributions for matrices γ and δ specific von Mises-Fisher distributions. Their method detailled in appendix §D remains to put priors onto the cells of the matrix $\overline{Y}$.

3 Defining a prior

One of the major difficulty to tackle the biadditive model is to ensure a prior on the parameters taking into account the over-parameterization and the associated constraints. As far as the constraints are linear, it is not too much problematic. But for the bilinear constraints on γ and δ matrices to be a set of Q orthonormal columns, no standard solution is at hand. In that section, we propose a straightforward solution having the great advantage to be implementable in the standard BUGS softwares.

[6] but no more uniform ones.

3.1 Using prior in an overparameterized framework

In the appendix (§C) we have considered the overparameterization difficulty with a reparameterization clearly separating the involved parameters into two transformed subsets $(\phi_1(\theta), \phi_2(\theta))$; the first is sufficient to define the data likelihood $[Y \mid \theta] = [Y \mid \phi_1(\theta)]$ and the second is left redundant once the first is taken into account $[\phi_2(\theta) \mid Y, \phi_1(\theta)] = [\phi_2(\theta) \mid \phi_1(\theta)]$. More, for multinormal priors, we achieved a clear separation getting the independence, i.e. $[\phi_1(\theta), \phi_2(\theta)] = [\phi_1(\theta)][\phi_2(\theta)]$ which implies that the prior and posterior distributions of $\phi_2(\theta)$ are identical, no information is given by the data onto the $\phi_2(\theta)$ parameters.

It can be also argued that the overparemeterization problem is identical to the missing value situation. Indeed, let us take the very simple example of one way factor without replicate and with known variance. Let the parameters be $\{\theta_1, \theta_2, ..., \theta_I\}$ and consider a prior-likelihood couple as

$$\begin{aligned} [\theta_i] &\sim N(50, 2) \text{ whatever is } i = 1, ..., I \text{ independent,} \\ [Y_i \mid \theta_i] &\sim N(\theta_i, 1) \text{ whatever is } i = 1, ..., I \text{ independent.} \end{aligned}$$

There is no overparameterization in the sense that every modification of any non constant function of the θ_i modifies the likelihood. But if we remove Y_1 from the data set, θ_1 can be modified without change in the likelihood inducing overparameterization. Nevertheless there is no harm for the Bayesian statistician, the joint distribution over parameters and data can be calculated and the posterior deduced. Of course, it is obvious that $\left[\theta_1 \mid (Y_i)_{i=1,...,I}\right] = [\theta_1]$, the posterior of θ_1 is equal to its prior, no information have been born onto this parameter by the data, logical enough.

The same can can be proposed for every overparameterized model. The functions of the parameters associated to the neccessary constraints of the type $f(\theta) = 0$ to prevent the indetermination does not influence the likelihood.

But we could imagine that some data be associated to them, producing complete identifiability of the parameter set. Why not to argue that these data are missing? And that only their prior will be identical to their posterior.

Taking the case of the $B(m, a, b, \pi_2)$ model with no replicates:

$$Y_{ij} \sim N(\mu + \alpha_i + \beta_j + \lambda_1\gamma_{i1}\delta_{j1} + \lambda_2\gamma_{i2}\delta_{j2}, 1) \text{ for } i = 1, ..., I \text{ and } j = 1, ..., J,$$

we can say that are missing some observations to replace the constraints, for instance

$$\begin{aligned} Y_{10} &\sim N(\mu + \alpha_1, 1) \\ Y_{01} &\sim N(\mu + \beta_1, 1) \\ Y_{-1,-1} &\sim N(\lambda_1) \\ Y_{i,-1} &\sim N(\lambda_1\gamma_{i1}, 1) \text{ for } i = 1, ..., 2 \\ Y_{-1,j} &\sim N(\lambda_1\delta_{j1}, 1) \text{ for } j = 1, ..., 2 \\ Y_{-2,-2} &\sim N(\lambda_2) \\ Y_{i,-2} &\sim N(\lambda_2\gamma_{i2}, 1) \text{ for } i = 1, ..., 3 \\ Y_{-2,j} &\sim N(\lambda_2\delta_{j2}, 1) \text{ for } j = 1, ..., 3 \end{aligned}$$

And it is clear that they ensure that all parameters (and every functions of them) are informed. So in principle a good algorithm to produce the posterior will work even with these values missing.

And at the moment of the interpretation we will be free to consider only functions of the parameters which are *identifiable*, that is with a posterior different from the prior. The first candidates are

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \lambda_1\gamma_{i1}\delta_{j1} + \lambda_2\gamma_{i2}\delta_{j2} \text{ for } i = 1, ..., I \text{ and } j = 1, ..., J$$

and we will use them (see the model code in §E) and base our interpretations from functions of them.

3.2 Standard complete proposal

We will use[7] the following independent normal distributions as priors plus a uniform for the error variance.

$$\begin{array}{rcl}
\mu & \sim & N\left(m, s_\mu^2\right) \\
\alpha_i & \sim & N\left(0, s_\alpha^2\right) \\
\beta_j & \sim & N\left(0, s_\beta^2\right) \\
\left(\lambda_q\right)_{q=1...Q} & \sim & \text{ordered sample of } Q \text{ independent } N\left(0, s_\lambda^2\right) \\
\gamma_{1q} & \sim & HN\left(0, 1\right) \\
\gamma_{iq} & \sim & N\left(0, 1\right) \text{ for } i > 1 \\
\delta_{1q} & \sim & HN\left(0, 1\right) \\
\delta_{jq} & \sim & N\left(0, 1\right) \text{ for } j > 1 \\
\sigma_E & \sim & U\left(0, S_{ME}\right)
\end{array} \qquad (5)$$

where $HN(0, 1)$ stands for the half-normal distribution (truncation of $N(0, 1)$ on the positive values). Only five hyparameters are used for the definition[8]; their interpretation[9] is:

- m the guessed mean value of the studied variable,
- s_μ quantifies the uncertainty we believe to have about m,
- s_α quantifies the variability we believe the genotype main effects have got,
- s_β quantifies the variability we believe the environment main effects have got,
- s_λ quantifies the variability we believe the GE interaction is on its different Q[10] components[11],

[7] Even if it does not matter, it would have been more consistent to use $\gamma_{iq} \sim N\left(0, I^{-1}\right)$ and $\delta_{jq} \sim N\left(0, J^{-1}\right)$ to stick to the choosen constraints (4).

[8] They are supposed to be numerical values, so expressed with latin letters.

[9] In the following we give a precise meaning to the terms *variability* and *uncertainty*. Variability is the variation linked to a random variable we are interested in. For instance, if we are interested in the yield of next year some variability has to be introduced as a consequence of the not known and influential climate effect. Uncertainty is related to a lack of knowledge about some value of interest, it could be decreased by acquiring additional relevant data. When a volume of grains is sampled to get the wetness of the crop, taking a greater volume will certainly reduce the uncertainty we have about the humidity of the total crop.

[10] The determination of the number of multiplicative components is a difficult question, we will see in Figure 7 how this can be approached from the posterior distributions to reduce their number.

[11] Of course, different distributions could have been used for the different λ_q, if some knowledge is available.

- S_{ME} quantifies both the remaining variability not taken into account previously and the uncertainty of the measurement[12].

Notice that

- we spoke about *uncertainty* only for the parameter μ since it represents the parameter insuring the translation invariance of the model (no need to have variability on it, it can be incorporated into the main effects).
- from our point of view, if we would like to introduce *uncertainty* on the other parameters, either an expectation has to be introduced as a random variable, and/or the standard deviation parameter has to be supposed random. Taking the example of the genotype main effect:
$$\alpha_i \mid \alpha_{0i} \sim N\left(\alpha_{0i}, s_\alpha^2\right) \text{ and } \alpha_{0i} \sim N\left(0, s_{u\alpha}^2\right)$$
or
$$\alpha_i \mid \varsigma_\alpha \sim N\left(0, \varsigma_\alpha^2\right) \text{ and } \varsigma_\alpha \sim U\left(0, S_{M\alpha}\right)$$
or both. The distribution of the newly introduced α_{0i} and/or ς_α could be interpreted as uncertainty.

Important remark When looking at the definition of the von Mises-Fisher distribution (§B), it is clear that the distributions induced by our proposal onto the orthonormal matrices γ and δ are indeed von Mises-Fisher distributions: the uniform distributions onto their spaces.

3.3 Prior marginal distribution of the data

We believe that it is quite important to look at what represents the information put in the prior definition step. In order to do that, the more natural way is to look, using the likelihood, at the induced prior distribution on the data set. This is done integrating the joint $[\theta, Y]$ distribution over θ.

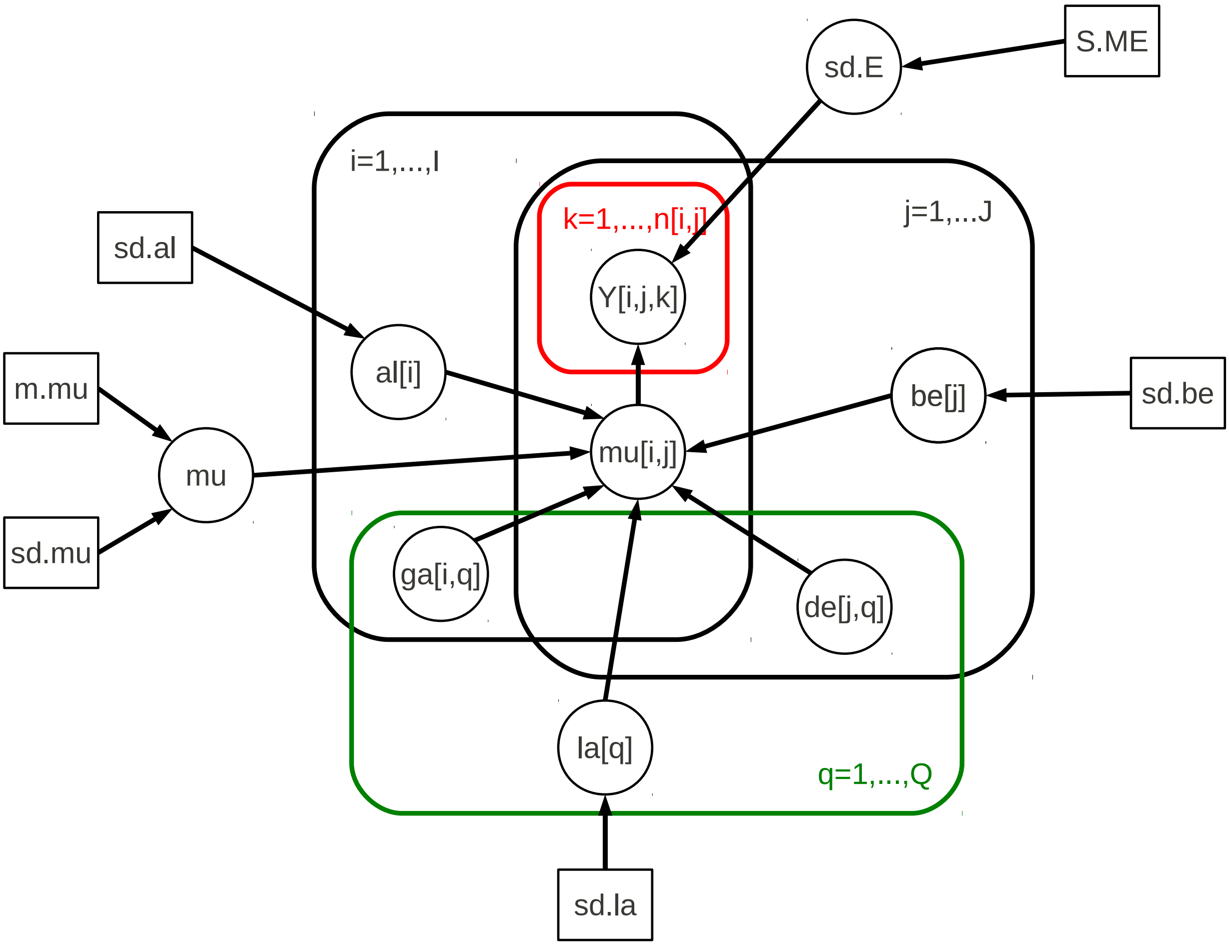
Due to the presence of the bilinear terms, it is not possible to obtain in close form the distribution of Y_{ijk} but it is possible to explicit expressions for their first two moments and also to simulate them. As defined in (2) and with the retained priors (5), one can check that (using results presented in §A.3 and §A.2) that

$$\begin{aligned}
E\left(Y_{ijk}\right) &= m \\
V\left(Y_{ijk}\right) &= s_\mu^2 + s_\alpha^2 + s_\beta^2 + s_\lambda^2 + \frac{1}{3} S_{ME}^2 \\
Cov\left(Y_{ijk}, Y_{ijk'}\right) &= s_\mu^2 + s_\alpha^2 + s_\beta^2 + s_\lambda^2 \\
Cov\left(Y_{ijk}, Y_{ij'h}\right) &= s_\mu^2 + s_\alpha^2 \\
Cov\left(Y_{ijk}, Y_{i'jh}\right) &= s_\mu^2 + s_\beta^2 \\
Cov\left(Y_{ijk}, Y_{i'j'h}\right) &= s_\mu^2
\end{aligned}$$

Notice that these results apply equally well when $i = 1$ or/and $j = 1$, even if the computation is slightly different. This is why we said in §2.2.2 that the symetry was not lost with the proposed constraints.

[12] Gamma distributions were often used for variance parameters, but recent works show that it is possible and more efficient to put uniform priors

Figure 1: Directed Acyclic Graph associated to the Bayesian approach. Rounded rectangles are used for loops. Circled nodes are random variables. Rectangular nodes are constant. Arcs indicate the direct relationships between the nodes. `mu` is used for μ, `al` for α,...



4 Accessing to the posterior distribution

4.1 Algebraically

No close form of posterior can be imagined[13], especially in the case of a non equally replicated design where even LS estimators are not known. Nevertheless, there is no difficulty (in all trials we did) to get simulated values from the posterior using Jags [Plummer.2011] one of the BUGS softwares.

4.2 Graphical presentation: the DAG

To better understand the mechanism of the modelling, it is of interest to build the underlying directed acyclic graph, so called DAG; it is presented in Figure 1. It emphasizes the central role

[13] as a consequence of the bilinear terms.

Table 1: Parameter values used for the simulation in §4.3

$\mu = 100$
$(\alpha_i) = (-1, -1, 0, 1, 1)$
$(\beta_j) = (-4, -3, -2, -1, 0, 1, 2, 3, 4)$
$\lambda = 12$
$(\gamma_i) = \left(\frac{2}{\sqrt{10}}, \frac{1}{\sqrt{10}}, 0, -\frac{1}{\sqrt{10}}, -\frac{2}{\sqrt{10}}\right)$
$(\delta_j) = \left(\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0, -\frac{1}{2}, -\frac{1}{2}\right)$
$\sigma_E = \frac{3}{2}$

Table 2: ANOVA table of the parameter values given in Table 1 only based on the expectation. The error variance being equal to 2.25.

	df	SS	MS
row effect	4	36	9.0
column effect	8	300	37.5
interaction effect	32	144	4.5

of μ_{ij}, the expectation of the observed variable in a frequentist approach. We will see in §5.7 that a Bayesian approach allows one to play between fixed and random effect affectation.

4.3 Simulation study

In order to check that in practical situations, a Bayesian approach is effective for biadditive models, we performed a small simulation study. From a $B(m, a, b, \pi)$ biadditive expectation[14] we added a normal random error with identical standard deviations (σ_E) and fitted a $B(m, a, b, \pi_2)$ model. This was done a number of times, to escape a favorable (or unfavorable) case.

More precisely, we did 49 simulations, with a posterior sample of 50 draws with a thinning coefficient of 200 to avoid autocorrelations; not because the computation are lengthy: they last 140 seconds but because we wanted to display the results graphically. The introduced parameter values are shown in Table 1 resulting in the ANOVA table given in Table 2.

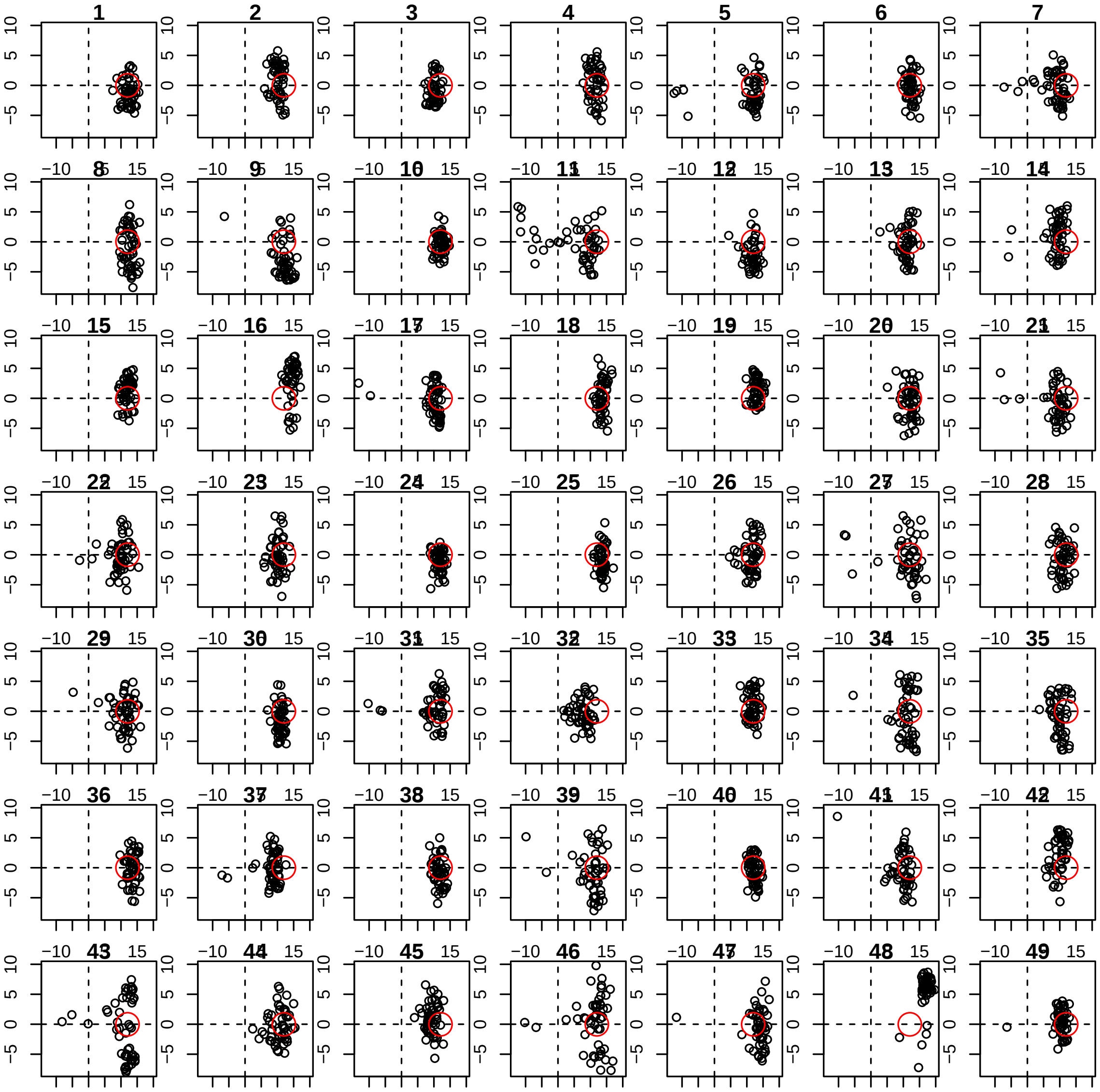
The results of the simulation study are given in Figures 2 (study of the suggested number of multiplicative terms), 3 (inferring the row main effects), 4 (inferring the row first interactive effects) and 5 (inferring the row second interactive effects).

Figure 2 displays the joint posterior distributions of (λ_1, λ_2). The values of λ_1 are positive for all the simulations, whereas λ_2 takes a continuum of values negative or positive. This is a clear indication that the credible region of λ_1 does not include the zero value, contrary to its of λ_2. Consequently, we can say that most often the Bayesian analysis correctly assess one multiplicative term (except in simulation 48 where the conclusion would be two multiplicative terms and a not so clear conclusion for simulation number 11 but still in favor of one multiplicative term). Here we can see an advantage to let the singular values being negative: it would have been less clear if the absolute value of the singular values have been used as commonly done.

The profile study of the row effects proposed in Figures 3, 4 and 5 would be much clearer if the bundle of the 50 profiles were replaced by the quantile profiles of a greater number of profiles. Straightforword (but tedious) algorithms can be implemented for that.

[14] additive part plus interaction with a unique multiplicative term.

Figure 2: Simulation study: number of multiplicative terms. For each of the 49 simulations, the simulated joint posterior density of λ_1 (abscissae) and λ_2 (ordinates) is displayed with 50 draws. The red circle indicates the true value ($\lambda_1 = 12, \lambda_2 = 0$). Due to the indetermination of the parameterization, negative values are equivalent to positive values.



Nevertheless α and γ_1 parameters seems well assessed. More heterogeneity occurs for γ_1 where particularly the already mentionned simulation number 11 has a few opposite profiles. Looking this case with attention, ones can notice that the γ_{11} parameter[15] is very close to zero and sometimes fautly; a free attribution would have given two sets of profile, one similar to the true profile and the other one opposite. The erratic behavior of γ_2 profiles (Figure 5) is consistent with the found unique dimension. Here too with the exception of simulation 48 which tended to propose two dimensions, where the bundle of profiles has got a systematic concave shape.

[15] The one forced to be positive due to our constraint system.

Figure 3: Simulation study: row main effects. For each of the 49 simulations, the simulated joint posterior density of the α_i profiles is displayed with 50 draws. A profile is defined as the lines linking (i, α_i); the red line indicates the true profile.

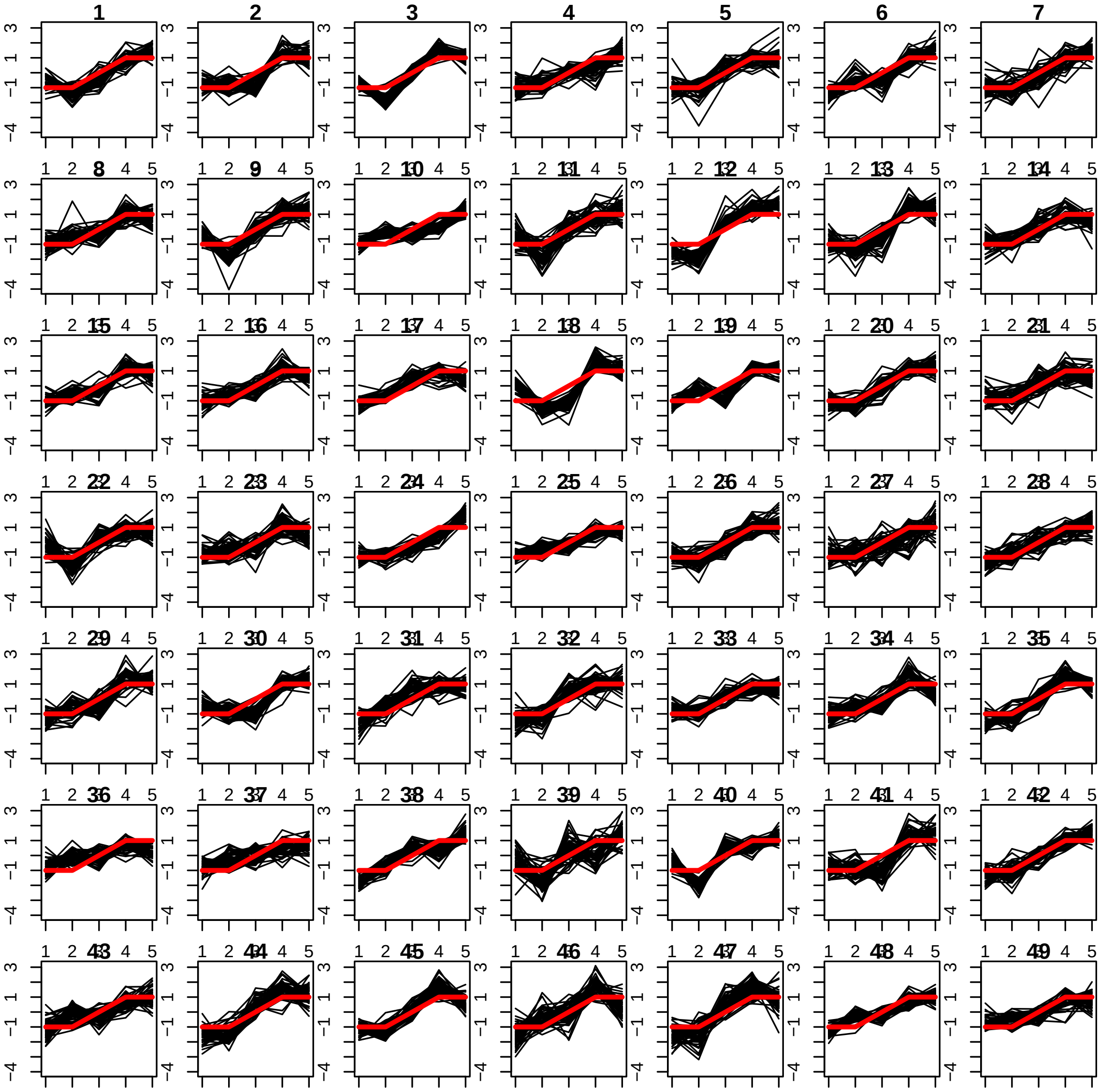


Figure 4: Simulation study: first interactive row effects. For each of the 49 simulations, the simulated joint posterior density of the γ_{i1} profiles is displayed with 50 draws. A profile is defined as the lines linking (i, γ_{i1}); the red line indicates the true profile.

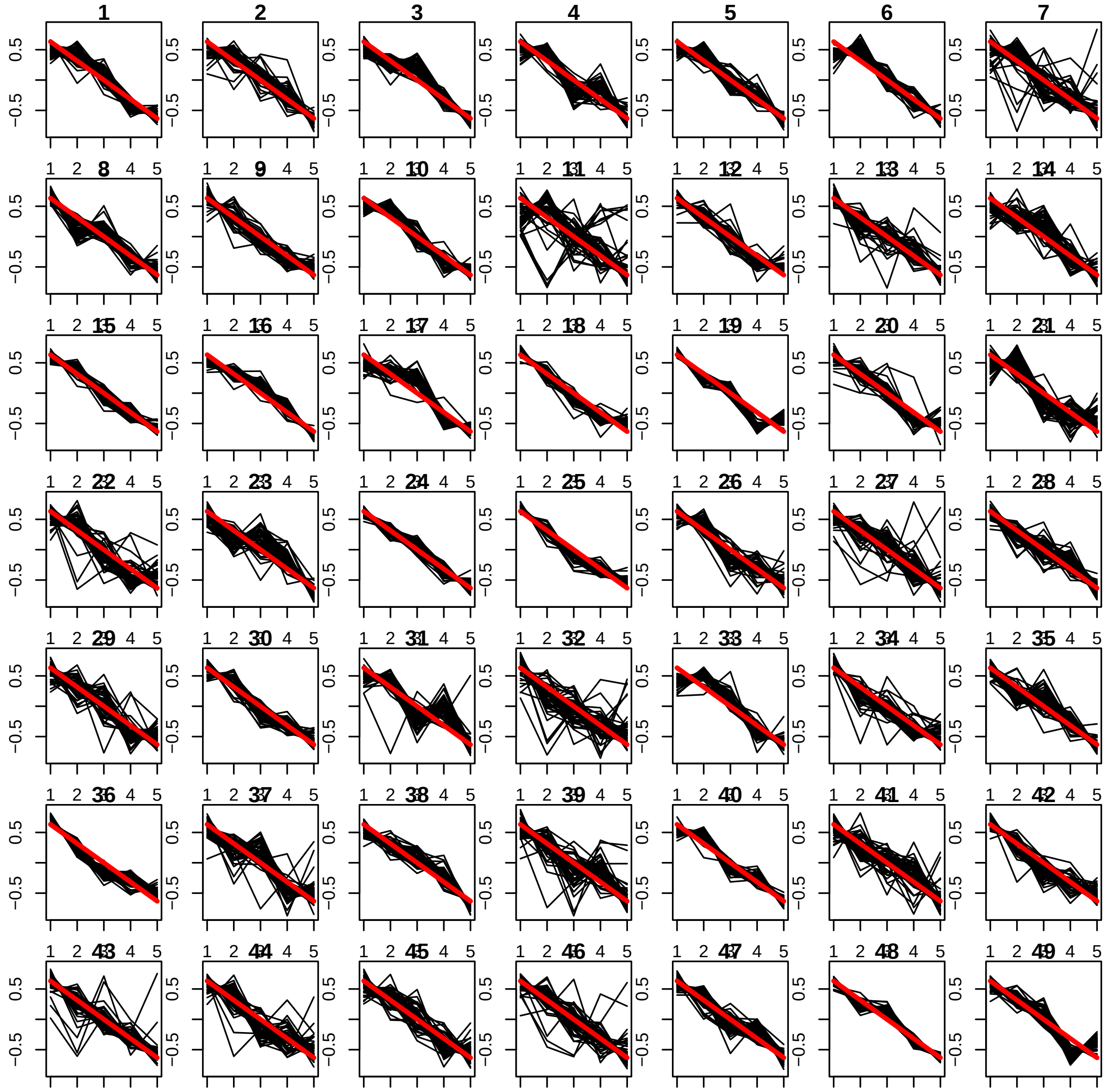
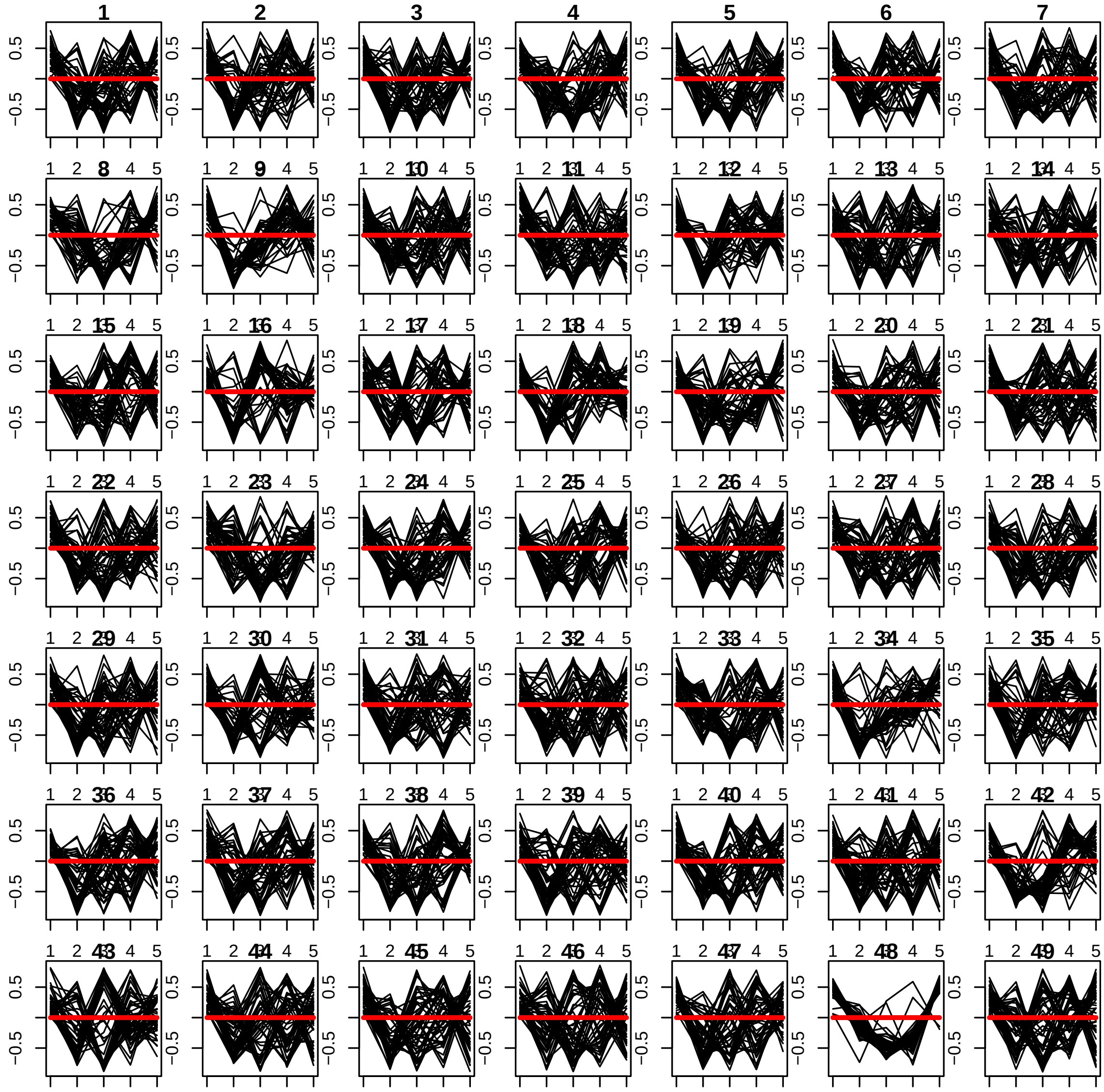


Figure 5: Simulation study: second interactive row effects. For each of the 49 simulations, the simulated joint posterior density of the γ_{i2} profiles is displayed with 50 draws. A profile is defined as the broken line linking (i, γ_{i2}); there is no true profile since there is no second multiplicative term in the proposed expectation; the red line is the null profile.



5 Using posterior simulations: a worked example

When analyzing a real genotype-environment data, five main questions[16] arise:

Q1 What is the genotype with the best performances across all the environments?

Q2 What is the genotype with the best performance for a specific environment?

Q3 Are the genotypes stable across all the environments?

Q4 Is it possible to rank the genotypes?

Q5 Could we assign a probability that a genotype will give more than a certain threshold?

Of course, the symetrical questions (considering genotype instead of environments) are also of value.

In this part, we try to address these questions. Moreover, in the classical statistical approach where estimates (and precision of them) of the biadditive model are obtained, it is not that easy to interpret the results. One can wonder if it would not be worst in a Bayesian context where distributions have to be handled instead of estimations? We would like to show that not.

To illustrate our purpose, we use a worked example using the data set proposed by [Crossa.1991] which comprises $I = 18$ genotypes and $J = 25$ environments, all combinations, no replicates. More precisely we will analyse it in four versions :

reduced using only the 40 combinations of the first 5 genotypes with the first 8 environments,

complete the complete data set,

reduced_missing the reduced data set where 4 values, randomly choosen, were considered missing,

complete_missing the complete data set where 45 values, randomly choosen, were considered missing.

Table 3 gives general features about the considered cases. From the ANOVA tables, one can conclude that important values have been eliminated when creating a 10% missing scheme in the reduced missing case; less obvious for the complete case.

The standard output of a MCMC algorithm is a simulation of the parameters from their posterior distribution. We will use a superscript s varying from 1 to S to indicate it. For instance α_i^s will be the sth simulation of the parameter α_i.

Because we are fan of them, we present the interpretation with graphs, but this also could be done with numerical values in tables.

5.1 Comparing prior versus posterior

As seen in §2.1.3, the major criticisms made to the Bayesian approach is the necessity to define a prior on the model parameters. Two precautions have to be taken with this respect.

Table 3: Some characteristics of the four data cases. The four computations were made with the same computer; 10^5 iterations were used for the burn-in phase; 10^4 were drawn with a systematic sampling of 10^{-2} giving 10^2 retained simulations.

case	ANOVA				duration (s)
		df	SS	MS	
reduced	Geno	4	1.0	0.25	3
	Envi	7	95.6	13.7	
	Geno:Envi	28	5.7	0.20	
complete	Geno	17	18.1	1.1	86
	Envi	24	2369.5	98.7	
	Geno:Envi	408	132.5	0.32	
reduced missing	Geno	4	3.9	0.96	2.7
	Envi	7	66.4	9.5	
	Geno:Envi	24	3.7	0.15	
complete missing	Geno	17	16.1	0.95	75
	Envi	24	2122.4	88.4	
	Geno:Envi	363	115.66	0.32	

Table 4: Prior definitions: used constant values referring to the distributions introduced in (5). Take care that implicitely all priors are supposed to be stochastically independent.

constants
$m = 5,\ s_\mu = 0.8$
$s_\alpha = 0.5$
$s_\beta = 0.5$
$s_\lambda = 0.5\sqrt{IJ}$
$S_{ME} = 2$

First the prior distribution has to be discussed with the experts of the field in order to incorporate to the priors as much as possible knowledge in them. Here, we directly used the rather vague priors shown in Table 4.

Second, a careful examination of the prior distributions has to be conducted on the quantities of interest. It can occur that prior on parameters lead to prior on observation out of relevance![17] This is a clear indication that something is wrong and must be changed. Also the comparison between priors and posteriors is needed to assess the level of involment of the data set in the posterior. In Figure 6 one can see that consistently with the used priors (Table 4) no effect is visible in the prior dimension, which is not the case for the posterior dimension where three distinct clusters appear. Probably, it would be logical to reduce the variability of the prior since negative yields are proposed for almost every μ_{ij} together with upper values also not sensible.

5.2 Looking for a sensible number of multiplicative terms

Figure 7 displays the joint posterior distributions of (λ_1, λ_2) for the reduced and complete cases. The reduced case is interesting because we can notice that the $S = 100$ simulations are distributed in two groups: the most numerous for negative values of λ_1, the other one for positive values of λ_2. In each group, λ_2 takes a continuum of values negative or positive. This is a clear indication that the credible region of λ_1 does not include the zero value, contrary to its of λ_2. The conclusion is that only one multiplicative is needed for the reduced case. For the complete case, a glance at the scales show that no doubt is left, both credible intervals are very far from zero and at least the first two dimensions have to be kept. As we didn't introduce a third term, nothing can be said about its necessity[18].

5.3 Interpreting the parameters

Let us concentrate our attention towards the genotypes; the same could be made for the environment factor.

The most straightforward use of the simulations is to draw scatter plots of a factor with the parameter depending on it. For instance[19] $\left(\alpha_i\sqrt{J}, \lambda_1\gamma_{i,1}\right)$, but instead of one point for genotypes, we will get S points providing a direct view of the variability/uncertainty: Bayesian statistics returns a distribution[20], not a point estimate with possibly a confidence interval. This is done in Figure 8 for the complete case and the first nine genotypes. They look impressively distinct. No overlapping for any pair of genotypes (this does not seem the consequence of the very reduced number of simulations: 100). Among these genotypes, the 6th looks attractive with the strongest main effect and limited interaction, then among the most stables.

[16] as suggested by Fred van Eeuwijk (Prof. at Wageningen University, NL) during a workshop on statistical interpretation of Genotype-Environment interaction, 2012, Inra, Jouy en Josas.

[17] Strongly negative or exagerated yields...

[18] Notice that in [Crossa.1991], three multiplicative terms have been introduced, we consider only two for the sake of the exposure simplicity. Nevertheless the model code provided in §E has got a `NQ` variable which can take any value. As underlined in §2.1.3, a flaw of Bayesian approaches is the lack of tools to select a model among a set of models. In case of nested models, credibility regions can be used to decide if some parameters take values associated to some submodels.

[19] The multiplication of α by $\sqrt{J}$ and of γ_q by λ_q provides them a squared norm equal to the sum of squares explained by the associated terms in the model which seems a fair way to relate them. More precisely $SS_{Geno} = J\sum_i \widehat{\alpha}_i^2$ and $SS_{G\times E(1)} = \lambda_1^2 = \lambda_1^2 \sum_i \widehat{\gamma_{i,1}}^2$, in the orthogonal case.

[20] Here with the means of simulations, then an empirical distribution.

Figure 6: Prior-posterior credible boxes for the μ_{ij} for the reduced data set. 2.5% and 97.5% quantiles were used to define the boxes around the median position. The red line is the first bissector.

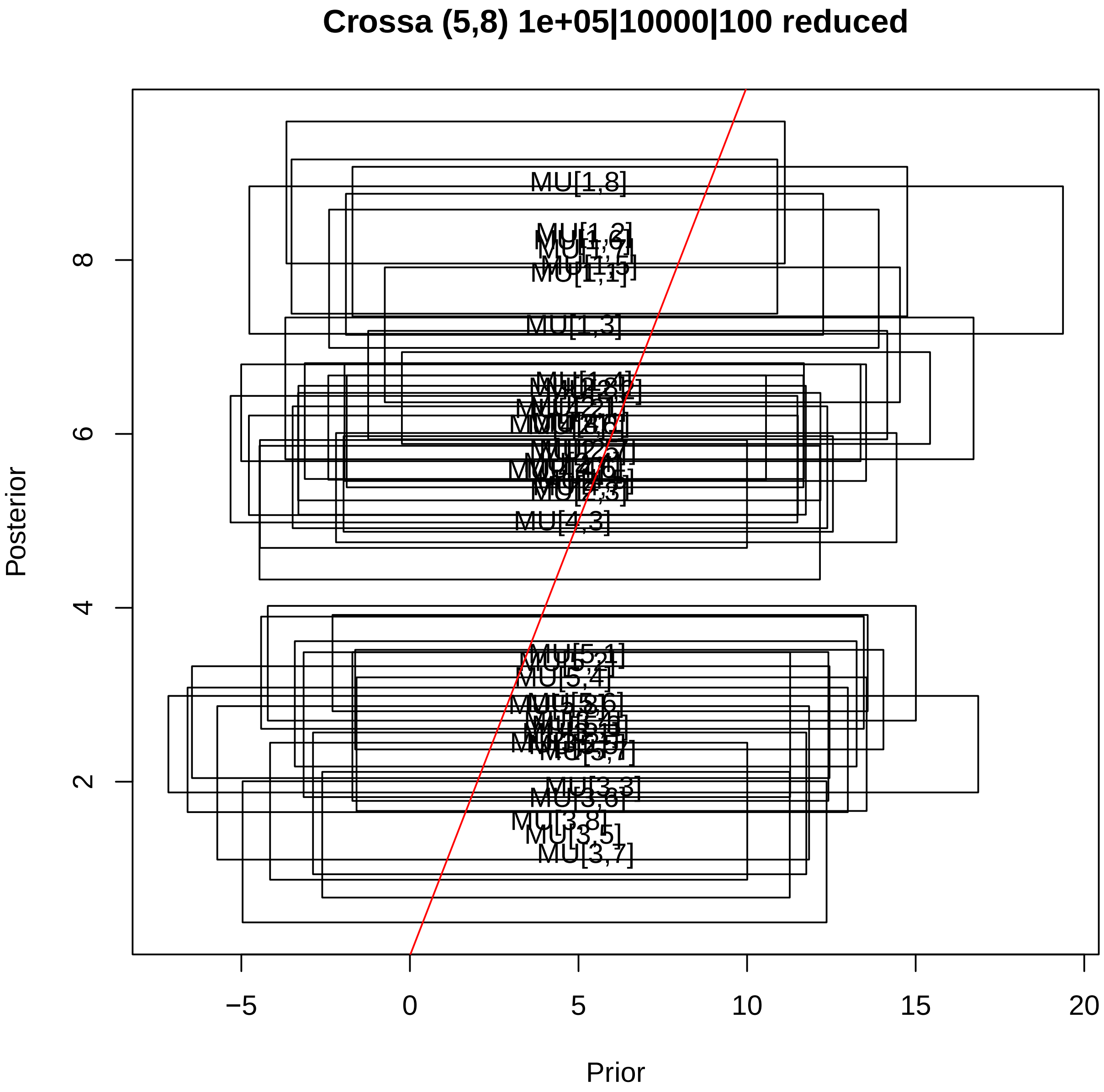
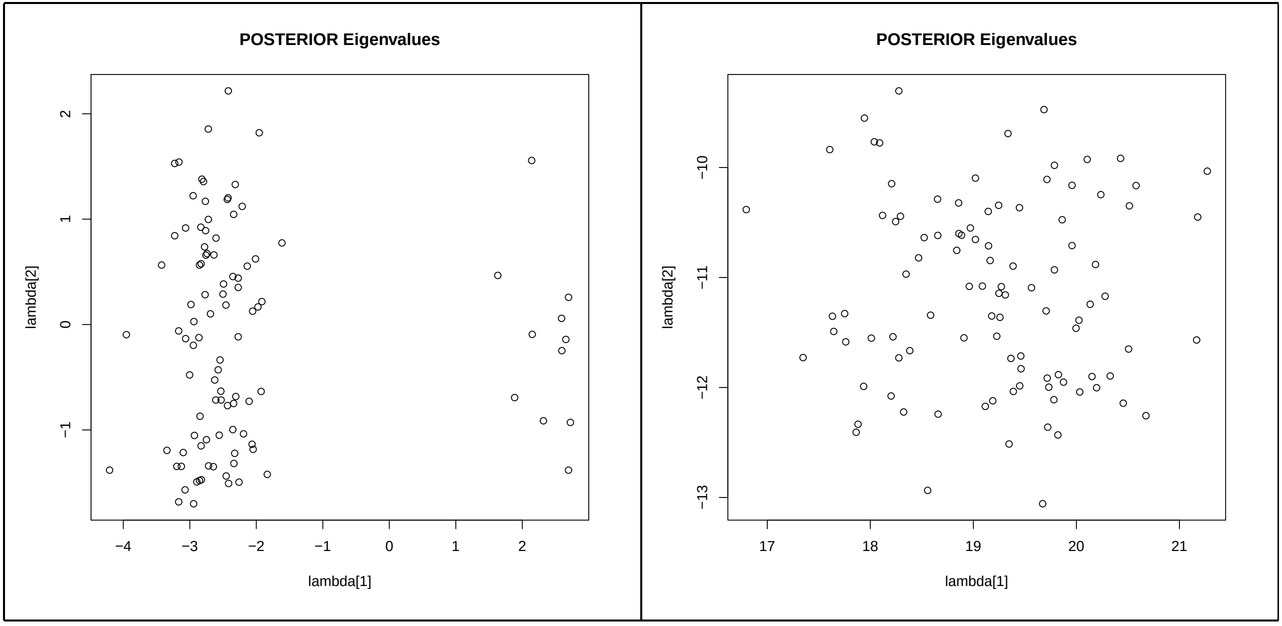


Figure 7: Posterior distributions of the λs for the reduced (left) and complete (right) cases.



Taking into account the easy interpretation given to the parameters associated to each genotype (since there is only one multiplicative term), it is possible to answer most of the questions risen at the beginning of this section. Among the nine displayed in Figure 8, genotype 6 has got the better performance across the set of all environments (**Q1**), it as also in the group of the stable genotypes {6,8,7} (**Q3**), certainly it deserves the first rank meanwhile genotype 3 is on the last position (**Q4**).

5.4 Interpreting the genotype responses across the environments

But, as the clear interpretation of the parameters identified in (2) is not always unambiguous, we suggest to always start from the matrix (μ_{ij}) defined in (3) and indicated in the DAG of Figure 1. The proposal is to show the genotype response for the spectrum of studied environments. For that, we think that a random curve for each genotype along the $E\left(\frac{1}{I}\sum_i \mu_{ij}\right)$ is of value.

To take into account the uncertainty linked with the posterior distribution, we propose to represent each genotype with a bundle of S curves (one for each simulation) obtained by joining the J points of coordinates $\left(\frac{1}{I}\sum_i \mu^s_{ij}, \mu_{ij} - \frac{1}{I}\sum_i \mu^s_{ij}\right)$. With the adopted constraints (§2.2.2), it is $\left(\mu^s + \beta^s_j, \alpha^s_i + \sum_{q=1}^{Q} \lambda^s_q \gamma^s_{iq} \delta^s_{js}\right)$. Other proposals can be think of, this one has the advantage of proposing non common terms between abscissae and ordinates, eliminating noise effects of systematic correlations.

This is proposed in Figures 9 and 10.

Indeed very strong differences appear and we retrieve the very good behavior of genotype 6. Using this diagram, we can also determine which are the best genotypes for a given environment (**Q2**). After genotype 6 which is definitively the best for all environments, the first one is the more advisable for poor environement necessarily placed on the left side of the abscissae.

Figure 8: Genotype effects (additive with α_i, interactive with $\gamma_{i,1}$) for the complete case. Each genotype is drawn in a different plot; only the first nine genotypes are shown (out of 18).

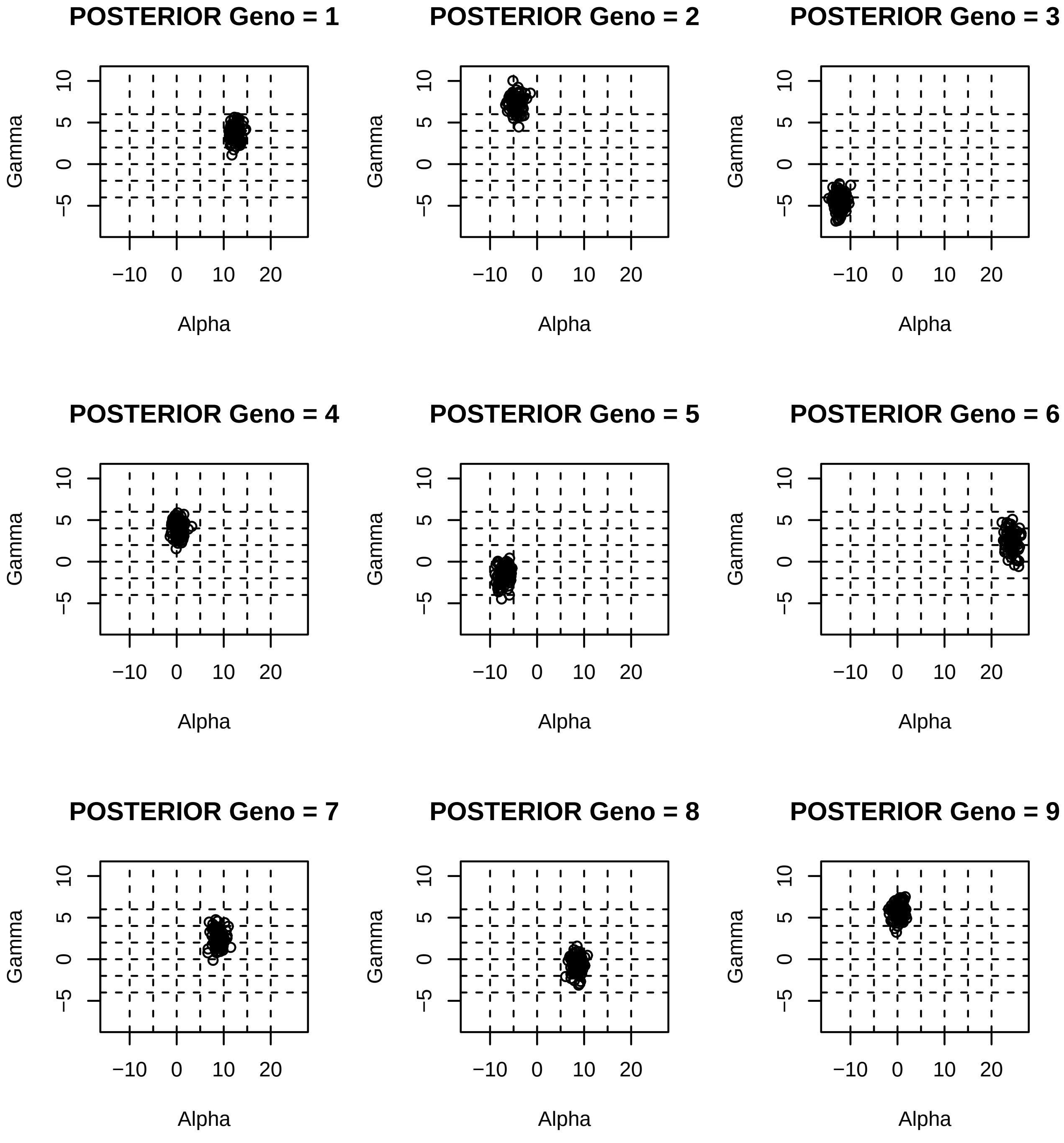


Figure 9: Posterior genotype profiles for the complete case (One to Nine)

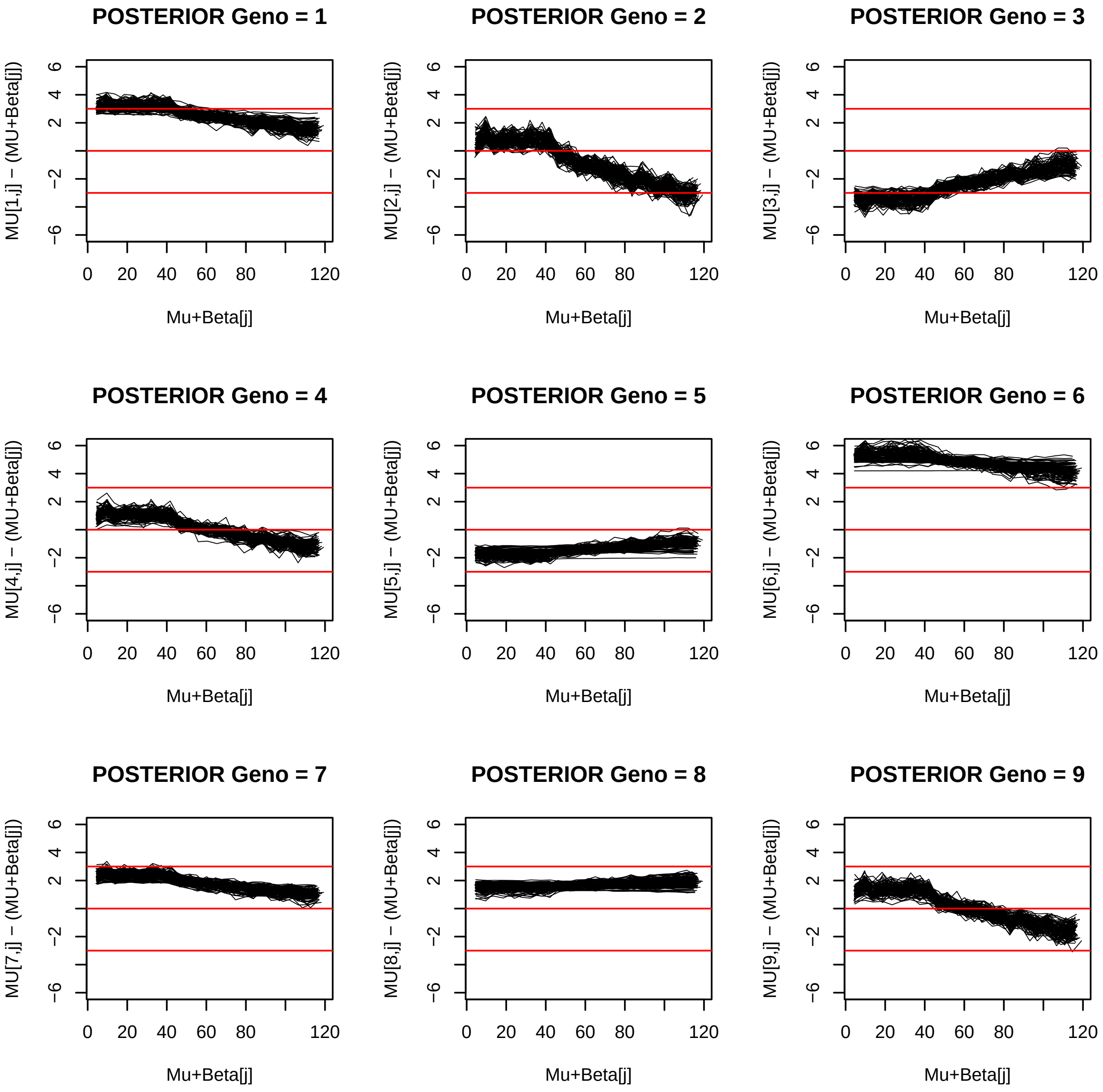


Figure 10: Posterior genotype profiles for the complete case (Ten to Eighteen)

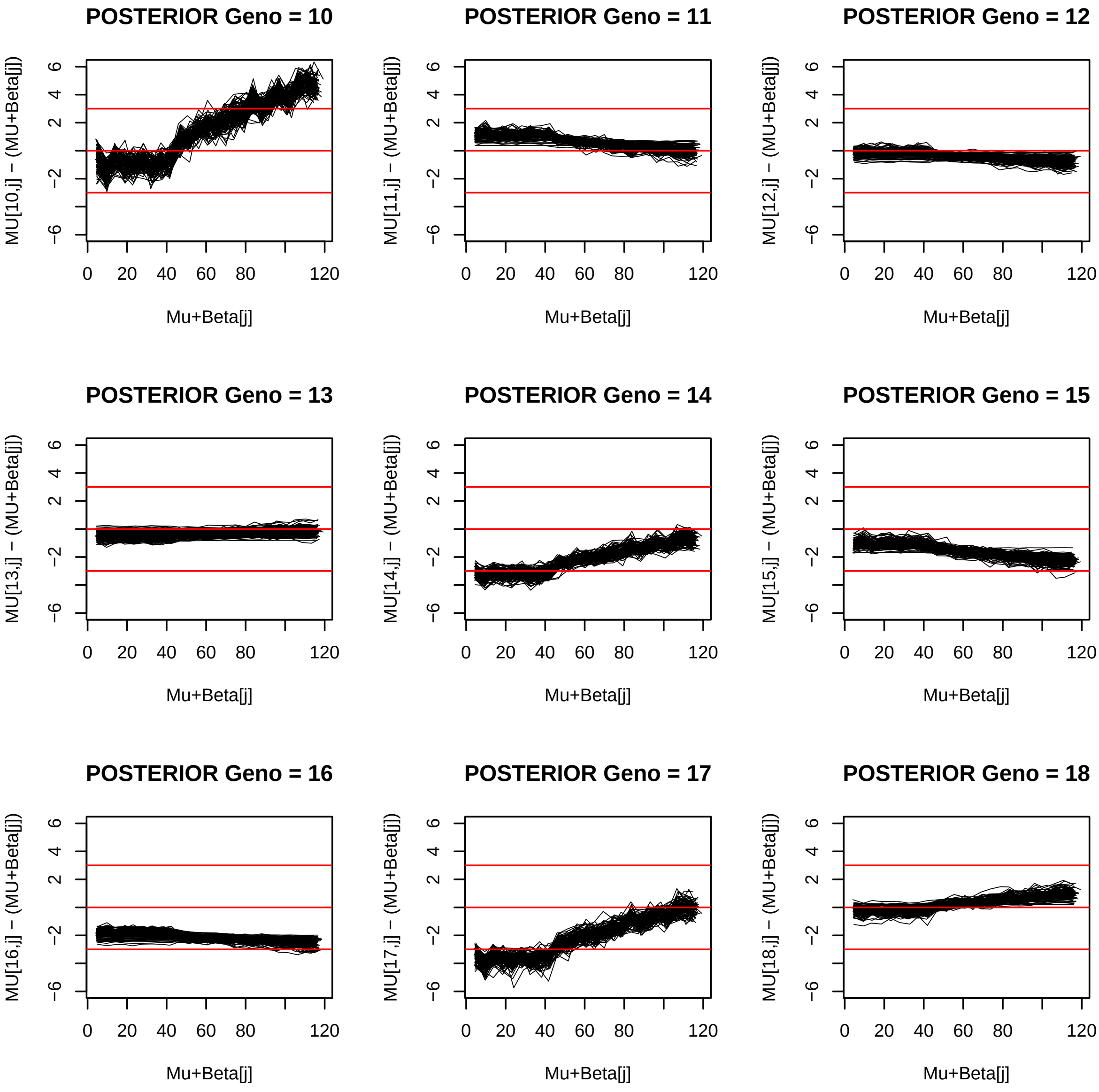
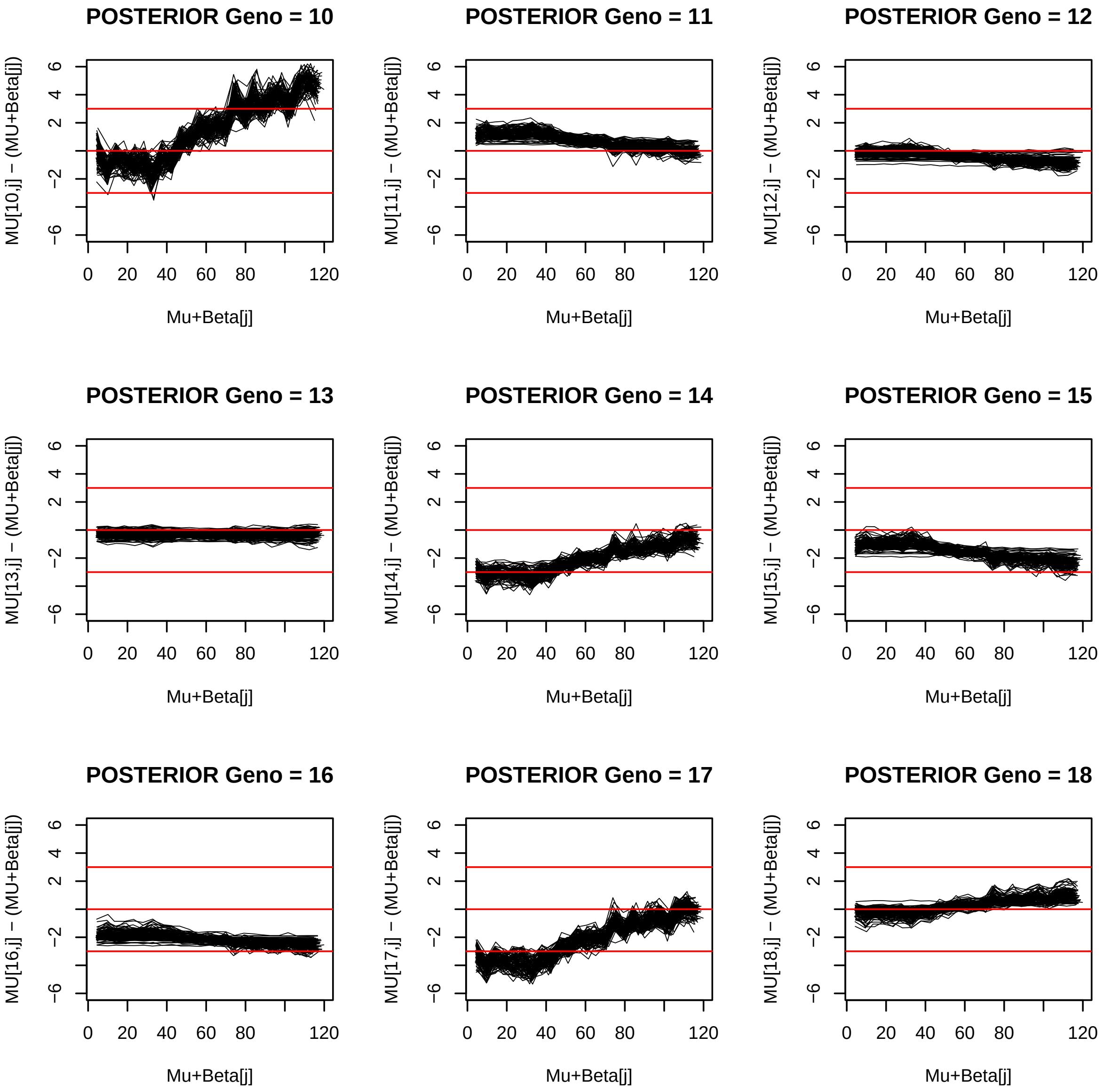


Figure 11: Posterior genotype profiles (Ten to Eighteen) in case of 45 missing values



5.5 Missing values

One of the strong points of the Bayesian approach is that missing values are no more a problem. Let's inspect the results obtained in such a situation. An intriguing fact appears in Table 3: the MCMC seems faster with missing values than without! In fact, this can be explained. The Bayesian statistical approach is no more than defining a joint probability distribution onto the whole set of parameters and data (prior and likelihood definitions) before conditioning with the observed data (posterior calculation). Missing values are not missing, they are not conditioning variable since not observed. One can then admit that more missing values there are, easier is the conditioning task!

The ten percent missing values does not alter the interpretation obtained from the complete data set, as it can be checked with Figure 11.

5.6 Exceeding a threshold

The previous diagrams does not allow to answer the last question (**Q5**): *could we assign a probability that a genotype will give more than a certain threshold?*, a quantitative question! Indeed, the bayesian framework is quite adapted to answer such a question!

First we have to be more precised about the question: is the yield for a given environment or for the range of all possible environments? If we are interested in a precise environment and that we have information about it, either because it was experimented or because we can predict its response through our modelling using pertinent covariables, then we can propose an answer to the question. But let us suppose that we are asking for a response across a set of environments which is not exactly represented by the experimental set. Let us also suppose that we are able to give relative weights (w_j) for each of our experimented environments to unbias the results, i.e. that

$$\pi_i = \sum_j w_j \mu_{ij}$$

is the performance of genotype i that we are looking for. One can see that if all w_j are zero except one of them, we will get the answer for a precised environment.

Then we can easily introduce this new variable into our model and be interested into the proportion of posterior simulations where the given threshold is exceeded, also it is of interest to look at the density of π_i, just to see where is the threshold. This was done for $w_{j=1,\dots,5} = \frac{1}{20}$, $w_{j=6,\dots,10} = \frac{1}{10}$, $w_{11} = \frac{1}{4}$ and $w_{j=12,\dots,25} = 0$ with a threshold of 8. Results are displayed in Figure 12 where it can be seen that only the first and sixth genotype can have a positive probability (respectively estimated to 0.1 and 1 by the posterior expectation of the proportions.)

5.7 Random versus fixed effects?

Genotype Environement data are often analyzed either by mixed models or by biadditive models (with a frequentist point of view). One of the most discussed point when defining such models is the choice between fixed or random status for the two factors. This is not so obvious. Indeed, when looking at the first question (**Q1**), the factor environement may be considered as random: we are not interested in one environement in particular and with the “subset” of environement we dispose, we want to tell what is the best genotype across all the possible environements (as done in §5.6). However, when regarding the second question, the factor environement must be considered as fixed: within this specific environment (and not another one) what is the performance of the genotypes? Using fixed effects model or random effects model is mainly the consequence of different questions[21] implying different models and then different results.

Within the Bayesian framework, there is no more unknown fixed parameters but only random variables and the inference could be denominated *the art of conditioning.* Consequently, we can pretend to keep with the general model and by some conditioning (for instance fixing the environment or not) adopt the right viewpoint to answer a precised question.

From a technical point of view, there are links between shrinking estimation, mixed models and Bayesian statistics: what is observed is not considered as unbiased truth.

[21] of course supposing that the data is pertinent to answer them.

Figure 12: Performances of the first nine genotype in a target set of environments

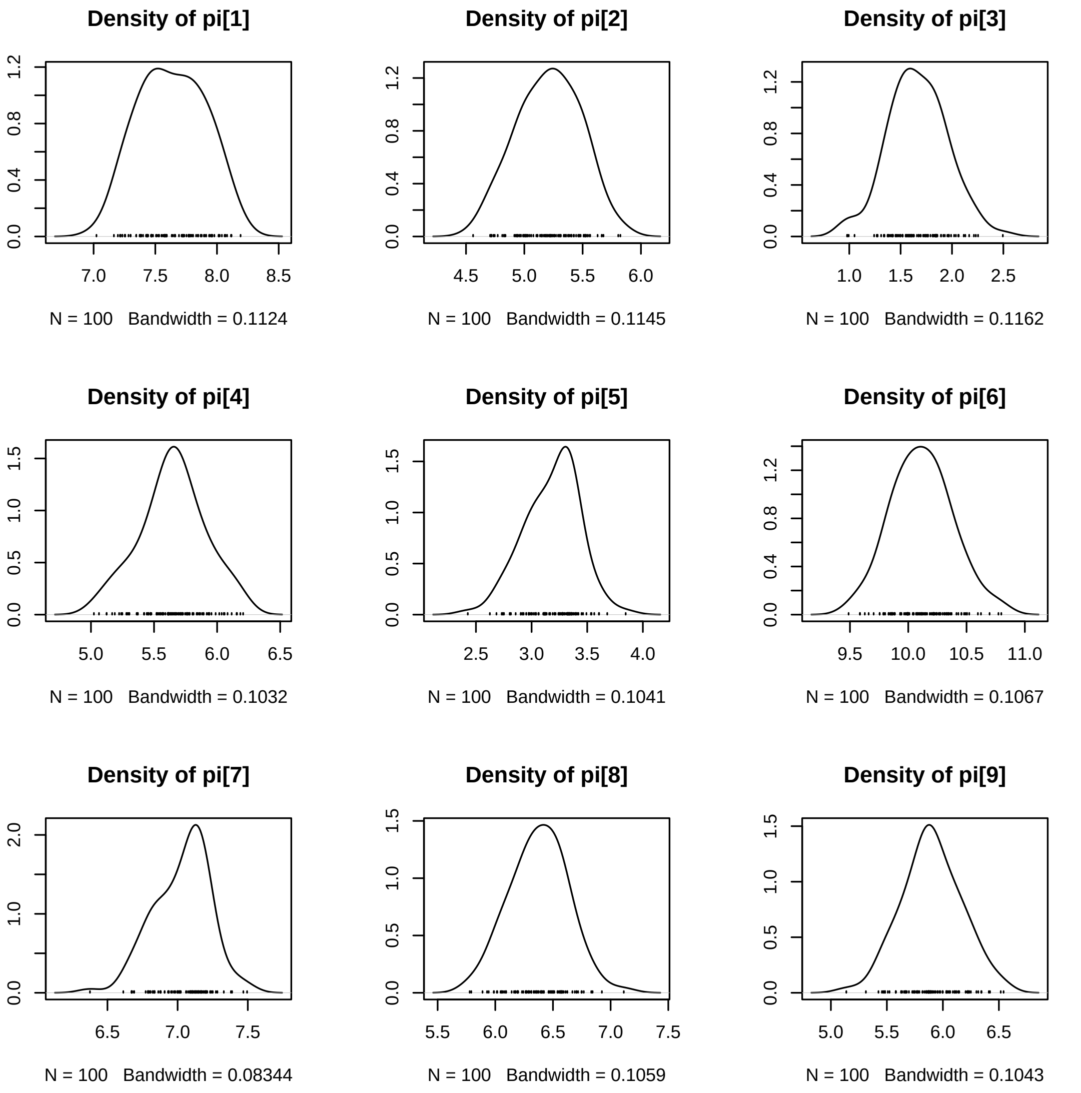
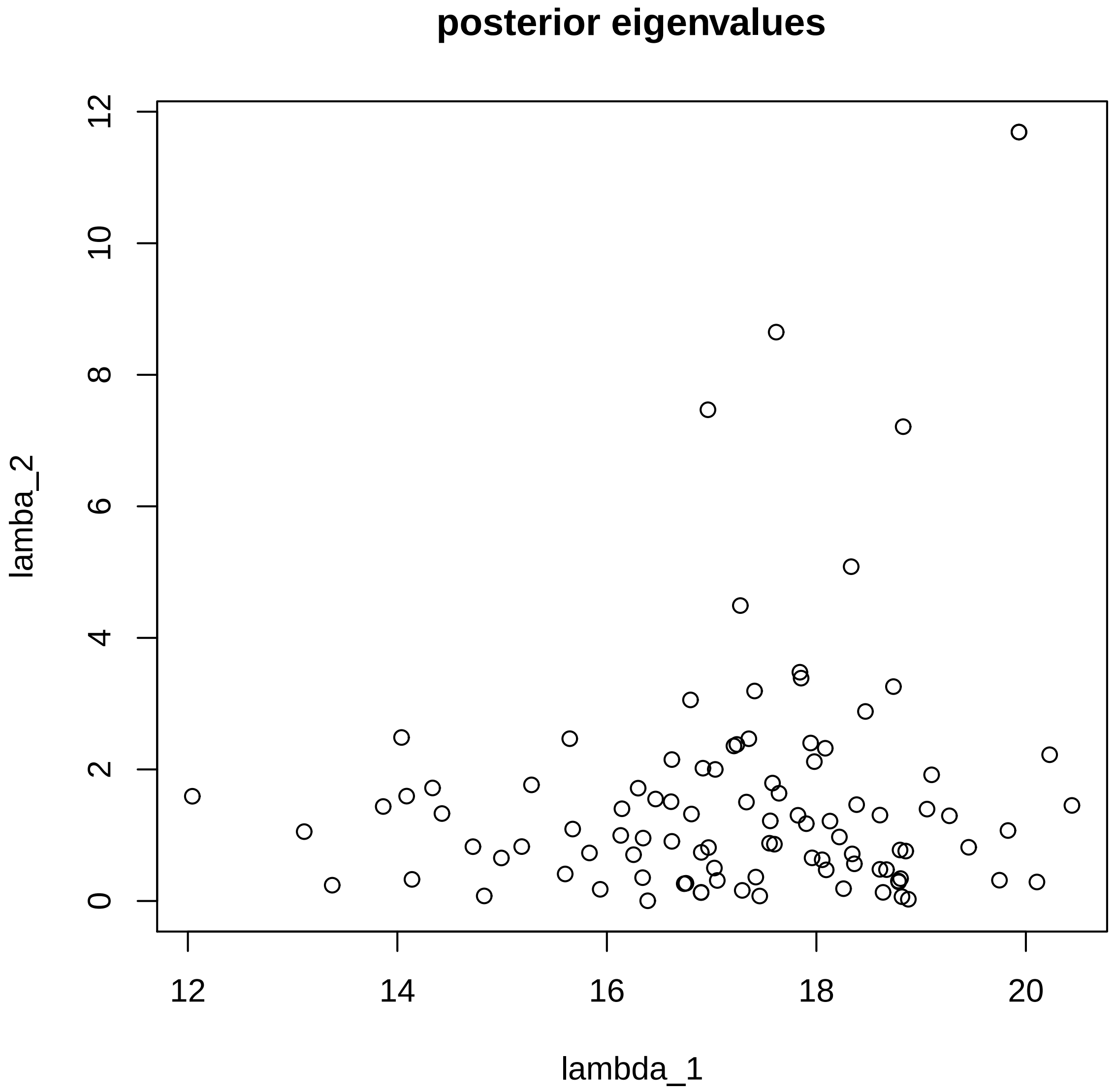


Figure 13: Similar Figure of Figure 7 with Perez et al.'s algorithm



5.8 Comparison with Perez et al. proposal

As explained in §2.3 other proposals already exist. In that section, we show the results obtained with the [Perez.2011] approach, using the R script of them.

For easier comparison the results obtained with [Perez.2011]'s algorithm are displayed with equivalent diagrams: Figures 13, 14, 15 and 16. Apart from the singular values, the results for the main effects ad the interactions seems very closed on this dataset. This is consistent with the fact that we used the same type of priors.

6 Conclusion

We have proposed an easy Bayesian treatment of the Biadditive model which seem to provide goods results based on a small simulation study and results comparable to the ones obtained

Figure 14: Similar Figure of Figure 8 with Perez et al.'s algorithm

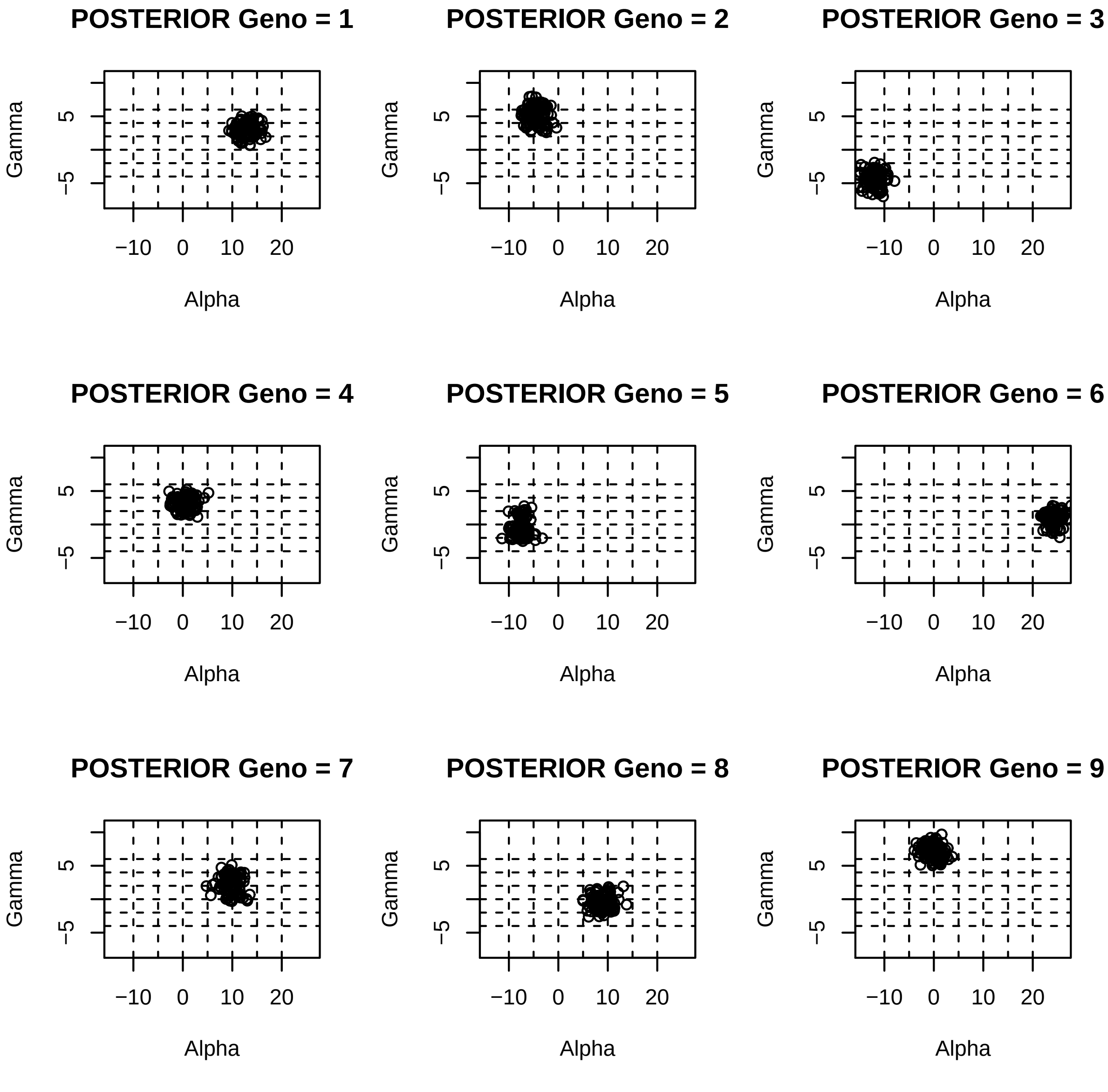


Figure 15: Similar Figure of Figure 9 with Perez et al.'s algorithm

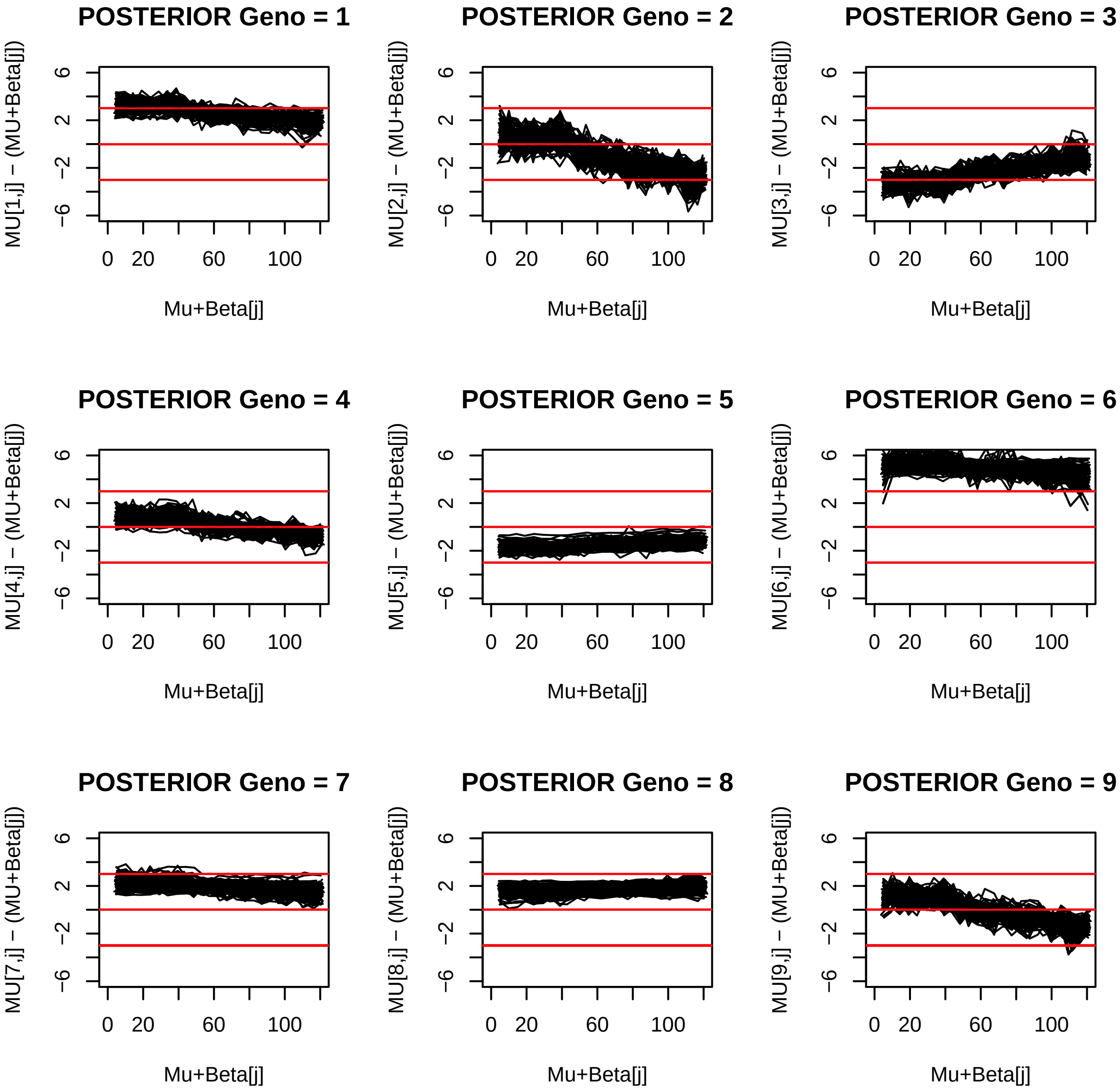
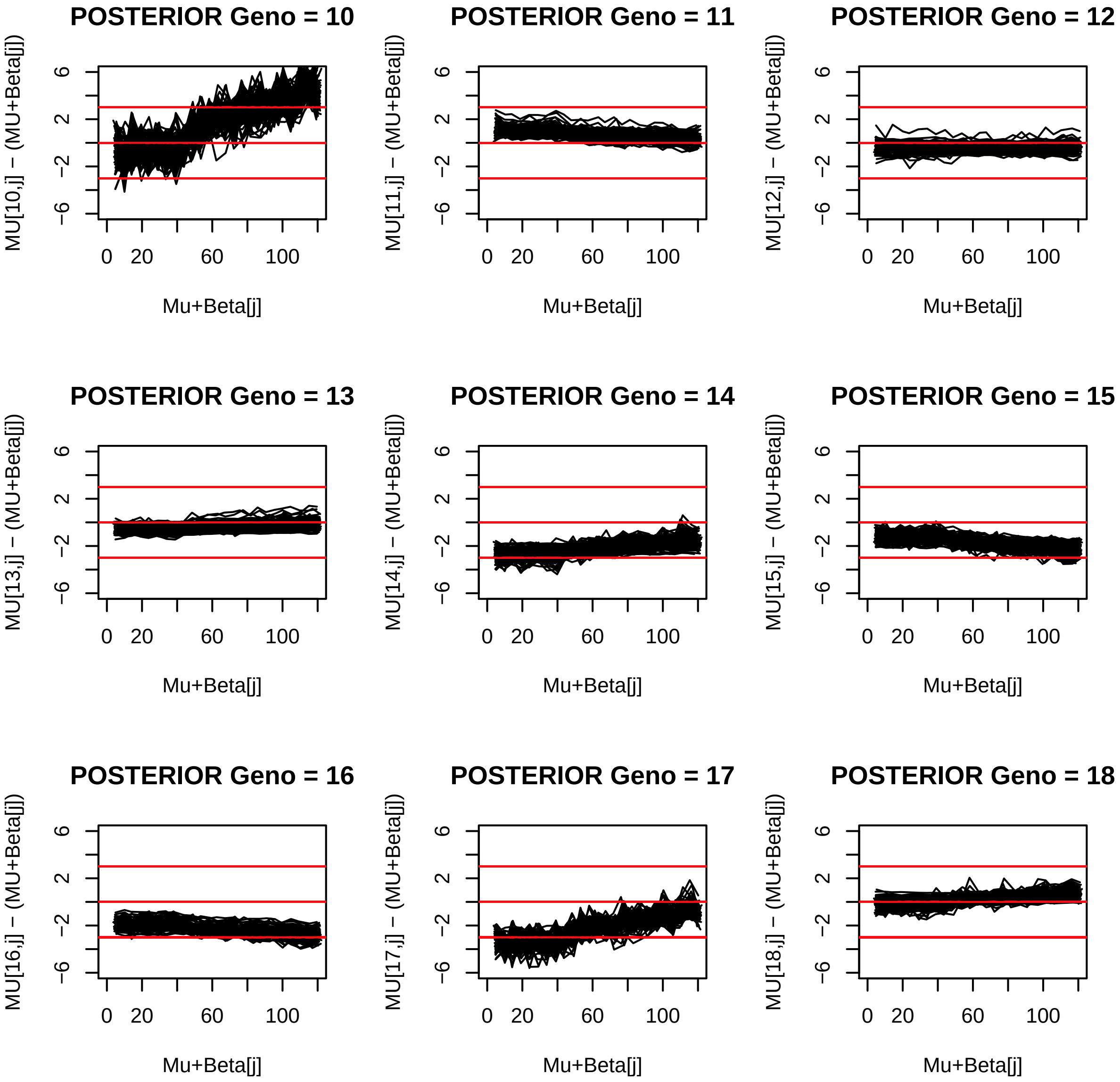


Figure 16: Similar Figure of Figure 10 with Perez et al.'s algorithm



by [Perez.2011] on a real dataset. The proposal solves the problem of overparametrization considering overparametrization as directly tractable with standard MCMC algorithms.

Regarding missing values, it would be interesting to compare the proposition to classical approaches where weighted least squares procedures are used to estimate the parameters from an incomplete dataset. One can argue that with the Bayesian point of view, we will obtain directly credible regions for the "estimated" data as well as for the factor effects which is very appealing.

Another point which deserve more research is the choice of the appropriate number of multiplicative terms which is a core issue in biadditive models. In our proposal, it is possible by inspecting the posterior distribution to decide how many terms to keep. More work has to be done using a simulation study for example in order to assess this procedure. Another path to investigate may be the use of "Bayesian Model Averaging", an empirical proposal to tackle models with different parametric dimensions in a Bayesian way. As the term suggests, perform as many Bayesian analyses as possible models and then take a synthetic posterior simply averaging each of them.

A Miscellaneous formulae

A.1 Conditional expectation and variance

It is well known that

$$\begin{aligned} E(Y) &= E_X\left(E_Y(Y \mid X)\right) \\ V(Y) &= E_X\left(V_Y(Y \mid X)\right) + V_X\left(E_Y(Y \mid X)\right) \end{aligned}$$

A.2 First two moments of E_{ijk}

When $E \mid \sigma_E \sim N(0, \sigma_E^2)$ and $\sigma_E \sim U(0, S_{ME})$, just applying formulae given in §A.1, one obtains

$$\begin{aligned} E(E) &= 0 \\ V(E) &= \frac{1}{3} S_{ME}^2 \end{aligned}$$

A.3 First two moments of a triple product

Let three independent variables A, B and C such that their expectations be respectively a, b, c and their standard deviation α, β, γ. Then

$$\begin{aligned} E(ABC) &= abc \\ V(ABC) &= \left(a^2 + \alpha^2\right)\left(b^2 + \beta^2\right)\left(c^2 + \gamma^2\right) - a^2 b^2 c^2 \end{aligned}$$

B von Mises-Fisher distribution

B.1 Definition

Following [De_Waal.2006], let X be a random matrix of size $I \times Q$ $(Q \leq I)$ be multinormally distributed with

$$\begin{aligned} E(X) &= \mu \\ V(\operatorname{vec}(X)) &= \Psi \otimes \Sigma \end{aligned}$$

where μ is any $I \times Q$ matrix, Σ is a $I \times I$ p.s.d. matrix, Ψ is a $Q \times Q$ p.s.d. matrix and the operator vec transforms a matrix into a vector by stacking its columns.

Then $X \mid X'X = s$ is distributed as a generalization of the von Mises-Fisher distribution with distribution given by the density

$$C \exp\left(\operatorname{tr}\left(\Sigma^{-1} x \Psi^{-1} \mu'\right) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1} x \Phi^{-1} x'\right)\right)$$
$$\text{when } x'x = s.$$

It is appealing that the constraint $x'x = s$ does not intervene in the density but for the definition of the support...

The **von Mises-Fisher distribution** occurs when $\Sigma = I_I, \Psi = s = I_Q$; its density is then defined to $\exp(\operatorname{tr}(x\mu'))$ for $x'x = I_Q$.

$$\begin{gathered} k \exp(\operatorname{tr}(x\mu')) \\ \text{when } x'x = I_Q. \end{gathered}$$

The normalizing constant, k, can be expressed as an infinite series of Hayakawa polynomials.

B.2 Interpretation of the von Mises-Fisher distribution

Some interesting facts can be emphasized:

- The matrix μ is the only parameter of the distribution.
- Matrix μ is no more than the expectation of the initial matrix X not restricted to lie on the hyper-sphere unity but due to curved space where is the restricted variable, its expectation does not belong to the hypersphere but is inside it. Nevertheless, μ plays the role of a centrality parameter for the direction of the normalized X.
- When $I = 2$ ou 3, and $Q = 1$ one can get hints, imagining the density function of a $N(\mu, I_I)$ intersected with the circle or sphere of radius one. The maximum density will be in the direction of μ; indeed what indicates $\operatorname{tr}(x\mu') = x'\mu$ since x and μ are vectors. The same reasonning applies when $I > 3$.
- Indeed for $I = 2$ and $Q = 1$, a simple reparameterization gives more insights. We can write $x' = (\cos(\theta), \sin(\theta))$ and $\mu = \lambda(\cos(\nu), \sin(\nu))$ then the density is proportional to $\exp(\lambda \cos(\theta - \nu))$ and can be easily drawn.

- When $I = 2$ and $Q = 2$, due to the orthogonality constraint, still the matrix x depends only of θ and reads
$$x = \left(\begin{array}{cc} \cos(\theta) & \cos\left(\theta + \frac{\pi}{2}\right) \\ \sin(\theta) & \sin\left(\theta + \frac{\pi}{2}\right) \end{array} \right)$$
so with obvious notation the density is proportional to
$$\exp\left(\lambda_1 \cos(\theta - \nu_1) + \lambda_2 \cos\left(\theta + \frac{\pi}{2} - \nu_2\right)\right)$$
which also can be easily represented with isocontour diagrams.

- When $I = 3$, similar considerations can be obtained with a parameterization with Q parameters by means of the Euler angles, for $Q = 3$, it reads
$$x = \left(\begin{array}{ccc} c_1 c_2 & -s_1 c_2 c_3 + s_2 s_3 & s_1 c_2 s_3 + s_2 c_3 \\ s_1 & c_1 c_3 & c_1 \\ -c_1 s_2 & s_1 s_2 c_3 + c_2 s_3 & -s_1 s_2 s_3 + c_2 c_3 \end{array} \right)$$
where $c_i = \cos(\theta_i)$ and $s_i = \sin(\theta_i)$.

- When $I > 3$, it is not that easy but, in principle, Euler angles can still be obtained by successive rotations around the canonical axis.

- The uniform distribution onto the hypersphere is obtained for $\mathrm{tr}(x'\mu) = 0$, which implies that $\mu = 0_{I \times Q}$.

- When μ is of rank $R < Q$, then it seems that R out of the Q vectors of x will be close to the directions comprised into μ and that the other ones will be let free on the orthogonal complement subspace? This rises the question if one cannot impose μ to be defined with orthogonal columns?

- It is clear that μ and $k\mu$ gives the same density, can other type of invariance be exhibited to allow a easier definition of the parameter μ?

C Prior/posterior with overparameterized models

C.1 Linear tiny case

C.1.1 Definition

In order to see the ideas and to be able to get all algebraical derivations, let us consider first a single case of overparameterization with two data values and three parameters to infer. Let be $Y = (Y_1, Y_2)$ and its expectation defined with three parameters $(\mu, \alpha_1, \alpha_2)$. The supposed independent priors are

$$\begin{array}{rcl} \mu & \sim & N(m, 1) \\ \alpha_1 & \sim & N(0, 1) \\ \alpha_2 & \sim & N(0, 1) \end{array}$$

(where m is a known numerical value) and the likelihood is given by

$$\left(\begin{array}{c} Y_1 \\ Y_2 \end{array} \middle| \mu, \alpha_1, \alpha_2 \right) \sim N\left(\left(\begin{array}{c} \mu+\alpha_1 \\ \mu+\alpha_2 \end{array} \right), \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right) \right)$$

Some easy computations due to the multinormal framework gives as joint distribution

$$\left(\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ Y_1 \\ Y_2 \end{array} \right) \sim N \left(\left(\begin{array}{c} m \\ 0 \\ 0 \\ m \\ m \end{array} \right), \left(\begin{array}{ccccc} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 3 & 1 \\ 1 & 0 & 1 & 1 & 3 \end{array} \right) \right).$$

C.1.2 Reparameterization

In order to distinguish what is reachable (or not) by the data, we can modify the parameter definition in the following linear way

$$\left(\begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \end{array} \right) = \left(\begin{array}{ccc} 2 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & -1 \end{array} \right) \left(\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \end{array} \right).$$

The transformation matrix being full rank, the two parameterizations are equivalent, in the sense that we can use one or the other, keeping the same modelling. One can notice that the first two θs can be expressed from the conditional expectation of Y:

$$\begin{array}{rcl} \theta_1 & = & E\left(Y_1 \mid (\mu, \alpha_1, \alpha_2)\right) + E\left(Y_2 \mid (\mu, \alpha_1, \alpha_2)\right) \\ \theta_2 & = & E\left(Y_1 \mid (\mu, \alpha_1, \alpha_2)\right) - E\left(Y_2 \mid (\mu, \alpha_1, \alpha_2)\right) \end{array}$$

and that there is no hope to express the last new parameter in such a way, since θ_3 is independent from (θ_1, θ_2), the new joint distribution reading

$$\left(\begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \\ Y_1 \\ Y_2 \end{array} \right) \sim N \left(\left(\begin{array}{c} 2m \\ 0 \\ 0 \\ m \\ m \end{array} \right), \left(\begin{array}{ccccc} 6 & 0 & 0 & 1 & 1 \\ 0 & 2 & 0 & 1 & -1 \\ 0 & 0 & 3 & 0 & 0 \\ 3 & 1 & 0 & 3 & 1 \\ 3 & -1 & 0 & 1 & 3 \end{array} \right) \right). \tag{6}$$

As a consequence, the new Bayesian reformulation of the model evidences the useless of parameter θ_3, indeed the new parameters are independent and

$$\begin{array}{rcl} \theta_1 & \sim & N(2m, 6) \\ \theta_2 & \sim & N(0, 2) \\ \theta_3 & \sim & N(0, 3) \end{array}$$

also the likelihood given by

$$\left(\begin{array}{c} Y_1 \\ Y_2 \end{array} \right) \mid (\theta_1, \theta_2, \theta_3) \sim N\left(\left(\begin{array}{c} \frac{\theta_1+\theta_2}{2} \\ \frac{\theta_1-\theta_2}{2} \end{array} \right), \left(\begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array} \right) \right),$$

does not depend on θ_3.

A noticeable point is that the two Ys are no more conditionally independent to the parameters. But why not? In fact, it is possible to linearly transform the data(Y_1, Y_2) in an equivalent pair (Z_1, Z_2) such the the Zs be conditionally independent of the parameters.

C.1.3 Posterior in the new parameterization

From (6), the posterior can be written applying the conditional formula of the multinormal distribution:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \mid (Y_1, Y_2) \sim N\left(\begin{pmatrix} \frac{Y_1+Y_2}{4} + \frac{3}{2}m \\ \frac{Y_1-Y_2}{2} - m \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}\right).$$

C.1.4 Conclusion

If the statistical analysis is made with the second parameterization, it is observed that $[\theta_3 \mid (Y_1, Y_2)]$ is no more than the prior $[\theta_3]$. This parameter can be eliminated and the overparameterization is no more present.

C.2 Additive case

Similar results are easily found with the additive model following the same steps that in §C.1.

C.2.1 Definition

$$E\left(Y_{ij} \mid \mu, \alpha_i, \beta_j\right) = \mu + \alpha_i + \beta_j$$

with prior[22] and likelihood:

$$\begin{array}{rcl} \mu & \sim & N(0,1) \\ \alpha_i & \sim & N(0,1) \;\; i = 1, ..., I \\ \beta_i & \sim & N(0,1) \;\; j = 1, ..., J \\ Y_{ij} \mid (\mu, \alpha, \beta) & \sim & N(\mu + \alpha_i + \beta_j, 1) \;\; i = 1, ..., I \;;\; j = 1, ..., J. \end{array}$$

Here, we have $1 + I + J$ parameters whose priors are defined independent and IJ data independent conditionally to the parameters. But it is well known that the parametric dimension of such additive model is $I + J - 1$ so that we have to find 2 (here linear) functions of the parameters which are independent of the data, and such that no information will be added when going from the prior to the posterior.

C.2.2 First reparameterization

It would be much more difficult to get a set of $1 + I + J$ independent orthogonal parameters $\theta_1, \theta_2, ..., \theta_{1+I+J}$ such that the last two be independent of the Y_{ij} as done in the tiny case. But we propose the following, let

$$\theta_{ij} = \mu + \alpha_i + \beta_j \text{ when } i = 1 \text{ and/or } j = 1.$$

The number of θs is $I + J - 1$, and

$$\mu + \alpha_i + \beta_j = \theta_{1j} + \theta_{i1} - \theta_{11}$$

[22] Assuming that $m = 0$ and all variances unity.

whatever is (i, j) showing that the θs generates all expectations of the data set. It now suffice to get two additional linear combinations of the (μ, α_i, β_j) independent from them. Following the indication given in §C.1.2, a possibility is

$$\begin{aligned} \rho_1 &= \mu - \sum_i \alpha_i \\ \rho_2 &= \mu - \sum_j \beta_j \end{aligned}$$

To better see what is behind the transformation, let us detail the case for $I = 3$ and $J = 4$. The new parameters are obtained from the basic ones from the following linear mapping

$$\begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \theta_{14} \\ \theta_{21} \\ \theta_{31} \\ \rho_1 \\ \rho_2 \end{pmatrix} = P_A \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

where

$$P_A = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{pmatrix}$$

giving

$$\begin{aligned} Var\begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \theta_{14} \\ \theta_{21} \\ \theta_{31} \\ \rho_1 \\ \rho_2 \end{pmatrix} &= P_A \times P_A' \\ &= \begin{pmatrix} 3 & 2 & 2 & 2 & 2 & 2 & 0 & 0 \\ 2 & 3 & 2 & 2 & 1 & 1 & 0 & 0 \\ 2 & 2 & 3 & 2 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 3 & 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 1 & 3 & 2 & 0 & 0 \\ 2 & 1 & 1 & 1 & 2 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 5 \end{pmatrix} \end{aligned}$$

C.2.3 Second reparameterization

Just to show that many reparameterizations are possible, again for $I = 3$ and $J = 4$. The new parameters are obtained from the basic ones from the following linear mapping

$$P_B \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 17 & 4 & 4 & 4 & 3 & 3 & 3 & 3 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & -4 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & -1 & -1 & -1 \end{pmatrix} \times \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

giving

$$P_B \times \quad P_B' = \begin{pmatrix} 373 & 21 & 21 & 20 & 20 & 20 & 0 & 0 \\ 21 & 2 & 1 & 1 & 1 & 1 & 0 & 0 \\ 21 & 1 & 2 & 1 & 1 & 1 & 0 & 0 \\ 20 & 1 & 1 & 2 & 1 & 1 & 0 & 0 \\ 20 & 1 & 1 & 1 & 2 & 1 & 0 & 0 \\ 20 & 1 & 1 & 1 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \end{pmatrix}$$

C.3 General formulation

As the derivation in §C.2.2 is done, one can think that it is mainly based on the linear relationship between the parameters and the expections of the data: in fact not. It is only based on the prior multinormality of the parameters. Let us now try to get the essence of the overparameterization behaviour.

C.3.1 Definition of overparametrization

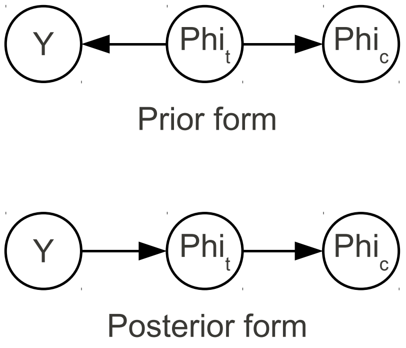
Let a model be defined with θ, the set of parameters and some data Y. Let $[\theta]$ be the prior and $[Y \mid \theta]$ the likelihood probability distributions. We will say that there is overparameterization when not all the parameters are necessary to define the likelihood, that is when it exists $\phi_t(\theta)$ a reparameterization with a smaller parametric dimension[23], equivalently defining the likelihood. More precisely:

$$\begin{aligned} [Y \mid \theta] &= [Y \mid \phi_t] \\ \mathrm{pd}(\phi_t) &< \mathrm{pd}(\theta). \end{aligned} \tag{7}$$

Intuitively, there is sufficient information into ϕ_t to provide the complete determination of the likelihood.

[23] We are not at ease of how define the parametric dimension of ϕ_t, we suspect that it is linked with the rank of the Jacobian $\frac{\partial \phi_t}{\partial \theta}$.

Figure 17: Prior and posterior forms of the joint distribution as DAGs



C.3.2 Dealing with overparameterization

There is at least two ways to prevent the overparameterization difficulty: (i) use a transformation of type ϕ_t with minimal parametric dimension restricting the parameter space; (ii) add additionnal constraints onto the initial parameter sets $\phi_c(\theta) = 0$ such $\theta \mid_{\phi_c(\theta)=0}$ be equivalent to some minimal ϕ_t. The first solution eliminates the difficulty but most of the time loosing some interesting symetrical properties. The interpretation of the parameters is then less easy and interesting, so most often the second way is preferred.

Assuming that for a given model, we are able to find $(\phi_t(\theta), \phi_c(\theta))$ such that ϕ_t is sufficient to define the likelihood and has got the minimum parametric dimension. So we have (7) and

$$(\theta) \Longleftrightarrow (\phi_t, \phi_c). \tag{8}$$

Then the initial model can be written into this new equivalent parameterization:

$$\begin{aligned} \text{prior} &: \quad [\phi_t, \phi_c] = [\phi_t][\phi_c \mid \phi_t], \\ \text{likelihood} &: \quad [Y \mid \phi_t]. \end{aligned}$$

The joint distribution is then the product of both:

$$[Y \mid \phi_t][\phi_t][\phi_c \mid \phi_t] = [Y][\phi_t \mid Y][\phi_c \mid \phi_t]. \tag{9}$$

Both forms are given in the two three-nodes Bayesian networks of Figure 17. The right hand side of Equation (9) gives the posterior form while the left one is the prior form. From (9), we obtain that the posterior distribution of the parameters as

$$[(\phi_t,\phi_c)|Y] = [\phi_t \mid Y][\phi_c \mid \phi_t]. \tag{10}$$

This does not mean that the posterior of ϕ_c does not depend on Y but that it depends on Y only through ϕ_t and more that this dependence is identical to the one defined at the prior level. It is the consequence of the d-separation operated by ϕ_t between the random variables Y and ϕ_c.

C.3.3 Conclusion

Use of an overparameterized model with a prior onto the complete set of parameters is without consequence. One is dealing with more random variables than necessary, no more. And experimental observations tends to think that it is a more efficient way, possibly due to a symmetrical repartition of the roles between the parameters.

When the prior on the complete set can be decomposed into independent (ϕ_c) and necessary (ϕ_t) transformed subsets, then the prior and the posterior of ϕ_c are identical. This is convenient but not required.

C.4 Biadditive case

Of course this applies to the biadditive case. Wihtout entering into a general treatment if we consider the model $B\left(*,*,*,\pi\right)$ provided with the following prior and likelihood:

$$\begin{aligned}
\lambda &\sim N\left(0,1\right) \\
\gamma_i &\sim N\left(0,1\right)\;\; i=1,...,I \\
\delta_j &\sim N\left(0,1\right)\;\; j=1,...,J \\
Y_{ij} \mid \left(\lambda,\gamma,\delta\right) &\sim N\left(\lambda\gamma_i\delta_j,1\right)\;\; i=1,...,I\;;\; j=1,...,J.
\end{aligned}$$

One can notice that the complete set of parameters is identical to the additive case. As the prior are similar and as necessary transformed subsets are identical (isomorphism between additive form and multiplicative form), it is possible to get the same kind of independant parameters as we obtained in §C.2.2 and §C.2.3. The two conditions to check are:

1. equivalent priors,
2. equivalence of the identifiability in the likelihood function.

This has to be generalized for any type of biadditive models, but we can wonder the utility (in term of data interpretation) of such results since we are interested into the active parameters not in the redundant ones.

D Gibbs sampling algorithms used in the literature proposals

D.1 Brief reminder about Gibbs sampling

When it is not possible to compute directly the posterior $[\theta \mid Y]$ most of the time the difficulty can be break down in smaller pieces, this is the Gibbs sampling technique. The parameter vector is partioned into convenient subsets $\theta = \{\theta_1, \theta_2, ..., \theta_P\}$ and an iterative scheme is applied by cycling over the P subsets to get draws into the global posterior.

$$\begin{gathered}
[\theta_1 \mid Y, \{\theta - \theta_1\}] \\
[\theta_2 \mid Y, \{\theta - \theta_2\}] \\
\qquad\qquad \cdots \\
[\theta_P \mid Y, \{\theta - \theta_P\}]
\end{gathered}$$

The only requirement is to be able to draw into the partial conditional distributions, in general a much easier task. Of course there are some drawbacks, mainly the non independence of the draws, and the safety of drawing into the posterior only asymptotically, then a burn-in phase of the algorithm is necessary.

D.2 Hoff's proposal

Hoff [Hoff.2009, Hoff.2012] is interested in multivariate data analysis techniques associated to the model:

$$Y_{I\times J} = UDV' + E,$$

where $U_{I\times Q}$ and $V_{J\times Q}$ are orthonormal matrices, D a diagonal matrix and E a matrix with independent components $E = \{\varepsilon_{i,j} \sim \mathcal{N}(0,\sigma^2)\}$. This model can be seen as an underlying model for Principal Component Analysis (PCA) and can be found under different names in the literature such as the fixed effect model [Caussinus.1986], the fixed factor scores model, etc.

Hoff [Hoff.2009, Hoff.2012] proposed a Bayesian treatment of this model. With such a treatment, he obtained posterior expectation of the singular values that are closer to the "true" ones compared to the maximum likelihood estimates; he also proposed a new way to select the number of dimensions. More precisely, his proposition is the following one.

The likelihood is given by:

$$\begin{aligned} L(Y,U,D,V,\sigma^2) &\propto \exp\left\{\frac{-1}{2\sigma^2}||Y-UDV'||^2\right\}, \\ & \exp\{\frac{-1}{2\sigma^2}(tr((Y-UDV')(Y-UDV')'))\}, \end{aligned}$$

$$\exp\{\frac{-1}{2\sigma^2}(tr(YY'-YVDU'-UDV'Y'+UDV'VDU'))\}.$$

The priors for the singular vectors U and V are defined as uniform on the Stiefel manifold. Uniform distributions on this manifold correspond to very simple cases of von Mises-Fisher distributions (cf. §B). The explicit form of these uniform distributions are given in [Smidl.2007].

The priors for the singular values as well as for the noise variance parameter are respectively:

$$\begin{aligned} \{d_1,..,d_Q\} &\sim \mathcal{N}(0,\tau^2) \quad , \\ 1/\tau^2 &\sim \text{gamma}(\eta_0/2, \eta_0\tau_0^2/2), \\ 1/\sigma^2 &\sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2). \end{aligned}$$

Then the model depends on the numerical constants ν_0, η_0, σ_0.

The joint posterior distribution for the parameters is:

$$f(U,D,V,\sigma^2|Y) \quad \propto \quad f(Y|U,D,V,\sigma^2) \times f(U,D,V,\sigma^2)$$

To make inferences on the quantities of interest, we need to consider the marginal posterior distributions. Since no close form is available for the joint posterior distribution for the parameters, a Gibbs sampler, which iteratively simulates each parameter from its full conditional distribution, is built.

Since the prior distribution for U is uniform, the conditional posterior distribution for U is completely determined by the likelihood (or the conditional likelihood):

$$\begin{aligned} f(U|V,D,Y,\sigma^2) & \propto & L(U,D,V,\sigma^2) \\ & & \exp\{\mathrm{tr}((YVD/\sigma^2)'U)\} \end{aligned}$$

This distribution is a von Mises-Fisher distribution denoted $MF(YVD/\sigma^2)$. Similarly, the condition posterior distribution for V is a $MF(Y'UD/\sigma^2)$. The conditional distributions for $d_q, 1/\tau^2$ and $1/\sigma^2$ follow respectively a normal, a gamma and a gamma distribution, with parameters given in [Hoff.2012]. Initial values are taken as the maximum likelihood estimates of the model (corresponding to the usual least squares estimates). On a simulated example, [Hoff.2012] showed that the posterior mean of $E\,(Y_{ij})$ is closer to true population values than the estimation obtained by maximum likelihood.

Hoff's method requires to simulate matrices from von Mises-Fisher distributions. To do so, he proposed a method [Hoff.2009]which is implemented in the **R** [RCRAN.2012] package `rstiefel`.

D.3 Perez et al.'s proposal

The method proposed by [Perez.2011] dedicated to linear-bilinear models also uses von Mises-Fisher distributions but in a different way. We let the linear terms out for the sake of simplicity. They have the same likelihood and conditional distributions for matrices U and V as defined in the previous section (§D.2), with an additional n corresponding to a constant number of replicates for each combination of genotype-interaction and τ instead of $1/\sigma^2$, to be in accordance with their notations. From these conditional likelihoods, [Perez.2011] define its priors as:

$$\pi(U|\tau) \propto \exp\{\mathrm{tr}(\tau n_0 Y_0 V_0 D_0 U')\},$$

with Y_0 define as the prior cell averages such that $Y_0 = U_0 D_0 V_0'$. They do the same thing for V. These priors are consequently distributed according to a von Mises-Fisher distribution. The prior for τ is a gamma distribution :

$$\pi(\tau) \sim \mathrm{gamma}(a/2, s_0^2/2)$$

With these priors, they only need to express their beliefs in the prior cell average Y_0, then U_0, V_0, D_0 follows from the SVD of Y_0, as well as their beliefs about a and s_0^2 . In order to draw samples from the marginal posteriors distribution a Gibbs sampler is also built. The conditional posterior distribution for U is:

$$\begin{aligned} \pi(U|V,D,Y,\tau) & \propto & f(U|V,D,Y,\tau) * \pi(U|\tau) \\ & & \exp\{\mathrm{tr}(\tau * (n_0 Y_0 V_0 D_0 + nYVD)U')\} \end{aligned}$$

This distribution is also a von Mises-Fisher distribution. The other posterior distributions are given in the paper [Perez.2011]. Their algorithm is available as an **R** function but the code is written for cases with replicates (it allows one to obtain an estimate for the variance σ^2); however, it is possible to easily adapt the code for cases without repetitions.

E BUGS coding of the model

Here is the BUGS coding of the model we used for the worked example of §5.

```
# 12_03_07
#
# A biadditive model with NQ multiplicative terms
# made as general as possible to get numerical
# example in the note written with Julie.
#
# constant to be defined are:
#   NI (number of genotypes)
#   NJ (number of environments)
#   NQ (number of multiplicative terms)
#    m.mu     (mu expectation)
#   sd.mu     (mu     standard deviation)
#   sd.alpha  (alpha  standard deviation)
#   sd.beta   (beta   standard deviation)
#   sd.lambda (lambda standard deviation)
#   sdm.E (maximum value of sigma.E)
#   w (weights for the targetted environments)
#   limit (threshold of interest)
#
model {
  #
  # prior on the additive part
  Mu ~ dnorm(m.mu,1/sd.mu^2);
  for (i in 1:NI) { alpha[i] ~ dnorm(0,1/sd.alpha^2);}
  for (j in 1:NJ) {  beta[j] ~ dnorm(0,1/sd.beta^2 );}
  #
  # prior on singular values
  for (q in 1:NQ) {
    lambda0[q] ~ dnorm(0,1/sd.lambda^2);
  }
  lambda[1:NQ] <- sort(lambda0);
  #
  # prior on row singular vectors
  for (q in 1:NQ) {
    gamma[1,q] ~ dnorm(0,1)T(0,);
    for (i in 2:NI) {
      gamma[i,q] ~ dnorm(0,1);
    }
  }
  # prior on column singular vectors
  for (q in 1:NQ) {
    delta[1,q] ~ dnorm(0,1)T(0,);
    for (j in 2:NJ) {
      delta[j,q] ~ dnorm(0,1);
    }
  }
  #
  # getting the expectation
  for (q in 1:NQ) {
    for (i in 1:NI) {
      gamma0[i,q] <- gamma[i,q] * lambda[q];
    }
  }
  INTE <- gamma0 %*% t(delta);
  for (i in 1:NI) { for (j in 1:NJ) {
    MU[i,j] <- Mu + alpha[i] +
               beta[j] + INTE[i,j];
  }}
  #
  # data variance
  sigma.E ~ dunif(0,sdm.E);
  #
  # likelihood of the data set
  for (i in 1:NI) { for (j in 1:NJ) {
    Y[i,j] ~ dnorm(MU[i,j],1/sigma.E^2);
    Mup[i,j] <- MU[i,j] * w[j];
  }}
  #
  # performance of the genotype in a
  # different set of environments
  for (i in 1:NI) {
    pi[i] <- sum(Mup[i,]);
    pr[i] <- step(pi[i] - limit);
  }
}
```

References

[Arminger.1998] G. Arminger. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. Psychometrika, 63(3):271-300, 1998.

[Bingham.1974] Ch. Bingham. An antipodally symmetric distribution on the sphere. The annals of statistics, 2(6): 1201-1225, 1974.

[Bishop.1999] Ch. M. Bishop. Bayesian PCA. in Advances in Neural Information Processing Systems (11)382:388, 1999.

[Caussinus.1986] H. Caussinus. Models and uses of principal component analysis (with discussions). In Multidimensional Data Analysis. Ed. J. De Leeuw, W. J. Heiser, J. J. Meulman & F. Critchley. 149-178pp. DSWO Press. 1986.

[Chikuse.2002] Y. Chikuse. Statistics on Special Manifolds. Springer, 2002.

[Crossa.1991] J. Crossa, P.N. Fox, W.H. Pfeiffer, S. Rajaram and H. G. Gauch, Jr.. AMMI adjustment for statistical analysis of an international wheat yield trial. Theor. Appl. Genet. , 81:27-37, 1991.

[Crossa.2011] J. Crossa, S. Perez-Elizalde, D. Jarquin, J. Miguel Cotes, K. Viele, G. Liu and P. L. Cornelius. Bayesian estimation of the additive main effects and multiplicative interaction model. Crop Science, 51:14581469, 2011.

[Denis.1992] J.-B. Denis and J. Gower. Biadditive models. Technical Report of the laboratoire de Biométrie, Inra-Versailles, 33pp, 1992.

[Denis.1994] .J.-B. Denis and J. Gower. Asymptotic covariances for the parameters of biadditive models. Utilitas Mathematica, 46: 193-205, 1994.

[Denis.1996] J.-B. Denis and J. C. Gower. Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. Applied Statistics, 45(4): 479-493, 1996.

[Denis.1998] J.-B. Denis. biareg, Splus functions to perform biadditive regressions. Technical Report of the laboratoire de Biométrie, Inra-Versailles, 29pp, 1998.

[De_Waal.2006] *D. J. De Waal.* The von Mises-Fisher matrix distribution (in the article about Matrix-Valued Distribution). *Encyclopedia of Statistical Sciences (2d ed., 7:4618), 2006.*

[Edwards.2006] J. W. Edwards. Bayesian modeling of heterogeneous error and genotype x environment interaction variances. Crop Science, 46:820-833, 2006.

[van_Eeuwijk.1996] F. A. van Eeuwijk, J.-B. Denis, and M. S. Kang. Genotype-by-environment interaction. CRC press. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables, pp. 15-49, 1996.

[Fisher.1923] R. A. Fisher and W. A. Mackenzie. Studies in crop variation II. The manurial response of different potato varieties. Journal of Agricultural Science XIII, 311-20, 1923.

[Freitas.2004] H. Freitas Lopes and M. West. Bayesian assessment in factor analysis. Statistica Sinica, 14: 41-67, 2004.

[Hoff.2012] P. H. Hoff. Bayesian analysis of matrix data with rstiefel. Technical Report, Departments of Statistics and Biostatistics, University of Washington, Seattle, 12pp, 2012.

[Hoff.2009] P. H. Hoff. Simulation of the matrix Bingham - von Mises - Fisher distribution, with applications to multivariate and relational data. Journal of Computational and Graphical Statistics, 18(2):438-456, 2009.

[Nounou.2002] M. N. Nounou, B. R. Bakshi, P. K. Goel and X. Shen. Bayesian Principal Component Analysis. The Ohio State University, Technical Report, 50pp, 2002.

[Perez.2011] S. Perez-Elizalde, D. Jarquin and J. Crossa. A general Bayesian estimation method of linear-bilinear models applied to plant breeding trials with genotype x environment interaction. Journal of Agricultural, Biological and Environmental Statistics, 17(1):15-37, 2011.

[Plummer.2011] M. Plummer. Jags version 3.1.0, user manual. Technical Report. sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/. 39pp. 2011.

[RCRAN.2012] R Development Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012.

[Smidl.2007] V. Smidl and A. Quinn. On Bayesian principal component analysis, Computational Statistics and Data Analysis, 51: 4101-4123, 2007.

[Tipping.1999] M. E. Tipping and Ch. M. Bishop. Probabilistic principal component analysis. J.R. Statist. Soc. B, 61(3):611-622, 1999.

[Viele.2000] K. Viele and C. Srinivasan. Parsimonious estimation of multiplicative interaction in analysis of variance using Kullback-Leibler information. Journal of Statistical Planning and Inference. 84:201-219, 2000.