Bayesian analysis of factorial experiments by mixture modelling

Agostino Nobile and Peter J. Green*
University of Bristol, UK.

November 1997

Summary

A Bayesian analysis for factorial experiments is presented, using finite mixture distributions to model the main effects and interactions. This allows both estimation and an analogue of hypothesis testing in a posterior analysis using a single prior specification. A detailed formulation based on this approach is provided for the case of the two-way model with replication, allowing interactions. Issues in formulating a suitable prior are discussed in detail, and, in the context of three illustrative applications, we discuss implementation, presentation of posterior distributions, sensitivity, and performance of the MCMC methods that are used.

Some key words: Analysis of Variance, Bayes linear model, Finite mixture distributions, Identifiability, Markov chain Monte Carlo, Partial exchangeability, Random partitions, Reversible jump, Sensitivity analysis.

1 Introduction

Faster computers and the increasing popularity of MCMC methods have allowed Bayesian statistics to become the tool of choice for a growing number of statistical practitioners wrestling with complex data analysis problems. Curiously, however, in respect of one of the areas of methodology most ubiquitous and centrally important in applications – namely the analysis of factorial experiments – the Bayesian approach has yet to provide a completely satisfactory answer. In this paper we try our hands at it.

One version of the classical theory of factorial experiments, going back to Fisher and further developed by Kempthorne (1955), completely avoids distributional assumptions, assuming only additivity, and uses randomisation to derive the standard tests of hypotheses about treatment effects. Here, we are interested in the more familiar classical approach via linear modelling and normal distribution theory. The corresponding Bayesian analysis has been developed mainly in the pioneering works of Box and Tiao (1973) and Lindley and Smith (1972). Box and Tiao (1973, Chapter 6) discuss Bayesian analysis of cross classified designs, including fixed, random and mixed effects models. They point out that in a Bayesian approach the distinction between fixed and random effects lies in "the nature of the prior distribution used to represent the behavior of the factors". They also show (Chapter 7) that shrinkage estimates of specific effects may result when

*Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: Ago.Nobile@bristol.ac.uk and P.J.Green@bristol.ac.uk.

a random effects model is assumed. Lindley and Smith (1972) use a hierarchically structured linear model built on multivariate normal components (special cases of the model are considered by Lindley, 1972 and Smith, 1973), with the focus on estimation of treatment effects. These are authoritative and attractive approaches, albeit with modest compromises to the Bayesian paradigm – in respect of the estimation of the variance components – necessitated by the computational limitations of the time. Nevertheless, the inference is almost entirely estimative: questions about the indistinguishability of factor levels, or more general hypotheses about contrasts, are answered indirectly trough their joint posterior distribution, e.g. by checking whether the hypothesis falls in a highest posterior density region. In blunter terms, little attempt is made, the notable exception being Dickey (1974), at answering the question a Bayesian would be likely to ask: what is the probability of the hypothesis?

Schervish (1992) moves closer to this goal by addressing questions of the form "how far is some linear function of the parameters away from some specified value?", in a (non-hierarchical) Bayesian linear model. Again, continuous (natural conjugate) priors are used, and the inference is summarised by the posterior distribution of a scalar measure of discrepancy between the data and the linear hypothesis of interest. The form of this measure is influenced by pedagogical concerns, to maintain comparability with the frequentist approach. It is calibrated to be in physical units corresponding to, for example, differences between treatment effects, so that an investigator would be able to derive a posterior probability that such differences exceed a specified physically-meaningful threshold.

Gopalan and Berry (1996) advocate an approach to multiple comparisons that more fully builds in the discrete character of the hypothesis-testing problem (see also Groenewald, 1991). A partition of the parameter space is pre-defined as part of the specification of the prior, each cell of which corresponds to some pattern of ties among the parameters; modelling of this aspect of the prior is semi-automated by the use of Dirichlet process priors. This approach allows direct inference about the partition, and posterior probabilities for the cells can be computed by MCMC methods.

The estimative and partition-based approaches co-exist in the recent paper by Bush and MacEachern (1996) on Bayesian analysis of the randomised block experiment; Dirichlet process priors are used for the block effects, and ordinary normal priors for the treatments. The flexibility of the Dirichlet process model is used implicitly, effectively robustifying the analysis to departures of the block effects from normality, and inference is made, not about the corresponding partitions, but just in terms of estimates of treatment effects.

Against this background, we can now state the approach of the present paper. The traditional dichotomy between estimation and testing in Bayesian statistics has recently blurred considerably. This is largely due to the research on model mixing and model averaging, where priors originally devised for testing are employed to provide inferences, and related measures of uncertainty, that take into account model uncertainty, e.g. see Kass and Raftery (1995). Consequently, we are not terribly innovative in using, for a Bayesian analysis of factorial experiments, a single prior specification suitable for both estimation and testing. In its detailed formulation, this prior will necessarily incorporate the analyst's view about what numerical differences between levels are considered practically significant. In our approach this judgement determines the amount of variation within clusters of effects. Posterior probabilities can then be computed that any subset of effects belongs to the same cluster, while "model-averaging" estimates of the effects are also produced automatically. This is all made possible by the use of finite mixture models for factorial effects, through the analysis of their underlying latent allocation variables. We choose to use explicitly-specified mixtures of normals, with unknown numbers of components, building on Richardson and Green (1997), rather than adopting the more restrictive Dirichlet process models. Comparisons between these classes of model will be found in Green and Richardson (1998). Our approach bears some resemblance with the one used by Consonni and Veronese (1995) for binomial experiments. Recast in the present

context, their model would assume a prior distribution on the partitions of levels and, conditional on the partition, exchangeability of the levels within each partition subset. Our model also defines a prior on the partitions of the levels: it is the distribution induced by the prior on the allocation variables. However, the two models are distinct, for instance we believe that the finite mixture approach is better equipped to deal with a random effects situation.

This paper is restricted to the case of the two-way ("row-plus-column") model with replications (possibly unequal and/or missing), and allowing interactions, but is intended to be extendible in an obvious way to more complicated designs. The formulation of the model, using a hierarchical structure, is inherently modular, and various other adaptations, for example to include covariates, would be straightforward. Computations are all done by Markov chain Monte Carlo (MCMC), making use of reversible jump moves (Green, 1995) where it is necessary to jump between parameter subspaces of differing dimension, as happens here when the numbers of components in the distributions of row, column or interaction effects change. Apart from the modelling flexibility permitted by MCMC, this approach leaves us particularly free to explore interesting aspects of the joint posterior distribution.

The paper is structured as follows. In Section 2, we introduce notation and describe our mixture-model-based formulation in detail. As is intuitively expected, and confirmed by pilot experiments, there are interesting patterns of sensitivity to prior specification: in Section 3 we provide a set of guidelines for the choice of prior hyperparameters. Three illustrative applications are then described in detail, in Section 4, where we cover implementational issues, many aspects of the posterior analysis, and give brief information about sensitivity and about MCMC performance. Details of the MCMC moves used are deferred to the Appendix.

2 A Bayesian two-way random effects model

We consider a two-way layout model. Denote by y_{ijk} the k-th observation collected at the i-th level of factor 1 and the j-th level of factor 2. There are m levels for factor 1 and n levels for factor 2. The number of observations corresponding to the pair of factor levels (i, j) is r_{ij}. Sometimes we will refer to these observations as "the cell (i, j)". Each observation is modelled as the sum of a systematic component (consisting of overall level, main effects and interaction) and a normal error component. Both main effects and the interaction are assumed random and drawn from finite mixtures of normal distributions.

A detailed description of the model follows. For the sake of notational simplicity we prefer to contravene traditional usage and employ σ_{ij}, σ_t^α, etc. to denote *variances* rather than standard deviations. All distributions are tacitly assumed conditional on the higher order parameters, even though only some of these appear on the right hand side of the $\sim$ symbol. Quantities for which a distribution is not specified are fixed constants and need to be assigned prior to the analysis.

It is assumed that

$$y_{ijk} = \theta_{ij} + \epsilon_{ijk} \qquad i = 1, \ldots, m; \;\; j = 1, \ldots, n; \;\; k = 1, \ldots, r_{ij}.$$

The systematic component θ_{ij} is the sum of the overall level μ, the main effects α_i and β_j and the interaction γ_{ij}:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad \forall i, j. \tag{1}$$

The error term ϵ_{ijk} is normally distributed

$$\epsilon_{ijk} \overset{ind.}{\sim} N(0, \sigma_{ij}) \qquad k = 1, \ldots, r_{ij}; \;\; \forall i, j$$

with zero mean and variance σ_{ij} allowed to differ from cell to cell according to the model

$$\sigma_{ij} \overset{iid}{\sim} I\Gamma(a, b) \tag{2}$$

$$b \sim \Gamma(q, h). \tag{3}$$

The overall level μ has normal prior distribution

$$\mu \sim N(\eta, \sigma^{\mu}),$$

while the remaining terms in the systematic component (1) are assumed to proceed from finite mixtures of unknown numbers of normal component distributions:

$$\alpha_i \overset{iid}{\sim} \sum_{t=1}^{k^{\alpha}} w_t^{\alpha} N(\mu_t^{\alpha}, \sigma_t^{\alpha}), \tag{4}$$

$$\beta_j \overset{iid}{\sim} \sum_{s=1}^{k^{\beta}} w_s^{\beta} N(\mu_s^{\beta}, \sigma_s^{\beta}), \tag{5}$$

$$\gamma_{ij} \overset{iid}{\sim} \sum_{u=1}^{k^{\gamma}} w_u^{\gamma} N(\mu_u^{\gamma}, \sigma_u^{\gamma}). \tag{6}$$

We assume that the prior on $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is given by (4)–(6) after conditioning on the classical identifying constraints

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_j \gamma_{ij} = \sum_i \gamma_{ij} = 0. \tag{7}$$

Next we specify the distributions for the parameters in the mixtures (4)-(6). We only give these distributions for the mixture of the first factor main effects, since similar structures are assumed, with obvious changes, for the other main effect and for the interactions.

For the number of components k^{α} we employ discrete uniform prior distributions on the integers from 1 to some maximum values k_{max}^{α}. See Section 3 for further discussion of this point. The mixture weights follow a Dirichlet distribution:

$$\boldsymbol{w}^{\alpha} \sim Dir(d_1^{\alpha}, \ldots, d_{k^{\alpha}}^{\alpha}).$$

We employ inverse gamma distributions as priors on the component variances

$$\sigma_t^{\alpha} \overset{ind.}{\sim} I\Gamma(a_t^{\alpha}, b_t^{\alpha}) \tag{8}$$

and normal distributions for the component means

$$\mu_t^{\alpha} \overset{ind.}{\sim} N(\xi_t^{\alpha}, 1/\tau^{\alpha}) \tag{9}$$

The prior precision of the component means is assumed to have prior gamma distribution:

$$\tau^{\alpha} \sim \Gamma(a^{\tau\alpha}, b^{\tau\alpha}).$$

The hyperparameters d_t^{α}, a_t^{α}, b_t^{α}, ξ_t^{α} are allowed to be different across components (albeit constant with respect to k^{α}) to permit prior specifications incorporating substantial information distinguishing the components. However, typically one may want to provide a unique common value for each of them, thus making the mixture components exchangeable. In Section 3 we discuss a practicable strategy for the choice of the hyperparameters. One main feature of the approach therein illustrated consists in choosing values that correspond to very well separated mixture components. This is necessary, as we wish to regard factor levels from the same mixture component as "practically indistinguishable".

Other specifications for the parameters in (4)-(6) are feasible. We only mention briefly two that we have considered and that lead to only minor modifications of the procedures described in Section 3 and in the Appendix. Initially we assumed the τ's to be fixed constants, on a par with the ξ's. Experimentation with the models with fixed and random τ's convinced us that there is merit in the added flexibility provided by random τ's. An alternative to (8)-(9) consists in using the conjugate prior which replaces (9) with

$$\mu_t^\alpha \overset{ind.}{\sim} N(\xi_t^\alpha, \sigma_t^\alpha/\tau^\alpha),$$

thus linking the variability of the component means to the variability within the components. We preferred the independent prior (8)-(9) as it seemed more intuitively appealing.

The mixture assumption on main effects and interactions in (4)-(6) can be restated by introducing latent quantities $\boldsymbol{z}^\alpha$, $\boldsymbol{z}^\beta$, $\boldsymbol{z}^\gamma$ which indicate from which components in the mixtures the main effects and interaction proceed. Thus, e.g., $z_i^\alpha = t$ means that α_i, the i-th level of factor 1, has been drawn from the t-th component of the finite mixture (4). We call these latent quantities allocations. Notice that while $\boldsymbol{z}^\alpha$ and $\boldsymbol{z}^\beta$ are vectors, of dimension m and n respectively, $\boldsymbol{z}^\gamma$ is a $m \times n$ matrix. On introducing the allocations $\boldsymbol{z}^\alpha$, equation (4) can be restated as follows. Conditional on the mixture weights $\boldsymbol{w}^\alpha$, each component in the allocation vector $\boldsymbol{z}^\alpha$ is assumed to have independent multinomial distribution with

$$\Pr\{z_i^\alpha = t\} = w_t^\alpha.$$

Once we condition on the $\boldsymbol{z}^\alpha$'s, the distributions of the α's reduce to normal:

$$\alpha_i \overset{ind.}{\sim} N(\mu_{z_i^\alpha}^\alpha, \sigma_{z_i^\alpha}^\alpha).$$

Similar expressions hold for the mixtures in (5) and (6). Introducing the allocations greatly facilitates computations. More importantly, it illuminates the partial exchangeability structures on main effects and interactions embedded in the prior distribution (for discussions and references on partial exchangeability see, e.g., Bernardo and Smith, 1994 and Schervish, 1995). Specifically, each allocation vector $\boldsymbol{z}^\alpha$ induces a partition of the α's in subsets, with exchangeability holding within each subset. Positive prior probability is assigned to each allocation vector, including those corresponding to exchangeable levels and to independent levels, thus affording great modelling flexibility. Moreover, the posterior distribution of the allocations sheds light on which partition patterns receive more support from the data.

Sampling from the posterior distribution of all the parameters and allocations is performed as described in the Appendix. The sample can be used for various inferential purposes including:

(i) estimation of main effects and interactions;

(ii) determination of most probable partition patterns of the main effects and interactions;

(iii) estimation of variance components;

(iv) prediction of future observables.

Several illustrations are provided in Section 4, with special emphasis on points (i) and (ii) above.

2.1 Parameter identifiability

Since the data $\boldsymbol{y}$ depends on the parameter $(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ only through $\boldsymbol{\theta}$ and the map from $(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ to $\boldsymbol{\theta}$ is not one to one, $(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is not identified. For instance, adding the same constant to each α_i and subtracting it from μ leaves the θ_{ij}'s unchanged.

In principle, lack of identifiability in the likelihood poses no problem to a Bayesian analysis as long as the prior distribution is proper (Lindley, 1971 page 46, Lindley and Smith, 1972). However, MCMC sampling of the resulting posterior distribution faces slow convergence problems: on contours of constant likelihood the posterior distribution is proportional to the prior and, as sample size increases, it will tend to concentrate on a lower dimensional manifold. Gelfand et al. (1995) suggested a centring reparameterisation for nested random effects models, while Vines et al. (1996) proposed a reparameterisation by sweeping (based on the classical constraints) for multiple random effects models. Another possibility is improving mixing by incorporating in the sampler Metropolis-Hastings moves that allow for swift changes along contours of constant likelihood, for an example see Nobile (1997).

An alternative approach consists in including identifying constraints in the prior distribution. This is the approach usually followed for fixed effects (see, e.g., Schervish, 1995 page 488). However, it has also been used for random effects models (Smith, 1973) and it is the approach we follow in the present paper by conditioning on the classical constraints (7).

2.2 Other models

In the above model we have assumed prior independence between the allocations $\boldsymbol{z}^\alpha$ and $\boldsymbol{z}^\beta$ and $\boldsymbol{z}^\gamma$. In some contexts it may be preferable to entertain more structured models, having the property that any two cells with the same allocations for both factor levels must have the same allocation for the interaction:

$$z_i^\alpha = z_t^\alpha, z_j^\beta = z_s^\beta \Longrightarrow z_{ij}^\gamma = z_{ts}^\gamma. \tag{10}$$

The finest partition corresponding to $\boldsymbol{z}^\gamma$ (for short, the partition of $\boldsymbol{z}^\gamma$) that satisfies (10) is the product partition induced by $\boldsymbol{z}^\alpha$ and $\boldsymbol{z}^\beta$. Thus, at one extreme one can assume that $\boldsymbol{z}^\alpha$ and $\boldsymbol{z}^\beta$ completely determine the partition of $\boldsymbol{z}^\gamma$, i.e., that the product partition is the partition of $\boldsymbol{z}^\gamma$. In this model $k^\gamma = k^\alpha k^\beta$ and

$$z_i^\alpha = z_t^\alpha, z_j^\beta = z_s^\beta \Longleftrightarrow z_{ij}^\gamma = z_{ts}^\gamma. \tag{11}$$

Condition (11) imposes much structure: for instance, interactions all from one component are inconsistent with any grouping of levels of either factor. A weaker model allows elements of the product partition to be grouped together to form the partition of $\boldsymbol{z}^\gamma$. This model is characterised by $k^\gamma \leq k^\alpha k^\beta$ and (10). The procedures presented in Section 3 and in the Appendix could be modified to deal with the estimation of both models, using Metropolis-Hastings draws to sample simultaneously all the allocations $\boldsymbol{z}^\alpha$, $\boldsymbol{z}^\beta$ and $\boldsymbol{z}^\gamma$. However, we have not pursued very far this line of research and have preferred instead to use the more flexible specification with prior independent allocations. For the three examples reported in Section 4 we have computed the frequency of times that the allocations sampled from the posterior distribution satisfy the conditions (10) and (11). They are reported below:

Data set	M.C. sample size	(10) satisfied	(11) satisfied
Survival, $\Delta = 1$	100,000	77,475	31
Survival, $\Delta = 0.25$	100,000	95,317	0
Fabric	100,000	8,111	1
Potato	100,000	99,140	0

These numbers seem to confirm that, at least for the data sets and priors we have considered, condition (11) is too restrictive. The adequacy of the additional structure implied by condition (10) seems instead to vary across data sets. In summary, we feel that the safest alternative consists in assuming prior independent allocations.

Conditions (10) and (11) add more structure to the model we entertain in this paper. We conclude this section by mentioning one modification that goes in the opposite direction of reducing structure. Rather than assuming a mixture distribution on the factor levels and the interactions, one could do away with (1) and instead require the cell means θ_{ij} to follow a mixture of normals with an unknown number of components. This model is easier to implement and is more flexible than the one we entertain: for instance, in a 2×2 design, it allows to consider directly the hypothesis $\theta_{11} = \theta_{12} = \theta_{21} \neq \theta_{22}$ that instead requires a much more complicate formulation in terms of α's, β's and γ's. This added flexibility may well provide the easiest approach to modelling, especially in the cases where requiring (1) would result in complicated patterns of interaction. However, it is achieved by losing the linear structure imposed by (1), which has a powerful explanatory role when it is satisfied and the main factors are dominant.

3 Choosing the hyperparameters

Several hyperparameters need to be specified. If prior information concerning the mechanism generating the data is available, it should be used in helping specifying these constants. In particular, prior information distinguishing the components is accommodated by our model and ought to be used whenever available. In this section we provide a set of guidelines that can be applied, as stated, when no such information is available. Nevertheless, the resulting prior distribution is far from uninformative. In the first instance, the hyperparameters are chosen in a way to make well separated mixture components very likely, as this is the basis to considering levels from distinct components as practically different. Secondly, the prior distribution incorporates the experimenter's judgement about what constitutes a practically significant difference between levels. Finally, we also make some use (in our opinion minimal use) of the data, specifically in equations (12) and (20).

The prior distributions of k^α, k^β, k^γ can be chosen as having support on small ranges of integer values. We suggest respective supports $\{1,\dots,m\}$, $\{1,\dots,n\}$, $\{1,\dots,mn\}$. In the examples of Section 4 discrete uniform distributions are used, other choices are also feasible. We emphasise the following difference with respect to the usual mixture analysis. Since the number of factor levels m and n is relatively small, and they play a role analogous to the number of data points in a mixture analysis, the posterior distributions of the number of mixture components will resemble the prior distributions. As a consequence, we are much less interested in, say, the posterior of k^α than in the posterior distribution of the partitions $\boldsymbol{\pi}^\alpha$ of the α's induced by the allocations $\boldsymbol{z}^\alpha$.

The mixture weights are chosen to have uniform distribution on the appropriate simplexes:

$$d_t^\alpha = d_s^\beta = d_u^\gamma = 1 \qquad \forall\, t, s, u.$$

The prior on k^α, $\boldsymbol{w}^\alpha$ and z^α induces a prior distribution on the partitions $\boldsymbol{\pi}^\alpha$ of the α's; similarly for the partitions $\boldsymbol{\pi}^\beta$ and $\boldsymbol{\pi}^\gamma$. In the examples in Sections 4.1 and 4.2, with $m = 3$ and $n = 4$, the prior specification adopted yielded the following prior distributions on $\boldsymbol{\pi}^\alpha$ and $\boldsymbol{\pi}^\beta$:

$\boldsymbol{\pi}^\alpha$	Prior probability
111	0.6
112, 121, 211	0.1222
123	0.0333

$\boldsymbol{\pi}^\beta$	Prior probability
1111	0.4286
1112, 1121, 1211, 2111	0.0714
1122, 1212, 1221	0.0476
1123, 1213, 1231, 2113, 2131, 2311	0.0226
1234	0.0071

These distributions can be used to check the appropriateness of, and possibly revise, the prior on the k's and $\boldsymbol{w}$'s and to aid in assessing the corresponding posterior distributions.

Next we consider the hyperparameters governing the prior distribution of the overall level μ. The mean η can be set equal to zero, $\eta = 0$. A large enough prior spread for μ is achieved by setting σ^μ equal to the square of the largest cell mean times a constant, say 100:

$$\sigma^\mu = 100 \cdot \max_{i,j} y_{ij.}^2. \tag{12}$$

As for the prior locations of the mixture components means, we set them all equal to 0:

$$\xi_t^\alpha = \xi_s^\beta = \xi_u^\gamma = 0 \qquad \forall\, t, s, u.$$

Our recipe for the remaining hyperparameters is a little more involved, so we prefer to organise it in subsections.

3.1 Mixture component variance hyperparameters

The hyperparameters a_t^α, b_t^α, a_s^β, b_s^β, a_u^γ, a_u^γ control the prior distributions of the variances of the normal components in the mixtures (4), (5) and (6). Our discussion is in terms of a_t^α, b_t^α only. Similar considerations apply to the hyperparameters in the distributions of σ_s^β and σ_u^γ, with m replaced by n and mn respectively.

Since we want to interpret the allocation of two factor levels in the same mixture component as an indication that they do not substantially differ, it is essential that the components variances be small. How small depends on a substantive judgement about what differences we are willing to consider as negligible. Suppose these judgements can be phrased as follows: "the effects of two factor levels, say α_i and α_j, are considered as essentially identical if they differ by less than a specified amount Δ". Then the problem becomes that of determining a_t^α and b_t^α such that the distribution of σ_t^α assigns most of the probability to the set of variances that make draws from the same component very likely to be less than Δ apart. Suppose we require

$$p_0 \leq \Pr\{\max|\alpha_i - \alpha_j| \leq \Delta\} = \Pr\left\{R \leq \frac{\Delta}{\sqrt{\sigma_t^\alpha}}\right\} \tag{13}$$

where p_0 is close to 1, the max is taken over $i, j = 1, \ldots, m$ and R is the range of m independent standard normal variates. Asymptotically,

$$\Pr\left\{R \leq \frac{\Delta}{\sqrt{\sigma_t^\alpha}}\right\} = \Pr\left\{U \geq 2m\Phi\left(-\frac{\Delta}{2\sqrt{\sigma_t^\alpha}}\right)\right\} \tag{14}$$

where U is a random variable with distribution function $F(u) = 1 - uK_1(u)$ and $K_1(\cdot)$ is the modified Bessel function of second kind and order 1 (see Gumbel, 1958 pages 312-315 and Kendall and Stuart, 1963 vol. 1 pages 340-341 for the details). Therefore, the inequality (13) can be rewritten as

$$p_0 \leq uK_1(u) \tag{15}$$

with

$$u = 2m\Phi(-\Delta/2\sqrt{\sigma_t^\alpha}). \tag{16}$$

Since the left hand side of (14) is decreasing in σ_t^α, (15) is equivalent to

$$\sigma_t^\alpha \leq \sigma_0, \tag{17}$$

where

$$\sigma_0 = \frac{\Delta^2}{4}\left[\Phi^{-1}\left(\frac{u_0}{2m}\right)\right]^{-2}$$

and u_0 solves $p_0 = u_0 K_1(u_0)$. To determine σ_0, first find u_0 (e.g., if we choose $p_0 = 0.9$ the solution is $u_0 = 0.3399779$) then substitute u_0 in (16) and solve for σ_0.

We choose the shape parameter $a_t^\alpha = 3$, in order to have finite second moments for σ_t^α. To specify the scale parameter b_t^α, we require that (17) holds with preassigned probability p_1. This yields the equation:

$$\begin{aligned} p_1 &= \Pr\{\sigma_t^\alpha \le \sigma_0\} \\ &= \Pr\{b_t^\alpha/\sigma_t^\alpha \ge b_t^\alpha/\sigma_0\} \\ &= 1 - F(b_t^\alpha/\sigma_0), \end{aligned} \tag{18}$$

where F is the cdf of a $\Gamma(a_t^\alpha, 1)$ random variable. Solving (18) for b_t^α gives:

$$b_t^\alpha = \sigma_0 F^{-1}(1 - p_1).$$

In our examples we chose $p_0 = p_1 = 0.9$.

3.2 Mixture component mean hyperparameters

The τ's govern the variability in the distributions of the mixture component means. Since we wish to interpret levels from the same mixture component as essentially equivalent, we are only interested in the case of well separated components, where the variability of the means is large relative to the variability in each component. This consideration leads to linking the prior variability of the components means to that within the means, which has already been determined in Section 3.1.

To set the hyperparameters $a^{\tau\alpha}$, $b^{\tau\alpha}$, $a^{\tau\beta}$, $b^{\tau\beta}$, $a^{\tau\gamma}$, $b^{\tau\gamma}$ we also make use of other information, provided by the data. We require that the expected value of $1/\tau$ is at least as large as some proportion, say 10%, of the overall variability in the data. This device provides a safeguard against completely missing the right magnitude of the τ's. It goes contrary to the likelihood principle; however, we are confident that only a minimal part of the information in the data is used in setting the prior. We only discuss the selection of $a^{\tau\alpha}$, $b^{\tau\beta}$: with obvious changes the same procedure applies to the hyperparameters for the other main effect and the interactions.

In Section 3.1 we have chosen a_t^α and b_t^α so that with high probability $\sigma_t^\alpha \le \sigma_0$. To ensure well separated components we are then led to require that

$$\frac{1}{\tau^\alpha} > \sigma_0$$

hold with high probability, say p_2. Specifically, we demand that

$$\begin{aligned} p_2 &= \Pr\{1/\tau^\alpha > \sigma_0\} \\ &= \Pr\{b^{\tau\alpha}\tau^\alpha < b^{\tau\alpha}/\sigma_0\} \\ &= F(b^{\tau\alpha}/\sigma_0) \end{aligned} \tag{19}$$

where p_2 is very close to 1 (we chose $p_2 = 0.999$), F is the cdf of a $\Gamma(a^{\tau\alpha}, 1)$ random variable and $a^{\tau\alpha} = 3$ to ensure finite second moments. Solving (19) yields

$$b^{\tau\alpha} = \sigma_0 F^{-1}(p_2).$$

Combining this with the above mentioned safeguard device yields

$$b^{\tau\alpha} = \sigma_0 F^{-1}(p_2) \vee (a^{\tau\alpha} - 1)\frac{\hat{\sigma}}{10}, \tag{20}$$

where $\hat{\sigma}$ is an estimate of the overall variability in the data.

3.3 Cell variance hyperparameters

We suggest choosing a, q and h so that the distribution of σ_{ij} is proper with finite first and second moments and approximately centred at the expected value of $1/\tau^{\gamma}$, the prior variance of the means in the interaction mixture components. Because of our choice of hyperparameters in (20), this centring value is at least as large as 10% of the data variability.

For the sake of clarity, we rewrite (2)-(3) as follows:

$$\sigma_{ij} = \frac{q}{h(a-1)} \frac{v}{u_{ij}}$$

$$v \sim \Gamma(q, q)$$

$$u_{ij} \overset{iid}{\sim} \Gamma(a, a-1), \text{ indep. of } v.$$

The above representation makes it clear that the σ_{ij}'s are, apart from the constant $q/h(a-1)$, products of two unit mean independent random variables: one, $1/u_{ij}$, specific to each σ_{ij}, the other, v, common to all of them. Choosing $a > 1$ and $q < 1$ corresponds to a prior on the σ_{ij}'s where they are approximately of the same unknown size. Once values of a and q are selected, one can set $h = \dfrac{q(a^{\tau\gamma}-1)}{(a-1)b^{\tau\gamma}}$ in order to have $E(\sigma_{ij}) = b^{\tau\gamma}/(a^{\tau\gamma}-1)$. As to the choice of a and q, some guide may be derived from the examination of the following tables, reporting the 0.01 and 0.99 quantiles of the distributions of $1/u_{ij}$ and v for various values of a and q respectively.

a	3	5	10	20	50	100	200	500
0.01 quantile	0.24	0.34	0.48	0.60	0.72	0.79	0.85	0.90
0.99 quantile	4.59	3.13	2.18	1.71	1.40	1.27	1.18	1.11

q	1	0.5	0.2	0.1	0.05
0.01 quantile	0.01	$2 \cdot 10^{-4}$	$3 \cdot 10^{-10}$	$6 \cdot 10^{-20}$	$1 \cdot 10^{-39}$
0.99 quantile	4.61	6.63	11.0	15.9	21.8

In our examples we used $a = 3$ and $q = 0.2$

We remark that our choice of the prior distributions on σ_{ij} and the τ's implies that a priori the contributions of main effects, interactions and error components to the overall variability are of comparable sizes.

4 Examples

We provide three illustrations of the use of our model. Two of them involve experiments with three levels of the first factor, four levels of the second factor and more than one replication in each cell. The third experiment has a larger number of levels on both factor, but only one replication in each cell. Even though in these experiments the number of replications r_{ij} is the same in each cell, we emphasise that our model can as easily be applied to unbalanced and incomplete designs.

In each case the sampler was run for 100,000 sweeps, with an initial 10,000 sweeps of burn-in. While all the simulated values of the allocations $\boldsymbol{z}$, after the burn-in period, were saved for further analysis, the other variables were only saved at the rate of 1 every 100. This was necessary because of the large size of the output. Since the priors employed are invariant with respect to a change in the labels of the allocations $\boldsymbol{z}$, we obtained a clearer and more economical presentation by mapping all the allocations corresponding to a given partition of the levels into one vector. We use the letter $\boldsymbol{\pi}$ to denote these partition vectors.

Examination of the simulated paths did not provide any evidence of obvious lack of convergence of the sampling Markov chain. Simulation times were close to 10-12 minutes on a Sun Sparcstation 4 for the first two examples. The third example required 23 minutes when fitted with no interactions and about 7 hours with interactions (this last run was done only for comparison purposes).

The hyperparameters were set as described in Section 3. The remaining control parameter Δ was set at values that we considered reasonable. In the first example, computations were repeated with a different value of Δ and the results were not dramatically different.

4.1 Surviving toxic agents

We consider the data on survival times analysed by Box and Cox (1964). The data is displayed in Figure 1 and it consists of survival times in hours of animals randomly assigned to each combination of three poisons and four treatments. Four animals were assigned to each combination.

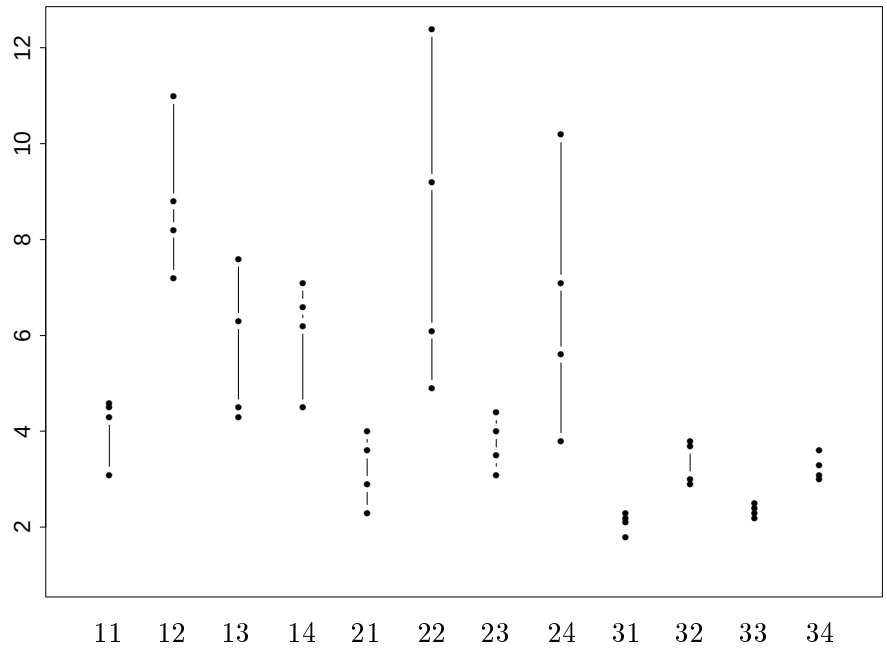


Figure 1: Survival time data set. Time in hours on the ordinates, abscissa are combinations of poisons and treatments. The only purpose of the lines is to facilitate viewing the plot "vertically".

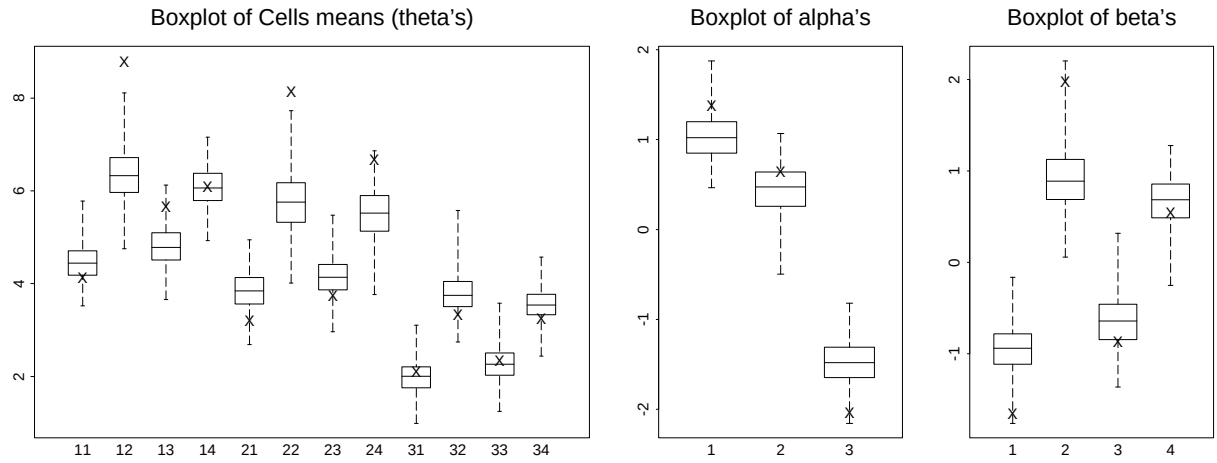
Classical two-way analysis of variance reveals very strong poison and treatment effects (the respective F statistics are $F_{2,36} = 23.2$ and $F_{3,36} = 13.8$) and mild interaction (p-value is 0.11). The analysis in terms of death rates, following from a reciprocal transformation of the response, is more sensitive: the main effects have increased significance while the interaction becomes much weaker (p-value is 0.39).

In effect, the borderline interaction in the analysis in terms of survival times arises because of heteroscedasticity in the error variances (approximately proportional to the cell means), which is not accounted for in the standard analysis of variance. In the model we consider, error variances are allowed to vary across poison/treatment combinations, so that this problem does not arise.

For these data, the control parameter Δ was chosen to be unity, meaning that we would consider two factor levels as essentially equivalent if their effects differed by less than an hour of survival

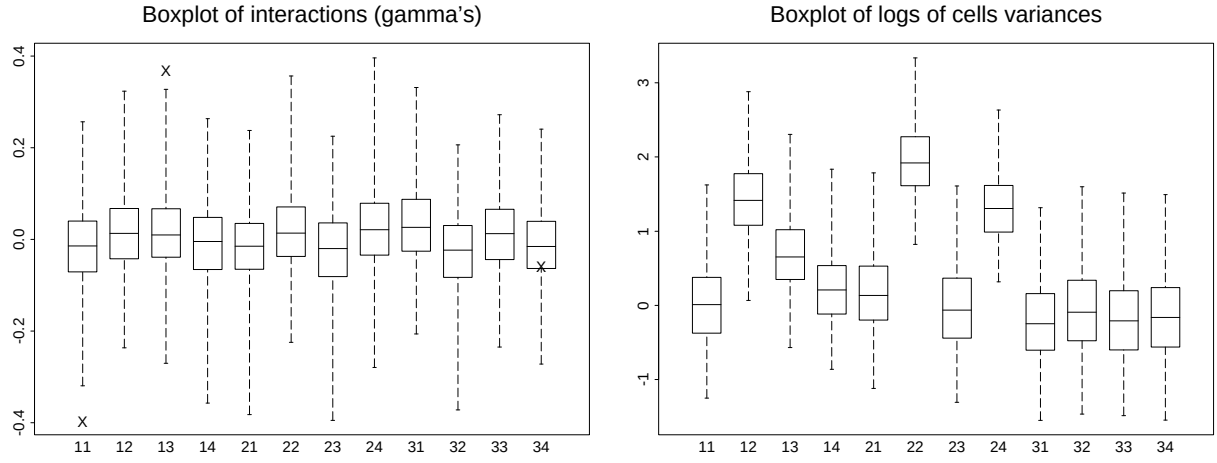
time, and similarly for the interactions. The values of the hyperparameters not explicitly stated in Section 3 were:

$$\sigma^{\mu} = 7744; \quad h = 0.160; \quad b_t^{\alpha} = 0.1099; \quad b_s^{\beta} = 0.0929; \quad b_u^{\gamma} = 0.0573; \quad b^{\tau\alpha} = b^{\tau\beta} = b^{\tau\gamma} = 1.252. \tag{21}$$



(a) (b)

Figure 2: Survival time data set. Panel (a): boxplots of θ_{ij}'s with superimposed the cell means marked as X's. Panel (b): boxplots of α_i's and β_j's. X's are classical estimates.



(a) (b)

Figure 3: Survival time data set. Panel (a): boxplots of interactions γ_{ij}'s, X's are classical estimates. Panel (b): boxplots of the logarithms of the cell variances σ_{ij}.

Figures 2 and 3 contain boxplots of the simulated samples from the posterior distributions for quantities that do not involve the mixture part of our model. Panel (a) in Figure 2 displays boxplots of the cell means θ_{ij}, with an "X" marking the cells sample averages. Clearly, posterior estimates afford much shrinkage, as most of the times the cell mean is outside the posterior interquartile range.

Similar conclusions can be drawn from panel (b) of Figure 2, containing boxplots of the posterior samples for the α's and the β's, and from panel (a) of Figure 3 containing boxplots of the interactions. In particular, the level of shrinkage for the interactions is very great, with only one

classical estimate in 12 falling within the whiskers (in fact most of the marks "X" fall outside of the plot). The distributions of the γ_{ij} are all similar and centred at 0, while clear differences between the α's and between the β's are visible from the boxplots in panel (b) of Figure 2. Posterior distributions of any contrast between the factor levels can be readily obtained from the simulation output. However, as we will detail shortly, our approach to judging whether two levels are the same is based on the posterior probability that the two levels are from the same mixture component, probability also readily estimated from the simulated allocations $\boldsymbol{z}$.

Panel (b) in Figure 3 contains the posterior distributions of the error variances σ_{ij}, on the log-scale. The variances of the observations in cells 12, 22 and 24 stand out as much larger than the rest. As the configuration of boxplots in panel (b) of Figure 3 does not perfectly mimic that in panel (a) of Figure 2, we feel justified in the use of the assumption of different variances, rather than the more structured assumption of proportionality between cell variances and means.

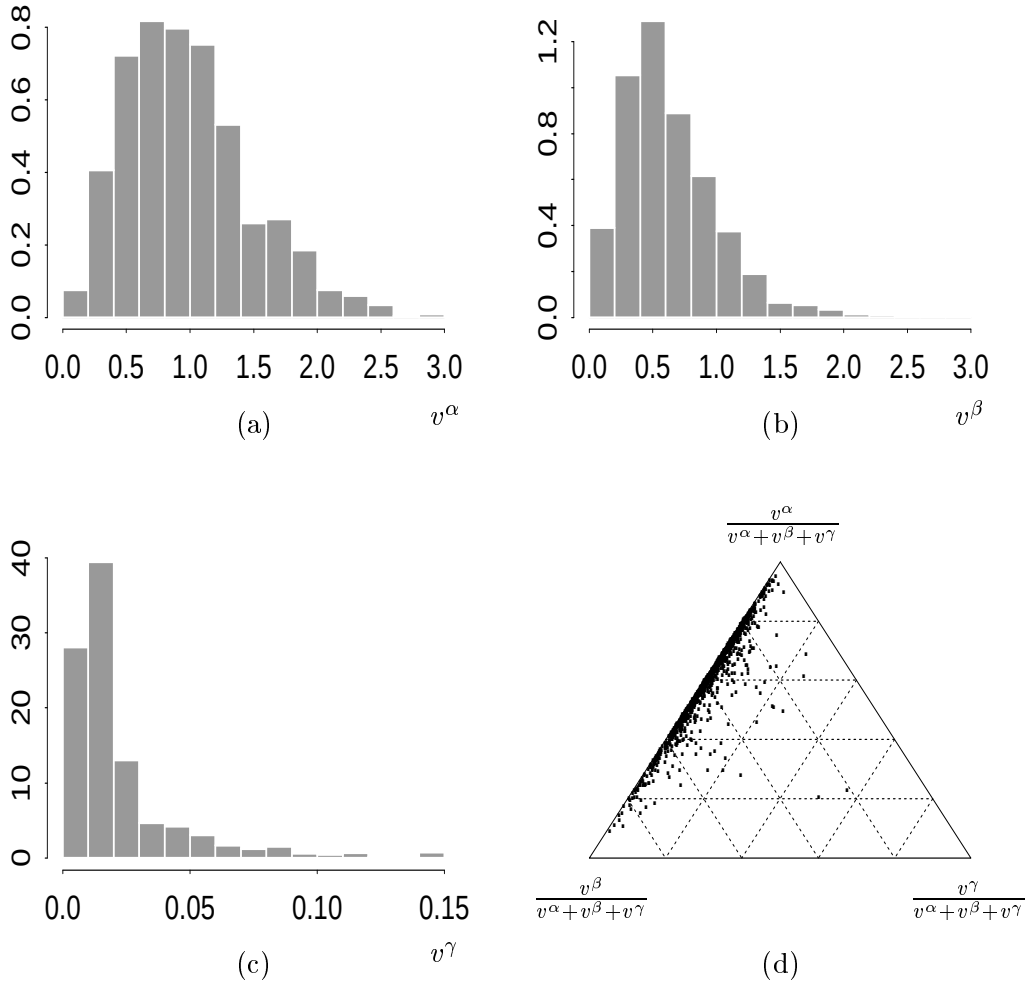


Figure 4: Survival time data set. Panels (a), (b), (c): histograms of samples from the posterior distributions of the variance components. Panel (d): trilinear plot of the posterior sample of the variance components, normalised to sum to unity.

Estimates of the posterior distributions of the variance components can obtained from the

simulation output in several ways, we only illustrate one. Denote $Var(\alpha)$ by v^α. Then, conditional on k^α, $\boldsymbol{w}^\alpha$, $\boldsymbol{\sigma}^\alpha$ and $\boldsymbol{\mu}^\alpha$, one has

$$v^\alpha = \sum_{t=1}^{k^\alpha} w_t^\alpha \sigma_t^\alpha + \sum_{t=1}^{k^\alpha} w_t^\alpha (\mu_t^\alpha)^2 - \left[\sum_{t=1}^{k^\alpha} w_t^\alpha \mu_t^\alpha\right]^2 .$$

Therefore a "sample" of v^α is easily computed from the simulation output. In a similar fashion "samples" of v^β and v^γ can be produced. Figure 4 displays histograms of the sampled v^α, v^β and v^γ, note the much smaller scale of the plot for v^γ. Also displayed is a trilinear plot of the variance components, normalised to sum to unity.

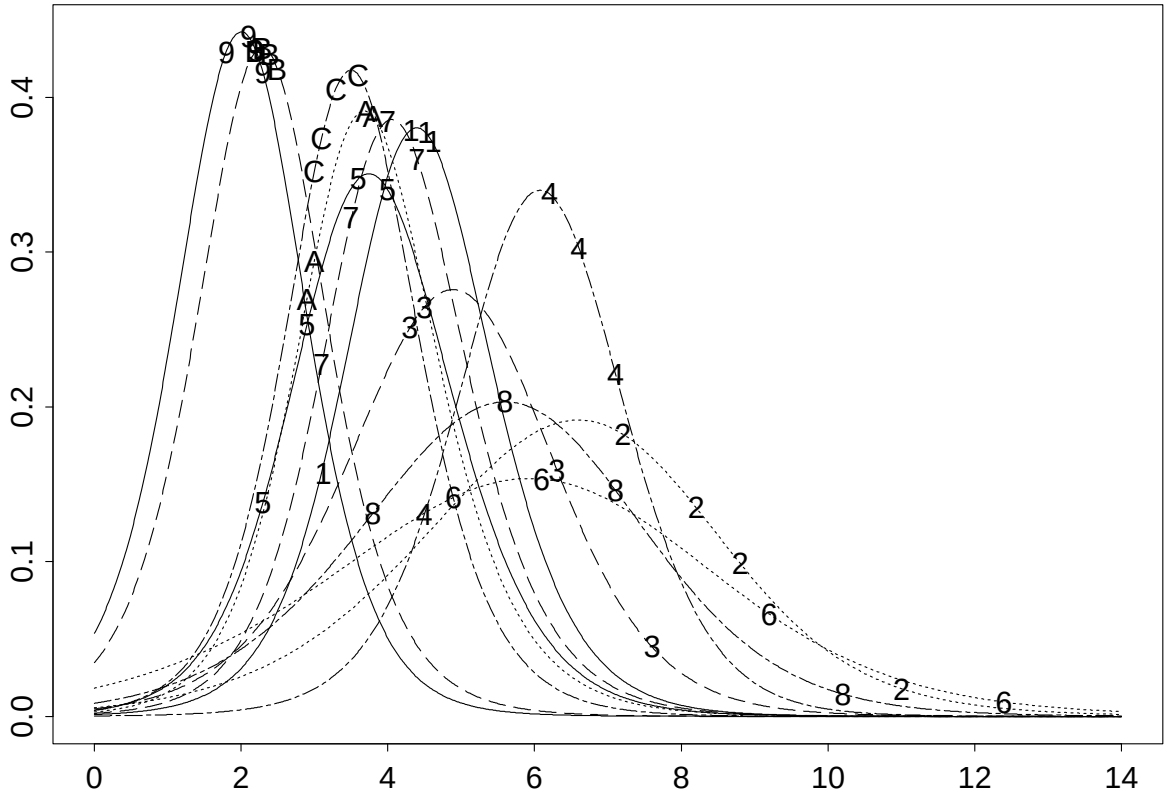


Figure 5: Survival time data set. Posterior predictive densities of the next observation conditioned on the poison/treatment combination. The labels 1,2,...,9,A,B,C denote poison/treatment combination, in the same order as in Figure 1. Each predictive density has four labels on it, placed at points with abscissae equal to the observed survival times.

Predictive distributions of future observations, conditional on the poison/treatment combination, are also easily computed from the simulation output, using the Rao-Blackwellised estimate:

$$p(y_{ij}) = \frac{1}{N} \sum_{l=1}^{N} \phi\left(\frac{y_{ij} - {\theta_{ij}}^{(l)}}{{\sigma_{ij}}^{(l)}}\right)$$

where $({\theta_{ij}}^{(l)}, {\sigma_{ij}}^{(l)})$, $l = 1, \ldots, N = 1000$, are drawn from the posterior and ϕ is the standard normal density. Estimates for the poison/treatment combinations in the data are reported in Figure 5. Predictive distributions of future observables corresponding to unobserved poison/treatment pairs could also be produced at the cost of some additional random variates generation.

As mentioned above, we make statements about which factor levels are alike based on the relative frequency of their being allocated to the same mixture component in the sample from the posterior distribution. For the poison factor, the frequency distribution of $\boldsymbol{\pi}^{\alpha}$ in the posterior sample was:

$\boldsymbol{\pi}^{\alpha}$	111	112	121	211	123
Frequency	1379	63663	287	3731	30940

From the above table we can conclude that the probability of no poison effect, i.e. that the three poisons affect survival time in the same way, is rather small. With probability 0.65 poisons 1 and 2 have the same effect, while with probability approximately 0.31 the three poisons all have different effects.

As for the treatment effects, the most frequent $\boldsymbol{\pi}^{\beta}$ were

$\boldsymbol{\pi}^{\beta}$	1212	2131	1213	1234	2111	2311	1211	1111	2113	1112	1231
Frequency	52673	15992	15934	3412	3166	2960	2015	1425	933	508	422

Thus, the probability of no treatment effect is approximately 0.014. With probability close to 0.53, $\beta_1 = \beta_3$ and $\beta_2 = \beta_4$; with probability close to 0.73, $\beta_1 = \beta_3$ while with probability close to 0.74 $\beta_2 = \beta_4$. Probability statements concerning the joint distribution of $\boldsymbol{\pi}^{\alpha}$ and $\boldsymbol{\pi}^{\beta}$ can as easily be made, based on the simulation output. To exemplify, the event $\{\alpha_1 = \alpha_2 \neq \alpha_3, \beta_1 = \beta_3 \neq \beta_2 = \beta_4\}$ has probability approximately 0.35.

As for the empirical distribution of $\boldsymbol{\pi}^{\gamma}$, its support included 9,299 partition vectors, of which 7,191 were only visited once while 9,035 were visited less than 10 times and accounted for 0.14 of the probability. The vector corresponding to the interactions all belonging to the same component had probability approximately 0.76, while vectors with all but one interaction from the same component accounted for an additional 0.04. These results are consistent with the marginal distributions displayed in Figure 3, panel (a).

The model was re-estimated changing Δ from 1 to 0.25, so that factor levels and interactions were considered as essentially equivalent if the difference of the corresponding survival times was less than 15 minutes. This change yielded the following modification to the list of hyperparameter values given in (21):

$$b_t^{\alpha} = 0.00687; \quad b_s^{\beta} = 0.00580; \quad b_u^{\gamma} = 0.00358.$$

The distributions of the sampled θ's, α's, β's and σ_{ij}'s were very similar to those displayed in Figures 2 and 3. The distribution of the interactions followed the same pattern as with unit Δ, but were considerably further shrunk towards 0.

The sample from the posterior distribution of $\boldsymbol{\pi}^{\alpha}$ had the following frequency distribution:

$\boldsymbol{\pi}^{\alpha}$	111	112	121	211	123
Frequency	0	57693	0	134	42173

As expected, the tighter definition of "being alike" imposed by the smaller value of Δ has the effect of shifting the distribution in favour of more heterogeneous allocation vectors. A similar change occurred in the distribution of $\boldsymbol{\pi}^{\beta}$:

$\boldsymbol{\pi}^{\beta}$	1212	2131	1213	1234	2111	2311	1211	1111	2113	1112	1231
Frequency	52264	18666	15730	3940	3247	3328	809	33	1519	259	172

where, to ease comparison, we maintained the same order of vectors as used above.

As for the empirical distribution of the sampled $\boldsymbol{\pi}^{\gamma}$, its support now contained 1,726 vectors, with 1,069 visited only once and 1,650 visited less than 10 times, the latter accounting for 0.03 of the probability. The vector corresponding to interactions all from the same component had probability 0.95, while the vectors that allocate all but one interaction to one component accounted for an additional 0.01. As for the two main factors, reducing Δ yielded a more concentrated distribution. Here, however, more homogeneous allocation vectors were favoured.

On the whole, the changes were consistent with a more stringent definition of equality between effects and they affect more the details than the overall picture. In the end it is the experimenter's responsibility to define what he/she considers as "essentially equivalent", or in other terms, the size of practically significant (differences between) effects.

4.2 Breaking the fabric

Figure 6 displays the results of an experiment on breaking strength of three fabrics at four temperatures. These data was analysed by Lindley (1972). For each combined level of fabric and temperature, two measurements of breaking strength are available.

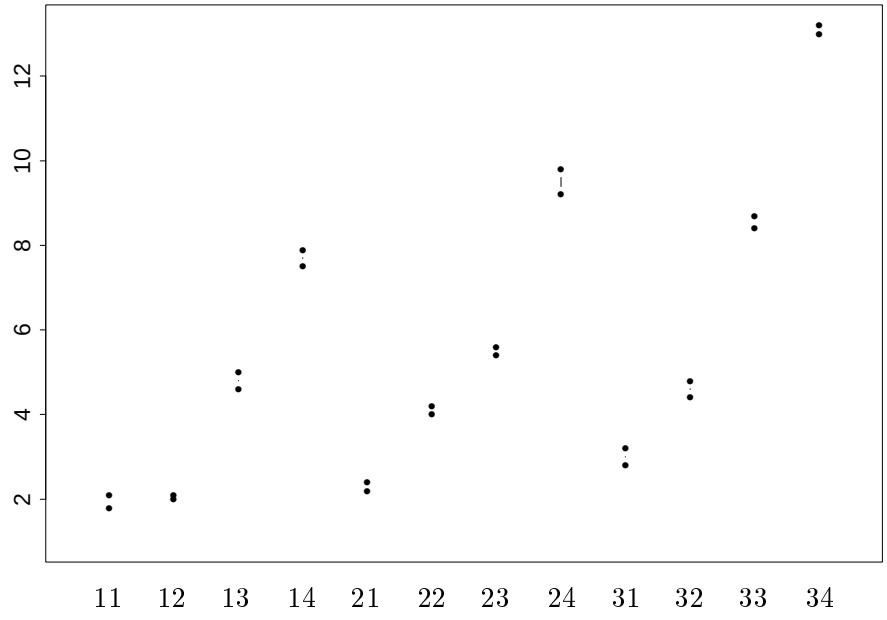


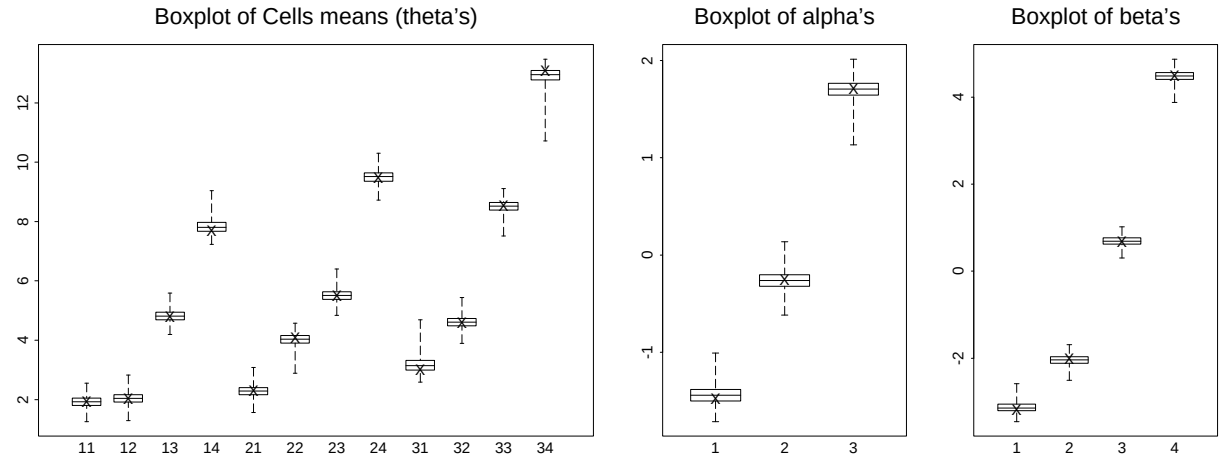
Figure 6: Fabric strength data set. On the abscissa are combinations of fabric and temperature level.

A classical two-way analysis of variance exhibits strong fabric effect ($F_{2,12} = 367.7$), temperature effect ($F_{3,12} = 1241.5$) and interaction ($F_{6,12} = 39.18$).

For these data we chose $\Delta = 2$, here is the list of hyperparameter values:

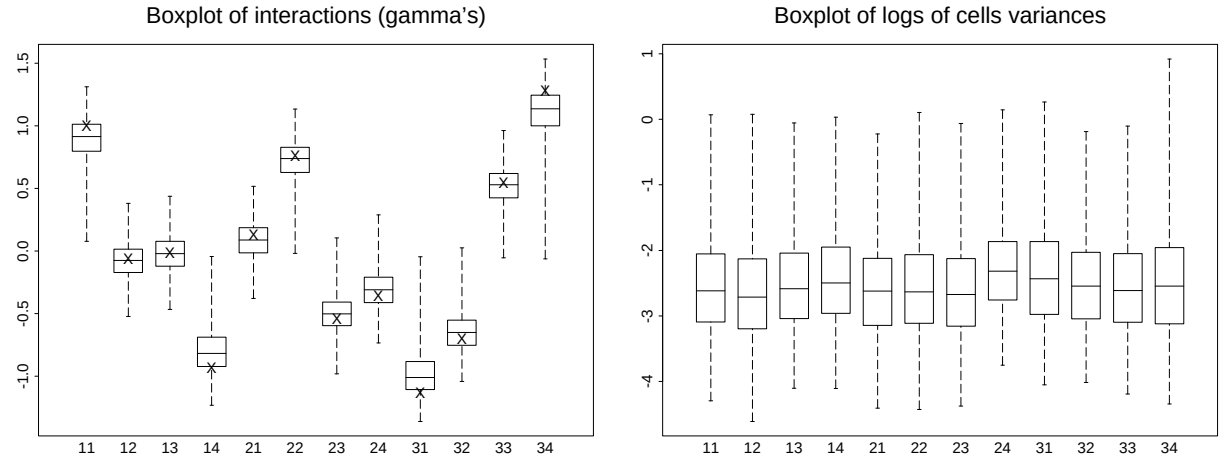
$$\sigma^{\mu} = 17161; \quad h = 0.0856; \quad b_t^{\alpha} = 0.439; \quad b_s^{\beta} = 0.371; \quad b_u^{\gamma} = 0.229;$$
$$b^{\tau\alpha} = 4.479; \quad b^{\tau\beta} = 3.785; \quad b^{\tau\gamma} = 2.336.$$

Figures 7 and 8 display boxplots of the sampled values of the parameters not relating to the



(a) (b)

Figure 7: Fabric strength data set. Panel (a): boxplots of θ_{ij}'s with superimposed the cell means marked as X's. Panel (b): boxplots of α_i's and β_j's. X's are classical estimates.



(a) (b)

Figure 8: Fabric strength data set. Panel (a): boxplots of interactions γ_{ij}'s, X's are classical estimates. Panel (b): boxplots of the logarithms of the cell variances σ_{ij}.

mixture part of the model. The distributions of θ's, α's, β's and γ's are all close to be classical estimates, denoted by an "X", even though some shrinkage can be detected in the extreme γ's, in particular γ_{31} and γ_{34}. The distributions of the cell variances σ_{ij} are all similar, as one would expect from the small number of replications in each cell. However, this common distribution depends on information coming from all the cells and, as a result, it is quite different from the prior distribution.

Consider next the distribution of the allocations. The distributions of $\boldsymbol{\pi}^{\alpha}$ and $\boldsymbol{\pi}^{\beta}$ are very concentrated, as seen from the following tables.

$\boldsymbol{\pi}^{\alpha}$	112	123	211	111	121
Frequency	51202	28469	10117	8622	1590

$\boldsymbol{\pi}^{\beta}$	1123	1234	1122	1111	1112	2113	2311	1121	1211	2111	1213
Frequency	73491	18652	1861	1839	1221	1025	606	461	264	187	173

For both factors, the most probable allocation vector is the one that puts together the first two levels, with the remaining levels in separate groups. The second most likely vector corresponds to all different levels for each factor. In both cases these two top vectors account for most of the probability. Next we note that the temperature levels are ordered and this information ought to be incorporated in the prior, if we wanted to provide more than just an example. One possible approach would be to give unit prior probability to the set of allocations $\boldsymbol{z}^{\beta}$ that group together only consecutive levels of temperature. It is reassuring to see that the posterior of $\boldsymbol{\pi}^{\beta}$, under no such restriction, assigns probability 0.99 to this set.

Figure 9: Fabric strength data set. A graphical display of the probability distribution of $\boldsymbol{\pi}^{\gamma}$.

The distribution of $\boldsymbol{\pi}^{\gamma}$ is much more spread. Its support contained 58,738 vectors, of which 45,209 were visited only once and 58,120 less than 10 times, for a total probability of 0.83. The

probability of all interactions coming from the same mixture component is rather small, 0.02. Also small, approximately 0.04, is the probability that more than 2/3 of the interactions are from the same component. The twelve most probable allocation vectors, excluding the one with all entries equal to 1, account for about 0.03 of the overall probability. They are reported in the following table, where they have been arranged as 3x4 tables.

$\boldsymbol{\pi}^\gamma$				Frequency	$\boldsymbol{\pi}^\gamma$				Frequency	$\boldsymbol{\pi}^\gamma$				Frequency
1	2	2	2		1	2	1	2		1	2	2	3	
2	1	2	2	1078	1	1	2	2	204	2	1	3	3	156
2	2	1	1		2	2	1	1		3	3	1	1	
1	2	2	3		1	1	1	3		1	2	2	2	
2	1	3	2	298	1	1	3	3	199	2	1	2	2	147
3	3	1	1		3	3	1	1		3	2	1	1	
1	2	2	2		1	1	1	3		1	2	2	2	
1	1	2	2	271	1	1	3	1	182	2	1	2	2	128
2	2	1	1		3	3	1	1		2	2	3	1	
1	2	2	2		1	2	2	3		1	2	2	2	
2	1	2	2	252	2	1	2	2	174	3	1	2	2	118
2	2	2	1		3	3	1	1		2	2	1	1	

Examining the above table in conjunction with the boxplots in panel (a) of Figure 8 we see that the most frequent vector corresponds to a partition of the interactions in two groups, with the first group comprising the four largest interactions and the other group the remaining ones. The second most frequent vector is a refinement of the most frequent one: it partitions interactions in positive, null and negative. The remaining ten vectors listed above are all variations of the top two vectors.

A more global idea of the distribution of $\boldsymbol{\pi}^\gamma$ can be obtained from Figure 9. This display is a multivariate analog of the quantile (inverse cdf) function. On the abscissa we report the probability scale, on the ordinate the components of $\boldsymbol{\pi}^\gamma$, the same grey-scale meaning that the components are equal. The plot was created by subsampling 10% of the sampled $\boldsymbol{\pi}^\gamma$, then ordering them to produce a picture with large patches. Therefore, it contains no information concerning the mixing of the sampling chain. The display suggests, in its left part, that the grouping pattern described above accounts for more than the 0.03 probability associated with the twelve vectors reported above.

Diagnosis of the convergence of a MCMC sampler is a growing area of research. Here we only provide some graphical evidence of the behaviour of our sampler. Considering the large number of variables sampled, one has to make a choice concerning which ones to present. In the present problem it seemed natural to display the sums of squares corresponding to the sampled main effects, interactions and residuals. These are reported in Figure 10, for the subsample of 1,000 saved iterates. No obvious non-stationary behavior is evident from the plots. We have resisted the temptation of reporting on the same plots the corresponding sums of squares of the classical analysis ($SS_\alpha = 41.4$, $SS_\beta = 209.5$, $SS_\gamma = 13.2$, $SS_e = 0.675$) for two reasons: the models are different under several respects, in particular, our model allows for unequal variances in each cell; the sum of squares of the simulated α, β, γ and residuals do not afford an orthogonal decomposition of the total sum of squares. The displays in Figure 11 are also based on the simulated sum of squares. At each sweep in the simulation, they give the cumulative probability of some quantiles of the distribution based on all the run. Thus, they allow us to see how some characteristics of the distribution of sampled values varies across the simulation. Again, no clear evidence of transient behaviour is apparent in the plots.

4.3 Plotting potatoes

Here we consider a data set of yields in tonnes/hectare of 7 varieties of potatoes tested at 16 different sites by the National Institute of Agricultural Botany in 1975. The data is reported in Patterson

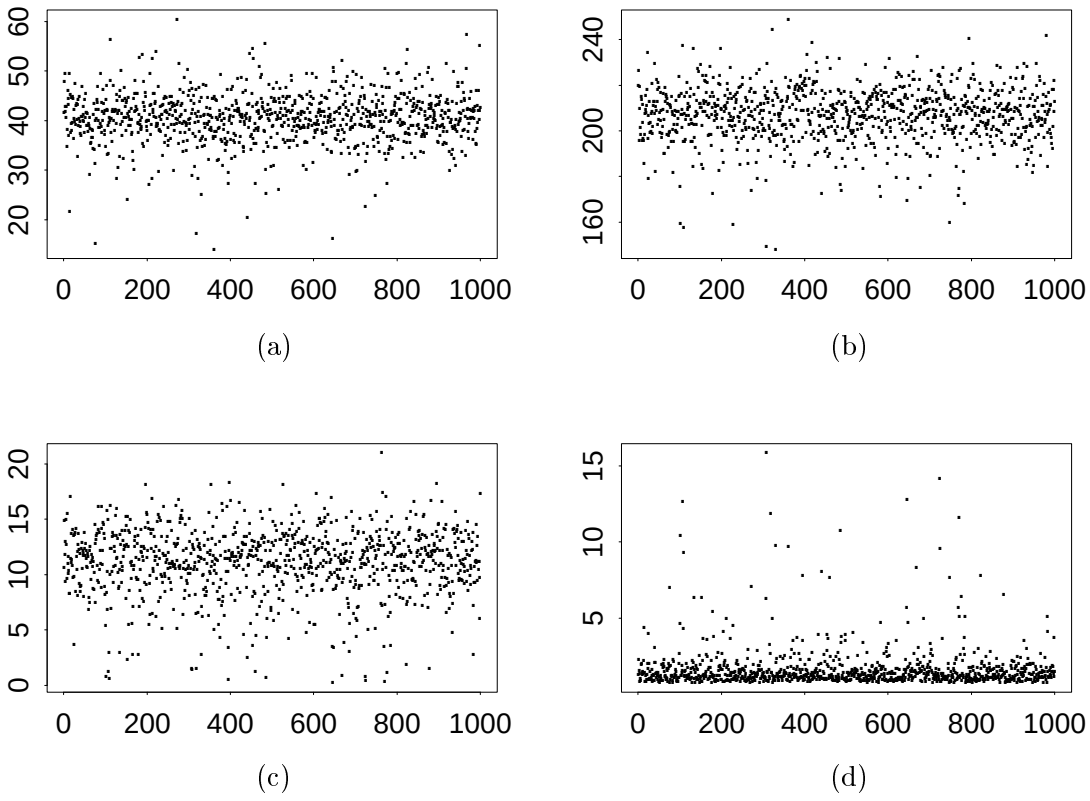


Figure 10: Fabric strength data set. Time series plots of the sum of squares of the α's (a), β's (b), γ's (c) and residual (d) across simulation (only a 1 in 100 subsample).

(1982), page 272. In this data set varieties are of interest and sites are a blocking variable.

The yields are displayed as "X"'s in Figure 12, panel (a), along with boxplots of the distributions of θ_{ij}. Despite the clutter, one can readily see the 16 clumps corresponding to the sites and within each clump the yields for the 7 varieties. There is only one observation for each combination of variety and site, thus a model with no interactions seems appropriate. The standard two-way ANOVA gives an extremely significant site effect ($F_{15,90} = 24.27$) and a very significant variety effect ($F_{6,90} = 3.62$).

However, in the present approach, one can also estimate a model with interaction. In such a model interactions and error components will be in competition to explain the variability which cannot be accounted for by the main effects. We estimated both the model with and without interactions, using $\Delta = 4$, hyperparameter values were as follows:

$$\sigma^{\mu} = 770884; \quad h = 0.00665; \quad b_t^{\alpha} = 0.831; \quad b_s^{\beta} = 1.133; \quad b_u^{\gamma} = 0.502; \quad b^{\tau\alpha} = b^{\tau\beta} = b^{\tau\gamma} = 30.09.$$

Results are very similar, since when interactions are present their posterior distributions are all centred at 0 and similarly distributed, see Figure 12, panel (b). Note, however, that this need not be so: strong prior opinion on small error variances, corresponding to a larger value of h, would yield more differentiated interactions. The following results are all based on the model with no interactions. Figure 13 contains, in panel (a), boxplots for the site and variety effects. The classical estimates of the main effect are all close to the central portions of the posterior distributions, even

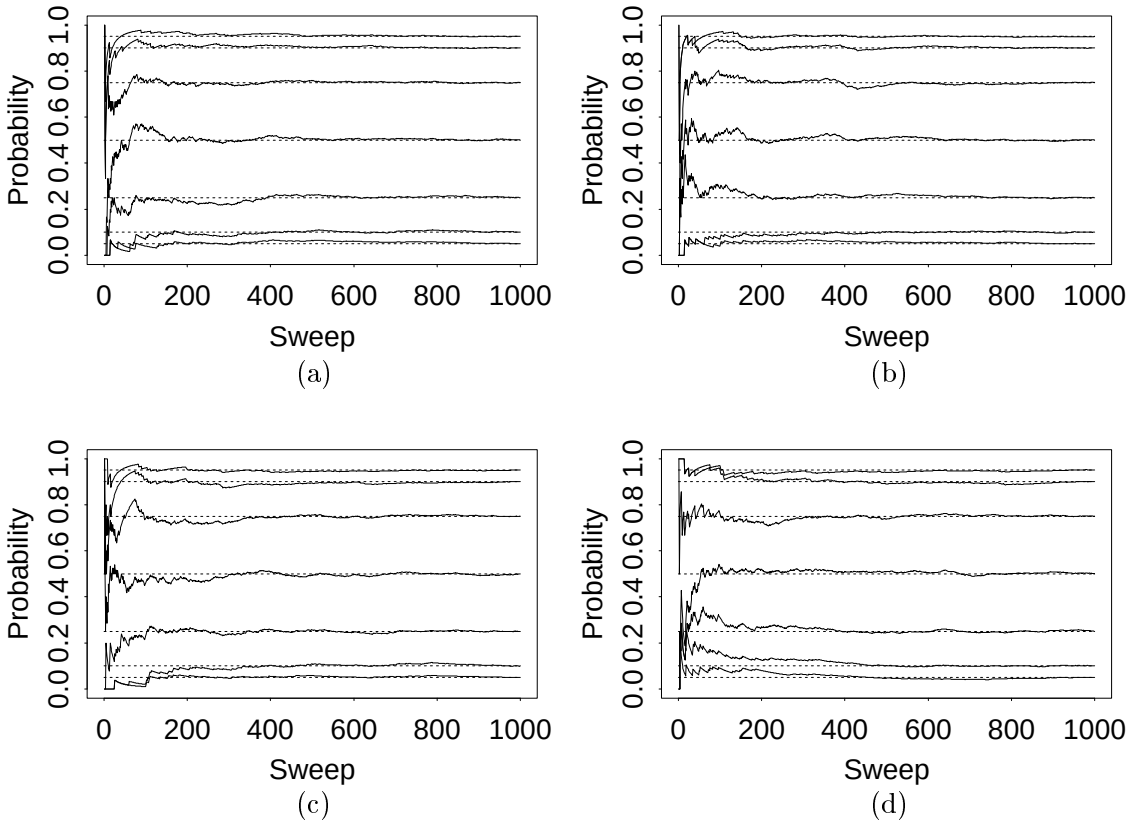


Figure 11: Fabric strength data set. Moving quantiles plots of the sum of squares of the α's (a), β's (b), γ's (c) and residual (d) across simulation (only a 1 in 100 subsample).

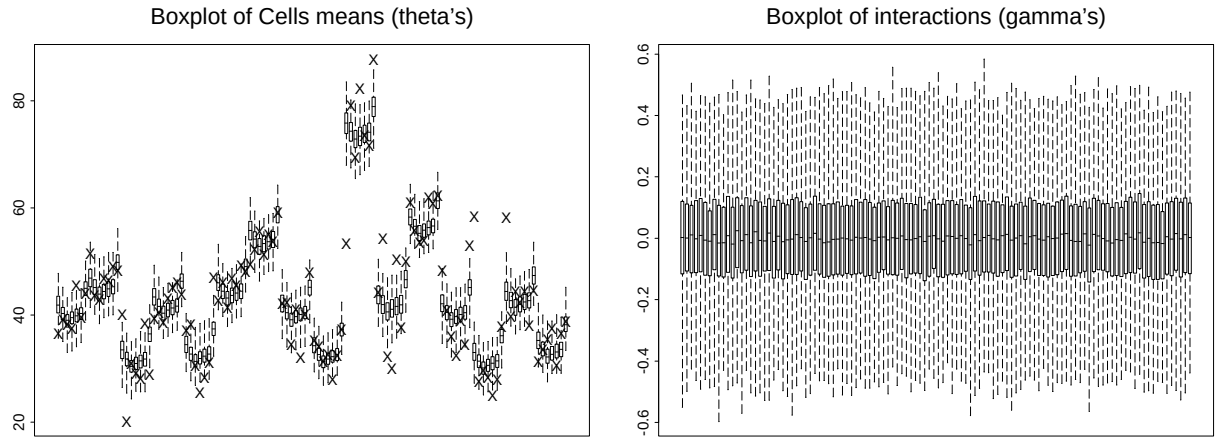
though some shrinkage is evident for the β's, especially β_3 and β_7. The boxplots of $\log(\sigma_{ij})$ in Figure 13 are all similar with the exception of a few, notably for $\sigma_{3,2}$, $\sigma_{10,1}$, $\sigma_{11,2}$, $\sigma_{11,4}$, $\sigma_{14,1}$ and $\sigma_{15,1}$, which assign probability mass to rather larger values. These all correspond to observations which deviate from the sum of the main effects, as one can check from Figure 12, panel (a).

Coming to the allocations, the distribution of $\boldsymbol{\pi}^{\alpha}$ is very spread: our dependent sample of 100,000 visited 54,341 different vectors. Of these, 39,102 were visited only once while 53,532 were visited less than 10 times, for a total probability of 0.83. The five most frequent vectors are

$\boldsymbol{\pi}^{\alpha}$	Frequency
1121213124131212	550
1521253124131212	461
1521253124161212	253
1521253124131252	239
1121213124151212	171

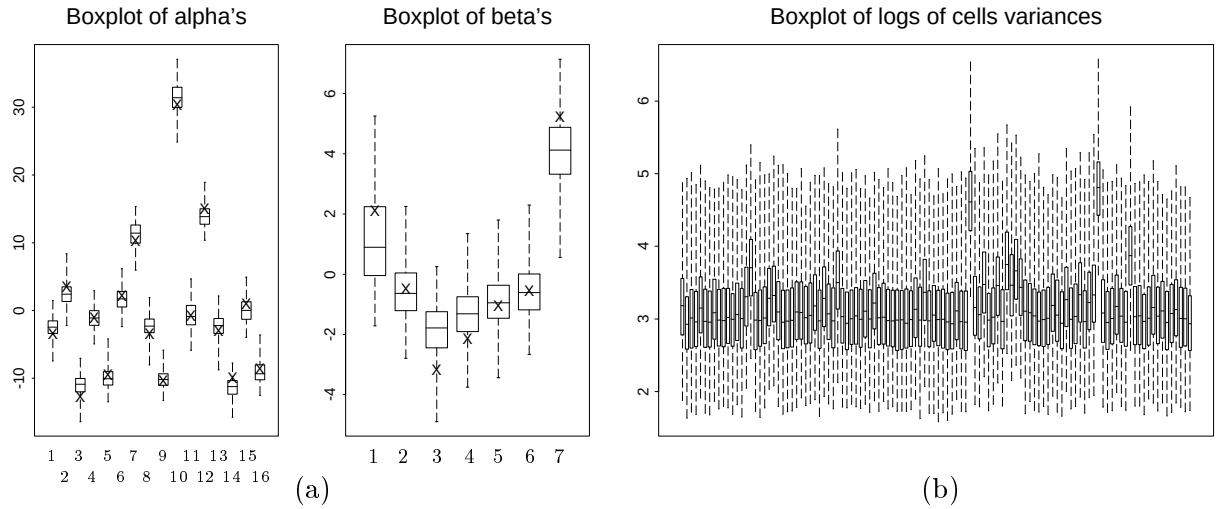
Estimates of probabilities of interest are readily derived from the output. For example, in all but 47 sampled vectors z^{α}_{10} was different from all other allocations, so that the probability that α_{10} is equal to any other level is negligible (in the sample, α_{10} was in the same component as α_{12} 47 times, of these the component also included α_7 once). We report, as other examples, the following estimates:

$$\Pr\{\alpha_7 = \alpha_{12} \neq \alpha_i, i \neq 7, 12\} = 0.45,$$



(a) (b)

Figure 12: Potato trial data set. Panel (a): boxplots of θ_{ij}'s with superimposed the cell means marked as X's, model with no interactions. Panel (b): boxplots of γ_{ij}'s, model with interactions.



(a) (b)

Figure 13: Potato trial data set. Panel (a): boxplots of α_i's and β_j's, X's are classical estimates, model with no interactions. Panel (b): boxplots of the logarithms of the cell variances σ_{ij}, model with no interactions.

$$\Pr\{\alpha_3 = \alpha_5 = \alpha_9 = \alpha_{14} = \alpha_{16}\} = 0.21,$$

$$\Pr\{\alpha_1 = \alpha_2 = \alpha_4 = \alpha_6 = \alpha_8 = \alpha_{11} = \alpha_{13} = \alpha_{15}\} = 0.01.$$

The distribution of $\boldsymbol{\pi}^\beta$ is more concentrated. Its support included 840 vectors, 59 visited only once, 347 less than 10 times, for a total probability of 0.01. With probability 0.01 all the levels are from the same mixture component. The six most probable vectors account for about 0.27 of the probability and are reported below:

$\boldsymbol{\pi}^\beta$	Frequency
1111112	10490
1222221	6764
1222223	4553
1121113	2481
1122113	1414
1232221	1285

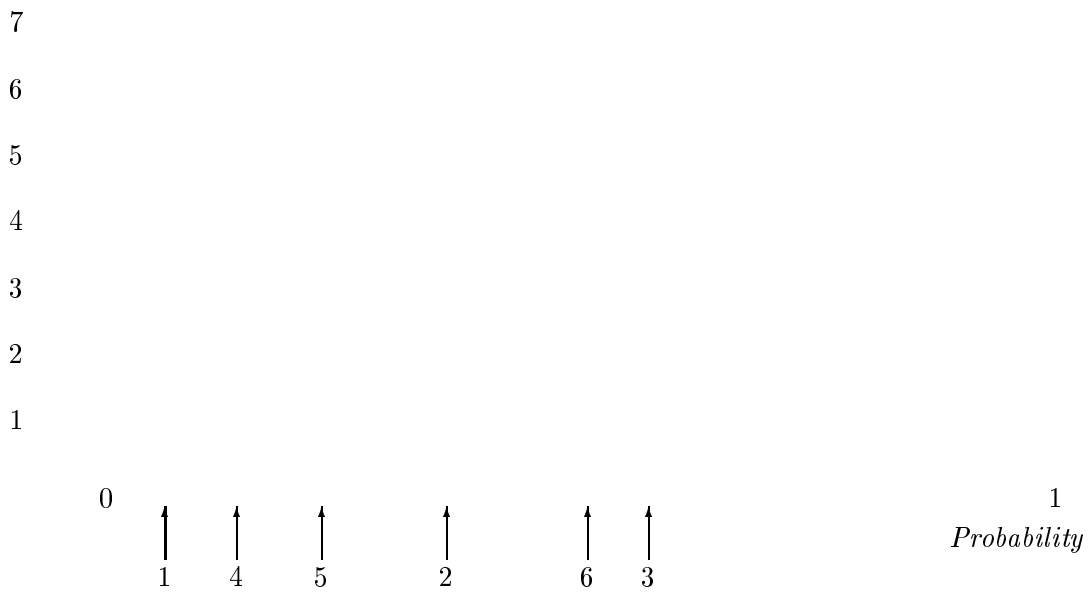


Figure 14: Potato trial data set. A graphical display of the probability distribution of $\boldsymbol{\pi}^\beta$.

These summaries are in agreement with the boxplots exhibited in Figure 13, panel (a). An overall view of the distribution of $\boldsymbol{\pi}^\beta$ is given in Figure 14. This is the same form of display as given in Figure 9 for the interactions in the fabric example. The plot suggests that the pattern where most of the levels in $\{\beta_2, \beta_3, \beta_4, \beta_5, \beta_6\}$ are grouped together accounts for much of the distribution. The six most probable partitions reported above are easily identified, even without the help of the arrows added to the plot.

5 Discussion

In this paper we provide a Bayesian approach to the analysis of factorial experiments that exploits finite mixtures to impose partially exchangeable prior distributions on the main effects and the

interactions. We consider hyperparameter choice, give examples of the computation of various posterior and predictive quantities and provide details of the MCMC sampler.

Although we have striven for it, at first it may seem that our model falls short of full generality in one important respect: its ability to fully accommodate the experimenter's prior beliefs. Consider the case when substantive information about some of the mixture components is available. This may take the form of a series of conditional statements given the number of components in the mixture. It is quite possible that the meaning of each component will depend on the number of components. Thus, the experimenter's beliefs, given $k^\alpha = 2$, about the second component in the mixture may well be different from his/her beliefs conditional on $k^\alpha = 3$. It thus seems that to accommodate these prior beliefs one needs to allow the hyperparameters to vary not only across components but also with respect to the number of components, as in Nobile (1994). This modification can be readily carried out and it would only involve a more complicated expression for the acceptance probability of the split/combine moves, as now changing the number of components may change the hyperparameters of all components.

However, one may counter-argue that if substantive prior information on $\tilde{k}^\alpha$ components is available, this will likely occur when some (possibly unobserved) attribute of the levels is the discriminating element. For instance, in the example of breaking strengths, the experimenter might want to distinguish between fabrics woven in natural and synthetic fibres. This case is accommodated within our model by placing a prior on k^α that assigns zero probability to the set $\{1, \ldots, \tilde{k}^\alpha - 1\}$ while using the available information to form a prior distribution for each component (in the example, one for natural and one for synthetic), thus identifying the labels of the the first $\tilde{k}^\alpha$ components. We emphasise that this does not rule out the possibility of high posterior probability on allocations z^α with much less components than $\tilde{k}^\alpha$, including the allocations $(t, t, \ldots, t)$ corresponding to exchangeable levels, since our mixture models allow for empty components.

Acknowledgements

Stimulating correspondance with Dennis Lindley in the early stages of this research is gratefully acknowledged. The work was supported by an EPSRC Stochastic Modelling in Science and Technology grant.

Appendix: Reversible Jump MCMC for Bayesian Two-way ANOVA

Simulation from the posterior distribution of the parameters and the latent variables is performed using the reversible jump algorithm of Green (1995). For the sake of clarity, we distinguish between moves that do not modify k^α, k^β or k^γ and moves that can change them, thus modifying the the dimension of the parameter space. The first group of moves is presently implemented by simulating from the full conditional distributions of allocations and parameters, except the numbers of mixture components, subject to the constraint (7). The second group of moves follows, with minor modifications, the approach of Richardson and Green (1997): they consist of generalised Metropolis-Hastings moves that either attempt to split one mixture component in two or attempt to combine two mixture components in one. A sweep in the algorithm is a deterministic sequence of moves comprising (i) a split/combine on the α's mixture; (ii) a split/combine on the β's mixture; (iii) a split/combine on the γ's mixture; (iv) draws from the full conditionals subject to (7). Random order sweeps could easily be implemented. Quantities of interest are output at the end of each sweep.

In order to write the full conditionals we need some additional notation. Let $N(x; \mu, \sigma)$ be the

normal density with mean μ and variance σ evaluated at x. Also let

$$y_{ij\cdot} = \frac{1}{r_{ij}} \sum_{k=1}^{r_{ij}} y_{ijk};$$

$$A_t = \{i : z_i^\alpha = t\}; \qquad m_t = \#A_t;$$

$$B_s = \{j : z_j^\beta = s\}; \qquad n_s = \#B_s;$$

$$C_u = \{(i,j) : z_{ij}^\gamma = u\}; \qquad r_u = \#C_u;$$

$$\overline{\alpha}_t = \frac{1}{m_t} \sum_{i \in A_t} \alpha_i; \qquad \overline{\beta}_s = \frac{1}{n_s} \sum_{j \in B_s} \beta_j; \qquad \overline{\gamma}_u = \frac{1}{r_u} \sum_{(i,j) \in C_u} \gamma_{ij}.$$

The following distributions are all conditional on the observed data $\boldsymbol{y}$ and the other parameters/latent variables.

$$\boldsymbol{w}^\alpha \sim Dir(d_1^\alpha + m_1, \ldots, d_{k^\alpha}^\alpha + m_{k^\alpha}), \tag{22}$$

$$\Pr\{z_i^\alpha = t\} = \frac{w_t^\alpha N(\alpha_i; \mu_t^\alpha, \sigma_t^\alpha)}{\sum_{v=1}^{k^\alpha} w_v^\alpha N(\alpha_i; \mu_v^\alpha, \sigma_v^\alpha)}, \tag{23}$$

$$\sigma_t^\alpha \sim I\Gamma\left(a_t^\alpha + \frac{m_t}{2}, b_t^\alpha + \frac{1}{2} \sum_{i \in A_t} (\alpha_i - \mu_t^\alpha)^2\right), \tag{24}$$

$$\mu_t^\alpha \sim N\left(\frac{\xi_t^\alpha \tau^\alpha + \overline{\alpha}_t m_t / \sigma_t^\alpha}{\tau^\alpha + m_t/\sigma_t^\alpha}, (\tau^\alpha + m_t/\sigma_t^\alpha)^{-1}\right), \tag{25}$$

$$\tau^\alpha \sim \Gamma\left(a^{\tau\alpha} + \frac{k^\alpha}{2}, b^{\tau\alpha} + \frac{1}{2} \sum_{t=1}^{k^\alpha} (\mu_t^\alpha - \xi_t^\alpha)^2\right). \tag{26}$$

The full conditional distributions of $\boldsymbol{w}^\beta$, $\boldsymbol{w}^\gamma$, z_j^β, z_u^γ, σ_s^β, σ_u^γ, μ_s^β, μ_u^γ, τ^β, τ^γ are obtained from (22)–(26) with obvious modifications. For the parameters in the error component we have:

$$\sigma_{ij} \sim I\Gamma\left(a + \frac{r_{ij}}{2}, b + \frac{1}{2} \sum_k (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2\right),$$

$$b \sim \Gamma\left(q + amn, h + \sum_{i,j} \sigma_{ij}^{-1}\right).$$

If it was not for the constraints in (7) we would have the following full conditionals of the parameters in the systematic component of the model:

$$\mu \sim N\left(\frac{\eta/\sigma^\mu + \sum_{i,j}(y_{ij\cdot} - \alpha_i - \beta_j - \gamma_{ij}) r_{ij} \sigma_{ij}^{-1}}{1/\sigma^\mu + \sum_{i,j} r_{ij}\sigma_{ij}^{-1}}, \left(1/\sigma^\mu + \sum_{i,j} r_{ij}\sigma_{ij}^{-1}\right)^{-1}\right), \tag{27}$$

$$\alpha_i \sim N\left(\frac{\mu_{z_i^\alpha}^\alpha / \sigma_{z_i^\alpha}^\alpha + \sum_j (y_{ij\cdot} - \mu - \beta_j - \gamma_{ij}) r_{ij} \sigma_{ij}^{-1}}{1/\sigma_{z_i^\alpha}^\alpha + \sum_j r_{ij}\sigma_{ij}^{-1}}, \left(1/\sigma_{z_i^\alpha}^\alpha + \sum_j r_{ij}\sigma_{ij}^{-1}\right)^{-1}\right), \tag{28}$$

$$\beta_j \sim N\left(\frac{\mu_{z_j^\beta}^\beta / \sigma_{z_j^\beta}^\beta + \sum_i (y_{ij\cdot} - \mu - \alpha_i - \gamma_{ij}) r_{ij} \sigma_{ij}^{-1}}{1/\sigma_{z_j^\beta}^\beta + \sum_i r_{ij}\sigma_{ij}^{-1}}, \left(1/\sigma_{z_j^\beta}^\beta + \sum_i r_{ij}\sigma_{ij}^{-1}\right)^{-1}\right), \tag{29}$$

$$\gamma_{ij} \sim N\left(\frac{\mu^{\gamma}_{z^{\gamma}_{ij}}/\sigma^{\gamma}_{z^{\gamma}_{ij}} + (y_{ij\cdot} - \mu - \alpha_i - \beta_j) r_{ij}\sigma_{ij}^{-1}}{1/\sigma^{\gamma}_{z^{\gamma}_{ij}} + r_{ij}\sigma_{ij}^{-1}}, \left(1/\sigma^{\gamma}_{z^{\gamma}_{ij}} + r_{ij}\sigma_{ij}^{-1}\right)^{-1}\right). \tag{30}$$

Simulation from the full conditionals is done conditional on the constraints in (7). Note that when simulating the α's one only need to make use of $\sum_i \alpha_i = 0$, as the β's and γ's are given and they already satisfy the respective constraints. Similar remarks apply to the simulation of the β's and γ's. The same argument implies that the constraint in (7) need not be considered explicitly when simulating from the other full conditionals.

Next we show how to simulate the α's, β's and γ's subject to (7). For the sake of clarity, rewrite (28) as

$$\alpha_i \sim N(\tilde{\mu}^{\alpha}_i, \tilde{\sigma}^{\alpha}_i).$$

It can shown that the joint distribution of $(\alpha_1, \ldots, \alpha_{m-1})^T$ conditional on $\overline{\alpha}$ is multivariate normal with mean vector

$$(\tilde{\mu}^{\alpha}_1, \ldots, \tilde{\mu}^{\alpha}_{m-1})^T + \frac{m}{\sum_{i=1}^m \tilde{\sigma}^{\alpha}_i}(\tilde{\sigma}^{\alpha}_1, \ldots, \tilde{\sigma}^{\alpha}_{m-1})^T \left(\overline{\alpha} - \frac{1}{m}\sum_{i=1}^m \tilde{\mu}^{\alpha}_i\right) \tag{31}$$

and covariance matrix with (r, s) element equal to

$$\tilde{\sigma}^{\alpha}_r \delta_{rs} - \frac{\tilde{\sigma}^{\alpha}_r \tilde{\sigma}^{\alpha}_s}{\sum_{i=1}^m \tilde{\sigma}^{\alpha}_i} \qquad r, s = 1, \ldots, m-1 \tag{32}$$

where δ_{rs} is the Kronecker delta. To simulate the α's, draw $(\alpha_1, \ldots, \alpha_{m-1})^T$ from the multivariate normal distribution with mean as in (31), with $\overline{\alpha} = 0$, and variance matrix as in (32). Then let $\alpha_m = -\sum_{i=1}^{m-1} \alpha_i$.

Simulating from the full conditional distribution of the β's is performed in a quite similar way. With obvious notation, draw $(\beta_1, \ldots, \beta_{n-1})^T$ from a multivariate normal distribution with mean

$$(\tilde{\mu}^{\beta}_1, \ldots, \tilde{\mu}^{\beta}_{n-1})^T + \frac{n}{\sum_{j=1}^n \tilde{\sigma}^{\beta}_j}(\tilde{\sigma}^{\beta}_1, \ldots, \tilde{\sigma}^{\beta}_{n-1})^T \left(-\frac{1}{n}\sum_{j=1}^n \tilde{\mu}^{\beta}_j\right)$$

and covariance matrix with (r, s) element equal to

$$\tilde{\sigma}^{\beta}_r \delta_{rs} - \frac{\tilde{\sigma}^{\beta}_r \tilde{\sigma}^{\beta}_s}{\sum_{j=1}^n \tilde{\sigma}^{\beta}_j} \qquad r, s = 1, \ldots, n-1. \tag{33}$$

Then set $\beta_n = -\sum_{j=1}^{n-1} \beta_j$.

Simulating the γ's is a little more involved. Rewrite (30) as

$$\gamma_{ij} \sim N(\tilde{\mu}^{\gamma}_{ij}, \tilde{\sigma}^{\gamma}_{ij}).$$

Denote by $\gamma_{\cdot+}$ the vector of m row means and by $\gamma_{+\cdot}$ the vector consisting of the first $n-1$ column means:

$$\gamma_{\cdot+} = (\gamma_{1+}, \ldots, \gamma_{m+})^T = \frac{1}{n}\left(\sum_{j=1}^n \gamma_{1j}, \ldots, \sum_{j=1}^n \gamma_{mj}\right)^T$$

$$\gamma_{+\cdot} = (\gamma_{+1}, \ldots, \gamma_{+,n-1})^T = \frac{1}{m}\left(\sum_{i=1}^m \gamma_{i1}, \ldots, \sum_{i=1}^m \gamma_{i,n-1}\right)^T.$$

It can be shown that the conditional distribution of the vector

$$(\gamma_{11},\ldots,\gamma_{1,n-1},\gamma_{21},\ldots,\gamma_{2,n-1},\ldots,\gamma_{m-1,1},\ldots,\gamma_{m-1,n-1})^T \tag{34}$$

given $\gamma_{\cdot+}$ and $\gamma_{+\cdot}$ is multivariate normal with mean vector

$$\chi_1 + A_{12}A_{22}^{-1}\left(\begin{bmatrix}\gamma_{\cdot+}\\ \gamma_{+\cdot}\end{bmatrix} - \chi_2\right) \tag{35}$$

and covariance matrix

$$A_{11} - A_{21}^T A_{22}^{-1} A_{21} \tag{36}$$

where

$$\chi_1 = (\tilde{\mu}_{11}^{\gamma},\ldots,\tilde{\mu}_{1,n-1}^{\gamma},\tilde{\mu}_{21}^{\gamma},\ldots,\tilde{\mu}_{2,n-1}^{\gamma},\ldots,\tilde{\mu}_{m-1,1}^{\gamma},\ldots,\tilde{\mu}_{m-1,n-1}^{\gamma})^T,$$

$$\chi_2 = \left(\frac{1}{n}\sum_{j=1}^{n}\tilde{\mu}_{1j}^{\gamma},\ldots,\frac{1}{n}\sum_{j=1}^{n}\tilde{\mu}_{mj}^{\gamma},\frac{1}{m}\sum_{i=1}^{m}\tilde{\mu}_{i1}^{\gamma},\ldots,\frac{1}{m}\sum_{i=1}^{m}\tilde{\mu}_{i,n-1}^{\gamma}\right)^T,$$

A_{11} is a diagonal matrix with diagonal entries given by the vector

$$(\tilde{\sigma}_{11}^{\gamma},\ldots,\tilde{\sigma}_{1,n-1}^{\gamma},\tilde{\sigma}_{21}^{\gamma},\ldots,\tilde{\sigma}_{2,n-1}^{\gamma},\ldots,\tilde{\sigma}_{m-1,1}^{\gamma},\ldots,\tilde{\sigma}_{m-1,n-1}^{\gamma})^T,$$

$$A_{12} = \begin{bmatrix} \frac{1}{n}\begin{bmatrix} \tilde{\sigma}_{11}^{\gamma} & \cdots & \tilde{\sigma}_{1,n-1}^{\gamma} & & & & \\ & & & \ddots & & & \\ & & & & \tilde{\sigma}_{m-1,1}^{\gamma} & \cdots & \tilde{\sigma}_{m-1,n-1}^{\gamma} \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \\ \frac{1}{m}\begin{bmatrix} \tilde{\sigma}_{11}^{\gamma} & & & & \tilde{\sigma}_{m-1,1}^{\gamma} & & \\ & \ddots & & \cdots & & \ddots & \\ & & \tilde{\sigma}_{1,n-1}^{\gamma} & & & & \tilde{\sigma}_{m-1,n-1}^{\gamma} \end{bmatrix} \end{bmatrix}$$

and

$$A_{22} = \begin{bmatrix} \frac{1}{n^2}\sum_{j=1}^{n}\tilde{\sigma}_{1j}^{\gamma} & & & \frac{1}{mn}\tilde{\sigma}_{11}^{\gamma} & \cdots & \frac{1}{mn}\tilde{\sigma}_{1,n-1}^{\gamma} \\ & \ddots & & & \cdots & \\ & & \frac{1}{n^2}\sum_{j=1}^{n}\tilde{\sigma}_{mj}^{\gamma} & \frac{1}{mn}\tilde{\sigma}_{m1}^{\gamma} & \cdots & \frac{1}{mn}\tilde{\sigma}_{m,n-1}^{\gamma} \\ \frac{1}{mn}\tilde{\sigma}_{11}^{\gamma} & \cdots & \frac{1}{mn}\tilde{\sigma}_{m1}^{\gamma} & \frac{1}{m^2}\sum_{i=1}^{m}\tilde{\sigma}_{i1}^{\gamma} & & \\ & \cdots & & & \ddots & \\ \frac{1}{mn}\tilde{\sigma}_{1,n-1}^{\gamma} & \cdots & \frac{1}{mn}\tilde{\sigma}_{m,n-1}^{\gamma} & & & \frac{1}{m^2}\sum_{i=1}^{m}\tilde{\sigma}_{i,n-1}^{\gamma} \end{bmatrix}.$$

Therefore, the γ's are simulated by drawing the vector (34) form a multivariate normal distribution with mean vector as given in (35) and covariance matrix as in (36). Then, the remaining components of γ are computed as follows:

$$\gamma_{i,n} = -\sum_{j=1}^{n-1}\gamma_{ij} \qquad i = 1,\ldots,m-1;$$

$$\gamma_{m,j} = -\sum_{i=1}^{m-1}\gamma_{ij} \qquad j = 1,\ldots,n.$$

The reversible jump moves follow, with minor changes, the ones proposed by Richardson and Green (1997, Section 3.2). Besides the exclusive use of split/combine moves (Richardson and Green also employ birth-and-death moves), the main difference is that we do not constraint the component means to be ordered. If the hyperparameters $(d_t^\alpha, a_t^\alpha, b_t^\alpha, \xi_t^\alpha)$ differ across components then the corresponding labels are uniquely identified. If a single value is specified for each hyperparameter, inference about the mixture parameters requires some identifying constraint on the labels to be imposed, we perform this exercise in a post-processing "after simulation" stage. Moreover, in the present context the most interesting quantities (main factor levels, interactions and their partitions) do not depend on the mixture labels.

In each simulation sweep either a split or a combine is attempted for the components in the mixtures (4), (5) and (6). We only discuss the moves for the mixture of the α's, and in doing so we drop the superscript $^\alpha$ from the relevant parameters and hyperparameters. We discuss first the case where hyperparameters differ across components, then the case where they are all identical. The split/combine move begins with the selection of a candidate new state. This is selected by first making a random choice between splitting, with probability s_k, and combining, with probability $c_k = 1 - s_k$, where $s_1 = c_{k_{max}} = 1$ and $s_k = 0.5$, $k = 2, \ldots, k_{max} - 1$. Suppose that there are currently k components in the mixture. If split is selected, we randomly choose one of these components, say j^*, and we split it in two components j_1 and j_2 according to the following recipe:

$$\begin{array}{cc} w_{j_1} = w_{j^*} u_1; & w_{j_2} = w_{j^*}(1-u_1); \\ \mu_{j_1} = \mu_{j^*} - u_2\sqrt{\sigma_{j^*}}\sqrt{w_{j_2}/w_{j_1}}; & \mu_{j_2} = \mu_{j^*} + u_2\sqrt{\sigma_{j^*}}\sqrt{w_{j_1}/w_{j_2}}; \\ \sigma_{j_1} = u_3(1-u_2^2)\sigma_{j^*}(w_{j^*}/w_{j_1}); & \sigma_{j_2} = (1-u_3)(1-u_2^2)\sigma_{j^*}(w_{j^*}/w_{j_2}). \end{array}$$

where $u_1 \sim Be(2,2)$, $u_2 \sim 2Be(2,2) - 1$ and $u_3 \sim Be(1,1)$. The candidate state is obtained by removing j^* and adding j_1 and j_2 to the list of existing components. We make the arbitrary convention that j_1 will take the place of j^* and j_2 will become the $(k+1)$-th component. In the candidate state, the observations presently allocated to j^* are reallocated to components j_1 and j_2 in accordance with:

$$\Pr[z_i = j_1] = \frac{p_1}{p_1 + p_2} \qquad \Pr[z_i = j_2] = \frac{p_2}{p_1 + p_2} \qquad i \in A_{j^*} \tag{37}$$

where

$$p_1 = \frac{w_{j_1}}{\sqrt{\sigma_{j_1}}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mu_{j_1})^2}{\sigma_{j_1}}\right\} \qquad \text{and} \qquad p_2 = \frac{w_{j_2}}{\sqrt{\sigma_{j_2}}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mu_{j_2})^2}{\sigma_{j_2}}\right\}.$$

The candidate state is then accepted with probability $\min(1, R)$ with

$$\begin{aligned} R &= \frac{\sigma_{j^*}^{\frac{m_{j_1}+m_{j_2}}{2}}}{\sigma_{j_1}^{\frac{m_{j_1}}{2}}\sigma_{j_2}^{\frac{m_{j_2}}{2}}} \exp\left\{-\frac{1}{2}\left[\sum_{i\in A_{j_1}}\frac{(\alpha_i-\mu_{j_1})^2}{\sigma_{j_1}} + \sum_{i\in A_{j_2}}\frac{(\alpha_i-\mu_{j_2})^2}{\sigma_{j_2}} - \sum_{i\in A_{j_1}\cup A_{j_2}}\frac{(\alpha_i-\mu_{j^*})^2}{\sigma_{j^*}}\right]\right\} \\ &\times \frac{p(k+1)}{p(k)}\frac{\Gamma(\sum d_j + d_{j_1} + d_{j_2})}{\Gamma(\sum d_j + d_{j^*})}\frac{\Gamma(d_{j^*})}{\Gamma(d_{j_1})\Gamma(d_{j_2})}\frac{w_{j_1}^{d_{j_1}+m_{j_1}-1} w_{j_2}^{d_{j_2}+m_{j_2}-1}}{w_{j^*}^{d_{j^*}+m_{j_1}+m_{j_2}-1}} \\ &\times \sqrt{\frac{\tau_{j_1}\tau_{j_2}}{(2\pi)\tau_{j^*}}} \exp\left\{-\frac{1}{2}\left[\tau_{j_1}(\mu_{j_1}-\xi_{j_1})^2 + \tau_{j_2}(\mu_{j_2}-\xi_{j_2})^2 - \tau_{j^*}(\mu_{j^*}-\xi_{j^*})^2\right]\right\} \\ &\times \frac{b_{j_1}^{a_{j_1}} b_{j_2}^{a_{j_2}}}{b_{j^*}^{a_{j^*}}}\frac{\Gamma(a_{j^*})}{\Gamma(a_{j_1})\Gamma(a_{j_2})}\frac{\sigma_{j^*}^{a_{j^*}+1}}{\sigma_{j_1}^{a_{j_1}+1}\sigma_{j_2}^{a_{j_2}+1}} \exp\left\{-\frac{b_{j_1}}{\sigma_{j_1}} - \frac{b_{j_2}}{\sigma_{j_2}} + \frac{b_{j^*}}{\sigma_{j^*}}\right\} \\ &\times \frac{c_{k+1}}{s_k P_{alloc}}\left\{\frac{1}{2}g_{2,2}(u_1)g_{2,2}\left(\frac{u_2+1}{2}\right)g_{1,1}(u_3)\right\}^{-1} w_{j^*}(1-u_2^2)\left[\frac{\sigma_{j^*}}{u_1(1-u_1)}\right]^{\frac{3}{2}} \end{aligned} \tag{38}$$

where the $\sum d_j$ is over the indexes of the components in the mixture that are not affected by the split/combine move, P_{alloc} is the probability of the reallocations in (37), $g_{1,1}$ and $g_{2,2}$ are the densities of $Be(1,1)$ and $Be(2,2)$ distributions.

The reverse of a split is a combine. If combine is selected, two components are selected as follows: j_1 is randomly chosen from the first k existing ones while j_2 is set equal to $(k+1)$, to ensure reversibility. Then a a new component j^* is formed according to

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2}; \\ \mu_{j^*} &= (w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2})/w_{j^*}; \\ \sigma_{j^*} &= \left[w_{j_1}(\mu_{j_1}^2 + \sigma_{j_1}) + w_{j_2}(\mu_{j_2}^2 + \sigma_{j_2})\right]/w_{j^*} - \mu_{j^*}^2. \end{aligned}$$

The candidate state results from removing j_1 and j_2 from the list of components and placing j^* in the place that was of j_1. The factor levels associated with j_1 and j_2 are also reallocated to j^*. The candidate is accepted with probability $\min(1, R^{-1})$, where R is given in (38).

The above split/combine moves are designed so that the "ejected" component in a split and the "absorbed" component in a combine are always the last component in the list. This is done to ensure that if component k is present, so are all the preceding ones (of course this makes complete sense only because of the differing hyperparameters). When the hyperparameters do not vary across components the above moves are still applicable. However, better mixing is achieved by a further randomisation: after the split of j^* into j_1 and j_2, randomly place j_2 in one of the $k+1$ possible locations (1-st, 2-nd, ..., $(k+1)$-th) in the list. The corresponding change in the combine consists in choosing j_1 and j_2 randomly from the $(k+1)$ existing components. It turns out that the ratio R used in the acceptance probabilities is unaffected by these modifications. Moreover, from a computational viewpoint, the random placement of j_2 needs not be done, so that the only modification consists in the random choice of j_1 and j_2 in the combine move.

References

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*, John Wiley & Sons, Chichester.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *J. Roy. Statist. Soc.* B, **26**, 211–243.

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Bush, C. A. and MacEachern, S. N. (1996) A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–285.

Consonni, G. and Veronese, P. (1995) A Bayesian method for combining results from several binomial experiments. *J. Amer. Statist. Assoc.*, **90**, 935–944.

Dickey, J. (1974) Bayesian alternatives to the F-test and least-squares estimate in the normal linear model. *Studies in Bayesian Econometrics and Statistics* S. E. Fienberg and A. Zellner, eds., Amsterdam: North-Holland, 515–554.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Gopalan, R. and Berry, D. A. (1996) *Bayesian multiple comparisons using Dirichlet process priors.* Technical report, ISDS, Duke University.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and Richardson, S. (1998) *Modelling heterogeneity with and without the Dirichlet process.* In preparation.

Groenewald, P. C. N. (1991) Multiple hypothesis and clustering for normal means - A Bayesian analysis. *South African Statist. J.*, **25**, 1–18.

Gumbel, E. J. (1958) *Statistics of Extremes*, Columbia University Press.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kass, R. E. and Raftery, A. E. (1995) Bayes Factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

Kempthorne, O. (1955) The randomization theory of experimental inference. *J. Amer. Statist. Assoc.*, **50**, 946–967.

Kendall, M. G. and Stuart, A. (1963) *The Advanced Theory of Statistics*, vol. 1, 2nd edition, Griffin & Co., London.

Lindley, D. V. (1971) *Bayesian Statistics: a review.* Philadelphia, SIAM.

Lindley, D. V. (1972) A Bayesian solution for two-way analysis of variance. Proceedings of the 9-th European Meeting of Statisticians, Budapest, Colloquia Mathematica Societatis János Bolyai, 475–496.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with Discussion). *J. Roy. Statist. Soc.* B, **34**, 1–42.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Nobile, A. (1994) *Bayesian analysis of finite mixture distributions*, Ph.D. thesis, Carnegie Mellon University.

Nobile, A. (1997) A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. Submitted for publication.

Patterson, H. D. (1982) FITCON and the analysis of incomplete varieties × trials tables. *Utilitas Mathematica*, **21A**, 267–289.

Richardson, S. and Green, P. J. (1997) On the Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. Roy. Statist. Soc.* B, **59**, 731–792.

Schervish, M. J. (1992) Bayesian analysis of linear models. *Bayesian Statistics*, **4**, 419–434, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds, Oxford: Oxford University Press.

Schervish, M. J. (1995) *Theory of Statistics*, Springer-Verlag.

Smith, A. F. M. (1973) Bayes estimates in one-way and two-way models. *Biometrika*, **60**, 319–329.

Vines, S. K., Gilks, W. R. and Wild, P. (1996) Fitting Bayesian multiple random effects models, *Statistics and Computing*, **6**, 337–346.

Tierney, L. (1994) Markov chains for exploring posterior distributions, *Annals of Statistics*, **22**, 1701–1762.