Bayesian analysis of factorial experiments by mixture modelling

Agostino Nobile* and Peter J. Green†

University of Glasgow, UK and University of Bristol, UK.

November 1997

Revised March 1999

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 two illustrative applications, we discuss implementation, presentation of posterior distributions, sensitivity, and performance of the Markov chain Monte Carlo methods that are used.

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

1 Introduction

Faster computers and the increasing popularity of Markov chain Monte Carlo methods have allowed Bayesian statistics to become the tool of choice for a growing number of statistical practitioners

*Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK. Email: agostino@stats.gla.ac.uk

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

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 & Tiao (1973) and Lindley & Smith (1972). Box & 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 appropriate inference procedure for fixed and random effects "depends upon 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 a random effects model is assumed. Lindley & 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. Gopalan & Berry (1998) advocate an approach to multiple comparisons that more fully builds in the discrete character of the hypothesis-testing problem: 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, and posterior probabilities for the cells are computed by Markov chain Monte Carlo methods. The estimative and partition-based approaches co-exist in the paper by Bush & MacEachern (1996) on Bayesian analysis of the randomised block experiment, with Dirichlet process priors used for the block effects and ordinary normal priors for the treatments.

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 & 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 incorporates the researcher'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 & Green (1997), rather than adopting the more restrictive Dirichlet process models. Comparisons between these classes of models will be found in Green & Richardson (1998). Our approach bears some resemblance with the one used by Consonni & 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. In our model, this is achieved via the prior distribution on the mixture allocation variables.

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 modular, and intended to be extendible to more complicated designs, and to experiments including covariates. Computations are all done by Markov chain Monte Carlo, 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 Markov chain Monte Carlo, 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. Two 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 Markov chain Monte Carlo performance. Details of the sampler are deferred to the Appendix.

2 A Bayesian two-way random effects model

We consider a two-way layout model. For $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, suppose there are r_{ij} replicate observations $\{y_{ijk}, k = 1, 2, \ldots, r_{ij}\}$ in cell (i, j), corresponding to the i-th level of factor 1 and the j-th level of factor 2. 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 notational simplicity we 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, although these are only rarely explicitly mentioned. Quantities for which a distribution is not specified are fixed constants and need to be assigned before 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}:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{1}$$

The error terms ϵ_{ijk} are independently normally distributed $\epsilon_{ijk} \sim N(0, \sigma_{ij})$, with zero means and variances σ_{ij} allowed to differ from cell to cell according to the model

$$\sigma_{ij} \stackrel{iid}{\sim} I\Gamma(a, b), \qquad b \sim \Gamma(q, h). \tag{2}$$

The overall level μ has normal prior distribution $\mu \sim N(\eta, \sigma^\mu)$. The remaining terms in the systematic component (1) are assumed to proceed from finite mixtures of unknown numbers of normal component distributions, subject to the classical identifying constraints:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_j \gamma_{ij} = 0, \quad \sum_i \gamma_{ij} = 0. \tag{3}$$

More precisely, we first consider

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

and then take, as the prior distribution on the α's, the conditional distribution of $(\alpha_1, \ldots, \alpha_m)^T$ given $\sum \alpha_i = 0$, where this is defined as the limit of the distribution given $|\sum_i \alpha_i| < \delta$ as $\delta \to 0$ (all similar conditionals in this paper should be interpreted in the same way). Thus the α's are dependent random variables. Similarly, the prior distributions of the β's and γ's are obtained by first considering

$$\beta_j \stackrel{iid}{\sim} \sum_{s=1}^{k^\beta} w_s^\beta N(\mu_s^\beta, \sigma_s^\beta), \qquad \gamma_{ij} \stackrel{iid}{\sim} \sum_{u=1}^{k^\gamma} w_u^\gamma N(\mu_u^\gamma, \sigma_u^\gamma) \tag{5}$$

and then conditioning on $\sum \beta_j = 0$, $\sum_j \gamma_{ij} = 0$, $\sum_i \gamma_{ij} = 0$.

Next we specify the distributions for the parameters in the mixtures (4)–(5). We only give these explicitly for the α's since similar structures are assumed for the β's and γ's. For the number of components k^α, the prior is uniform on the integers from 1 to some maximum value k^α_{max}. See Section 3 for further discussion of this point. The mixture weights follow a Dirichlet distribution: $w^\alpha \sim Dir(d_1^\alpha, \ldots, d_{k^\alpha}^\alpha)$. We employ normal and inverse gamma distributions as priors on the component means and variances

$$\mu_t^\alpha \stackrel{ind.}{\sim} N(\xi_t^\alpha, 1/\tau^\alpha), \qquad \sigma_t^\alpha \stackrel{ind.}{\sim} I\Gamma(a_t^\alpha, b_t^\alpha).$$

The prior precision of the component means is assumed to have a 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 to permit prior specifications incorporating substantial information distinguishing the components. However, typically one may want to provide a common value for each of them, making the mixture components exchangeable. In Section 3 we discuss a practicable strategy for hyperparameter choice which selects values corresponding to very well separated mixture components, to meet the requirement that factor levels from the same component are "practically indistinguishable".

The mixture assumption on main effects and interactions in (4)–(5) can be restated by introducing latent *allocation variables* z^α, z^β, z^γ 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). On introducing the allocations, equation (4) can be restated as follows. Conditional on the mixture weights w^α, each component in the allocation vector z^α is independently drawn from the multinomial distribution with $\text{pr}(z_i^\alpha = t) = w_t^\alpha$. Once we condition on the z^α's, the distribution of the α's reduces to a singular m-variate normal with covariance of rank $m-1$. Analogous distributions hold for the mixtures in (5), see the Appendix for further details. Introducing the allocations greatly facilitates computations. More importantly, it illuminates the partial exchangeability structures on main effects and interactions embedded in the prior (for discussions and references on partial exchangeability see, e.g., Bernardo & Smith, 1994 and Schervish, 1995). Specifically, each allocation vector z^α induces a partition of the α's into subsets, with exchangeability holding within each. Positive prior probability is assigned to each allocation vector, including those corresponding to only one subset (all exchangeable α's) and to m subsets, thus affording great modelling flexibility.

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 y depends on the parameter $(\mu, \alpha, \beta, \gamma)$ only through θ and the map from $(\mu, \alpha, \beta, \gamma)$ to θ is not one to one, $(\mu, \alpha, \beta, \gamma)$ is not identified.

In principle, lack of identifiability in the likelihood poses no problem to the Bayesian provided the prior distribution is proper (Lindley, 1971 page 46, Lindley & Smith, 1972), although in such a situation inference may be very sensitive to prior assumptions. In practice, Markov chain Monte Carlo sampling of the resulting posterior faces slow convergence problems: on contours of constant likelihood the posterior is proportional to the prior and, as sample size increases, it will tend to concentrate on a lower dimensional manifold. Gelfand, Sahu & Carlin (1995) suggested a centering reparameterisation for nested random effects models, while Vines, Gilks & Wild (1996) proposed a reparameterisation for multiple random effects models by sweeping, based on the classical constraints. Another possibility is improving mixing by incorporating Metropolis-Hastings moves that allow for swift changes along contours of constant likelihood; for an example, see Nobile (1998).

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.

2.2 Other models

In the above model we have assumed prior independence between the allocations z^α, z^β and z^γ. In some contexts it may be preferable to entertain more structured models, having the property that $z^\alpha_{i_1} = z^\alpha_{i_2}$ and $z^\beta_{j_1} = z^\beta_{j_2}$ imply $z^\gamma_{i_1 j_1} = z^\gamma_{i_2 j_2}$. At one extreme one can assume that the product partition induced by z^α and z^β is the partition of z^γ. In this model, 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 z^γ. 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 z^α, z^β and z^γ. However, we have preferred to use the more flexible specification with prior independent allocations.

We conclude this section by mentioning one modification going towards *reducing* structure. Rather than assuming mixture distributions on the factor levels and the interactions, one could directly model the cell means θ_{ij} with a normal mixture. This model is easier to implement and is more flexible than the one we entertain: for instance, in a 2×2 design, it allows directly considering the hypothesis $\theta_{11} = \theta_{12} = \theta_{21} \neq \theta_{22}$ that requires a much more complicated formulation in terms of α's, β's and γ's. This added flexibility may well provide the easiest approach to modelling, but 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 this specification. 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. We also make minimal use of the data, specifically in equation (6).

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 numbers of factor levels m and n, which play a role analogous to the number of data points in a mixture analysis, is typically small, 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 π^α of the α's induced by the allocations z^α.

The mixture weights are chosen to have uniform distribution on the appropriate simplexes: $d_t^\alpha = d_s^\beta = d_u^\gamma = 1$. The prior on k^α, w^α and z^α induces a prior distribution on the partitions π^α of the α's; similarly for the partitions π^β and π^γ. In the example in Section 4.1, with $m = 3$ and $n = 4$, the prior specification adopted yielded the following prior distributions on π^α and π^β:

π^α	Prior probability
111	0.6
112, 121, 211	0.1222
123	0.0333

π^β	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 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:

$$\sigma^\mu = 100 \cdot \max_{i,j} y_{ij\cdot}^2. \tag{6}$$

As for the prior locations of the mixture components means, we set them all equal to 0: $\xi_t^\alpha = \xi_s^\beta = \xi_u^\gamma = 0$. Our recipe for the remaining hyperparameters is a little more involved, so we prefer to organise it in subsections.

3.1 Variability between and within mixture components

Two sets of hyperparameters control the variability of the normal components in the mixtures (4), (5). The variability within components is controlled by the hyperparameters a_t^α, b_t^α, a_s^β, b_s^β, a_u^γ, b_u^γ in the prior distributions of σ_t^α, σ_s^β and σ_u^γ. The variability between component means depends on the hyperparameters $a^{\tau\alpha}$, $b^{\tau\alpha}$, $a^{\tau\beta}$, $b^{\tau\beta}$, $a^{\tau\gamma}$, $b^{\tau\gamma}$ through the prior precisions τ^α, τ^β, τ^γ. Our discussion is only in terms of the hyperparameters governing the mixture (4), the same considerations apply to the hyperparameters in the distributions of σ_s^β, σ_u^γ, τ^β, τ^γ. In order to lighten the notation, in this section and the following one we denote σ_t^α, a_t^α, b_t^α τ^α, $a^{\tau\alpha}$, $b^{\tau\alpha}$ by σ_t, a_t, b_t τ, a^τ, b^τ 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 and b_t such that the distribution of σ_t 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

$$p_0 = \mathrm{pr}(|\alpha_i - \alpha_j| \leq \Delta) \tag{7}$$

where p_0 is close to 1. After integrating σ_t out, $\alpha_i - \alpha_j$ has a t distribution with $2a_t$ degrees of freedom, location 0 and precision $a_t/(2b_t)$, i.e., $(\alpha_i - \alpha_j)\{a_t/(2b_t)\}^{1/2} \sim t_{2a_t}$. Thus, (7) becomes

$$p_0 = 2F_{2a_t}\left\{\Delta\left(\frac{a_t}{2b_t}\right)^{1/2}\right\} - 1 \tag{8}$$

where F_{2a_t} is the distribution function of a t_{2a_t} distribution. Solving (8) for b_t yields

$$b_t = \frac{a_t}{2}\Delta^2\left\{F_{2a_t}^{-1}\left(\frac{1+p_0}{2}\right)\right\}^{-2}.$$

We choose the shape parameter $a_t = 3$, in order to have finite second moments for σ_t. The selection of p_0 is discussed at the end of the section.

Consider next the hyperparameters in the distribution of τ, governing the spread of the mixture component means μ_t^α. Here too we choose $a^\tau = 3$ to ensure finite second moments. Since we wish to interpret differences between component means as practically significant differences, their prior distribution should assign little probability to $(-\Delta, \Delta)$. We choose of doing this by requiring that for any two component means μ_t^α and μ_r^α, the ratio between the probability densities of $\mu_t^\alpha - \mu_r^\alpha$ and $\alpha_i - \alpha_j$ be less than 1 on the interval $(-\Delta, \Delta)$, while the opposite holds on $(-\infty, -\Delta) \cup (\Delta, \infty)$. After integrating out τ, $\mu_t^\alpha - \mu_r^\alpha$ has a t distribution with $2a^\tau$ degrees of freedom, location 0 and precision $a^\tau/(2b^\tau)$. Therefore, the above requirement leads to the equation

$$t_{2a_t}\left\{\Delta\left(\frac{a_t}{2b_t}\right)^{1/2}\right\}\left(\frac{a_t}{2b_t}\right)^{1/2} = t_{2a^\tau}\left\{\Delta\left(\frac{a^\tau}{2b^\tau}\right)^{1/2}\right\}\left(\frac{a^\tau}{2b^\tau}\right)^{1/2} \tag{9}$$

where t_{2a_t} denotes the probability density of a standard t distribution with $2a_t$ degrees of freedom. Since $a_t = a^\tau$, equation (9) has only one solution in b^τ (beside the trivial one $b^\tau = b_t$), which can be easily determined numerically, e.g., using bisection.

In this procedure, p_0 controls both b_t and b^τ. Increasing p_0 tightens the distribution of $\alpha_i - \alpha_j$ around 0, thus lowering b_t; it also lowers the density of $\alpha_i - \alpha_j$ at Δ, with the result of a larger spread for the distribution of $\mu_t^\alpha - \mu_r^\alpha$, i.e., larger b^τ. The following table contains, for some levels of p_0, the values of b_t and b^τ determined by our procedure.

p_0	0.8	0.9	0.95	0.99
b_t/Δ^2	0.7236	0.3973	0.2505	0.1091
b^τ/Δ^2	3.619	10.23	30.04	454.4

As already explained when introducing (7), p_0 is close to 1. However, values very close to 1 should be avoided as they correspond to prior distributions that assign extremely small probability to σ_t and $1/\tau$ having about the same magnitude. Given p_0, the distribution of $(a^\tau b_t)(a_t b^\tau)^{-1}(\sigma_t\tau)^{-1}$ is $F_{2a_t,2a^\tau}$. In the following table we provide $\mathrm{pr}\{(\sigma_t\tau)^{-1} \in I\}$ for selected intervals I, corresponding to few values of p_0.

		I				
	pr(I)	$(0,1)$	$(1,10)$	$(10,10^2)$	$(10^2,10^3)$	$(10^3,\infty)$
	0.8	0.035	0.75	0.21	0.0010	$1.2\cdot 10^{-6}$
p_0	0.9	0.00049	0.14	0.80	0.061	0.00015
	0.95	$5.6\cdot 10^{-6}$	0.0040	0.41	0.57	0.010
	0.99	$1.4\cdot 10^{-10}$	$1.4\cdot 10^{-7}$	0.00012	0.053	0.95

It seems to us that sensible values of p_0 lie close 0.95 and in our examples we used $p_0 = 0.95$.

3.2 Within cells variability

We suggest choosing a, q and h so that the distribution of σ_{ij} is proper with finite second moments and approximately centred at the expected value of $1/\tau$, the prior variance of the means in the mixture components. For the sake of clarity, we rewrite (2) as follows:

$$\sigma_{ij} = \frac{q}{h(a-1)}\frac{v}{u_{ij}}, \qquad v \sim \Gamma(q,q), \qquad u_{ij} \overset{iid}{\sim} \Gamma(a,a-1), \text{ indep. of } v.$$

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 $h = q(a^\tau - 1)/\{(a-1)b^\tau\}$ in order to have $E(\sigma_{ij}) = E(1/\tau) = b^\tau/(a^\tau - 1)$. 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.

From the previous discussion one observes that b_t, b^τ and $1/h$ are all proportional to Δ^2. This suggests an Empirical Bayes variant of our recipe which does not require explicitly specifying Δ: follow the recipe as described with $\Delta = 1$, then multiply the resulting b_t, b^τ by $s_y^2(a^\tau - 1)/b^\tau$ and divide h by the same quantity. The effect is to set $E(\sigma_{ij})$ and $E(1/\tau)$ equal to s_y^2, the sample variance of the observations, while implicitly selecting a value of Δ.

4 Examples

We provide two illustrations of the use of our model. One of them involves a 3×4 experiment with replication, the other has a larger number of levels on both factors, but only one observation per cell. Even though the designs of these experiments are balanced, 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. With the exception of the allocations, simulated values were only recorded at the rate of 1 every 100, to save space. Since the priors employed are invariant with respect to relabelling the allocations z, we obtained a clearer and more economical presentation in terms of the partitions, denoted by π in place of z. The simulated paths did not display any obvious lack of convergence of the sampling Markov chain. Simulation times were close to 10 minutes on a Sun Sparcstation

4 for the first example. The second 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 A small design with replication

We consider the data on survival times analysed by Box & 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.

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

In effect, the borderline-significant interaction in the analysis of survival times arises because of heteroscedasticity in the error variances, which is not accounted for in the standard analysis. In the model we consider, error variances are allowed to vary between cells, avoiding this problem.

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:

$$\sigma^{\mu} = 7744; \quad h = 0.006658; \quad b_t^{\alpha} = b_s^{\beta} = b_u^{\gamma} = 0.2505; \quad b^{\tau\alpha} = b^{\tau\beta} = b^{\tau\gamma} = 30.04. \tag{10}$$

Panel (a) of Figure 2 displays boxplots of the cell means θ_{ij}, with "X" marking the cells sample averages. Clearly, posterior estimates afford much shrinkage, as the cell sample average is usually outside the posterior interquartile range. Similar conclusions can be drawn from panels (b) and (c) of Figure 2, containing boxplots of the posterior samples for the main effects and interactions. 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. 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. Panel (d) in Figure 2 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.

Estimates of the posterior distributions of the variance components can be obtained from the simulation output in several ways; we only illustrate one. Denote var(α) by v^α. Then, conditional on k^α, w^α, σ^α and μ^α, 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. Figure 3 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.

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

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

where $\{\theta_{ij}{}^{(l)}, \sigma_{ij}{}^{(l)}\}$, $l = 1, \ldots, N = 1000$, are drawn from the posterior and $\phi(y; \theta, \sigma)$ is the normal density with mean θ and variance σ evaluated at y. Estimates for the poison/treatment combinations in the data are reported in Figure 4.

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 posterior sample. For the poison factor, the frequency distribution of π^α in the posterior sample was:

π^α	111	112	121	211	123
Frequency	2703	75148	221	5403	16525

From the above table we can conclude that the probability of no poison effect is about 0.03. With probability 0.78 poisons 1 and 2 have the same effect, while with probability approximately 0.17 the three poisons all have different effects. As for the treatment effects, the most frequent π^β were

π^β	1212	1213	1211	2131	1111	2311	2111	1234	1231	1112	2113
Frequency	47540	15873	9217	8665	5423	4730	4393	1994	663	490	396

Thus, the probability of no treatment effect is approximately 0.05. With probability close to 0.48, $\beta_1 = \beta_3$ and $\beta_2 = \beta_4$; with probability close to 0.79, $\beta_1 = \beta_3$ while with probability close to 0.66

$\beta_2 = \beta_4$. Probability statements concerning the joint distribution of π^α and π^β 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 0.37. Regarding the empirical distribution of π^γ, its support included 4,336 partition vectors, of which 3,036 were only visited once while 4,192 were visited less than 10 times and accounted for 0.07 of the probability. The interactions all belonged to the same component with probability approximately 0.88, while vectors with all but one interaction from the same component accounted for an additional 0.03. These results are consistent with the marginal distributions displayed in Figure 2, panel (c).

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 (10):

$$h = 0.1065; \quad b_t^\alpha = b_s^\beta = b_u^\gamma = 0.01566; \quad b^{\tau\alpha} = b^{\tau\beta} = b^{\tau\gamma} = 1.877.$$

The distributions of the sampled θ's, α's, β's and σ_{ij}'s were very similar to those displayed in Figure 2. 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 π^α had the following frequency distribution:

π^α	111	112	121	211	123
Frequency	1	58978	0	306	40715

As expected, the tighter definition of "being alike" imposed by the smaller value of Δ leads to lower posterior probability of allocation to the same component. A similar change occurred in the distribution of π^β:

π^β	1212	1213	1211	2131	1111	2311	2111	1234	1231	1112	2113
Frequency	53191	19905	848	17388	92	2125	1665	3886	90	159	590

where, to ease comparison, we maintained the same order of vectors as used above. The empirical distribution of the sampled π^γ did not differ much from the one obtained for $\Delta = 1$: all interactions came from the same component with probability 0.90, and all but one from the same component with probability 0.02. 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.

Diagnosis of the convergence of Markov chain Monte Carlo methods is a growing area of research. Here we only provide some graphical evidence of the behaviour of our sampler. Among the many variables sampled, in the present context it seemed natural to display the sums of squares corresponding to the sampled main effects, interactions and residuals. These are reported in Figure 5, for the subsample of 1,000 saved iterates. No obvious non-stationary behaviour 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 ($SS_\alpha = 103.3$, $SS_\beta = 92.12$, $SS_\gamma = 25.01$, $SS_e = 80.07$) for two reasons: the models are different in 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 6 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.

4.2 A larger unreplicated experiment

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 are reported in Patterson (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 7, panel (a), along with boxplots of the posterior 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 no replication, thus a model with no interactions seems appropriate. The standard two-way Analysis of Variance 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 compete 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:

$$\sigma^{\mu} = 770884; \quad h = 0.0004161; \quad b_t^{\alpha} = b_s^{\beta} = b_u^{\gamma} = 4.008; \quad b^{\tau\alpha} = b^{\tau\beta} = b^{\tau\gamma} = 480.6.$$

Results are very similar, since when interactions are present their posterior distributions are all centred at 0 and similarly distributed, see Figure 7, panel (b). Note, however, that this need not be so: strong prior opinion on small error variances would yield more differentiated interactions. The following results are all based on the model with no interactions. Figure 7 contains, in panel (c), boxplots for the site and variety effects. The classical estimates of the main effects are all close to the central portions of the posterior distributions, even though some shrinkage is evident for the β's. The boxplots of $\log \sigma_{ij}$ in panel (d) of Figure 7 are all similar with the exception of a few which assign probability mass to rather larger values. These all correspond to observations which deviate from the sum of the main effects.

Coming to the allocations, the distribution of π^{α} is very spread: our dependent sample of 100,000 visited 64,089 different vectors. Of these, 52,302 were visited only once while 63,483 were visited less than 10 times, for a total probability of 0.85. The five most frequent vectors are

π^{α}	Frequency
1121213124131212	997
1121213124151212	411
1521213124131212	333
1521253124131212	309
1521253124161212	251

Estimates of probabilities of interest are readily derived from the output. For example, in all but 243 sampled vectors z_{10}^{α} was different from all other allocations, so that the probability that α_{10} is equal to any other level is rather small. We report, as other examples, the following estimates:

$$\mathrm{pr}(\alpha_7 = \alpha_{12} \neq \alpha_i, i \neq 7, 12) = 0.48, \qquad \mathrm{pr}(\alpha_3 = \alpha_5 = \alpha_9 = \alpha_{14} = \alpha_{16}) = 0.22,$$

$$\mathrm{pr}(\alpha_1 = \alpha_2 = \alpha_4 = \alpha_6 = \alpha_8 = \alpha_{11} = \alpha_{13} = \alpha_{15}) = 0.03.$$

The distribution of π^{β} is more concentrated. Its support included 797 vectors, 85 visited only once, 376 less than 10 times, for a total probability of 0.01. The five most probable vectors account for about 0.57 of the probability and are reported below:

π^β	1111111	1111112	2111112	3111112	1131112
Frequency	22371	21361	8619	2761	1897

An overall view of the distribution of π^β is given in Figure 8. This display is a multivariate analogue of the quantile function. On the abscissa we report the probability scale, on the ordinate the components of π^β, the same grey-scale meaning that the components are equal. The plot was created by subsampling 10% of the sampled π^β, then ordering them to produce a picture with large patches. Therefore, it contains no information concerning the mixing of the sampling chain. 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 five 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 Markov chain Monte Carlo 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 accommodate fully 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 reversible jump 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. 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 (characterised by a different value of the attribute), 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 correspondence with Dennis Lindley in the early stages of this research is gratefully acknowledged. Comments from an associate editor and two anonymous referees have led to a much improved version. The work was supported by a grant from the EPSRC initiative in Stochastic Modelling in Science and Technology, and the major part of it was completed while the first author was also at the University of Bristol.

Appendix: The reversible jump Markov chain Monte Carlo sampler

Simulation from the posterior distribution of the parameters and the latent variables is performed using the reversible jump algorithm of Green (1995), which is an extension of the method of Hastings (1970) that allows variable-dimension parameters. 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. The first group of moves consists of draws from the full conditional distributions, while the second group follows, with minor modifications, the approach of Richardson & Green (1997).

In order to write the full conditionals we need some additional notation. Let

$$y_{ij\cdot} = \frac{1}{r_{ij}} \sum_{k=1}^{r_{ij}} y_{ijk}; \qquad A_t = \{i : z_i^\alpha = t\}; \qquad m_t = \# A_t; \qquad \overline{\alpha}_t = \frac{1}{m_t} \sum_{i \in A_t} \alpha_i.$$

The following distributions are all conditional on the observed data y and the other parameters/latent variables.

$$w^\alpha \sim Dir(d_1^\alpha + m_1, \ldots, d_{k^\alpha}^\alpha + m_{k^\alpha}), \tag{11}$$

$$\mathrm{pr}(z_i^\alpha = t) = \frac{w_t^\alpha \phi(\alpha_i; \mu_t^\alpha, \sigma_t^\alpha)}{\sum_{v=1}^{k^\alpha} w_v^\alpha \phi(\alpha_i; \mu_v^\alpha, \sigma_v^\alpha)}, \tag{12}$$

$$\sigma_t^\alpha \sim I\Gamma\left\{a_t^\alpha + \frac{m_t}{2}, b_t^\alpha + \frac{1}{2}\sum_{i\in A_t}(\alpha_i - \mu_t^\alpha)^2\right\}, \tag{13}$$

$$\mu_t^\alpha \sim N\left\{\frac{\xi_t^\alpha \tau^\alpha + \overline{\alpha}_t m_t/\sigma_t^\alpha}{\tau^\alpha + m_t/\sigma_t^\alpha}, (\tau^\alpha + m_t/\sigma_t^\alpha)^{-1}\right\}, \tag{14}$$

$$\tau^\alpha \sim \Gamma\left\{a^{\tau\alpha} + \frac{k^\alpha}{2}, b^{\tau\alpha} + \frac{1}{2}\sum_{t=1}^{k^\alpha}(\mu_t^\alpha - \xi_t^\alpha)^2\right\}. \tag{15}$$

The full conditional distributions of w^β, w^γ, 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 (11)–(15) 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 (3) we would have the following full conditionals of the parameters in the systematic component of the model:

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

$$\alpha_i \sim N\left\{\frac{\mu_{z_i^\alpha}^\alpha/\sigma_{z_i^\alpha}^\alpha + \sum_j(y_{ij\cdot} - \mu - \beta_j - \gamma_{ij})r_{ij}\sigma_{ij}^{-1}}{1/\sigma_{z_i^\alpha}^\alpha + \sum_j r_{ij}\sigma_{ij}^{-1}}, \left(1/\sigma_{z_i^\alpha}^\alpha + \sum_j r_{ij}\sigma_{ij}^{-1}\right)^{-1}\right\}, \tag{16}$$

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

$$\gamma_{ij} \sim N\left\{\frac{\mu_{z_{ij}^\gamma}^\gamma/\sigma_{z_{ij}^\gamma}^\gamma + (y_{ij\cdot} - \mu - \alpha_i - \beta_j)r_{ij}\sigma_{ij}^{-1}}{1/\sigma_{z_{ij}^\gamma}^\gamma + r_{ij}\sigma_{ij}^{-1}}, \left(1/\sigma_{z_{ij}^\gamma}^\gamma + r_{ij}\sigma_{ij}^{-1}\right)^{-1}\right\}. \tag{17}$$

Simulation from the full conditionals is done subject to the constraints in (3). Note that when simulating the α's one only need to make use of $\sum \alpha_i = 0$, 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 constraints in (3) 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 (3). Rewrite (16) as $\alpha_i \sim N(\tilde{\mu}_i^\alpha, \tilde{\sigma}_i^\alpha)$ or, in matrix notation,

$$\alpha \sim N_m(\tilde{\mu}^\alpha, D^\alpha)$$

where $D^\alpha = \mathrm{diag}(\tilde{\sigma}_1^\alpha, \ldots, \tilde{\sigma}_m^\alpha)$. Let $S^\alpha = \sum \alpha_i$ and denote by 1_m a column vector of m 1's. Then the conditional distribution of α given $S^\alpha = 0$ is singular multivariate normal with covariance matrix of rank $m-1$:

$$\alpha \mid S^\alpha = 0 \ \sim \ N_m\left(\tilde{\mu}^\alpha - \frac{D^\alpha 1_m 1_m^T \tilde{\mu}^\alpha}{1_m^T D^\alpha 1_m}, D^\alpha - \frac{D^\alpha 1_m 1_m^T D^\alpha}{1_m^T D^\alpha 1_m}\right). \tag{18}$$

To simulate α, draw from the nonsingular multivariate normal of the first $m-1$ components in (18), then set α_m equal to minus their sum. Draws from the full conditional distribution of the β's are performed similarly. As for the γ's, rewrite (17) as $\gamma_{ij} \sim N(\tilde{\mu}_{ij}^\gamma, \tilde{\sigma}_{ij}^\gamma)$ or, in matrix notation,

$$\gamma \sim N_{mn}(\tilde{\mu}^\gamma, D^\gamma)$$

where $\gamma = [\gamma_{11}, \ldots, \gamma_{1n}, \ldots, \gamma_{m1}, \ldots, \gamma_{mn}]^T$ and $D^\gamma = \mathrm{diag}(\tilde{\sigma}_{11}^\gamma, \ldots, \tilde{\sigma}_{1n}^\gamma, \ldots, \tilde{\sigma}_{m1}^\gamma, \ldots, \tilde{\sigma}_{mn}^\gamma)$. Let

$$S_R^\gamma = \left(\sum_{j=1}^n \gamma_{1j}, \ldots, \sum_{j=1}^n \gamma_{mj}\right)^T = (I_m \otimes 1_n^T)\gamma, \qquad S_C^\gamma = \left(\sum_{i=1}^m \gamma_{i1}, \ldots, \sum_{i=1}^m \gamma_{i,n-1}\right)^T = (1_m^T \otimes J_n)\gamma$$

where J_n is the matrix consisting of the first $n-1$ rows of I_n. Then, the conditional distribution of γ given $S_R^\gamma = 0_m$ and $S_C^\gamma = 0_{n-1}$ is singular multivariate normal:

$$\gamma \mid S_R^\gamma = 0_m, S_C^\gamma = 0_{n-1} \ \sim \ N_{mn}\left(\tilde{\mu}^\gamma - A_{12}A_{22}^{-1}\begin{bmatrix} I_m \otimes 1_n^T \\ 1_m^T \otimes J_n \end{bmatrix}\tilde{\mu}^\gamma, \ D^\gamma - A_{12}A_{22}^{-1}A_{12}^T\right) \tag{19}$$

where

$$A_{12} = \left[D^\gamma(I_m \otimes 1_n) \ \vdots \ D^\gamma(1_m \otimes J_n^T)\right]$$

and

$$A_{22} = \begin{bmatrix} (I_m \otimes 1_n^T)D^\gamma(I_m \otimes 1_n) & (I_m \otimes 1_n^T)D^\gamma(1_m \otimes J_n^T) \\ (1_m^T \otimes J_n)D^\gamma(I_m \otimes 1_n) & (1_m^T \otimes J_n)D^\gamma(1_m \otimes J_n^T) \end{bmatrix}.$$

To simulate γ, draw the subvector $(\gamma_{11}, \ldots, \gamma_{1,n-1}, \gamma_{21}, \ldots, \gamma_{2,n-1}, \ldots, \gamma_{m-1,1}, \ldots, \gamma_{m-1,n-1})^T$ from its nonsingular marginal distribution in (19). Then, compute the remaining components as follows:

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

The reversible jump moves follow, with minor changes, the ones proposed by Richardson & Green (1997, Section 3.2). Besides the exclusive use of split/merge moves (Richardson & Green used "combine" for "merge", and 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 merge is attempted for the components in the mixtures (4), (5). 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/merge 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:

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

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:

$$\text{pr}(z_i = j_1) = \frac{p_1}{p_1 + p_2} \qquad \text{pr}(z_i = j_2) = \frac{p_2}{p_1 + p_2} \qquad i \in A_{j^*} \tag{20}$$

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

$$R = \frac{\sigma_{j^*}^{\frac{m_{j_1}+m_{j_2}}{2}}}{\sigma_{j_1}^{\frac{m_{j_1}}{2}} \sigma_{j_2}^{\frac{m_{j_2}}{2}}} \exp\left[-\frac{1}{2}\left\{\sum_{i \in A_{j_1}} \frac{(\alpha_i - \mu_{j_1})^2}{\sigma_{j_1}} + \sum_{i \in A_{j_2}} \frac{(\alpha_i - \mu_{j_2})^2}{\sigma_{j_2}} - \sum_{i \in A_{j_1} \cup A_{j_2}} \frac{(\alpha_i - \mu_{j^*})^2}{\sigma_{j^*}}\right\}\right]$$

$$
\begin{aligned}
\times & \quad \frac{p(k+1)}{p(k)} \frac{\Gamma(\sum d_j + d_{j_1} + d_{j_2})}{\Gamma(\sum d_j + d_{j^*})} \frac{\Gamma(d_{j^*})}{\Gamma(d_{j_1})\Gamma(d_{j_2})} \frac{w_{j_1}^{d_{j_1}+m_{j_1}-1} w_{j_2}^{d_{j_2}+m_{j_2}-1}}{w_{j^*}^{d_{j^*}+m_{j_1}+m_{j_2}-1}} \\
\times & \quad \left\{ \frac{\tau_{j_1}\tau_{j_2}}{(2\pi)\tau_{j^*}} \right\}^{1/2} \exp\left[-\frac{1}{2} \left\{ \tau_{j_1}(\mu_{j_1}-\xi_{j_1})^2 + \tau_{j_2}(\mu_{j_2}-\xi_{j_2})^2 - \tau_{j^*}(\mu_{j^*}-\xi_{j^*})^2 \right\} \right] \\
\times & \quad \frac{b_{j_1}^{a_{j_1}} b_{j_2}^{a_{j_2}}}{b_{j^*}^{a_{j^*}}} \frac{\Gamma(a_{j^*})}{\Gamma(a_{j_1})\Gamma(a_{j_2})} \frac{\sigma_{j^*}^{a_{j^*}+1}}{\sigma_{j_1}^{a_{j_1}+1}\sigma_{j_2}^{a_{j_2}+1}} \exp\left\{ -\frac{b_{j_1}}{\sigma_{j_1}} - \frac{b_{j_2}}{\sigma_{j_2}} + \frac{b_{j^*}}{\sigma_{j^*}} \right\} \\
\times & \quad \frac{c_{k+1}}{s_k P_{alloc}} \left\{ \frac{1}{2} g_{2,2}(u_1) g_{2,2}\left(\frac{u_2+1}{2}\right) g_{1,1}(u_3) \right\}^{-1} w_{j^*}(1-u_2^2) \left\{ \frac{\sigma_{j^*}}{u_1(1-u_1)} \right\}^{\frac{3}{2}} \qquad (21)
\end{aligned}
$$

where the $\sum d_j$ is over the indexes of the components in the mixture that are not affected by the split/merge move, P_{alloc} is the probability of the reallocations in (20), $g_{1,1}$ and $g_{2,2}$ are the densities of $Be(1,1)$ and $Be(2,2)$ distributions.

The reverse of a split is a merge. If merge 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

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

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 (21).

The above split/merge moves are designed so that the "ejected" component in a split and the "absorbed" component in a merge 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 merge 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 merge move.

References

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

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

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

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

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

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

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

Gopalan, R. & Berry, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *J. Am. Statist. Assoc.* **93**, 1130–1139.

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

Green, P. J. & Richardson, S. (1998). *Modelling heterogeneity with and without the Dirichlet process.* Submitted for publication.

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

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

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

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

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

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

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

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing* **8**, 229–242.

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

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

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

Schervish, M. J. (1995). *Theory of Statistics.* New York: 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. & Wild, P. (1996). Fitting Bayesian multiple random effects models. *Statistics and Computing* **6**, 337–346.

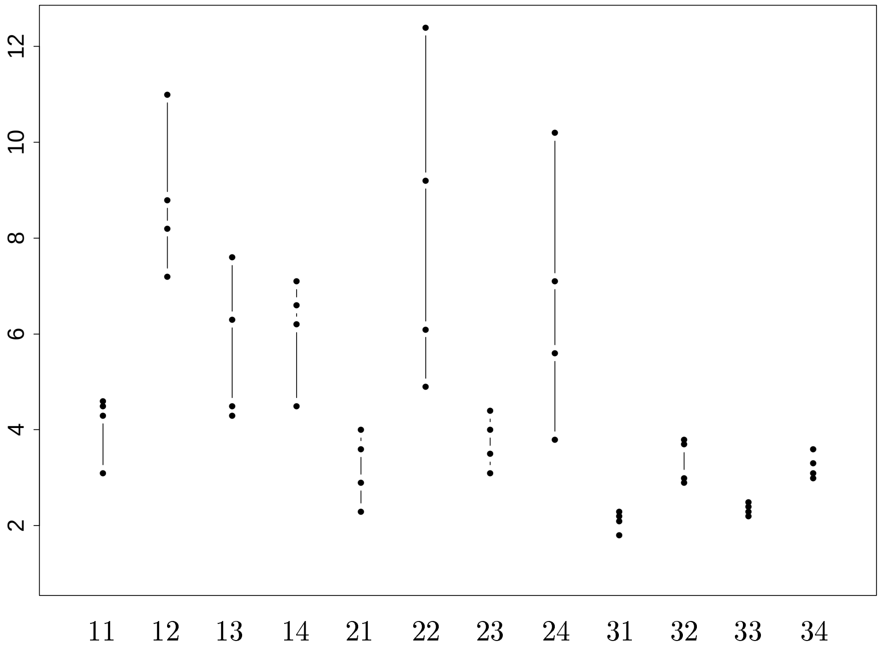

Figure 1: Survival time data set. Time in hours on the ordinates, abscissa are combinations of poisons and treatments. The only purpose of the lines is to facilitate viewing the plot "vertically".

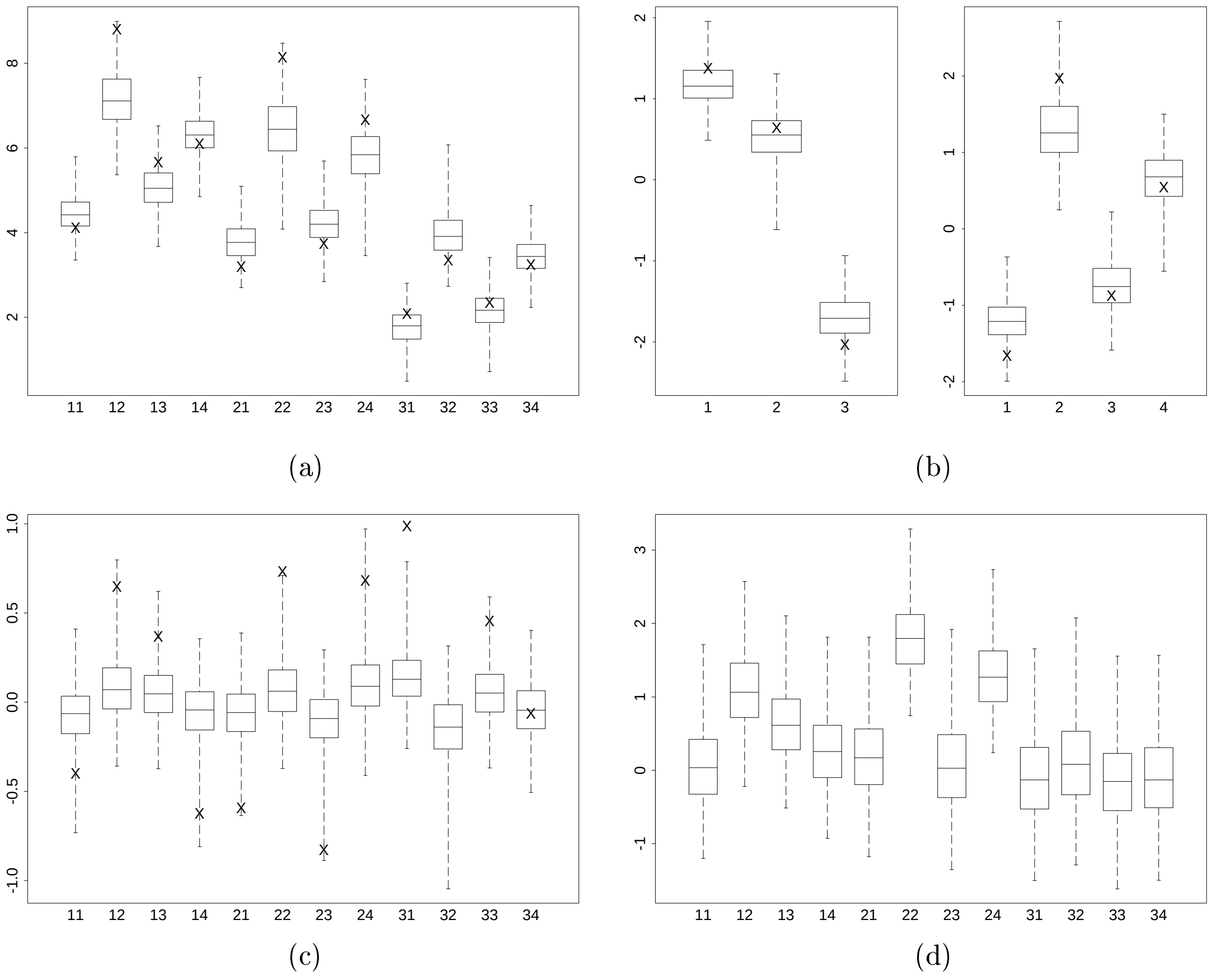

Figure 2: Survival time data set. Panel (a): boxplots of θ_{ij}'s with superimposed the cell means marked as X's. Panel (b): boxplots of α_i's and β_j's. X's are classical estimates. Panel (c): boxplots of interactions γ_{ij}'s, X's are classical estimates. Panel (d): boxplots of the logarithms of the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

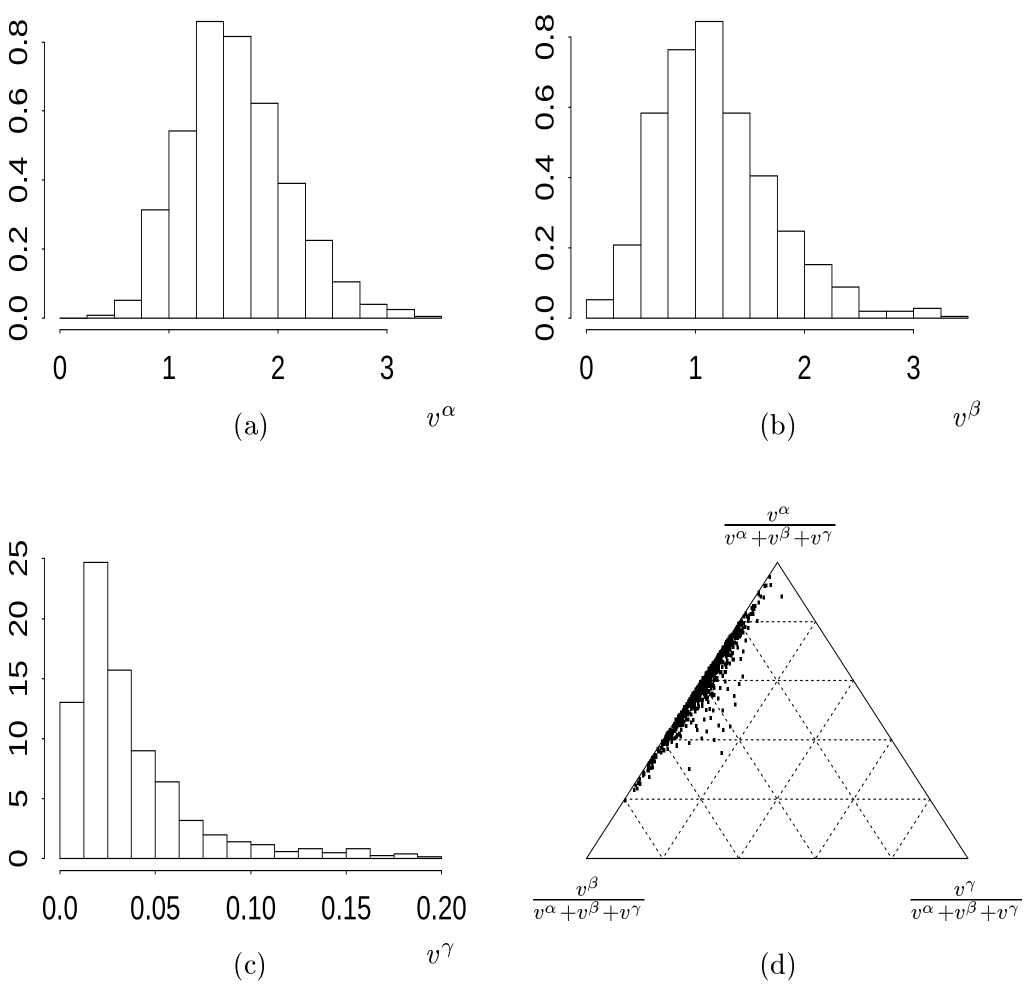

Figure 3: Survival time data set. Panels (a), (b), (c): histograms of samples from the posterior distributions of the variance components. Panel (d): trilinear plot of the posterior sample of the variance components, normalised to sum to unity.

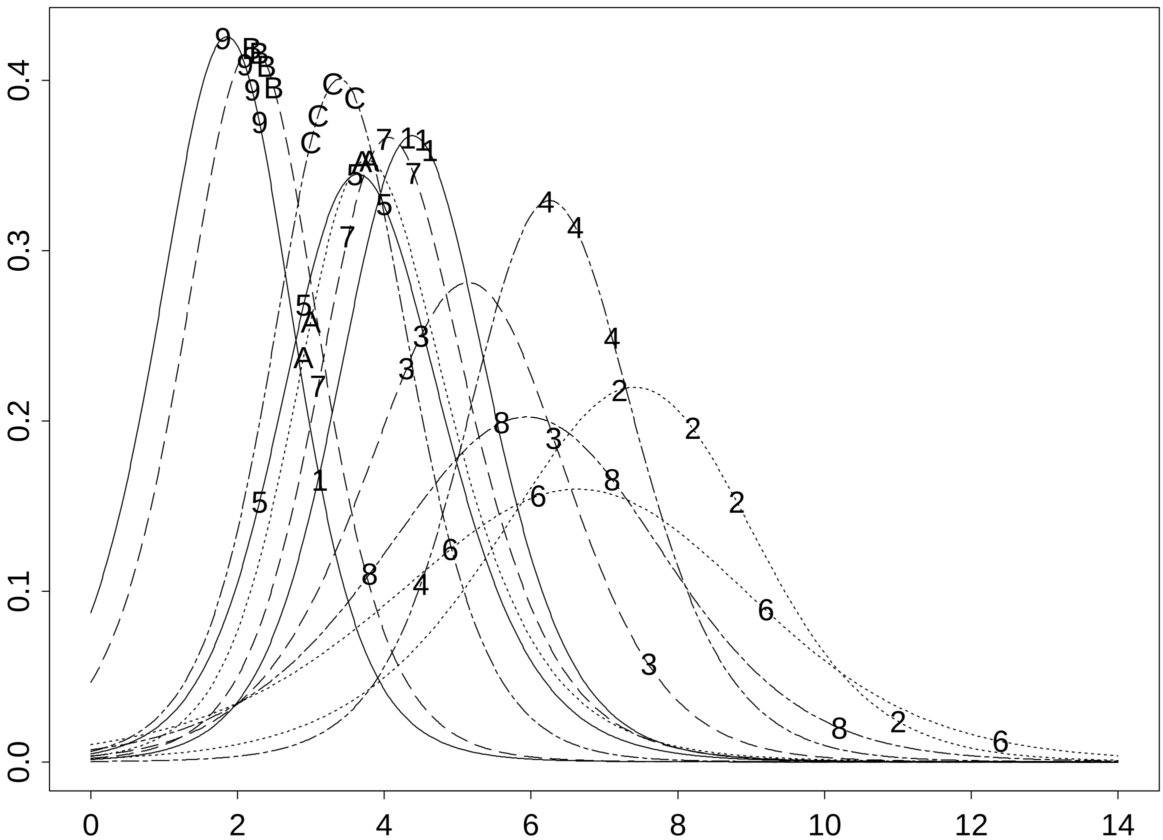

Figure 4: Survival time data set. Posterior predictive densities of the next observation conditioned on the poison/treatment combination. The labels 1,2,. . . ,9,A,B,C denote poison/treatment combination, in the same order as in Figure 1. Each predictive density has four labels on it, placed at points with abscissae equal to the observed survival times.

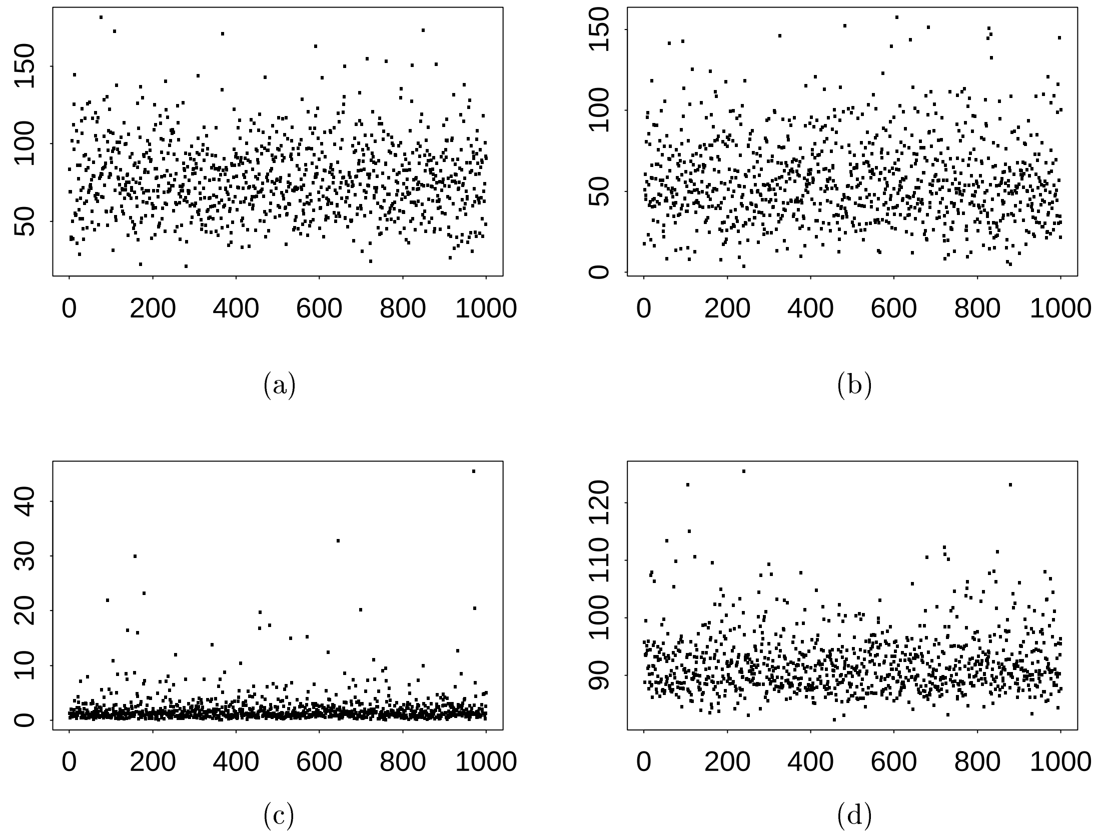

Figure 5: Survival time data set. Time series plots of the sum of squares of the α's (a), β's (b), γ's (c) and residual (d) across simulation (only a 1 in 100 subsample).

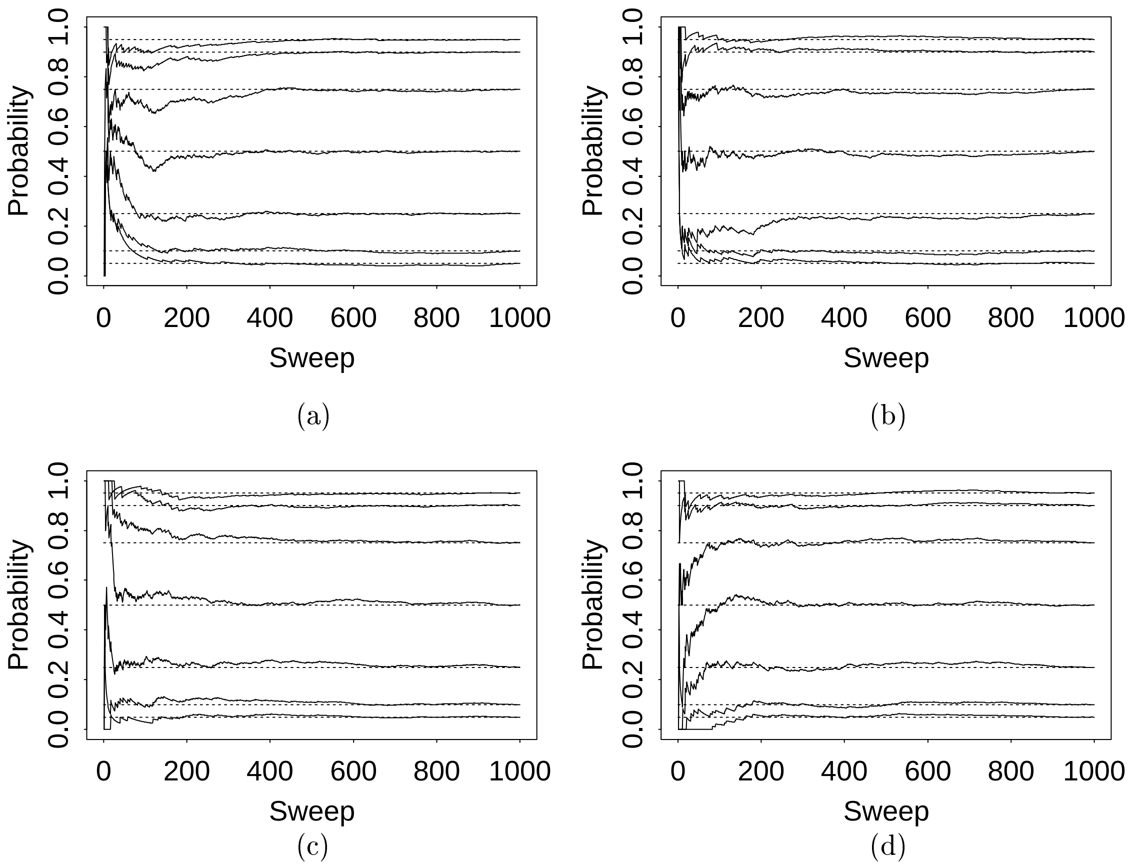

Figure 6: Survival time data set. Plots of the cumulative occupancy fractions (solid lines) for selected quantiles (dotted lines) of the Markov chain Monte Carlo sample empirical distribution of the sum of squares of α's (a), β's (b), γ's (c) and residual (d).

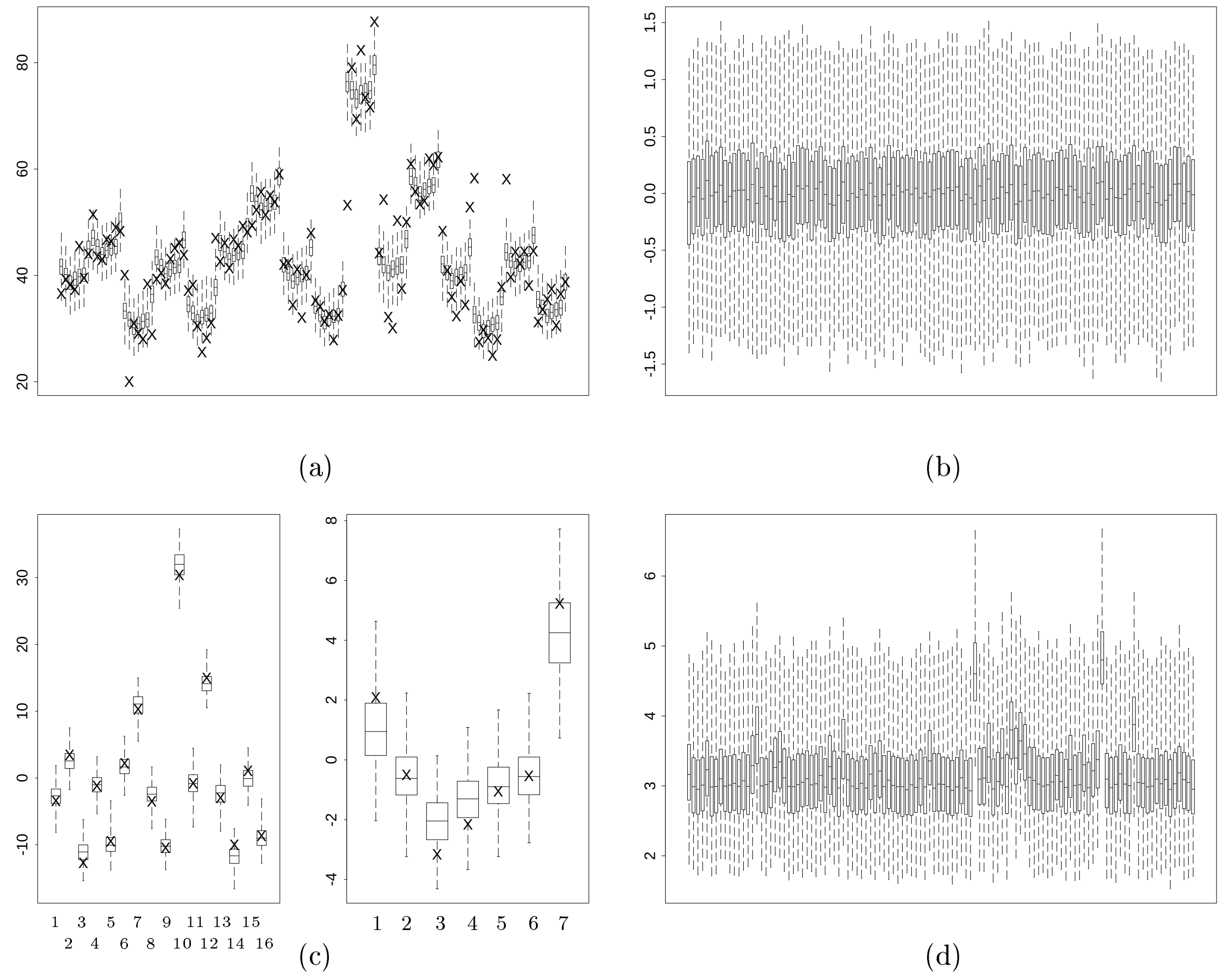

Figure 7: Potato trial data set. Panel (a): boxplots of θ_{ij}'s with superimposed the cell means marked as X's, model with no interactions. Panel (b): boxplots of γ_{ij}'s, model with interactions. Panel (c): boxplots of α_i's and β_j's, X's are classical estimates, model with no interactions. Panel (d): boxplots of the logarithms of the cell variances σ_{ij}, model with no interactions.

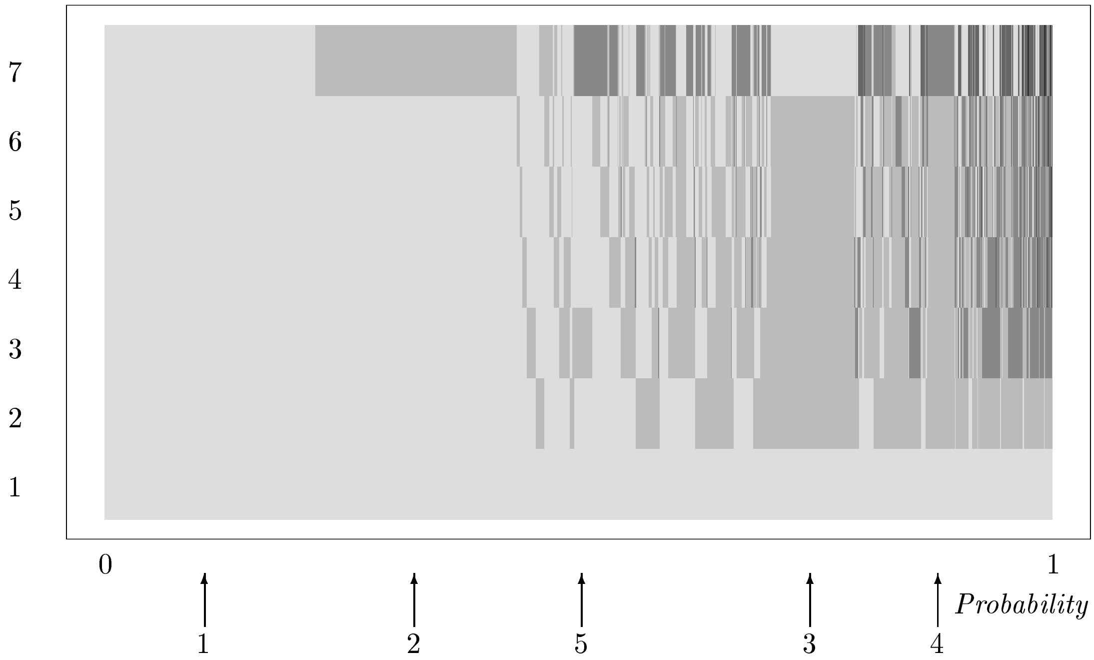

Figure 8: Potato trial data set. A graphical display of the probability distribution of π^{β}.