Statistical Methods for the Analysis of Covariance and Spatio-Temporal Models

Orestis Papasouliotis

Doctor of Philosophy

University of Edinburgh

2000

To my parents, Dimitrios and Fani Papasouliotis

Abstract

A new hierarchical model is proposed for the analysis of covariance with random, unequal variances. Bayesian inference for the new model depends upon uncertainty in the degrees of freedom for a chi-squared distribution. The convergence of MCMC is validated by extremely accurate Laplacian approximations. Practical applications include neuropsychological tests in offender profiling and nutrition data. As a special case, the parallel line model with equal variances is considered, and alternatives to the F-test for the equality of the group effects, including Bayes factors, are investigated.

In the final chapter, a self-similarity spatio-temporal model for the pressures of oil wells in an Alaskan oil field is considered, and our Bayesian techniques are extended to this situation with a view towards future interdisciplinary research.

Acknowledgements

I am grateful to my supervisor, Thomas Leonard, for his advice and support.

This thesis was motivated by Brian Yandell's lecture notes on the Analysis of Covariance at the University of Wisconsin-Madison. I would like to thank both Brian Yandell and Norman Draper for their advice and for supporting my transition from Madison to Edinburgh. I acknowledge the University of Wisconsin-Madison and the U.S. National Institutes of Health for funding my Ph.D. studies at the University of Wisconsin.

A number of people helped me complete my Ph.D. thesis by offering me data sets. They include Keith Ashcroft and Tony Busuttil who provided the Scottish prisoner data, James A. Moses Jr. with his Stanford student and medical data sets, Brian Yandell who offered the animal nutrition data, and Ian Main and Kes Heffer who provided the BP Exploration oil pressure data set. I am grateful to them all.

Colin Aitken has given substantial advice on Bayes factors. I would also like to thank Chris Theobald, who has provided useful background information on applications of MCMC, Angelika van der Linde for advice on the Kalman filter and Sergei Zatsepin for helpful discussions on self-similarity. John Hsu kindly advised me regarding Theorem 1 and equivalence between different Laplacian approximations in the literature. James Smith made available to me an initial version of the BP data, set up in Splus. Steve Law, Jessica Gaines and Noel Smyth were always available to answer my computing queries.

Finally I would like to express my gratitude to my parents, Dimitrios and Fani, and my brother, George, and to Elaine Borghi for her proof reading of the thesis, advice and support.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

I also declare that from 1993 to 1996 I was registered as a Ph.D. student at the University of Wisconsin-Madison. During this period I obtained my M.S. in Statistics and completed my course credits for a Ph.D. (15 single semester courses with G.P.A. 3.77/4, including Mathematical Statistics 709/710, taught from both Lehmann books). I also worked as a teaching and project assistant for Professor Thomas Leonard. My project assistantship has led to five publications in journals and a book on the Madison Drug and Alcohol Abuse Study. I transferred my Ph.D. studies to Edinburgh in June 1996, where I am employed as a Research Associate in the University of Edinburgh Statistical laboratory, and have completed substantial research further to my thesis topics.

(*Orestis Papasouliotis*)

Table of Contents

Chapter 1

Analysis of covariance (parallel lines model)

1.1 Introduction

The analysis of covariance (ANCOVA) method is due to Sir Ronald A. Fisher, who in the 1920's recognized the importance of accounting for additional potential sources of variation when comparing different treatments in agricultural experiments, and first described the method in 1932. The first occurrence of the term is attributed by David, (1995) to Bailey for his 1931 paper. Nowadays the idea of combining regression and analysis of variance in linear and generalized linear models is very common and applied in several different ways, (Cox and McCullagh, 1982), to increase the precision in designed experiments, to reduce the bias in observational studies, for adjustments for missing values in balanced designs and for adjustment for historical controls in clinical trials. Its analysis, that enables the comparison of factor levels while adjusting for the association between response and covariate, in situations where such a comparison is meaningful, can be found in many standard statistical texts, for example see Searle, (1987), and for a more applied perspective, Yandell, (1997).

The most simple form of the model is

$$y_{ij} = \theta_i + \beta(x_{ij} - x_{i.}) + \epsilon_{ij} \tag{1.1}$$

for $i = 1, 2, ..., m$, and $j = 1, ..., n_i$, expressing that the mean response for each group depends on the group, via θ_i, as well as the covariate, through $x_{ij} - x_{i.}$, with the two effects being additive. We use y_{ij} and x_{ij} to denote the values of the response and the

covariate respectively for the jth member of the ith group and the dot symbol to denote average with respect to the corresponding subscript. The slope β is the change in y's associated with a unit change of the covariate, and the error terms ϵ_{ij} are assumed to be independently normally distributed with mean 0 and variance σ_ϵ^2, or $\epsilon_{ij} \sim IN(0, \sigma_\epsilon^2)$.

This model is the most elementary form of a class of models appearing in the literature. Cochran (1957) and Cox and McCullagh (1982) discuss theory and applications of the fixed effects model with equal variances. Fairfield Smith (1957) gives emphasis on different interpretations of the adjusted slope and factor means assuming various relationships between the covariate and the factor. Urquhart (1982) also analyzes adjusted group means when a factor affects the covariate. Zelen (1957) discusses applications of ANCOVA models to incomplete block designs with the slopes dependent or independent on the blocks. Federer (1957) studies two-way classifications with several covariates and unequal number of observations per cell. Wilkinson (1957) presents a method of performing analysis of covariance when a set of responses is missing while the covariate information is complete. Finney (1957) examines difficulties associated with attempting to balance treatment allocation with respect to the covariates. Coons (1957) describes further techniques for missing data.

The previous authors all provided ANCOVA results related to the linear model and its various fixed effects applications. More general versions of the ANCOVA model include these of Koch, Amara, Davis and Gillings (1982), who investigate a variety of procedures for categorical data, Quade (1982), who suggests a non-parametric matching based analysis for measurement data or their ranks, Lane and Nelder (1982), who extend the fixed effects analysis to generalized linear models, Henderson (1982), who discusses random regressions, and Hendrix, Carter and Scott (1982), who assume fixed effects for different slopes. Theobald et al (1999) discuss applications of ANCOVA models with mixed and nested effects to crop variety trials. Related work in a Bayesian context, placing emphasis on combining regressions than comparing factor effects, is that of Lindley and Smith (1972), Smith (1973a), Miller and Fortney (1984) and Blattberg and George (1991).

The current research was motivated by a study of the scores of five different groups of people (three different kinds of offenders in Scottish prisons and two different kinds of controls from Stanford, U.S.A.) on twelve different neuropsychological tests. The main question of interest was whether different groups were associated with higher scores for certain tests. These tests were also thought to be related to the participants in the

study age.

Our initial primary target, was to establish whether we could find improved alternatives to the standard (F test) procedures for assessing equality of the group effects. This will ultimately lead us into an alternative type of Bayesian inference based on a continuous prior distribution.

1.2 Thesis plan

A brief introduction to the analysis of the random effects parallel lines ANCOVA model is first given in the continuation of this chapter, followed by a discussion of the standard F tests and an extensive study of the frequency properties of a number of alternative tests statistics, including Bayes factors, for testing equality of the group means. The second part of the chapter deals with the Bayesian inference for this model, further procedures for assessing equality of the means, model adequacy and a few practical testing suggestions, which are introductory to ideas presented in Chapter 2.

Chapter 2 is a purely theoretical chapter, introducing a new hierarchical analysis of covariance model with unequal variances, a presentation of MCMC procedures and their application in the full Bayesian analysis of this model, and concluding with extensions to models with many covariates.

In the third Chapter, we analyze a variety of data sets, including the complete neuropsychological test data and an animal nutrition data set, and test our Bayesian interpretations against two groups of simulated data sets.

Laplacian approximations are discussed in Chapter 4. Following two initial relatively simple examples, we present a special case of such an approximation, demonstrate how to apply it in order to achieve very accurate inferences in our proposed model, and hence help validate the convergence of the MCMC methods.

The thesis continues with Chapter 5, where a multiple response self-similarity spatio-temporal model, used for the analysis of oil well pressures, is proposed, together with extensions of our preceding Bayesian techniques to this situation. The analysis for a simpler version, single well, model via MCMC and its application on a number of oil wells is also presented. Some suggestions for future work are included in the final section of this chapter.

The final chapter of the thesis (Chapter 6) provides a brief description of our conclusions and their implications.

1.3 Classical inference for the random effects ANCOVA model

The random effects parallel line ANCOVA model, for m groups and n_i, for $i = 1, \ldots, m$, observations per group is described by the following formulation:

(a) Conditionally on the random effects, $\theta_1, \theta_2, \ldots, \theta_m$, the slope, β, and the variance, σ_ϵ^2, the observations y_{ij} are independent, for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, and normally distributed with respective means

$$\theta_i + \beta(x_{ij} - x_{i.}) \qquad (i = 1, \ldots, m), \tag{1.2}$$

and variance σ_ϵ^2, with $x_{i.}$ denoting the ith group covariate mean.

(b) Conditionally on σ_ϵ^2, μ_θ and λ_θ, the θ_i are independent and normally distributed with common mean μ_θ and variance $\sigma_\theta^2 = \lambda_\theta \sigma_\epsilon^2$, for $i = 1, \ldots, m$.

Following Box and Tiao, (1968), we introduce the unknown parameter λ_θ which measures deviations of the random effects θ_i from their mean and will be used for testing the group mean equality hypothesis.

Define

$$\begin{aligned} S_R^2 &= \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \left\{ y_{ij} - y_{i.} - \widehat{\beta}\,(x_{ij} - x_{i.}) \right\}^2 \\ &= \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 - \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - x_{i.})\,(y_{ij} - y_{i.}) \right\}^2 / s^2, \end{aligned} \tag{1.3}$$

with

$$\widehat{\beta} = \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2 \right\}^{-1} \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - x_{i.})\,(y_{ij} - y_{i.}), \tag{1.4}$$

and

$$s^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2, \tag{1.5}$$

with S_R^2 the residual sum of squares, $\widehat{\beta}$ the least squares estimate of the slope coefficient, and s^2 the within groups sum of squares for the covariate. Set also

$$N^* = N - \sum_{i=1}^m n_i^2 / N, \tag{1.6}$$

with $N = \sum_{i=1}^m n_i$, the total sample size.

The ANOVA table for this model, which forms the basis for frequentist inference, is

Table 1.1: *ANOVA table for the random effects parallel line model.*

Source of variation	SS	Expected SS
Between groups	$\sum_{i=1}^{m} n_i(y_{i.} - y_{..})^2$	$N^*\sigma_\theta^2 + (m-1)\sigma_\epsilon^2$
Regression	$\widehat{\beta}^2 s^2$	$\beta^2 s^2 + \sigma_\epsilon^2$
Residual	S_R^2	$(N-m-1)\sigma_\epsilon^2$
Total	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - y_{..})^2$	$N^*\sigma_\theta^2 + \beta^2 s^2 + (N-1)\sigma_\epsilon^2$

presented in Table 1.1. It is an extension of the ANOVA table of the random effects one-way model to include the covariate and of the standard fixed effects ANCOVA model to accommodate the random effects, and hence, slightly non standard. To derive the expected between groups sum of squares we first need to observe, taking expectations only with respect to the distribution of the data y_{ij}, that

$$E(y_{i.} - y_{..}) = \theta_i - \sum_{k=1}^{m} \frac{n_k \theta_k}{N} = u_i, \tag{1.7}$$

and

$$var(y_{i.} - y_{..}) = \sigma_\epsilon^2 \left(\frac{1}{n_i} - \frac{1}{N} \right), \tag{1.8}$$

and subsequently that

$$E\left(\sum_{i=1}^{m} n_i (y_{i.} - y_{..})^2 \right) = \sum_{i=1}^{m} n_i u_i^2 + (m-1)\sigma_\epsilon^2. \tag{1.9}$$

Since, by taking expectations with respect to the distribution of θ_i, the following two equalities hold

$$E_{\boldsymbol{\theta}}(u_i) = 0, \tag{1.10}$$

and

$$var_{\boldsymbol{\theta}}(u_i) = \left(\sum_{k=1}^{m} \frac{n_k^2}{N^2} + 1 - 2\frac{n_i}{N} \right) \sigma_\theta^2, \tag{1.11}$$

combining (1.6), (1.9), (1.10) and (1.11) provides the result appearing in Table 1.1. For the slope parameter β, we have that

$$E(y_{ij} - y_{i.}) = \beta(x_{ij} - x_{i.}), \tag{1.12}$$

and

$$cov(y_{ij} - y_{i.}, y_{kl} - y_{k.}) = \begin{cases} (1 - \frac{1}{n_i})\sigma_\epsilon^2, & \text{if } i = k,\ j = l \\ -\frac{\sigma_\epsilon^2}{n_i}, & \text{if } i = k,\ j \neq l \\ 0, & \text{if } i \neq k \end{cases} . \quad (1.13)$$

Hence by (1.4), (1.12) and (1.13), we obtain, after some algebraic manipulation,

$$E(\widehat{\beta}) = \beta, \quad (1.14)$$

and

$$var(\widehat{\beta}) = \sigma_\epsilon^2 / s^2. \quad (1.15)$$

Both previous results are identical to these of the slope of a simple linear regression model. The same holds for the expected sum of squares for the regression term. Finally, for the residual sum of squares, we have that

$$S_R^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 - \widehat{\beta}^2 s^2, \quad (1.16)$$

and by (1.12), (1.13), (1.14) and (1.15), we obtain the expectation of (1.16) to be equal to $(N - m - 1)\sigma_\epsilon^2$. This is another well known result related to linear models.

A standard way of testing the null hypothesis $H_0 : \theta_1 = \theta_2 = ... = \theta_m$, or equivalently $H_0 : \lambda_\theta = 0$, is via the F statistic,

$$F = \frac{\sum_{i=1}^{m} n_i (y_{i.} - y_{..})^2 / (m - 1)}{S_R^2 / (N - m + 1)}. \quad (1.17)$$

Under H_0, the statistic in (1.17) has an F distribution with $m - 1$ and $N - m + 1$ degrees of freedom. For the hypothesis $H_0' : \beta = 0$, the corresponding test statistic is

$$F' = \frac{\widehat{\beta}^2 s^2}{S_R^2 / (N - m + 1)}, \quad (1.18)$$

which has an F distribution with 1 and $N - m + 1$ degrees of freedom when H_0' is true. A slightly different approach is that of adjusted tests (Type III F tests) of hypotheses, that condition on all other parameters included in the model, before constructing the appropriate F test for the parameter of interest. For relevant details on the fixed effects ANCOVA model see Yandell (1997, pp. 263-269). A subsequent step, when the overall equality test provides a significant result, would be to perform individual pairwise comparisons, using Fisher's least significant differences (LSD) or some other

method, or test other required contrasts. In this situation, it may be of interest to consider the adjusted least squares means, which are defined as

$$\widehat{\xi}_i = y_{i.} - \widehat{\beta}(x_{i.} - x_{..}), \tag{1.19}$$

and permit the comparison of different group means at the same level of the covariate. This comparison, however, may not always be reasonable, see Cox and McCullagh (1982, pp. 550-551) for further details.

1.3.1 *F* test properties

The standard F statistic optimality properties (see Arnold, 1981), for testing hypotheses in the linear models depend largely on whether the data are balanced or not. Hence, for the slightly simpler one-way random effects model, the standard F statistic constitutes an optimal test (Khuri et al, 1998, p. 13), which means it is uniformly most powerful similar (UMPS), unbiased (UMPU) and invariant (UMPI). However, while the F statistic can still be employed as a fixed size test, it possesses neither of the three previous properties in the unbalanced case (unequal number of replications in different groups).

The lack of an optimal test for testing the mean equality hypothesis will motivate an investigation of the frequency properties of a number of alternative test statistics, some of them Bayesian, based on the posterior density of λ_θ, that will be presented in the following sections. For example, the F, the likelihood ratio, and Bayesian statistics can yield different significance probabilities for observed data sets. Which significance probability is appropriate? A final suggestion will be made in section 1.13, after extensive study.

1.4 Test statistics for equality of group means

We will study the Bayes factor for testing the hypothesis of interest. Bayes factors are known to be associated with problems in Bayesian interpretation, but nevertheless possess a power optimality property. Definitions and a relevant discussion will be presented in section 1.5.

We will also consider a number of other Bayesian test statistics, based on the posterior distribution of λ_θ, namely the posterior mean, the posterior mode, the posterior

median, and the posterior mean of the quantity η_θ, that will be defined later.

Firstly, however, we will obtain the likelihood ratio test statistic, a modified version of it, based on a ratio unbiased estimator of λ_θ, as well as consider the actual maximum likelihood estimator (MLE) of λ_θ as possible test statistics.

1.4.1 Likelihood ratio tests

Given the model formulation described in section 1.3, the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ conditionally on β, σ_ϵ^2, μ_θ and λ_θ is

$$\begin{aligned} p(\boldsymbol{y},\boldsymbol{\theta}|\beta,\sigma_\epsilon^2,\mu_\theta,\lambda_\theta) &= p(\boldsymbol{y}|\boldsymbol{\theta},\beta,\sigma_\epsilon^2)p(\boldsymbol{\theta}|\sigma_\epsilon^2,\mu_\theta,\lambda_\theta) \\ &= \textstyle\prod_{i=1}^m \prod_{j=1}^{n_i} (2\pi\sigma_\epsilon^2)^{-1/2} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\{y_{ij}-\theta_i-\beta(x_{ij}-x_{i.})\}^2\right] \\ &\times \textstyle\prod_{i=1}^m (2\pi\lambda_\theta\sigma_\epsilon^2)^{-1/2} \exp\left\{-\frac{1}{2\lambda_\theta\sigma_\epsilon^2}(\theta_i-\mu_\theta)^2\right\} . \end{aligned} \tag{1.20}$$

For the first exponent in (1.20) we have that

$$\sum_{i=1}^m \sum_{j=1}^{n_i} \{y_{ij}-\theta_i-\beta(x_{ij}-x_{i.})\}^2 = S_R^2 + \sum_{i=1}^m n_i(y_{i.}-\theta_i)^2 + (\widehat{\beta}-\beta)^2 s^2, \tag{1.21}$$

with S_R^2, $\widehat{\beta}$ and s^2 defined in (1.3), (1.4) and (1.5) respectively. To obtain the MLE for λ_θ we need to integrate out the random effects θ_i from (1.20) and subsequently maximize with respect to the four parameters β, σ_ϵ^2, μ_θ and λ_θ. To perform the integration, the following lemma is useful.

Lemma 1.1. Let $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ be $(p\times 1)$ vectors, and let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $(p\times p)$ symmetric matrices such that $\boldsymbol{A}+\boldsymbol{B}$ is nonsingular. Then the identity

$$(\boldsymbol{\gamma}-\boldsymbol{\alpha})^T\boldsymbol{A}(\boldsymbol{\gamma}-\boldsymbol{\alpha})+(\boldsymbol{\gamma}-\boldsymbol{\beta})^T\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\beta}) = (\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)^T(\boldsymbol{A}+\boldsymbol{B})(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)+(\boldsymbol{\alpha}-\boldsymbol{\beta})^T\boldsymbol{H}(\boldsymbol{\alpha}-\boldsymbol{\beta})$$

holds, where $\boldsymbol{\gamma}^* = (\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{A}\boldsymbol{\alpha}+\boldsymbol{B}\boldsymbol{\beta})$ and $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}$.

Proof. See Box and Tiao, 1992, pp. 418-419.

Corollary 1.1. If α, β, θ, A and B are scalars with A and $B \neq 0$, then

$$A(\theta-\alpha)^2 + B(\theta-\beta)^2 = (A+B)(\theta-\theta^*)^2 + (A^{-1}+B^{-1})^{-1}(\alpha-\beta)^2,$$

where $\theta^* = (A+B)^{-1}(A\alpha+B\beta)$.

Hence, by Corollary 1.1, for the quadratic term in θ_i of (1.20), we have that

$$n_i\,(y_{i.}-\theta_i)^2+\lambda_\theta^{-1}\,(\theta_i-\mu_\theta)^2=\left(n_i+\lambda_\theta^{-1}\right)(\theta_i-\theta_i^*)^2+\left(n_i^{-1}+\lambda_\theta\right)^{-1}(y_{i.}-\mu_\theta)^2 \quad (1.22)$$

with

$$\theta_i^*=(n_i y_{i.}+\lambda_\theta^{-1}\mu_\theta)/(n_i+\lambda_\theta^{-1}). \quad (1.23)$$

Integrating out the θ_i, using that

$$\int \exp\left\{-\frac{1}{2\sigma^2}\,(x-\mu)^2\right\}dx\propto\sigma, \quad (1.24)$$

we can obtain the likelihood of β, σ_ϵ^2, μ_θ and λ_θ, which is

$$\begin{aligned}\ell(\beta,\sigma_\epsilon^2,\mu_\theta,\lambda_\theta|\boldsymbol{y})\propto{}&(\sigma_\epsilon^2)^{-N/2}\textstyle\prod_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1/2}\\&\times\exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{S_R^2+\textstyle\sum_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1}(y_{i.}-\mu_\theta)^2+(\widehat{\beta}-\beta)^2s^2\right\}\right].\end{aligned} \quad (1.25)$$

The MLE's of β, σ_ϵ^2, μ_θ and λ_θ, under the alternative hypothesis, $H_1:\lambda_\theta\neq 0$, are obtained by maximizing (1.25) with respect to these parameters and are found to be equal to

$$\widehat{\beta}_1 = \widehat{\beta}, \quad (1.26)$$

$$\widehat{\sigma}_{\epsilon,1}^2 = \frac{S_R^2+\sum_{i=1}^m\left(n_i^{-1}+\widehat{\lambda}_\theta\right)^{-1}(y_{i.}-\widehat{\mu}_{\theta,1})^2}{N}, \quad (1.27)$$

$$\widehat{\mu}_{\theta,1} = \frac{\sum_{i=1}^m\left(n_i^{-1}+\widehat{\lambda}_\theta\right)^{-1}y_{i.}}{\sum_{i=1}^m\left(n_i^{-1}+\widehat{\lambda}_\theta\right)^{-1}}, \quad (1.28)$$

with $\widehat{\lambda}_\theta$, maximizing the profile likelihood of λ_θ, i.e.

$$\ell_p(\lambda_\theta|\boldsymbol{y})\propto\prod_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1/2}\left\{S_R^2+\sum_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1}(y_{i.}-\widehat{\mu}_{\theta,1})^2\right\}^{-N/2}. \quad (1.29)$$

The latter maximization does not have an algebraically explicit solution, so we will use a suitable iterative algorithm to perform it. Under the null hypothesis, $H_0:\lambda_\theta=0$, the MLE's of β, σ_ϵ^2, μ_θ in (1.26), (1.27) and (1.28), reduce to

$$\widehat{\beta}_0 = \widehat{\beta}, \quad (1.30)$$

$$\widehat{\sigma}^2_{\epsilon,0} = \frac{S^2_R + \sum_{i=1}^m n_i(y_{i.} - y_{..})^2}{N}, \tag{1.31}$$

$$\widehat{\mu}_{\theta,0} = y_{..}. \tag{1.32}$$

Consequently, the likelihood ratio test (LRT) statistic for testing H_0 is

$$LRT = \left(\frac{\widehat{\sigma}^2_{\epsilon,0}}{\widehat{\sigma}^2_{\epsilon,1}}\right)^{-N/2} \prod_{i=1}^m \left\{ n_i^{1/2} \left(n_i^{-1} + \widehat{\lambda}_\theta\right)^{1/2} \right\}. \tag{1.33}$$

Notice that the unbiased estimates of σ^2_θ and σ^2_ϵ, denoted by $\widetilde{\sigma}^2_\theta$ and $\widetilde{\sigma}^2_\epsilon$ respectively, can be determined using the expectations in Table 1.1. In this way, a ratio unbiased estimate of λ_θ, denoted by $\widetilde{\lambda}_\theta$, equals

$$\widetilde{\lambda}_\theta = \frac{\widetilde{\sigma}^2_\theta}{\widetilde{\sigma}^2_\epsilon} = \frac{m-1}{N^*} \max\left(F - 1, 0\right), \tag{1.34}$$

with N^* defined in (1.6), since

$$\widetilde{\sigma}^2_\theta = \frac{m-1}{N^*} \max\left\{ \frac{\sum_{i=1}^m n_i(y_{i.} - y_{..})^2}{m-1} - \widetilde{\sigma}^2_\epsilon, 0 \right\}. \tag{1.35}$$

The modified LRT statistic, which replaces $\widehat{\lambda}_\theta$ in (1.33) with $\widetilde{\lambda}_\theta$ will also be studied.

1.4.1.1 Equally replicated case

In the equally replicated case, $n_1 = n_2 = \ldots = n_m = n$, the MLE's of μ_θ under the two hypotheses coincide (they are equal to $y_{..}$) and the value of λ_θ that maximizes the profile likelihood in (1.29) can be expressed in an algebraically explicit fashion. For, since the MLE of σ^2_ϵ under H_1 is equal to

$$\widehat{\sigma}^2_\epsilon = \frac{S^2_R + (n^{-1} + \widehat{\lambda}_\theta)^{-1} \sum_{i=1}^m (y_{i.} - y_{..})^2}{N}, \tag{1.36}$$

maximizing minus twice the log of the LRT statistic with respect to λ_θ, we obtain that

$$-m + (\widehat{\sigma}^2_\epsilon)^{-1}(n^{-1} + \widehat{\lambda}_\theta)^{-1} \sum_{i=1}^m (y_{i.} - y_{..})^2 = 0, \tag{1.37}$$

and after some algebraic manipulation that

$$\widehat{\lambda}_\theta = \frac{1}{n} \max\left\{ \frac{(m-1)(N-m)}{m(N-m-1)} F - 1, 0 \right\}. \tag{1.38}$$

In this case, the *LRT* statistic is a strictly increasing function of $\widehat{\lambda}_\theta$ and hence of the F statistic. As a result, the significance probabilities and power associated with the two tests coincide. The same conclusion is valid for the relationship of the F statistic with the modified *LRT* statistic, by equation (1.34).

1.4.2 Statistics based upon the posterior density of λ_θ

The likelihood of β, σ_ϵ^2, μ_θ and λ_θ is given by expression (1.25). Assume a uniform prior distribution for β, σ_ϵ^2, μ_θ and λ_θ. Then the posterior distribution of the four parameters is

$$\pi(\beta,\sigma_\epsilon^2,\mu_\theta,\lambda_\theta|\boldsymbol{y}) \propto \ell(\beta,\sigma_\epsilon^2,\mu_\theta,\lambda_\theta|\boldsymbol{y}) \propto (\sigma_\epsilon^2)^{-N/2}\prod_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1/2} \times \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{S_R^2+\sum_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1}(y_{i\cdot}-\mu_\theta)^2+(\widehat{\beta}-\beta)^2 s^2\right\}\right]. \tag{1.39}$$

Integrating out β, using the normal integral, we have that

$$\pi(\sigma_\epsilon^2,\mu_\theta,\lambda_\theta|\boldsymbol{y}) \propto (\sigma_\epsilon^2)^{-(N-1)/2}\prod_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1/2} \times \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{S_R^2+\sum_{i=1}^m\left(n_i^{-1}+\lambda_\theta\right)^{-1}(y_{i\cdot}-\mu_\theta)^2\right\}\right]. \tag{1.40}$$

To integrate (1.40) with respect to μ_θ the following Lemma is necessary.

Lemma 1.2. Let $\boldsymbol{x_i}$, for $i = 1,2,...,m$, and $\boldsymbol{\alpha}$ be $(p \times 1)$ vectors, and let $\boldsymbol{F_i}$ for $i = 1,2,...,m$ be $(p \times p)$ matrices such that $\sum_{i=1}^m \boldsymbol{F_i}$ is nonsingular. Then the identity

$$\sum_{i=1}^m(\boldsymbol{x_i}-\boldsymbol{\alpha})^T\boldsymbol{F_i}(\boldsymbol{x_i}-\boldsymbol{\alpha}) = \sum_{i=1}^m(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}})^T\boldsymbol{F_i}(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}}) + \sum_{i=1}^m(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}})^T\boldsymbol{F_i}(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}})$$

holds, where $\widetilde{\boldsymbol{x}} = (\sum_{i=1}^m \boldsymbol{F_i})^{-1}\sum_{i=1}^m \boldsymbol{F_i}\boldsymbol{x_i}$.

Proof.

$$\begin{aligned}
&\sum_{i=1}^m(\boldsymbol{x_i}-\boldsymbol{\alpha})^T\boldsymbol{F_i}(\boldsymbol{x_i}-\boldsymbol{\alpha}) = \sum_{i=1}^m(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}}+\widetilde{\boldsymbol{x}}-\boldsymbol{\alpha})^T\boldsymbol{F_i}(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}}+\widetilde{\boldsymbol{x}}-\boldsymbol{\alpha})\\
&= \sum_{i=1}^m(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}})^T\boldsymbol{F_i}(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}}) + \sum_{i=1}^m(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}})^T\boldsymbol{F_i}(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}}) - 2(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}})^T\sum_{i=1}^m\boldsymbol{F_i}(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}})\\
&= \sum_{i=1}^m(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}})^T\boldsymbol{F_i}(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}}) + \sum_{i=1}^m(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}})^T\boldsymbol{F_i}(\boldsymbol{x_i}-\widetilde{\boldsymbol{x}}) - 2(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}})^T\sum_{i=1}^m\boldsymbol{F_i}\boldsymbol{x_i}\\
&+ 2(\boldsymbol{\alpha}-\widetilde{\boldsymbol{x}})^T\sum_{i=1}^m\boldsymbol{F_i}\left(\sum_{i=1}^m\boldsymbol{F_i}\right)^{-1}\sum_{i=1}^m\boldsymbol{F_i}\boldsymbol{x_i}
\end{aligned}$$

$$= \sum_{i=1}^{m} (\boldsymbol{\alpha} - \widetilde{\boldsymbol{x}})^T \boldsymbol{F_i} (\boldsymbol{\alpha} - \widetilde{\boldsymbol{x}}) + \sum_{i=1}^{m} (\boldsymbol{x_i} - \widetilde{\boldsymbol{x}})^T \boldsymbol{F_i} (\boldsymbol{x_i} - \widetilde{\boldsymbol{x}}) .$$

Corollary 1.2. If x_i, f_i, for $i = 1, 2, ..., m$, and α are scalars, with $\sum_{i=1}^{m} |f_i| \neq 0$, then

$$\sum_{i=1}^{m} f_i (x_i - \alpha)^2 = \sum_{i=1}^{m} f_i (\alpha - \widetilde{x})^2 + \sum_{i=1}^{m} f_i (x_i - \widetilde{x})^2 ,$$

where $\widetilde{x} = \sum_{i=1}^{m} f_i x_i / \sum_{i=1}^{m} f_i$.

Applying Corollary 1.2, we directly derive that

$$\sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1} (y_{i.} - \mu_\theta)^2 = \sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1} (y_{i.} - \bar{y}_\theta)^2 + \sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1} (\bar{y}_\theta - \mu_\theta)^2 , \tag{1.41}$$

with

$$\bar{y}_\theta = \frac{\sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1} y_{i.}}{\sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1}}. \tag{1.42}$$

A second integration using the normal integral provides that

$$\pi(\sigma_\epsilon^2, \lambda_\theta, |\boldsymbol{y}) \propto (\sigma_\epsilon^2)^{-(N-2)/2} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(S_R^2 + \mathrm{A}_\theta\right)\right\} \Upsilon(\lambda_\theta), \tag{1.43}$$

with

$$\Upsilon(\lambda_\theta) = \prod_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1/2} \left\{\sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1}\right\}^{-1/2}, \tag{1.44}$$

and

$$\mathrm{A}_\theta = \sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1} (y_{i.} - \bar{y}_\theta)^2 . \tag{1.45}$$

The previous density has, as a function of σ_ϵ^2, the same parametric form as an inverted chi-squared density with $N-4$ degrees of freedom and scale parameter $(S_R^2 + \mathrm{A}_\theta)/(N-4)$. Using this observation, the integration with respect to the variance σ_ϵ^2 can be performed. Hence the posterior density of λ_θ is

$$\pi(\lambda_\theta|\boldsymbol{y}) \propto \left\{S_R^2 + \sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1} (y_{i.} - \bar{y}_\theta)^2\right\}^{-(N-4)/2} \Upsilon(\lambda_\theta). \tag{1.46}$$

Using (1.46), it is possible to perform a numerical integration in order to obtain the constant of proportionality and, subsequently, use another one to evaluate the posterior mean of λ_θ. Additionally we can maximize (1.46) numerically to obtain its

posterior mode, or combine numerical integration and linear interpolation to compute its posterior median.

Another quantity we will consider as a possible test statistic for equality of the group means is the posterior mean of

$$\eta_\theta = m^{-1} \sum_{i=1}^{m} \frac{n_i}{n_i + \lambda_\theta^{-1}}. \tag{1.47}$$

Since, as we show later, in (1.67), we have that

$$E(\theta_i|\lambda_\theta, \boldsymbol{y}) = \frac{n_i}{n_i + \lambda_\theta^{-1}} y_{i\cdot} + \left(1 - \frac{n_i}{n_i + \lambda_\theta^{-1}}\right) \bar{y}_\theta, \tag{1.48}$$

$E(\eta_\theta|\boldsymbol{y}) \leq 0.5$ suggests equality of the θ_i, while $E(\eta_\theta|\boldsymbol{y}) > 0.5$, inequality, though this specific comparison with 0.5 incorporates the concept of practical significance and is not a fixed size test for statistical significance.

1.4.2.1 Equally replicated case

In the equally replicated case, $\bar{y}_\theta$ in (1.42) reduces to $y_{\cdot\cdot}$, and hence the posterior density of λ_θ can be rewritten as

$$\pi(\lambda_\theta|\boldsymbol{y}) \propto \left\{ S_R^2 + (n^{-1} + \lambda_\theta)^{-1} \sum_{i=1}^{m} (y_{i\cdot} - y_{\cdot\cdot})^2 \right\}^{-(N-4)/2} (n^{-1} + \lambda_\theta)^{-(m-1)/2}. \tag{1.49}$$

Using the transformation $z = (m-1)F/\{(m-3)(n^{-1} + \lambda_\theta)n\}$, with F the standard F statistic, defined in (1.17), we obtain that the posterior distribution of z satisfies

$$\pi(z|\boldsymbol{y}) \propto \left(1 + \frac{m-3}{N-m-1} z\right)^{-(N-4)/2} z^{(m-3)/2-1}, \tag{1.50}$$

and hence z has an F distribution with $m-3$ and $N-m-1$ degrees of freedom, but truncated at $z = (m-1)F/(m-3)$. We assume here a uniform $(0, \infty)$ prior for λ_θ. Notice that, under $H_0 : \lambda_\theta = 0$, z is equal to the standard F statistic multiplied by the adjustment $(m-1)/(m-3)$ due to our choice of prior distribution. The final test statistic we consider is the Bayes factor, with a detailed discussion of its properties and characteristics following immediately.

1.5 Bayes factors

Definition 1.1. Given two hypotheses, H_0 and H_1, the Bayes factor in favour of H_0 is given by the ratio of the posterior to the prior odds of the two hypotheses, or if models M_0 and M_1 correspond to H_0 and H_1 for data $\boldsymbol{y}$,

$$\text{Bayes factor} = \frac{p(\boldsymbol{y}|M_0)}{p(\boldsymbol{y}|M_1)} = \frac{p(M_0|\boldsymbol{y})}{p(M_1|\boldsymbol{y})} \Big/ \frac{p(M_0)}{p(M_1)}, \tag{1.51}$$

(Bernardo and Smith, 1994, p.390).

Hence, the Bayes factor, which is first attributed to Jeffreys (1935), is equivalent to the ratio of the integrated likelihoods. It reduces to the likelihood ratio when the two hypotheses are simple and requires an integration over the parameter space when one of the hypotheses is composite. In this way, it differs from the likelihood ratio statistic that involves maximization with respect of the nuisance parameters. Unlike classical significance tests, the Bayes factor quantifies the evidence in favour of a null hypothesis and can be used for non-nested alternative hypotheses without theoretical complications.

By its definition, comparing the Bayes factor with the neutral value of 1 can be used as a first indication of whether the data favour the null hypothesis. Kass and Raftery (1995), slightly modifying earlier suggestions by Jeffreys (1961), provide a table of ranges of values of the Bayes factor and the corresponding strength of evidence. However, their guidelines appear not to be universally accepted.

There are a number of difficulties associated with the Bayes factor. The first obvious one is its dependence on the prior distribution. O'Hagan (1994) discusses additional difficulties associated with flat prior distributions. In particular, for testing simple versus composite hypotheses, a proper prior of increasing range results in increasing values of the Bayes factor. Taking this situation to the limit, by assuming an improper uniform prior on the real line, would give a value of the Bayes factor equaling ∞, with the obvious difficulty in interpreting it. Hence, the Bayes factors based on improper priors cannot be interpreted using a standard table, as described above, since the normalizing constants are missing.

Berger and Mortera (1999) discuss the properties of a number of variations of the Bayes factor used in the absence of subjective prior information, like the fractional Bayes factor (O'Hagan, 1995) and the intrinsic Bayes factor (Berger and Perrichi, 1996), which

present their own challenges as well. For example, the fractional Bayes factor uses a part of the data to obtain the posterior distribution of the quantity of interest. This posterior distribution subsequently plays the role of a (proper) prior for the evaluation of the Bayes factor based on the remaining of the data, with the obvious question of how to choose the training part of the data for the first step of that application.

Lindley (1957) discusses the statistical paradox of rejecting a null hypothesis at level of significance α, while the posterior probability that the same hypothesis is true is as high as $100(1-\alpha)\%$ for sufficiently large sample size, under a uniform prior, in which case the posterior probability of interest is an increasing function of the Bayes factor. This situation, called Lindley's paradox, was studied by numerous authors since then; see Bernardo and Smith (1994, p. 394). Atkinson (1978) discusses further problems when interpreting Bayes factors, that can lead to counterintuitive properties.

Despite the difficulties in its interpretation, the Bayes factor is associated with an optimality property discussed in the following sections.

1.5.1 A Bayes factor optimality property

Definition 1.2. For any set A, let I_A be its indicator function defined by

$$I_A(x) = 1 \quad \text{or} \quad 0, \qquad \text{as} \quad x \in A \quad \text{or} \quad x \notin A. \tag{1.52}$$

Crook and Good (1982) present the definition of the strength of a test and a Neyman-Pearson type theorem for testing a simple null hypothesis. They adapt the standard proof (e.g. Lehmann, 1994, pp. 74-76) of the Neyman-Pearson lemma. A generalization of that definition and theorem, covering the composite null hypothesis case, is presented below.

Definition 1.3. Let $\boldsymbol{y}=(y_1,\ldots,y_N)^T$ an $(N \times 1)$ vector of observations possessing specified sampling density or probability mass function $f(\boldsymbol{y}|\boldsymbol{\phi})$, given an unknown vector of parameters $\boldsymbol{\phi} = (\boldsymbol{\phi_1}, \boldsymbol{\phi_2})^T$, for $\boldsymbol{y} \in \mathcal{R}^N$, $\boldsymbol{\phi_1} \in \Phi_1 \subseteq \mathcal{R}^{p_1}$ and $\boldsymbol{\phi_2} \in \Phi_2 \subseteq \mathcal{R}^{p_2}$. Consider the composite null hypothesis $H_0 : \boldsymbol{\phi_1} = \boldsymbol{\phi_1^0}$ versus the alternative hypothesis $H_1 : \boldsymbol{\phi_1} \neq \boldsymbol{\phi_1^0}$. Let $\Phi = \Phi_1 \times \Phi_2$, G denote a probability distribution concentrated on Φ, Φ_0 the subset of Φ for which $\boldsymbol{\phi_1} = \boldsymbol{\phi_1^0}$ and Φ_0^c the complement of Φ_0. Then the strength or average power (against G) of a significance test of size α, is

$$\beta_G^* = \underset{\boldsymbol{\phi}|G}{E}\, I_{\Phi_0^c}(\boldsymbol{\phi})\beta^*(\boldsymbol{\phi}) \tag{1.53}$$

where $\beta^*(\boldsymbol{\phi})$ is the power function of the test, and the expectation is with respect to a random vector $\boldsymbol{\phi}$ possessing distribution G.

Note that in the previous definition, as the null hypothesis is composite, we define the size α of the test to itself be an average, that is

$$\alpha = \underset{\boldsymbol{\phi}|G}{E}\ I_{\Phi_0}(\boldsymbol{\phi})\beta^*(\boldsymbol{\phi}). \tag{1.54}$$

This differs from the standard definition that uses a supremum. The current definition might be regarded as more appealing. The decision theoretic proof of the following theorem extends an idea by John Hsu, and a proof, related to two simple hypotheses, reported by Rice (1988, pp. 524-525).

Theorem 1. Consider the test ψ^* which accepts H_0, whenever

$$\lambda(\boldsymbol{y}) > K, \tag{1.55}$$

where

$$\lambda(\boldsymbol{y}) = \frac{\underset{\boldsymbol{\phi}|G}{E}\ I_{\Phi_0}(\boldsymbol{\phi})f(\boldsymbol{y}|\boldsymbol{\phi_1},\boldsymbol{\phi_2})}{\underset{\boldsymbol{\phi}|G}{E}\ I_{\Phi_0^c}(\boldsymbol{\phi})f(\boldsymbol{y}|\boldsymbol{\phi_1},\boldsymbol{\phi_2})} \tag{1.56}$$

and K is determined by the size α of the test. Then ψ^* maximizes the strength against G among all tests of size α.

Proof. Let π, $(\pi \neq 0)$, and π^* denote respectively the prior and posterior probabilities that H_0 is correct. Applying Bayes' theorem, we find that

$$\pi^* = \frac{\pi \underset{\boldsymbol{\phi}|G}{E}\ I_{\Phi_0}(\boldsymbol{\phi})f(\boldsymbol{y}|\boldsymbol{\phi_1},\boldsymbol{\phi_2})}{\pi \underset{\boldsymbol{\phi}|G}{E}\ I_{\Phi_0}(\boldsymbol{\phi})f(\boldsymbol{y}|\boldsymbol{\phi_1},\boldsymbol{\phi_2}) + (1-\pi)\underset{\boldsymbol{\phi}|G}{E}\ I_{\Phi_0^c}(\boldsymbol{\phi})f(\boldsymbol{y}|\boldsymbol{\phi_1},\boldsymbol{\phi_2})} = \frac{\pi\lambda(\boldsymbol{y})}{\pi\lambda(\boldsymbol{y}) + 1 - \pi}. \tag{1.57}$$

Also consider the usual zero-one loss function, defined by

$$\begin{aligned} L(H_0,\boldsymbol{\phi}) &= 0 \quad \text{for} \quad \boldsymbol{\phi_1} = \boldsymbol{\phi_1^0}, & L(H_0,\boldsymbol{\phi}) &= 1 \quad \text{for} \quad \boldsymbol{\phi_1} \neq \boldsymbol{\phi_1^0}, \\ L(H_1,\boldsymbol{\phi}) &= 1 \quad \text{for} \quad \boldsymbol{\phi_1} = \boldsymbol{\phi_1^0}, \quad \text{and} & L(H_1,\boldsymbol{\phi}) &= 0 \quad \text{for} \quad \boldsymbol{\phi_1} \neq \boldsymbol{\phi_1^0}. \end{aligned} \tag{1.58}$$

Then

$$\underset{\boldsymbol{\phi}|\boldsymbol{y},G}{E}\ L(H_0,\boldsymbol{\phi}) = 1 - \pi^*, \qquad \text{and} \qquad \underset{\boldsymbol{\phi}|\boldsymbol{y},G}{E}\ L(H_1,\boldsymbol{\phi}) = \pi^*. \tag{1.59}$$

The test that accepts H_0 when $\pi^* > 1/2$ is Bayes, (Casella and Berger, 1990, theorem 10.3.3, p. 477). Hence the Bayes rule accepts H_0 when $\lambda(\boldsymbol{y}) > 1/\pi - 1$, which means that ψ^* is a Bayes test against G with $K = 1/\pi - 1$. Notice that π does not need to be defined in advance, but is defined once K is obtained based on the size of the test.

The risk function of any test ψ with size α and power function $\beta(\boldsymbol{\phi})$ satisfies $r_\psi(\boldsymbol{\phi}) = 1 - \beta(\boldsymbol{\phi})$, for $\boldsymbol{\phi_1} \neq \boldsymbol{\phi}_1^0$, and averages α on Φ_0. The average risk of test ψ under the prior distribution G on $\boldsymbol{\phi}$ is

$$R_\psi = \pi\alpha + (1-\pi)(1-\beta_G), \tag{1.60}$$

where β_G is the strength of ψ. Then, if R_ψ^* is the average risk of ψ^*,

$$\begin{aligned} R_\psi^* &= \pi\alpha + (1-\pi)(1-\beta_G^*) \\ &\leq \pi\alpha + (1-\pi)(1-\beta_G), \end{aligned} \tag{1.61}$$

since ψ^* is Bayes. Consequently $\beta_G^* > \beta_G$ and the proof is complete.

The immediate consequence of the previous lemma is that the Bayes factor, as a test statistic, must have better power properties than F for some values of λ_θ in the parameter space. Also, by changing the interval upper bound of the uniform prior of λ_θ, it is possible to improve the power in areas of particular interest in the parameter space. Hence, from a Fisherian point of view, the prior distribution need not be based upon prior beliefs, but rather upon the required power properties.

Crook and Good (1982) compared the power function of the Bayes factor with the power functions of three other statistics, the likelihood ratio, the chi-squared and the type II likelihood (hyperlikelihood) ratio, the distributions of which are studied in detail in Good and Crook (1974), for testing equiprobability for multinomial distributions and association for contingency tables. The results found confirmed Theorem 1, however the differences in strength were small, all less than 1%, as the authors expected, due to the functional relationships between the four test statistics.

For the hypothesis $H_0 : \lambda_\theta = 0$, the Bayes factor (BF) is defined as

$$BF = \frac{p(\boldsymbol{y}|\lambda_\theta = 0)}{p(\boldsymbol{y})} = \left.\frac{\pi(\lambda_\theta|\boldsymbol{y})}{\pi(\lambda_\theta)}\right|_{\lambda_\theta=0}, \tag{1.62}$$

with $\pi(\lambda_\theta)$ and $\pi(\lambda_\theta|\boldsymbol{y})$ denoting the prior and posterior densities of λ_θ respectively. Assuming a uniform $(0,1)$ prior for λ_θ, corresponding to a prior sample size of at least one, the Bayes factor becomes

$$BF = \pi(\lambda_\theta = 0|\boldsymbol{y}), \tag{1.63}$$

and hence it is equal to the value of the full posterior density of λ_θ evaluated at 0. Assume that the posterior density of λ_θ, including the normalizing constant, has already been computed under the uniform (0,1) prior and it is required to obtain the Bayes factor under a restricted range for the uniform prior, e.g. $0 < \lambda_\theta < \lambda$, over which optimal strength is required. Then, if we denote by $\pi(\lambda_\theta|\boldsymbol{y})$ and $\widetilde{\pi}(\lambda_\theta|\boldsymbol{y})$ the posterior density of λ_θ under the former and latter priors respectively, and BF and $\widetilde{BF}$ the corresponding values of the Bayes factor, since

$$\widetilde{\pi}(\lambda_\theta|\boldsymbol{y}) = \frac{\pi(\lambda_\theta|\boldsymbol{y})}{\int_0^\lambda \pi(\lambda_\theta|\boldsymbol{y})d\lambda_\theta}, \qquad \text{for} \quad 0 < \lambda_\theta < \lambda, \tag{1.64}$$

the following equalities hold

$$\widetilde{BF} = \frac{\widetilde{\pi}(\lambda_\theta|\boldsymbol{y})}{\widetilde{\pi}(\lambda_\theta)}\Big|_{\lambda_\theta=0} = \frac{\pi(\lambda_\theta = 0|\mathbf{y})}{\lambda^{-1}\int_0^\lambda \pi(\lambda_\theta|\boldsymbol{y})d\lambda_\theta} = \frac{BF}{\lambda^{-1}\int_0^\lambda \pi(\lambda_\theta|\boldsymbol{y})d\lambda_\theta}. \tag{1.65}$$

Hence the Bayes factor under the second prior can be computed with negligible further computational effort, given the fact that a one dimensional numerical integration had already been used for the calculation of the normalizing constant of $\pi(\lambda_\theta|\boldsymbol{y})$.

1.6 A simulation study

Throughout this chapter we will be using a small part of the neuropsychological study data, that motivated our research, to illustrate properties of the model we are studying and the different test statistics. We also produced three additional data sets by altering the observed group means, so that our simulation study covers a broad range of significance probabilities. We selected small sample sizes to enhance possible power differences. A more complete analysis of the data, with more general forms of the ANCOVA model will follow in Chapters 3 and 4.

The data sufficient statistics are presented in Table 1.2 and correspond to sample sizes $n_1 = 8$, $n_2 = n_3 = 5$, $n_4 = 4$ and $n_5 = 14$.

Table 1.2: *Sufficient statistic values for the four data sets.*

Sufficient statistics	Data set 1	2	3	4
$y_{1.}$	6.375	16.625	5.000	6.875
$y_{2.}$	3.800	19.200	4.000	3.800
$y_{3.}$	5.400	13.400	6.800	5.600
$y_{4.}$	3.000	4.750	0.250	2.750
$y_{5.}$	8.643	19.429	7.000	8.786
$\widehat{\beta}$	-0.902	-0.608	0.404	-0.074
S_R^2	301.976	2025.408	517.307	681.704

We proceed by presenting the proposed test statistic values and significance probabilities for the four data sets in Tables 1.3 and 1.4.

Table 1.3: *Test statistic values for the four data sets.*

Statistic	Data set 1	Data set 2	Data set 3	Data set 4
F	3.8612	2.7979	2.3596	1.8802
MLE of λ_θ	0.2965	0.1910	0.1254	0.0899
$-2\log(LRT)$	4.4528	1.2219	0.7808	0.7376
$-2\log($"Modified" $LRT)$	4.3133	1.1591	0.6822	0.6780
Posterior Mean	0.5066	0.4643	0.4300	0.3863
Posterior Mode	0.3499	0.2530	0.1741	0.1196
Posterior Median	0.4920	0.4406	0.3935	0.3337
$\mathrm{E}(\eta_\theta \mid \mathbf{y})$	0.7089	0.6754	0.6517	0.6217
$-2\log($Bayes Factor$)$, $\lambda_\theta \sim U(0,1)$	4.0662	1.1126	0.4278	-0.0621
$-2\log($Bayes Factor$)$, $\lambda_\theta \sim U(0,0.5)$	4.1043	1.3632	0.8482	0.5557

For the likelihood ratio test, we did not employ the well known asymptotic result stating that the distribution of minus twice the logarithm of the ratio of the likelihood under the null and the alternative hypothesis has a chi-squared distribution with, in our case for $H_0 : \lambda_\theta = 0$, one degree of freedom. Rather, and similarly with all the other test statistics studied, apart from the F statistic, we found the true test statistic 0.95 and 0.99 percentiles, under the null hypothesis, based on Monte Carlo simulation. For computational economy, we performed all posterior density calculations assuming a uniform (0,1) prior on λ_θ. The maximizations required for the posterior mode and the MLE of λ_θ were performed by the regula falsi (false position) method. For the numerical integrations, Simpson's rule was used, in an adaptive manner, successively

Table 1.4: *Significance probabilities for the four data sets.*

Statistic	Data set 1	Data set 2	Data set 3	Data set 4
F	0.01192	0.04375	0.07573	0.13961
MLE of λ_θ	0.01764	0.04484	0.08261	0.11678
$-2\log(LRT)$	0.00719	0.06360	0.09238	0.09610
$-2\log$("Modified" LRT)	0.00742	0.06366	0.09561	0.09598
Posterior Mean	0.01729	0.04072	0.07498	0.15086
Posterior Mode	0.01961	0.04168	0.07949	0.12736
Posterior Median	0.01839	0.04002	0.07472	0.15166
$\mathrm{E}(\eta_\theta \mid \mathbf{y})$	0.01500	0.04297	0.07686	0.14191
$-2\log$(Bayes Factor), $\lambda_\theta \sim U(0,1)$	0.00776	0.05371	0.08350	0.11477
$-2\log$(Bayes Factor), $\lambda_\theta \sim U(0,0.5)$	0.00667	0.05758	0.08589	0.10784

doubling the grid of points employed, until the normalization constant's five first decimal digits stabilized. One million Monte Carlo simulations were used to obtain the significance probabilities and power values. Despite the reduced range we assumed for λ_θ, we found that the integrations for the posterior values were very time consuming. This has to be combined with the fact that additional simulations are needed for every combination of sample sizes, in order to obtain critical values for different statistics, to realize at least the convenience provided by the standard F statistic.

Table 1.3 demonstrates that different test statistics can provide different significance probabilities. The between statistic variability is increasing when being further away from the rejection of the null hypothesis. For data set 4, for example, the maximum difference was about 6% while for data set 1, which presents more evidence of group mean inequality, the same difference was only 1.3%. Despite the previous fact, it is clear that for the first two data sets, the significance probabilities can lie on both sides of the standard 5% or 1% cutoffs for both classical and Bayesian test statistics.

Whether the cutoff is the 5% or the 1% level of significance, the decision whether to accept the null hypothesis of the equality of the group means could potentially depend on what test statistic one uses. This fact combined with the lack of optimality of the F statistic makes the subsequent power function analysis even more important.

1.6.1 Power and conditional power

We selected the Bayes factor, the F statistic, the likelihood ratio statistic, the posterior mode and the Bayes factor under uniform $(0, 0.5)$ prior, in order to avoid pairs of test

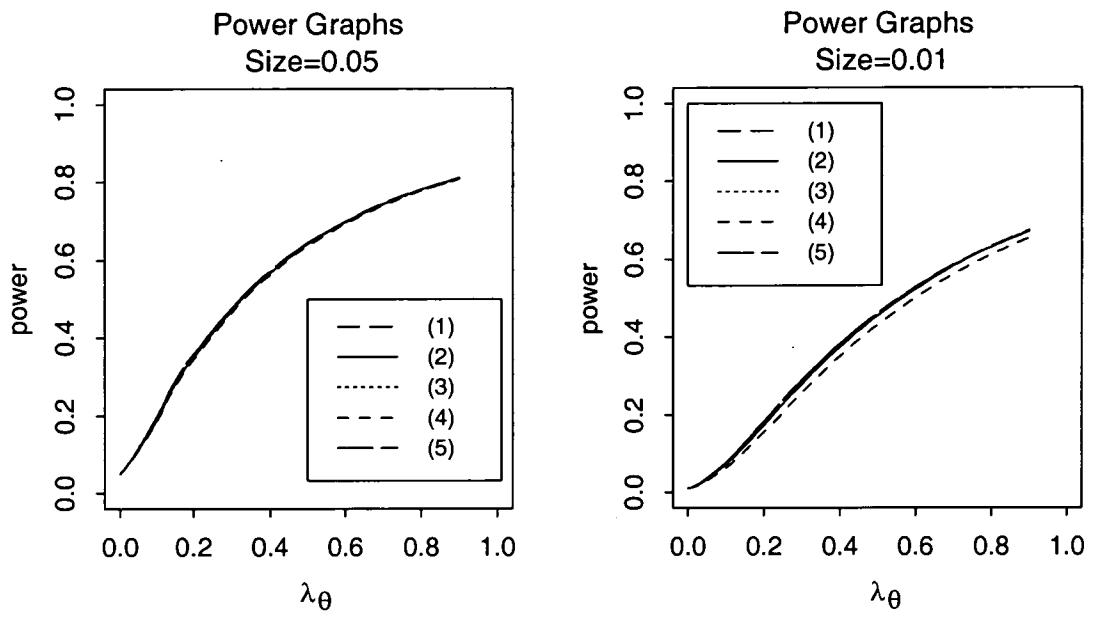


Figure 1.1: Power function for (1) Bayes factor, (2) F statistic, (3) likelihood ratio statistic, (4) posterior mode, (5) Bayes factor under uniform $(0, 0.5)$ prior. Left: 0.05 test size. Right: 0.01 test size.

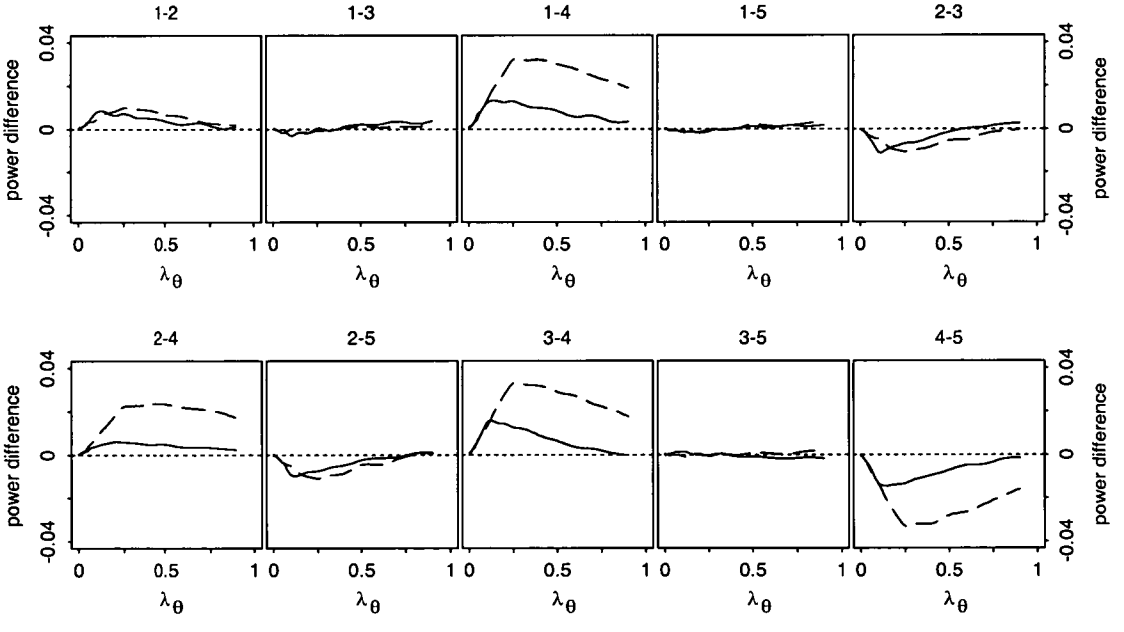


Figure 1.2: Pairwise power function differences for (1) Bayes factor, (2) F statistic, (3) likelihood ratio statistic, (4) posterior mode, (5) Bayes factor under uniform $(0, 0.5)$ prior. The solid lines correspond to 0.05 size tests. The dashed lines correspond to 0.01 size tests.

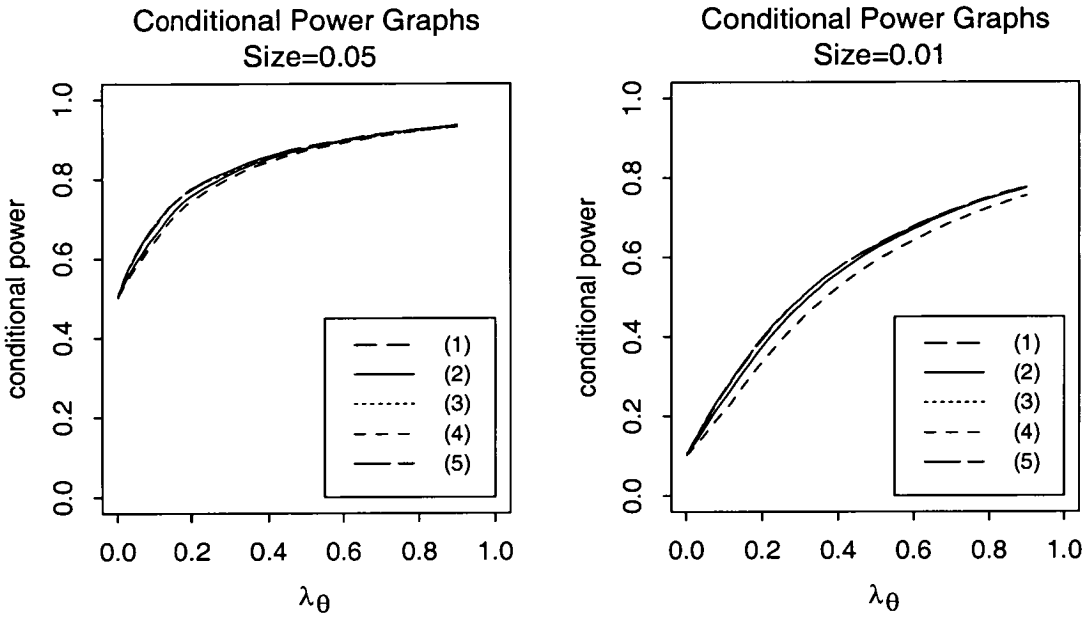


Figure 1.3: Conditional (on $F > F_{m-1,N-m-1;0.90}$) power function for (1) Bayes factor, (2) F statistic, (3) likelihood ratio statistic, (4) posterior mode, (5) Bayes factor under uniform $(0, 0.5)$ prior. Left: 0.05 test size. Right: 0.01 test size.

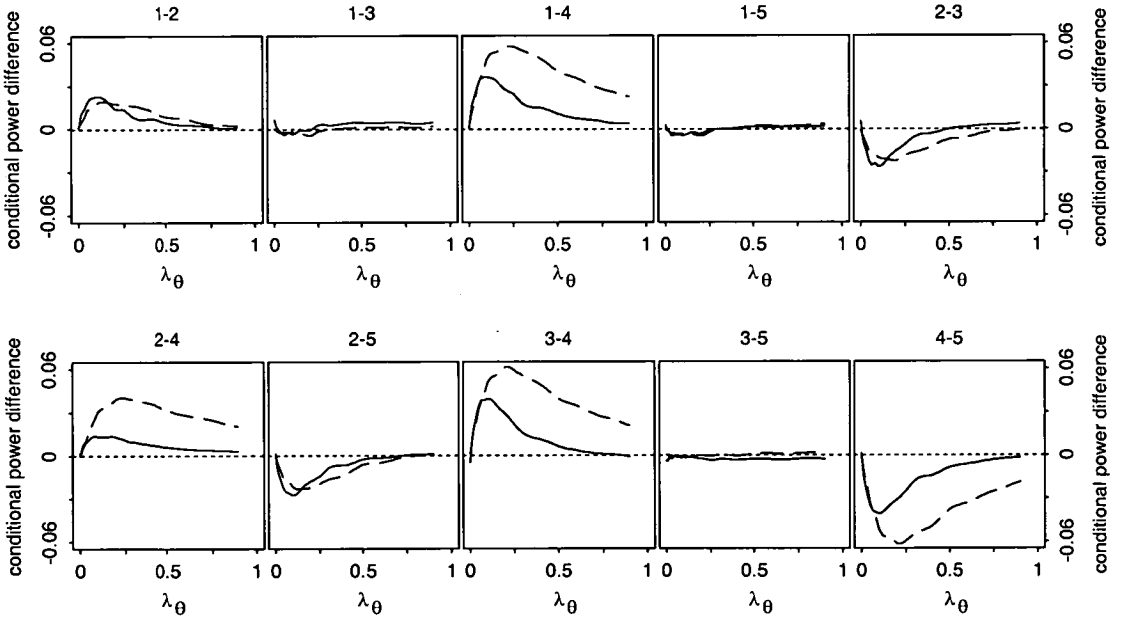


Figure 1.4: Pairwise conditional (on $F > F_{m-1,N-m-1;0.90}$) power function differences for (1) Bayes factor, (2) F statistic, (3) likelihood ratio statistic, (4) posterior mode, (5) Bayes factor under uniform $(0, 0.5)$ prior. The solid lines correspond to 0.05 size tests. The dashed lines correspond to 0.01 size tests.

statistics that potentially had similar properties, according to Table 1.3. Subsequently, we obtained their power for a variety of values of λ_θ in the range 0 to 0.9.

Figure 1.1 presents the power curves of the five selected test statistics for test size 5% and 1%. Because the differences appeared to be quite small, we also obtained the corresponding pairwise power differences, showed in Figure 1.2.

Although the differences do not appear to be great, the Bayes factor has larger power than F over the whole parameter space, and, by shortening the range of integration of λ_θ from 0 to 1, to 0 to twice the reciprocal of the minimum sample size, in this case (0, 0.5), it is possible to get slightly larger improvements in that interval, where the posterior densities are concentrated (see Figure 1.5). The maximum power difference between the Bayes factors and the F statistic is about 1%. The performance of the posterior mode is rather poor. The LRT does better than F in almost the whole range of λ_θ. It also has higher power than BF for small λ_θ but slightly lower for larger ones.

The same conclusions were reached when the power properties of these statistics were studied conditionally on $F > F_{m\text{-}1,N\text{-}m\text{-}1;0.90}$. This situation is important because it represents the only situations when somebody would realistically consider to reject the null hypothesis. The corresponding graphs are presented in Figures 1.3 and 1.4. Moreover, the conditional power differences appeared to be roughly twice the size of the power ones.

The differences in the significance probabilities of the various test statistics, together with the power differences detected, cast a fair amount of doubt on the appropriateness of the use of the F test. Admittedly, these differences were not that great and could be considered to be compensated by the easiness by which the F statistic and its p-value are obtained by standard statistical packages. Hence, one could claim that the obtained results are quite discouraging in practical terms. Following these considerations, we will concentrate for the next few sections of the chapter on Bayesian inference for the current model.

1.7 Posterior density of λ_θ

The functional form of the posterior density of λ_θ was presented in (1.46). Since, from a Bayesian point of view, the inference about the parameter λ_θ is in its posterior distribution, it should be possible to make an applied judgment about the null hypothesis, either looking at its graph or some summary of it.

Figure 1.5 represents the posterior density of λ_θ for the four data sets of Table 1.2. As the null hypothesis, $H_0 : \lambda_\theta = 0$, lies on the boundary of the parameter space, the

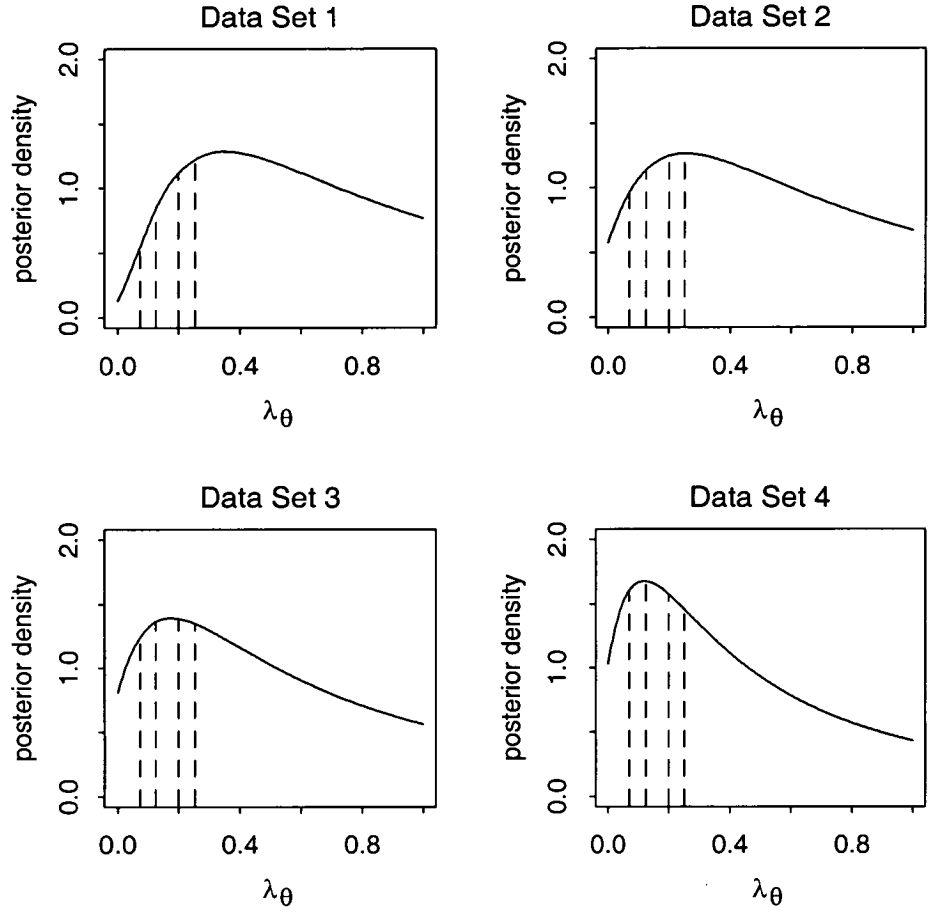


Figure 1.5: Posterior densities of λ_θ for four data sets. The vertical lines are the straight lines with equation λ_θ=1/n_i, i=1,...,5.

posterior density is difficult to interpret. However, as λ_θ^{-1} is the prior sample size, and $\lambda_\theta^{-1} + n_i$ the posterior sample size, it is useful to contrast the posterior density with the reciprocals of the actual sample sizes, n_i^{-1}. We would consider rejecting H_0 whenever most of the posterior probability exceeds n_i^{-1}. A more detailed discussion of this idea is presented in section 2.4.

1.8 Posterior densities of group means

By (1.20), (1.21) and (1.22), conditionally on β, σ_ϵ^2, μ_θ and λ_θ, which are all assumed to have uniform prior distribution, the posterior distribution of θ_i, for $i = 1, \ldots, m$, is

normal with mean θ_i^*, defined in (1.23), and variance $\left(n_i + \lambda_\theta^{-1}\right)^{-1} \sigma_\epsilon^2$, or

$$\frac{n_i + \lambda_\theta^{-1}}{\lambda_\theta^{-1}} \left(\theta_i - \frac{n_i y_{i.}}{n_i + \lambda_\theta^{-1}} \right) \Bigg| \beta, \sigma_\epsilon^2, \mu_\theta, \lambda_\theta, \boldsymbol{y} \sim N\left(\mu_\theta, \lambda_\theta^2 \left(n_i + \lambda_\theta^{-1}\right) \sigma_\epsilon^2\right), \tag{1.66}$$

which is evidently independent of β. Additionally, by (1.40) and (1.41), the posterior distribution of μ_θ, conditionally on σ_ϵ^2 and λ_θ, is normal with mean $\bar{y}_\theta$, defined in (1.42), and variance $\sigma_\epsilon^2 / \sum_{i=1}^m \left(n_i^{-1} + \lambda_\theta\right)^{-1}$. The posterior density of θ_i conditionally only on σ_ϵ^2 and λ_θ can be obtained by the following Lemma.

Lemma 1.3. If the distribution of $X|\mu$ is normal with mean μ and variance σ^2 and the distribution of $\mu|a$ is normal with mean a and variance b^2, then the distribution of $X|a$ is normal with mean a and variance $\sigma^2 + b^2$.

Proof. By the distribution of the sum of two independent normal variates.

Combining (1.66) with the conditional distribution of μ_θ, using Lemma 1.3, and performing the standard location and scale transformation for a normal random variable, we obtain that, conditionally on σ_ϵ^2 and λ_θ, the posterior density of θ_i is normal with mean $\theta_i^\dagger$ and variance $\omega_{\theta_i} \sigma_\epsilon^2$, where

$$\theta_i^\dagger = \left(n_i y_{i.} + \lambda_\theta^{-1} \bar{y}_\theta\right) / \left(n_i + \lambda_\theta^{-1}\right), \tag{1.67}$$

and

$$\omega_{\theta_i} = 1 / \left(n_i + \lambda_\theta^{-1}\right) + \lambda_\theta^{-2} \left\{ \sum_{i=1}^m \left(n_i^{-1} + \lambda_\theta\right)^{-1} \right\}^{-1} / \left(n_i + \lambda_\theta^{-1}\right)^2. \tag{1.68}$$

Notice that, by (1.67), where n_i^{-1} lies compared with λ_θ can be a useful way of judging whether the group means are equal. In particular, $n_i/(n_i + \lambda_\theta^{-1}) \leq 0.5$, or $\lambda_\theta \leq n_i^{-1}$, provides evidence in support of the null hypothesis $H_0 : \lambda_\theta = 0$, since then, $\theta_i^\dagger$ is closer to $\bar{y}_\theta$ than to $y_{i.}$.

The posterior distribution of θ_i can subsequently be obtained as

$$\pi(\theta_i|\boldsymbol{y}) = \iint \pi(\theta_i, \sigma_\epsilon^2, \lambda_\theta|\boldsymbol{y}) \mathrm{d}\sigma_\epsilon^2 \mathrm{d}\lambda_\theta = \iint \pi(\theta_i|\sigma_\epsilon^2, \lambda_\theta, \boldsymbol{y}) \pi(\sigma_\epsilon^2, \lambda_\theta|\boldsymbol{y}) \mathrm{d}\sigma_\epsilon^2 \mathrm{d}\lambda_\theta. \tag{1.69}$$

The integration with respect to σ_ϵ^2 in (1.69), collecting terms in σ_ϵ^2 from (1.43) and the previous normal distribution, then using the inverted chi-squared integral, provides that

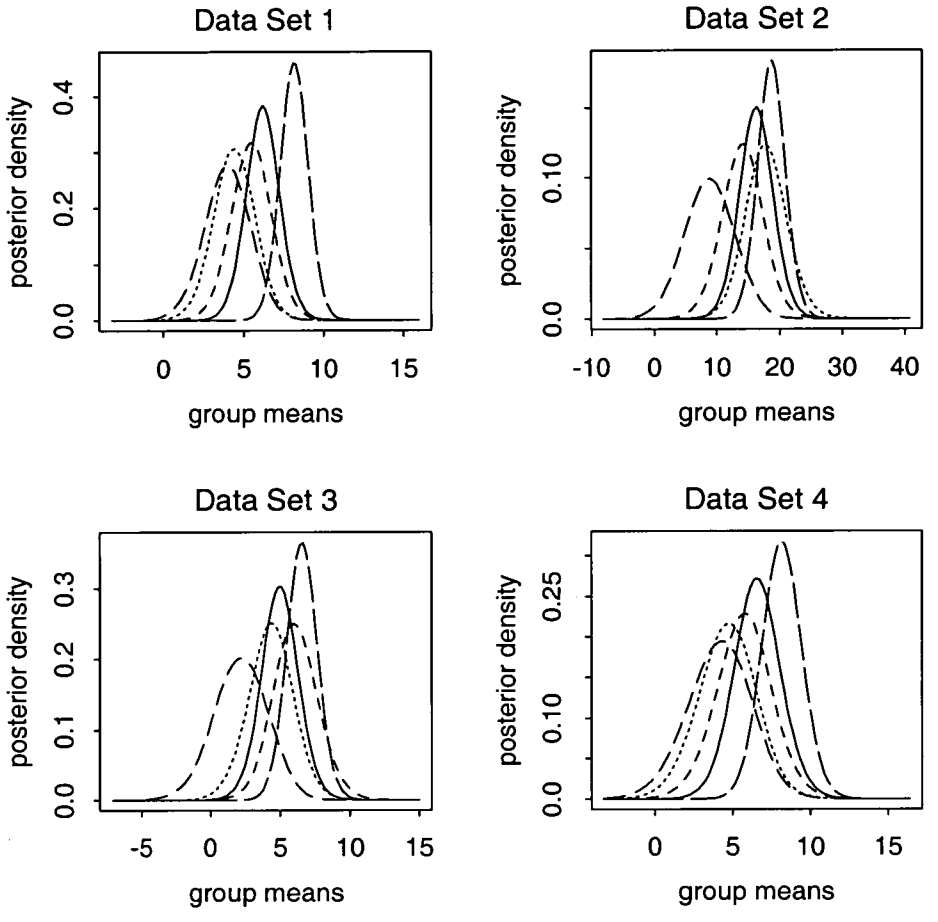


Figure 1.6: Posterior densities of group means (θ_i) for four data sets.

$$\pi(\theta_i|\boldsymbol{y}) \propto \int \left\{ \frac{\left(\theta_i - \theta_i^{\dagger}\right)^2}{\omega_{\theta_i}} + S_R^2 + \mathrm{A}_\theta \right\}^{-(N-3)/2} \Upsilon(\lambda_\theta)\omega_{\theta_i}^{-1/2} \mathrm{d}\lambda_\theta, \qquad (1.70)$$

or using the definition of $\pi(\lambda_\theta|\boldsymbol{y})$ in (1.46), that

$$\pi(\theta_i|\boldsymbol{y}) \propto \int \left\{ 1 + \frac{\left(\theta_i - \theta_i^{\dagger}\right)^2}{\omega_{\theta_i}(S_R^2 + \mathrm{A}_\theta)} \right\}^{-N/2+3/2} (S_R^2 + \mathrm{A}_\theta)^{-1/2}\omega_{\theta_i}^{-1/2}\pi(\lambda_\theta|\boldsymbol{y})\mathrm{d}\lambda_\theta. \qquad (1.71)$$

A numerical integration is hence required for its evaluation at each point.

Figure 1.6 represents the posterior densities of the group means for the four data sets. They can be used, parallel to the classical ANCOVA, for individual comparisons, e.g. by considering the area of overlapping tails, once an overall group mean difference (via λ_θ) has been concluded.

1.9 Posterior density of slope

Combining (1.39) and (1.41), it is possible to integrate out this time the parameter μ_θ first and obtain the posterior density of β, σ_ϵ^2, and λ_θ as

$$\pi(\beta, \sigma_\epsilon^2, \lambda_\theta | \boldsymbol{y}) \propto (\sigma_\epsilon^2)^{-(N-1)/2} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{S_R^2 + (\widehat{\beta}-\beta)^2 s^2 + \mathrm{A}_\theta\right\}\right] \Upsilon(\lambda_\theta). \quad (1.72)$$

A subsequent integration of σ_ϵ^2, gives that

$$\pi(\beta, \lambda_\theta | \boldsymbol{y}) \propto \left\{S_R^2 + (\widehat{\beta}-\beta)^2 s^2 + \mathrm{A}_\theta\right\}^{-(N-3)/2} \Upsilon(\lambda_\theta). \quad (1.73)$$

Hence the posterior density of β satisfies

$$\begin{aligned} \pi(\beta|\boldsymbol{y}) &= \int \pi(\beta|\lambda_\theta, \boldsymbol{y}) \pi(\lambda_\theta|\boldsymbol{y}) \mathrm{d}\lambda_\theta \\ &\propto \int \left\{1 + \frac{(\widehat{\beta}-\beta)^2 s^2}{S_R^2 + \mathrm{A}_\theta}\right\}^{-(N-3)/2} \left(S_R^2 + \mathrm{A}_\theta\right)^{-1/2} \pi(\lambda_\theta|\boldsymbol{y}) \mathrm{d}\lambda_\theta, \end{aligned} \quad (1.74)$$

requiring an one dimensional numerical integration for its computation.

Figure 1.7 (page 31) represents the posterior densities of the slopes for the four data sets. Hypotheses concerning β can be tested using the posterior probability that it is less than a certain value. Hence, to test whether the slope is 0, one could use the posterior probability that $\beta \leq 0$ and reject the null hypothesis when this probability is too small or large. For the four data sets, the corresponding values were 0.971, 0.693, 0.252 and 0.543, suggesting that possibly only for the first data set the slope is significantly negative, as can be seen in the corresponding graph.

1.10 Posterior densities of adjusted means

Conditionally on σ_ϵ^2 and λ_θ, the posterior density of θ_i is normal with mean $\theta_i^\dagger$ and variance $\omega_{\theta_i}\sigma_\epsilon^2$, with $\theta_i^\dagger$ and ω_{θ_i} defined in (1.67) and (1.68) respectively. Additionally, the joint posterior density of β, σ_ϵ^2 and λ_θ is given by (1.72), hence the conditional posterior of β is normal with mean $\widehat{\beta}$ and variance σ_ϵ^2/s^2. Also, β is conditionally independent of θ_i.

Using standard properties of the normal distribution, we conclude that the conditional posterior distribution of the adjusted mean, $\xi_i = \theta_i - \beta(x_{i.} - x_{..})$, is normal with mean ξ_i^* and variance $\omega_{\xi_i}\sigma_\epsilon^2$, where

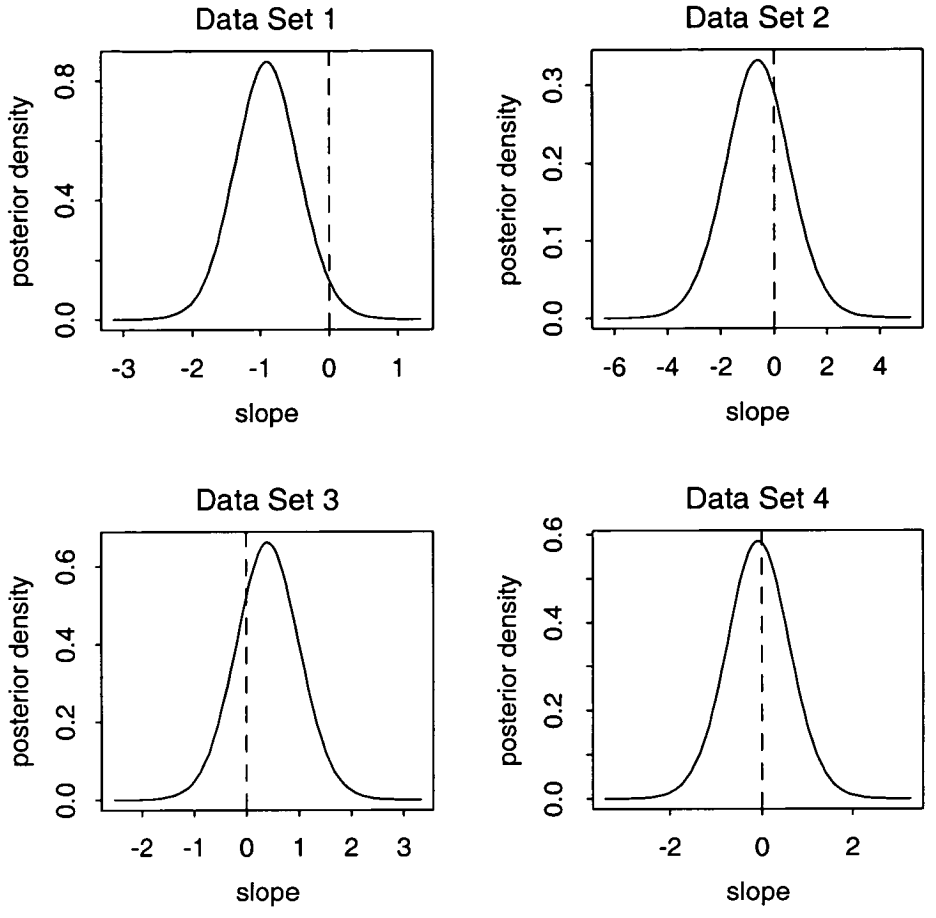


Figure 1.7: Posterior densities of slope (β) for four data sets. The dashed vertical lines correspond to $\beta = 0$.

$$\xi_i^* = \theta_i^\dagger - \hat{\beta}(x_{i.} - x_{..}), \tag{1.75}$$

and

$$\omega_{\xi_i} = \omega_{\theta_i} + (x_{i.} - x_{..})^2 s^{-2}. \tag{1.76}$$

Using comparable algebraic derivations, with the ones used for the posterior density of the group means θ_i, firstly integrating with respect to σ_ϵ^2 and secondly slightly rearranging the integrand, we obtain that

$$\begin{aligned}
\pi(\xi_i|\boldsymbol{y}) &= \iint \pi(\xi_i|\sigma_\epsilon^2, \lambda_\theta, \boldsymbol{y})\pi(\sigma_\epsilon^2, \lambda_\theta|\boldsymbol{y})\mathrm{d}\sigma_\epsilon^2\mathrm{d}\lambda_\theta \\
&\propto \int \left\{ \frac{(\xi_i - \xi_i^*)^2}{\omega_{\xi_i}} + S_R^2 + \mathrm{A}_\theta \right\}^{-(N-3)/2} \Upsilon(\lambda_\theta)\omega_{\xi_i}^{-1/2}\mathrm{d}\lambda_\theta \\
&\propto \int \left\{ 1 + \frac{(\xi_i - \xi_i^*)^2}{\omega_{\xi_i}(S_R^2 + \mathrm{A}_\theta)} \right\}^{-(N-3)/2} (S_R^2 + \mathrm{A}_\theta)^{-1/2}\omega_{\xi_i}^{-1/2}\pi(\lambda_\theta|\boldsymbol{y})\mathrm{d}\lambda_\theta,
\end{aligned}$$

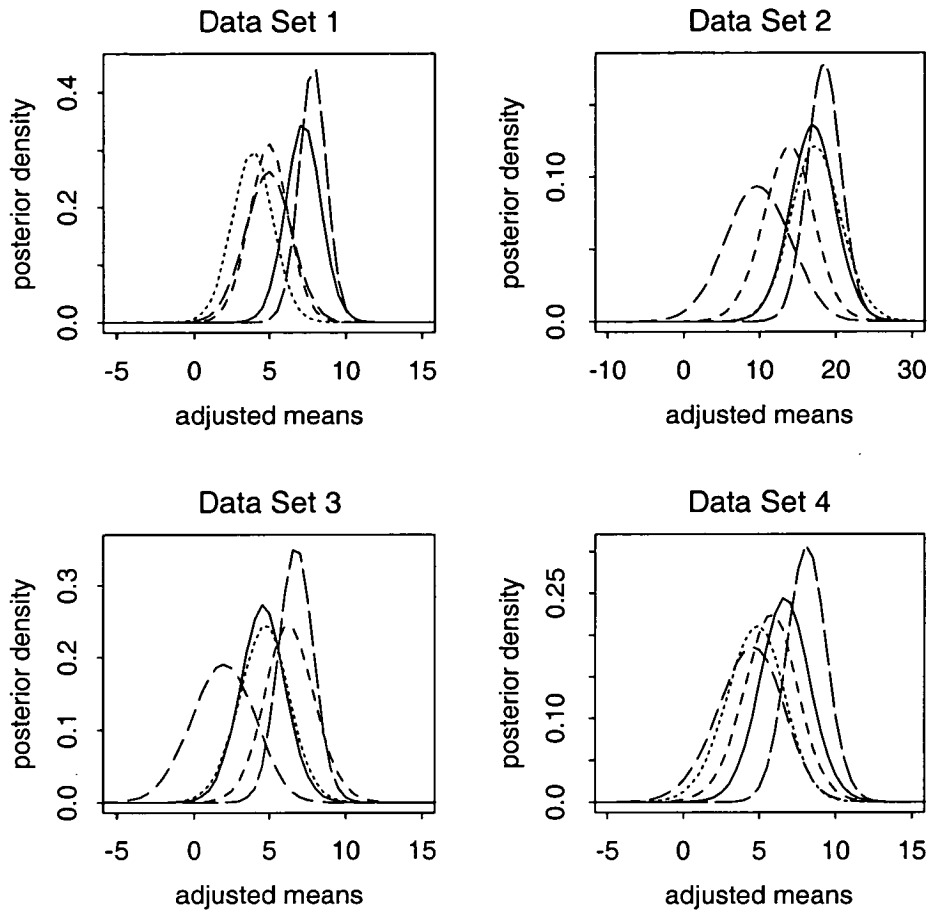


Figure 1.8: Posterior densities of adjusted means (ξ_i) for four data sets.

(1.77)

requiring the usual one dimensional numerical integration. Figure 1.8 represents the posterior densities of the adjusted group means for the four data sets. These can be used to compare group means of different groups at the same level of the covariate. They can also be contrasted with the unadjusted group means (Figure 1.6) for an indirect assessment of the influence of the covariate. Hence, contrasting the posterior densities of Figure 1.8 with the ones in Figure 1.6, we conclude that the adjustment was mainly noticeable for data set 1, the only one for which the slope appears to be non zero (see Figure 1.7).

1.11 Parametric residuals

In order to judge the fit of the model in greater detail, we could use a method introduced by Leonard and Novick (1986), in the context of log-linear models for contingency tables. They considered the posterior density of the parametric residuals, which in our formulation are defined as $\rho_i = \theta_i - \mu_\theta$, for $i = 1, 2, \ldots, m$.

To compute their posterior density, we have to observe that conditionally on σ_ϵ^2, μ_θ and λ_θ, the posterior distribution of the θ_i is normal with mean

$$\frac{n_i y_{i\cdot} + \lambda_\theta^{-1}\mu_\theta}{n_i + \lambda_\theta^{-1}}, \tag{1.78}$$

and variance $\sigma_\epsilon^2 \left(n_i + \lambda_\theta^{-1}\right)^{-1}$. Combining this result with the posterior distribution of μ_θ given σ_ϵ^2 and λ_θ, as described in section 1.8, we can obtain that, conditionally on σ_ϵ^2 and λ_θ, ρ_i is normally distributed with mean ρ_i^* and variance $\omega_{\rho_i}\sigma_\epsilon^2$, with

$$\rho_i^* = \frac{n_i \left(y_{i\cdot} - \bar{y}_\theta\right)}{n_i + \lambda_\theta^{-1}}, \tag{1.79}$$

and

$$\omega_{\rho_i} = \left\{\sum_{i=1}^{m} \left(n_i^{-1} + \lambda_\theta\right)^{-1}\right\}^{-1} \frac{n_i^2}{\left(n_i + \lambda_\theta^{-1}\right)^2} + \left(n_i + \lambda_\theta^{-1}\right)^{-1}, \tag{1.80}$$

with $\bar{y}_\theta$ defined in (1.42). Hence the unconditional posterior density of ρ_i satisfies

$$\begin{aligned}
\pi(\rho_i|\boldsymbol{y}) &= \iint \pi(\rho_i|\sigma_\epsilon^2, \lambda_\theta, \boldsymbol{y})\pi(\sigma_\epsilon^2, \lambda_\theta|\boldsymbol{y})\mathrm{d}\sigma_\epsilon^2\mathrm{d}\lambda_\theta \\
&\propto \int \left\{\frac{(\rho_i - \rho_i^*)^2}{\omega_{\rho_i}} + S_R^2 + \mathrm{A}_\theta\right\}^{-(N-3)/2} \Upsilon(\lambda_\theta)\omega_{\rho_i}^{-1/2}\mathrm{d}\lambda_\theta \\
&\propto \int \left\{1 + \frac{(\rho_i - \rho_i^*)^2}{\omega_{\rho_i}(S_R^2 + \mathrm{A}_\theta)}\right\}^{-(N-3)/2} (S_R^2 + \mathrm{A}_\theta)^{-1/2}\omega_{\rho_i}^{-1/2}\pi(\lambda_\theta|\boldsymbol{y})\mathrm{d}\lambda_\theta.
\end{aligned} \tag{1.81}$$

Figure 1.9 represents the posterior densities of the parametric residuals for the four data sets. Large areas of a posterior density to the left or right of zero would indicate an individual group departure from the null model. Unusual posterior density shapes can be anticipated due to the multiplicative nature of the integrand in (1.81). For example, bimodal features and bumps (e.g. bumps for data sets 2 and 3), meaningfully parallel a phenomenon noticed by Aitken et al (1997) in another context. Furthermore, those curves with single modes at zero (e.g. data sets 3 and 4), strongly indicate agreement

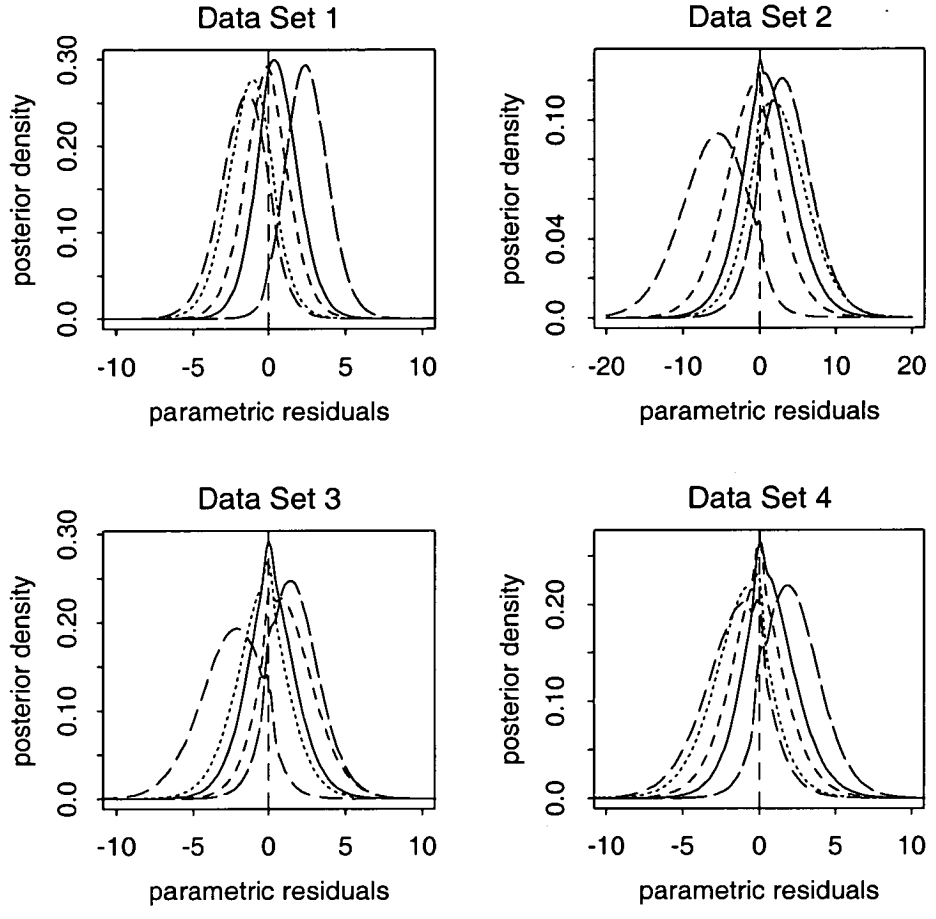


Figure 1.9: Posterior densities of parametric residuals (ρ_i) for four data sets. The dashed vertical lines correspond to $\rho_i = 0$.

with the null hypothesis for the appropriate group.

1.12 Shrinkage estimators

Our analysis of shrinkage estimators is intended to intuitively justify, the comparisons in Figure 1.5 of the posterior density of λ_θ with the reciprocals of the sample sizes, n_i^{-1}. The posterior mean of θ_i, conditionally on λ_θ as presented in (1.67) can also be expressed as

$$\theta_i^\dagger = E(\theta_i|\lambda_\theta, \boldsymbol{y}) = \gamma_i^* y_{i\cdot} + \zeta_i^*, \tag{1.82}$$

with

$$\gamma_i^* = E_{\lambda_\theta|\boldsymbol{y}}\left(\frac{n_i}{n_i + \lambda_\theta^{-1}}\right), \tag{1.83}$$

and

$$\zeta_i^* = E_{\lambda_\theta|\boldsymbol{y}}\left(\frac{\lambda_\theta^{-1}}{n_i + \lambda_\theta^{-1}}\bar{y}_\theta\right). \tag{1.84}$$

Since ζ_i^* is close to $(1-\gamma_i^*)y_{..}$, comparing γ_i^* with 0.5 can be a criterion for judging group mean equality, with $\gamma_i^* \leq 0.5$ suggesting equality of the group means. More precisely, consider the preliminary test estimators (e.g. Cohen, 1974),

$$\tilde{\theta}_i = \begin{cases} y_{i.}, & \boldsymbol{y} \in C \\ \eta, & \boldsymbol{y} \notin C \end{cases} \tag{1.85}$$

with C and η unknown. Consider also the quadratic loss function

$$L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = m^{-1}\sum_{i=1}^{m}(\tilde{\theta}_i - \theta_i)^2. \tag{1.86}$$

The posterior expected loss of $\tilde{\boldsymbol{\theta}}$, under (1.86), becomes

$$q(\tilde{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^{m}(\tilde{\theta}_i - \gamma_i^* y_{i.} - \zeta_i^*)^2 + m^{-1}\sum_{i=1}^{m} var(\theta_i|\boldsymbol{y}). \tag{1.87}$$

Using definition (1.85), the posterior expected loss can be expressed as

$$q(\tilde{\boldsymbol{\theta}}) = \begin{cases} m^{-1}\sum_{i=1}^{m}(y_{i.} - \gamma_i^* y_{i.} - \zeta_i^*)^2 + m^{-1}\sum_{i=1}^{m} var(\theta_i|\mathbf{y}), & \boldsymbol{y} \in C \\ m^{-1}\sum_{i=1}^{m}(\eta - \gamma_i^* y_{i.} - \zeta_i^*)^2 + m^{-1}\sum_{i=1}^{m} var(\theta_i|\mathbf{y}), & \boldsymbol{y} \notin C \end{cases}. \tag{1.88}$$

Since the selected loss function is quadratic, (1.88) is minimized, for $\boldsymbol{y} \notin C$, when $\eta = \eta^*$, with

$$\eta^* = m^{-1}\sum_{i=1}^{m}(\gamma_i^* y_{i.} + \zeta_i^*). \tag{1.89}$$

Some rearrangement, combined with the definition of $\bar{y}_\theta$ in (1.42), gives that

$$\eta^* = E_{\lambda_\theta|\boldsymbol{y}}\left(\bar{y}_\theta\right). \tag{1.90}$$

Thus, the minimum posterior expected loss has the form

$$q_{\min}(\tilde{\boldsymbol{\theta}}) = \begin{cases} Q_1, & \boldsymbol{y} \in C \\ Q_2, & \boldsymbol{y} \notin C \end{cases}, \tag{1.91}$$

with

$$Q_1 = m^{-1}\sum_{i=1}^{m}(y_{i.} - \gamma_i^* y_{i.} - \zeta_i^*)^2 + m^{-1}\sum_{i=1}^{m} var(\theta_i|\boldsymbol{y}), \tag{1.92}$$

and

$$Q_2 = m^{-1}\sum_{i=1}^{m}(\eta^* - \gamma_i^* y_{i\cdot} - \zeta_i^*)^2 + m^{-1}\sum_{i=1}^{m} var(\theta_i|\boldsymbol{y}). \tag{1.93}$$

The Bayes rule, under the constrained class in (1.89), indicates $\boldsymbol{y} \in C$ and hence rejects $H_0 : \lambda_\theta = 0$ whenever $Q_1 < Q_2$.

Table 1.5: *Shrinkage estimators for the four data sets.*

Data Set	Group	n_i	$y_{i\cdot}$	γ_i^*	ζ_i^*	$\theta_i^\dagger$	var(θ_i/y)
1	1	8	6.375	0.754	1.415	6.220	1.135
	2	5	3.800	0.667	1.907	4.443	1.757
	3	5	5.400	0.667	1.907	5.510	1.646
	4	4	3.000	0.621	2.168	4.032	2.151
	5	14	8.643	0.835	0.950	8.168	0.788
			Q_1/Q_2= 0.505				
2	1	8	16.625	0.720	4.298	16.269	7.401
	2	5	19.200	0.632	5.630	17.773	10.859
	3	5	13.400	0.632	5.630	14.105	10.787
	4	4	4.750	0.587	6.323	9.111	16.613
	5	14	19.429	0.805	2.996	18.644	4.920
			Q_1/Q_2= 0.672				
3	1	8	5.000	0.697	1.498	4.981	1.819
	2	5	4.000	0.608	1.931	4.361	2.635
	3	5	6.800	0.608	1.931	6.062	2.658
	4	4	0.250	0.562	2.153	2.293	3.945
	5	14	7.000	0.785	1.065	6.562	1.239
			Q_1/Q_2= 0.738				
4	1	8	6.875	0.667	2.002	6.587	2.269
	2	5	3.800	0.576	2.539	4.727	3.467
	3	5	5.600	0.576	2.539	5.764	3.204
	4	4	2.750	0.530	2.809	4.267	4.245
	5	14	8.786	0.760	1.449	8.127	1.632
			Q_1/Q_2= 0.762				

For the four data sets already described, the values of γ_i^*, ζ_i^*, $\theta_i^\dagger$, and the quantity Q_1/Q_2, that is, the expected loss ratio of rejection over not rejection, are presented in Table 1.5. Apart from the interest the smoothed estimates of the θ_i's present, it is quite clear, by the values of the γ_i^*'s and the expected loss ratios, that different conclusions about H_0 are reached, compared with the test statistics presented in previous sections,

and one would tend to refute H_0 in all four data sets. In general, consideration of Q_1/Q_2 together with the posterior densities of the θ_i permit to incorporate the concept of practical significance into the decision making process.

1.12.1 Equally replicated case

In section 1.8 we showed that conditionally on σ_ϵ^2 and λ_θ, the group mean, θ_i, for $i = 1, 2, \ldots, n_i$, is normal, with mean defined in (1.67) and independent of σ_ϵ^2. In the equally replicated case, with $n_i = n$, (1.67) takes a simpler form, which is

$$E(\theta_i|\lambda_\theta, \boldsymbol{y}) = y_{i.} - \frac{\lambda_\theta^{-1}}{n + \lambda_\theta^{-1}} (y_{i.} - y_{..}) . \tag{1.94}$$

The unconditional posterior mean of θ_i can then be rewritten as

$$E(\theta_i|\boldsymbol{y}) = y_{i.} - E_{\lambda_\theta|\boldsymbol{y}} \left(\frac{n^{-1}}{n^{-1} + \lambda_\theta} |\mathbf{y} \right) (y_{i.} - y_{..}) , \tag{1.95}$$

with the expectation in the right hand side of (1.95) taken with respect to the posterior distribution of λ_θ. In section 1.4.2.1, the quantity $z = (m-1)F/\{(m-3)(n^{-1} + \lambda_\theta)n\}$ was shown to have a truncated F distribution, a posteriori, with $m-3$ and $N-m-1$ degrees of freedom.

Let $B_x(\cdot, \cdot)$, be the incomplete Beta function, defined as

$$B_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt, \;\; \alpha, \beta > 0, \; 0 < x < 1, \tag{1.96}$$

and

$$\mathcal{I}_x(\alpha, \beta) = \frac{B_x(\alpha, \beta)}{B(\alpha, \beta)}, \tag{1.97}$$

the cumulative distribution function (c.d.f.) of the Beta distribution, where

$$B(\alpha, \beta)^{-1} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}, \tag{1.98}$$

using the standard definition of the gamma function, $\Gamma(\cdot)$. Then, if the random variable Z is distributed as F with α and β degrees of freedom, the following formula relates its survivor function with the c.d.f. of a Beta variate,

$$P(Z > Z_0) = \mathcal{I}_x \left(\beta/2, \alpha/2 \right) , \tag{1.99}$$

where

$$x = \frac{\beta}{\beta + \alpha Z_0}. \tag{1.100}$$

Setting $m-3=\nu_1$, $N-m-1=\nu_2$, and $A=\{z : z \leq \frac{\nu_1+2}{\nu_1}F\}$, we find

$$E_{\lambda_\theta|\boldsymbol{y}}\left(\frac{n^{-1}}{n^{-1}+\lambda_\theta}\right) = E\left(\frac{\nu_1}{\nu_1+2}F^{-1}z\right) = \frac{E\left(\frac{\nu_1}{\nu_1+2}F^{-1}z\ I_A(z)\right)}{E\left(I_A(z)\right)}, \tag{1.101}$$

with $I(\cdot)$ the indicator function. The expectation in the denominator of (1.101) can be computed by an application of (1.99). For the numerator, a transformation to $t=\nu_2/(\nu_2+\nu_1 z)$ is first needed. These computations lead to the following result

$$E(\theta_i|\boldsymbol{y}) = y_{i.} - F^{-1}g(F)\left(y_{i.} - y_{..}\right) \tag{1.102}$$

with

$$g(F) = \frac{\nu_1\nu_2\mathcal{I}_{b(F)}\left(\frac{\nu_1+2}{2}, \frac{\nu_2-2}{2}\right)}{(\nu_1+2)(\nu_2-2)\mathcal{I}_{b(F)}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}, \tag{1.103}$$

and

$$b(F) = \frac{(\nu_1+2)F}{\nu_2+(\nu_1+2)F}. \tag{1.104}$$

Using the identity

$$\mathcal{I}_x(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)}x^\alpha(1-x)^{\beta-1} + \mathcal{I}_x(\alpha+1,\beta-1), \tag{1.105}$$

(see Abramowitz and Stegun, 1965, p. 944), (1.103) reduces to

$$g(F) = \frac{\nu_1\nu_2}{(\nu_1+2)(\nu_2-2)}\left\{1 - \frac{\Gamma(\frac{\nu_1+\nu_2}{2})b(F)^{\frac{\nu_1}{2}}(1-b(F))^{\frac{\nu_2}{2}-1}}{\Gamma(\frac{\nu_1+2}{2})\Gamma(\frac{\nu_2}{2})\mathcal{I}_{b(F)}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}\right\}, \tag{1.106}$$

which can be contrasted with equation (2.3) appearing in Leonard and Ord (1976) that corresponds to the random effects ANOVA model. The latter is incorrectly stated.

The ζ_i^* and γ_i^*, that were defined in (1.83) and (1.84), are now independent of the group i, and satisfy

$$\zeta_i^* = y_{..}E_{\lambda_\theta|\boldsymbol{y}}\left(\frac{n^{-1}}{\lambda_\theta+n^{-1}}\right) = \zeta^*, \tag{1.107}$$

and

$$\gamma_i^* = 1 - E_{\lambda_\theta|\boldsymbol{y}}\left(\frac{n^{-1}}{\lambda_\theta+n^{-1}}\right) = 1 - \frac{\zeta^*}{y_{..}} = \gamma^*. \tag{1.108}$$

As $\lambda_\theta/(\lambda_\theta+n^{-1}) \leq 1/2$ is equivalent to $\lambda_\theta \leq n^{-1}$, and the Bayesian decision, in this

equally replicated case tells us to prefer H_0 whenever $E\left(\lambda_\theta/(\lambda_\theta + n^{-1})\right) \leq 1/2$, our argument gives added justification for the comparison of the posterior density of λ_θ (see Figure 1.5) with the reciprocal of the sample size, and also motivates the final choice of Bayes factor, considered in the last section of this chapter.

Substituting (1.107) and (1.108) in (1.91) using that $\bar{y}_\theta = y_{..}$, we obtain that

$$q_{\min}(\widetilde{\boldsymbol{\theta}}) = \begin{cases} (\zeta^*/y_{..})^2 m^{-1} \sum_{i=1}^m (y_{i.} - y_{..})^2 + m^{-1} \sum_{i=1}^m var(\theta_i|\boldsymbol{y}), & \boldsymbol{y} \in C \\ (\gamma^*)^2 m^{-1} \sum_{i=1}^m (y_{i.} - y_{..})^2 + m^{-1} \sum_{i=1}^m var(\theta_i|\boldsymbol{y}), & \boldsymbol{y} \notin C \end{cases} . \quad (1.109)$$

Table 1.6 presents the minimum values of F for which the function $F^{-1}g(F) < 0.5$, hence rejecting the null hypothesis. The conclusion is that $F > 2$ always rejects, however, for not particularly small sample sizes, this value can be quite smaller. These results can be contrasted with those of Leonard and Ord (1976) and Stone (1977), who also consider frequency mean squared error, and cross-validatory justifications. Stone recommends the critical value of 2, for all values of m and n. Our critical values can be much closer to unity.

Table 1.6: *F values for rejecting group mean equality hypothesis.*

Sample	Number of groups (m)							
size (n)	5	10	15	20	25	50	75	100
5	<0.1000	1.3607	1.6677	1.7882	1.8479	1.9369	1.9590	1.9695
10	<0.1000	1.3723	1.6649	1.7790	1.8363	1.9266	1.9518	1.9640
15	<0.1000	1.3751	1.6635	1.7759	1.8326	1.9237	1.9497	1.9625
20	<0.1000	1.3764	1.6627	1.7743	1.8308	1.9223	1.9487	1.9617
25	<0.1000	1.3771	1.6623	1.7734	1.8297	1.9214	1.9482	1.9613
30	<0.1000	1.3776	1.6620	1.7728	1.8290	1.9209	1.9478	1.9610
35	<0.1000	1.3779	1.6617	1.7724	1.8285	1.9205	1.9475	1.9608
40	<0.1000	1.3782	1.6616	1.7720	1.8281	1.9202	1.9473	1.9607
45	<0.1000	1.3783	1.6614	1.7718	1.8278	1.9200	1.9472	1.9605
50	<0.1000	1.3785	1.6613	1.7716	1.8276	1.9198	1.9471	1.9604

1.13 Concluding remarks

Our study was originally motivated by the power optimality property of the Bayes factor. This provided positive, but quite disappointing results, given the computational effort required for the simplest ANCOVA model. Additionally, we discovered an am-

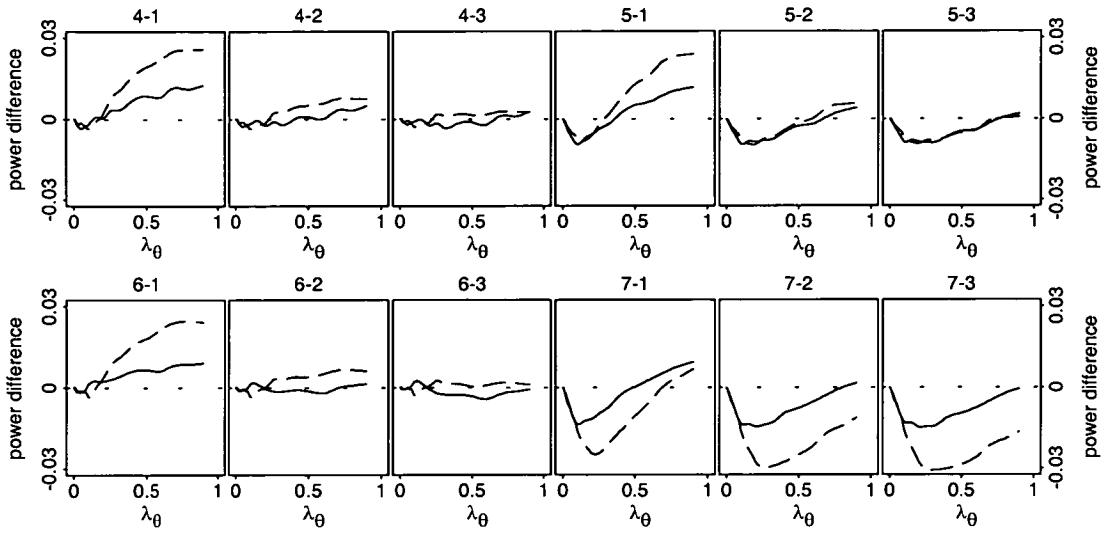


Figure 1.10: Pairwise power function differences for (1) Bayes factor of $\lambda_\theta = 0$ vs $\lambda_\theta = \min n_i^{-1}$, (2) Bayes factor of $\lambda_\theta = 0$ vs $\lambda_\theta =$ average n_i^{-1}, (3) Bayes factor of $\lambda_\theta = 0$ vs $\lambda_\theta = \max n_i^{-1}$, (4) Bayes factor, (5) F statistic, (6) likelihood ratio statistic, (7) posterior mode. The solid lines correspond to 0.05 size tests. The dashed lines correspond to 0.01 size tests. The five group sample sizes were $n_1 = 8$, $n_2 = 5$, $n_3 = 4$, $n_4 = 4$, $n_5 = 14$.

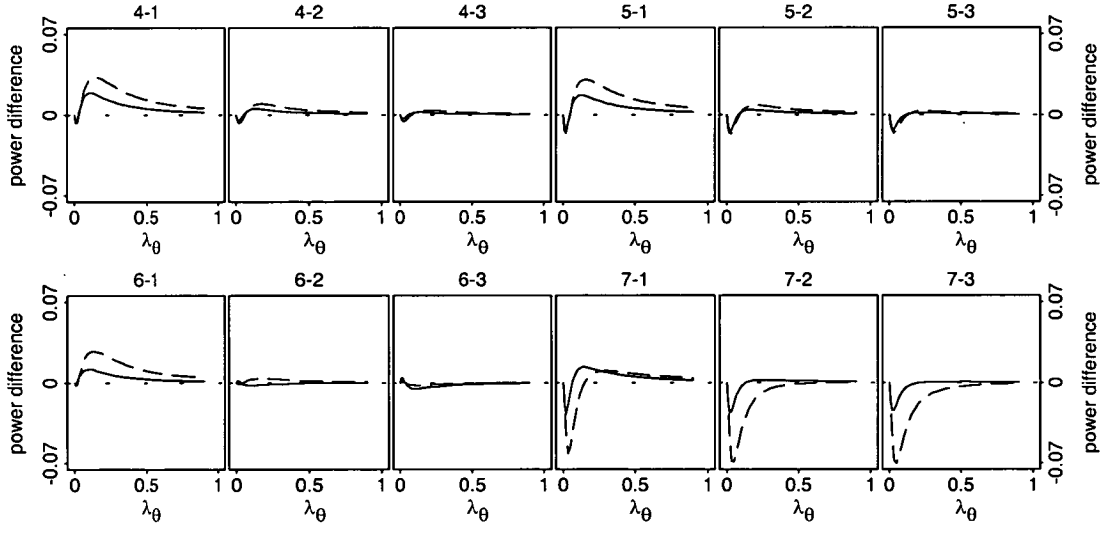


Figure 1.11: Pairwise power function differences for (1) Bayes factor of $\lambda_\theta = 0$ vs $\lambda_\theta = \min n_i^{-1}$, (2) Bayes factor of $\lambda_\theta = 0$ vs $\lambda_\theta =$ average n_i^{-1}, (3) Bayes factor of $\lambda_\theta = 0$ vs $\lambda_\theta = \max n_i^{-1}$, (4) Bayes factor, (5) F statistic, (6) likelihood ratio statistic, (7) posterior mode. The solid lines correspond to 0.05 size tests. The dashed lines correspond to 0.01 size tests. The five group sample sizes were $n_1 = 40$, $n_2 = 22$, $n_3 = 20$, $n_4 = 67$, $n_5 = 128$.

biguity as to what is the best statistic for testing the hypothesis $H_0 : \lambda_\theta = 0$. On the other hand, as we discussed in previous sections, comparing the reciprocal of the sample sizes with the posterior density of the parameter λ_θ can provide a test statistic that is intuitively appealing.

Following these ideas, we are going to make a final suggestion of test statistics for testing $H_0 : \lambda_\theta = 0$, that avoid the computational effort, while not losing the previously described intuitive appeal. These statistics are the Bayes factors for simple versus simple hypotheses, with null hypothesis corresponding to $\lambda_\theta = 0$ and alternative corresponding to (a) $\lambda_\theta = \min n_i^{-1}$, (b) $\lambda_\theta =$ average n_i^{-1}, and (c) $\lambda_\theta = \max n_i^{-1}$. The power difference results for these three Bayes factors and the F statistic, the likelihood ratio statistic, the posterior mode and the standard Bayes factor for the sample sizes of the data sets used throughout Chapter 1 are presented in Figure 1.10, and for the larger sample sizes of the full neuropsychological test data are presented in Figure 1.11.

By Theorem 1, we knew that the three new test statistics would maximize the average power at λ_θ corresponding to the simple alternative hypothesis. However, as Figures 1.10 and 1.11 demonstrate, these statistics can provide slightly better power than the F statistic for small sample sizes and values of λ_θ close to zero. Additionally, especially the Bayes factor corresponding to the alternative $H_1 : \lambda_\theta = \min n_i^{-1}$, can be slightly better than the standard Bayes factor for small values of λ_θ, i.e. those lying in a demonstrably key region of interest.

These considerations lead us to attempt the full Bayesian analysis, but under continuous prior distributions, when using more complicated ANCOVA models, to seek methods that help us interpret the corresponding posterior densities using comparisons of λ_θ with the reciprocals of the sample sizes and to try to develop similar comparisons for other parameters. This is the route we will follow in the next three chapters.

Chapter 2

Bayesian inference for the analysis of covariance with random variances

Following the conclusions of Chapter 1, we will study the Bayesian inference of a new sampling model for the Analysis of Covariance, which permits the investigation of equality of the group means and slopes, when the variances are taken to be unequal. An unknown parameter ν will be introduced, which measures the degree of equality of the variances. The inferences will be based on Markov chain Monte Carlo methods, of which a summary will be presented. This model will be used later in this thesis to analyze the data from the Scottish offender study, which motivated the undertaking of this research. In the latter sections of this chapter we will present the Bayesian inference for further generalizations of ANCOVA models that include several covariates. In all cases, our inferences for the sampling parameter ν will provide one key novelty of our procedures.

2.1 Introduction to sampling model

Consider m groups of observations $\{y_{ij};\ j = 1,\ldots,n_i\}$, for $i = 1,\ldots m$, $(m \geq 4)$, where a scalar value x_{ij}, of an explanatory, or confounding, variable is assigned to each observation y_{ij}. It is frequently unreasonable to assume equality of the regression slopes in the relationship of the explanatory variable and the observations $\boldsymbol{y}$, as well as equality of the fixed effects variances across the m groups. Therefore a standard

ANCOVA, as already presented, can be inappropriate.

We consider, hence, a model with three sets of random effects, firstly a set of conditional group means $\{\theta_1, \theta_2, \ldots, \theta_m\}$, secondly a set of conditional regression slopes $\{\beta_1, \beta_2, \ldots, \beta_m\}$, and finally, a set of conditional variances $\{\phi_1, \phi_2, \ldots, \phi_m\}$. Three unknown parameters in the model, μ_θ, μ_β, and ζ^{-1}, will respectively represent the common means of the three sets of random effects, the θ_i, the β_i, and the ϕ_i^{-1}. Three further unknown parameters λ_θ, λ_β, and ν^{-1} will measure departures of the random effects from these means, i.e. from three equality hypotheses.

Our sampling model is defined by the following hierarchical structure:

(a) Conditionally on the three sets of random effects, the y_{ij} are independent, for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, and normally distributed with respective means

$$\theta_i + \beta_i(x_{ij} - x_{i.}) \qquad (i = 1, \ldots, m), \tag{2.1}$$

and variances ϕ_i, with $x_{i.}$ denoting the ith group covariate mean. The variable

$$\xi_i = \theta_i - \beta_i(x_{i.} - x_{..}) \tag{2.2}$$

is known as the (conditional) "adjusted group mean" for the ith group, where $x_{..}$ denotes the grand covariate mean. The expression in (2.2) for ξ_i adjusts the group mean θ_i in the same way as in the constant slope and variance model.

(b) Conditionally on the ϕ_i and two model parameters μ_θ and λ_θ, the θ_i and β_i are mutually independent, and the θ_i are normally distributed with common mean μ_θ and respective variances $\lambda_\theta\phi_i$, for $i = 1, \ldots, m$.

(c) Conditionally on the ϕ_i and two further model parameters μ_β and λ_β, the β_i are normally distributed with common mean μ_β and respective variances $\lambda_\beta\phi_i$, for $i = 1, \ldots, m$.

(d) The ϕ_i, given ζ and ν, the two final model parameters, are independent with respective densities

$$\pi(\phi_i|\nu,\zeta) = K(\nu)\zeta^{\nu/2}\phi_i^{-\frac{1}{2}(\nu+2)}\exp(-\frac{\nu\zeta}{2\phi_i}) \qquad (0 < \phi_i < \infty; i = 1, \ldots, m), \tag{2.3}$$

with

$$K(\nu) = (\nu/2)^{\nu/2}/\Gamma(\nu/2), \tag{2.4}$$

that is, $\nu\zeta/\phi_i$ possesses a chi-squared distribution with ν degrees of freedom.

Box and Tiao (1992, p. 219) compare several unequal variances in a random effects ANOVA model, but using log-uniform distributions for the sampling variances ϕ_i. O'Hagan (1979) indicates that their apparently robustifying sampling distribution is in fact still "outlier prone", while random sampling from a univariate $t-$distribution is "outlier resistant". Precise definitions of these intuitively appealing concepts can be found in O'Hagan's paper. Lindley (1971) proposes a different random effects model in a Bayesian ANOVA context ($\beta_i \equiv 0$) which takes the θ_i and ϕ_i to be independent. His model yields a complicated joint distribution for the observations, which is not generalized multivariate t. Leonard (1975) and Leonard and Hsu (1992) assume multivariate normal distributions for sets of log-variances and for log-covariance matrices. Our independent inverted chi-squared distributions for the ϕ_i give simpler results in the current sampling situation. We are using an alternative formulation of $m-$group regression models (Lindley and Smith, 1972, Miller and Fortney, 1984, and Blattberg and George, 1991), but for the purposes of ANCOVA (e.g. for noisy data), rather than regression, i.e. it is of less critical importance for the regression surface to provide a good fit to the data points.

Under this choice of sampling model, the joint distribution of the observations and the three sets of random effects, given the six model parameters, using the standard notation for joint, $p(\cdot)$, and conditional, $p(\cdot|\cdot)$, densities, is

$$\begin{aligned} p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi}|\mu_\theta,\mu_\beta,\lambda_\theta,\lambda_\beta,\nu,\zeta) &= p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi},\mu_\theta,\lambda_\theta)p(\boldsymbol{\beta}|\boldsymbol{\phi},\mu_\beta,\lambda_\beta)p(\boldsymbol{\phi}|\nu,\zeta) \\ &= \textstyle\prod_{i=1}^m \prod_{j=1}^{n_i} (2\pi\phi_i)^{-1/2} \exp\left[-\frac{1}{2\phi_i}\{y_{ij}-\theta_i-\beta_i(x_{ij}-x_{i.})\}^2\right] \\ &\times \textstyle\prod_{i=1}^m (2\pi\lambda_\theta\phi_i)^{-1/2} \exp\left\{-\frac{1}{2\lambda_\theta\phi_i}(\theta_i-\mu_\theta)^2\right\} \\ &\times \textstyle\prod_{i=1}^m (2\pi\lambda_\beta\phi_i)^{-1/2} \exp\left\{-\frac{1}{2\lambda_\beta\phi_i}(\beta_i-\mu_\beta)^2\right\} \\ &\times \textstyle\prod_{i=1}^m K(\nu)\zeta^{\nu/2}\phi_i^{-(\nu/2+1)} \exp\left\{-\frac{\nu\zeta}{2\phi_i}\right\}. \end{aligned} \tag{2.5}$$

Setting

$$\begin{aligned} U_i &= \textstyle\sum_{j=1}^{n_i}\left\{y_{ij}-y_{i.}-\hat{\beta}_i(x_{ij}-x_{i.})\right\}^2 \\ &= \textstyle\sum_{j=1}^{n_i}(y_{ij}-y_{i.})^2 - \left\{\sum_{j=1}^{n_i}(x_{ij}-x_{i.})(y_{ij}-y_{i.})\right\}^2/s_i^2, \end{aligned} \tag{2.6}$$

with

$$\hat{\beta}_i = \left\{ \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2 \right\}^{-1} \sum_{j=1}^{n_i} (x_{ij} - x_{i.}) (y_{ij} - y_{i.}), \tag{2.7}$$

and

$$s_i^2 = \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2, \tag{2.8}$$

where U_i is the part of the residual part sum of squares, $\hat{\beta}_i$ the least squares estimate of the slope, and, s_i^2 the covariate within group sum of squares, corresponding to group i, after some rearrangement, (2.5) can be expressed as

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \phi | \mu_\theta, \mu_\beta, \lambda_\theta, \lambda_\beta, \nu, \zeta) = (2\pi)^{-\frac{N}{2}-m} K(\nu)^m \zeta^{\frac{\nu m}{2}} \textstyle\prod_{i=1}^m \phi_i^{-\frac{1}{2}(\nu+n_i+4)} (\lambda_\theta \lambda_\beta)^{-\frac{m}{2}} \\ \times \exp\left[-\tfrac{1}{2} \textstyle\sum_{i=1}^m \phi_i^{-1} \left\{ U_i + n_i (y_{i.} - \theta_i)^2 + s_i^2 \left(\hat{\beta}_i - \beta_i \right)^2 \right\} \right] \\ \times \exp\left[-\tfrac{1}{2} \textstyle\sum_{i=1}^m \phi_i^{-1} \left\{ \lambda_\theta^{-1} (\theta_i - \mu_\theta)^2 + \lambda_\beta^{-1} (\beta_i - \mu_\beta)^2 + \nu\zeta \right\} \right] \end{aligned} \tag{2.9}$$

with $N = \sum_{i=1}^m n_i$. Using Corollary 1.1, the two quadratic terms in θ_i and β_i in the exponent of (2.9) are equal to

$$n_i (y_{i.} - \theta_i)^2 + \lambda_\theta^{-1} (\theta_i - \mu_\theta)^2 = \left(n_i + \lambda_\theta^{-1} \right) (\theta_i - \theta_i^*)^2 + \left(n_i^{-1} + \lambda_\theta \right)^{-1} (y_{i.} - \mu_\theta)^2 \tag{2.10}$$

and

$$s_i^2 \left(\hat{\beta}_i - \beta_i \right)^2 + \lambda_\beta^{-1} (\beta_i - \mu_\beta)^2 = \left(s_i^2 + \lambda_\beta^{-1} \right) (\beta_i - \beta_i^*)^2 + \left(s_i^{-2} + \lambda_\beta \right)^{-1} \left(\hat{\beta}_i - \mu_\beta \right)^2 \tag{2.11}$$

with

$$\theta_i^* = (n_i y_{i.} + \lambda_\theta^{-1} \mu_\theta) / (n_i + \lambda_\theta^{-1}), \tag{2.12}$$

and

$$\beta_i^* = (s_i^2 \hat{\beta}_i + \lambda_\beta^{-1} \mu_\beta) / (s_i^2 + \lambda_\beta^{-1}), \tag{2.13}$$

for $i = 1, \ldots m$. Integrating out the θ_i and β_i from (2.9) using (2.10)-(2.13) together with the normal integral, we find that the joint density of the observations and the m group variances is given by

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\phi} | \mu_\theta, \mu_\beta, \lambda_\theta, \lambda_\beta, \nu, \zeta) = A \prod_{i=1}^m \phi_i^{-\frac{1}{2}(\nu+n_i+2)} \times \\ \exp\left[-\frac{1}{2} \sum_{i=1}^m \phi_i^{-1} \left\{ \nu\zeta + U_i + \left(n_i^{-1} + \lambda_\theta \right)^{-1} (y_{i.} - \mu_\theta)^2 + \left(s_i^{-2} + \lambda_\beta \right)^{-1} \left(\hat{\beta}_i - \mu_\beta \right)^2 \right\} \right], \end{aligned} \tag{2.14}$$

where

$$A = (2\pi)^{-\frac{N}{2}} K(\nu)^m \zeta^{\frac{\nu m}{2}} (\lambda_\theta \lambda_\beta)^{-\frac{m}{2}} \prod_{i=1}^{m} \left\{ \left(n_i + \lambda_\theta^{-1}\right)^{-1/2} \left(s_i^2 + \lambda_\beta^{-1}\right)^{-1/2} \right\}. \qquad (2.15)$$

A final integration of the ϕ_i, provides the joint density of the y_{ij}, unconditional upon the random effects to be

$$p(\boldsymbol{y}|\mu_\theta, \mu_\beta, \lambda_\theta, \lambda_\beta, \nu, \zeta) = A \prod_{i=1}^{m} \left\{ \Gamma\left(\tfrac{\nu+n_i}{2}\right) 2^{\frac{\nu+n_i}{2}} \right\} \times$$
$$\prod_{i=1}^{m} \left\{ \nu\zeta + U_i + \left(n_i^{-1} + \lambda_\theta\right)^{-1} (y_{i\cdot} - \mu_\theta)^2 + \left(s_i^{-2} + \lambda_\beta\right)^{-1} \left(\hat{\beta}_i - \mu_\beta\right)^2 \right\}^{-\frac{1}{2}(\nu+n_i)}. \qquad (2.16)$$

As the previous expression in brackets can be rearranged as a positive constant plus a positive definite quadratic form in the $y_{i\cdot}$, (2.16) is a product of generalized multivariate t-densities for the $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{i,n_i})$. The marginal distribution of each observation, and each observed group mean, is a generalized univariate t-distribution with ν degrees of freedom, though the observations are not independent. While the sufficient statistics U_i, $y_{i\cdot}$, and $\hat{\beta}_i$ appearing in this joint distribution are "outlier prone", this property simply encourages us to carefully consider whether or not to include outlying observations, and should not be regarded as undesirable. Conversely, an outlier resistant sampling distribution might conceal outlying observations in an unreasonable manner.

A modification to Schwarz's information criterion, BIC, (Schwarz, 1978) for this model (see Leonard and Hsu, 1999, Chapter 6, for a justification of the 2π adjustment), is

$$BIC^* = \log p(\boldsymbol{y}|\hat{\mu}_\theta, \hat{\mu}_\beta, \hat{\lambda}_\theta, \hat{\lambda}_\beta, \hat{\nu}, \hat{\zeta}) - 3 \max\left\{\log\left(N/2\pi\right), 2\right\}, \qquad (2.17)$$

where the first term denotes the supremum of (2.16), and $\hat{\mu}_\theta$, $\hat{\mu}_\beta$, $\hat{\lambda}_\theta$, $\hat{\lambda}_\beta$, $\hat{\nu}$, $\hat{\zeta}$ are the corresponding maximum likelihood estimates. The expression in (2.17) may be compared with BIC^* for any competing model, e.g., reduced forms of the current model (equal group means, parallel lines, equal variances) and models with further explanatory variables. This criterion avoids the overdependence on prior assumptions for the model parameters, inherent in Bayes factors and can perform better than Akaike's information criterion, AIC, (Akaike, 1978) for finite sample sizes (see Katz, 1981, Koehler and Murphree, 1988, Leonard and Hsu, 1999, Chapter 1). Our modification reduces to AIC whenever $\log(N/2\pi) < 2$, i.e. $N < 47$. While neither BIC nor BIC^* provide particularly good finite sample size approximations to the logarithms of prior predictive

densities, BIC^* can nevertheless be an excellent criterion for model comparison, in its own right.

Bayes factors can imply quite different prior distributions for the observations, under different models. Suppose in general that we assign prior probability ϕ to model M_1, with sampling density $p_1(\boldsymbol{y}|\boldsymbol{\theta_1})$, and prior probability $1-\phi$ to model M_2, with sampling density $p_2(\boldsymbol{y}|\boldsymbol{\theta_2})$. Then the posterior probability of M_1 is

$$\phi^* = \frac{\phi B}{\phi B + (1-\phi)}, \tag{2.18}$$

where

$$B = p_1(\boldsymbol{y})/p_2(\boldsymbol{y}), \tag{2.19}$$

the Bayes factor, and $p_1(\boldsymbol{y})$ and $p_2(\boldsymbol{y})$ are the prior predictive densities of the observation vector $\boldsymbol{y}$ under models M_1 and M_2 respectively. If B is substantially different from its neutral value ($B = 1$) for almost every realization of $\boldsymbol{y}$, then this unfortunately implies that the prior distribution for $\boldsymbol{y}$ under M_1 is substantially different from the prior distribution for $\boldsymbol{y}$ under M_2. One possible interpretation (Michael Evans, Irwin Guttman, Tom Leonard, personal communication) is that such formulations are, in a real world sense, incoherent, and we can therefore appreciate that there will be inherent difficulties with interpreting the posterior probability ϕ^* (e.g O'Hagan 1994, pp. 187-199). We here interpret B as a "value of evidence" (Good, 1991). Certainly, if B is instead used as a test statistic, then Lindley's paradox is less relevant (see Chapter 1).

Given the difficulties associated with hypothesis tests, even in the single slope and variance ANCOVA model of Chapter 1, and the additional difficulties related to the computations of the corresponding test statistics for the current more complex model, we proceed under the philosophy "the inference is in the posterior distribution", and will seek to fully interpret the marginal posterior densities of the random variables θ_i, β_i, and ϕ_i, and of the six parameters μ_θ, μ_β, ζ, λ_θ, λ_β, and ν. We will not place positive probabilities on the key hypotheses $H_\theta : \lambda_\theta = 0$, $H_\beta : \lambda_\beta = 0$, and $H_\phi : \nu = \infty$, since these would lead to the evaluation of Bayes factors in the posterior analysis. We instead consider continuous prior and posterior distributions. Careful interpretations will be required (see section 2.4). The parameter ν will also be considered with regard to prior and posterior probability mass functions which are concentrated on the positive integers.

2.2 Prior assumptions

In the prior assessment, it is assumed that μ_θ, μ_β, λ_θ, λ_β, and the pair (ζ, ν) are mutually independent. Furthermore, $\omega_1\tau_1/\lambda_\theta$ and $\omega_2\tau_2/\lambda_\beta$ possess chi-squared distributions with respective degrees of freedom ω_1 and ω_2, so that τ_1^{-1} and τ_2^{-1} are the respective prior means of λ_θ^{-1} and λ_β^{-1}. Consider the parameter λ_θ. A possible interpretation for its prior is that it provides information based on a sample of ω_1 observations with mean $\omega_1\tau_1/(\omega_1 - 2)$ for λ_θ, if $\omega_1 > 2$, while the degrees of freedom ω_1 measure the sureness of the prior belief about λ_θ. As ω_1 tends to ∞, λ_θ converges almost surely to its prior estimate τ_1. An analogous interpretation holds for the prior of λ_β.

For analytic convenience, $\nu\psi\zeta$ is taken to possess a conditional distribution, given ν, which is chi-squared with $\nu\psi\zeta_0$ degrees of freedom, so that ζ has mean ζ_0 and variance $2\zeta_0/\psi\nu$. Finally $b\nu$ has a chi-squared distribution with a degrees of freedom, so that ν has mean a/b and variance $2a/b^2$, and

$$\pi(\nu) \propto \nu^{\frac{1}{2}a-1} \exp(-\frac{1}{2}\nu b). \tag{2.20}$$

Alternatively ν can be taken to possess the discrete distribution which assigns probabilities $q_1, \ldots, q_M$ to the positive integers $1, 2, \ldots, M$.

Either prior distribution requires the specification of eight prior parameters ω_1, τ_1, ω_2, τ_2, ζ_0, ψ, a and b. We do not assume proper distributions for μ_θ and μ_β since these parameters are typically strongly controlled by the data. If a discrete prior distribution is assumed for ν, then the posterior probabilities will depend upon Bayes factors. However, in this very specific situation, our computations will justify the Bayes factors, since the solution will closely approximate the solution under a continuous prior for ν.

2.3 Posterior inference

Under the assumptions of sections 2.1 and 2.2, using (2.9) and the standard notation, $\pi(\cdot)$ and $\pi(\cdot|\boldsymbol{y})$, to denote prior and posterior densities, the joint posterior density of all 3m unobserved random variables and six unknown parameters is

$$
\begin{aligned}
&\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi},\mu_\theta,\mu_\beta,\lambda_\theta,\lambda_\beta,\nu,\zeta|\boldsymbol{y}) \propto p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi}|\mu_\theta,\mu_\beta,\lambda_\theta,\lambda_\beta,\nu,\zeta)\pi(\mu_\theta,\mu_\beta,\lambda_\theta,\lambda_\beta,\nu,\zeta) \\
&\propto \quad \pi(\nu)K(\nu)^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}\zeta^{\frac{1}{2}\nu(m+\psi\zeta_0)-1}\textstyle\prod_{i=1}^m \phi_i^{-\frac{1}{2}(\nu+n_i+4)} \\
&\times \quad \lambda_\theta^{-\frac{1}{2}(\omega_1+m+2)}\lambda_\beta^{-\frac{1}{2}(\omega_2+m+2)}\exp(-\tfrac{1}{2}\lambda_\theta^{-1}\omega_1\tau_1 - \tfrac{1}{2}\lambda_\beta^{-1}\omega_2\tau_2 - \tfrac{1}{2}\nu\psi\zeta) \\
&\times \quad \exp\left[-\tfrac{1}{2}\textstyle\sum_{i=1}^m \phi_i^{-1}\left\{U_i + n_i\left(y_{i\cdot}-\theta_i\right)^2 + s_i^2\left(\hat{\beta}_i-\beta_i\right)^2\right\}\right] \\
&\times \quad \exp\left[-\tfrac{1}{2}\textstyle\sum_{i=1}^m \phi_i^{-1}\left\{\lambda_\theta^{-1}\left(\theta_i-\mu_\theta\right)^2 + \lambda_\beta^{-1}\left(\beta_i-\mu_\beta\right)^2 + \nu\zeta\right\}\right]
\end{aligned}
\tag{2.21}
$$

where $\pi(\nu)$ denotes our chi-squared prior density for ν. Additionally, $K(\cdot)$, U_i, $\hat{\beta}_i$ and s_i^2 are defined in (2.4), (2.6), (2.7), and (2.8).

It is impossible to obtain algebraically explicit expressions for the marginal posterior densities of each of the $\theta_i, \beta_i, \phi_i, \mu_\theta, \mu_\beta, \lambda_\theta, \lambda_\beta, \nu$, and ζ. In fact, we can only obtain the joint posterior density of the six model parameters, applying Bayes theorem to (2.16), and an approximation to the posterior density of the $3m$ random effects, (see section 4.4), by performing the appropriate integrations analytically. The latter two results will be used to obtain accurate approximations to the posterior densities of the quantities of interest in Chapter 4. The remainder of this section, however, will concentrate on Markov chain Monte Carlo (MCMC) simulation methods and their application to ANCOVA models.

2.3.1 MCMC - An overview

2.3.1.1 Introduction and theoretical background

In a nutshell, Markov chain Monte Carlo, (see Gilks et al, 1995, Gelfand and Smith, 1990), is a combination of Monte Carlo integration with Markov chain theory used for performing high dimensional integrations, impossible to achieve analytically or numerically in a different way. Bayesian inference, where such integrations are routinely required to obtain marginal posterior and predictive distributions is its principal, but not exclusive, area of application. Monte Carlo integration is applicable when it is possible to draw independent samples from the required distribution (target distribution, hereafter). It forms sample averages of the quantities of interest to approximate expectations. In this situation, the Strong Law of Large Numbers (see Billingsley, 1986, p. 290) ensures that by increasing the sample size the approximation can be as accurate as desired. An application of Monte Carlo integration is presented in section 4.3.2.

MCMC provides a crucial generalization that replaces the independent draws by dependent ones drawn throughout the support of the target distribution in the correct

proportions by a long running Markov chain whose stationary distribution is the target distribution. This idea was first introduced in the literature in 1953 by Metropolis et al and in the statistical literature by Hastings (1970). However its full potential for Bayesian inference was established by Gelfand and Smith (1990), whose paper seemed to trigger the substantial future research activity that followed.

Assume that π is the stationary distribution of a Markov process $\{X_t\}$, $t = 0, 1, \ldots$, where X_t is a vector for every t. Discrete time has only been assumed for the simplification of the relevant notation. If X_t is aperiodic, irreducible and positive recurrent, then regardless of the starting state, X_0, π is the limiting distribution of successive iterates from that process and the ergodic theorem (see Roberts, 1995) ensures that averages of the form

$$f_N = \sum_{t=m+1}^{N} \frac{f(X_t)}{N-m}, \tag{2.22}$$

converge to their expectations under π, after typically discarding a sufficient burn-in of m iterations. Hence, once a Markov chain with the desired stationary distribution has been found and has been attained, it is possible to obtain the required expectations as sample averages, exactly as with simple Monte Carlo integration.

The Metropolis-Hastings algorithm, which is based on a modification by Hastings of Metropolis's method, addresses the problem of constructing a Markov chain with the desired stationary distribution π. In particular at every t, X_{t+1} is chosen between X_t and a candidate Y, where Y is drawn from a proposal distribution $q(\cdot|X_t)$ and accepted with probability $\alpha(X_t, Y)$, that is, accepted if a randomly generated uniform (0,1) deviate is less than $\alpha(X_t, Y)$, where

$$\alpha(X_t, Y) = \min\left\{1, \frac{\pi(Y)q(X_t|Y)}{\pi(X_t)q(Y|X_t)}\right\}. \tag{2.23}$$

Gilks et al (1995) provide a simple proof of the fact that the stationary distribution of $\{X_t\}$ constructed according to (2.23) is π, whatever $q(\cdot)$ is. X_t may have different dimensions at different iterations, it can be for example the vector of parameter values of a variable number of components mixture of distributions (Richardson and Green, 1997).

2.3.1.2 Implementation aspects

The original Metropolis algorithm assumed $q(X_t|Y) = q(Y|X_t)$, for all X_t and Y, and h steps to complete the transition from X_t to X_{t+1}, with h the number of blocks the vector X_t is divided into, the corresponding number of proposal distributions, candidates drawn and acceptance probabilities of the form (2.23).

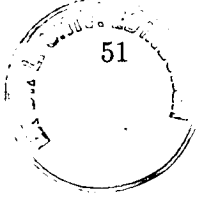
The latter blockwise implementation forms the basis of the most widely applied type of MCMC, the Gibbs sampler. This was first suggested by Geman and Geman (1984) in the context of image reconstruction related to Gibbs distributions. If the vector parameter of interest has dimension p, and has been initialized at X_0, then the Gibbs sampler constructs a Markov chain by drawing the j-th component at time $t+1$, $X_{t+1}(j)$, from its full conditional distribution, that is, the distribution of that component given the remaining ones, or

$$\pi(X_{t+1}(j)|X_{t+1}(1),\ldots,X_{t+1}(j-1),X_t(j+1),\ldots,X_t(p)). \tag{2.24}$$

Hence, the proposal distribution of the j-th component is (2.24) and substituting into (2.23) provides an acceptance probability of 1. The Gibbs sampler uses the property that π is uniquely determined by the set of its full conditional distributions (see Gelman and Speed, 1993 and 1999, for a relevant discussion), and substantially reduces the computational complexities by replacing a high dimensional problem by p unidimensional ones.

The Gibbs sampler is typically applied to hierarchical Bayes linear models because of their simple full conditionals, (e.g. Gelfand et al, 1990). However, Hobert and Casella (1996), indicate that if the prior distribution is improper, the posterior distribution can be improper too, even though the full conditionals are all proper. For example they present the one-way random effects model, with uniform prior for the grand mean and log uniform priors for the variance components. The full conditional distributions of this model are all proper, however, Hill (1965) showed that the posterior distribution is improper. If the Gibbs sampler is applied to this type of model, the positive recurrency assumption breaks down, resulting in infinite expected time of return to a Markov chain state. Therefore, either poor practical convergence or apparent inferences based upon a non-existing distribution is envisioned.

In our application of MCMC to ANCOVA models, we will use Gibbs sampling with a modification for one of the full conditional distributions, rejection/acceptance sampling



(see Ripley, 1987, pp. 60-71, Zeger and Karim, 1991). Rejection sampling is applicable in situations where it is difficult to sample a point Y from a distribution q, but easy to do so from another distribution Q, such that $q \leq AQ$, with A a known constant. It involves drawing points Y from Q and U from uniform (0,1), until $U \leq q(Y)/Q(Y)$, in which case Y is accepted. To implement rejection sampling, the normalizing constants don't need to be known, however, for computational efficiency, A has to be close to unity, because the overall acceptance probability is equal to $A^{-1} \int q(x)dx$, or A^{-1} for normalized $q(\cdot)$. Further methods for sampling from complex distributions, such as the ratio of uniforms and adaptive rejection sampling, are discussed by Gilks (1995).

In order to obtain marginal posterior densities, we will use a method first suggested by Tanner and Wong (1987) in a data augmentation context. Therefore, instead of using histograms to group successively generated values of the Gibbs sampler, or a density smoothing method, we will average the conditional densities over the simulated values. This method provides superior estimates, as indicated by Lehmann and Casella (1998, pp. 47, 258, 292), who use the Rao-Blackwell theorem to demonstrate this result. According to this well-known theorem, it is possible to obtain an estimator of reduced risk as the conditional, on a sufficient statistic, expectation of the original finite expectation and risk estimator for strictly convex loss functions. In our application the conditioning is inherent in the full conditional distributions used by the Gibbs sampler.

2.3.1.3 Convergence assessment

The Gibbs sampler is just one of the abundance of variations of MCMC. In practice, MCMC is widely used in many areas of statistics, a fact reflected in the multitude of forms it is encountered in. Although the different applications can be quite problem specific, their common characteristic is the demand for fast convergence of the created chains to the target distribution. The convergence rate depends on the relationship between the proposal and target distributions and can be possibly accelerated by variable transformations, reparametrizations (see Gelfand et al, 1995) or sampling from modified proposal distributions, that can provide faster mixing, that is, movement around the support of the target distribution (see Gilks and Roberts, 1995).

Detecting convergence, or more frequently lack of it, is of fundamental importance and has obvious repercussions to the inferences drawn. A variety of convergence assessment criteria have been and continue being proposed in the literature. Two relevant reviews are these of Cowles and Carlin (1996) and a slightly more mathematical and

up to date one by Brooks and Roberts (1998). Perhaps, Laplacian approximations (see Chapter 4) will eventually replace the practical need for many of these procedures.

Two of the most popular convergence criteria are the one of Gelman and Rubin (1992), which is based on output from many parallel chains and an ANOVA type criterion of between and within chain variances for each parameter of interest and an estimate of a shrink factor that reduces to one as the number of iterations increases at which point convergence is concluded, and that of Raftery and Lewis (1992), who use results from two-state Markov chain theory to determine the total number of iterations, number of burn-in iterations, distance of successive iterations to be included in the computations (thinning) for selected quantiles of interest, accuracy, probability of obtaining the desired accuracy and convergence tolerance.

Further convergence assessment criteria are the ones of: a) Geweke (1992), who obtained the estimated standard error of the mean of a parameter of interest based on spectral analysis arguments, but didn't provide explicitly a method of using it to assess convergence. b) Heidelberger and Welch (1983), who used earlier developments by Schruben (1982) and Schruben et al (1983) and combined Brownian bridge theory and spectral analysis for obtaining estimated confidence intervals for parameters of interest from parts of the chain that passed a stationarity test. c) Ritter and Tanner (1992), who proposed an importance sampling based method. d) Zellner and Min (1995), who suggested a criterion applicable to posterior distributions that can be factored into two parts the conditional distributions of which are easy to be sampled. e) Liu, Liu and Rubin (1992) that inferred convergence by constructing a global control variable and applying Gelman and Rubin's criterion on a number of independent sequences. f) Roberts (1992), that assessed convergence of the full posterior distribution, rather than some univariate statistic, using several independent replications of the chain. g) Yu (1994), which like Roberts's, proposed convergence assessment of the full posterior, but from a single replication of the chain, however with questionable applicability to high dimension problems. h) Brooks et al (1997), who obtained a total variation statistic aimed at producing an upper bound to the L^1 distance between full conditional kernel estimates from different chains. i) Johnson (1994), who suggested a modification of Gelman and Rubin's criterion, assessing convergence when the path of a number of independent chains have, up to a certain tolerance, converged. j) Mykland et al (1995), who relied on regeneratative simulation to assess convergence of the full posterior distribution. k) Garren and Smith (1993), who proposed a criterion based on the closeness

of the second largest eigenvalue of the kernel density transition matrix to unity, and finally, l) Yu and Mykland (1998), and its mathematical rigorization by Brooks (1998), who suggested a cusum plot based criterion applied to univariate statistics of the target distribution.

The common feature of all convergence assessment criteria is their use of MCMC output for their implementation, because there exist no useful theoretical convergence bounds of wide applicability, even for moderately complicated statistical models. However, despite their convenience, arising from their model independence, Cowles and Carlin (1996) demonstrated that in general automated convergence monitoring is unrealistic and that some of the proposed convergence assessment criteria even fail to detect the type of lack of convergence they were designed to detect originally. The problem of typically not knowing the stationary distribution appears to be insurmountable , since even if somebody elects to use many parallel chains, which is a practice criticized by many because of its computational inefficiency, only ensures that the parallel chains eventually converge to the same distribution, which is not necessarily the true one, while the criteria that are founded on some univariate function of the MCMC output give no guarantees about the behavior of other, to a certain extent arbitrary, functions that may give contradictory convergence information. A recent attempt of tackling MCMC convergence is that of Guinenneuc-Jouyaux and Robert (1998), who advocate a discretization method for continuous Markov chains that simplifies the theoretical results and accelerates the convergence.

2.3.2 Gibbs sampling application to ANCOVA

The full conditional density of any particular parameter of interest, required for the Gibbs sampling iterations, is proportional, as a function of the parameter of interest, to the joint posterior density (2.21). Therefore, using (2.21) and (2.10), and adopting the notation $\pi^*(\cdot)$ to denote a full conditional density, we can conclude that the full conditional density of θ_i is

$$\pi^*(\theta_i) \propto \exp\left\{-\frac{1}{2}(n_i + \lambda_\theta^{-1})\phi_i^{-1}(\theta_i - \theta_i^*)^2\right\}, \tag{2.25}$$

and consequently θ_i is conditionally normal with mean θ_i^* and variance $(n_i + \lambda_\theta^{-1})^{-1}\phi_i$. Similarly,

$$\pi^*(\beta_i) \propto \exp\left\{-\frac{1}{2}(s_i^2 + \lambda_\beta^{-1})\phi_i^{-1}(\beta_i - \beta_i^*)^2\right\}, \tag{2.26}$$

so that β_i is also conditionally normal. Conditionally on the ϕ_i and the six model parameters, the θ_i and β_i are independent in the posterior, by factorizing (2.21). Thus, the full conditional distribution of the adjusted group mean ξ_i, defined in (2.2), is also normal with mean ξ_i^* and variance $V_{\xi_i^*}\phi_i$, where

$$\xi_i^* = \theta_i^* - \beta_i^*(x_{i.} - x_{..}), \tag{2.27}$$

and

$$V_{\xi_i^*} = (n_i + \lambda_\theta^{-1})^{-1} + (s_i^2 + \lambda_\beta^{-1})^{-1}(x_{i.} - x_{..})^2. \tag{2.28}$$

It is not necessary to simulate from this particular full conditional distribution. However, this result is needed when calculating the marginal posterior of ξ_i by the density averaging method described at the end of section 2.3.1.2. The full conditional density of ϕ_i is

$$\pi^*(\phi_i) \propto \phi_i^{-\frac{1}{2}(\nu_i^*+2)} \exp\left(-\frac{\nu_i^*\Lambda_i}{2\phi_i}\right), \tag{2.29}$$

with

$$\nu_i^* = \nu + n_i + 2, \tag{2.30}$$

and

$$\nu_i^*\Lambda_i = \nu\zeta + U_i + W_{i1} + W_{i2}, \tag{2.31}$$

where

$$W_{i1} = n_i\,(y_{i.} - \theta_i)^2 + s_i^2\left(\hat{\beta}_i - \beta_i\right)^2, \tag{2.32}$$

and

$$W_{i2} = \lambda_\theta^{-1}\,(\theta_i - \mu_\theta)^2 + \lambda_\beta^{-1}\,(\beta_i - \mu_\beta)^2\,. \tag{2.33}$$

Thus, the density of ϕ_i has the same parametric form as the inverted chi-squared density (2.3), and hence, $\nu_i^*\Lambda_i/\phi_i$ is, given ν_i^* and Λ_i, chi-squared with ν_i^* degrees of freedom. Consider now the full conditional density of μ_θ. By Corollary 1.2, as a function of μ_θ,

$$\sum_{i=1}^{m} \phi_i^{-1}\,(\theta_i - \mu_\theta)^2 \propto \sum_{i=1}^{m} \phi_i^{-1}\,(\mu_\theta - \mu_\theta^*)^2\,, \tag{2.34}$$

with

$$\mu_\theta^* = \sum_{i=1}^{m} \phi_i^{-1}\theta_i / \sum_{i=1}^{m} \phi_i^{-1}, \tag{2.35}$$

and consequently, from (2.21),

$$\pi^*(\mu_\theta) \propto \exp\left\{-\frac{1}{2}\lambda_\theta^{-1}\sum_{i=1}^{m}\phi_i^{-1}(\mu_\theta - \mu_\theta^*)^2\right\}, \tag{2.36}$$

so that μ_θ is normally distributed with mean μ_θ^* and precision $\lambda_\theta^{-1}\sum_{i=1}^{m}\phi_i^{-1}$. Similarly,

$$\pi^*(\mu_\beta) \propto \exp\left\{-\frac{1}{2}\lambda_\beta^{-1}\sum_{i=1}^{m}\phi_i^{-1}(\mu_\beta - \mu_\beta^*)^2\right\}, \tag{2.37}$$

with

$$\mu_\beta^* = \sum_{i=1}^{m}\phi_i^{-1}\beta_i / \sum_{i=1}^{m}\phi_i^{-1}, \tag{2.38}$$

which gives that μ_β is normally distributed with mean μ_β^* and precision $\lambda_\beta^{-1}\sum_{i=1}^{m}\phi_i^{-1}$. From (2.21) it follows that $\pi^*(\lambda_\theta)$, $\pi^*(\lambda_\beta)$, and $\pi^*(\zeta)$, are of the same inverted chi-squared form as (2.3), but with different parameters (see section 2.3.3). Finally, the full conditional posterior density of ν, in the continuous case, is

$$\pi^*(\nu) \propto \pi(\nu)K(\nu)^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}\zeta^{\frac{1}{2}\nu(m+\psi\zeta_0)}\prod_{i=1}^{m}\phi_i^{-\nu/2}\exp\left\{-\frac{1}{2}\nu\zeta(\sum_{i=1}^{m}\phi_i^{-1}+\psi)\right\}, \tag{2.39}$$

where $K(\cdot)$ is defined in (2.4). The application of Stirling's approximation, $\Gamma(\nu) \sim \sqrt{2\pi}e^{-\nu}\nu^{\nu-1/2}$, gives $K(\nu) \sim (\nu/2)^{1/2}e^{\nu/2}/\sqrt{2\pi}$. Therefore (2.39) may be approximated by the gamma density

$$\widetilde{\pi}^*(\nu) \propto \nu^{\frac{1}{2}(m+a+1)-1}\exp(-\frac{1}{2}\nu B) \qquad (0 < \nu < \infty), \tag{2.40}$$

where

$$B = -m(1+\log\zeta) + \zeta(\psi + \sum_{i=1}^{m}\phi_i^{-1}) + \psi\zeta_0\{\log(\zeta_0/\zeta) - 1\} + \sum_{i=1}^{m}\log\phi_i + b. \tag{2.41}$$

A similar device was employed by Lindley (1971) when approximating joint modal estimates under his alternative formulation.

Subject to the approximation (2.40), it is straightforward to employ Markov Chain Monte Carlo techniques to calculate the posterior densities of all unknown quantities of interest. It is possible to extend these techniques to incorporate the exact density for ν in (2.39) by acceptance sampling from (2.40). In the latter case, the envelope function is the same as (2.39), with the gamma function terms replaced by the corresponding Stirling approximations, and envelope constant equal to one. However, in our numerical investigations, we have found that this makes negligible difference to our marginal

posterior densities unless ν is estimated to be very large, so that the ϕ_i are close to equal. In such situations we recommend either referring to a parallel, less complex analysis, where the ϕ_i are assumed equal, or to the methodology described as follows.

As an alternative to the preceding acceptance sampling, we assign discrete prior probabilities for the parameter ν on the integers $\{1, 2, \ldots, M\}$, and then calculate all marginals using straightforward MCMC, and the full conditional distribution for ν as described in the following section. This method works well, even if the preceding acceptance sampling fails, and gives very similar results when the acceptance sampling does not fail, if the prior probabilities are taken to be proportional to the previously specified prior density. Additionally, the difference in computer time is small. Our treatment of the parameter ν, either as a continuous or a discrete variable, and its full conditional distribution, can be contrasted with Besag's and Higdon's (1999) of the two parameters of an inverted gamma distribution of conditionally independent variances corresponding to different plots in a hierarchical t model exploring variety and spatial effects in agricultural experiments. In their applications, they adopt a discretization, on very few integer values of the parameters, that greatly simplifies their MCMC iterations. Treating ν as a discrete variable can also be considered a version of the griddy Gibbs sampler (Ritter and Tanner, 1992). When it is difficult to sample from a univariate full conditional distribution, the previous authors obtain an approximation to the inverse cumulative distribution function by evaluating the full conditional distribution on a grid. Then, they draw a random uniform (0,1) deviate and transform via the inverse c.d.f. to a draw from the full conditional distribution of interest. We use a quite natural discretization for the parameter ν, which provides us with a simple grid.

In our case studies, we will be using long single chains, comprising of 10,000 burn-in iterations and a further 50,000 iterations from which we will be averaging the full conditional densities to obtain our final answers. The long chain provides higher chance of reaching stationarity. The burn-in computational time is negligible (less than 5 seconds for all $3m + 6$ random effects and parameters), and its adequate length will be graphically confirmed by displays of the trajectories of five number summaries (minimum, first quartile, median, third quartile, maximum) for all parameters of interest, across a number of short parallel chains. These percentiles typically settle down after few tens of iterations. For the last 50,000 iterations we will be checking the five successive 10,000 iterations averages for consistency, and average the five partial averages to conclude the computation of every marginal posterior distribution of interest. Following McEachern

and Berliner (1994), who showed that it results to poorer estimates, thinning won't be used. Typically, 1,000 burn-in iterations followed by averaging 3,000 densities give results very close to the true (10,000+50,000 iterations) ones. No previously adopted formal convergence assessment criteria will be used, but we will confirm convergence by obtaining estimates of all the marginal posterior distributions using Laplacian approximations and contrasting pairs of graphical displays.

2.3.3 Full conditional distributions

Each of the following statements is made conditionally upon the data, and all other unknown random variables and parameters in the model:

A1: For $i = 1, \ldots, m$, the θ_i are independent and normally distributed with respective means θ_i^*, defined in (2.12), and variances $(n_i + \lambda_\theta^{-1})^{-1}\phi_i$.

A2: For $i = 1, \ldots, m$, the β_i are independent and normally distributed with respective means β_i^*, defined in (2.13), and respective variances $(s_i^2 + \lambda_\beta^{-1})^{-1}\phi_i$, where s_i^2 is defined in (2.8).

A3: For $i = 1, \ldots, m$, the ϕ_i are independent, and $\nu_i^* \Lambda_i / \phi_i$ possesses a chi-squared distribution with ν_i^* degrees of freedom, with ν_i^* and Λ_i defined in (2.30) and (2.31).

N.B. All the full conditional distributions in A1-A3 can be regarded as relating to the sampling model described in section 2.1. They do not refer to the prior assumptions of section 2.2, which are now addressed.

A4: The mean μ_θ is normally distributed with mean μ_θ^*, defined in (2.35), and variance $\lambda_\theta / \sum_{i=1}^m \phi_i^{-1}$.

A5: The mean μ_β is normally distributed with mean μ_β^*, defined in (2.38), and variance $\lambda_\beta / \sum_{i=1}^m \phi_i^{-1}$.

A6: For the parameter λ_θ, the quantity $(\omega_1 + m)\lambda_\theta^*/\lambda_\theta$ has a chi-squared distribution with $\omega_1 + m$ degrees of freedom, where

$$\lambda_\theta^* = \left\{ \omega_1 \tau_1 + \sum_{i=1}^m \phi_i^{-1} (\theta_i - \mu_\theta)^2 \right\} / (\omega_1 + m). \tag{2.42}$$

A7: For the parameter λ_β, the quantity $(\omega_2 + m)\lambda_\beta^*/\lambda_\beta$ has a chi-squared distribution with $\omega_2 + m$ degrees of freedom, where

$$\lambda_\beta^* = \left\{ \omega_2 \tau_2 + \sum_{i=1}^m \phi_i^{-1} (\beta_i - \mu_\beta)^2 \right\} / (\omega_2 + m). \tag{2.43}$$

A8: For the parameter ζ, the quantity $\nu(\sum_{i=1}^{m} \phi_i^{-1} + \psi)\zeta$ has a chi-squared distribution with $\nu(\psi\zeta_0 + m)$ degrees of freedom.

A9: The full conditional distribution of ν is described in equation (2.39), for situations where the prior density of ν is continuous. If instead ν possesses the discrete distribution already indicated in section 2.3.2, then for $1, 2, \ldots, M$, the posterior probability, that $\nu = i$, is $Q(i)/\sum_{\nu=1}^{N} Q(\nu)$, where $Q(\nu)$ represents the right hand side of equation (2.39), but with $\pi(\nu)$ replaced by the prior probability q_ν. One possible choice is to take the prior probabilities q_i, that $\nu = i$, to be proportional to $\pi(i)$, where $\pi(\nu)$ is the density of a gamma distribution with parameters $a/2$ and $b/2$.

A10: For $i = 1, \ldots, m$, the ξ_i are independent and normally distributed with respective means ξ_i^* defined in (2.27), and variances $V_{\xi_i^*}\phi_i$, with $V_{\xi_i^*}$ defined in (2.28).

2.4 Interpretation of scale parameters

Consider the parameter λ_θ initially. It has a posterior density concentrated on the positive real numbers, $(0, \infty)$. Suppose we are investigating the hypothesis $\lambda_\theta = \lambda_\theta^0$ against $\lambda_\theta \neq \lambda_\theta^0$, with $\lambda_\theta^0 > 0$. It is a well known procedure in the Bayesian literature (e.g. and Leonard and Hsu pp. 109-110), though usually conflicting with Bayes factors, to use a_p, the posterior probability that $\lambda_\theta \leq \lambda_\theta^0$. Then refute H_0, if a_p is too small or too large, for example, if $a_p \leq 0.005$ or $a_p \geq 0.995$, for two-sided tests. That is, a_p plays the role of a "Bayesian significance probability". An analogous method can be employed for one-sided tests.

A problem would arise if the value we wish to test is a boundary point of the parameter space. In the λ_θ situation, if we want to test the null hypothesis H_0: $\lambda_\theta = 0$, against H_1: $\lambda_\theta > 0$, then a_p always equals 0. One possibility for interpreting the posterior density is by reference to some important, in the context of the problem, fixed value d together with a_p, the posterior probability that $\lambda_\theta \leq d$, (see, for example, Carlin and Louis, 1996, p. 45). Again, we would tend to reject the null hypothesis, if the previous probability is too small. However, there is no reason for comparing a_p with 1% or 5% (see also the developments in Chapter 1). The decision about the validity of the tested hypothesis is rather the product of experience from analyzing lots of data sets. In the ANCOVA problem, intuitive choices for λ_θ are $\min n_i^{-1}$ and $\max n_i^{-1}$, for λ_β are $\min s_i^{-2}$ and $\max s_i^{-2}$, and for ν are $\max n_i$ and $\min n_i$, as the discussion of the remainder of this section illustrates. We remember that we cannot compare

the subsequent Bayesian significance probabilities with 1% and 5%. We instead use interpretations based on simulations to suggest what to do. Our later computations suggested that the first of each of the preceding pair of choices for the three ANCOVA model parameters are preferable. In general we suggest either using practical experience or simulated data, to judge these Bayesian significance probabilities.

The parameters λ_θ, λ_β and ν may be interpreted by reference to the first three conditional distributions of section 2.3.3. The conditional median of the θ_i, given $y_{i.}$, θ_i^*, as described in (2.12), is the weighted average of the mean μ_θ and the observed value $y_{i.}$, with weights proportional to the corresponding precisions. Hence, θ_i^* will be closer to the parameter μ_θ than to $y_{i.}$, if and only if $\lambda_\theta < n_i^{-1}$. This can be interpreted as indicating that "if $\lambda_\theta < \min_{i=1,\ldots,m} n_i^{-1}$ then our model is more supportive of a hypothesis H_θ, which takes all the θ_i to be equal to μ_θ, than a hypothesis H_θ^*, which takes the θ_i to be mutually unequal fixed effects. If $\lambda_\theta > \max_{i=1,\ldots,m} n_i^{-1}$, then the reverse is true. A value of λ_θ between these limits suggests that some subset of the θ_i may be equal, and the remainder unequal".

Consideration of (2.13) suggests a similar form of weighted average for the conditional median of β_i, β_i^*, and a similar interpretation of the parameter λ_β, but with θ_i, n_i, μ_θ, H_θ, and H_θ^* respectively replaced by β_i, s_i^2, μ_β, H_β, and H_β^*, where H_β^* denotes the hypothesis that the β_i are mutually unequal fixed effects. In particular the equality hypothesis H_β is preferred to H_β^* if $\lambda_\beta < \min_{i=1,\ldots,m} s_i^{-2}$.

Equations (2.31)-(2.33) tell us that the conditional mean of ϕ_i, given $y_{i.}$ is

$$\phi_i^* = (\nu\zeta + n_i U_i^*)/(\nu + n_i), \tag{2.44}$$

where U_i^* denotes the conditional expectation, given $y_{i.}$, of $n_i^{-1}(U_i + W_{i1} + W_{i2})$, and U_i, W_{i1}, and W_{i2} satisfy (2.6), (2.32) and (2.33). Since the U_i^* depend upon all our random effects assumptions, the following interpretation is needed: "If $\nu > \max_{i=1,\ldots,m} n_i$ then our model is more supportive of H_ϕ: all ϕ_i equal to ζ, than H_ϕ^*: our entire random effects model is true, with $\nu < \infty$. If $\nu < \min_{i=1,\ldots,m} n_i$, then H_ϕ is preferred. A value of ν between those limits suggests that some subset of the ϕ_i may be equal, and the remainder unequal".

2.5 Multivariate generalizations of the ANCOVA model

2.5.1 Sampling model

Consider m groups of observations $\{y_{ij};\ j = 1,\ldots,n_i\}$, for $i = 1,\ldots m$, $(m \geq 4)$, each assigned two vectors of explanatory variables $\boldsymbol{u_{ij}}$ and $\boldsymbol{x_{ij}}$ of dimension d_1 and d_2 respectively. Consider also the two sets of random effects $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_m\}$ and $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m\}$, with $\boldsymbol{\gamma}_i$ $(d_1 \times 1)$, for $i = 1,\ldots m$, the main parameters of interest (means, slopes, adjusted means), and $\boldsymbol{\beta}_i$ $(d_2 \times 1)$, for $i = 1,\ldots m$, the secondary ones, and finally a third set random effects, the conditional variances $\{\phi_1, \phi_2, \ldots, \phi_m\}$. The means of the three sets of random effects are $\boldsymbol{\mu_\gamma}$, $\boldsymbol{\mu_\beta}$ and ζ^{-1}. Three further parameters $\boldsymbol{Q_\gamma}$ $(d_1 \times d_1)$, $\boldsymbol{Q_\beta}$ $(d_2 \times d_2)$ and ν^{-1} measure departures of the random effects from their corresponding means.

The following hierarchy describes our sampling model:

(a) Conditionally on $\boldsymbol{\gamma}_i$, $\boldsymbol{\beta}_i$ and ϕ_i, the observations y_{ij} are independent, for $i = 1,\ldots,m$ and $j = 1,\ldots,n_i$, and normally distributed with means

$$\boldsymbol{u_{ij}}^T\boldsymbol{\gamma}_i + \boldsymbol{x_{ij}}^T\boldsymbol{\beta}_i \qquad (i = 1,\ldots,m), \tag{2.45}$$

and variances ϕ_i, with the rows of the specified design vectors $\boldsymbol{u_{ij}}$ and $\boldsymbol{x_{ij}}$ corresponding to main effects appropriately centered, and chosen to ensure that all matrices appearing in (2.52) and (2.53) exist.

(b) Conditionally on the ϕ_i and the model parameters $\boldsymbol{\mu_\gamma}$, $\boldsymbol{\mu_\beta}$, $\boldsymbol{Q_\gamma}$ and $\boldsymbol{Q_\beta}$, the $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}_i$ are mutually independent and normally distributed, with the $\boldsymbol{\gamma}_i$ having mean $\boldsymbol{\mu_\gamma}$ and variance $\boldsymbol{Q_\gamma}\phi_i$, for $i = 1,\ldots m$, and the $\boldsymbol{\beta}_i$ mean $\boldsymbol{\mu_\beta}$ and variance $\boldsymbol{Q_\beta}\phi_i$, for $i = 1,\ldots,m$. A special case of this model is the mixed effects model, where $|\boldsymbol{Q_\gamma}| \to \infty$ and hence the $\boldsymbol{\gamma}_i$ are fixed effects and only the $\boldsymbol{\beta}_i$ are random.

(c) The ϕ_i, given ζ and ν, are independent with densities identical to the ones in (2.3), that is scaled inverse chi-squared distributions with ν degrees of freedom.

The joint distribution of the observations and the random effects is

$$\begin{aligned}
p(\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\phi}|,\boldsymbol{\mu_\gamma},\boldsymbol{\mu_\beta},\boldsymbol{Q_\gamma},\boldsymbol{Q_\beta},\nu,\zeta) = \\
&= \textstyle\prod_{i=1}^m \prod_{j=1}^{n_i} (2\pi\phi_i)^{-1/2} \exp\left\{-\frac{1}{2\phi_i}\left(y_{ij} - \boldsymbol{u_{ij}}^T\boldsymbol{\gamma}_i - \boldsymbol{x_{ij}}^T\boldsymbol{\beta}_i\right)^2\right\} \\
&\times \textstyle\prod_{i=1}^m (2\pi\,|\phi_i\boldsymbol{Q_\gamma}|)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}_i - \boldsymbol{\mu_\gamma})^T(\phi_i\boldsymbol{Q_\gamma})^{-1}(\boldsymbol{\gamma}_i - \boldsymbol{\mu_\gamma})\right\} \\
&\times \textstyle\prod_{i=1}^m (2\pi\,|\phi_i\boldsymbol{Q_\beta}|)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{\mu_\beta})^T(\phi_i\boldsymbol{Q_\beta})^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\mu_\beta})\right\} \\
&\times \textstyle\prod_{i=1}^m K(\nu)\zeta^{\nu/2}\phi_i^{-(\nu/2+1)} \exp\left\{-\frac{\nu\zeta}{2\phi_i}\right\},
\end{aligned} \tag{2.46}$$

with $K(\cdot)$ defined in (2.4). Let

$$\boldsymbol{U_i} = \sum_{j=1}^{n_i} \boldsymbol{u_{ij}} \boldsymbol{u_{ij}}^T, \tag{2.47}$$

$$\boldsymbol{X_i} = \sum_{j=1}^{n_i} \boldsymbol{x_{ij}} \boldsymbol{x_{ij}}^T, \tag{2.48}$$

$$\boldsymbol{U} \circ \boldsymbol{X_i} = \sum_{j=1}^{n_i} \boldsymbol{u_{ij}} \boldsymbol{x_{ij}}^T \tag{2.49}$$

$$\text{and} \quad \boldsymbol{X} \circ \boldsymbol{U_i} = \sum_{j=1}^{n_i} \boldsymbol{x_{ij}} \boldsymbol{u_{ij}}^T. \tag{2.50}$$

Maximizing

$$\sum_{j=1}^{n_i} \left(y_{ij} - \boldsymbol{u_{ij}}^T \boldsymbol{\gamma_i} - \boldsymbol{x_{ij}}^T \boldsymbol{\beta_i} \right)^2 \tag{2.51}$$

with respect to $\boldsymbol{\gamma_i}$ and $\boldsymbol{\beta_i}$, we obtain the respective least squares estimates, $\hat{\boldsymbol{\gamma}}_{\boldsymbol{i}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$, which are

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{i}} = \left(\boldsymbol{I} - \boldsymbol{U_i}^{-1} \boldsymbol{U} \circ \boldsymbol{X_i} \boldsymbol{X_i}^{-1} \boldsymbol{X} \circ \boldsymbol{U_i} \right)^{-1} \boldsymbol{U_i}^{-1} \left(\sum_{j=1}^{n_i} \boldsymbol{u_{ij}} y_{ij} - \boldsymbol{U} \circ \boldsymbol{X_i} \boldsymbol{X_i}^{-1} \sum_{j=1}^{n_i} \boldsymbol{x_{ij}} y_{ij} \right), \tag{2.52}$$

and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \left(\boldsymbol{I} - \boldsymbol{X_i}^{-1} \boldsymbol{X} \circ \boldsymbol{U_i} \boldsymbol{U_i}^{-1} \boldsymbol{U} \circ \boldsymbol{X_i} \right)^{-1} \boldsymbol{X_i}^{-1} \left(\sum_{j=1}^{n_i} \boldsymbol{x_{ij}} y_{ij} - \boldsymbol{X} \circ \boldsymbol{U_i} \boldsymbol{U_i}^{-1} \sum_{j=1}^{n_i} \boldsymbol{u_{ij}} y_{ij} \right), \tag{2.53}$$

with $\boldsymbol{I}$ denoting the identity matrix of the appropriate dimensions. Setting

$$S_i^2 = \sum_{j=1}^{n_i} \left(y_{ij} - \boldsymbol{u_{ij}}^T \hat{\boldsymbol{\gamma}}_{\boldsymbol{i}} - \boldsymbol{x_{ij}}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \right)^2, \tag{2.54}$$

and

$$\boldsymbol{V_{i,1}} = \begin{bmatrix} \boldsymbol{U_i} & \boldsymbol{U} \circ \boldsymbol{X_i} \\ \boldsymbol{X} \circ \boldsymbol{U_i} & \boldsymbol{X_i} \end{bmatrix}, \tag{2.55}$$

(2.51) can be written as

$$S_i^2 + \boldsymbol{w_{i,1}}^T \boldsymbol{V_{i,1}} \boldsymbol{w_{i,1}}, \tag{2.56}$$

with

$$w_{i,1} = \left(\gamma_i - \hat{\gamma}_i,\, \beta_i - \hat{\beta}_i\right)^T . \tag{2.57}$$

Completing the square for γ_i and β_i in (2.46) using Lemma 1.1 and (2.56), and setting

$$V_2 = \begin{bmatrix} {Q_\gamma}^{-1} & 0 \\ 0 & {Q_\beta}^{-1} \end{bmatrix} \tag{2.58}$$

and

$$w_{i,2} = (\gamma_i - \mu_\gamma,\, \beta_i - \mu_\beta)^T , \tag{2.59}$$

we find that

$${w_{i,1}}^T V_{i,1} w_{i,1} + {w_{i,2}}^T V_2 w_{i,2} = {w_{i,3}}^T (V_{i,1} + V_2) w_{i,3} + (w_{i,2} - w_{i,1})^T V_{i,3} (w_{i,2} - w_{i,1}), \tag{2.60}$$

with

$$V_{i,3} = V_{i,1}(V_{i,1} + V_2)^{-1} V_2 \tag{2.61}$$

and

$$w_{i,3} = (\gamma_i - {\gamma_i}^*,\, \beta_i - {\beta_i}^*)^T , \tag{2.62}$$

where

$$\begin{pmatrix} {\gamma_i}^* \\ {\beta_i}^* \end{pmatrix} = (V_{i,1} + V_2)^{-1} \begin{pmatrix} U \circ X_i \hat{\beta}_i + U_i \hat{\gamma}_i + {Q_\gamma}^{-1} \mu_\gamma \\ X_i \hat{\beta}_i + X \circ U_i \hat{\gamma}_i + {Q_\beta}^{-1} \mu_\beta \end{pmatrix} . \tag{2.63}$$

The matrix inversion of (2.61) can be performed using the following Lemma, found in Seber (1977, pp. 390-391).

Lemma 2.1. If A is $(p \times p)$ and D is $(r \times r)$ symmetric matrices, B is $(p \times r)$ matrix, and all inverses exist, then

$$\begin{bmatrix} A & B \\ B^T & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + F E^{-1} F^T & -F E^{-1} \\ -E^{-1} F^T & E^{-1} \end{bmatrix} , \tag{2.64}$$

with $E = D - B^T A^{-1} B$ and $F = A^{-1} B$.

This result can be verified by performing the multiplication of the matrix times its inverse and of the inverse times the original matrix, which are equal to the identity matrix, and then using the uniqueness of the inverse to complete the proof.

Applying Lemma 2.1 we can obtain that

$$(\boldsymbol{V}_{i,1}+\boldsymbol{V}_2)^{-1}=\begin{bmatrix} \boldsymbol{A}_i{}^{-1}+\boldsymbol{F}_i\boldsymbol{E}_i{}^{-1}\boldsymbol{F}_i{}^T & -\boldsymbol{F}_i\boldsymbol{E}_i{}^{-1} \\ -\boldsymbol{E}_i{}^{-1}\boldsymbol{F}_i{}^T & \boldsymbol{E}_i{}^{-1} \end{bmatrix}, \tag{2.65}$$

with

$$\boldsymbol{A}_i = \boldsymbol{U}_i+\boldsymbol{Q}_\gamma{}^{-1}, \tag{2.66}$$

$$\boldsymbol{B}_i = \boldsymbol{U}\circ\boldsymbol{X}_i, \tag{2.67}$$

$$\boldsymbol{D}_i = \boldsymbol{X}_i+\boldsymbol{Q}_\beta{}^{-1}, \tag{2.68}$$

$$\boldsymbol{E}_i = \boldsymbol{D}_i-\boldsymbol{B}_i{}^T\boldsymbol{A}_i{}^{-1}\boldsymbol{B}_i, \tag{2.69}$$

$$\text{and}\quad \boldsymbol{F}_i = \boldsymbol{A}_i{}^{-1}\boldsymbol{B}_i. \tag{2.70}$$

Similarly to the model of section 2.1, it is possible to integrate out, using the multivariate normal integral definition, the γ_i and β_i, and subsequently the variances, ϕ_i, to obtain the joint density of the observations, y_{ij}, unconditionally upon the random effects. The joint density is thus equal to

$$\begin{aligned} &p(\boldsymbol{y}|\boldsymbol{\mu}_\gamma,\boldsymbol{\mu}_\beta,\boldsymbol{Q}_\gamma,\boldsymbol{Q}_\beta,\nu,\zeta) = \\ &\quad = (2\pi)^{-\frac{N}{2}}K(\nu)^m\zeta^{\frac{\nu m}{2}}\prod_{i=1}^m\left\{\Gamma\left(\tfrac{\nu+n_i}{2}\right)2^{\frac{\nu+n_i}{2}}\right\}|\boldsymbol{Q}_\gamma\boldsymbol{Q}_\beta|^{-m/2}\prod_{i=1}^m|\boldsymbol{V}_{i,1}+\boldsymbol{V}_2|^{-1/2} \\ &\quad \times\prod_{i=1}^m\left\{\nu\zeta+S_i^2+(\boldsymbol{w}_{i,2}-\boldsymbol{w}_{i,1})^T\boldsymbol{V}_{i,3}(\boldsymbol{w}_{i,2}-\boldsymbol{w}_{i,1})\right\}^{-\frac{1}{2}(\nu+n_i)}, \end{aligned} \tag{2.71}$$

with $N=\sum_{i=1}^m n_i$. Having obtained the maximum likelihood estimates of $\boldsymbol{\mu}_\gamma$, $\boldsymbol{\mu}_\beta$, $\boldsymbol{Q}_\gamma$, $\boldsymbol{Q}_\beta$, ν and ζ using a suitable maximization routine, it is possible to construct a Bayesian information criterion, similar to (2.17) to choose between competing models (equal variances, further covariates). The number of parameters to be estimated in the current model is $d_1(d_1+3)/2+d_2(d_2+3)/2+2$.

2.5.2 Prior to posterior inference

In the prior assessment, we assume that $\boldsymbol{\mu}_\gamma$, $\boldsymbol{\mu}_\beta$, $\boldsymbol{Q}_\gamma$, $\boldsymbol{Q}_\beta$ and the pair (ζ,ν) are independent. For the inverse of the matrices $\boldsymbol{Q}_\gamma$ and $\boldsymbol{Q}_\beta$ we will assume Wishart distributions. The inverse of a symmetric positive definite matrix $\boldsymbol{U}$ $(p\times p)$, has a Wishart distribution with k degrees of freedom and scale matrix $\boldsymbol{\Sigma}$, also positive definite, if and

only if the probability density function of $\boldsymbol{U}$ is

$$p(\boldsymbol{U}) = \frac{|\boldsymbol{U}|^{-\frac{k+p+1}{2}} \exp\left\{-\frac{1}{2}\text{trace}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{-1}\right)\right\}}{2^{kp/2}\pi^{p(p-1)/4}\,|\boldsymbol{\Sigma}|^{k/2}\prod_{i=1}^{p}\Gamma\left\{\frac{1}{2}(k+1-i)\right\}}, \tag{2.72}$$

(Mardia et al, 1979, p. 85). We assume that the inverse of $\boldsymbol{Q}_\gamma$ has a Wishart distribution with k_γ degrees of freedom and scale matrix $\boldsymbol{\Sigma}_\gamma$ and the inverse of $\boldsymbol{Q}_\beta$ has the same distribution with parameters k_β and $\boldsymbol{\Sigma}_\beta$, which can be altered as part of a sensitivity analysis. We will also assume flat priors for $\boldsymbol{\mu}_\gamma$ and $\boldsymbol{\mu}_\beta$ and the same chi-squared priors as in the single covariate model for the multiples of ν and ζ (section 2.2). Using this formulation, the posterior density of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and the parameters $\boldsymbol{\mu}_\gamma$, $\boldsymbol{\mu}_\beta$, $\boldsymbol{Q}_\gamma$, $\boldsymbol{Q}_\beta$, ν and ζ is

$$\begin{aligned}
&\pi(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\mu}_\gamma,\boldsymbol{\mu}_\beta,\boldsymbol{Q}_\gamma,\boldsymbol{Q}_\beta,\nu,\zeta|\boldsymbol{y}) \propto \\
&\propto \pi(\nu)K(\nu)^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}\zeta^{\frac{1}{2}\nu(m+\psi\zeta_0)-1}\exp(-\tfrac{1}{2}\nu\psi\zeta)\textstyle\prod_{i=1}^{m}\phi_i^{-\frac{1}{2}(\nu+n_i+2+d_1+d_2)} \\
&\times |\boldsymbol{Q}_\gamma|^{-\frac{1}{2}(k_\gamma+d_1+m+1)}\,|\boldsymbol{Q}_\beta|^{-\frac{1}{2}(k_\beta+d_2+m+1)}\exp\left\{-\tfrac{1}{2}\text{trace}\left(\boldsymbol{\Sigma}_\gamma{}^{-1}\boldsymbol{Q}_\gamma{}^{-1}+\boldsymbol{\Sigma}_\beta{}^{-1}\boldsymbol{Q}_\beta{}^{-1}\right)\right\} \\
&\times \exp\left[-\tfrac{1}{2}\textstyle\sum_{i=1}^{m}\phi_i^{-1}\left\{\nu\zeta+S_i^2+\boldsymbol{w}_{i,1}{}^T\boldsymbol{V}_{i,1}\boldsymbol{w}_{i,1}+\boldsymbol{w}_{i,2}{}^T\boldsymbol{V}_2\boldsymbol{w}_{i,2}\right\}\right],
\end{aligned} \tag{2.73}$$

with $\pi(\nu)$ denoting the prior of ν, which may be either discrete or the continuous chi-squared already described.

Obtaining the marginal posterior distributions of every quantity of interest involves another application of the Gibbs sampler. To derive the full conditional distributions, we need to observe that the following results hold:

$$\sum_{i=1}^{m}\phi_i^{-1}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_\gamma)^T\boldsymbol{Q}_\gamma{}^{-1}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_\gamma) = \text{trace}\left\{\boldsymbol{Q}_\gamma{}^{-1}\sum_{i=1}^{m}\phi_i^{-1}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_\gamma)(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_\gamma)^T\right\}, \tag{2.74}$$

and, by Lemma 1.2,

$$\sum_{i=1}^{m}\phi_i^{-1}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_\gamma)^T\boldsymbol{Q}_\gamma{}^{-1}(\boldsymbol{\gamma}_i-\boldsymbol{\mu}_\gamma) \propto \text{trace}\left\{\boldsymbol{Q}_\gamma{}^{-1}\sum_{i=1}^{m}\phi_i^{-1}(\boldsymbol{\mu}_\gamma-\boldsymbol{\mu}_\gamma{}^*)(\boldsymbol{\mu}_\gamma-\boldsymbol{\mu}_\gamma{}^*)^T\right\}, \tag{2.75}$$

as a function of $\boldsymbol{\mu}_\gamma$, with

$$\boldsymbol{\mu}_\gamma{}^* = \sum_{i=1}^{m}\phi_i^{-1}\boldsymbol{\gamma}_i / \sum_{i=1}^{m}\phi_i^{-1}. \tag{2.76}$$

Analogous to (2.74) and (2.75) formulae hold, if $\boldsymbol{\gamma}_i$, $\boldsymbol{\mu}_\gamma$ and $\boldsymbol{Q}_\gamma$ are replaced by $\boldsymbol{\beta}_i$, $\boldsymbol{\mu}_\beta$

and $\boldsymbol{Q}_\beta$ respectively, and $\boldsymbol{\mu}_\gamma{}^*$ by

$$\boldsymbol{\mu}_\beta{}^* = \sum_{i=1}^{m} \phi_i^{-1} \boldsymbol{\beta}_i / \sum_{i=1}^{m} \phi_i^{-1}. \tag{2.77}$$

2.5.3 Full conditional distributions

The following statements are made conditionally on the data and the rest of the unknown random effects and parameters in the model:

B1: For $i = 1, \ldots, m$, the $\boldsymbol{\gamma}_i$ are independent and normally distributed with respective means $\boldsymbol{\gamma}_i{}^*$, and variances $(\boldsymbol{U}_i + \boldsymbol{Q}_\gamma{}^{-1})^{-1}\phi_i$.

B2: For $i = 1, \ldots, m$, the $\boldsymbol{\beta}_i$ are independent and normally distributed with respective means $\boldsymbol{\beta}_i{}^*$, and respective variances $(\boldsymbol{X}_i + \boldsymbol{Q}_\beta{}^{-1})^{-1}\phi_i$.

B3: For $i = 1, \ldots, m$, the ϕ_i are independent, and $\nu_i^* \mathrm{M}_i / \phi_i$ possesses a chi-squared distribution with ν_i^* degrees of freedom, with $\nu_i^* = \nu + n_i + d_1 + d_2$, and

$$\nu_i^* \mathrm{M}_i = \nu\zeta + S_i^2 + \boldsymbol{w}_{i,1}{}^T \boldsymbol{V}_{i,1} \boldsymbol{w}_{i,1} + \boldsymbol{w}_{i,2}{}^T \boldsymbol{V}_2 \boldsymbol{w}_{i,2}. \tag{2.78}$$

B4: The mean $\boldsymbol{\mu}_\gamma$ is normally distributed with mean $\boldsymbol{\mu}_\gamma{}^*$ and variance $\boldsymbol{Q}_\gamma / \sum_{i=1}^m \phi_i^{-1}$.

B5: The mean $\boldsymbol{\mu}_\beta$ is normally distributed with mean $\boldsymbol{\mu}_\beta{}^*$ and variance $\boldsymbol{Q}_\beta / \sum_{i=1}^m \phi_i^{-1}$.

B6: The inverse of the matrix $\boldsymbol{Q}_\gamma$ has a Wishart distribution with $k_\gamma + m$ degrees of freedom and scale matrix equal to

$$\left\{ \sum_{i=1}^{m} \phi_i^{-1} (\boldsymbol{\gamma}_i - \boldsymbol{\mu}_\gamma)(\boldsymbol{\gamma}_i - \boldsymbol{\mu}_\gamma)^T + \boldsymbol{\Sigma}_\gamma{}^{-1} \right\}^{-1}. \tag{2.79}$$

B7: The matrix $\boldsymbol{Q}_\beta{}^{-1}$ has a Wishart distribution with $k_\beta + m$ degrees of freedom and scale matrix equal to

$$\left\{ \sum_{i=1}^{m} \phi_i^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\mu}_\beta)(\boldsymbol{\beta}_i - \boldsymbol{\mu}_\beta)^T + \boldsymbol{\Sigma}_\beta{}^{-1} \right\}^{-1}. \tag{2.80}$$

B8: The full conditional distribution of ζ is identical to that of **A8**, p. 59.

B9: The full conditional distribution of ν is identical to that of **A9**, p. 59.

In order to generate matrices of dimension p from a Wishart distribution with k degrees of freedom and scale matrix $\boldsymbol{\Sigma}$, it is sufficient to add the crossproducts of k independent multivariate normal vectors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Each multivariate normal vector can be generated by multiplying the lower triangular matrix

of the Cholesky decomposition of $\boldsymbol{\Sigma}$ by a vector of p independent random numbers drawn from the standard normal distribution. A previous implementation of the Gibbs sampler with Wishart matrices by the author of this thesis can be found in Izenman et al (1998). The Cholesky decomposition of a non-negative definite symmetric matrix can be obtained by standard numerical routines, (Press et al, 1994, p. 97).

For testing hypotheses of the form $H_0 : \boldsymbol{\alpha}^T\boldsymbol{\beta_i} = 0$, $(i = 1, \ldots, m)$, we will need to consider the posterior densities of the quantities $\boldsymbol{\alpha}^T\boldsymbol{Q_\beta}\boldsymbol{\alpha}$. These can be computed using the fact that if a matrix $\boldsymbol{U}$ has the distribution (2.72), then the distribution of $\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}/\boldsymbol{\alpha}^T\boldsymbol{U}\boldsymbol{\alpha}$ is chi-squared with $k-p+1$ degrees of freedom and hence the required posterior density can be computed by averaging the successive full conditional posterior densities, that is, using the "Rao-Blackwellization" method described in section 2.3.1.2.

The previous algebraic calculations can be greatly simplified, if we assume orthogonality of the $\boldsymbol{u_{ij}}$ and $\boldsymbol{x_{ij}}$ vectors., i.e., if the following condition holds:

$$\sum_{j=1}^{n_i} \boldsymbol{x_{ij}}\boldsymbol{u_{ij}}^T = 0 \qquad (i = 1, \ldots, m). \tag{2.81}$$

In this case, the matrices $\boldsymbol{U} \circ \boldsymbol{X_i}$ and $\boldsymbol{X} \circ \boldsymbol{U_i}$ in (2.49) and (2.50) are equal to $\mathbf{0}$, the least squares estimates of $\boldsymbol{\gamma_i}$ and $\boldsymbol{\beta_i}$ in (2.51) and (2.52) reduce to their usual standard form and are uncorrelated with each other, and the $\boldsymbol{\gamma_i}^*$ and $\boldsymbol{\beta_i}^*$ are the same as in (2.63), with the matrix $\boldsymbol{V_{i,1}} + \boldsymbol{V_2}$ being inverted in a straightforward fashion, since it is of block diagonal form. Also, $\boldsymbol{V_{i,3}} = \text{diag}(\boldsymbol{U_i}(\boldsymbol{U_i}+\boldsymbol{Q_\gamma}^{-1})^{-1}\boldsymbol{Q_\gamma}^{-1}, \boldsymbol{X_i}(\boldsymbol{X_i}+\boldsymbol{Q_\beta}^{-1})^{-1}\boldsymbol{Q_\beta}^{-1})$. The preceding full conditional distribution statements **B1-B9** hold without any modifications.

2.5.4 An ANCOVA model with several covariates

Perhaps the most important special case of the general ANCOVA model, in practical terms, is described by the following hierarchy:

(a) Conditionally on the random effects θ_i, $\boldsymbol{\beta_i}$ and ϕ_i, the observations y_{ij} are independent, and normally distributed with means

$$\theta_i + \boldsymbol{x_{ij}}^T\boldsymbol{\beta_i} \qquad (i = 1, \ldots, m), \tag{2.82}$$

and variances ϕ_i.

(b) Conditionally on the ϕ_i and the model parameters μ_θ, $\boldsymbol{\mu_\beta}$, λ_θ and $\boldsymbol{Q_\beta}$, the θ_i and $\boldsymbol{\beta_i}$

are mutually independent and normally distributed. The θ_i have mean μ_θ and variance $\lambda_\theta \phi_i$, and the $\boldsymbol{\beta_i}$ mean $\boldsymbol{\mu_\beta}$ and variance $\boldsymbol{Q_\beta}\phi_i$, for $i = 1, \ldots, m$.

(c) The ϕ_i, conditionally on ζ and ν, are independent and have scaled inverse chi-squared distributions, with ν degrees of freedom and scale parameter ζ, exactly as in (2.3).

This formulation is a combination of the two models already described in this chapter, and implies that it is desirable to test for equality of the group means (θ_i), in the presence of a number of covariates ($\boldsymbol{x_{ij}}$), with $\boldsymbol{x_{ij}}$ being a $d_2 \times 1$ centered vector. If this vector corresponds to a set of main effects then the following (orthogonality) condition holds:

$$\sum_{j=1}^{n_i} \boldsymbol{x_{ij}} = \mathbf{0} \qquad (i = 1, \ldots, m). \tag{2.83}$$

If, on the other hand, it contains second or higher order terms (including quadratic, multiplicative interaction ones etc), (2.83) may not be true.

In this case the joint distribution of the observations, given the model parameters reduces to

$$\begin{aligned} p(\boldsymbol{y}|\mu_\theta, \boldsymbol{\mu_\beta}, \lambda_\theta, \boldsymbol{Q_\beta}, \nu, \zeta) = \\ = (2\pi)^{-\frac{N}{2}} K(\nu)^m \zeta^{\frac{\nu m}{2}} \prod_{i=1}^m \left\{ \Gamma\left(\tfrac{\nu+n_i}{2}\right) 2^{\frac{\nu+n_i}{2}} \right\} \left(|\lambda_\theta \boldsymbol{Q_\beta}|\right)^{-m/2} \prod_{i=1}^m |\boldsymbol{V_{i,1}} + \boldsymbol{V_2}|^{-1/2} \\ \times \prod_{i=1}^m \left\{ \nu\zeta + S_i^2 + (\boldsymbol{w_{i,2}} - \boldsymbol{w_{i,1}})^T \boldsymbol{V_{i,3}} (\boldsymbol{w_{i,2}} - \boldsymbol{w_{i,1}}) \right\}^{-\frac{1}{2}(\nu+n_i)}, \end{aligned} \tag{2.84}$$

with the least squares estimates of θ_i and $\boldsymbol{\beta_i}$, respectively

$$\hat{\theta}_i = \left(1 - n_i \boldsymbol{x_{i.}}^T \boldsymbol{X_i}^{-1} \boldsymbol{x_{i.}}\right)^{-1} \left(y_{i.} - \boldsymbol{x_{i.}}^T \boldsymbol{X_i}^{-1} \sum_{j=1}^{n_i} \boldsymbol{x_{ij}} y_{ij} \right), \tag{2.85}$$

and

$$\hat{\boldsymbol{\beta_i}} = \left(\boldsymbol{X_i} - n_i \boldsymbol{x_{i.}} \boldsymbol{x_{i.}}^T\right)^{-1} \left(\sum_{j=1}^{n_i} \boldsymbol{x_{ij}} y_{ij} - n_i \boldsymbol{x_{i.}} y_{i.} \right), \tag{2.86}$$

with

$$S_i^2 = \sum_{j=1}^{n_i} \left(y_{ij} - \hat{\theta}_i - \boldsymbol{x_{ij}}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \right)^2, \tag{2.87}$$

with $\boldsymbol{V_{i,1}}$, $\boldsymbol{V_2}$, $\boldsymbol{V_{i,3}}$, $\boldsymbol{w_{i,1}}$ and $\boldsymbol{w_{i,2}}$ defined as in (2.55), (2.58), (2.62), (2.57) and (2.59) respectively, with $\boldsymbol{U_i} = n_i$, $\boldsymbol{X} \circ \boldsymbol{U_i} = n_i \boldsymbol{x_{i.}}$, and $\boldsymbol{U} \circ \boldsymbol{X_i} = n_i \boldsymbol{x_{i.}}^T$.

Notice that in the main effects centered covariate situation (2.83) holds. This condition is a simplification of the orthogonality condition (2.81), and its effect reduces

$\hat{\theta}_i$ to $y_{i\cdot}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ to its standard form. Obtaining the maximum likelihood estimates and constructing a Bayesian information type criterion for model comparison, requires estimating $d_2(d_2+3)/2+4$ parameters.

Assuming the same prior distributions for μ_θ, λ_θ, ν and ζ as in section 2.2 and for $\boldsymbol{\mu_\beta}$ and $\boldsymbol{Q_\beta}$ as in section 2.5.2, we obtain the posterior distribution of the three sets of random effects and model parameters to be

$$\begin{aligned}
&\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi},\mu_\theta,\boldsymbol{\mu_\beta},\lambda_\theta,\boldsymbol{Q_\beta},\nu,\zeta|\boldsymbol{y}) \propto \\
&\propto \pi(\nu)K(\nu)^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}\zeta^{\frac{1}{2}\nu(m+\psi\zeta_0)-1}\exp(-\tfrac{1}{2}\nu\psi\zeta)\prod_{i=1}^m \phi_i^{-\frac{1}{2}(\nu+n_i+d_2+3)} \\
&\times\lambda_\theta^{-\frac{1}{2}(\omega_1+m+2)}\,|\boldsymbol{Q_\beta}|^{-\frac{1}{2}(k_\beta+d_2+m+1)}\exp\left[-\tfrac{1}{2}\left\{\lambda_\theta^{-1}\omega_1\tau_1+\text{trace}\left(\boldsymbol{\Sigma_\beta}^{-1}\boldsymbol{Q_\beta}^{-1}\right)\right\}\right] \\
&\times\exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^m\phi_i^{-1}\left(\nu\zeta+S_i^2+\boldsymbol{w_{i,1}}^T\boldsymbol{V_{i,1}}\boldsymbol{w_{i,1}}+\boldsymbol{w_{i,2}}^T\boldsymbol{V_2}\boldsymbol{w_{i,2}}\right)\right\}.
\end{aligned} \tag{2.88}$$

Hence the full conditional distributions for the applications of the Gibbs sampler, conditionally on the data and the remaining random effects and parameters in the model, become:

C1: For $i=1,\ldots,m$, the θ_i are independent and normally distributed with respective means and variances identical to **A1**, p. 58.

C2: For $i=1,\ldots,m$, the $\boldsymbol{\beta_i}$ are independent and normally distributed with respective means and variances identical to **B2**, p. 66.

C3: For $i=1,\ldots,m$, the ϕ_i are independent, and $\nu_i^*\mathrm{M}_i/\phi_i$ possesses a chi-squared distribution with ν_i^* degrees of freedom, with $\nu_i^*=\nu+n_i+d_2+1$ and M_i defined in (2.78) with the adjustments described in page 68.

C4: The mean μ_θ is normally distributed with mean μ_θ^*, defined in (2.35), and variance $\lambda_\theta/\sum_{i=1}^m\phi_i^{-1}$.

C5: The mean $\boldsymbol{\mu_\beta}$ is normally distributed with mean $\boldsymbol{\mu_\beta}^*$, defined in (2.77), and variance $\boldsymbol{Q_\beta}/\sum_{i=1}^m\phi_i^{-1}$.

C6: The quantity $(m+\omega_1)\lambda_\theta^*/\lambda_\theta$ has a chi-squared distribution with $m+\omega_1$ degrees of freedom, with λ_θ^* as in (2.42).

C7: The matrix $\boldsymbol{Q_\beta}^{-1}$ has a Wishart distribution with $k_\beta+m$ degrees of freedom and scale matrix equal to that in (2.80).

C8: The full conditional distribution of ζ is identical to that of **A8**, p. 59.

C9: The full conditional distribution of ν is identical to that of **A9**, p. 59.

C10: For $i=1,\ldots,m$, the adjusted means $\xi_i=\theta_i-\boldsymbol{\beta_i}^T(\boldsymbol{x_{i\cdot}}-\boldsymbol{x_{\cdot\cdot}})$ are independent

and normally distributed with respective means ξ_i^* and variances $V_{\xi_i^*}\phi_i$, where

$$\xi_i^* = \theta_i^* - (\boldsymbol{\beta_i}^*)^T(\boldsymbol{x_{i.}} - \boldsymbol{x_{..}}), \tag{2.89}$$

and

$$V_{\xi_i^*} = (n_i + \lambda_\theta^{-1})^{-1} + (\boldsymbol{X_i} + \boldsymbol{Q_\beta}^{-1})^{-1}(\boldsymbol{x_{i.}} - \boldsymbol{x_{..}})^2. \tag{2.90}$$

If the vector $\boldsymbol{x_{ij}}$ contains only main effect terms, then the multivariate normal density of the $\boldsymbol{\beta_i}$ can be replaced by a set of independent normal distributions with mean $\mu_{\beta,k}$ and variances $\lambda_{\beta,k}\phi_i$, for $k = 1, \ldots, d_2$, with the prior distribution for the $\lambda_{\beta,k}$, scaled inverse chi-squared with $\omega_{2,k}$ degrees of freedom and scale parameters $\tau_{2,k}$. In this case, because $\sum_{j=1}^{n_i} \boldsymbol{x_{ij}} = \boldsymbol{0}$, the aforementioned steps, **C2**, **C5**, and **C7** involving generating a d_2 dimensional vector the first two, and a $d_2 \times d_2$ symmetric matrix the third, can be replaced by **C2'**, **C5'**, and **C7'**, involving generating d_2 scalar quantities each as, follows:

C2': For $i = 1, \ldots, m$, and for $k = 1, \ldots, d_2$ the $\beta_{i,k}$ are independent and normally distributed with respective means $\beta_{i,k}^*$, and respective variances $(s_{i,k}^2 + \lambda_{\beta,k}^{-1})^{-1}\phi_i$.

C5': For $k = 1, \ldots, d_2$, the means $\mu_{\beta,k}$ are independent normally distributed with respective means $\mu_{\beta,k}^*$ and respective variances $\lambda_{\beta,k}/\sum_{i=1}^m \phi_i^{-1}$.

C7': For $k = 1, \ldots, d_2$, the quantities $(m + \omega_{2,k})\lambda_{\beta,k}^*/\lambda_{\beta,k}$ are independent and have a chi-squared distribution with $m + \omega_{2,k}$ degrees of freedom.

The remaining **C** steps hold as already presented, with the usual simplification because of the orthogonality condition. The new algebraic symbols appearing in the **C2'**, **C5'**, and **C7'** have the same definitions as in **A1-A10**, (pp. 58-59), with the k index denoting which element of the $\boldsymbol{\beta_i}$ vector they refer to.

The models already presented should cover any random means, slopes and variances ANCOVA, whatever its complexity, and allow to test for the equality of any particular set of random effects or other hypotheses of interest. Practical considerations regarding their application and resulting inference will be explored in detail in the following chapter.

Chapter 3

Case studies

In this chapter we will present the application of the random effects ANCOVA model for the analysis of a number of data sets. We will begin with the comprehensive analysis of the visual functions neuropsychological test, which was completed by three groups of Scottish prison inmates and two control groups in Stanford, USA, and proceed with the illustration of the main results and conclusions drawn from the entire set of tests, twelve in total, taken by the same five groups. The chapter will conclude with a study of food additives on the weight gain in animals, and a study of the performance of the proposed, in section 2.4, test statistic for testing equality of the group variances by using simulated data.

3.1 Analysis of neuropsychological tests

The current analysis is motivated by an ongoing investigation by Scottish forensic scientists, in the area of offender profiling. One possible objective is to be able to use the results of neuropsychological tests to predict type of offender.

A comprehensive review of offender profiling techniques is presented by Jackson and Bekerian (1997). Daeid et al (1998) have previously investigated inmates of Irish prisons for the purposes of offender profiling. Neuropsychological tests have been studied by Moses et al (1992), with the aim of providing normal scores and decision rules for mentally healthy and unhealthy people.

Our main objective will not be validating these previous results, but rather providing comparisons of the scores of the five groups using the appropriate model for the responses and covariates available. Our preliminary investigations, using standard constant variance models, suggested that most of the neuropsychological tests did not

yield significant differences, between the three offender groups. Furthermore, the one or two tests yielding apparently significant results were shown to be less significant when confounding variables such as age and intelligence were considered (paedophile prisoners tend to be more intelligent and older). As a result of our subsequent analyses we conclude that neuropsychological tests appear more likely to be able to predict the presence of a medical condition than type of offender, based on age and neuropsychological test score information, the sole pieces of information available for all five groups. This is however a classic situation involving many important confounding variables (see Box, Hunter and Hunter, 1978, pp. 8, 493-495).

3.1.1 Visual functions test

We now report the analysis of the scores y_{ij} on a neuropsychological test (for visual functions) which was completed by (a) $n_1 = 67$ Stanford students, (b) $n_2 = 22$ Scottish rapists, (c) $n_3 = 40$ Scottish paedophiles, (d) $n_4 = 20$ Scottish murderers, and (e) $n_5 = 128$ Stanford medical patients. The explanatory variable x_{ij} represents age of participant. The offenders in the middle three groups were all interviewed in Scottish prisons. The values of the sufficient statistics $\hat{\theta}_i = y_{i\cdot}$, $\hat{\beta}_i$, and $\hat{\phi}_i = U_i/(n_i - 2)$, provide unbiased estimators for θ_i, β_i, and ϕ_i, and their estimated standard errors are reported in Table 3.2. Furthermore, numerical values of the quantities in (2.8) are $s_1^2 = 9236.99$, $s_2^2 = 795.27$, $s_3^2 = 4390.40$, $s_4^2 = 944.20$ and $s_5^2 = 20850.62$. An initial ANOVA of the dependent variable, assuming equal variances indicated a statistically significant difference between the group means ($F = 12.16$ on 4 and 272 d.f.) This statistically significant difference was also observed using ANCOVA to incorporate the age variable ($F = 10.97$ on 4 and 271 d.f.). Two of the slopes (β_1 and β_2) were significantly greater than zero. A test for equality of the slopes gave $F = 2.66$ on 4 and 267 d.f.. For these data, the information criterion in (2.17) gave $BIC^* = -768.47$, very favourably comparing with $BIC^* = -777.78$ for the equal variance (unequal θ_i and β_i) model.

Our analysis proceeds under the choices of prior parameters $\omega_1 = 4$, $\tau_1 = 1$, $\omega_2 = 4$, $\tau_2 = 0.005$, $\zeta_0 = 10$, $\psi = 0.02$, $a = 3$, and $b = 0.04$. These were chosen to ensure that the posterior distribution roughly matched the posterior distribution under uniform prior distributions for λ_θ, λ_β, ζ, and ν. We however confine attention in the current example to proper distributions for these parameters, in the prior assessment, together with a sensitivity analysis (see Dickey, 1973), described below.

To assess the adequacy of the length of the burn-in, we use a suggestion by Ritter

(1992, p. 103) and plot the trajectories of the 0th, 25th, 50th, 75th and 100th percentile of all random effects and model parameters from a number of parallel realizations of the chain, thus using some aspects of the Gelman and Rubin convergence assessment criterion mentioned in section 2.3.1.3. Similarly to Ritter, we used one hundred parallel chains, however unlike his non-linear regression study, where the settling down of the trajectories happened after about 2,000 simulations, in our situation it takes place after less than 10.

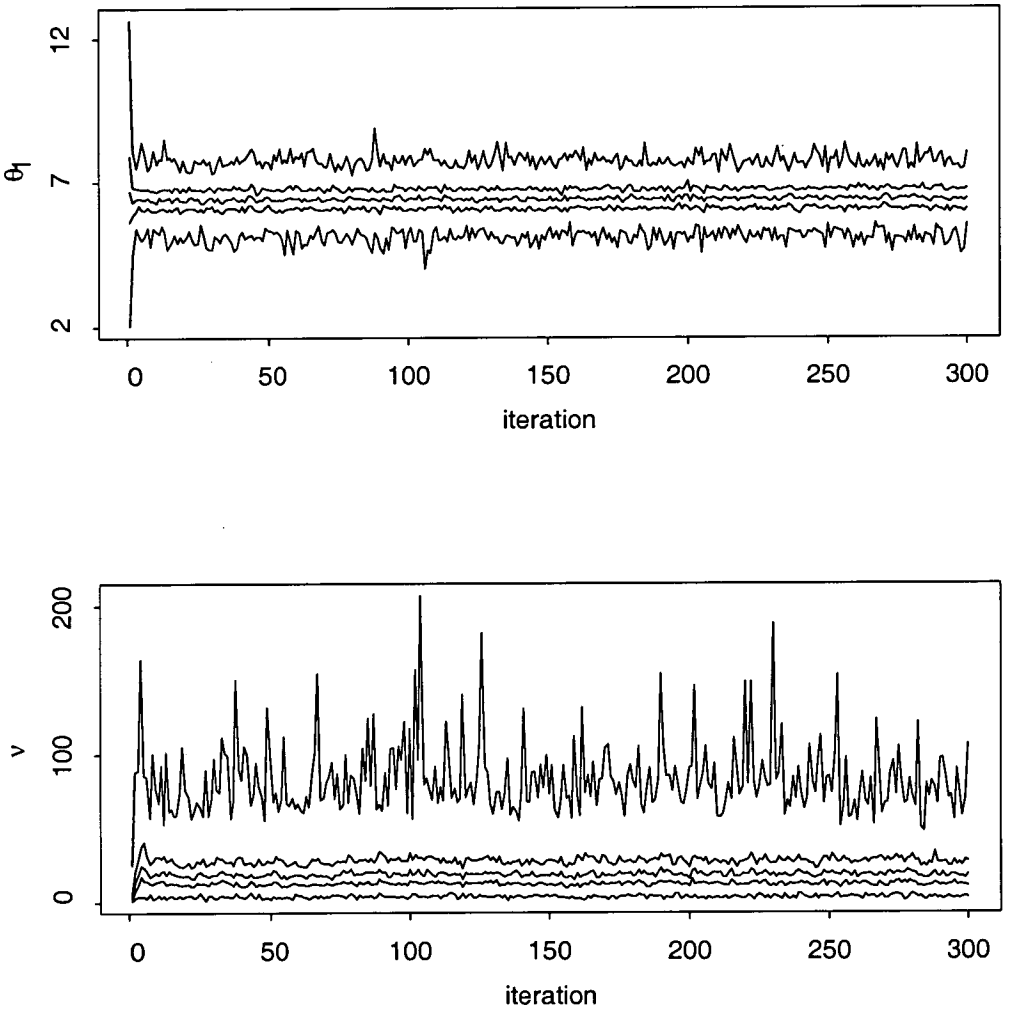


Figure 3.1: Visual functions test. Quantile (0th, 25th, 50th, 75th, 100th percentile) traces for θ_1 (Top) and ν (Bottom). The traces stabilize after about ten iterations.

The two plots of Figure 3.1 represent the trajectories of θ_1 and ν, for the first 300 iterations. They were obtained using overdispersed and uniformly distributed starting values for all random effects and model parameters. We omitted the remaining 19 corresponding plots to avoid being repetitious, however the conclusion reached would remain the same: Even when using starting values which are away from reasonable

estimates, only a very short burn-in period is necessary. If the starting values are realistic initial estimates of the corresponding parameters, then it may be possible that no burn-in is required at all.

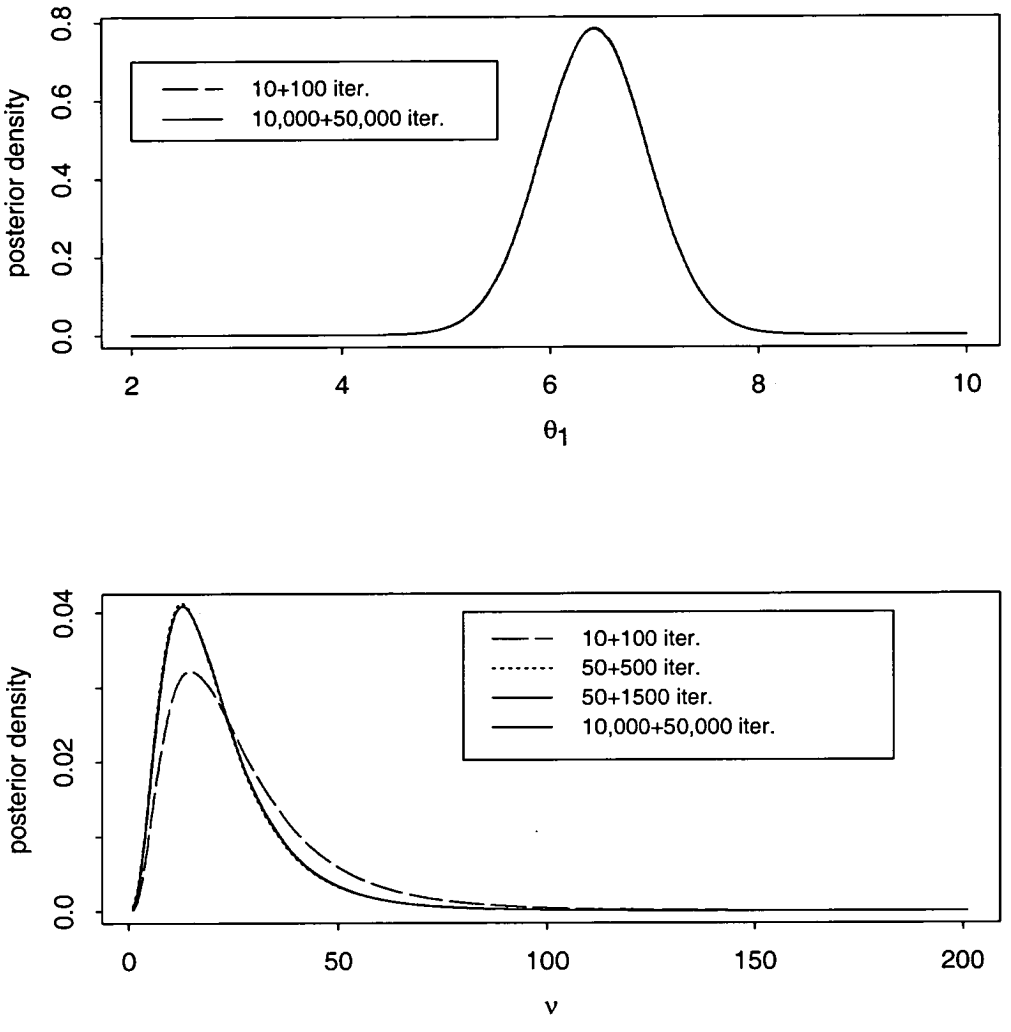


Figure 3.2: Visual functions test. Posterior density of θ_1 (Top) and ν (Bottom) for different numbers of iterations. The two θ_1 and longest iteration two ν curves are indistinguishable.

Motivated by the previous conclusion, and assuming that the marginal posterior distributions produced using 10,000 burn-in and 50,000 density averages are the "true" ones (an assumption validated in Chapter 4 using Laplacian approximations), we next attempted to find the minimum number of iterations needed to obtain posterior inferences practically identical to the "true" ones. For all quantities whose full conditional distribution is normal, using unrealistic starting values, 10 burn-in iterations followed by 500 density averages provided marginal posterior distributions visually indistinguishable from the "true" ones, although in some situations 10+100 iterations proved to be

adequate. The parameter requiring the maximum number of iterations was ν with 50+1500. The two plots of Figure 3.2 illustrate these results for various choices of chain lengths.

The posterior densities of $\theta_1, \ldots, \theta_5$ are described in Figure 3.3. These were calculated by each of the three simulation procedures (approximate MCMC, acceptance sampling for ν, a discrete point prior for ν matching the continuous prior) described in section 2.3.2, giving results which are identical up to visual accuracy. The convergence of the simulations will be validated by Laplacian procedures in Chapter 4.

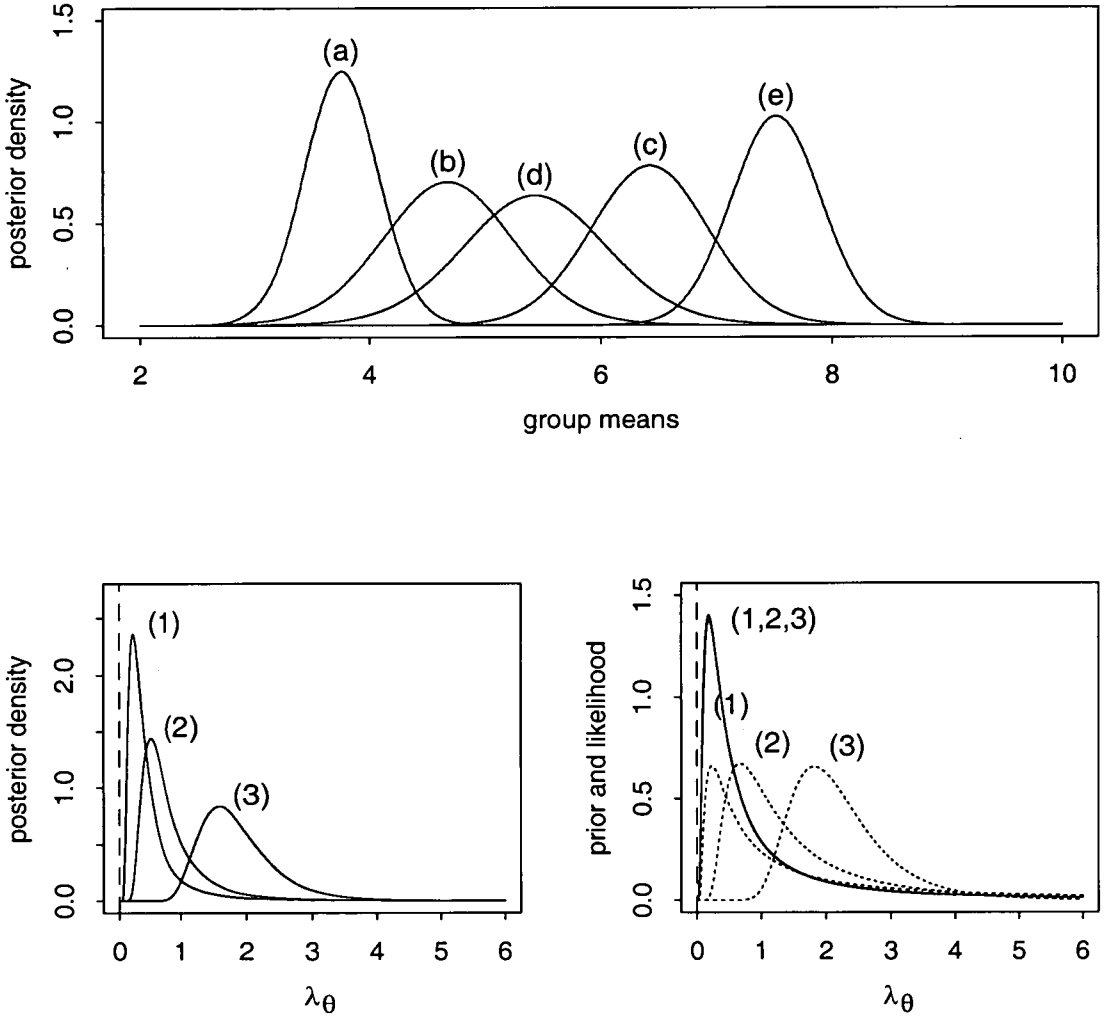


Figure 3.3: Visual functions test. Top: Posterior density of five, (a)-(e), group means (θ_i). Bottom left: Posterior density of λ_θ under three choices, (1)-(3), of prior distribution. Bottom right: Integrated likelihood (solid line) and prior density (dotted line) of λ_θ, under three choices, (1)-(3), of prior distribution. The dashed vertical lines correspond to $\lambda_\theta = \widetilde{\lambda_\theta}$.

The curves in the top plot of Figure 3.3 suggest the ordering (a), (b), (d), (c), (e) of the five groups, according to the magnitude of the test results, and take into

account the unequal variances, and the uncertainty about the six model parameters. A visual inspection suggests that the differences between the three groups of Scottish offenders, (b), (c), and (d), are not of practical significance. However the Stanford medical patients (e) give substantially higher scores than both the Stanford students (a) and the Scottish rapists (b). The Stanford students (a) also have substantially lower scores than the paedophiles (c). We can only make these comparisons if the overall investigation of λ_θ, as described in the bottom left plot of Figure 3.3, suggests that λ_θ is very different from zero. The posterior densities of the adjusted group means in (2.2) and the contrasts $\theta_i - \theta_.$ can be similarly calculated, and give similar results to our interpretations for the θ_i.

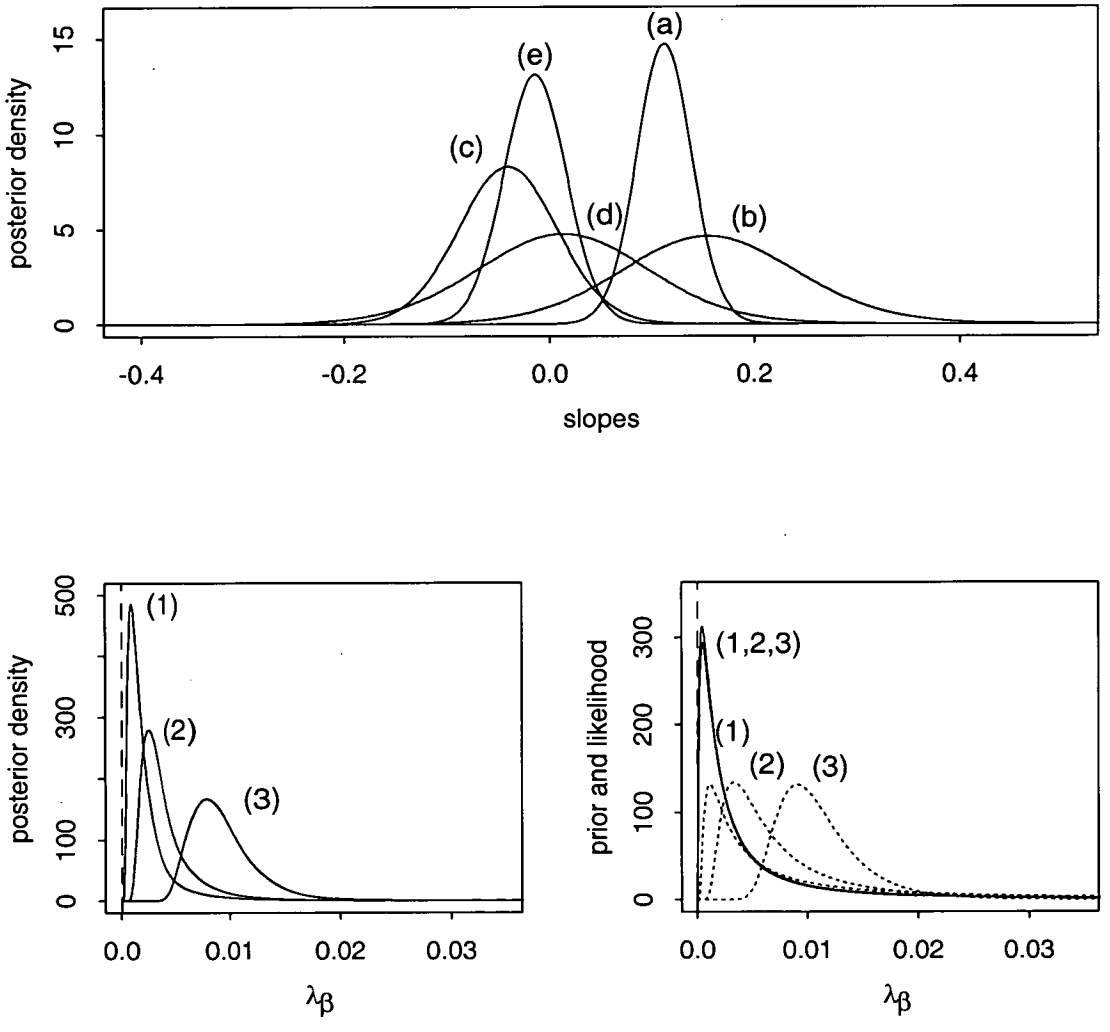


Figure 3.4: Visual functions test. Top: Posterior density of five, (a)-(e), group slopes (β_i). Bottom left: Posterior density of λ_β under three choices, (1)-(3), of prior distribution. Bottom right: Integrated likelihood (solid line) and prior density (dotted line) of λ_β, under three choices, (1)-(3), of prior distribution. The dashed vertical lines correspond to $\lambda_\beta = \widetilde{\lambda_\beta}$.

The posterior densities of the conditional slopes β_i are described by the curves in the top plot of Figure 3.4. It is not obvious whether these densities refute a parallel line model with $\beta_1 = \ldots = \beta_5$. This hypothesis will be further investigated by considering the posterior density of λ_β. Posterior densities of the contrasts $\beta_i - \beta_.$ can also be computed.

The posterior densities of the conditional variances ϕ_i are described in the top plot of Figure 3.5. The curves indicate that the variance for the Stanford medical patients (e) is substantially higher than the Stanford students (a) and the rapists (b). An overall evaluation of equality of the variances may be obtained from the posterior density of ν.

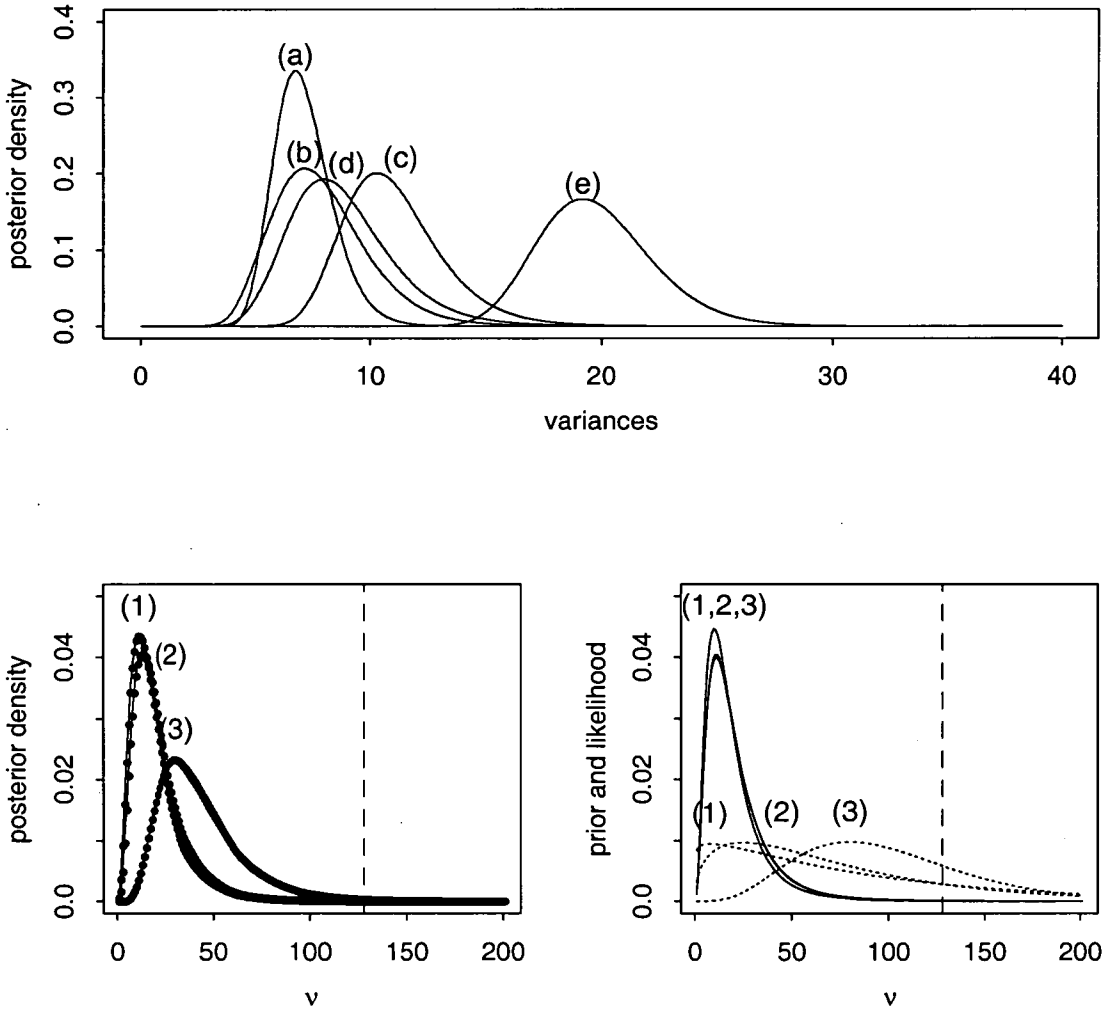


Figure 3.5: Visual functions test. Top: Posterior density of five, (a)-(e), group variances (ϕ_i). Bottom left: Posterior density/probability mass function of ν (solid line under continuous prior, $\cdots$ under discrete prior) under three choices, (1)-(3), of prior distribution. Bottom right: Integrated likelihood (solid line) and prior density (dotted line) of ν, under three choices, (1)-(3), of prior distribution. The dashed vertical lines correspond to $\nu = \nu_\phi$.

It is straightforward to demonstrate that the posterior densities for the θ_i, β_i, and ϕ_i contain noticeable differences when compared with the t-densities and chi-squared densities based upon a fixed effects analysis with uniform distributions for the θ_i, β_i, and $\log \phi_i$, in the prior assessment. For example, the inferences for β_2 and β_4 are quite substantially different.

The posterior density of λ_θ, curve (2) in the bottom left plot of Figure 3.3, gives the initial, incorrect, impression, that λ_θ is close to zero. No Bayesian significance probability is available for the hypothesis $H_\theta : \lambda_\theta = 0$, as this value also provides an extreme value of the parameter space. The interpretation of this density therefore possesses an interesting problem. We answer this problem, by considering the point $\widetilde{\lambda_\theta} = \min_{i=1,\ldots,5} n_i^{-1}$, and the posterior probability, P_θ, that $\lambda_\theta \leq \widetilde{\lambda_\theta}$, (see vertical line in the same plot). If P_θ is small then the data refute H_θ. In our example P_θ is less than 10^{-15}, strongly refuting H_θ. For an interpretation of $\widetilde{\lambda_\theta}$, see section 2.4.

In connection with the vertical line in the bottom left plot of Figure 3.4, we see that P_β, the posterior probability that $\lambda_\beta \leq \widetilde{\lambda_\beta} = \min_{i=1,\ldots,5} s_i^{-2}$, for curve (2) corresponding to the prior distribution already described, is again less than 10^{-15}, thus refuting the hypothesis H_β that the regression slopes are all equal. The vertical line in the bottom left plot of Figure 3.5 should be interpreted by noting that P_ϕ, the posterior probability that $\nu > \nu_\phi = \max_{i=1,\ldots,5} n_i$, is equal to 0.000598, thus refuting the hypothesis H_ϕ of equality of the variances. For interpretations of $\widetilde{\lambda_\beta}$ and ν_ϕ, again see section 2.4. The posterior densities of ν under the continuous and discrete priors described in section 2.3.3 very closely match each other for all sets of priors.

3.1.1.1 Sensitivity analysis

The dotted curve (2) in the bottom right plot of Figure 3.3 denotes the current prior density for λ_θ. Two further choices of parameters were made. The second choice was $\omega_1 = 1$, $\tau_1 = 0.7$, $\omega_2 = 1$, $\tau_2 = 0.0035$, $\zeta_0 = 10.5$, $\psi = 0.029$, $a = 2$, and $b = 0.025$ and third choice was $\omega_1 = 20$, $\tau_1 = 2$, $\omega_2 = 20$, $\tau_2 = 0.01$, $\zeta_0 = 15$, $\psi = 0.013$, $a = 10$, and $b = 0.1$. Dotted curves (1) and (3) of the same plot denote the corresponding prior densities for λ_θ. Curves (1), (2), and (3) of the bottom left plot of Figure 3.3 describe the posterior density of λ_θ under our three choices of prior distribution. They are quite sensitive to the choice of prior distribution. Moreover, the solid curves in the bottom right plot of Figure 3.3, describe three integrated likelihoods for λ_θ, each obtained by dividing the corresponding posterior densities in the same figure, by the

prior densities for λ_θ, and then renormalizing to ensure that the integrated likelihood integrates to unity. The integrated likelihoods are remarkably insensitive to the choices of prior distribution, and should also be considered as part of our inferential procedure. Similar results are presented in Figure 3.4 for the parameter λ_β and in Figure 3.5, for the important parameter ν. The dots in the latter figure describe the posterior probabilities for ν under the discretization for the prior probabilities indicated in section 2.3.2.

While the marginal posterior density of λ_θ, λ_β and ν are influenced by the choice of prior, the posterior distributions of the θ_i, β_i, and ϕ_i, (reported in Figure 3.6 for the three sets of priors, are quite insensitive to these choices, with the posterior densities of the variances slightly more sensitive. This stability is expected since the random effects are further apart from the specified prior parameters in the model hierarchy than the three parameters.

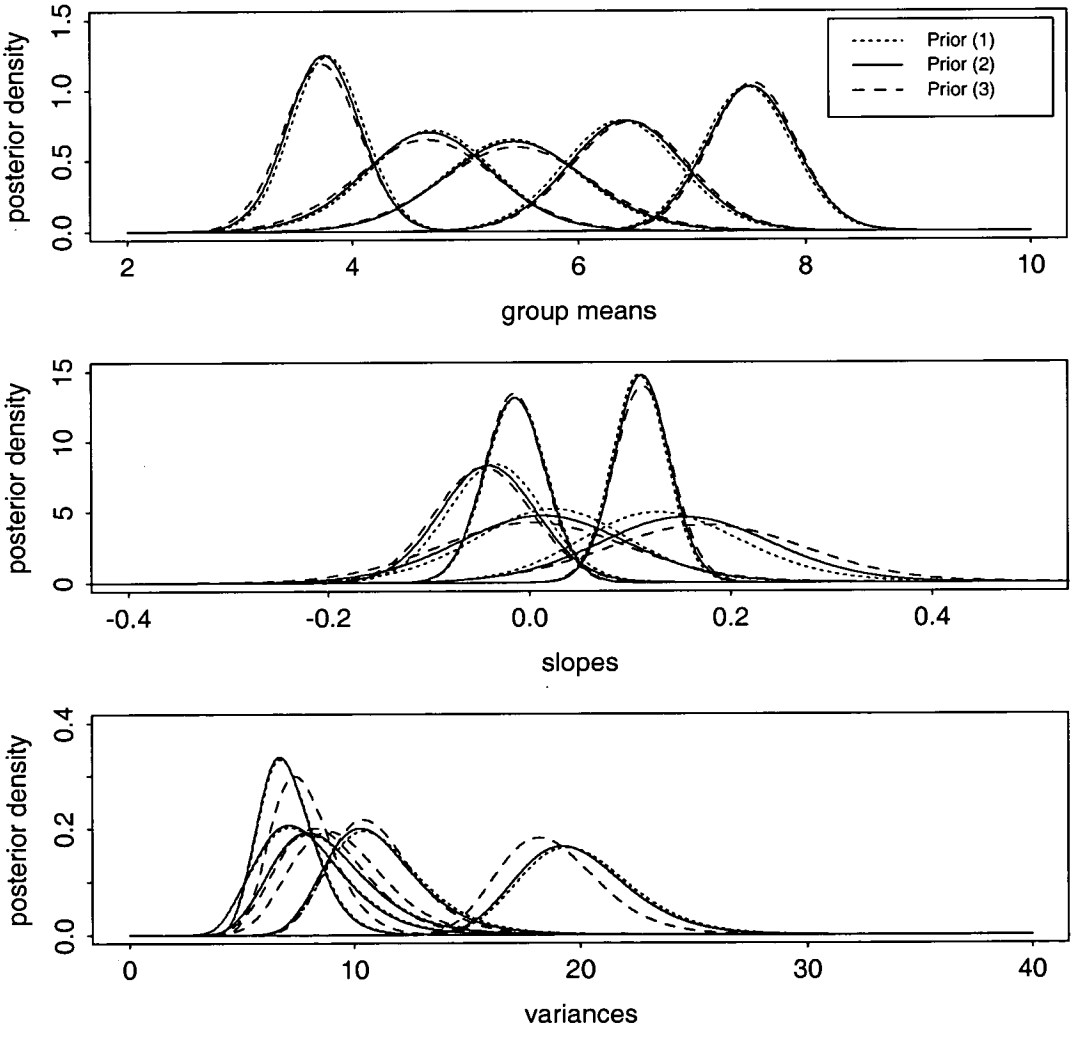


Figure 3.6: Visual functions test. Top: Posterior density of group means (Top), slopes (Middle), and variances (Bottom), under three choices, (1)-(3), of prior distribution.

3.1.2 Analysis of twelve neuropsychological tests

In Tables 3.1-3.4 we describe the results of the twelve neuropsychological tests, including these of the visual functions test described in section 3.1.1. In all cases, the results of the tests are compared for all five groups (a)-(e), with age as covariate, and hence common values for s_i^2, $(i = 1, \ldots, 5)$. Additionally, for every test, higher score signifies more pathological condition. Our analysis used the three prior distributions as already described. We computed the posterior probabilities P_θ, P_β, and P_ϕ, for λ_θ, λ_β, and ν, to assess whether the means, slopes and variances, respectively, can be considered to be equal or not.

The numerical values presented in the four tables include the posterior means and standard deviations of all three sets of random effects (statistics A and B, respectively), under the prior distribution (2) of the previous section, and the corresponding unbiased estimators and estimated standard errors (statistics C and D). The posterior means and standard deviations were computed by numerical integration of the marginal posterior densities obtained by MCMC and the discrete version of the full conditional distribution of ν. The unbiased estimates and estimated standard errors of the group means and slopes as well as the unbiased estimates of the group variances are included in the output of standard statistical packages, like Splus. Noticing that the quantities U_i/ϕ_i have chi-squared distributions with $n_i - 2$ degrees of freedom, the estimated standard error for the ith variance can be derived to be equal to

$$\sqrt{\frac{2U_i^2}{(n_i-2)^3}}, \tag{3.1}$$

with U_i defined in (2.6) and n_i the ith group sample size.

The initial conjecture of the forensic scientists who undertook this study, was that the Scottish offenders of groups (b) and (c) would have significantly higher scores than the general population, hence included group (d) in their study as a control group. Given that murderers can hardly be considered to be representative of a general population, it was suggested to include another control group. In this manner, the two groups from a study that used the same neuropsychological tests in Stanford were used. The different origin of the various groups in the study and different interviewer would automatically raise questions as to whether any differences that might be observed are indeed real, although the direct manner in which the test scores are obtained would tend to diminish these reservations.

Table 3.1: *Statistical analysis of results of twelve neuropsychological tests. A: Posterior mean, B: Posterior standard deviation, C: Unbiased estimate, D: Estimated standard error.*

Parameter	Statistic	Group (a)	(b)	(c)	(d)	(e)
		Test: Motor functions				
Group Mean	A	10.186	17.296	17.335	13.954	16.899
	B	(0.885)	(1.899)	(1.190)	(1.487)	(1.033)
	C	10.075	17.500	17.450	13.900	16.930
	D	(0.887)	(1.962)	(1.200)	(1.521)	(1.040)
Slope	A	0.225	0.153	-0.118	0.183	-0.091
	B	(0.075)	(0.276)	(0.112)	(0.194)	(0.081)
	C	0.230	0.177	-0.135	0.213	-0.094
	D	(0.076)	(0.326)	(0.115)	(0.221)	(0.082)
Variance	A	53.751	85.094	58.932	48.062	138.281
	B	(9.363)	(26.392)	(13.368)	(15.699)	(17.355)
	C	52.739	84.725	57.638	46.269	138.569
	D	(9.251)	(26.792)	(13.223)	(15.423)	(17.458)
		Test: Rhythm				
Group Mean	A	2.160	5.835	5.400	4.934	5.823
	B	(0.261)	(0.696)	(0.688)	(0.817)	(0.405)
	C	2.119	5.955	5.450	5.000	5.844
	D	(0.257)	(0.698)	(0.701)	(0.841)	(0.407)
Slope	A	0.083	0.171	-0.010	-0.014	-0.064
	B	0.022)	(0.101)	(0.064)	(0.107)	(0.032)
	C	0.083	0.215	-0.015	-0.038	-0.066
	D	0.022)	(0.116)	(0.067)	(0.122)	(0.032)
Variance	A	4.646	11.365	19.637	14.331	21.271
	B	(0.810)	(3.528)	(4.453)	(4.679)	(2.670)
	C	4.413	10.713	19.656	14.146	21.247
	D	(0.774)	(3.388)	(4.509)	(4.715)	(2.677)
		Test: Tactile functions				
Group Mean	A	4.034	6.244	6.258	6.497	6.746
	B	(0.338)	(0.960)	(0.620)	(1.500)	(0.566)
	C	4.000	6.318	6.300	6.600	6.766
	D	(0.336)	(0.992)	(0.628)	(1.578)	(0.571)
Slope	A	0.079	0.114	0.087	-0.080	0.007
	B	(0.029)	(0.139)	(0.058)	(0.195)	(0.044)
	C	0.080	0.134	0.088	-0.133	0.006
	D	(0.029)	(0.165)	(0.060)	(0.230)	(0.045)
Variance	A	7.855	21.840	16.046	49.051	41.603
	B	(1.368)	(6.774)	(3.639)	(16.007)	(5.222)
	C	7.559	21.629	15.800	49.786	41.700
	D	(1.326)	(6.840)	(3.625)	(16.595)	(5.254)

Table 3.2: *Statistical analysis of results of twelve neuropsychological tests (Cont'd). A: Posterior mean, B: Posterior standard deviation, C: Unbiased estimate, D: Estimated standard error.*

Parameter	Statistic	Group (a)	(b)	(c)	(d)	(e)
		Test: Visual functions				
Group Mean	A	3.749	4.667	6.421	5.423	7.516
	B	(0.305)	(0.526)	(0.531)	(0.616)	(0.406)
	C	3.716	4.636	6.475	5.450	7.547
	D	(0.305)	(0.526)	(0.538)	(0.630)	(0.409)
Slope	A	0.112	0.158	-0.040	0.014	-0.014
	B	(0.026)	(0.077)	(0.050)	(0.081)	(0.032)
	C	0.114	0.192	-0.048	-0.003	-0.015
	D	(0.026)	(0.088)	(0.051)	(0.092)	(0.032)
Variance	A	6.396	6.527	11.767	8.217	21.408
	B	(1.114)	(2.027)	(2.669)	(2.684)	(2.687)
	C	6.251	6.092	11.573	7.941	21.387
	D	(1.097)	(1.926)	(2.655)	(2.647	(2.694)
		Test: Receptive speech				
Group Mean	A	3.600	7.921	9.211	8.428	8.441
	B	(0.364)	(0.834)	(0.853)	(0.855)	(0.537)
	C	3.537	8.000	9.300	8.550	8.461
	D	(0.358)	(0.847)	(0.865)	(0.867)	(0.540)
Slope	A	0.120	0.049	-0.064	0.002	-0.066
	B	(0.031)	(0.121)	(0.080)	(0.111)	(0.042)
	C	0.123	0.052	-0.072	-0.011	-0.068
	D	(0.031)	(0.141)	(0.083)	(0.126)	(0.042)
Variance	A	9.044	16.210	30.091	15.686	37.256
	B	(1.576)	(5.029)	(6.824)	(5.124)	(4.676)
	C	8.597	15.794	29.940	15.046	37.300
	D	(1.508)	(4.995)	(6.869)	(5.015)	(4.699)
		Test: Expressive speech				
Group Mean	A	5.066	11.253	12.446	8.662	12.884
	B	(0.516)	(1.059)	(1.067)	(1.187)	(0.923)
	C	4.985	11.409	12.575	8.650	12.930
	D	(0.511)	(1.069)	(1.082)	(1.219)	(0.929)
Slope	A	0.145	0.083	-0.005	0.053	-0.127
	B	(0.044)	(0.153)	(0.100)	(0.153)	(0.072)
	C	0.148	0.088	-0.011	0.045	-0.131
	D	(0.044)	(0.178)	(0.103)	(0.177)	(0.073)
Variance	A	18.209	26.265	47.229	30.256	110.280
	B	(3.173)	(8.152)	(10.711)	(9.881)	(13.841)
	C	17.501	25.161	46.821	29.700	110.531
	D	(3.070)	(7.957)	(10.742)	(9.900)	(13.926)

Table 3.3: *Statistical analysis of results of twelve neuropsychological tests (Cont'd). A: Posterior mean, B: Posterior standard deviation, C: Unbiased estimate, D: Estimated standard error.*

		Group				
Parameter	Statistic	(a)	(b)	(c)	(d)	(e)
		Test: Writing				
Group Mean	A	3.913	7.556	9.877	7.515	8.399
	B	(0.293)	(0.870)	(0.843)	(0.735)	(0.475)
	C	3.866	7.636	10.000	7.600	8.422
	D	(0.288)	(0.892)	(0.853)	(0.745)	(0.477)
Slope	A	0.103	0.042	-0.040	0.037	-0.014
	B	(0.025)	(0.125)	(0.079)	(0.095)	(0.037)
	C	0.105	0.037	-0.047	0.030	-0.015
	D	(0.025)	(0.148)	(0.081)	(0.108)	(0.037)
Variance	A	5.870	17.620	29.373	11.555	29.122
	B	(1.023)	(5.465)	(6.661)	(3.774)	(3.655)
	C	5.559	17.500	29.112	11.107	29.145
	D	(0.975)	(5.534)	(6.679)	(3.702)	(3.672)
		Test: Reading				
Group Mean	A	1.761	2.235	3.189	2.210	6.153
	B	(0.217)	(0.503)	(0.445)	(0.478)	(0.511)
	C	1.746	2.227	3.225	2.200	6.203
	D	(0.216)	(0.517)	(0.452)	(0.489)	(0.514)
Slope	A	0.087	0.069	-0.033	0.069	-0.068
	B	(0.018)	(0.073)	(0.042)	(0.062)	(0.040)
	C	0.088	0.074	-0.040	0.073	-0.070
	D	(0.018)	(0.086)	(0.043)	(0.071)	(0.040)
Variance	A	3.227	5.978	8.267	4.962	33.819
	B	(0.562)	(1.854)	(1.875)	(1.621)	(4.245)
	C	3.134	5.875	8.161	4.785	33.785
	D	(0.550)	(1.858)	(1.872)	(1.595)	(4.257)
		Test: Arithmetic				
Group Mean	A	1.984	4.090	7.432	6.763	8.104
	B	(0.321)	(0.652)	(0.920)	(1.029)	(0.662)
	C	1.940	4.09	7.550	6.950	8.148
	D	(0.316)	(0.654)	(0.933)	(1.042)	(0.666)
Slope	A	0.117	0.061	-0.082	-0.119	-0.066
	B	(0.027)	(0.095)	(0.086)	(0.134)	(0.052)
	C	0.120	0.068	-0.091	-0.172	-0.067
	D	(0.027)	(0.109)	(0.089)	(0.152)	(0.052)
Variance	A	7.029	9.903	34.995	22.740	56.658
	B	(1.225)	(3.074)	(7.936)	(7.427)	(7.111)
	C	6.698	9.405	34.820	21.725	56.728
	D	(1.175)	(2.974)	(7.988)	(7.242)	(7.147)

Table 3.4: *Statistical analysis of results of twelve neuropsychological tests (Cont'd). A: Posterior mean, B: Posterior standard deviation, C: Unbiased estimate, D: Estimated standard error.*

Parameter	Statistic	Group (a)	(b)	(c)	(d)	(e)
		Test: Memory				
Group Mean	A	5.156	7.060	8.452	7.760	9.927
	B	(0.407)	(0.682)	(0.613)	(0.969)	(0.454)
	C	5.104	7.045	8.500	7.800	9.961
	D	(0.406)	(0.692)	(0.622)	(1.007)	(0.457)
Slope	A	0.186	0.034	0.040	0.003	-0.000
	B	(0.034)	(0.099)	(0.058)	(0.127)	(0.036)
	C	0.191	0.020	0.038	-0.019	-0.001
	D	(0.035)	(0.115)	(0.059)	(0.147)	(0.036)
Variance	A	11.375	10.985	15.658	20.364	26.774
	B	(1.982)	(3.409)	(3.551)	(6.649)	(3.360)
	C	11.042	10.531	15.466	20.269	26.720
	D	(1.937)	(3.330)	(3.548)	(6.756)	(3.366)
		Test: Intellectual processes				
Group Mean	A	10.464	15.634	17.910	18.987	19.229
	B	(0.805)	(1.497)	(1.381)	(1.957)	(0.813)
	C	10.343	15.636	18.000	19.250	19.273
	D	(0.801)	(1.530)	(1.404)	(2.006)	(0.818)
Slope	A	0.193	-0.046	-0.213	-0.345	-0.140
	B	(0.068)	(0.219)	(0.130)	(0.257)	(0.064)
	C	0.203	-0.038	-0.224	-0.436	-0.142
	D	(0.068)	(0.254)	(0.134)	(0.292)	(0.064)
Variance	A	44.496	52.628	79.203	82.833	85.675
	B	(7.752)	(16.327)	(17.962)	(27.049)	(10.753)
	C	42.984	51.496	78.851	80.467	85.549
	D	(7.540)	(16.285)	(18.090)	(26.822)	(10.778)
		Test: Intermediate memory				
Group Mean	A	3.130	4.644	5.494	5.027	6.269
	B	(0.300)	(0.536)	(0.378)	(0.466)	(0.365)
	C	3.090	4.636	5.525	5.050	6.289
	D	(0.301)	(0.551)	(0.382)	(0.472)	(0.367)
Slope	A	0.060	0.059	-0.000	0.063	0.018
	B	(0.025)	(0.077)	(0.035)	(0.060)	(0.028)
	C	0.061	0.066	-0.004	0.070	0.018
	D	(0.026)	(0.092)	(0.036)	(0.069)	(0.029)
Variance	A	6.197	6.788	5.968	4.710	17.237
	B	(1.079)	(2.106)	(1.354)	(1.539)	(2.163)
	C	6.079	6.681	5.840	4.461	17.269
	D	(1.066)	(2.113)	(1.340)	(1.487)	(2.176)

The posterior distributions of λ_θ, λ_β, and ν were somewhat sensitive to the choice of prior, as already illustrated in the analysis of the visual functions test. The posterior probabilities P_θ, P_β, and P_ϕ, even if very sensitive quantitatively, in the overwhelming majority of situations were very small apart from P_ϕ for the memory test and prior (3) that gave value 0.078, hence indicating less validity of the equal mean, slope and variance hypotheses. The exact values of the Bayesian significance probabilities are not reported here since they were very low.

The values of the posterior means of the random effects are very stable to the choice of prior distribution, especially if the corresponding posterior standard deviations are considered, and quite close to their unbiased estimates. It is worth noticing the shrinking of the posterior means towards an overall mean, compared with the respective unbiased estimates, since they incorporate the information on the distribution of the θ_i, β_i and ϕ_i of the model hierarchy. The shrinkages are quite small and this is related to the sample sizes, which are quite high. The saving in the estimated standard errors compared to the unbiased estimates are also very limited, since we don't actually know the hyperparameters of the random effects.

The means of the student group are the ones with the smallest posterior means for all twelve tests and three sets of priors, as expected. However, contrary to forensic scientist expectations, we didn't observe the highest means being those of groups ((b) and (c), rather, in eight out of the twelve tests, the medical patient group (e) had the highest posterior mean, although only for the reading test the difference was significant. For the remaining seven tests, the difference of the means of group (e) and at least two of the three offender groups was not of practical significance. For the motor functions test, offender group (c) had the highest posterior mean closely followed by groups (b) and (e), with the murderer group (d) having the second lowest mean. For the rhythm test the (descending) group ordering was (b), (e), (c), (b) and (a), for the receptive speech test it was (c), (e), (b), (b) and (a), and for the the writing test the ordering was (c), (e), (b), (d) and (a). The ordering depended on the choice of prior for the last four tests, however in all of them there were no practical differences between the three offender groups and the medical one, while all four groups appeared to have quite higher scores than the students (a).

In all cases, the posterior means of the slopes are very small in absolute value. As a result, combined with the values of $x_{i.} - x_{..}$, the deviation of the ith group covariate mean from the overall one, the slopes would tend to change the posterior distribution

of the unadjusted group means, θ_i, very slightly to obtain the posterior density of the adjusted means, ξ_i, and hence the values of the latter are not reported here. For most of the tests the posterior means are positive suggesting higher scores for older people. This appear to be reversed in groups (c) and (e), where for most tests the posterior means are negative.

The posterior means of the variances of the five groups present a stable pattern across all tests for all three priors. The estimate for the medical patients group is always the highest and significantly higher than the estimate for the student group, which is always the smallest apart from the motor functions test. The variances of the three offender groups tend to be close to each other with variable orderings, just as observed with the posterior means of the group means.

The results suggest that none of the neuropsychological tests can be used to distinguish any one of the offender groups (b), (c), (d) from both the other offender groups, and the Stanford medical group (e). Age being the only confounding variable for which data are available for all five groups, our results are a bit restricted. Moreover, as with all observational studies, we cannot rule out the existence of another confounder that unlike age has a very substantial effect on the test scores, according to which the current data are highly imbalanced and which could potentially dramatically change the current conclusions, after adjusting the scores for this confounder. It is however not obvious that the current neuropsychological tests with the data available can be used to profile offenders in a meaningful way. Additionally the results obtained definitely contradict any initial guesses.

3.2 Nutrition data example

We proceed by reporting the results of our analysis of the data coming from an experiment in food microbiology and toxicology. It was conducted to study the effects of different food additives on animal weight gain. Two levels (0.25% and 0.5%) of an additive (CLA), a control and a single level (0.5%) of a second additive (LA) were used. We label them 1 to 4, in the same order. Feed intake was a possible covariate, with quite different values across the four groups. Table 3.5 presents a summary of the data.

A fixed effects constant variance ANCOVA analysis is described by Yandell (1997, pp. 256-273). By using a constant slope model and adjusted for the rest of the parameters in the model F (Type III) tests, significant differences between the four different

additives were obtained, as well as a positive slope for feed intake relating higher weight gain to more feed intake. In a further step, a test for a statistically significant interaction between factor and covariate, corresponding to different slopes for different levels of the factor, was not significant ($F = 3.60$ on 3 and 4 d.f.), while an attempt to use two different slopes, one for the two high levels of additives and one for the control and the low level of additive CLA gave a rather significant result, ($F = 15.14$ on 1 and 6 d.f.), indicating the intuitive conclusion that equal increases in feed intake result in steeper weight gain for animals taking smaller concentration additives.

Table 3.5: *Data summary for nutrition example.*

Group (i)	n_i	s_i^2	$\hat{\theta}_i$	$\hat{\beta}_i$	$\hat{\phi}_i$
1	3	21110.542	1457.567	0.459	96.344
2	3	2875.227	1479.033	0.057	519.133
3	3	93219.447	1409.667	0.496	243.430
4	3	18906.887	1338.900	0.145	17.394

The information criterion in (2.17) was $BIC^* = -65.18$, against $BIC^* = -90.74$ for the constant variance, unequal slope and mean model, suggesting the plausibility of the unequal variance assumption. Note however that, due to the extremely small sample sizes, there is little information in the data about the variances. BIC^* uses point estimates for the ϕ_i, but ignores their high standard errors. We will proceed with the analysis assuming we have definite prior information. If this is not true, the subsequent analysis would be rather implausible, since the data contain little information.

We performed the analysis using three sets of prior parameters $\omega_1 = 4.5$, $\tau_1 = 1.2$, $\omega_2 = 1$, $\tau_2 = 0.0035$, $\zeta_0 = 60$, $\psi = 2/15$, a =2.2 , and $b = 0.4$ for set (1), $\omega_1 = 4.5$, $\tau_1 = 2$, $\omega_2 = 4.5$, $\tau_2 = 0.005$, $\zeta_0 = 120$, $\psi = 1/60$, a =18, and $b = 2$ for set (2), and $\omega_1 = 4.5$, $\tau_1 = 5$, $\omega_2 = 20$, $\tau_2 = 0.01$, $\zeta_0 = 198$, $\psi = 1/90$, a=30 , and $b = 1$ for set (3). We report the posterior densities of the θ_i, β_i and ϕ_i under the second choice of prior distribution. As in all previous cases studies, the posterior densities of the random effects are quite insensitive to the choice of prior and result to identical inference in practice.

The posterior densities of the group means, in the top left plot of Figure 3.7, suggest the ordering (d), (c), (a) and (b), with significant pairwise differences, apart the one of groups (a) and (b). Hence there is an indication that the first additive (CLA), at both

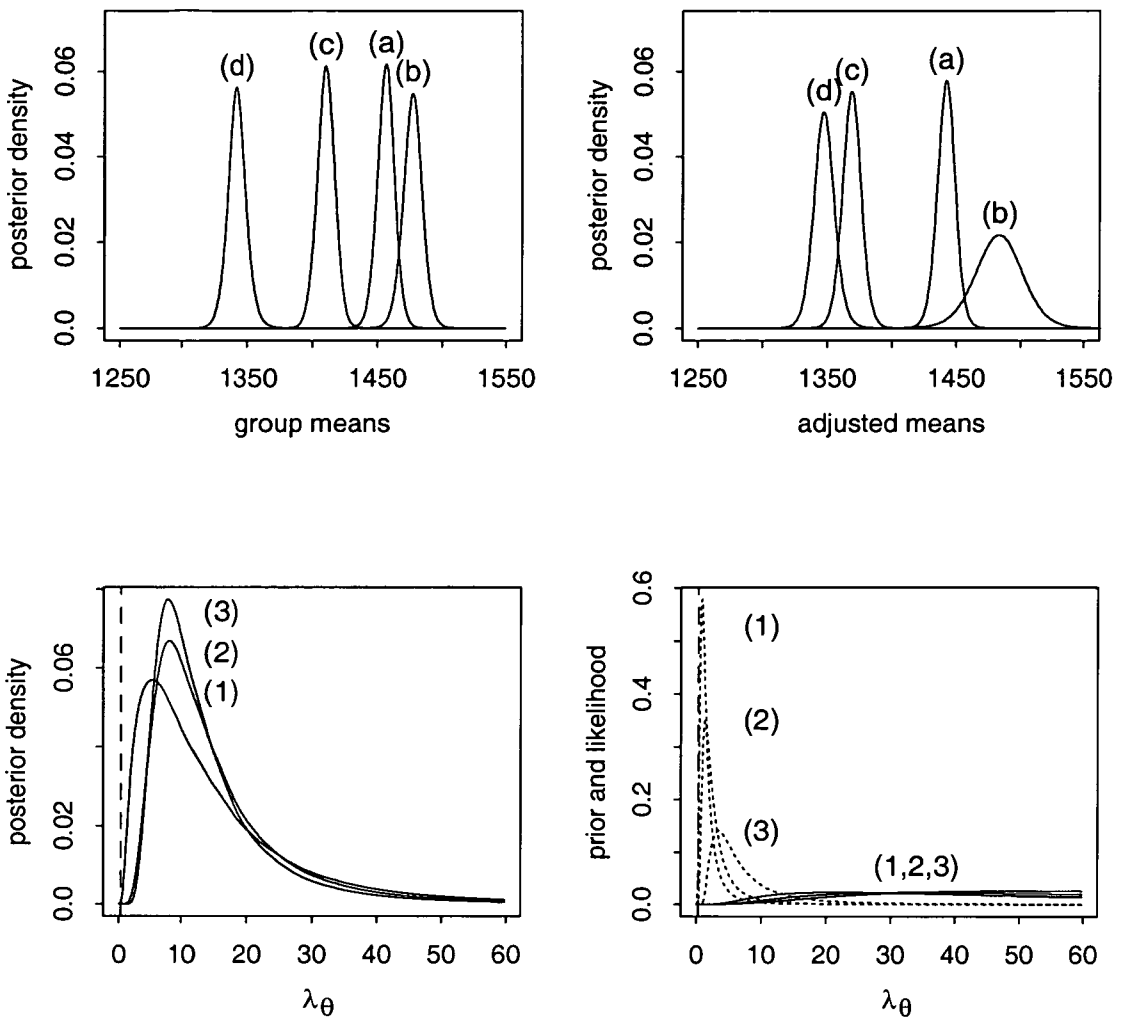


Figure 3.7: Nutrition data. Top left: Posterior density of four, (a)-(d), group means (θ_i). Top right: Posterior density of four, (a)-(d), adjusted group means (ξ_i). Bottom left: Posterior density of λ_θ under three choices, (1)-(3), of prior distribution. Bottom right: Integrated likelihood (solid line) and prior density (dotted line) of λ_θ, under three choices, (1)-(3), of prior distribution. The dashed vertical lines correspond to $\lambda_\theta = \widetilde{\lambda_\theta}$.

concentrations, is associated with higher weight gain than the control and the other additive (LA). The ξ, in the top right plot of Figure 3.7, that compare the different group means at the same level of the covariate, $x_{..}$, give slightly different conclusions. Although the ordering of the four groups has remained the same, the difference of groups (c) and (d) no longer is of statistical significance, while the adjusted mean of group (b) has quite larger variance, compared to the mean, θ_2, has much higher posterior variance, reflecting the high posterior variance of the corresponding slope parameter. We address the question of overall equality of the group means by considering the posterior density of λ_θ, in the bottom left plot of Figure 3.7, and P_θ, the posterior probability that $\lambda_\theta \leq \widetilde{\lambda_\theta} = 1/3$. The posterior density of λ_θ is quite sensitive to the choice of prior

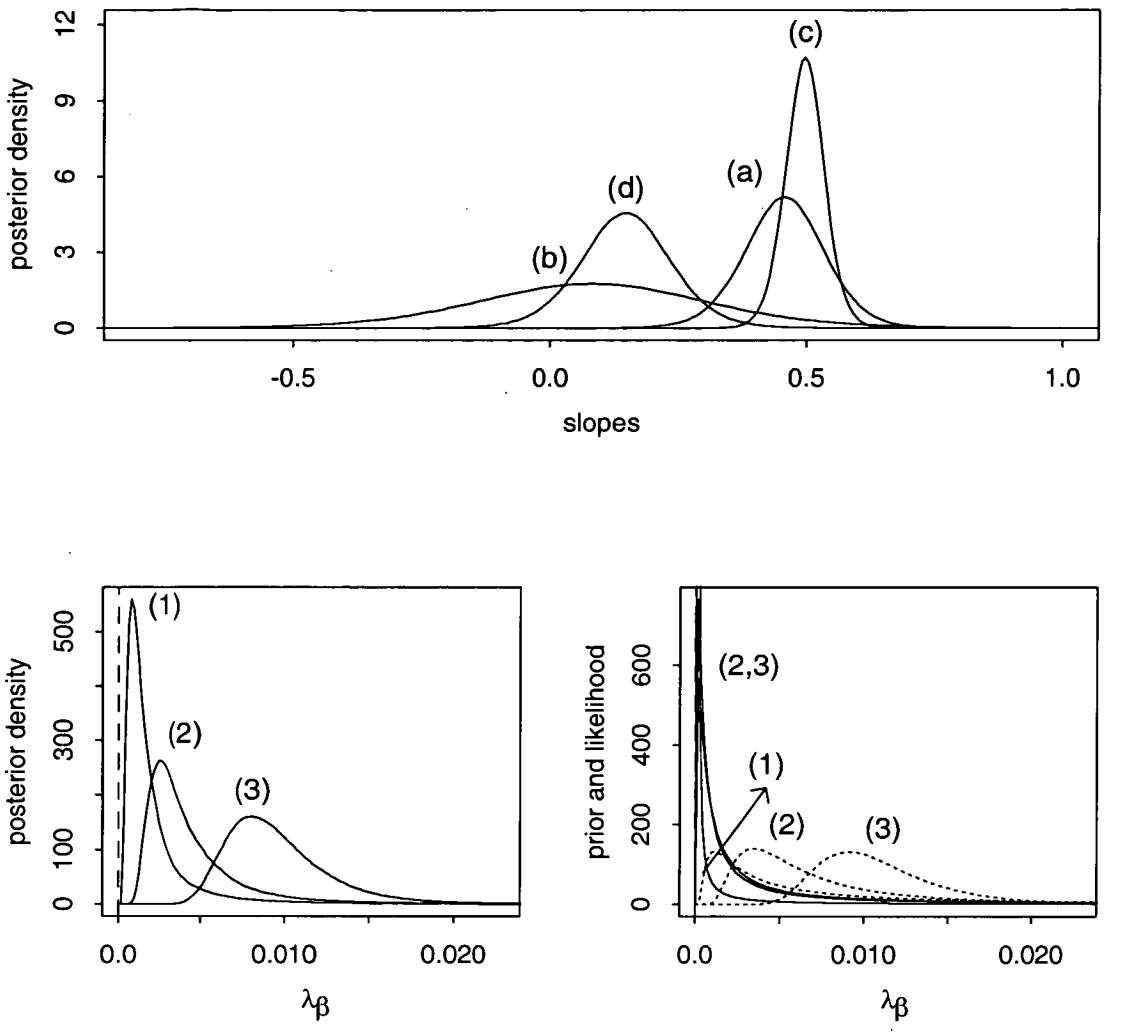


Figure 3.8: Nutrition data. Top: Posterior density of four, (a)-(d), group slopes (β_i). Bottom left: Posterior density of λ_β under three choices, (1)-(3), of prior distribution. Bottom right: Integrated likelihood (solid line) and prior density (dotted line) of λ_β, under three choices, (1)-(3), of prior distribution. The dashed vertical lines correspond to $\lambda_\beta = \widetilde{\lambda_\beta}$.

distribution, but in all three cases, P_θ is negligible, less than 10^{-5}, strongly refuting $H_\theta : \lambda_\theta = 0$, the hypothesis of equality of the group means. The integrated likelihood for λ_θ, in the bottom right plot of Figure 3.7, is very stable to the choice of prior but quite flat. Adopting the criterion already described for testing the hypothesis H_θ, this time using the integrated likelihood, would result in contradictory inference to the one already obtained based on the posterior density of λ_θ. This an indication of the substantial part played by the prior in the posterior inference, in conjunction with the small sample sizes. Our findings, however, seem to confirm the practical results of the analysis by Yandell.

The posterior densities of the slopes, in the top plot of Figure 3.8, suggest that (b)

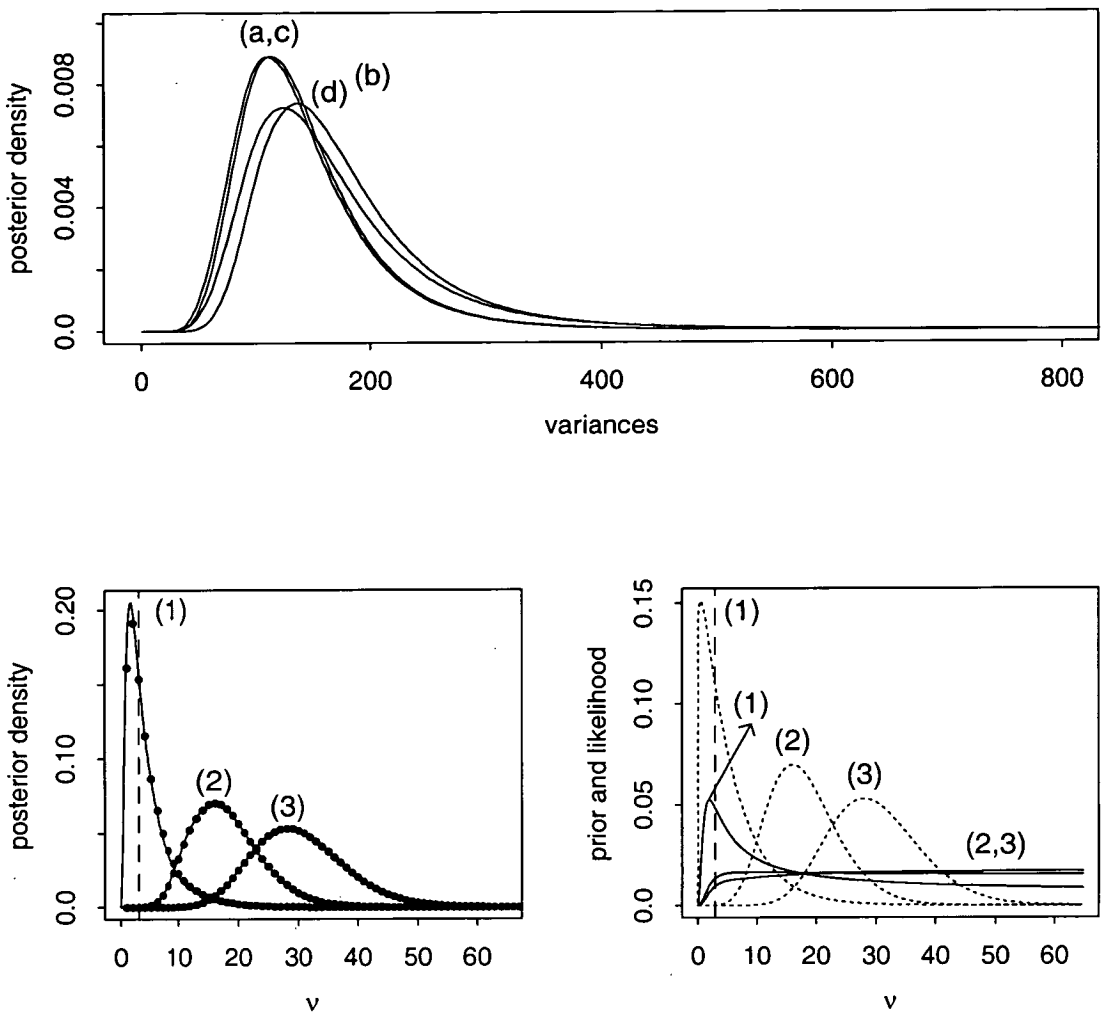


Figure 3.9: Nutrition data. Top: Posterior density of four, (a)-(d), group variances (ϕ_i). Bottom left: Posterior density/probability mass function of ν (solid line under continuous prior, $\cdots$ under discrete prior) under three choices, (1)-(3), of prior distribution. Bottom right: Integrated likelihood (solid line) and prior density (dotted line) of ν, under three choices, (1)-(3), of prior distribution. The dashed vertical lines correspond to $\nu = \nu_\phi$.

and (d), the two 0.5% treatments, are associated with slower increases in weight gain with feed intake, than additives (a) and (c), confirming results already discussed. The posterior density of λ_β is, in the bottom left plot of Figure 3.8, is typically sensitive to the choice of prior and again P_β, the probability that $\lambda_\beta \leq \widetilde{\lambda_\beta}$, is very small, less than 10^{-10} in all three cases, strongly refuting the slope equality hypothesis, H_β. The integrated likelihood of λ_β is quite stable and would also indicate rejection of H_β, according to our proposed criterion.

The posterior densities of the variances ϕ_i under prior distribution (2), are described in Figure 3.9. Based on that figure and without studying the posterior density of ν,

or calculating the posterior probability that $\nu \geq \nu_\phi$, we can conclude that the data do not provide enough evidence for the rejection of H_ϕ, the variance equality hypothesis. Comparing the posterior means of the variances with their unbiased estimates (in Table 3.5), we can see that they are very far apart and their ordering hasn't been preserved. This ordering reversal can be explained by the dependence of the posterior density of the variances on s_i^2 and the big discrepancies of the latter across different groups, as depicted in Table 3.5. In fact, the original suggestion, based on the same table, that the variances are unequal, is not true. To see this, observe that by equation (3.1), for $n_i = 3$, the estimated standard error of the variance is equal to its unbiased estimate multiplied by $\sqrt{2}$. The fact that the variances do not appear to be unequal, however, doesn't mean that we should set them to be equal. Inference about the variances on such small sample sizes is not reliable, a conclusion supported by the flat integrated likelihoods in Figure 3.9, and the close agreement of the prior and posterior distributions of ν.

3.3 Simulated data examples

So far, in the practical examples we studied, the decision about whether the groups variances were equal or not, was quite clear, given the graphs of the corresponding posterior densities. We will proceed by presenting examples of simulated data sets to examine whether our suggestion, of considering the posterior probability that $\nu \geq \nu_\phi$ as a Bayesian significance probability for testing the variance equality hypothesis, is useful in practice.

We considered two sets of simulated data, both of which with $m = 10$ and, for simplicity, $\beta_i = 0$, essentially reducing our model to a random effects ANOVA one. The data for the first set were generated using $\theta_i = 10i$ and $\phi_i = 4 + 25i$ for $i = 1, \ldots, m$, and used $K = 10$ runs. For each run k, with $k = 1, \ldots, K$, the ten sample sizes were equal to $n_i = 5k$, for $i = 1, \ldots, m$, hence we were just increasing the sample sizes between each run, while keeping them equal within each run. For the second set of data we used $\theta_i = 0$ and sample sizes randomly generated and equal to 23, 46, 23, 39, 28, 45, 21, 41, 31, and 22, for all $k = 1, \ldots, K$ simulations, then randomly generated the observations using the variance vector $\phi_i = 10$, for $i = 1, \ldots, 5$, and $\phi_i = 10 + (k-6) \times 8$, for $i = 6, \ldots, 10$, hence increasingly separating the variances. As we mentioned in the previous example, because of the large sampling variation of the observed residual sums

of squares, for the second set of data, we replaced them by their expectations, hence being able to observe the gradual separation of the posterior densities of the variances, since in this case, the posterior densities were more settled..

For each data set, we report the posterior densities of the variances, the posterior density of the parameter ν, the suggested Bayesian significance probability, and the value of Bartlett's test statistic (see Draper and Smith, 1999, p. 56), together with the corresponding exact significance probability for the hypothesis of variance equality. The exact significance probability was very close to the approximate one, based on the chi-squared distribution with 9 degrees of freedom. The results for the first set of data are displayed in Figure 3.10 and Table 3.6 and for the second one in Figure 3.11 and Table 3.7.

The significance probability of Bartlett's test does not numerically agree with our suggested significance probability, P_ϕ, so strictly speaking, if we wanted to base a decision only on a significance probability, we would get contradictory answers in some situations, although the orderings of the significance probabilities seem to roughly agree. On the other hand, our method enables us to make an applied decision regarding variance equality looking at their posterior densities. This applied decision, in both sets of simulated data, very well matches the posterior density of the ϕ_i with the posterior density of ν and the corresponding interpretation via the posterior probability that $\nu \geq \nu_\phi = \max n_i$.

Based upon applied considerations of the posterior densities of the variances in Figures 3.10 and 3.11, e.g. overlapping tail areas, we judge that it is reasonable to infer variance inequality for $k = 5, \ldots, 10$ for the first data set and $k = 4$ or $5, \ldots, 10$ for the second. Hence Bayesian significance probabilities as small as 0.05 (see Figure 3.10 and Table 3.6), and in the range 0.01 to 0.08 (see Figure 3.11 and Table 3.7) correspond to our applied judgement of inequality. More generally, we have found that if the Bayesian significance probability is less than 0.05, then this corresponds well with an applied judgement of the posterior densities of the variances and gives an intuitive justification for the choice $\nu_\phi = \max n_i$. This can be extremely useful for an applied statistician.

For the first data set, as we demonstrated in the nutrition example, inference about the variances with small sample sizes is quite difficult to obtain, hence although the variances from which we generated the data were quite different, the small sample sizes together with the large sampling errors conceal these differences. For the second data

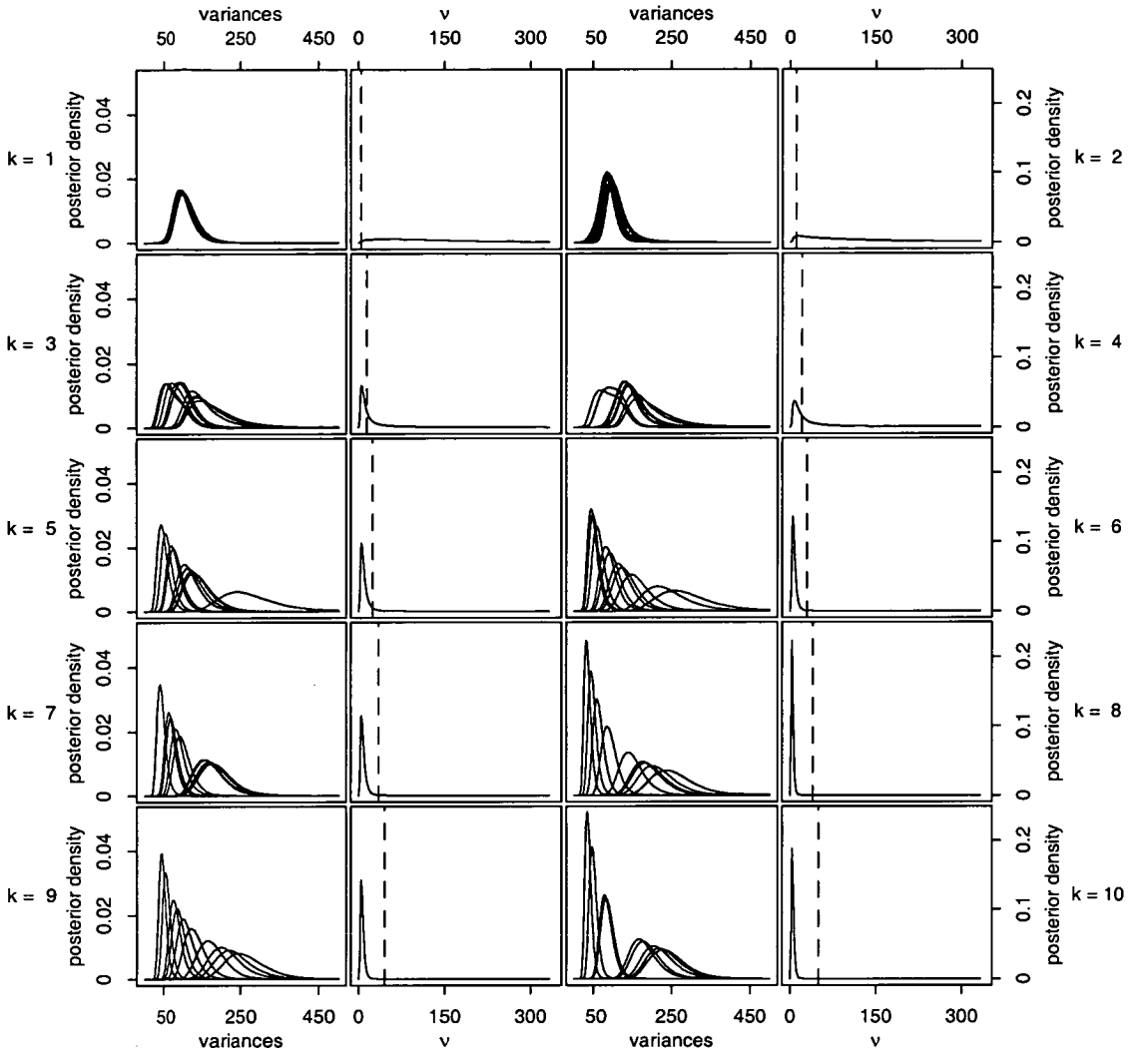


Figure 3.10: Posterior density of the ten group variances (ϕ_i) and posterior density of ν for the first set of simulated data. The dashed vertical lines correspond to $\nu = \nu_\phi = 10$.

Table 3.6: *Variance equality test results for first group of simulated data sets. A: Bayesian significance probability, B: Bartlett's test statistic value, C: Bartlett's test significance probability.*

Data Set		1	2	3	4	5
	A	1.000	0.992	0.693	0.786	0.504×10^{-1}
Statistic	B	6.653	15.354	31.181	30.026	48.157
	C	0.674	0.816×10^{-1}	0.254×10^{-3}	0.424×10^{-3}	$< 10^{-10}$

Data Set		6	7	8	9	10
	A	0.240×10^{-2}	0.552×10^{-2}	0.427×10^{-5}	0.152×10^{-4}	0.209×10^{-5}
Statistic	B	64.352	57.506	94.376	82.274	106.147
	C	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$

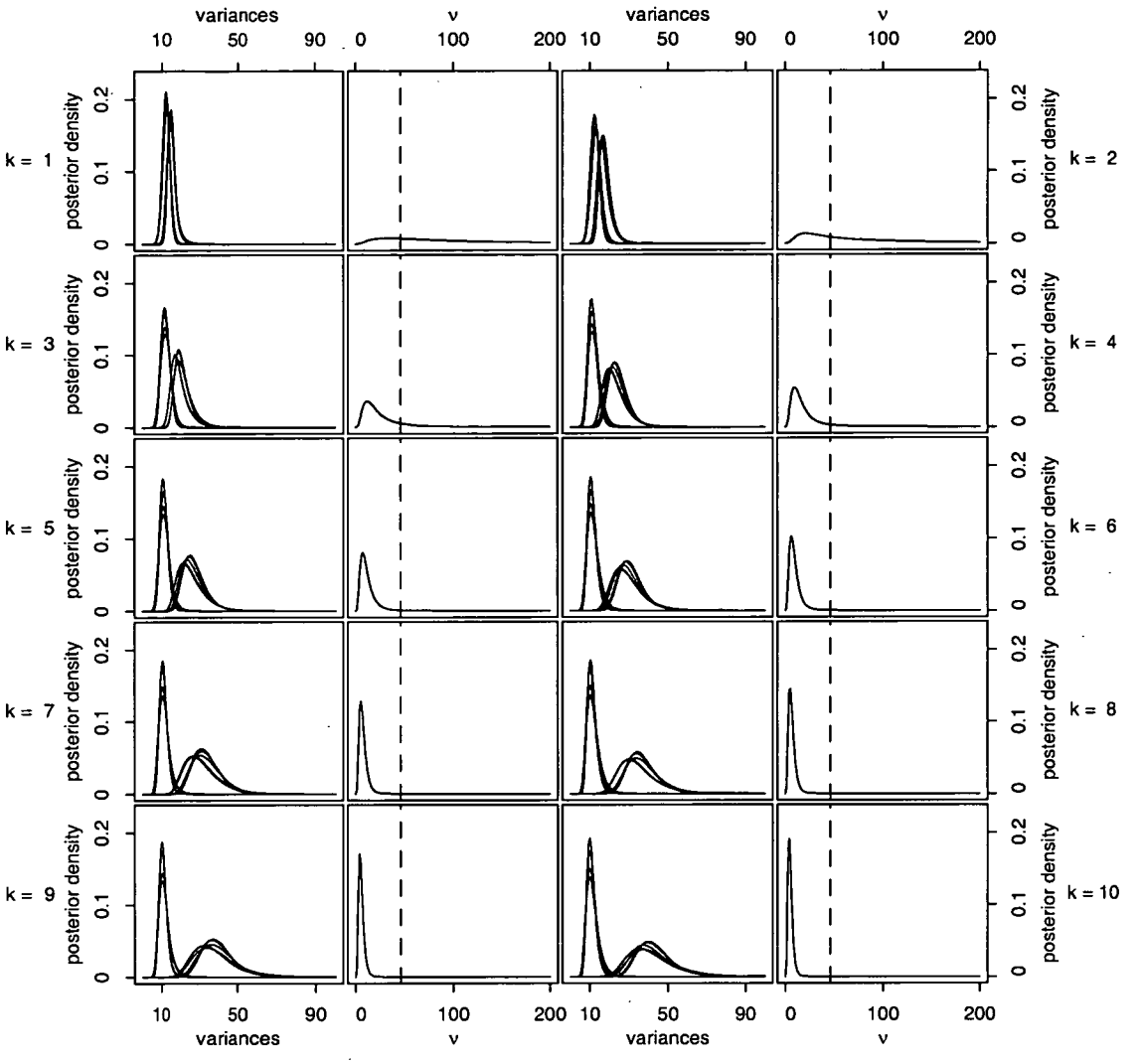


Figure 3.11: Posterior density of the ten group variances (ϕ_i) and posterior density of ν for the second set of simulated data. The dashed vertical lines correspond to $\nu = \nu_\phi = 46$.

Table 3.7: *Variance equality test results for second group of simulated data sets. A: Bayesian significance probability, B: Bartlett's test statistic value, C: Bartlett's test significance probability.*

Data Set		1	2	3	4	5
	A	0.937	0.776	0.302	0.817×10^{-1}	0.111×10^{-1}
Statistic	B	15.435	23.094	30.921	38.711	46.353
	C	0.795×10^{-1}	0.588×10^{-2}	0.269×10^{-3}	0.130×10^{-4}	$< 10^{-10}$

Data Set		6	7	8	9	10
	A	0.197×10^{-2}	0.130×10^{-3}	0.261×10^{-4}	0.808×10^{-5}	0.194×10^{-5}
Statistic	B	53.792	61.002	67.972	74.702	81.198
	C	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$

set, where we used the expected values of the variances and quite large sample sizes, the decrease of the P_ϕ very well reflects the gradual separation of the two subsets of five variances.

We have established the usefulness of our proposed models in the analysis of several data sets, however, all the results presented were obtained assuming the convergence of the MCMC simulations, having taken precautionary measures (i.e. long chains), but without having validated this assumption. In the next chapter we will address this problem, not by using some formal convergence criterion, whose weaknesses have been previously discussed, but by obtaining posterior inferences through completely different methods and comparing the relevant results.

Chapter 4

Applications of Laplacian methods

The main subject of this chapter is the description and application of Laplacian type approximations to obtain the marginal posterior density of each of the random effects and parameters of the random variance ANCOVA model studied in sections 2.1 to 2.4. In the first part of this chapter we will present a brief review of Laplacian methods leading to the derivation of the approximation we will be applying to the ANCOVA problem. They will be followed by two simple introductory examples and the full application in all algebraic detail.

4.1 Background

Laplacian methods do not only provide simple initial approximations useful for starting points for further exact computations, e.g. starting points for MCMC simulations, but can produce extremely accurate results when compared to those of the exact methods and substantially more accurate than results produced by normal approximations, while being computationally less intensive than simulation methods. Hence, their computational efficiency and potential accuracy provide an appealing alternative to MCMC and other methods, like the quadrature based one by Naylor and Smith (1982), for obtaining marginal posterior distributions. Hence their application will have a twofold purpose, both confirming the convergence of the MCMC procedures of Chapter 2 and saving computer time.

The central idea in Laplace's method for evaluating integrals is approximating the

integrand by a normal curve centered at its mode and having variance equal to minus the inverse of the Hessian matrix of the log of the integrand evaluated at its mode.

Leonard (1982), having justified Laplacian approximations to predictive distributions, recommended using a conditional version of Laplace's method to approximate marginal posterior densities for subsets of a vector parameter by using a normal approximation to the posterior density of the nuisance parameters conditional on the parameters of interest. He similarly obtained a modified expression for the profile likelihood function, the modification being the inverse of the square root of the determinant of the likelihood information matrix of the parameters of interest, conditionally on the remaining parameters.

Tierney and Kadane (1986) suggested approximations to posterior moments, in addition to marginal posterior densities, and provided detailed derivations of their asymptotic errors. They concluded that these Laplacian approximations are of the same order as the errors of saddlepoint approximations. In particular they showed that the error of approximating a marginal density is asymptotically of the order of n^{-1} in some fixed neighborhood of the parameter of interest, where n is the sample size, and hence smaller than the error of the normal approximation, which is of the order $O(n^{-1/2})$. This error is further reduced to become of the order $O(n^{-3/2})$, if one elects to renormalize the resulting marginal posterior density, and hence remove the error in the constant of integration. It needs to be emphasized, however, that these error magnitude results are asymptotic and may not be that useful for finite sample inference. Nevertheless, when carefully applied, conditional Laplacian procedures can be very accurate.

Let a $p \times 1$ vector $\boldsymbol{\theta}$ possess posterior density $\pi_{\boldsymbol{y}}(\boldsymbol{\theta})$, and $\eta = \boldsymbol{\alpha}^T\boldsymbol{\theta}$ denote a linear transformation of $\boldsymbol{\theta}$, with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)^T$ fixed and specified. Assume that preliminary transformations adjust the conditional density of $\boldsymbol{\theta}$, given η, towards either a multivariate normal or generalized multivariate $t-$density. Following the developments by Leonard (1982), and Tierney and Kadane (1986), and further advances by Leonard and Novick (1986), who used Laplacian approximations to marginalize posterior densities in the context of log-linear models for contingency tables, and Leonard, Hsu and Tsui (1989), Tierney, Kass and Kadane (1989a) show that the posterior density of η can be represented, to saddle point accuracy, by

$$\pi^*_{\boldsymbol{y}}(\eta) \propto \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_\eta) / \left\{ |\boldsymbol{R}_\eta|^{1/2} (\boldsymbol{\alpha}^T \boldsymbol{R}_\eta^{-1} \boldsymbol{\alpha})^{1/2} \right\}, \tag{4.1}$$

where $\boldsymbol{\theta}_\eta$ maximizes $\pi_{\boldsymbol{y}}(\boldsymbol{\theta})$ subject to $\boldsymbol{\alpha}^T\boldsymbol{\theta} = \eta$, and

$$\boldsymbol{R}_\eta = -\partial^2 \log \pi_{\boldsymbol{y}}(\boldsymbol{\theta})/\partial(\boldsymbol{\theta}\boldsymbol{\theta}^T)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\eta}. \tag{4.2}$$

This approximation is not completely accurate for hierarchical models and only applies to positive functions η. Extensions to expectations of nonpositive functions, by applying the approximation to moment generating functions, thus ensuring positive integrands, and then taking the appropriate derivatives, or adding a large constant to the nonpositive function and then subtracting it from the approximation, as well as similar approaches for approximating variances of nonpositive functions are discussed by Tierney, Kass and Kadane, (1989b).

4.2 A Laplacian approximation

We can however refer to a modification, which is a special case of suggestions by Leonard et al (1989), and Sun et al (1996), and is briefly reported by Leonard and Hsu (1999). The modification is

$$\pi^*_{\boldsymbol{y}}(\eta) \propto \pi_{\boldsymbol{y}}(\boldsymbol{\theta}^*_\eta) \exp(\frac{1}{2}\boldsymbol{\ell}^T_\eta \boldsymbol{G}_\eta \boldsymbol{\ell}_\eta)/\left(|\boldsymbol{R}^*_\eta|^{1/2}\omega^{1/2}_\eta\right), \tag{4.3}$$

where

$$\boldsymbol{\ell}_\eta = \partial \log \pi_{\boldsymbol{y}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_\eta}, \tag{4.4}$$

$$\boldsymbol{R}^*_\eta = -\partial^2 \log \pi_{\boldsymbol{y}}(\boldsymbol{\theta})/\partial(\boldsymbol{\theta}\boldsymbol{\theta}^T)|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_\eta}, \tag{4.5}$$

$$\boldsymbol{G}_\eta = (\boldsymbol{R}^*_\eta)^{-1} - \omega^{-1}_\eta (\boldsymbol{R}^*_\eta)^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T(\boldsymbol{R}^*_\eta)^{-1}, \tag{4.6}$$

and

$$\omega_\eta = \boldsymbol{\alpha}^T(\boldsymbol{R}^*_\eta)^{-1}\boldsymbol{\alpha}, \tag{4.7}$$

where $\boldsymbol{\theta}^*_\eta$ denotes some representative of the region $D = \{\boldsymbol{\theta} : \boldsymbol{\alpha}^T\boldsymbol{\theta} = \eta\}$, with high conditional posterior density, given η. This approximation reduces to (4.1) when $\boldsymbol{\theta}^*_\eta$ is replaced by the conditional mode $\boldsymbol{\theta}_\eta$. A direct derivation of approximation (4.3) follows.

Let $\pi_{\boldsymbol{y}}(\boldsymbol{\theta})$ denote the posterior density of a $p \times 1$ vector $\boldsymbol{\theta}$, and $\eta = \boldsymbol{\alpha}^T\boldsymbol{\theta}$ be some

linear function of $\boldsymbol{\theta}$. Expanding $\log \pi_{\boldsymbol{y}}(\boldsymbol{\theta})$ in a Taylor series about $\boldsymbol{\theta} = \boldsymbol{\theta}_\eta^*$ and neglecting cubic and higher order terms in the series, gives the second order approximation

$$\begin{aligned} \log \pi_{\boldsymbol{y}}^*(\boldsymbol{\theta}) &= \log \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_\eta^*) + \boldsymbol{\ell}_\eta^T(\boldsymbol{\theta} - \boldsymbol{\theta}_\eta^*) - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\eta^*)^T \boldsymbol{R}_\eta^*(\boldsymbol{\theta} - \boldsymbol{\theta}_\eta^*) \\ &= \log \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_\eta^*) + \tfrac{1}{2}\boldsymbol{\ell}_\eta^T(\boldsymbol{R}_\eta^*)^{-1}\boldsymbol{\ell}_\eta - \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_\eta)^T \boldsymbol{R}_\eta^*(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_\eta) \end{aligned} \tag{4.8}$$

where $\boldsymbol{\ell}_\eta$ and $\boldsymbol{R}_\eta^*$ satisfy (4.4) and (4.5) and

$$\tilde{\boldsymbol{\theta}}_\eta = \boldsymbol{\theta}_\eta^* + (\boldsymbol{R}_\eta^*)^{-1}\boldsymbol{\ell}_\eta. \tag{4.9}$$

Consequently, when $\boldsymbol{\alpha}^T\boldsymbol{\theta} = \eta$, and in some neighborhood of $\boldsymbol{\theta} = \boldsymbol{\theta}_\eta^*$, $\pi_{\boldsymbol{y}}(\boldsymbol{\theta})$ is approximated by

$$\begin{aligned} \pi_{\boldsymbol{y}}^*(\boldsymbol{\theta}) &= \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_\eta^*) \exp\left\{\tfrac{1}{2}\boldsymbol{\ell}_\eta^T(\boldsymbol{R}_\eta^*)^{-1}\boldsymbol{\ell}_\eta - \tfrac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_\eta)^T \boldsymbol{R}_\eta^*(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_\eta)\right\} \\ &\propto \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_\eta^*)|\boldsymbol{R}_\eta^*|^{-1/2} \exp\left\{\tfrac{1}{2}\boldsymbol{\ell}_\eta^T(\boldsymbol{R}_\eta^*)^{-1}\boldsymbol{\ell}_\eta\right\} \boldsymbol{\Psi}_{\boldsymbol{\theta}}\left(\tilde{\boldsymbol{\theta}}_\eta, (\boldsymbol{R}_\eta^*)^{-1}\right), \end{aligned} \tag{4.10}$$

where $\boldsymbol{\Psi}_{\boldsymbol{\theta}}(\boldsymbol{\mu}, \boldsymbol{C})$ denotes a multivariate normal density for $\boldsymbol{\theta}$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$. The next step can be rigorized via the $D-$region described by Leonard, Hsu, and Tsui (1989). In particular, the following statement holds,

$$\pi_{\boldsymbol{y}}(\eta) = \lim_{\gamma \to 0} \gamma^{-1} \int_{D_\gamma} \pi_{\boldsymbol{y}}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad \eta \in \Omega \subseteq R, \tag{4.11}$$

with D_γ denoting the region, $D_\gamma = D(\eta, \gamma) = \{\boldsymbol{\theta} : \eta \leq \boldsymbol{\alpha}^T\boldsymbol{\theta} \leq \eta + \gamma\}$, since the following probabilistic argument is true in the posterior:

$$\pi_{\boldsymbol{y}}(\eta) = \lim_{\gamma \to 0} \gamma^{-1} P(\eta \leq \boldsymbol{\alpha}^T\boldsymbol{\theta} \leq \eta + \gamma) = \lim_{\gamma \to 0} \gamma^{-1} P(\boldsymbol{\theta} : \boldsymbol{\theta} \in D) = \lim_{\gamma \to 0} \gamma^{-1} \int_D \pi_{\boldsymbol{y}}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{4.12}$$

Combining (4.10) with (4.11), we obtain that the marginal posterior density of η is approximated by

$$\pi_{\boldsymbol{y}}^*(\eta) \propto \pi_{\boldsymbol{y}}(\boldsymbol{\theta}_\eta^*)|\boldsymbol{R}_\eta^*|^{-1/2} \exp\left\{\frac{1}{2}\boldsymbol{\ell}_\eta^T(\boldsymbol{R}_\eta^*)^{-1}\boldsymbol{\ell}_\eta\right\} \Psi_\eta\left(\boldsymbol{\alpha}^T\tilde{\boldsymbol{\theta}}_\eta, \boldsymbol{\alpha}^T(\boldsymbol{R}_\eta^*)^{-1}\boldsymbol{\alpha}\right), \tag{4.13}$$

where $\Psi_\eta(\mu, \sigma^2)$ denotes a normal $N(\mu, \sigma^2)$ density for η. Using the definition of the normal density and equation (4.7), the previous expression can be rewritten as

$$\pi^*_{\boldsymbol{y}}(\eta) \propto \pi_{\boldsymbol{y}}(\boldsymbol{\theta}^*_\eta)|\boldsymbol{R}^*_\eta|^{-1/2}\omega_\eta^{-1/2} \exp\left\{\frac{1}{2}\boldsymbol{\ell}_\eta^T(\boldsymbol{R}^*_\eta)^{-1}\boldsymbol{\ell}_\eta - \frac{1}{2}\omega_\eta^{-1}(\eta - \boldsymbol{\alpha}^T\tilde{\boldsymbol{\theta}}_\eta)^2\right\}. \quad (4.14)$$

By equality (4.9), we have that

$$(\eta - \boldsymbol{\alpha}^T\tilde{\boldsymbol{\theta}}_\eta)^2 = \boldsymbol{\ell}_\eta^T(\boldsymbol{R}^*_\eta)^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T(\boldsymbol{R}^*_\eta)^{-1}\boldsymbol{\ell}_\eta. \quad (4.15)$$

A final substitution of (4.15) into (4.14), using definition (4.6), gives the expression (4.3).

The previous result does not suggest a simplification to (4.1) when $\boldsymbol{\theta}^*_\eta$ is replaced by $\boldsymbol{\theta}_\eta$ for general form of the function $\eta = g(\boldsymbol{\theta})$. However, the two approximations are algebraically equivalent when $\eta = \boldsymbol{\alpha}^T\boldsymbol{\theta}$, and $\boldsymbol{\theta}^*_\eta = \boldsymbol{\theta}_\eta$. This result can be demonstrated using the fact that if $\boldsymbol{\theta}^*_\eta = \boldsymbol{\theta}_\eta$, then from the application of the Lagrange multiplier method for obtaining the conditional maximum, we have that

$$\boldsymbol{\ell}_\eta = \Lambda\boldsymbol{\alpha}, \quad (4.16)$$

with Λ the Lagrange multiplier. Using (4.16) and (4.9), we can derive that

$$\omega_\eta\boldsymbol{\ell}_\eta^T(\boldsymbol{R}^*_\eta)^{-1}\boldsymbol{\ell}_\eta = (\eta - \boldsymbol{\alpha}^T\tilde{\boldsymbol{\theta}}_\eta)^2 = \Lambda^2\omega_\eta^2. \quad (4.17)$$

Substituting the last equality in (4.14), the exponent becomes 0, and the expression reduces to (4.1), since $\boldsymbol{R}^*_\eta = \boldsymbol{R}_\eta$.

Although, (4.3) possesses inferior asymptotic properties to (4.1), the quadratic term within the exponent can provide an essential extra contribution to this approximation for finite sample sizes and thus make it more useful, as demonstrated by Leonard et al (1989) with several relevant examples, all with general form of $\eta = g(\boldsymbol{\theta})$.

In case the suggested approximation failed to give sufficiently accurate results, we could consider further possibilities, like the Laplacian t approximation (Leonard et al, 1994), which can be rather more accurate in the tails of the required distributions, but involve the additional complication of having to estimate the degrees of freedom of the multivariate t distribution, and saddlepoint approximations (Daniels, 1954, Reid, 1988), which are generalizations of Edgeworth expansions, and can be considered a complex analysis and rather more complicated analogue of Laplacian approximations.

4.3 Two examples of Laplacian methods

4.3.1 A simple hierarchical model

Suppose that $y_1, y_2, \ldots, y_m$ are independent and normally distributed given respective means $\theta_1, \theta_2, \ldots, \theta_m$, and known variances $\tau^2/n_1, \tau^2/n_2, \ldots, \tau^2/n_m$, that $\theta_1, \theta_2, \ldots, \theta_m$ are independent and normally distributed with common mean, μ, and common variance σ^2, and that $\nu\zeta/\sigma^2$ has a chi-squared distribution with ν degrees of freedom. The joint posterior density of $\theta_1, \theta_2, \ldots, \theta_m$, and σ^2 is

$$\pi_{\boldsymbol{y}}(\boldsymbol{\theta}, \sigma^2) \propto (\sigma^2)^{-\frac{1}{2}(m+\nu+2)} \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{m} n_i(\theta_i - y_i)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\theta_i - \mu)^2 - \frac{\nu\zeta}{2\sigma^2}\right\}. \tag{4.18}$$

Integrating out σ^2 from (4.18), we find that the posterior density of $\theta_1, \theta_2, \ldots, \theta_m$, is

$$\pi_{\boldsymbol{y}}(\boldsymbol{\theta}) \propto \left\{\nu\zeta + \sum_{i=1}^{m}(\theta_i - \mu)^2\right\}^{-\frac{1}{2}(\nu+m)} \exp\left\{-\sum_{i=1}^{m} n_i(\theta_i - y_i)^2/2\tau^2\right\}. \tag{4.19}$$

Consider an example with $m = 10$, $\mu = 0$, $\tau^2 = 1$, $n_i = 20$, for $i = 1, \ldots, m$, and $\nu = \zeta = 1$. Suppose the observations are -1.7, -0.5, -0.8, -1.11, -1,14, 1.3, 1.5, 0.7, 0.9, and 1.11. Curve (i) in Figure 4.1 gives the exact posterior density of $\eta = -\frac{1}{5}(\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5) + \frac{1}{5}(\theta_6 + \theta_7 + \theta_8 + \theta_9 + \theta_{10})$. This was computed by numerically integrating the posterior density of η, given σ^2, with respect to the posterior density of σ^2,

$$\pi_{\boldsymbol{y}}(\sigma^2) \propto (\sigma^2)^{-\frac{1}{2}(\nu+2)} \prod_{i=1}^{m}(n_i^{-1}\tau^2 + \sigma^2)^{-1/2} \exp\left\{-\frac{\nu\zeta}{2\sigma^2} - \sum_{i=1}^{m}\frac{(y_i - \mu)^2}{n_i^{-1}\tau^2 + \sigma^2}\right\}. \tag{4.20}$$

The latter can be obtained from (4.18), by applying Corollary 1.1 and integrating out the θ_i. Observe also that the posterior density of η, given σ^2, is normal with mean μ_η and variance σ_η^2, with

$$\mu_\eta = \sum_{i=1}^{m} \frac{\alpha_i(\sigma^{-2}\mu + n_i\tau^{-2}y_i)}{\sigma^{-2} + n_i\tau^{-2}}, \tag{4.21}$$

and

$$\sigma_\eta^2 = \sum_{i=1}^{m} \frac{\alpha_i^2}{\sigma^{-2} + n_i\tau^{-2}}. \tag{4.22}$$

Curve (ii) gives the Laplacian approximation (4.1) when applied to the joint posterior density of the θ_i and $\gamma = \log\sigma^2$, and when $\boldsymbol{\theta}_\eta$ represents the joint posterior modes of the θ_i and γ, given η. The computation of $\boldsymbol{\theta}_\eta$ involves application of Lagrange's multiplier method and the solution of a cubic equation in σ^2, which can be achieved using Cadran's formulae (see Abramowitz and Stegun, 1965, p. 17). A curve of similar accuracy may be observed when the Laplacian approximation (4.1) is instead applied to the posterior density (4.19) and when $\boldsymbol{\theta}_\eta$ represents the joint posterior modes of the θ_i, given η.

There is clearly room for improvement. Following O'Hagan (1976) and Sun et al (1996), it is beneficial to consider other sets of modes. In particular, the joint posterior modes of the θ_i, given $\eta = \alpha_1\theta_1 + \ldots + \alpha_m\theta_m$, and σ^2, are

$$\tilde{\theta}_i = (\sigma^{-2}\mu + n_i\tau^{-2}y_i - \lambda\alpha_i)/(\sigma^{-2} + n_i\tau^{-2}) \qquad (i = 1, \ldots, m), \tag{4.23}$$

where

$$\lambda = \left\{\sum_{i=1}^{m} \frac{\alpha_i(\sigma^{-2}\mu + n_i\tau^{-2}y_i)}{\sigma^{-2} + n_i\tau^{-2}} - \eta\right\} \Big/ \sum_{i=1}^{m} \frac{\alpha_i^2}{\sigma^{-2} + n_i\tau^{-2}}. \tag{4.24}$$

We base our approximations upon quantities θ_i^* which replace σ^2 in (4.23) by e^{γ^*}, where γ^* is the marginal posterior mode of $\gamma = \log\sigma^2$. The marginal mode of γ should be obtained numerically by maximizing its marginal posterior density. However, while $\boldsymbol{\theta}_\eta^* = (\theta_1^*, \ldots, \theta_m^*)^T$ satisfies $\eta = \boldsymbol{\alpha}^T\boldsymbol{\theta}_\eta^*$, $\boldsymbol{\theta}_\eta^*$ no longer maximizes any particular density. Unfortunately (4.1) does not possess a precise theoretical derivation, if $\boldsymbol{\theta}_\eta$ is replaced by $\boldsymbol{\theta}_\eta^*$.

Curve (iii) of Figure 4.1 gives the approximation (4.3) when applied to (4.19) using the preceding definition of $\boldsymbol{\theta}_\eta^*$, and $\boldsymbol{\ell}_\eta = (\ell_i)$, $\boldsymbol{R}_\eta^* = (r_{ij})$, with

$$\ell_i = n_i\tau^{-2}(y_i - \theta_i) - (m + \nu)(\theta_i - \mu)\left\{\nu\zeta + \sum_{i=1}^{m}(\theta_i - \mu)^2\right\}^{-1}, \qquad (i = 1, \ldots, m) \tag{4.25}$$

and for $i, j = 1, \ldots, m$,

$$r_{ij} = \begin{cases} n_i\tau^{-2} - 2(m+\nu)(\theta_i - \mu)^2\left\{\nu\zeta + \sum_{i=1}^{m}(\theta_i - \mu)^2\right\}^{-2} \\ \quad +(m+\nu)\left\{\nu\zeta + \sum_{i=1}^{m}(\theta_i - \mu)^2\right\}^{-1}, & \text{when} \quad i = j \\ \\ -2(m+\nu)(\theta_i - \mu)(\theta_j - \mu)\left\{\nu\zeta + \sum_{i=1}^{m}(\theta_i - \mu)^2\right\}^{-2}, & \text{when} \quad i \neq j, \end{cases} \tag{4.26}$$

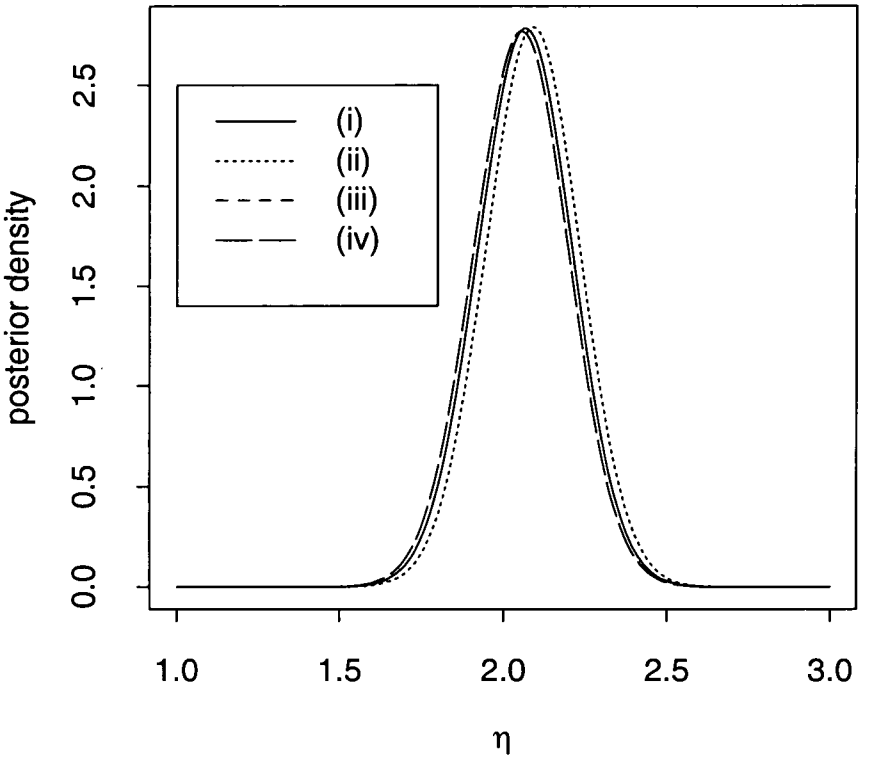


Figure 4.1: First example of Laplacian methods: Posterior density of η. (i) Exact result; (ii) initial Laplacian approximation; (iii) recommended Laplacian approximation; (iv) alternative Laplacian approximation.

obtained by straightforward differentiations of the log of (4.19). It is identical, to visual accuracy, to curve (i). However, considerable care should also be taken when applying approximation (4.3). Curve (iv) of Figure 4.1 gives the same approximation, when applied to the posterior density of $\theta_1, \ldots, \theta_m$ and γ, with the elements of $\boldsymbol{\theta}^*_\eta$ now replaced by the previous elements, together with γ^*. This does not give as good accuracy as our recommended approximation (iii).

4.3.2 A single stage model

Extreme care is also needed when choosing $\boldsymbol{\theta}^*_\eta$, as demonstrated by a further example. Suppose that given $\phi_1, \phi_2, \ldots, \phi_m$, the statistics $U_1, U_2, \ldots, U_m$ are independent, and that, for $i = 1, \ldots, m$, U_i/ϕ_i has chi-squared distribution with n_i degrees of freedom, and the distribution of $\theta_i = \log \phi_i$ in the prior assessment is uniform over $(-\infty, \infty)$. Then the posterior density of the θ_i, given the U_i, is

$$\pi(\boldsymbol{\theta}|\boldsymbol{U}) \propto \exp\left(-\frac{1}{2}\sum_i n_i\theta_i - \frac{1}{2}\sum_i e^{-\theta_i}U_i\right). \tag{4.27}$$

We seek an approximation to the posterior density of the contrast $\eta = \sum_i \alpha_i\theta_i$, where $\sum_i \alpha_i = 0$. The exact conditional posterior modes of the θ_i, given that $\sum_i \alpha_i\theta_i = \eta$, satisfy the equations

$$\theta_i^* = \log\{U_i/(n_i + \alpha_i\lambda)\} \qquad (i = 1, \ldots, m), \tag{4.28}$$

where λ satisfies the non-linear equation

$$\eta = \sum_{i=1}^{m} \alpha_i \log\{U_i/(n_i + \alpha_i\lambda)\}. \tag{4.29}$$

It is possible to invert (4.29) so that λ can be expressed in terms of η. Consider a numerical example with $U_1 = 439.77, U_2 = 121.83, U_3 = 142.94, U_4 = 406.32, U_5 = 2694.74, n_1 = 39, n_2 = 21, n_3 = 19, n_4 = 66,$ and $n_5 = 127$. Curve (i) of Figure 4.2 denotes the exact posterior density of $\eta = -\frac{1}{3}(\theta_1+\theta_2+\theta_3)+\frac{1}{2}(\theta_4+\theta_5)$. It was computed by $N = 10,000$ Monte Carlo simulations. In particular, using that the posterior density of $\eta, \phi_2, \ldots, \phi_m$ is

$$\pi(\eta, \phi_2, \ldots, \phi_m|\boldsymbol{U}) \propto \phi_\eta^{-n_1/2}\exp(-U_1/2\phi_\eta)\prod_{i=2}^{m}\phi_i^{-(n_i/2+1)}\exp(-U_i/2\phi_i), \tag{4.30}$$

with

$$\phi_\eta = \exp\left\{\alpha_1^{-1}\left(\eta - \sum_{i=2}^{m}\alpha_i\log\phi_i\right)\right\}, \tag{4.31}$$

the marginal posterior density of η can be obtained as

$$\pi(\eta|\boldsymbol{U}) = N^{-1}\sum_{k=1}^{N}\phi_{\eta,k}^{-n_1/2}\exp(-U_1/2\phi_{\eta,k}), \tag{4.32}$$

where $\phi_{\eta,k}$ is the ϕ_η of equation (4.31) corresponding to the kth independent realization of ϕ_i, for $i = 2, \ldots, m$, from a scaled inverse chi-squared distribution with n_i degrees of freedom and U_i/n_i scale parameter.

Curve (ii) gives our Laplacian approximation (4.3) based upon the preceding conditional posterior modes and with $\boldsymbol{\ell}_\eta = (\ell_i)$, $\boldsymbol{R}_\eta^* = (r_{ij})$, where

$$\ell_i = -n_i/2 + U_i e^{-\theta_i}/2, \qquad (i = 1, \ldots, m) \tag{4.33}$$

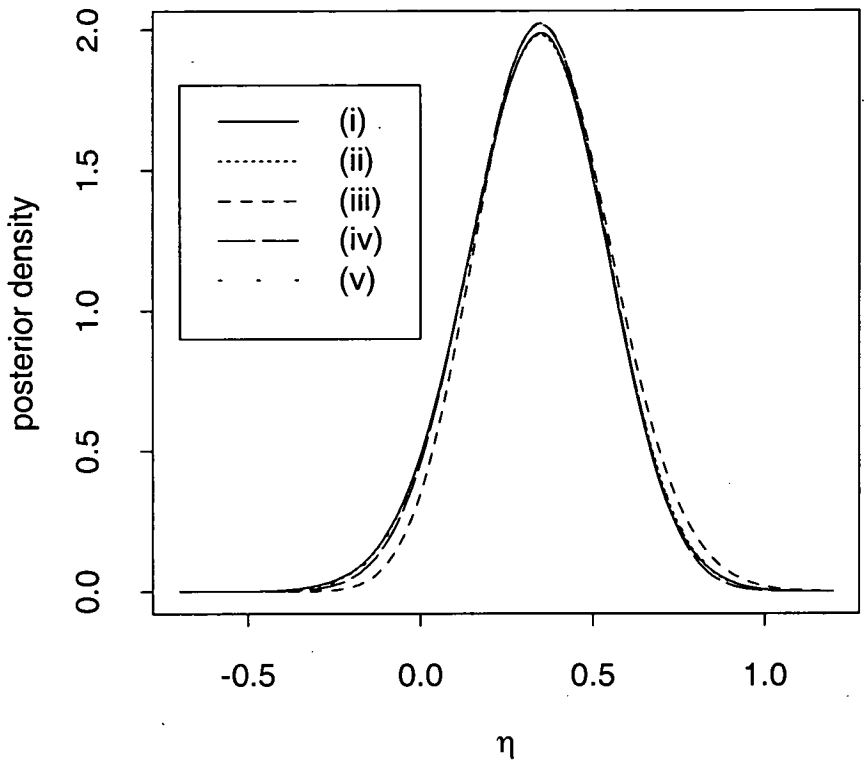


Figure 4.2: Second example of Laplacian methods: Posterior density of η. (i) Exact result; (ii) recommended Laplacian approximation; (iii) alternative Laplacian approximation (linear case); (iv) alternative Laplacian approximation (quadratic case); (v) alternative Laplacian approximation (cubic case).

and for $i, j = 1, \ldots, m$,

$$r_{ij} = \begin{cases} U_i e^{-\theta_i}/2, & \text{when} \quad i = j \\ 0, & \text{when} \quad i \neq j. \end{cases} \tag{4.34}$$

This approximation is effectively identical to the exact curve.

We might however be tempted to approximate λ in terms of η, with the objective of obtaining an algebraically explicit approximation to the posterior density. The terms of the expression

$$\log(n_i + \alpha_i \lambda) = \log n_i + n_i^{-1}\alpha_i \lambda - n_i^{-2}\alpha_i^2 \lambda^2/2 + n_i^{-3}\alpha_i^3 \lambda^3/3, \tag{4.35}$$

provide successive approximations to the right hand side of (4.29), which are linear, quadratic, and cubic in λ. The Laplacian solutions, based upon the linear and quadratic

approximations are given by curves (iii) and (iv) of Figure 4.2, and are less accurate. However (v) gives the Laplacian solution based upon the cubic approximation to (4.29), (choose the root which is closest to the linear solution), and is quite accurate.

Based upon those two examples, we conclude that Laplacian approximations should not be unequivocally applied, but if employed with care, can provide remarkably accurate results. Asymptotic properties may not particularly help when comparing finite sample approximations (Leonard, Hsu and Tsui, 1989) because substantial finite terms can vanish as $n \to \infty$.

4.4 Application of Laplacian methods to ANCOVA models

The aim of this section is the development of Laplacian approximations for the posterior density of the random effects and six parameters of the random variance ANCOVA model studied in Chapter 2. In particular, following the results obtained in the preceding two examples, we will seek to approximate the marginal posterior density of each of the random effects, θ_i, β_i, and ϕ_i, by applying approximation (4.3) on the joint posterior density of the θ_i, β_i, and $\gamma_i = \log \phi_i$, evaluated at the conditional modes of θ_i, β_i, and γ_i, given the six model parameters, for values of these parameters maximizing their joint posterior density. To obtain approximations to the marginal posterior density of each of the six model parameters μ_θ, μ_β, λ_θ, λ_β, ν and ζ, we will employ approximation (4.3) directly on their joint posterior density.

4.4.1 Approximations of random effects and parameter marginal densities

The joint posterior density of θ_i, β_i, and $\gamma_i = \log \phi_i$ can be obtained from (2.21) by integrating out the six model parameters. In particular, having modified the quadratic forms in μ_θ and μ_β of the exponent of (2.21) to obtain similar expressions to (2.36) and (2.37), we can integrate out both μ_θ and μ_β to obtain the joint posterior density of θ_i, β_i, ϕ_i, λ_θ, λ_β, ν, and ζ, which, after some rearrangement, is

$$
\begin{aligned}
&\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi},\lambda_\theta,\lambda_\beta,\nu,\zeta|\boldsymbol{y}) \\
&\propto \pi(\nu)\left\{K(\nu)\right\}^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}\zeta^{\frac{1}{2}\nu(m+\psi\zeta_0)-1}e^{-\frac{1}{2}\nu\zeta(\psi+\sum_{i=1}^m\phi_i^{-1})} \\
&\times (\textstyle\sum_{i=1}^m\phi_i^{-1})^{-1}\prod_{i=1}^m\phi_i^{-\frac{1}{2}(\nu+n_i+4)}\exp(-\tfrac{1}{2}B_1) \\
&\times \lambda_\theta^{-\frac{1}{2}(\omega_1+m+1)}\exp(-\tfrac{1}{2}\lambda_\theta^{-1}B_2)\lambda_\beta^{-\frac{1}{2}(\omega_2+m+1)}\exp(-\tfrac{1}{2}\lambda_\beta^{-1}B_3),
\end{aligned}
\tag{4.36}
$$

with

$$
B_1=\sum_{i=1}^m e^{-\gamma_i}\left\{U_i+n_i\left(y_{i\cdot}-\theta_i\right)^2+s_i^2\left(\hat{\beta}_i-\beta_i\right)^2\right\}, \tag{4.37}
$$

$$
B_2=\omega_1\tau_1+\sum_{i=1}^m e^{-\gamma_i}(\theta_i-\mu_\theta^*)^2, \tag{4.38}
$$

and

$$
B_3=\omega_2\tau_2+\sum_{i=1}^m e^{-\gamma_i}(\beta_i-\mu_\beta^*)^2, \tag{4.39}
$$

with $K(\cdot)$ defined in (2.4), and μ_θ^* and μ_β^* in (2.35) and (2.38), with the obvious transformation for ϕ_i. A further step involves the integration of λ_θ and λ_β which results in the following expression for the posterior density of θ_i, β_i, ϕ_i, ν, and ζ:

$$
\begin{aligned}
&\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi},\nu,\zeta|\boldsymbol{y}) \\
&\propto \pi(\nu)\left\{K(\nu)\right\}^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}\zeta^{\frac{1}{2}\nu(m+\psi\zeta_0)-1}e^{-\frac{1}{2}\nu\zeta(\psi+\sum_{i=1}^m\phi_i^{-1})} \\
&\times (\textstyle\sum_{i=1}^m\phi_i^{-1})^{-1}\prod_{i=1}^m\phi_i^{-\frac{1}{2}(\nu+n_i+4)}\exp(-\tfrac{1}{2}B_1)B_2^{-\frac{1}{2}(\omega_1+m-1)}B_3^{-\frac{1}{2}(\omega_2+m-1)}.
\end{aligned}
\tag{4.40}
$$

Subsequent integration of ζ, provides the posterior density of the model $3m$ random effects and ν as

$$
\begin{aligned}
&\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\phi},\nu|\boldsymbol{y}) \\
&\propto \pi(\nu)\left\{K(\nu)\right\}^m K(\nu\psi\zeta_0)\zeta_0^{-\frac{1}{2}\nu\psi\zeta_0}(\textstyle\sum_{i=1}^m\phi_i^{-1})^{-1}\prod_{i=1}^m\phi_i^{-\frac{1}{2}(\nu+n_i+4)}\exp(-\tfrac{1}{2}B_1) \\
&\times B_2^{-\frac{1}{2}(\omega_1+m-1)}B_3^{-\frac{1}{2}(\omega_2+m-1)}\Gamma\left(\tfrac{1}{2}\nu m+\tfrac{1}{2}\nu\psi\zeta_0\right)\left(\tfrac{1}{2}\nu\psi+\tfrac{1}{2}\nu\textstyle\sum_{i=1}^m\phi_i^{-1}\right)^{-\frac{1}{2}\nu(m+\psi\zeta_0)}.
\end{aligned}
\tag{4.41}
$$

Employing Stirling's approximation for the Gamma function, (4.41) as a function of ν is proportional to

$$
\nu^{\frac{m}{2}+\frac{a}{2}-1}\exp(-\frac{\nu}{2}B_4), \tag{4.42}
$$

where

$$
B_4=\sum_{i=1}^m\gamma_i+\psi\zeta_0\log\zeta_0+(m+\psi\zeta_0)\log\left(\frac{\psi+\sum_{i=1}^m e^{-\gamma_i}}{\psi\zeta_0+m}\right)+b. \tag{4.43}
$$

Hence integrating out ν from (4.41) and transforming to $\gamma_i = \log \phi_i$, provide the posterior density of the θ_i, β_i, and γ_i, unconditionally on the six model parameters

$$\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}) \propto$$
$$\prod_{i=1}^{m} e^{-\frac{1}{2}(n_i+2)\gamma_i} \left(\sum_{i=1}^{m} e^{-\gamma_i}\right)^{-1} \exp\left(-\tfrac{1}{2}B_1\right) B_2^{-\frac{1}{2}(\omega_1+m-1)} B_3^{-\frac{1}{2}(\omega_2+m-1)} B_4^{-\frac{1}{2}(a+m)}. \tag{4.44}$$

The preceding posterior density forms the basis for the application of approximation (4.3) to obtain the marginal posterior density of each of the random effects in the model. The approximation also requires the computation of the vector $\boldsymbol{\ell}_\eta$ and the matrix $\boldsymbol{R}^*_\eta$, for which explicit expressions are presented in Appendix 4.5.1. They both need to be evaluated at the values of θ_i, β_i and γ_i described next.

The conditional density of the θ_i, β_i, and ϕ_i, given the $y_{i\cdot}$ and the six model parameters, is proportional to (2.21). Consider for example that the parameter of interest is $\eta = \theta_j$, for a single j in $1, \ldots, m$. Maximizing (2.21), with respect to θ_i, β_i, and ϕ_i, conditionally on the values of the six model parameters and $\theta_j = \eta$, gives the maximizing values $\theta_i = \theta_i^*$ for $i = 1, \ldots, m$, $i \neq j$, and $\beta_i = \beta_i^*$ for $i = 1, \ldots, m$, where θ_i^* and β_i^* satisfy (2.12) and (2.13). Similarly for $i = 1, \ldots, m$,

$$\phi_i = \begin{cases} (\nu + n_i + 2)^{-1} \left\{\nu\zeta + U_i + (n_i + \lambda_\theta^{-1})(\eta - \theta_i^*)^2 + \Omega_i\right\}, & \text{when} \quad i = j \\ (\nu + n_i + 2)^{-1} \left(\nu\zeta + U_i + \Omega_i\right), & \text{when} \quad i \neq j, \end{cases} \tag{4.45}$$

where

$$\Omega_i = (n_i^{-1} + \lambda_\theta)^{-1}(y_{i\cdot} - \mu_\theta)^2 + (s_i^{-2} + \lambda_\beta)^{-1}(\hat{\beta}_i - \mu_\beta)^2 \qquad (i = 1, \ldots, m). \tag{4.46}$$

If the parameter of interest is $\eta = \beta_j$, the corresponding values are $\theta_i = \theta_i^*$ for $i = 1, \ldots, m$, $\beta_i = \beta_i^*$ for $i = 1, \ldots, m$, $i \neq j$,

$$\phi_j = (\nu + n_j + 2)^{-1} \left\{\nu\zeta + U_j + (s_j^2 + \lambda_\beta^{-1})(\eta - \beta_j^*)^2 + \Omega_j\right\}, \tag{4.47}$$

and ϕ_i same as in (4.45) for $i \neq j$. When we wish to approximate the marginal posterior density of ϕ_j, the maximizing values are $\theta_i = \theta_i^*$ and $\beta_i = \beta_i^*$ for $i = 1, \ldots, m$, and ϕ_i as appearing in (4.45) for $i \neq j$.

Furthermore, the joint posterior density of the six model parameters, μ_θ, μ_β, $\epsilon_\theta = \log \lambda_\theta$, $\epsilon_\beta = \log \lambda_\beta$, $\epsilon_\nu = \log \nu$ and $\epsilon_\zeta = \log \zeta$ is

$$\pi(\mu_\theta, \mu_\beta, \epsilon_\theta, \epsilon_\beta, \epsilon_\nu, \epsilon_\zeta | \boldsymbol{y})$$
$$\propto \exp\left[-\tfrac{1}{2}\omega_1\epsilon_\theta - \tfrac{1}{2}\omega_1\tau_1 e^{-\epsilon_\theta} - \tfrac{1}{2}\omega_2\epsilon_\beta - \tfrac{1}{2}\omega_2\tau_2 e^{-\epsilon_\beta} + \tfrac{1}{2}\psi\zeta_0 e^{\epsilon_\nu}\{\epsilon_\nu + \log(\psi/2) + \epsilon_\zeta\}\right]$$
$$\times \exp\left(-\tfrac{1}{2}\psi e^{\epsilon_\nu} e^{\epsilon_\zeta} + \tfrac{1}{2}a\epsilon_\nu - \tfrac{1}{2}be^{\epsilon_\nu}\right)\left\{\Gamma(\tfrac{1}{2}e^{\epsilon_\nu}\psi\zeta_0)\right\}^{-1} p(\boldsymbol{y}|\mu_\theta, \mu_\beta, \epsilon_\theta, \epsilon_\beta, \epsilon_\nu, \epsilon_\zeta), \tag{4.48}$$

where the last contribution to the right hand side may be obtained, by the obvious substitution, from (2.16). When any single random effect is the parameter of interest, we employ the values $\widetilde{\mu}_\theta$, $\widetilde{\mu}_\beta$, $\widetilde{\epsilon}_\theta$, $\widetilde{\epsilon}_\beta$, $\widetilde{\epsilon}_\nu$ and $\widetilde{\epsilon}_\zeta$ unconditionally maximizing (4.48), and substituting these for the parameter values in the conditional modes described in the previous paragraph.

O'Hagan (1976), concluded that marginal modes provide better approximations to posterior means than joint modes with nuisance parameters, when using a hierarchical linear multiple regression model and wanting to estimate the variances with the slopes as nuisance parameters. Kass and Steffey (1989), obtain very accurate asymptotically expressions for the mean and variance of functions of first stage parameters for conditionally independent hierarchical models, using marginal posterior modes for the second stage parameters, and prove the insensitivity of these results to the choice of prior distribution.

Hence, this procedure is preferred to joint modes of the θ_i, β_i, and ϕ_i together with the model parameters, because of our results in the first example of section 4.3, and the suggestions by O'Hagan and Kass and Steffey. The maximization just described was done by a modification of Powell's direction set method (Press et al, 1994, pp. 412-420).

We approximate the marginal posterior of θ_j by application of the Laplacian approximation (4.3) to the joint posterior density of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (4.44) (i.e. the $\boldsymbol{\theta}$ vector in (4.3) represents the three vectors $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (4.44)) . The conditional mode vector $\boldsymbol{\theta}^*_\eta$ appearing in (4.3) should be replaced by the values for the elements of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, defined in the third paragraph of the current section. Similar procedures are available for approximating the marginal posterior density of any particular θ_i, β_i or γ_i. Laplacian approximations for the densities of linear combinations of these parameters, involve slightly more complex algebra. It is important to refer our Laplacian approximation to (4.44), rather than the joint posterior density of the θ_i, β_i, γ_i, and six model parameters to avoid the problems introduced by the first example of section 4.3. The latter procedure can again lead to numerical inaccuracy.

In the numerical example of section 3.1.1, our Laplacian approximations (dotted

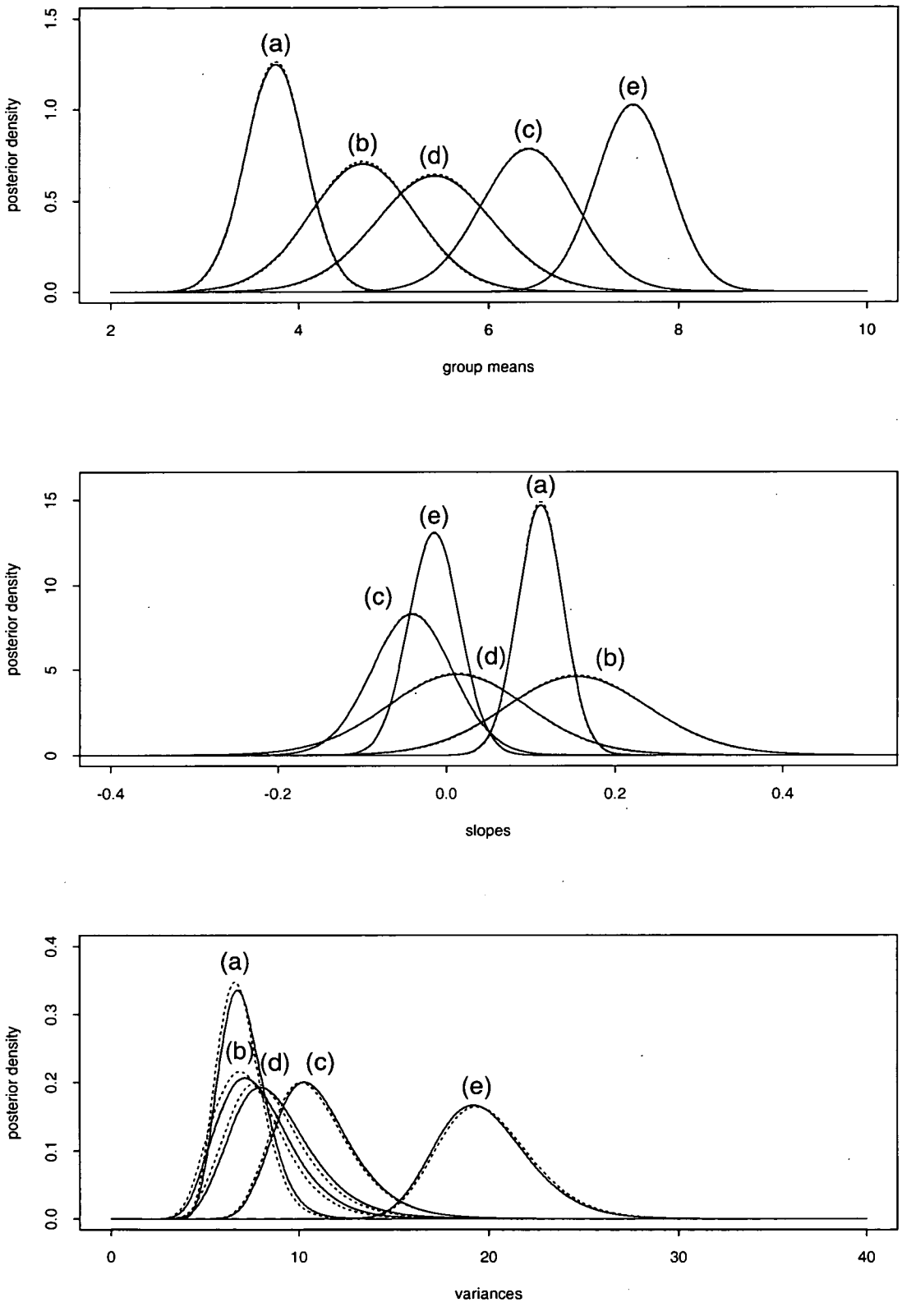


Figure 4.3: Visual functions test. Posterior density of five, (a)-(e), group means (θ_i), group slopes (β_i), and group variances (ϕ_i), by MCMC (solid line) and Laplacian approximation (dotted line).

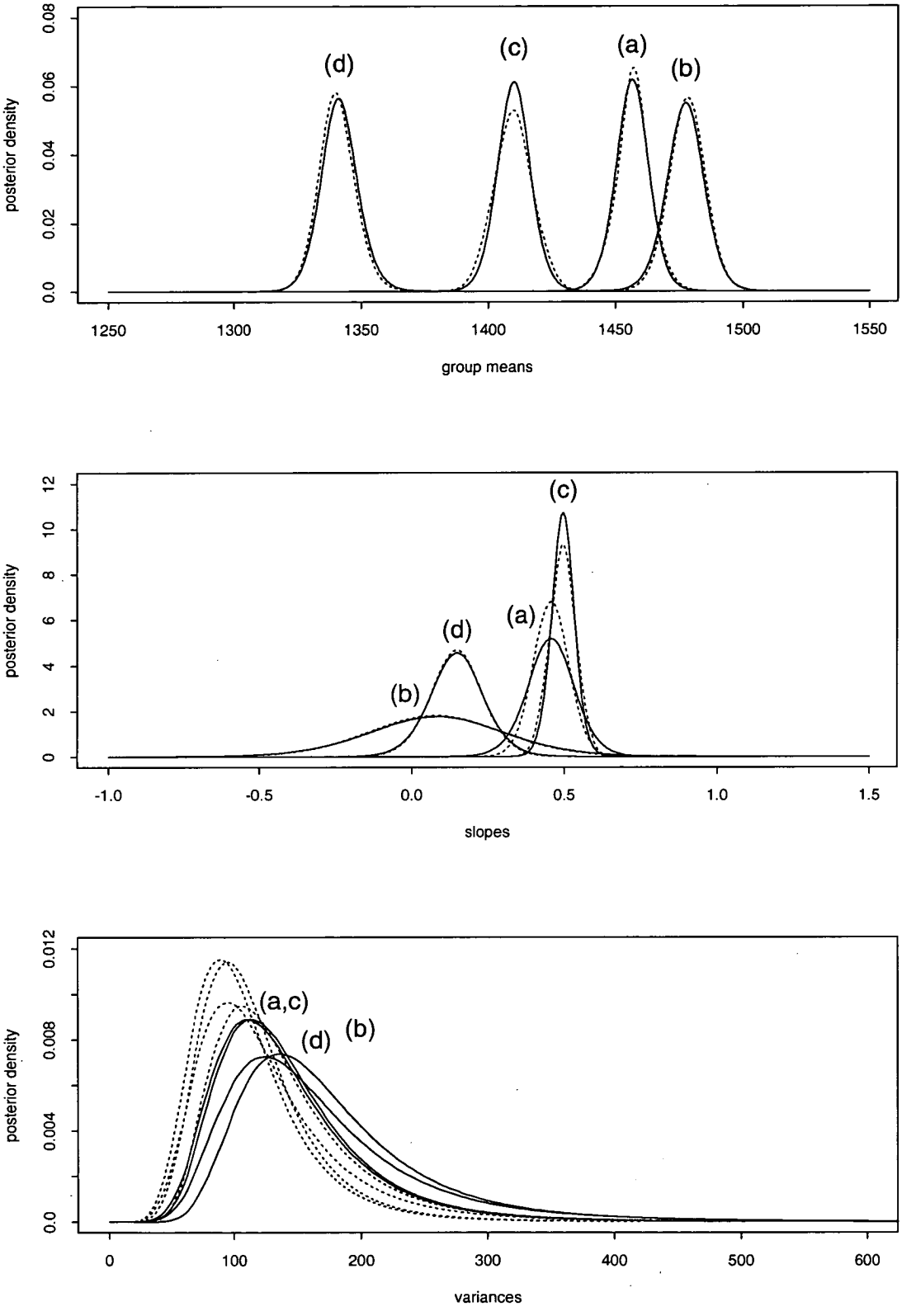


Figure 4.4: Nutrition data. Posterior density of five, (a)-(e), group means (θ_i), group slopes (β_i), and group variances (ϕ_i), by MCMC (solid line) and Laplacian approximation (dotted line).

curves) to the posterior densities of the θ_i and β_i are extremely close to the exact curves (solid curves), calculated by MCMC, as represented in the top two plots of Figure 4.3, and seem to have validated the convergence of MCMC. Our approximations to the posterior densities of the ϕ_i are described by the dotted curves of the bottom plot of Figure 4.3. These are again reassuringly close to the exact densities, represented by the solid curves. All posterior densities correspond to choice (2) of prior distribution, described in section 3.1.1, with similar results obtained for choices (1) and (3). All Laplacian approximations for the θ_i, β_i, and ϕ_i were calculated in 11.8 seconds of CPU time on a Sun Ultra 1 workstation. For the exact procedure, 10,000 MCMC iterations, after 10,000 iterations on burn-in (4.6 seconds) took 238 seconds. This is typically enough for virtually exact calculations. The closeness of the approximate results to the exact ones together with the computational time saving justifies the use of Laplacian approximations to assess MCMC convergence and improve time efficiency, when computing the posterior densities of the random effects in the model.

In Figure 4.4 we present the corresponding results for the nutrition data example of section 3.2. The agreement of the MCMC and the Laplacian approximation results is not as close as before, not surprisingly though, given the extremely small sample sizes. The Laplacian approximation confirms the reversal of the ordering of the group variances, as we already discussed in Chapter 3.

The first of the preceding benefits, i.e. validation of the MCMC convergence, is also achieved when computing the posterior densities of the six model parameters. The marginal posterior density of any of the six model parameters, may be obtained by direct application of (4.3) to the posterior density (4.48). The corresponding $\boldsymbol{\ell}_\eta$ vector and $\boldsymbol{R}^*_\eta$ matrix are described in Appendix 4.5.2. The Laplacian approximations to the posterior densities of μ_θ, μ_β, λ_θ, λ_β, ν and ζ, under the three different choices of prior distribution, are described by the dotted curves in Figure 4.5. They are identical to visual accuracy to the exact curves (solid lines), computed by MCMC, despite the high skewness of some of the posterior densities. However, because of the high number of numerical maximizations needed, one maximization of a five variable function for each point of each marginal posterior density, this procedure is not much faster computationally than MCMC, nevertheless it is valuable for assessing its convergence. For obtaining marginal posterior densities of the model parameters, our approximation is identical to (4.3).

We also need to point out that if uniform distributions are instead assumed for λ_θ, λ_β, ν, and ζ in the prior assessment, then the Laplacian approximations for the

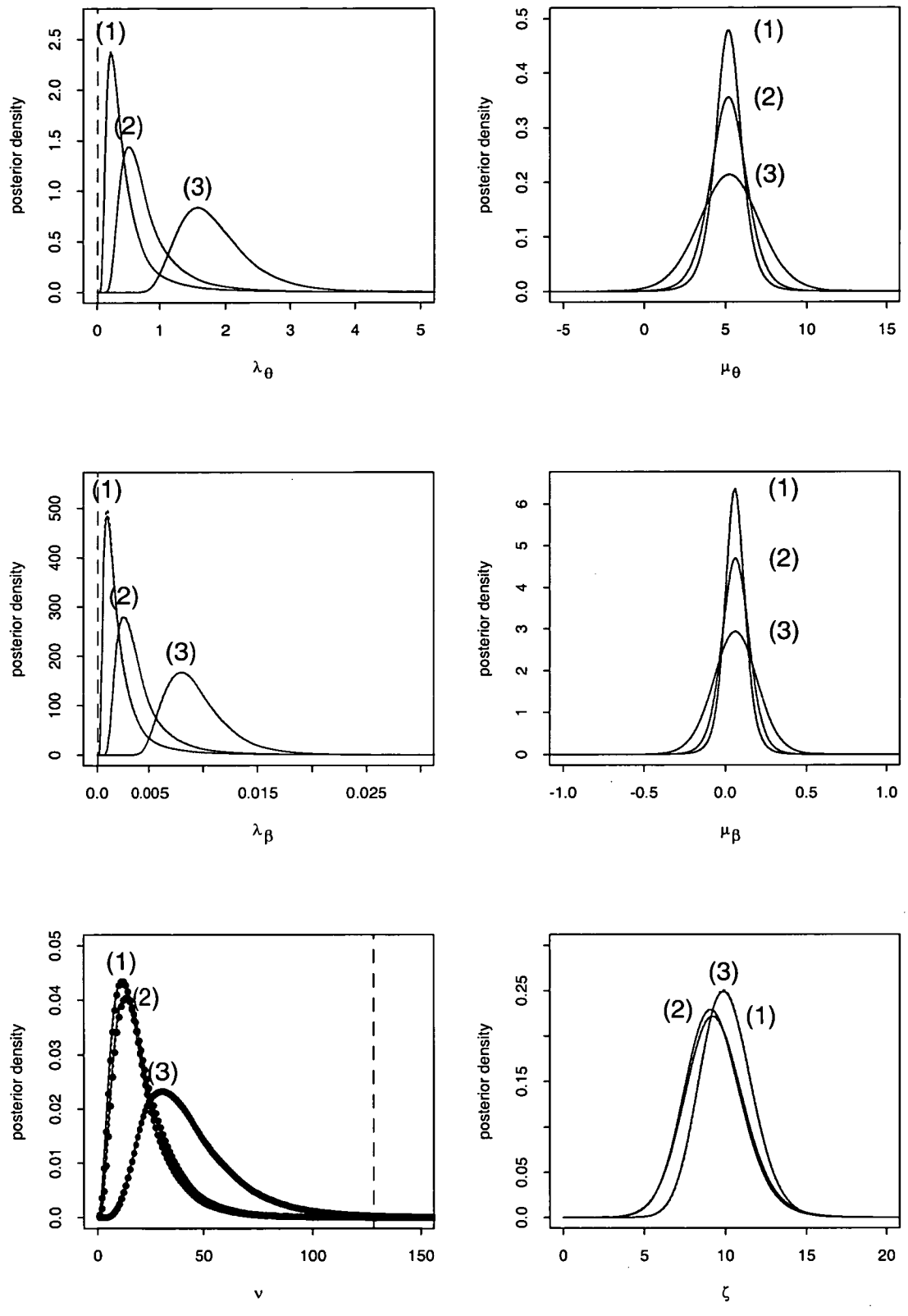


Figure 4.5: Visual functions test. Posterior density of λ_θ, λ_β, ν, μ_θ, μ_β, and ζ, by MCMC (solid line) and Laplacian approximation (dotted line), under three choices, (1)-(3), of prior distribution.

random effects can work extremely well. If the quadratic term in the exponent of (4.3) is omitted and instead (4.3) is used, with the conditional modes of the random effects, then this method may lead to substantial inaccuracies (see Figure 4.6), certifying that this quadratic term can potentially play a dominant part in the accuracy of the results. However, this modification, which was suggested by Sun (1992), provided very accurate results when using proper priors for our ANCOVA models.

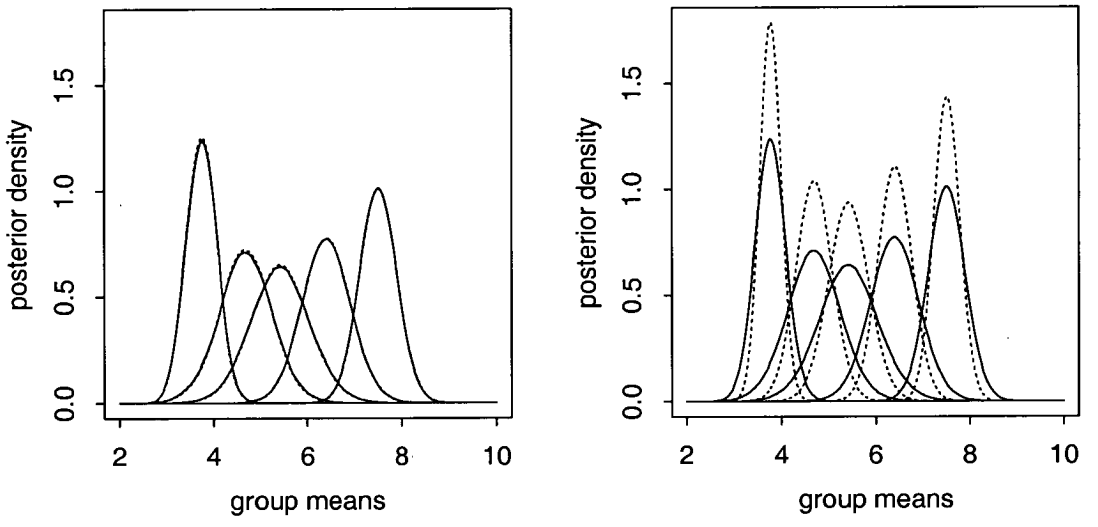


Figure 4.6: Visual functions test. Posterior density of five, (a)-(e), group means (θ_i), under prior distribution (2). Left: Using approximation (4.3). Right: Using approximation (4.1).

Concluding, we should mention that although applying Laplacian procedures is not as simple as applying MCMC, they can provide substantial computational efficiency and numerical accuracy. In some situations they can also provide algebraically closed expressions for marginal posterior densities, which is another characteristic that makes them appealing. Clearly, however, their application involves pitfalls, for example trying to apply an approximation on the full posterior distribution of all the random effects and parameters in our case would lead to very inaccurate results. Hence they potentially require more mathematical derivations than MCMC. Nevertheless, in the ANCOVA models we studied, the availability of both methods very effectively complemented each other's shortcomings.

4.5 Appendix: Derivatives of log-posterior density

4.5.1 Posterior density of random effects

The joint posterior density of the $3m$ random effects, $\gamma_i = \log \phi_i$, θ_i, β_i, for $i = 1, \ldots m$, is given by expression (4.44). Let $\boldsymbol{\theta} = (\gamma_1, \ldots, \gamma_m, \theta_1, \ldots, \theta_m, \beta_1, \ldots, \beta_m)^T$. The first derivative of the log of the posterior density of the the $3m$ random effects, with respect to $\boldsymbol{\theta}$ is $\boldsymbol{\ell}_1 = (\ell_1(1), \ldots, \ell_1(3m))^T$, where

$$\begin{aligned}
\ell_1(j) &= -\frac{1}{2}n_j - 1 + e^{-\gamma_j}B_7^{-1} + \frac{1}{2}e^{-\gamma_j}(U_j + W_{j1}) \\
&+ \frac{1}{2}(m + \omega_1 - 1)e^{-\gamma_j}(\theta_j - \mu_\theta^*)B_2^{-1}\left(\theta_j - \mu_\theta^* - 2B_5B_7^{-1}\right) \\
&+ \frac{1}{2}(m + \omega_2 - 1)e^{-\gamma_j}(\beta_j - \mu_\beta^*)B_3^{-1}\left(\beta_j - \mu_\beta^* - 2B_6B_7^{-1}\right) \\
&- \frac{1}{2}(m + a)B_4^{-1}(1 - e^{-\gamma_j}B_8) \qquad (4.49) \\
\ell_1(m + j) &= n_j e^{-\gamma_j}(y_{j\cdot} - \theta_j) - (m + \omega_1 - 1)e^{-\gamma_j}B_2^{-1}\left(\theta_j - \mu_\theta^* - B_5B_7^{-1}\right) \qquad (4.50) \\
\ell_1(2m + j) &= s_j^2 e^{-\gamma_j}(\hat{\beta}_j - \beta_j) - (m + \omega_2 - 1)e^{-\gamma_j}B_3^{-1}\left(\beta_j - \mu_\beta^* - B_6B_7^{-1}\right) \qquad (4.51)
\end{aligned}$$

for $j = 1, \ldots m$, with

$$\begin{aligned}
B_5 &= \sum_{i=1}^{m} e^{-\gamma_i}(\theta_i - \mu_\theta^*), \qquad (4.52) \\
B_6 &= \sum_{i=1}^{m} e^{-\gamma_i}(\beta_i - \mu_\beta^*), \qquad (4.53) \\
B_7 &= \sum_{i=1}^{m} e^{-\gamma_i}, \qquad (4.54) \\
B_8 &= (\psi + B_7)^{-1}(m + \psi\zeta_0), \qquad (4.55)
\end{aligned}$$

and B_2, B_3, B_4, W_{j1}, μ_θ^*, μ_β^*, U_j, $\hat{\beta}_j$, s_j^2 defined in (4.38), (4.39), (4.43), (2.32), (2.35), (2.38), (2.6), (2.7), and (2.8) respectively.

The negative of the Hessian matrix of the log of the posterior density of the $3m$ random effects with respect to the vector $\boldsymbol{\theta}$ is the symmetric matrix $\boldsymbol{R}_1 = (-r_1(i,j))$, with

$$r_1(i,j) = \partial^2 \log \pi(\gamma_1, ..., \gamma_m, \theta_1, ..., \theta_m, \beta_1, ..., \beta_m | \boldsymbol{y}) / \partial\boldsymbol{\theta}(i)\partial\boldsymbol{\theta}(j) \quad (i,j = 1, \ldots, 3m), \qquad (4.56)$$

where the elements of its upper triangular part are

$$
\begin{aligned}
r_1(j,j) &= -e^{-\gamma_j}B_7^{-1} + e^{-2\gamma_j}B_7^{-2} - \frac{1}{2}e^{-\gamma_j}(U_j + W_{j1}) \\
&+ (m+\omega_1-1)e^{-\gamma_j}(\theta_j-\mu_\theta^*)B_2^{-1}B_5B_7^{-1}\left(1-2e^{-\gamma_j}B_7^{-1}\right) \\
&+ \frac{1}{2}(m+\omega_1-1)e^{-\gamma_j}(\theta_j-\mu_\theta^*)^2B_2^{-1}\left(2e^{-\gamma_j}B_7^{-1}-1\right) \\
&+ \frac{1}{2}(m+\omega_1-1)B_2^{-2}\left\{2e^{-\gamma_j}(\theta_j-\mu_\theta^*)B_5B_7^{-1} - e^{-\gamma_j}(\theta_j-\mu_\theta^*)^2\right\}^2 \\
&+ (m+\omega_2-1)e^{-\gamma_j}(\beta_j-\mu_\beta^*)B_3^{-1}B_6B_7^{-1}\left(1-2e^{-\gamma_j}B_7^{-1}\right) \\
&+ \frac{1}{2}(m+\omega_2-1)e^{-\gamma_j}(\beta_j-\mu_\beta^*)^2B_3^{-1}\left(2e^{-\gamma_j}B_7^{-1}-1\right) \\
&+ \frac{1}{2}(m+\omega_2-1)B_3^{-2}\left\{2e^{-\gamma_j}(\beta_j-\mu_\beta^*)B_6B_7^{-1} - e^{-\gamma_j}(\beta_j-\mu_\beta^*)^2\right\}^2 \\
&+ \frac{1}{2}(m+a)B_4^{-1}\left\{e^{-\gamma_j}B_8 - (m+\psi\zeta_0)^{-1}e^{-2\gamma_j}B_8^2\right\} \\
&+ \frac{1}{2}(m+a)B_4^{-2}\left(1-e^{-\gamma_j}B_8\right)^2 \qquad (4.57) \\
r_1(j,k) &= -e^{-\gamma_j-\gamma_k}B_7^{-2} + \frac{1}{2}(m+a)B_4^{-2}\left(1-e^{-\gamma_j}B_8\right)\left(1-e^{-\gamma_k}B_8\right) \\
&- (m+\omega_1-1)e^{-\gamma_j-\gamma_k}\left(\theta_j+\theta_k-2\mu_\theta^*\right)B_2^{-1}B_5B_7^{-2} \\
&+ (m+\omega_1-1)e^{-\gamma_j-\gamma_k}(\theta_j-\mu_\theta^*)(\theta_k-\mu_\theta^*)B_2^{-1}B_7^{-1} \\
&+ \frac{1}{2}e^{-\gamma_j-\gamma_k}(\theta_j-\mu_\theta^*)(\theta_k-\mu_\theta^*)B_2^{-2} \\
&\times \left(\theta_j-\mu_\theta^*-2B_5B_7^{-1}\right)\left(\theta_k-\mu_\theta^*-2B_5B_7^{-1}\right) \\
&- (m+\omega_2-1)e^{-\gamma_j-\gamma_k}\left(\beta_j+\beta_k-2\mu_\beta^*\right)B_3^{-1}B_6B_7^{-2} \\
&+ (m+\omega_2-1)e^{-\gamma_j-\gamma_k}(\beta_j-\mu_\beta^*)(\beta_k-\mu_\beta^*)B_3^{-1}B_7^{-1} \\
&+ \frac{1}{2}e^{-\gamma_j-\gamma_k}(\beta_j-\mu_\beta^*)(\beta_k-\mu_\beta^*)B_3^{-2} \\
&\times \left(\beta_j-\mu_\beta^*-2B_6B_7^{-1}\right)\left(\beta_k-\mu_\beta^*-2B_6B_7^{-1}\right) \\
&+ \frac{1}{2}(m+a)(m+\psi\zeta_0)e^{-\gamma_j-\gamma_k}B_4^{-1}B_7^{-2} \qquad (4.58) \\
r_1(j,m+j) &= -n_je^{-\gamma_j}(y_{j\cdot}-\theta_j) \\
&- (m+\omega_1-1)(1-e^{-\gamma_j}B_7^{-1})e^{-\gamma_j}B_2^{-1}\left(B_5B_7^{-1}-\theta_j+\mu_\theta^*\right) \\
&+ (m+\omega_1-1)e^{-2\gamma_j}(\theta_j-\mu_\theta^*)B_2^{-2} \\
&\times \left(-\theta_j+\mu_\theta^*+2B_5B_7^{-1}\right)\left(\theta_j-\mu_\theta^*-B_5B_7^{-1}\right) \qquad (4.59) \\
r_1(j,m+k) &= (m+\omega_1-1)e^{-\gamma_j-\gamma_k}B_2^{-1}B_7^{-1}\left(B_5B_7^{-1}-\theta_j+\mu_\theta^*\right) \\
&+ (m+\omega_1-1)e^{-\gamma_j-\gamma_k}(\theta_j-\mu_\theta^*)B_2^{-2} \\
&\times \left(-\theta_j+\mu_\theta^*+2B_5B_7^{-1}\right)\left(\theta_k-\mu_\theta^*-B_5B_7^{-1}\right) \qquad (4.60)
\end{aligned}
$$

$$
\begin{aligned}
r_1(j,2m+j) &= -s_j^2 e^{-\gamma_j}(\hat{\beta}_j - \beta_j) \\
&- (m+\omega_2-1)(1-e^{-\gamma_j}B_7^{-1})e^{-\gamma_j}B_3^{-1}\left(B_6B_7^{-1} - \beta_j + \mu_\beta^*\right) \\
&+ (m+\omega_2-1)e^{-2\gamma_j}(\beta_j - \mu_\beta^*)B_3^{-2} \\
&\times \left(-\beta_j + \mu_\beta^* + 2B_6B_7^{-1}\right)\left(\beta_j - \mu_\beta^* - B_6B_7^{-1}\right) \qquad (4.61) \\
r_1(j,2m+k) &= (m+\omega_2-1)e^{-\gamma_j-\gamma_k}B_3^{-1}B_7^{-1}\left(B_6B_7^{-1} - \beta_j + \mu_\beta^*\right) \\
&+ (m+\omega_2-1)e^{-\gamma_j-\gamma_k}(\beta_j - \mu_\beta^*)B_3^{-2} \\
&\times \left(-\beta_j + \mu_\beta^* + 2B_6B_7^{-1}\right)\left(\beta_k - \mu_\beta^* - B_6B_7^{-1}\right) \qquad (4.62) \\
r_1(m+j,m+j) &= -n_j e^{-\gamma_j} - (m+\omega_1-1)e^{-\gamma_j}B_2^{-1}(1-e^{-\gamma_j}B_7^{-1}) \\
&+ 2(m+\omega_1-1)e^{-2\gamma_j}B_2^{-2}(\theta_j - \mu_\theta^* - B_5B_7^{-1})^2 \qquad (4.63) \\
r_1(m+j,m+k) &= (m+\omega_1-1)e^{-\gamma_j-\gamma_k}B_2^{-1}B_7^{-1} + 2(m+\omega_1-1)e^{-\gamma_j-\gamma_k}B_2^{-2} \\
&\times (\theta_j - \mu_\theta^* - B_5B_7^{-1})(\theta_k - \mu_\theta^* - B_5B_7^{-1}) \qquad (4.64) \\
r_1(m+j,2m+j) &= 0 \qquad (4.65) \\
r_1(m+j,2m+k) &= 0 \qquad (4.66) \\
r_1(2m+j,2m+j) &= -s_j^2 e^{-\gamma_j} - (m+\omega_2-1)e^{-\gamma_j}B_3^{-1}(1-e^{-\gamma_j}B_7^{-1}) \\
&+ 2(m+\omega_2-1)e^{-2\gamma_j}B_3^{-2}(\beta_j - \mu_\beta^* - B_6B_7^{-1})^2 \qquad (4.67) \\
r_1(2m+j,2m+k) &= (m+\omega_2-1)e^{-\gamma_j-\gamma_k}B_3^{-1}B_7^{-1} + 2(m+\omega_2-1)e^{-\gamma_j-\gamma_k}B_3^{-2} \\
&\times (\beta_j - \mu_\beta^* - B_6B_7^{-1})(\beta_k - \mu_\beta^* - B_6B_7^{-1}) \qquad (4.68)
\end{aligned}
$$

for $j = 1, \ldots, m$, and $k = j+1, \ldots, m-1$.

4.5.2 Posterior density of six model parameters

The joint posterior density of the six parameters μ_θ, μ_β, $\epsilon_\theta = \log\lambda_\theta$, $\epsilon_\beta = \log\lambda_\beta$, $\epsilon_\zeta = \log\zeta$ and $\epsilon_\nu = \log\nu$ is given by expression (4.48). Let $\boldsymbol{\theta} = (\mu_\theta, \mu_\beta, \epsilon_\theta, \epsilon_\beta, \epsilon_\zeta, \epsilon_\nu)^T$. The first derivative of the log of the posterior density of the six parameters, with respect to $\boldsymbol{\theta}$ is $\boldsymbol{\ell}_2 = (\ell_2(1), \ldots, \ell_2(6))^T$, with

$$
\ell_2(i) = \partial \log \pi(\mu_\theta, \mu_\beta, \epsilon_\theta, \epsilon_\beta, \epsilon_\zeta, \epsilon_\nu | \boldsymbol{y}) / \partial\theta_i \qquad (i = 1, \ldots, 6), \qquad (4.69)
$$

and

$$
\ell_2(1) = \sum_{i=1}^{m} \frac{(e^{\epsilon_\nu} + n_i)(y_{i\cdot} - \mu_\theta)}{\Xi_i(n_i^{-1} + e^{\epsilon_\theta})}, \qquad (4.70)
$$

$$\ell_2(2) = \sum_{i=1}^{m} \frac{(e^{\epsilon_\nu} + n_i)(\hat{\beta}_i - \mu_\beta)}{\Xi_i(s_i^{-2} + e^{\epsilon_\beta})}, \tag{4.71}$$

$$\ell_2(3) = \frac{1}{2}\sum_{i=1}^{m} \frac{e^{-\epsilon_\theta}}{n_i + e^{-\epsilon_\theta}} + \frac{1}{2}e^{\epsilon_\theta}\sum_{i=1}^{m} \frac{(e^{\epsilon_\nu} + n_i)(y_{i\cdot} - \mu_\theta)^2}{\Xi_i(n_i^{-1} + e^{\epsilon_\theta})^2} - \frac{1}{2}\omega_1 - \frac{1}{2}m$$
$$+ \frac{1}{2}\omega_1\tau_1 e^{-\epsilon_\theta}, \tag{4.72}$$

$$\ell_2(4) = \frac{1}{2}\sum_{i=1}^{m} \frac{e^{-\log\lambda_\beta}}{s_i^2 + e^{-\epsilon_\beta}} + \frac{1}{2}e^{\epsilon_\beta}\sum_{i=1}^{m} \frac{(e^{\epsilon_\nu} + n_i)(\hat{\beta}_i - \mu_\beta)^2}{\Xi_i(s_i^{-2} + e^{\epsilon_\beta})^2} - \frac{1}{2}\omega_2 - \frac{1}{2}m$$
$$+ \frac{1}{2}\omega_2\tau_2 e^{-\epsilon_\beta}, \tag{4.73}$$

$$\ell_2(5) = -\frac{1}{2}e^{\epsilon_\zeta + \epsilon_\nu}\sum_{i=1}^{m} \frac{(e^{\epsilon_\nu} + n_i)}{\Xi_i} + \frac{1}{2}me^{\epsilon_\nu} + \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu} - \frac{1}{2}\psi e^{\epsilon_\zeta + \epsilon_\nu}, \tag{4.74}$$

$$\ell_2(6) = -\frac{1}{2}e^{\epsilon_\nu}\sum_{i=1}^{m}\log(\Xi_i/2) - \frac{1}{2}e^{\epsilon_\zeta+\epsilon_\nu}\sum_{i=1}^{m}\frac{(e^{\epsilon_\nu} + n_i)}{\Xi_i} + \frac{1}{2}me^{\epsilon_\nu}\log(e^{\epsilon_\zeta+\epsilon_\nu}/2)$$
$$+ \frac{1}{2}me^{\epsilon_\nu} + \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu}\log(\psi e^{\epsilon_\zeta+\epsilon_\nu}/2) + \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu} + \frac{1}{2}a - \frac{1}{2}be^{\epsilon_\nu} - \frac{1}{2}\psi e^{\epsilon_\zeta+\epsilon_\nu}$$
$$- \frac{1}{2}me^{\epsilon_\nu}\Omega(e^{\epsilon_\nu}/2) - \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu}\Omega(\psi\zeta_0 e^{\epsilon_\nu}/2) + \frac{1}{2}e^{\epsilon_\nu}\sum_{i=1}^{m}\Omega\left(\frac{e^{\epsilon_\nu} + n_i}{2}\right), \tag{4.75}$$

where

$$\Omega(z) = d\log\Gamma(z)/dz, \tag{4.76}$$

(see Abramowitz and Stegun, 1965, pp. 258-260, for computational details),

$$\Xi_i = U_i + \frac{(y_{i\cdot} - \mu_\theta)^2}{n_i^{-1} + e^{\epsilon_\theta}} + \frac{(\hat{\beta}_i - \mu_\beta)^2}{s_i^{-2} + e^{\epsilon_\beta}} + e^{\epsilon_\zeta + \epsilon_\nu}. \tag{4.77}$$

and U_i, $\hat{\beta}_i$ and s_i^2 defined in (2.6), (2.7), and (2.8) respectively.

The negative of the Hessian matrix of the log of the posterior density of the six model parameters with respect to the vector $\boldsymbol{\theta}$ is the symmetric matrix $\boldsymbol{R}_2 = (-r_2(i,j))$, with

$$r_2(i,j) = \partial^2 \log\pi(\mu_\theta, \mu_\beta, \epsilon_\theta, \epsilon_\beta, \epsilon_\zeta, \epsilon_\nu|\boldsymbol{y})/\partial\theta_i\partial\theta_j \qquad (i,j = 1,\ldots,6), \tag{4.78}$$

where the elements of its upper triangular part are

$$r_2(1,1) = -\sum_{i=1}^{m}\frac{e^{\epsilon_\nu} + n_i}{(n_i^{-1} + e^{\epsilon_\theta})\Xi_i} + 2\sum_{i=1}^{m}\frac{(y_{i\cdot} - \mu_\theta)^2(e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^2\Xi_i^2} \tag{4.79}$$

$$r_2(1,2) = +2\sum_{i=1}^{m}\frac{(y_{i\cdot} - \mu_\theta)(\hat{\beta}_i - \mu_\beta)(e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})(s_i^{-2} + e^{\epsilon_\beta})\Xi_i^2} \tag{4.80}$$

$$r_2(1,3) = -e^{\epsilon_\theta}\sum_{i=1}^{m}\frac{(y_{i\cdot} - \mu_\theta)(e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^2\Xi_i} + e^{\epsilon_\theta}\sum_{i=1}^{m}\frac{(y_{i\cdot} - \mu_\theta)^3(e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^3\Xi_i^2} \tag{4.81}$$

$$r_2(1,4) = e^{\epsilon_\beta} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)(\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i^2} \tag{4.82}$$

$$r_2(1,5) = -e^{\epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)(e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta}) \Xi_i^2} \tag{4.83}$$

$$r_2(1,6) = e^{\epsilon_\nu} \sum_{i=1}^{m} \frac{y_{i.} - \mu_\theta}{(n_i^{-1} + e^{\epsilon_\theta}) \Xi_i} - e^{\epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)(e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta}) \Xi_i^2} \tag{4.84}$$

$$r_2(2,2) = -\sum_{i=1}^{m} \frac{e^{\epsilon_\nu} + n_i}{(s_i^{-2} + e^{\epsilon_\beta}) \Xi_i} + 2 \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i^2} \tag{4.85}$$

$$r_2(2,3) = e^{\epsilon_\theta} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)(y_{i.} - \mu_\theta)^2 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})(n_i^{-1} + e^{\epsilon_\theta})^2 \Xi_i^2} \tag{4.86}$$

$$r_2(2,4) = -e^{\epsilon_\beta} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)(e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i} + e^{\epsilon_\beta} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^3 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^3 \Xi_i^2} \tag{4.87}$$

$$r_2(2,5) = -e^{\epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)(e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta}) \Xi_i^2} \tag{4.88}$$

$$r_2(2,6) = e^{\epsilon_\nu} \sum_{i=1}^{m} \frac{\hat{\beta}_i - \mu_\beta}{(s_i^{-2} + e^{\epsilon_\beta}) \Xi_i} - e^{\epsilon_\zeta + \epsilon_\nu} \sum\nolimits_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)(e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta}) \Xi_i^2} \tag{4.89}$$

$$\begin{aligned} r_2(3,3) &= -\frac{1}{2}\omega_1 \tau_1 e^{-\epsilon_\theta} - \frac{1}{2} \sum_{i=1}^{m} \frac{e^{-\epsilon_\theta}}{n_i + e^{-\epsilon_\theta}} + \frac{1}{2} \sum_{i=1}^{m} \left(\frac{e^{-\epsilon_\theta}}{n_i + e^{-\epsilon_\theta}} \right)^2 \\ &+ \frac{1}{2} e^{\epsilon_\theta} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^2 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^2 \Xi_i} - e^{2\epsilon_\theta} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^2 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^3 \Xi_i} \\ &+ \frac{1}{2} e^{2\epsilon_\theta} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^4 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^4 \Xi_i^2} \end{aligned} \tag{4.90}$$

$$r_2(3,4) = e^{\epsilon_\theta + \epsilon_\beta} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^2 (\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^2 (s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i^2} \tag{4.91}$$

$$r_2(3,5) = -e^{\epsilon_\theta + \epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^2 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^2 \Xi_i^2} \tag{4.92}$$

$$r_2(3,6) = \frac{1}{2} e^{\epsilon_\theta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^2}{(n_i^{-1} + e^{\epsilon_\theta})^2 \Xi_i} - \frac{1}{2} e^{\epsilon_\theta + \epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(y_{i.} - \mu_\theta)^2 (e^{\epsilon_\nu} + n_i)}{(n_i^{-1} + e^{\epsilon_\theta})^2 \Xi_i^2} \tag{4.93}$$

$$\begin{aligned} r_2(4,4) &= -\frac{1}{2}\omega_2 \tau_2 e^{-\epsilon_\beta} - \frac{1}{2} \sum_{i=1}^{m} \frac{e^{-\epsilon_\beta}}{s_i^2 + e^{-\epsilon_\beta}} + \frac{1}{2} \sum_{i=1}^{m} \left(\frac{e^{-\epsilon_\beta}}{s_i^2 + e^{-\epsilon_\beta}} \right)^2 \\ &+ \frac{1}{2} e^{\epsilon_\beta} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i} - e^{2\epsilon_\beta} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^3 \Xi_i} \\ &+ \frac{1}{2} e^{2\epsilon_\beta} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^4 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^4 \Xi_i^2} \end{aligned} \tag{4.94}$$

$$r_2(4,5) = -e^{\epsilon_\beta + \epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i^2} \tag{4.95}$$

$$r_2(4,6) = \frac{1}{2} e^{\epsilon_\theta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^2}{(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i} - \frac{1}{2} e^{\epsilon_\beta + \epsilon_\zeta + \epsilon_\nu} \sum_{i=1}^{m} \frac{(\hat{\beta}_i - \mu_\beta)^2 (e^{\epsilon_\nu} + n_i)}{(s_i^{-2} + e^{\epsilon_\beta})^2 \Xi_i^2}$$

$$(4.96)$$

$$r_2(5,5) = -\frac{1}{2}e^{\epsilon_\zeta+\epsilon_\nu}\sum_{i=1}^{m}\frac{e^{\epsilon_\nu}+n_i}{\Xi_i} + \frac{1}{2}e^{2\epsilon_\zeta+2\epsilon_\nu}\sum_{i=1}^{m}\frac{e^{\epsilon_\nu}+n_i}{\Xi_i^2} - \frac{1}{2}\psi e^{\epsilon_\zeta+\epsilon_\nu} \quad (4.97)$$

$$\begin{aligned} r_2(5,6) &= -\frac{1}{2}e^{\epsilon_\zeta+2\epsilon_\nu}\sum_{i=1}^{m}\frac{1}{\Xi_i} - \frac{1}{2}e^{\epsilon_\zeta+\epsilon_\nu}\sum_{i=1}^{m}\frac{e^{\epsilon_\nu}+n_i}{\Xi_i} + \frac{1}{2}e^{2\epsilon_\zeta+2\epsilon_\nu}\sum_{i=1}^{m}\frac{e^{\epsilon_\nu}+n_i}{\Xi_i^2} \\ &+ \frac{1}{2}me^{\epsilon_\nu} + \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu} - \frac{1}{2}\psi e^{\epsilon_\zeta+\epsilon_\nu} \end{aligned} \quad (4.98)$$

$$\begin{aligned} r_2(6,6) &= -e^{\epsilon_\zeta+2\epsilon_\nu}\sum_{i=1}^{m}\frac{1}{\Xi_i} - \frac{1}{2}e^{\epsilon_\zeta+\epsilon_\nu}\sum_{i=1}^{m}\frac{e^{\epsilon_\nu}+n_i}{\Xi_i} + \frac{1}{2}e^{2\epsilon_\zeta+2\epsilon_\nu}\sum_{i=1}^{m}\frac{e^{\epsilon_\nu}+n_i}{\Xi_i^2} \\ &- \frac{1}{2}e^{\epsilon_\nu}\sum_{i=1}^{m}\log(\Xi_i/2) - \frac{1}{2}\psi e^{\epsilon_\zeta+\epsilon_\nu} + \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu}\log(\psi e^{\epsilon_\zeta+\epsilon_\nu}/2) - \frac{1}{2}be^{\epsilon_\nu} \\ &+ \frac{1}{2}me^{\epsilon_\nu}\log(e^{\epsilon_\zeta+\epsilon_\nu}/2) + me^{\epsilon_\nu} + \psi\zeta_0 e^{\epsilon_\nu} - \frac{1}{2}me^{\epsilon_\nu}\Omega(e^{\epsilon_\nu}/2) \\ &- \frac{1}{4}me^{2\epsilon_\nu}\Omega'(e^{\epsilon_\nu}/2) - \frac{1}{2}\psi\zeta_0 e^{\epsilon_\nu}\Omega(\psi\zeta_0 e^{\epsilon_\nu}/2) - \frac{1}{4}\psi^2\zeta_0^2 e^{2\epsilon_\nu}\Omega'(\psi\zeta_0 e^{\epsilon_\nu}/2) \\ &+ \frac{1}{2}e^{\epsilon_\nu}\sum_{i=1}^{m}\Omega\left(\frac{e^{\epsilon_\nu}+n_i}{2}\right) + \frac{1}{4}e^{2\epsilon_\nu}\sum_{i=1}^{m}\Omega'\left(\frac{e^{\epsilon_\nu}+n_i}{2}\right), \end{aligned} \quad (4.99)$$

where

$$\Omega'(z) = d^2\log\Gamma(z)/dz^2. \quad (4.100)$$

Chapter 5

Spatio-temporal models for oil well pressures

In this final chapter we will extend some of the Bayesian techniques already used to spatio-temporal models for oil well pressures. Initially, we will give a detailed description of the problem and discuss some data characteristics. An exploratory data analytic method that detects potentially good predictor wells and its application on a homogeneous sector of the oil field under study which contains twelve producer wells will then be presented. A self-similarity spatio-temporal multiple response model for pressures with discussion of its Bayesian analysis will then be indicated. A reduced model for a single well, and its analysis via an application of the Gibbs sampler will then be proposed, followed by a confirmatory study of the results of the previously mentioned predictive method on the twelve wells. We conclude with some indication of directions for future research.

5.1 Background

We will analyze data obtained from the British Petroleum oil field at Kuparuk, in Alaska. They consist of monthly measurements of flow rate taken over a period of 171 months, beginning in January 1981. These measurements denote flow rates (volume in barrels per day), averaged over 1 month. For injectors they are the flow rates of the water injected, while for producers, the oil production rate is one component of the total volume flow rate. If we denote by V_C the volume of each compound, C, produced, then the average overall production in a given well over a certain period can be obtained

via the following standard formula:

$$\text{Flow Rate} = \frac{1.25 \times V_{Oil} + 0.8 \times (V_{Gas} - 0.45 \times V_{Oil}) + V_{Water}}{\text{Days in operation}}, \qquad (5.1)$$

which accounts for differences in each component volume between the rock formation underground and the surface pipes. The flow rates are proportional to the well pressure. Injection is needed to maintain pressure and hence production in producer wells.

The data set consists of measurements recorded from all the 840 wells in the field; in units of barrels per day.. For each well, there is a number of months during which no measurements are available, because of closure due to maintenance, insufficient data recording and other reasons. Additionally, during the period of study, some wells were operated as both injectors during some months and producers during some other months. The oil field is divided into two sectors, north east and south west, by a fault across which it is thought that there is no oil, gas and water flow.

It is of major interest to be able to predict the output of producer wells, given past and present values of the pressures of injector and producer wells that have been identified as strongly related to the producers of interest. Current values can be important predictors because of the time interval it takes an injector to actually have an effect on production levels. The development of such a procedure can potentially lead to huge financial savings associated with the optimal operation of existing oil wells and the targeted drilling of new ones. In this context, it is important to investigate the hypothesis of long-term correlation between pairs of wells. Geologists use the term long-term correlation to denote correlation between wells at distances greater than can be accounted for by simple Darcian flow, due to non-linear mechanical effects.

Earlier work by Banks (1995), concentrated on computing Spearman's rank correlation coefficient for all possible pairs of injector/producer wells and trying, quite successfully, to match the directions of higher correlations implying good communication between wells parallel to the orientations of the maximum horizontal stress. The correlation with the maximum stress implies that the non-linear mechanism probably involves coupling between the fluid permeability, the local pore fluid pressure, and the tectonic stress field.

The work we will present in this chapter, is an initial attempt at identifying injectors and other producers whose pressures can help predict the actual pressure of a given producer well and at modelling the production pressures of wells in a relatively small

homogeneous area. Such studies will help detect the physical reason for the correlations, as well as potentially form the basis for the design and operation of production fields. A map of the part of the south west section of the oil field, which we will use in our analysis, is displayed in Figure 5.1 (page 125).

5.2 Exploratory analysis

5.2.1 Methods

In order to identify wells with pressures associated with the pressures of a chosen producer well of interest at time t, y_t, $t = 2, \ldots, 171$, we will use a predictive mean squared error criterion of the form

$$\sum_{t=2}^{171} \left(y_t - \boldsymbol{\beta}_1^T \boldsymbol{x}_t - \boldsymbol{\beta}_2^T \boldsymbol{x}_{t-1} - \boldsymbol{\beta}_3^T \boldsymbol{y}_t - \boldsymbol{\beta}_4^T \boldsymbol{y}_{t-1} \right)^2, \tag{5.2}$$

with $\boldsymbol{x}_t$ the injection at selected wells at time t, $\boldsymbol{x}_{t-1}$ the injection at selected, possibly different than before, wells at time $t-1$, $\boldsymbol{y}_t$ the production at selected wells at time t, and $\boldsymbol{y}_{t-1}$ the production at selected wells, including possibly the chosen producer well, at time $t-1$, and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, and $\boldsymbol{\beta}_4$ unknown vector parameters. This set up is completely general, and can be modified to include further lags of producers and injectors, or possibly less terms that the ones appearing in (5.2).

Using this method, it is easy to identify wells that substantially increase the coefficient of determination, R^2, and hence are potentially useful for prediction. Although we are using the least squares method and standard statistical packages to obtain estimated values for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, and $\boldsymbol{\beta}_4$, we do not focus on the inference for these parameters, since clearly the independent error assumption of the linear model theory is violated, e.g. by including $\boldsymbol{y}_t$ as predictors and because the y_t are time correlated. Rather, at this exploratory stage, we just use R^2 and the Bayesian criterion described below to obtain some initial suggestions of good predictors.

Adding new predictors may reduce the number of observations because of the missing values, so we will use at this stage a modification of the Bayesian criterion presented in (2.17) to compare different models. This criterion is defined as

$$BIC^{\dagger} = -\log 2\pi - \log(\frac{S_R^2}{N}) - 1 - \frac{k}{N} \max \left\{ \log \left(\frac{N}{2\pi} \right), 2 \right\}, \tag{5.3}$$

where k is the number of estimated parameters, including the error variance, N is the number of observations in the model, and S_R^2 the standard residual sum of squares. This modification can be derived by obtaining the maximized log likelihood of the linear model, penalizing it by the same factor as in (2.17) and then dividing by $N/2$. In this manner, the resulting criterion gives a value per observation and can be applied for comparing models with different number of observations by selecting the model with higher criterion value.

5.2.2 Results

For the twelve producers wells enclosed in the rectangle of Figure 5.1, we applied the criterion (5.2) in order to identify corresponding sets of good predictors. Due to the absence of expert geological advice concerning potential wells highly correlated with any of the producers of interest, we proceeded using a parallel of forward selection regression, adding one predictor at a time. For this and all subsequent analyses we standardized each well time series to have sample mean 0 and standard deviation 1.

Table 5.1: *Exploratory analysis summary for twelve oil wells. I and P refer to injectors and producers respectively, the numeric subscript to well numbers and a superscript l to lagged values of the corresponding wells.*

Producer Well	Predictor Wells										R^2
P_{280}	P^l_{280}	I_{301}	P^l_{552}	P^l_{361}	P_{514}						0.9057
P_{281}	P^l_{281}	I^l_{498}	I_{334}	P_{280}	P_{518}	I^l_{269}					0.9278
P_{282}	P^l_{282}	I^l_{394}	I_{271}	P^l_{550}							0.9161
P_{283}	P^l_{283}	I_{347}	P_{552}	P_{292}	I^l_{396}	I^l_{297}	I_{529}	I_{467}			0.9077
P_{518}	P^l_{518}	P_{497}	P^l_{548}	P_{517}							0.9325
P_{517}	P^l_{517}	P^l_{548}									0.9108
P_{304}	P^l_{304}	P_{354}									0.8647
P_{292}	P^l_{292}	I_{508}	I_{506}	I^l_{175}	I^l_{271}	I_{540}	I^l_{494}	I_{509}	I^l_{569}	I_{274}	0.7262
P_{515}	P^l_{515}	P^l_{553}	P^l_{318}	P_{295}	P_{514}	P_{331}					0.9026
P_{516}	P^l_{516}	P^l_{548}	I_{561}	P_{393}	P_{294}	I_{567}	I^l_{364}	I_{372}	I^l_{394}	I_{273}	0.9149
P_{305}	P^l_{305}	I_{506}	I_{372}	I^l_{372}	P_{552}	P^l_{549}	I^l_{467}				0.8535
P_{306}	P^l_{306}	P_{497}	P_{294}								0.8677

At every step, out of all south-east sector wells (injectors, producers and their one month lagged values), the next best predictor was selected, i.e. the one that maximized (5.3) and increased R^2, while, at the same time, its introduction didn't result in a

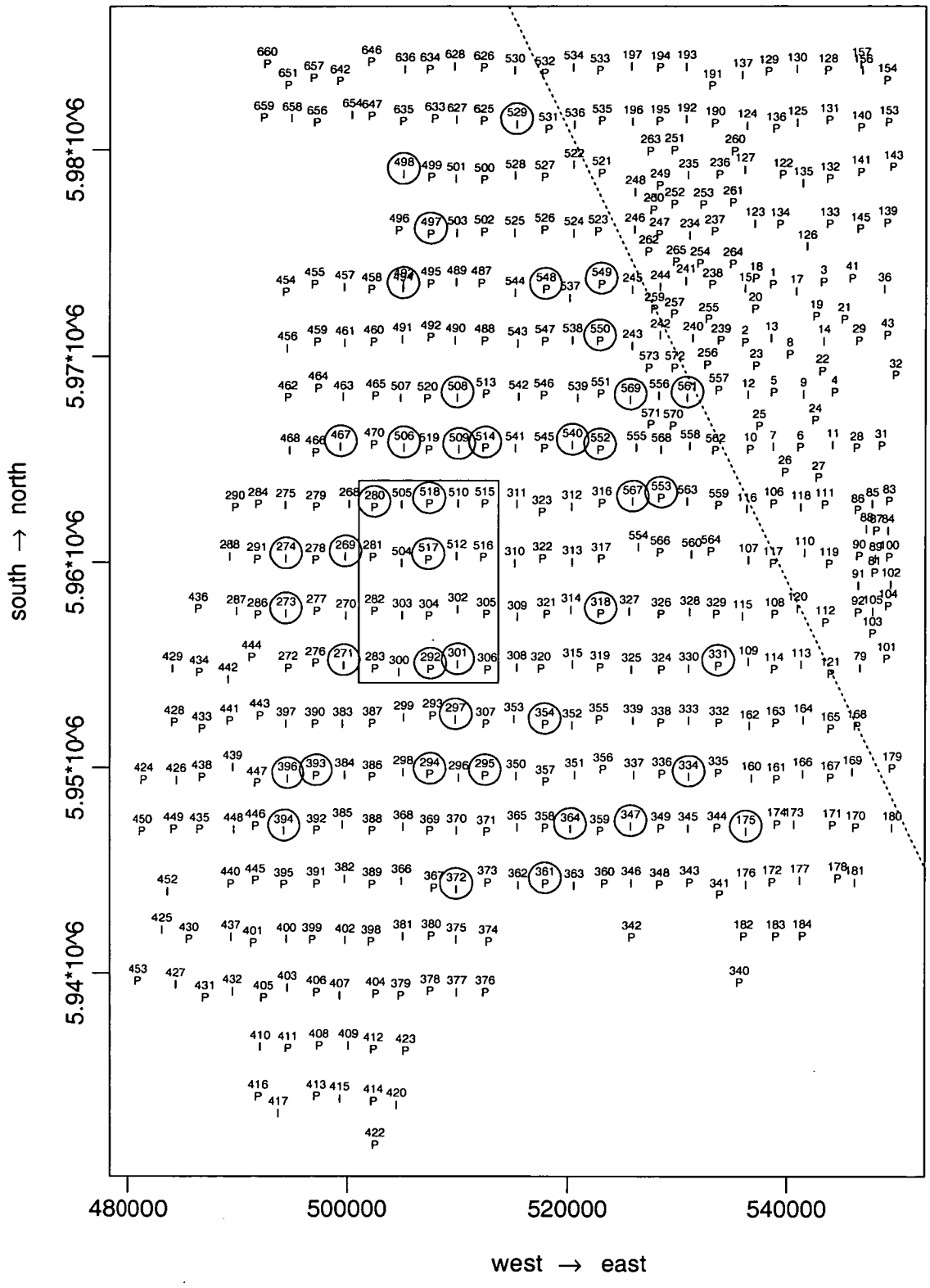


Figure 5.1: Map of the Kuparuk oil field. The numbers represent the number of each oil well. I and P refer to injector and producer wells respectively. The straight line is the sector dividing boundary. The rectangle encloses the set of producers under study. The circles denote good predictor wells according to our exploratory analysis.

greater than 10% loss of observations due to missing values. Our stopping rule was either a value of R^2 greater than 0.90 or the reduction of R^2, that can be explained by the possible sample size reduction, when introducing a new predictor.

Table 5.1 contains the predictor wells in the order they were introduced and the final value of R^2. For all the twelve wells of the region, the first predictor was the one month lagged value. There was no consistent pattern for the following steps, however, there seemed to be a high number of relatively distant wells in the oil field that led to substantial increases of R^2, as the high number of circles away from the rectangle in Figure 5.1 illustrate. Although at this preliminary stage we are not focusing on modelling these relationships, but just identifying possible relationships, the hypothesis of long-term correlation appears to be reasonable, based on these data.

5.3 Multiple well model

The proposed model for multiple wells can be expressed as

$$
\begin{aligned}
y_{it} &= \boldsymbol{x}_{it}^T \boldsymbol{\beta}_i + \theta_{it} + \epsilon_{it}, && (5.4)\\
\theta_{it} &= \theta_{i,t-1} + g_{it} + \xi_t + a_{it} + \eta_{it}, && (5.5)\\
g_{it} &= g_{i,t-1} + h_{it} + \alpha_{it}, && (5.6)\\
h_{it} &= h_{i,t-1} + \tau_{it}, && (5.7)\\
\xi_t &= \xi_{t-1} + \lambda_t + \omega_t, && (5.8)\\
\lambda_t &= \lambda_{t-1} + \gamma_t + \zeta_t, && (5.9)\\
\gamma_t &= \gamma_{t-1} + \kappa_t, && (5.10)
\end{aligned}
$$

for $i = 1, 2, \ldots, r$ and $t = 1, 2, \ldots, N$, with the error terms ϵ_{it}, η_{it}, α_{it}, τ_{it}, ω_t, ζ_t and κ_t mutually independent and normally distributed with corresponding variances $V_{i\epsilon}$, $V_{i\eta}$, $V_{i\alpha}$, $V_{i\tau}$ V_ω V_ζ and V_κ. Additionally the a_{it} are independent for different t and $(a_{1t}, \ldots, a_{rt})$ have a multivariate normal distribution with zero mean and covariance matrix satisfying

$$cov(a_{it}, a_{jt}) = ve^{-b<i,j>}, \qquad (5.11)$$

with v and b unknown parameters and $< i, j >$ the distance measure between wells i and j.

This model is a version of the normal dynamic linear model (DLM) studied by Harrison and Stevens (1976), Pole et al (1994), and West and Harrison (1997), with constant variances over time (constant DLM). According to this formulation, the production of well i at time t, as expressed by the observation equation (5.4), depends on the past and current pressures of some good predictor wells through the regression function $\boldsymbol{x}_{it}^T\boldsymbol{\beta}_i$ and the level of a well specific underlying process θ_{it}, with ϵ_{it} the corresponding observational error. There are also two quadratic growth elements (the quadratic growth model is discussed in the following section), one for individual wells, equations (5.5)-(5.7), and an overall one, equations (5.8)-(5.10). Additionally, a_{it} is the spatial component of the model and ξ_t the self-similarity component. Many geological and geophysical processes exhibit self-similarity, e.g. Turcotte (1992) and Main (1996). Hence a statistical model with this property is reasonable and desirable. West and Harrison (1997) mainly present results with all variances specified or only the observational variance unknown and the ratios of the system variance(s), as all the variances excluding $V_{i\epsilon}$ are referred to, to the observational one specified. We assume instead that all variances are unknown and will have to be estimated, following Leonard and Hsu's (1999) analysis for a simple case and improving on Mehra (1979) who obtains point estimates for variances.

This is the model proposed for future research. Its full Bayesian analysis can be performed by MCMC, along the lines we will demonstrate for the single well model in the subsequent sections. The only difficult parameter to simulate is b, which we can discretize over a range, similarly with the parameter ν in the ANCOVA models. It is thought that by introducing the regressors, we will be able to explain the spatial correlation and hence render the spatial component terms insignificant ($b \equiv 0$).

5.4 Single well model

5.4.1 Sampling model

A simplified version of the full model for the analysis of pressures of individual wells is

$$y_t = \boldsymbol{x}_t^T\boldsymbol{\beta} + \theta_t + \epsilon_t, \tag{5.12}$$

$$\theta_t = \theta_{t-1} + g_t + \eta_t, \tag{5.13}$$

$$g_t = g_{t-1} + h_t + \alpha_t, \tag{5.14}$$

$$h_t = h_{t-1} + \tau_t, \tag{5.15}$$

for $t = 1, 2, \ldots, N$, with the error terms ϵ_t, η_t, α_t and τ_t mutually independent and normally distributed with mean 0 and variances V_ϵ, V_η, V_α and V_τ, respectively. For analytic convenience, we also assume θ_0, g_0, h_0, to be mutually independent and normally distributed with mean 0 and respective variances $k_\eta V_\eta$, $k_\alpha V_\alpha$ and $k_\tau V_\tau$, with the k_η, k_α and k_τ specified.

This is related to the quadratic growth model described by West and Harrison (1997, p. 226) with the additional regression terms in (5.12) and the four variances V_ϵ, V_η, V_α and V_τ unknown. The terms θ_t, g_t and h_t correspond to level, growth and change of growth of the underlying process at time t. The quadratic growth model is considered adequate for short and long term forecasting in many practical situations (see West and Harrison, 1997, pp. 208 and 225).

This model without covariates can be expressed in an ARIMA(0,3,3) form,

$$(1 - B)^3 y_t = \epsilon_t - \xi_1 q_{t-1} - \xi_2 q_{t-2} - \xi_3 q_{t-3}, \tag{5.16}$$

with the errors, q_t, independent and normally distributed with zero mean and variance $\xi_3^{-1} V_\epsilon$, and B the backward shift operator, defined by $B y_t = y_{t-1}$, provided that the ξ_i, for $i = 1, 2, 3$, satisfy the following equations

$$\begin{aligned} \xi_3^{-1}(1 + \xi_1^2 + \xi_3^2 + \xi_3^2) &= (20 + 6\gamma_1 + 2\gamma_2 + \gamma_3) \\ \xi_3^{-1}(\xi_1 - \xi_1\xi_2 - \xi_2\xi_3) &= (15 + 4\gamma_1 + \gamma_2) \\ \xi_3^{-1}(-\xi_2 + \xi_1\xi_3) &= (6 + \gamma_1), \end{aligned} \tag{5.17}$$

where $\gamma_1 = V_\eta / V_\epsilon$, $\gamma_2 = V_\alpha / V_\epsilon$ and $\gamma_3 = V_\tau / V_\epsilon$. Additionally, the roots of the equation $1 - \xi_1 z - \xi_2 z^2 - \xi_3 z^3 = 0$ must be greater than one in absolute value, so that the series satisfies the invertability condition for moving average processes.

Reduced versions of the proposed single well model can be obtained if some of the variance components are found to equal zero. In particular, if V_τ=0, then all the h_t are zero, and we effectively have the linear growth model of West and Harrison (1997, p. 218) plus the regression terms. This model, can also be expressed in the ARIMA(0,2,2) form

$$(1 - B)^2 y_t = \epsilon_t - \xi_1 q_{t-1} + \xi_2 q_{t-2}, \tag{5.18}$$

with the q_t independent and normally distributed with zero mean and variance $\xi_2^{-1}V_\epsilon$, if the following equations are satisfied,

$$\begin{aligned} \xi_2^{-1}(1+\xi_1^2+\xi_3^2) &= (6+2\gamma_1+\gamma_2) \\ \xi_2^{-1}(\xi_1+\xi_1\xi_2) &= (4+\gamma_1), \end{aligned} \tag{5.19}$$

and the roots of the equation $1-\xi_1 z+\xi_2 z^2=0$ are all greater than one in absolute value.

If additionally V_α=0, then h_t and g_t are all zero, (5.13) is the last evolution equation, and the reduced model is a Markovian one with superimposed random noise plus the regressors. The equivalent ARIMA representation is a (0,1,1), that is, of the form

$$(1-B)y_t = \epsilon_t - \xi q_{t-1}, \tag{5.20}$$

with the q_t, independent and normally distributed with zero mean and variance $\xi^{-1}V_\epsilon$, if $\xi^{-1}+\xi = 2+\gamma_1$ and with ξ_0, the smallest root of the previous equation, satisfying $\xi_0 \leq 1$ (Leonard and Hsu, 1999, p. 233). The equivalence between the constant DML and ARIMA forms of the previous three models are obtained by matching the autocovariance structures under the two formulations. Finally, if V_τ, V_α and V_η all equal zero, the reduced model is a linear regression one, since all the h_t, g_t and θ_t equal zero, for $t=0,1,\ldots,N$.

Using the described formulation for the quadratic growth model including the regressors, the likelihood of the slope vector, $\boldsymbol{\beta}$, and the four variance components is equal to

$$\begin{aligned} & p(\boldsymbol{y}_t, \boldsymbol{\theta}_t, \boldsymbol{g}_t, \boldsymbol{h}_t | \boldsymbol{\beta}, V_\epsilon, V_\eta, V_\alpha, V_\tau) = \\ &= (2\pi V_\epsilon)^{-N/2} \exp\left\{-\tfrac{1}{2}V_\epsilon^{-1}\textstyle\sum_{t=1}^N (y_t - \boldsymbol{x}_t^T\boldsymbol{\beta} - \theta_t)^2\right\} \\ &\times (2\pi V_\eta)^{-N/2}(2\pi k_\eta V_\eta)^{-1/2} \exp\left\{-\tfrac{1}{2}V_\eta^{-1}\textstyle\sum_{t=1}^N (\theta_t - \theta_{t-1} - g_t)^2 - \tfrac{1}{2}k_\eta^{-1}V_\eta^{-1}\theta_0^2\right\} \\ &\times (2\pi V_\alpha)^{-N/2}(2\pi k_\alpha V_\alpha)^{-1/2} \exp\left\{-\tfrac{1}{2}V_\eta^{-1}\textstyle\sum_{t=1}^N (g_t - g_{t-1} - h_t)^2 - \tfrac{1}{2}k_\alpha^{-1}V_\alpha^{-1}g_0^2\right\} \\ &\times (2\pi V_\tau)^{-N/2}(2\pi k_\tau V_\tau)^{-1/2} \exp\left\{-\tfrac{1}{2}V_\tau^{-1}\textstyle\sum_{t=1}^N (h_t - h_{t-1})^2 - \tfrac{1}{2}k_\tau^{-1}V_\tau^{-1}h_0^2\right\}. \end{aligned} \tag{5.21}$$

5.4.2 Prior to posterior inference

We assume that, a priori, $\boldsymbol{\beta}$ and the four variance components V_ϵ, V_η, V_α and V_τ are independent. We also assume that $\boldsymbol{\beta}$ is normally distributed with mean $\boldsymbol{\beta}_0$ and

covariance matrix $\boldsymbol{C}$, and, for the prior distribution of the four variances, that $\omega_1\lambda_1/V_\epsilon$, $\omega_2\lambda_2/V_\eta$, $\omega_3\lambda_3/V_\alpha$, and $\omega_4\lambda_4/V_\tau$ possess chi-squared distributions with ω_1, ω_2, ω_3 and ω_4 degrees of freedom respectively.

Using the prior distribution already described, the joint posterior density of the $\boldsymbol{\theta}_t$, $\boldsymbol{g}_t$, $\boldsymbol{h}_t$, the vector of slopes and the four variances, becomes

$$\begin{aligned}
&\pi(\boldsymbol{\theta}_t, \boldsymbol{g}_t, \boldsymbol{h}_t, \boldsymbol{\beta}, V_\epsilon, V_\eta, V_\alpha, V_\tau | \boldsymbol{y}_t) \propto p(\boldsymbol{y}_t, \boldsymbol{\theta}_t, \boldsymbol{g}_t, \boldsymbol{h}_t | \boldsymbol{\beta}, V_\epsilon, V_\eta, V_\alpha, V_\tau)\pi(\boldsymbol{\beta}, V_\epsilon, V_\eta, V_\alpha, V_\tau) \\
&\propto p(\boldsymbol{y}_t, \boldsymbol{\theta}_t, \boldsymbol{g}_t, \boldsymbol{h}_t | \boldsymbol{\beta}, V_\epsilon, V_\eta, V_\alpha, V_\tau) \left|\boldsymbol{C}\right|^{-1/2} \exp\left\{\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^T \boldsymbol{C}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\} \\
&\times V_\epsilon^{-\frac{1}{2}(\omega_1+2)} \exp(-\tfrac{1}{2}V_\epsilon^{-1}\omega_1\lambda_1) V_\eta^{-\frac{1}{2}(\omega_2+2)} \exp(-\tfrac{1}{2}V_\eta^{-1}\omega_2\lambda_2) \\
&\times V_\alpha^{-\frac{1}{2}(\omega_3+2)} \exp(-\tfrac{1}{2}V_\alpha^{-1}\omega_3\lambda_3) V_\tau^{-\frac{1}{2}(\omega_4+2)} \exp(-\tfrac{1}{2}V_\tau^{-1}\omega_4\lambda_4).
\end{aligned} \tag{5.22}$$

Similarly with section 2.3.2, we can use the joint posterior density in (5.22) to obtain the full conditional densities of each of its parameters and subsequently apply the Gibbs sampler to obtain the marginal posterior density of every quantity of interest. A minor algebraic complication relates to the fact that for $t = 1, 2, \ldots, N-1$, each of the θ_t, g_t and h_t appear in three different terms of (5.21) and thus of (5.22). For example the full conditional posterior density of θ_t, for $t = 1, 2, \ldots, N-1$, is

$$\begin{aligned}
\pi^*(\theta_t) \;&\propto\; \exp\left\{-\tfrac{1}{2}V_\epsilon^{-1}(y_t - \boldsymbol{x}_t^T\boldsymbol{\beta} - \theta_t)^2\right\} \\
&\times\; \exp\left\{-\tfrac{1}{2}V_\eta^{-1}(\theta_t - \theta_{t-1} - g_t)^2 - \tfrac{1}{2}V_\eta^{-1}(\theta_{t+1} - \theta_t - g_{t+1})^2\right\}.
\end{aligned} \tag{5.23}$$

The previous expression is a normal density for θ_t. This can be shown by using the following lemma.

Lemma 5.1. If a, b, c, x, A, B and C are scalars with A, B or $C \neq 0$, then

$$A(x-a)^2 + B(x-b)^2 + C(x-c)^2 = k + (A+B+C)\left(x - \frac{Aa+Bb+Cc}{A+B+C}\right)^2, \tag{5.24}$$

with k constant in x.

For a proof see Lemma 1.2 and Corollary 1.2 of which the current lemma is a special case, or expand the squared terms on the left and right hand side of (5.24) and equate the coefficients of x^2 and x.

Using Lemma 5.1, the full conditional posterior density of θ_t becomes

$$\pi^*(\theta_t) \propto \exp\left\{-\frac{1}{2}(V_\epsilon^{-1} + 2V_\eta^{-1})\left(\theta_t - \theta_t^*\right)^2\right\}, \tag{5.25}$$

with

$$\theta_t^* = \left(V_\epsilon^{-1} + 2V_\eta^{-1}\right)^{-1} \left\{V_\epsilon^{-1}\left(y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}\right) + V_\eta^{-1}\left(\theta_{t-1} + \theta_{t+1} + g_t - g_{t+1}\right)\right\}, \quad (5.26)$$

so it is normal with mean θ_t^* and variance $\left(V_\epsilon^{-1} + 2V_\eta^{-1}\right)^{-1}$. Similar algebraic derivations can demonstrate that the full conditional densities of the g_t and h_t, for $t = 1, 2, \ldots, N-1$ are also normal. The full conditional densities of g_0, g_N, h_0, h_N, the slope vector, $\boldsymbol{\beta}$, and the four variance components can be derived using computations analogous to these of sections 2.3.2 and 2.5.1. The relevant results for all parameters are presented in the next section.

5.4.3 Full conditional distributions

Each of the following statements is made conditionally upon the data, and all other unknown random variables and parameters in the model:

S1: θ_0 is normally distributed with mean θ_0^*, where

$$\theta_0^* = \left(1 + k_\eta^{-1}\right)^{-1}(\theta_1 - g_1), \quad (5.27)$$

and variance $V_\eta\left(1 + k_\eta^{-1}\right)^{-1}$.

S2: For $t = 1, 2, \ldots, N-1$, θ_t is normally distributed with mean θ_t^*, defined in (5.26) and variance $\left(V_\epsilon^{-1} + 2V_\eta^{-1}\right)^{-1}$.

S3: θ_N is normally distributed with mean θ_N^*, where

$$\theta_N^* = \left(V_\epsilon^{-1} + V_\eta^{-1}\right)^{-1} \left\{V_\epsilon^{-1}\left(y_N - \boldsymbol{x}_N^T\boldsymbol{\beta}\right) + V_\eta^{-1}\left(\theta_{N-1} + g_N\right)\right\}, \quad (5.28)$$

and variance $\left(V_\epsilon^{-1} + V_\eta^{-1}\right)^{-1}$.

S4: g_0 is normally distributed with mean g_0^*, where

$$g_0^* = \left(1 + k_\alpha^{-1}\right)^{-1}(g_1 - h_1), \quad (5.29)$$

and variance $V_\alpha\left(1 + k_\alpha^{-1}\right)^{-1}$.

S5: For $t = 1, 2, \ldots, N-1$, g_t is normally distributed with mean g_t^*, where

$$g_t^* = \left(V_\eta^{-1} + 2V_\alpha^{-1}\right)^{-1} \left\{V_\eta^{-1}\left(\theta_t - \theta_{t-1}\right) + V_\alpha^{-1}\left(g_{t-1} + g_{t+1} + h_t - h_{t+1}\right)\right\}, \quad (5.30)$$

and variance $\left(V_\eta^{-1} + 2V_\alpha^{-1}\right)^{-1}$.

S6: g_N is normally distributed with mean g_N^*, where

$$g_N^* = \left(V_\eta^{-1} + V_\alpha^{-1}\right)^{-1} \left\{V_\eta^{-1}\left(\theta_N - \theta_{N-1}\right) + V_\alpha^{-1}\left(g_{N-1} + h_N\right)\right\}, \tag{5.31}$$

and variance $\left(V_\eta^{-1} + V_\alpha^{-1}\right)^{-1}$.

S7: h_0 is normally distributed with mean h_0^*, where

$$h_0^* = \left(1 + k_\tau^{-1}\right)^{-1} h_1, \tag{5.32}$$

and variance $V_\tau \left(1 + k_\tau^{-1}\right)^{-1}$.

S8: For $t = 1, 2, \ldots, N-1$, h_t is normally distributed with mean h_t^*, where

$$h_t^* = \left(V_\alpha^{-1} + 2V_\tau^{-1}\right)^{-1} \left\{V_\alpha^{-1}\left(g_t - g_{t-1}\right) + V_\tau^{-1}\left(h_{t-1} + h_{t+1}\right)\right\}, \tag{5.33}$$

and variance $\left(V_\alpha^{-1} + 2V_\tau^{-1}\right)^{-1}$.

S9: h_N is normally distributed with mean h_N^*, where

$$h_N^* = \left(V_\alpha^{-1} + V_\tau^{-1}\right)^{-1} \left\{V_\alpha^{-1}\left(g_N - g_{N-1}\right) + V_\tau^{-1} h_{N-1}\right\}, \tag{5.34}$$

and variance $\left(V_\alpha^{-1} + V_\tau^{-1}\right)^{-1}$.

S10: For the variance V_ϵ, the quantity $(\omega_1 + N)V_\epsilon^*/V_\epsilon$ has a chi-squared distribution with $\omega_1 + N$ degrees of freedom, where

$$V_\epsilon^* = \left\{\omega_1\lambda_1 + \sum_{t=1}^{N}\left(y_t - \boldsymbol{x}_t^T\boldsymbol{\beta} - \theta_t\right)^2\right\}/(\omega_1 + N). \tag{5.35}$$

S11: For the variance V_η, the quantity $(\omega_2 + N + 1)V_\eta^*/V_\eta$ has a chi-squared distribution with $\omega_2 + N + 1$ degrees of freedom, where

$$V_\eta^* = \left\{\omega_2\lambda_2 + \sum_{t=1}^{N}\left(\theta_t - \theta_{t-1} - g_t\right)^2 + k_\eta^{-1}\theta_0^2\right\}/(\omega_2 + N + 1). \tag{5.36}$$

S12: For the variance V_α, the quantity $(\omega_3 + N + 1)V_\alpha^*/V_\alpha$ has a chi-squared distribution with $\omega_3 + N + 1$ degrees of freedom, where

$$V_\alpha^* = \left\{\omega_3\lambda_3 + \sum_{t=1}^{N}\left(g_t - g_{t-1} - h_t\right)^2 + k_\alpha^{-1}g_0^2\right\}/(\omega_3 + N + 1). \tag{5.37}$$

S13: For the variance V_τ, the quantity $(\omega_4+N+1)V_\tau^*/V_\tau$ has a chi-squared distribution with $\omega_4 + N + 1$ degrees of freedom, where

$$V_\tau^* = \left\{ \omega_4\lambda_4 + \sum_{t=1}^{N} (h_t - h_{t-1})^2 + k_\tau^{-1} h_0^2 \right\} / (\omega_4 + N + 1). \tag{5.38}$$

S14: The vector $\boldsymbol{\beta}$ is normally distributed with mean $\boldsymbol{\beta}^*$, where

$$\boldsymbol{\beta}^* = \left(V_\epsilon^{-1} \sum_{i=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^T + C^{-1} \right)^{-1} \left(V_\epsilon^{-1} \sum_{i=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^T \hat{\boldsymbol{\beta}} + C^{-1} \boldsymbol{\beta}_0 \right), \tag{5.39}$$

and variance $\left(V_\epsilon^{-1} \sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^T + C^{-1} \right)^{-1}$.

To obtain the corresponding full conditional posterior densities under the linear growth model, we only need **S1-S6**, **S10-S12** and **S14**, with $h_t = 0$, for $t = 0, 1, \ldots, N$, Finally, for the two-stage model, we need to iterate between **S1-S3**, **S10**, **S11** and **S14**, with $g_t = 0$, for $t = 0, 1, \ldots, N$, in order to obtain the marginal posterior distributions of all parameters of interest. Unlike the Kalman filter (e.g. Kalman, 1960, Kalman and Bucy, 1961, Harrison and Stevens, 1976, and Meinhold and Singpurwalla, 1983), the Gibbs sampling iterations use both past and future values of the underlying unobservable quantities θ_t, g_t, and h_t.

5.4.4 Results of analysis

Using the quadratic growth and the reduced DLM models already described, we repeated the analysis for the response and predictor wells of section 5.2.2, aiming to confirm the relationships observed earlier, but using more formal statistical methodology this time.

The missing values, that were omitted in the exploratory analysis, were tackled in two different ways. If predictor well values were missing, we imputed them as the weighted average of the immediately preceding and following available values with weights inversely proportional to their distance from the missing value. This is, of course, a rather simplistic imputation method that could be changing some of the results. Particularly, since it is flattening the predictor time series the significance of the corresponding slope coefficient may be downweighted. On the other hand, if a predictor is found to be important using the exploratory predictive method of section 5.2.1, as well as a constant DLM model despite the imputation inefficiency, then this is an indication that this predictor is quite good. If only a response value was missing, then

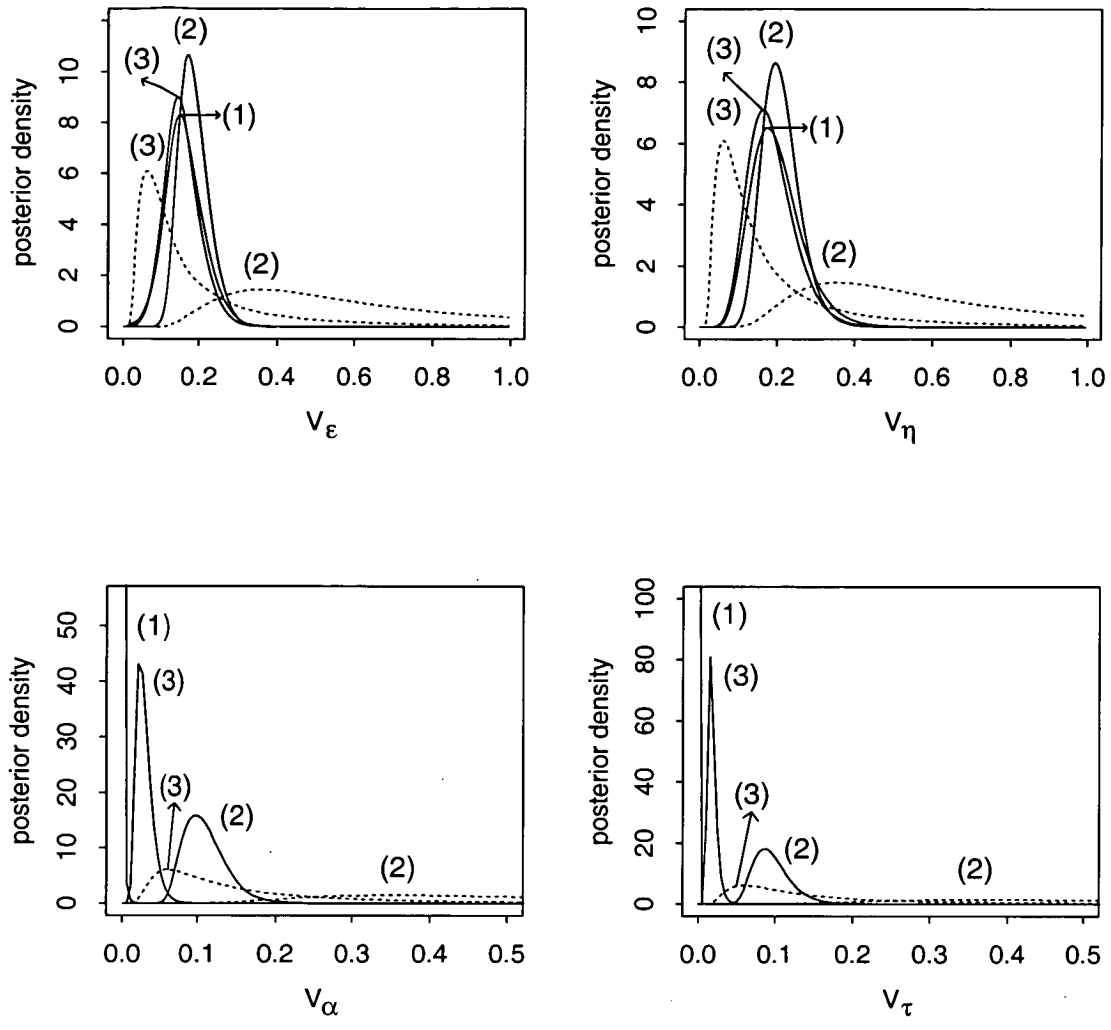


Figure 5.2: Top left: Posterior and prior density of V_ϵ for the two-stage Markovian model. Top right: Posterior and prior density of V_η for the two-stage Markovian model. Bottom left: Posterior density of V_α for the linear growth model. Bottom right: Posterior density of V_τ for the quadratic growth model. All results correspond to models with producer well 280 as response and its predictor wells of Table 5.1 as covariates and are presented for three choices, (1)-(3), of prior distribution. The solid curves denote the posterior densities and the dotted curves the proper prior densities.

it was generated as part of the Gibbs sampling iterations according to equation (5.12). The latter situation was rather rare for the twelve wells analyzed, a maximum of 10 values needed to be generated this way for well 517, however the percentage of imputation for each of the predictors of the twelve wells ranged from 0% to 25%. Typically, for the first 30 to 40 months there were no response or predictor observations available for neither of the twelve wells we studied. For modelling the response of each producer well its lagged value is no longer included in its predictors.

The analysis was performed under three sets of priors for the variance components.

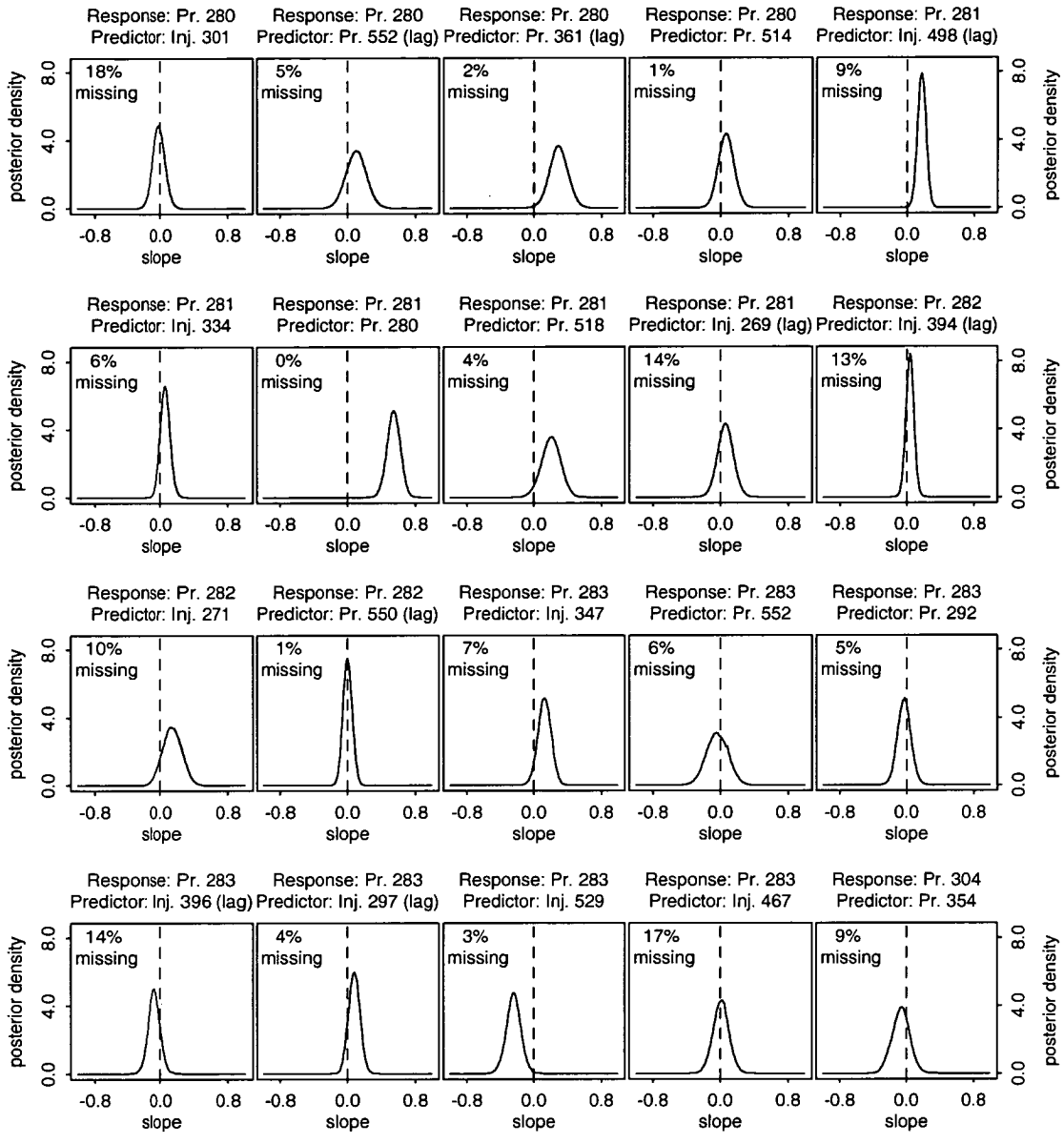


Figure 5.3: Posterior densities of slope coefficients of predictor wells for producer wells 280, 281, 282, 283 and 304 under the two stage Markovian model and uniform prior for the variance components. The stated percentages express proportion of imputed predictor values for the corresponding response.

For the first choice, prior (1), $\omega_i = -2$ and $\lambda_i = 0$, $i = 1, \ldots, 4$, that is, we assumed uniform priors. For the second, prior (2), $\omega_i = 5$ and $\lambda_i = 0.5$, $i = 1, \ldots, 4$, and for the third, prior (3), $\omega_i = 3$ and $\lambda_i = 0.1$. For the slope vector $\boldsymbol{\beta}$, we assumed always the same prior with mean $\mathbf{0}$ and diagonal covariance matrix with the variances all equal to 10, hence a rather uninformative prior given the scale of the observations (mean 0 and standard deviation 1). Finally we set, $k_\eta = k_\alpha = k_\tau = 2$.

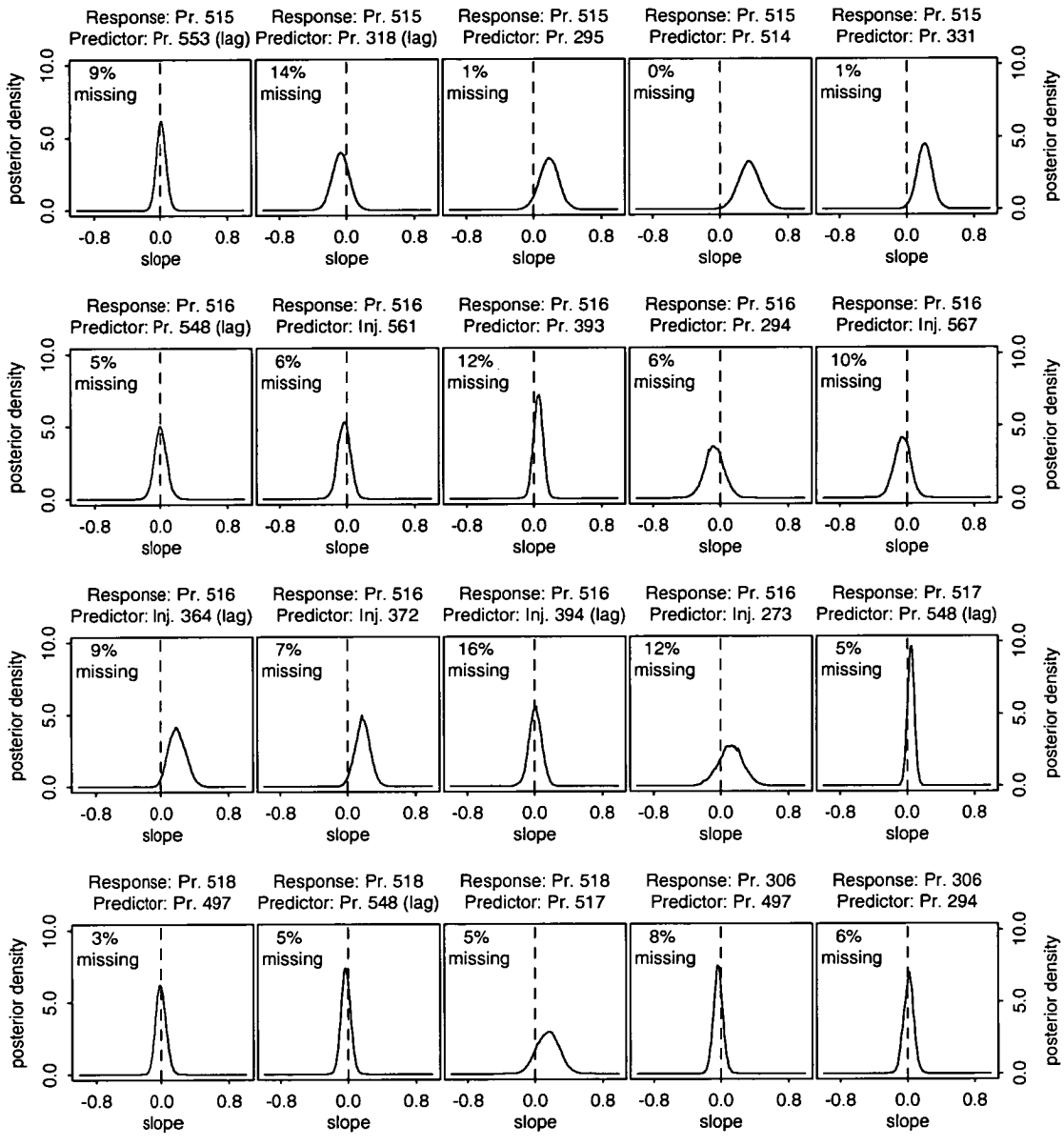


Figure 5.4: Posterior densities of slope coefficients of predictor wells for producer wells 515, 516, 517, 518 and 306 under the two stage Markovian model and uniform prior for the variance components. The stated percentages express proportion of imputed predictor values for the corresponding response.

Figure 5.2 contains graphs of the posterior densities and the corresponding priors of the variance components V_ϵ, V_η, V_α, V_τ for three different models for well 280. The results we describe were identical to these of the variance components of the corresponding models for the other 11 wells studied. The graphs for V_τ relate to the full quadratic growth model including the predictors. The posterior density of V_τ was quite insensitive to the choice of prior, however, it was always shifted towards zero for all

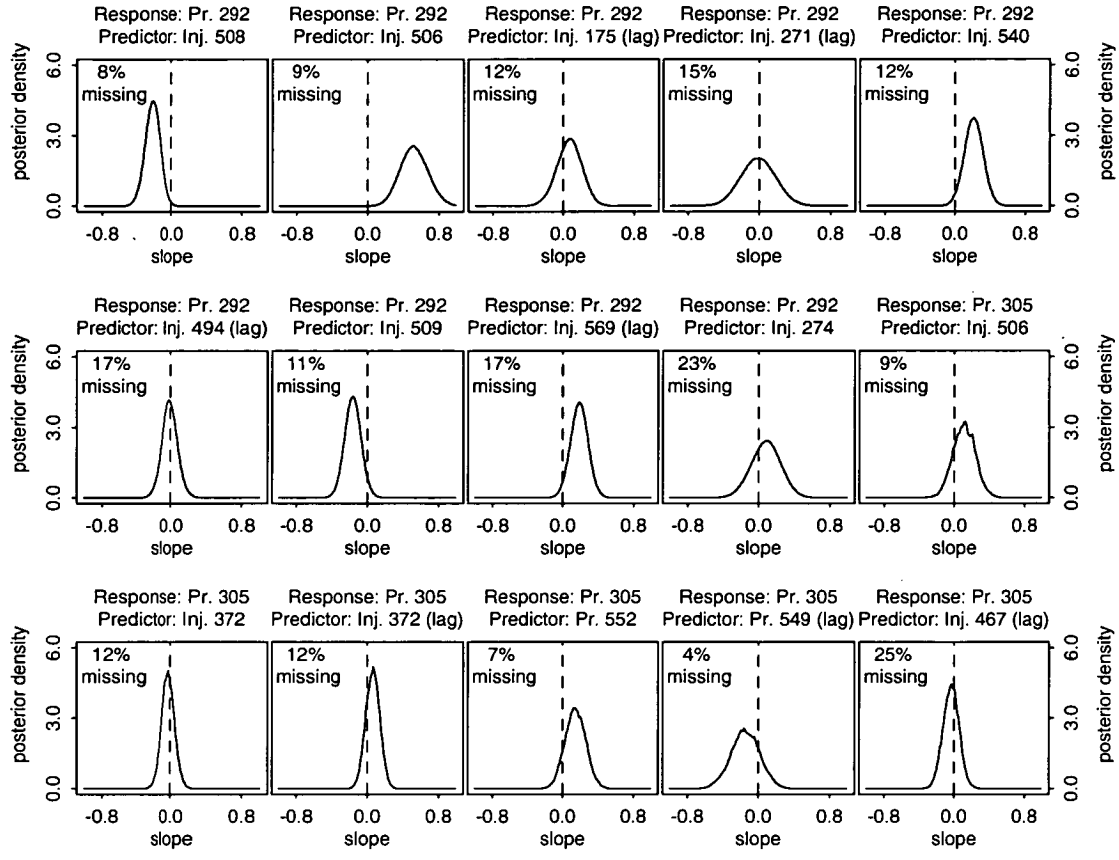


Figure 5.5: Posterior densities of slope coefficients of predictor wells for producer wells 292 and 305 under the two stage Markovian model and uniform prior for the variance components. The stated percentages express proportion of imputed predictor values for the corresponding response.

the priors tried, culminating at a spike at zero under uniform prior distribution. This suggested that the model had to be reduced, thus omitting (5.15). The graphs for V_α were obtained using the linear growth model with covariates and lead to the same conclusion as before, that V_α equals zero, hence omitting the last state equation (5.14). The graphs for V_ϵ and V_η correspond to the two stage Markovian model with covariates that includes only the θ_t. Clearly, for these two variances the posterior densities are less sensitive to the choice of prior and definitely not zero. These results, together with the ARIMA/DLM correspondence, seem to verify initial indications of the predictive method, which led us to use only up to one month lagged values in the analysis. The statistical confirmation of an one month lag will form a strong constraint for any physical models for the observed correlation.

Hence, for producer 280 and the remaining eleven ones studied, we subsequently

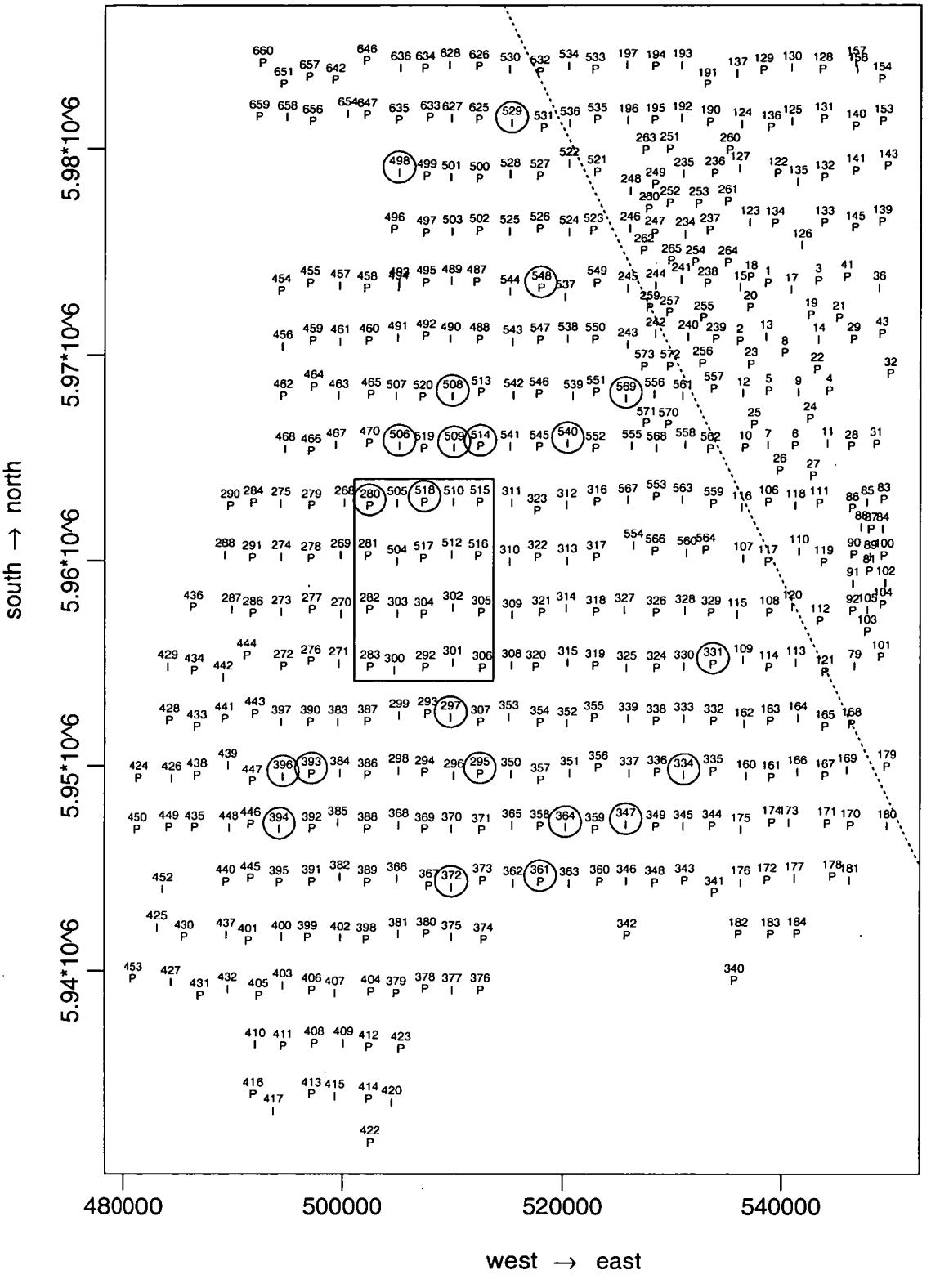


Figure 5.6: Map of the Kuparuk oil field. The numbers represent the number of each oil well. I and P refer to injectors and producers wells respectively. The straight line is the sector dividing boundary. The rectangle encloses the chosen set of producers. The circles denote good predictor wells according to the exploratory predictive analysis with non-zero slope according to the DLM analysis.

display the posterior densities of the slope coefficient for all their predictors under the current model (Figures 5.3-5.5). Our computations demonstrated that these posterior densities were very stable under the three different models. All the results presented were obtained by averaging 50,000 densities after 50,000 iterations of burn-in.

The absolute value of all the coefficients is smaller than 1 generally, a fact reflecting the scale of the data. Many of the wells that were considered good predictors in the exploratory analysis have insignificant slope coefficient. We did not only observe injectors with significantly positive slope, in which case their effect on the corresponding producer is easily understandable, but also injectors with significantly negative slope. In this situation, it is possible that this injector obstructs the flow towards the producer of interest from another injector, which is positively associated with that producer, so it may actually need to be shut off to increase the production. We have observed positive slope coefficients for producer predictor wells, possibly signifying pairs of wells benefiting from the same injectors, but not negative slopes, signifying pairs of competing producers.

Figure 5.6 represents a second map of the southeast section of Kuparuk field, this time with the predictor wells having significantly different than zero slope coefficients, according to the DLM analysis, encircled. Contrasting Figure 5.1 with Figure 5.6, we can observe that although a total of 20 wells lost their good predictor status with the DLM analysis, there is still a big number of wells, situated away from the twelve well rectangle, with significantly different than zero slope. It is also encouraging to notice that 14 of these wells are injectors and to remember the potential downweighting of the significance due to missing values.

5.5 Suggestions for future work (Chapter 5)

The convergence of the MCMC applications needs to be validated using Laplacian approximations, similarly to the ANCOVA computations and Chapter 4. To do this, the equivalence between the ARIMA and constant DLM models could be used to obtain estimates of the variance components.

Better imputation methods need to be developed, possibly after consulting more detailed records from the Kuparuk site and obtaining expert advice from petroleum engineers about the missing value meaning (failure to record, no production, maintenance, and so on).

Having obtained a fuller data set using a geologically sound imputation method the analysis of the multiple well model will become feasible. The analysis via MCMC and Laplacian approximations could be repeated, in order to examine whether a model without the spatial component is adequate and to examine the hypothesis of self-similarity, which is of major importance.

A desirable final stage would be to examine the predictive performance of the multiple well model against future observations, when these become available.

Chapter 6

Concluding remarks

In this final chapter we will describe some of the main conclusions obtained in the different subject areas covered in this thesis and indicate possibilities for future research.

6.1 ANCOVA models

Motivated by a related result of Crook and Good (1982), a more general theorem involving the optimal average power property of the Bayes factor was proved in section 1.5.1. The latter, combined with the lack of an optimal test statistic for testing equality of the group means in the parallel line ANCOVA model, triggered a search of alternative statistics and a study of their frequency properties. The main conclusion was that although different statistics, including the Bayes factor and the likelihood ratio statistic, could provide slight power improvements over the standard F test and differences in the significance probabilities that could result in opposing decisions, their computational complexity, which could only increase in more complicated models, rendered them less appealing.

In all situations, the Bayes factor was treated as a test statistic. We did not seek to interpret its values, due to the well established relevant problems. On the other hand, we used the Bayes factor in its most elementary form. More advanced forms studied by O'Hagan (1995), Berger and Perrichi (1996) and Berger and Mortera (1999) could also be considered in future research, however our opinion is that their appeal as possible test statistics would be questionable, as a result of the computational complications already described.

The last few sections of Chapter 1, included some initial suggestions concerning inference of mean equality by interpreting the posterior densities of the appropriate

parameter through carefully selected Bayesian significance probabilities and simplified Bayes factors.

The main thrust of the work in ANCOVA models was the proposal of a hierarchical model with random means, slopes and variances (section 2.1) and extensions to multi-covariate situations (section 2.5), that extended previous models in the literature and provided a novel way of addressing the question of variance equality, as well as a prior specification ensuring algebraic tractability (section 2.2). This provided a very comprehensive finite sample analysis of an ANCOVA model with unequal random variances. Based on results of Chapter 1 concerning the lack of an optimal test statistic, the idea of judging different equality hypotheses through Bayesian significance probabilities was fully developed in section 2.4. Using the standard idea in Bayesian statistics, of obtaining the posterior means of quantities of interest as the weighted averages of frequentist estimates and prior means, key cut off points used for the interpretation of posterior densities through posterior probabilities were suggested.

These suggestions were validated using several real data sets in Chapter 3. The magnitude of the Bayesian significance probabilities corresponding to acceptance/rejection of the hypotheses of interest was left to the experience of the data analyst and can provide stimulation for future work through lengthy simulations. A conclusion drawn from the study of two sets of simulated data (section 3.3) was that for the parameter ν, used for judging variance equality across groups for essentially ANOVA models with unequal variances, inferences based on visual inspection of the posterior densities of the variances were in close agreement with Bayesian significance probabilities with cutoff in the region of 0.01 to 0.05. Similar methods of inference could prove very valuable in variance components models in general for testing whether different variances could be considered to be equal to zero.

6.1.1 MCMC and Laplacian approximations

Inference for all random variances ANCOVA models, whether these included one or several covariates, was based on an application of the Gibbs sampler. The full conditional distributions of all random effects and parameters of the models proposed were relatively easy to obtain, a typical situation with hierarchical linear models. The only exception was the parameter ν. For the latter, three alternative methods were proposed, including a discretization one, similar to the griddy Gibbs sampler, and an approximate one, based on a suggestion of Lindley (1971), all of which gave extremely consistent results.

The implementation of the Gibbs sampler for the single covariate model, which was applied to the twelve neuropsychological test results, showed that the stabilization of the

marginal posterior densities occurred after a maximum total of less than 2,000 iterations for all quantities generated, and after a few hundreds for some of them. The facility and speed of the implementation of the MCMC procedures was thus clearly demonstrated. However, the word "stabilization" was just mentioned instead of "convergence", because it is known that this visual stabilization does not necessarily imply convergence (Cowles and Carlin, 1996). The well established weakness of MCMC methodology is that it is not possible to know at which point of the iterations the generated values come from the targeted distribution. In other words, the lack of relevant theoretical results, make it impossible to confirm that the apparent convergence is also actual convergence.

To compensate for this weakness, a study aimed at reproducing the MCMC based inferences using Laplacian approximations, a completely different method, was undertaken. Having as starting point the impossibility of obtaining accurate approximations using joint modes, and two rather simple relevant examples, we used a special case of an approximation suggested by Leonard et al (1989) to achieve practically identical inferences compared to those obtained with MCMC. These approximations used the joint posterior distribution of the model parameters and the corresponding modes for the marginal posterior distributions of all model parameters, and the joint posterior density of all random effects and conditional modes of the random effects given the model parameters to obtain approximations to the marginal posterior distributions of each of the random effects.

The implementation of Laplacian approximations confirmed first and foremost that MCMC is easier to apply and that subsequently one would most probably consider it as the first choice method of inference. However, it also confirmed that Laplacian approximations can be more time efficient and provide extremely accurate results even for small sample sizes. Tierney, Kass, and Kadane (1989a) regard Laplacian approximations as involving asymptotic justifications and develop their saddlepoint properties as the sample size tends to infinity. We have, however, demonstrated by the proof of section 4.2 and our computations, that they can provide excellent finite sample approximations, i.e. whenever a multivariate normal approximation to a conditional posterior density is accurate. Hence, when applied carefully, these approximations can be used both as a confirmatory tool of MCMC methods but also as the main method of inference for people with relevant experience, especially if the corresponding software was made publicly available. Although, we did not obtain any theoretical results for more general situations, the results of sections 4.3.1 and 4.4.1 provide a good indication of how to accurately approximate the marginal posterior distributions of random effects in hierarchical linear models in general. The methodology we presented greatly simplifies the Laplacian t approximation (Leonard, Hsu and Ritter, 1994, and Sun et al 1996), by avoiding the need for a complicated choice of the degrees of freedom. Hence, the latter

methodology would appear to be overcomplex.

As our implementation of the Gibbs sampler was only studied in a rather simple case, we did not address possible problems such as slow mixing chains and inefficient parameterizations. These could be studied, as topics of future research, with the general ANCOVA models of section 2.5, where it is possible that such problems might occur, together with the corresponding extensions of our Laplacian approximations, when appropriate data sets become available.

6.2 Neuropsychological tests

Exploring the relationships between neuropsychological test scores and offender type was one of the principal motivations for undertaking this research. The general single covariate random variances model of Chapter 2 was used to analyze the data available. It provided a major improvement to previous standard ANCOVA models used for this data, as the unequal variance assumption was proved to be valid in the analysis presented in section 3.1.

From a practical viewpoint the obtained results provided a major surprise. Contrary to expectations of our forensic pathology collaborators, who anticipated higher scores, corresponding to more pathological conditions, for the two sex offender groups, the mean score of the medical patient control group was the highest, though usually not significantly higher compared to the scores of some offender groups, for eight out of the twelve available tests. For the remaining four tests, their mean scores were not significantly different from those of most of the offender groups. Hence, our data analytic results certified that, at least using a single response and age as the only covariate, discriminating between the different groups is impossible. Therefore, we do not believe that neuropsychological test scores can be used to predict offender type, given the information currently available, with the obvious implications.

Nevertheless, our conclusions could be regarded as rather preliminary. Future studies could be designed to include a carefully selected random sample from the general population as a control group and to record other possible confounders for all groups. In this situation, a model of the form studied in section 2.5.4 could be used in the analysis. However, as the common characteristic of all the ANCOVA models studied in this thesis was the single response, no questions pertaining to the correlation of the scores in different tests were addressed. In that context, the extension of our ANCOVA model to multiresponse situations could be desirable, or at an additional stage, the use of an established multivariate analysis technique, such as discriminant analysis, or possible extensions of it.

6.3 Models for oil well pressures

The last subject tackled in this thesis was modelling output pressures of oil wells in a BP oilfield in Alaska. The work completed had an exploratory character, however it led to the important initial conclusion, confirmation of expert geological beliefs, of the existence of long-term correlation of the pressures of different oil wells over time.

This preliminary conclusion was reached through two distinct routes, a purely data analytic one and a second one that involved formal statistical modelling. The first method detected injector and producer wells whose pressures were highly correlated with the pressures of a given producer well of interest, according to a proposed modified information criterion, defined in (5.3), that takes into consideration the varying number of data points available as a result of the introduction of new predictors. It used a subjective application of forward selection using the coefficient of determination of the linear regression model to identify good predictors. The second method involved the development of time series (constant DLM) models, that were analyzed using MCMC, which had the same regressors as the ones identified using the data analytic method, and served as a confirmatory study for about half of the relationships detected by the latter. The combined good performance of the MCMC procedures for temporal models as well as for the ANCOVA promises well for the analysis of any linear random effects model with unknown variance components. For nonlinear random effects models, the application of MCMC methodology is more complicated (see Zeger and Karim, 1996).

Finally, to address the full complexity of the complete Alaskan data set, which is by nature multivariate, a general multiresponse spatio-temporal model was proposed in section 5.3. Its study is intended to be part of a two year externally funded research effort that should cover a number of statistical issues that arose from the single well implementations and rest of the thesis. Namely, the development of imputation methods for the missing responses and regressors, the reduction of the model by the elimination of specific system equations and related variance components testing, the simplification of the spatial process to a constant over time or climatic seasons one, the efficient applications of the Gibbs sampler, and the verification of MCMC convergence using Laplacian approximations, are all envisioned to be covered, with long term aim the development of an automated system for reservoir management.

Appendix A

Publications by O. Papasouliotis

Papasouliotis, O., and Leonard, T. (1999). The Madison Drug and Alcohol Abuse Study. Chapter 4 of *A Course in Categorical Data Analysis*, by Leonard T., with contributions from Papasouliotis O.. New York: Chapman & Hall/CRC.

Papasouliotis, O. (1999). Contributions to *A Course in Categorical Data Analysis*, by Leonard T.. New York: Chapman & Hall/CRC.

Main, I. G., Leonard, T., Papasouliotis, O., Hatton, C. G., and Meredith, P. G. (1999). One slope or two? - Detecting statistically significant breaks of slope in geophysical data, with application to fracture scaling relationships. Geophysical Research Letters, **26**, 2801-2804.

Brown, R. L., Leonard, T., Saunders, L. A., and Papasouliotis, O. (1998). The prevalence and detection of substance use disorders in inpatients of ages 18 to 49: an opportunity for prevention. Preventive Medicine, **27**,101-110.

Aitken, C. G. G., Bring, J., Leonard, T., and Papasouliotis, O. (1997). Estimation of quantities handled and the burden of proof. Journal of the Royal Statistical Society, Ser. A, **160**, 333-350.

Parton, R. M., Fischer, S., Mahlo, R., Papasouliotis, O., Jelitto, T. C., Leonard, T., and Read, N. D. (1997). Pronounced cytoplasmic pH gradients are not required for tip growth in plant and fungal cells. Journal of Cell Science, **110**, 1187-1198.

Brown, R. L., Brown, R. L., Saunders, L. A., Castelaz, C. A., and Papasouliotis, O. (1997). Physicians decisions to prescribe benzodiazepines for nervousness and insomnia. Journal of General Internal Medicine, **12**, 44-52.

Brown, R. L., Leonard, T., Rounds, L. A., and Papasouliotis, O. (1997). A two-item screen for alcohol and other drug problems. Journal of Family Practice, **44**, 151-160.

Brown, R. L., Patterson, J. J., Rounds, L. A., and Papasouliotis, O. (1996). Substance abuse among primary care patients with chronic back pain. Journal of Family Practice, **43**, 152-160.

References

Abramowitz, M., and Stegun, I. (1965). *Handbook of Mathematical Functions.* New York: Dover Publications.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, **30(A)**, 9-14.

Aitken, C. G. G., Bring, J., Leonard, T., and Papasouliotis, O. (1997). Estimation of quantities handled and the burden of proof. Journal of the Royal Statistical Society, Ser. A, **160**, 333-350.

Aitkin, M. (1997). The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, **7**, 253-261.

Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis.* New York: Wiley.

Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39-48.

Bailey, A. L. (1931). The analysis of covariance. *Journal of the American Statistical Association*, **26**, 424-435.

Banks, J. (1995). Correlation analysis of Kuparuk production data. British Petroleum internal report.

Berger, J. O., and Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, **94**, 542-554.

Berger, J. O., and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

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

Besag, J., and Higdon D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **61**, 691-746.

Billingsley, C. (1986). *Probability and Measure.* New York: Wiley.

Blattberg, R. C., and George, E. I. (1991). Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *Journal of the American Statistical Association*, **86**, 304-315.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters.* New York: Wiley.

Box, G. E. P., and Tiao, G. C. (1968). Bayesian estimation of means for the random effects model. *Journal of the American Statistical Association*, **63**, 174-181.

Box, G. E. P., and Tiao, G., C. (1992). *Bayesian Inference in Statistical Analysis.* New York: Wiley.

Brooks, S. P. (1998). Quantitative convergence assessment for Markov chain Monte Carlo via cusums. *Statistics and Computing*, **8**, 267-274.

Brooks, S. P., Dellaportas, P., and Roberts, G. O. (1997). A total variation method for diagnosing convergence of MCMC algorithms. *Journal of Computational and Graphical Statistics*, **6**, 251-265.

Brooks, S. P., and Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, **8**, 319-335.

Carlin, B. P., and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall.

Casella, G., and Berger, R. L. (1990). *Statistical Inference.* California: Wadsworth.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, **13**, 261-281.

Cohen, A. (1974). To pool or not to pool in hypothesis testing. *Journal of the American Statistical Association*, **69**, 721-725.

Coons I. (1957). The analysis of covariance as a missing plot technique. *Biometrics*, **13**, 387-405.

Cowles, M. C., and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, 883-904.

Cox, D. R., and McCullagh, P. (1982). Some aspects of analysis of covariance. *Biometrics*, **38**, 541-561.

Crook, J. F., and Good, I. J. (1982). The powers and strengths of tests for multinomials and contingency tables. *Journal of the American Statistical Association*, **77**, 793-802.

Daeid, N. N., Lynch, J., and Wideman, D. A. (1998). Statistical differences between offender groups. *Forensic Science International*, **98**, 151-156.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, **25**, 631-650.

David, H. A. (1995). First(?) occurrence of common terms in mathematical statistics. *The American Statistician*, **49**, 121-133.

Dickey, J. M. (1973). Scientific reporting. *Journal of the Royal Statistical Society, Ser. B*, **35**, 285-305.

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis* (third edition). New York: Wiley.

Evans, M., Hastings, N. A. J., and Peacock, B. (1993). *Statistical Distributions.* New York: Wiley.

Fairfield Smith, H. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics*, **13**, 282-308.

Federer, W. T. (1957). Variance and covariance analysis for unbalanced classifications. *Biometrics*, **13**, 333-362.

Finney, D. J. (1957). Stratification, balance and covariance. *Biometrics*, **13**, 373-386.

Fisher, R. A. (1932). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Garren, S. T., and Smith, R. L. (1993). Convergence diagnostics for Markov chain samplers. Technical report, University of North Carolina, Department of Statistics.

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

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelfand, A. E., Hills, S. E., Racine-Poon A., and Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association*, **85**, 972-985.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman & Hall.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with discussion). *Statistica Sinica*, **6**, 733-807.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

Gelman, A., and Speed, T. P., (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society, Ser. B*, **55**, 185-188.

Gelman, A., and Speed, T. P., (1999). Characterizing a joint probability distribution by conditionals (correction). *Journal of the Royal Statistical Society, Ser. B*, **61**, 483.

Geman, A., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith eds), pp. 169-193. Oxford: Oxford University Press.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473-511.

Gilks, W. R. (1995). Full conditional distributions. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds), pp. 75-88. London: Chapman & Hall.

Gilks, W. R., Richardson, S., and Spiegelhalter D. J. (1995). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds), pp. 1-19. London: Chapman & Hall.

Gilks, W. R., and Roberts, G. O. (1995). Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds), pp. 89-114. London: Chapman & Hall.

Good, I. J. (1991). Weight of evidence and the likelihood ratio. In *The Use of Statistics in Forensic Science* (C. G. G. Aitken and D. A. Stoney, eds), pp. 85-106. Chichester: Ellis Horwood.

Good, I. J., and Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of the American Statistical Association*, **69**, 711-720.

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

Guihenneuc-Jouyaux, C., and Robert, C. P. (1998). Discretization of continuous Markov chains and Markov chain Monte Carlo convergence assessment. *Journal of the American Statistical Association*, **93**, 1055-1067.

Harrison, J. P., and Stevens, C. (1976). Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **38**, 205-247.

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

Henderson, C. R. (1982). Analysis of covariance in the mixed model: Higher-level, nonhomegeneous, and random regressions. *Biometrics*, **38**, 623-640.

Heidelberger, P., and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, **31**, 1109-1144.

Hendrix, L. J., Carter, M. W., and Scott, D. T. (1982). Covariance analyses with heterogeneity of slopes in fixed effects. *Biometrics*, **38**, 641-650.

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association*, **60**, 806-825.

Hobert, J. P., and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**, 1461-1473.

Hsu, J. S. J. (1990). *Bayesian Inference and Marginalization.* Ph.D. thesis, University of Wisconsin-Madison.

Hsu, J. S. J. (1995). Generalized Laplacian approximations in Bayesian inference. *The Canadian Journal of Statistics*, **23**, 399-410.

Izenman A. J., Papasouliotis, O., Leonard, T., and Aitken, C. G. G. (1998). Bayesian predictive evaluation of measurement error with application to the assessment of illicit drug quantity. *Technical Report 3*, Statistical Laboratory, University of Edinburgh.

Jackson, J. L., and Bekerian, D. A. (1997). Offender Profiling: Theory, Research and Practice. Chichester: Wiley.

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, **91**, 154-166.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, **31**, 203-222.

Jeffreys, H. (1961). *Theory of Probability*, third edition. Oxford: Oxford University Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering, Series D*, **82**, 35-45.

Kalman, R. E., and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME, Journal of Basic Engineering, Series D*, **83**, 95-108.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kass, R. E., and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, **84**, 717-726.

Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 243-249.

Khuri, A. I., Mathew, T., and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models.* New York: Wiley.

Koch, G. G., Amara, I. A., Davis G. W., and Gillings, D. B. (1982). A review of some statistical methods for covariance analysis of categorical data. *Biometrics*, **38**, 563-595.

Koehler, A. B., and Murphree, E. S. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics*, **37**, 187-195.

Lane, P. W., and Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, **38**, 613-621

Lehmann, E. L. (1991). *Theory of Point Estimation.* California: Wadsworth.

Lehmann, E. L., and Casella G. (1998). *Theory of Point Estimation* (second edition). New York: Springer-Verlag.

Lehmann, E. L. (1994). *Testing Statistical Hypotheses.* New York: Chapman & Hall.

Leonard, T. (1975). A Bayesian approach to the linear model with unequal variances. *Technometrics*, **17**, 95-102.

Leonard, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika*, **63**, 69-75.

Leonard, T. (1982). Comment on "A simple predictive density function." *Journal of the American Statistical Association*, **77**, 657-658.

Leonard, T., and Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, **20**, 1669-1696.

Leonard, T., and Hsu, J. S. J. (1999). *Bayesian Methods.* New York: Cambridge University Press.

Leonard, T., Hsu, J. S. J., and Ritter, C. (1994). The Laplacian t-approximation in Bayesian inference. *Statistica Sinica*, **4**, 127-142.

Leonard, T., Hsu, J. S. J., and Tsui, K. W. (1989). Bayesian marginal inference. *Journal of the American Statistical Association*, **84**, 1051-1058.

Leonard, T., and Novick, M. R. (1986). Bayesian full rank marginalization for two-way contingency tables. *Journal of Educational and Behavioural Statistics*, **11**, 33-56.

Leonard, T., and Ord, K. (1976). An investigation of the F-test procedure as an estimation short-cut. *Journal of the Royal Statistical Society, Ser. B*, **38**, 95-98.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187-192.

Lindley, D. V. (1971). The estimation of many parameters. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 435-455. Toronto: Holt, Rinehart and Winston.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **34**, 1-41.

Liu, C., Liu, J., and Rubin, D. B. (1992). A variational control variable for assessing the convergence of the Gibbs sampler. In *Proceedings of the American Statistical Association, Statistical Computing Section*, pp. 74-78.

MacEachern, S. N., and Berliner, M. L. (1994). Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188-190.

Main, I. G. (1996). Statistical physics, seismogenesis, and seismic hazard. *Reviews in Geophysics*, **34**, 433-462.

Mardia, K. V., Kent J. T., and Bibby J. M. (1979). *Multivariate Analysis.* London: Academic Press.

Mehra, R. K. (1979). Kalman filters and their application to forecasting. In *Forecasting. Studies in the Management Sciences, Volume 12*, pp. 75-94. New York: Elsevier.

Meinhold, R. J., and Singpurwalla, N. D. (1983). Understanding the Kalman filter. *The American Statistician*, **37**, 123-127.

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

Miller, R. B., and Fortney, W. G. (1984). Industry-wide expense standards using random coefficient regression. *Insurance Mathematics and Economics*, **3**, 19-33.

Moses, J. A. Jr., Schefft, B. K., Wong, J. L., and Berg, R. A. (1992). Revised norms and decision rules for the Luria-Nebraska neuropsychological battery, form III. *Archives of Clinical Neuropsychology*, **7**, 251-269.

Mykland, P., Tierney, L., and Yu, B. (1995). Representation in Markov chain samplers. *Journal of the American Statistical Association*, **90**, 233-241.

Naylor, J. C., and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214-225.

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika*, **63**, 329-333.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society, Ser. B*, **41**, 358-367.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics. Volume 2B: Bayesian inference.* London: Edward Arnold.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **56**, 99-138.

Pole, A., West, M., and Harrison J. (1994). *Applied Bayesian Forecasting and Time Series Analysis.* New York: Chapman & Hall.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1994). *Numerical recipes in C: the art of scientific computing*, second edition. Cambridge: Cambridge University Press.

Quade, D. (1982). Nonparametric analysis of covariance by matching. *Biometrics*, **38**, 597-611.

Raftery, A. E., and Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith eds), pp. 763-773. Oxford: Oxford University Press.

Reid, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statistical Science*, **3**, 213-238.

Richardson, S., and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **59**, 731-792.

Rice, J. A. (1988). *Mathematical Statistics and Data Analysis.* California: Wadsworth.

Ripley, B. D. (1987). *Stochastic Simulation.* New York: Wiley.

Ritter, C. (1992). *Modern Inference in Nonlinear Least Squares Regression.* Unpublished doctoral thesis, University of Wisconsin-Madison.

Ritter, C., and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, **87**, 861-868.

Roberts G. (1992). Convergence diagnostics of the Gibbs sampler. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith eds), pp. 775-782. Oxford: Oxford University Press.

Roberts G. (1995). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds), pp. 45-57. London: Chapman & Hall.

Roberts G. (1996). Methods for estimating L^2 convergence of Markov chain Monte Carlo. In *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner* (D. Berry, I. Chaloner, and J. Geweke eds), pp. 373-384. Amsterdam: North-Holland.

Schruben, L. (1982). Detecting initialization bias in simulation output. *Operations Research*, **30**, 569-590.

Schruben, L., Singh, H., and Tierney, L. (1983). Optimal tests for initialization bias in simulation output. *Operations Research*, **31**, 1167-1178.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Searle, S. R. (1987). *Linear models for unbalanced data.* New York: Wiley.

Seber, G. A. F. (1977). *Linear regression analysis.* New York: Wiley.

Smith, A. F. M. (1973a). A general Bayesian linear model. *Journal of the Royal Statistical Society, Ser. B*, **35**, 65-75.

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

Smith, A. F. M., and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Ser. B*, **42**, 213-220.

Spiegelhalter, D. J., and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society, Ser. B*, **44**, 377-387.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, **64**, 29-35.

Sun, L. (1992). *Bayesian Estimation Procedures for One- and Two- Way Hierarchical Models.* Unpublished doctoral thesis, University of Toronto.

Sun, L., Hsu, J. S. J., Guttman, I., and Leonard, T. (1996). Bayesian methods for variance components models. *Journal of the American Statistical Association*, **91**, 743-752.

Tanner, M. A., and Wong W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528-550.

Theobald, C. M., Nabugoomu, F., and Talbot, M. (1999). A Bayesian approach to regional and local-area prediction from crop variety trials. Submitted to *Biometrics.*

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

Tierney, L., and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82-86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989a). Approximate marginal densities of nonlinear functions. *Biometrika*, **76**, 425-433.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989b). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**, 710-716.

Turcotte, D. L. (1992). *Fractals and Chaos in Geology and Geophysics.* New York: Cambridge University Press.

Venables, W. N., and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus.* New York: Springer-Verlag.

Urquhart, N. S. (1982). Adjustment in covariance when one factor affects the covariate. *Biometrics*, **38**, 651-660.

West, M., and Harrison J. (1997). *Bayesian Forecasting and Dynamic Models*, second edition. New York: Springer-Verlag.

Wilkinson G. N. (1957). The analysis of covariance with incomplete data. *Biometrics*, **13**, 363-372.

Yandell, B. S. (1997). *Practical data analysis for designed experiments.* London: Chapman & Hall.

Yu, B. (1994). Monitoring the convergence of Markov samplers based on estimated L^1 error. Technical Report 409, University of California at Berkeley, Department of Statistics.

Yu, B., and Mykland, P. (1994). Looking at Markov samplers through cusum path plots: A simple diagnostic idea. Technical Report 413, University of California at Berkeley, Department of Statistics.

Yu, B., and Mykland, P. (1998). Looking at Markov samplers through cusum path plots: A simple diagnostic idea. *Statistics and Computing*, **8**, 275-286.

Zeger, S. L. and Karim M. R. (1991). Generalized linear-models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79-86.

Zelen, M. (1957). The analysis of covariance for incomplete block designs. *Biometrics*, **13**, 309-332.

Zellner, A., and Min, C.-K. (1995). Gibbs sampler convergence criteria. *Journal of the American Statistical Association*, **90**, 921-927.