Inequality Constrained Hierarchical Models

ISBN: 90-393-2553-7

Inequality Constrained Hierarchical Models

Ongelijkheidsgerestricteerde Hierarchische Modellen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op het gezag van de Rector Magnificus, Prof. dr. W. H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op
donderdag 12 mei 2005, des middags te 14.30 uur

door

Bernet Sekasanvu Kato

geboren op 20 april 1969, te Kampala, Uganda

Promotor: Prof. dr. H. J. A. Hoijtink
Faculteit der Sociale Wetenschappen
Universiteit Utrecht

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.

Francis Galton, Natural Inheritance

Man is flying too fast for a world that is round. Soon he will catch up with himself in a great rear-end collision and Man will never know that what hit him from behind was Man.

James Thurber

Research is the process of going up alleys to see if they are blind.

Marston Bates

Education is what survives when what has been learnt has been forgotten.

Cornelius Otis Skinner

It is a good morning exercise for a research scientist to discard a pet hypothesis everyday before breakfast. It keeps him young.

Konrad Laurenz

Acknowledgements

This work could not have been accomplished without the assistance of a number of people.

First of all, I owe my deep gratitude to my supervisor Herbert Hoijtink. It has been an honour and a pleasure to work with him. I am very grateful to him for his invaluable guidance, great insight and stimulating discussions during the time I have been working on this thesis. Thank you Herbert!

I thank the Department of Methodology and Statistics at the University of Utrecht for providing me with the environment, equipment and other support conducive to working on the thesis. It has been great working with all the members of the department. In particular, I am grateful to Ardo van den Hout for assistance in the Dutch language.

I gratefully acknowledge members of the Bayesian reading group (Herbert, Irene and Olav) for those stimulating discussions over the years. Not to forget the many good times we had together out there at conferences and dinners.

I am indebted to my family and friends who have shown me their constant and unending support. Their love and support have kept me going and kept me grounded. Thanks are due to the 'London group' for allowing me to drop in from time to time (sometimes on short notice!). Special thanks are due to Annette for being a blessing.

To the 'good' people who meet at the home of Izaak and Dorottya. You were very encouraging. Izaak and Dorottya, many thanks for opening your doors to us. Gerald (formerly) in Groningen and Grace in Enschedde, it was nice sharing our experiences on our respective journeys to 'PhD land'. To all of you, I say *mwebale nyo banange*!

Finally, I gratefully acknowledge the financial support from the Dutch Organisation for Scientific Research (NWO) that allowed me to stay at the University of Utrecht and carry out this research.

Contents

Chapter 1

Introduction

1.1 Hierarchical Data

In applied statistics, after some empirical data have been gathered, the next step is to analyze the data and extract the maximum amount of information. This is accomplished through the construction and use of one or several statistical models. The purpose of the model(s) may be to explain how an outcome variable of particular interest is related to a set of covariates. Usually a single observation on the outcome variable is measured for each subject and then there is independence between observations. Such data are often analysed using single-level models e.g., standard regression analysis.

However, there are many types of studies in which data, both outcome and covariates are gathered in a nested or hierarchical fashion. The hierarchy implies that units at one level (also called micro level units) are grouped into or nested within higher level (or macro) units. This yields the so-called hierarchical or multilevel data. Multilevel data arise frequently in many kinds of applications. For instance, in fields such as health and education data are often collected in a nested fashion, e.g., patients within hospitals or pupils within classrooms which are in turn within schools. Over the last two decades, there has been increasing interest in developing suitable techniques and software for the statistical modelling and analysis of multilevel data, resulting in a broad class of models known under the generic name of *multilevel* models (Goldstein, 1995).

On the other hand, there are also situations where subjects are sampled cross-sectionally but then studied longitudinally, with outcomes observed at multiple time points for each subject. In this case a hierarchical data structure arises, with the measurements at different time points nested within each individual subject. Studies that involve repeated measurements or longitudinal data on each of a number of subjects are frequently used in many social, behavioural and biomedical applications. For such

data, mixed effects (that is having both fixed and random effects) models provide a useful and flexible framework in which population characteristics are modelled as fixed effects, between individual variation is modelled as random effects, and within-subject variations are accounted for by an error process. Mixed effects longitudinal models have a long history in bio-statistical theory and practice, dating at least to the seminal work of Laird and Ware (1982). Essentially the analysis of longitudinal data is a special case of multilevel modelling in which the lowest level units arise solely from the passing of time. The purpose of this thesis is to make a contribution to some issues in statistical models for the analysis of hierarchical data.

The analysis of empirical data using the mixed effects model relies on several assumptions. For example, it is usually assumed that level one residuals have a constant variance. In Chapter 2 of this thesis, test quantities which a data analyst could use to test this assumption are proposed and their performance is evaluated. Furthermore, test quantities which can be used to test the homogeneity of regression slopes in multilevel models are also presented and evaluated.

1.2 Translating Theories into Statistical Models

In some situations, researchers have some expectations or several theories concerning the outcome of their empirical research. These (competing) theories may come from the opinions of subject-area experts or may have accumulated from past studies.

Based on the theories about the outcome of a research, two approaches can be adopted. The first one is to use conventional (statistical) methods to investigate the tenability of the expectations or theories presented by the subject-area experts. This is based on the premise that statistical procedures are not mechanistic ends in themselves (i.e., fixed and rule-bound), but rather are flexible tools that should be adjusted and adapted into an appropriate means for testing a given substantive theory. This is the approach used in Chapter 3.

The second approach is to translate the competing theories into statistical models. As a result, several models are obtained for a particular data set rendering a set of competing models. The competing models are obtained by putting constraints on the model parameters. Having obtained a set of competing models, the most natural question for the researcher is: which model (or models) should I ultimately choose for the final presentation of my results? Therefore a formal model selection procedure is needed to select a model or smaller sets of models which are best supported by the data at hand.

Developments in Bayesian statistics have rendered a framework for model selection which can be used as an alternative to hypothesis testing. The main difference between (frequentist) hypothesis testing and Bayesian model selection is that for the former, p-values are used while for the latter posterior model probabilities are computed for each model under consideration. In the literature, there is a huge list of articles on procedures for model selection. This is largely due to the advances in computer technology over the past few decades which have enabled researchers to consider an increasingly wider variety of statistical models for a given data set. That model selection is a hot issue is expressed by Bayarri and Berger (2004, Section 4.2) as follows: "The story here is far from settled, in that there is no agreement on the immediate horizon as to even a reasonable method of model selection. It seems highly likely, however, that any such agreement will be based on a mixture of frequentist and Bayesian arguments".

Order restricted parameter problems arise in many settings within the social, behavioural and medical sciences. A great deal of (frequentist) literature exists on the analysis of such problems (see for example, Barlow et al., 1972; Agresti and Coull, 1996; Robertson et al., 1988 and the Journal of Statistical Planning and Inference, 2002, Volume 107, Issues 1-2 and references therein). However, frequentist analysis of these problems is often difficult or even impossible since the constraints make the usual approaches for estimation and testing break down. For example, calculation of the (restricted) maximum likelihood estimator subject to order restrictions in the parameters is often difficult, requiring specialized algorithms that are not easily generalizable. This has necessitated a vast literature on order-restricted estimators for different situations. Few articles have focused on Bayesian analysis of Order restricted parameter problems. And more so, order restricted parameter problems in hierarchical data settings have received relatively little attention.

In the Bayesian approach to data analysis the likelihood function of the data is combined with the *prior* distribution of the model parameters rendering the *posterior* distribution of the model parameters. Subsequently the *posterior* distribution is used for all the inferences. For instance, a sample from the *posterior* distribution can be used to obtain (model) parameter estimates, (posterior) standard deviations and credibility intervals. For order restricted parameter problems, a Bayesian analysis proceeds by building the constraints into the *prior* distribution (see for example, Gelfand, Smith and Lee, 1992). In Chapters 4 and 5, we will focus on Bayesian selection of competing models in the context of Inequality constrained hierarchical linear models. In Chapter 4, estimation of model parameters is also discussed.

1.3 Outline of the Thesis

The thesis is based on four papers, of which one is published, another is accepted for publication and the rest are submitted for publication. The chapters can be read separately and in any order. However there maybe some overlap in the content and notation.

In Chapter 2, we propose statistics and discrepancy measures that can be used to investigate homogeneity in a random intercept model. The proposed test quantities are evaluated using asymptotic, posterior predictive and plug-in p-values.

In Chapter 3, we focus on statistical models for structured repeated measurements. More specifically the aim of this chapter is to describe and develop models which can be used to analyse structured repeated measurements. The proposed models are illustrated using data from two experiments in Clinical psychology.

Chapter 4 is devoted to Inequality constrained hierarchical linear models. In particular these models stem from translating theories into statistical models rendering a set of competing models. Subsequently, a Bayesian approach is used for estimation and model selection. In particular a novel method, namely the *method of encompassing priors* is used for model selection.

In Chapter 5, we return to the issue of model selection. To this end, we use a posterior predictive model selection approach proposed in the literature. Subsequently, we examine the performance of this approach when applied to Inequality constrained hierarchical linear models.

Chapter 2

Testing Homogeneity in a Random Intercept Model using Asymptotic, Posterior predictive and Plug-In P-Values

Abstract[1]

In this paper three statistics and three discrepancy measures with which homogeneity in the random intercept model can be investigated will be evaluated. The first two can be used to test the homogeneity of level one residual variances across level two units and the third can be used to test whether effects should be fixed or random. Each statistic and discrepancy measure will be evaluated using asymptotic (if available), posterior predictive and plug in p-values. A simulation study will be used to investigate the frequency properties of these p-values. In the discussion it will be indicated how the results obtained for the random intercept model with one explanatory variable can be useful during the construction of general two level models.

[1]Published as: Kato and Hoijtink (2004), Testing Homogeneity in a Random Intercept Model using Asymptotic, Posterior predictive and Plug-in p-values. *Statistica Neerlandica*, **58**, 179–196.

2.1 Introduction

A crucial aspect of statistical analysis is checking the fit of models and generalizing the models to improve the fit. The main intent of this paper is to evaluate three statistics and three discrepancy measures for testing homogeneity in the random intercept model. Each statistic and discrepancy measure will be evaluated using asymptotic (if available), posterior predictive and plug-in p-values. The frequency properties of these p-values will be determined.

A random intercept model with one explanatory variable will be introduced in Section 2.2. Furthermore two methods, the Gibbs sampler and the EM algorithm, that can be used to estimate the parameters of the model will be presented. Section 2.3 introduces and explains the p-values that can be used to evaluate statistics and discrepancy measures. It will be explained how these can be computed using simulation procedures. Section 2.4 presents two statistics and two discrepancy measures that can be used to test homogeneity of level one residual variances in a random intercept model. In Section 2.5 a statistic and a discrepancy measure that can be used to test whether or not effects are random are presented. In section 2.6 a simulation study used to evaluate the frequency properties of the various p-values will be presented. Section 2.7 presents the main results of the simulation study. In Section 2.8 it will be explained how the results obtained for a random intercept model can be useful during the construction of general two level models.

2.2 The Random Intercept Model

A random intercept model is a simple multilevel model. Our motivation for using a random intercept model for the evaluation of various statistics, discrepancy measures and p-values is the feasibility of frequency evaluations of computer intensive procedures like the computation of posterior predictive and plug-in p-values (see Section 2.3).

Suppose that we have two levels of data, level one and level two, where level one units are grouped within level two units. Examples of these levels could be pupils within schools and measurements within subjects. In the sequel level one and level two units will be referred to as "persons" and "groups", respectively.

Let the random variable y_{jk} denote the response of interest for the k-th person in the j-th group and x_{jk} denote the value of the covariate for the k-th person in the j-th group. Then the random intercept model is given by

$$y_{jk} = \alpha_j + \beta x_{jk} + \varepsilon_{jk}, \tag{2.1}$$

where $j = 1, \ldots, J$ (number of level two groups) and $k = 1, \ldots, K_j$, where K_j denotes the number of persons in group j. The α_js are random intercepts that are assumed to have a normal distribution with mean μ and variance τ^2. It is assumed that the level one residuals ε_{jk} are normally distributed with mean zero and variance σ^2. The regression coefficient β (also called the slope) relates x_{jk} to y_{jk}. The data matrices containing x_{jk} and y_{jk} for $k = 1, \ldots, K_j$ and $j = 1, \ldots, J$ will be denoted by $\mathbf{x}$ and $\mathbf{y}$ respectively. The parameters of the random intercept model can be estimated by sampling from the posterior distribution of the model using Markov Chain Monte Carlo methods (Browne and Draper, 2000; Hoijtink, 2000) which renders the Expected A Posteriori (EAP) estimates of the model parameters, or by using the EM algorithm (Dempster, Laird and Rubin, 1977; Hoijtink, 2000) which renders posterior modes of the model parameters. These procedures can be found in Appendices A and B, respectively.

2.3 p-values

Investigating the compatibility of a model with the data is an important aspect of statistical analysis. In this paper three statistics and three discrepancy measures will be used to investigate the compatibility of the homogeneity assumptions of a random intercept model with the data. Each will be evaluated using p-values. When the null model has unknown parameters (for the model in (2.1) $\theta = [\mu, \beta, \tau, \sigma]$) there are several ways to compute a p-value. Bayarri and Berger (2000a) summarize the various p-values for such composite null models (but only for statistics, not for discrepancy measures) giving the main strengths and weaknesses of each p-value. In particular they introduce the conditional predictive p-value (p_{cpred}) and the partial posterior predictive p-value (p_{ppost}) and indicate their advantages from both Bayesian and frequentist perspectives. Using a number of simple examples they showed that for checking the adequacy of a parametric model, these new p-values are often superior to the plug-in (p_{plug}) and the posterior predictive p-value (p_{post}). Robins, Van Der Vaart and Ventura (2000) investigated the large sample properties of various p-values and find that the asymptotic results indeed confirm the superiority of the conditional predictive and partial posterior predictive p-values.

The primary goal of statistics must be to provide relevant practical methods for assessing the fit of complex models. Much as the new p-values proposed by Bayarri and Berger (2000a) have advantages over the existing ones, they are difficult to compute. In particular computation of the conditional predictive and partial posterior predictive p-values for the goodness of fit tests that will be considered in the framework of a random intercept model (let alone more complex multilevel models) will be very difficult. However, as will be shown, p_{plug} and p_{post} have a more than reasonable

performance in the situations considered in this paper. Indeed in their rejoinder, Bayarri and Berger (2000b, p. 1170) write "all we are really suggesting is that often p_{ppost} can be computed and should then be used, but often it will not be easily computable, and then one should then try p_{post} or p_{plug} as the best available options".

Given a null hypothesis and a test statistic the classical p-value is defined as

$$p_c = P\{T(\mathbf{y}^{rep}, \mathbf{x}) \geq T(\mathbf{y}, \mathbf{x}) \mid \mathbf{x}, H_0\}, \tag{2.2}$$

where $\mathbf{y}^{rep}$ denotes a data matrix from the null population and $T(\mathbf{y}, \mathbf{x})$ denotes the observed value of the test statistic. For the model in (2.1) (but also for more general multilevel models) $\mathbf{x}$ is ancillary. That is, both for the observed and replicated data $\mathbf{x}$ is the same. The probability in (2.2) is computed with respect to the sampling distribution of $T(\mathbf{y}^{rep}, \mathbf{x})$ under the null hypothesis H_0. The p-value in (2.2) is defined only if the null hypothesis completely specifies the parameter vector, or if the test statistic is a pivotal quantity (Berkhof, Hoijtink and van Mechelen, 2000; Meng, 1994).

The Bayesian solution (Meng, 1994) to deal with composite null hypotheses (hypotheses where not all parameters are specified) is to condition on H_0 and the observed data when a p-value has to be computed. For a statistic the posterior predictive p-value is defined as

$$p_{post} = P\{T(\mathbf{y}^{rep}, \mathbf{x}) \geq T(\mathbf{y}, \mathbf{x}) \mid \mathbf{y}, \mathbf{x}, H_0\}. \tag{2.3}$$

A discrepancy measure (Meng, 1994) is a 'test statistic' that is a function of both the data and the unknown parameter vector θ. The posterior predictive p-value for a discrepancy measure is defined as

$$p_{post} = P\{D(\mathbf{y}^{rep}, \mathbf{x}, \theta) \geq D(\mathbf{y}, \mathbf{x}, \theta) \mid \mathbf{y}, \mathbf{x}, H_0\}. \tag{2.4}$$

The probability in (2.3) is taken over the posterior predictive distribution of $\mathbf{y}^{rep}$ conditional on H_0 and $\mathbf{x}$ while the probability in (2.4) is taken over the joint posterior distribution of $\mathbf{y}^{rep}$ and θ given H_0 and $\mathbf{x}$.

The classical solution to deal with composite null-hypotheses is to replace the unknown parameters θ by their maximum likelihood estimates or posterior mode $\hat{\theta}$. Although this is usually the point of departure for the derivation of asymptotic distributions of test statistics, it can also be used to compute finite sample distributions both for statistics and discrepancy measures using the plug-in p-value (Bayarri and Berger, 2000a) which is defined as

$$p_{plug} = P\{T(\mathbf{y}^{rep}, \mathbf{x}) \geq T(\mathbf{y}, \mathbf{x}) \mid \hat{\theta}, \mathbf{x}, H_0\}, \tag{2.5}$$

in the case of a statistic T and by

$$p_{plug} = P\{D(\mathbf{y}^{rep}, \mathbf{x}, \hat{\theta}) \geq D(\mathbf{y}, \mathbf{x}, \hat{\theta}) \mid \hat{\theta}, \mathbf{x}, H_0\}, \tag{2.6}$$

in the case of a discrepancy measure D. The probabilities in (2.5) and (2.6) are taken over the distribution of $\mathbf{y}^{rep}$ given $\hat{\boldsymbol{\theta}}$ and $\mathbf{x}$.

Both p_{post} and p_{plug} can be computed using simulation. For p_{post} sample $\boldsymbol{\theta}_r$ for $r = 1, \ldots, R$ from the posterior distribution $g(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x})$. This can be done using the Gibbs sampler, and the details can be found in Appendix A. For p_{plug} compute $\hat{\boldsymbol{\theta}}$. This can be done using the EM-algorithm, and the details can be found in Appendix B. Subsequently a two-step procedure is used to obtain p_{post}:

1. Sample $\mathbf{y}_r^{rep}$ from $f(\mathbf{y}_r^{rep} \mid \boldsymbol{\theta}_r, \mathbf{x})$ for $r = 1, \ldots, R$.
2. For the statistics, the p-value is the proportion of $T(\mathbf{y}_r^{rep}, \mathbf{x})$ exceeding $T(\mathbf{y}, \mathbf{x})$. For discrepancies measures, the p-value is the proportion of $D(\mathbf{y}_r^{rep}, \mathbf{x}, \boldsymbol{\theta}_r)$ exceeding the corresponding $D(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_r)$ for $r = 1, \ldots, R$.

The same procedure can be used to obtain p_{plug} by plugging in the estimate $\hat{\boldsymbol{\theta}}$ for the unknown parameter $\boldsymbol{\theta}_r$ for $r = 1, \ldots, R$.

2.4 Testing Homogeneity of Level One Residual Variances

In this section two statistics and two discrepancy measures that can be used to test the homogeneity of level one residual variances across the groups at level two will be presented. Let σ_j^2 denote the residual variance for group j, then the null hypothesis to be tested can be written as:

$$H_0 : \sigma_1^2 = \ldots = \sigma_J^2. \tag{2.7}$$

The alternative hypothesis is that: at least one of the variances is significantly different from the rest. Bryk and Raudenbush (1992, p. 208) and Snijders and Bosker (1999, p. 126) present the following test statistic:

$$T_1(\mathbf{y}, \mathbf{x}) = \sum_{j=1}^{J} \left\{ \left[\log(S_j^2) - \frac{\sum_{j'=1}^{J} (K_{j'} - 2) \times \log(S_{j'}^2)}{\sum_{j'=1}^{J} (K_{j'} - 2)} \right]^2 \frac{K_j - 2}{2} \right\}, \tag{2.8}$$

where S_j^2 is the estimated residual variance for group j, obtained by performing an ordinary least squares regression within each group j. It should be noted that our way of presenting equation (2.8), differs from the way in which the aforementioned authors present it but it is basically the same test quantity. Equation (2.8) is obtained by adapting our notation to the test statistic given by the afore-mentioned authors. According to the afore mentioned authors $T_1(\mathbf{y}, \mathbf{x})$ has a large sample chi-square distribution with $J - 1$ degrees of freedom under the null hypothesis (2.7).

The discrepancy measure counter part of $T_1(\mathbf{y}, \mathbf{x})$ is

$$D_1(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{J} \left\{ \left[\log(S_j^2) - \log(\sigma^2) \right]^2 \frac{K_j - 2}{2} \right\}. \tag{2.9}$$

As can be seen in (2.9), the estimate of $\log(\sigma^2)$ in (2.8) is now replaced by the unknown $\log(\sigma^2)$. Since the logarithm of a variance is approximately normally distributed, then the quantity

$$A_j = \left\{ \left[\log(S_j^2) - \frac{\sum_{j'=1}^{J} (K_{j'} - 2) \times \log(S_{j'}^2)}{\sum_{j'=1}^{J} (K_{j'} - 2)} \right] \sqrt{\frac{K_j - 2}{2}} \right\} \tag{2.10}$$

is a standardized residual measure. Under the null hypothesis (2.7), the distribution of each A_j in (2.10), $j = 1, \ldots, J$ is close to a standard normal distribution. Consequently the null distribution of $T_1(\mathbf{y}, \mathbf{x}) = \sum_{j=1}^{J} A_j^2$ depends mainly on the values of K_j and not on any of the unknown parameters (see for instance, Snijders and Bosker, 1999, Section 9.4.1). It is therefore likely that in finite samples both (2.8) and (2.9) are almost pivotal quantities. The consequence is that any p-value obtained using (2.8) and (2.9) is likely to have good frequency properties.

In addition, a non pivotal statistic and a non-pivotal discrepancy measure that can be used to test (2.7) are also included in the study. The reason for doing this is two-fold. First, this is aimed at illustrating the flexibility of posterior predictive and plug-in p-values in dealing with virtually any statistic and discrepancy measure that a researcher considers to be appropriate. Second, we want to be able to differentiate between the performance of statistics, discrepancy measures, and both types of p-values in the context of a random intercept model. The non-pivotal statistic is:

$$T_2(\mathbf{y}, \mathbf{x}) = \hat{\tau}^2 \left[\log \left(\frac{S_{(1)}^2}{S_{(J)}^2} \right) \right]^2, \tag{2.11}$$

where $S_{(1)}^2$ and $S_{(J)}^2$ are respectively the smallest and largest estimated residual variances. The EM algorithm is used to obtain $\hat{\tau}^2$, the posterior mode of τ^2. The Discrepancy measure counterpart of $T_2(\mathbf{y}, \mathbf{x})$ which will be called $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ is obtained by replacing the estimate $\hat{\tau}^2$ in (2.11) by the unknown τ^2. It is clear that the sampling distributions of both test quantities $T_2(\mathbf{y}, \mathbf{x})$ and $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ depend on the parameter τ^2 and thus they are non pivotal quantities. Large values of both the statistic and the discrepancy are indicative of heterogeneity in the level one residual variances.

2.5 Testing the Homogeneity of the Slope

In multilevel models, random effects represent the variability in intercepts and slopes not explained by the covariates included in the model. It is

therefore of interest to test whether effects should be fixed or random. For example if you consider students within schools, the aim of the analysis would be, say, to establish whether some schools are more 'effective' than others in promoting students' learning and development, accounting for variations in the characteristics of students when they start school. In this case it is of interest to test whether there is a variation between schools in their intercepts and/or slopes.

Here it will only be tested whether the slope β in a random intercept model should be fixed or random. In the discussion (Section 2.8) the approach proposed will be extended to general two level models. Investigating whether the slope should be fixed or random requires to test whether the variance of the slope across level two units lies on the boundary of the parameter space:

$$H_0 : \mathrm{Var}(\beta) = 0. \tag{2.12}$$

It follows from classical likelihood theory (Cox and Hinkley, 1990, Chapter 9) that under some regularity conditions the distribution of a parameter can be approximated by a normal distribution. However, the performance of the normal approximation strongly depends on the null value of the parameter, with larger samples needed for values relatively close to the boundary of the parameter space. In the case that the null-value is a boundary value (as it is in (2.12)), the normal approximation completely fails and will not yield valid inferences (Verbeke and Molenberghs, 2000). As indicated in (2.12), here the parameter of interest is a variance and, the null value is a boundary value of the parameter space of the variance.

A solution to the problem in the previous section has been suggested by Snijders and Bosker (1999, p. 90): use a likelihood ratio test to investigate (2.12), but halve the p-value that is obtained. However it is not clear how effective this method is and so the frequency properties of their approach still have to be investigated. Indeed investigation of the frequency properties appears to be necessary because in the framework of linear mixed models for longitudinal data, Stram and Lee (1994, 1995) have shown that the asymptotic null distribution for testing hypotheses of the type where the null-value of the parameter of interest is on the boundary of the parameter space is often a mixture of chi-squared distributions rather than the classical single chi-squared distribution.

The approach to be proposed in this section does not rely on asymptotic arguments, does not have a problem with the fact that the null-value of the parameter of interest is on the boundary of the parameter space, and as is shown later has good finite sample frequency properties. To test (2.12) a statistic and a discrepancy measure that are both an estimate of $\mathrm{Var}(\beta)$ will be used:

$$T_3(\mathbf{y}, \mathbf{x}) = \frac{\sum_{j=1}^{J} (\hat{\beta}_{j,LS} - \hat{\beta})^2}{J - 1} - \frac{\sum_{j=1}^{J} SE^2(\hat{\beta}_{j,LS})}{J}, \tag{2.13}$$

and

$$D_3(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \frac{\sum_{j=1}^{J}(\hat{\beta}_{j,LS} - \beta)^2}{J-1} - \frac{\sum_{j=1}^{J} SE^2(\hat{\beta}_{j,LS})}{J}, \tag{2.14}$$

where $\hat{\beta}_{j,LS}$ is the least squares estimate of the slope in group j and $\hat{\beta}$ is the posterior mode of the common slope obtained from the EM algorithm (see appendix B). The first term on the right hand side of (2.13) and (2.14) is an estimate of $\text{Var}(\hat{\beta}_{j,LS})$. To obtain an estimate of $\text{Var}(\beta)$, the error in the estimate of β_j has to be accounted for. The latter is achieved by subtraction of the average of the variances of the estimates:

$$SE^2(\hat{\beta}_{j,LS}) = \frac{S_j^2}{\sum_{k=1}^{K_j}(x_{jk} - \bar{x}_j)^2}, \tag{2.15}$$

for $j = 1, \ldots, J$, where $\bar{x}_j$ denotes the average of the covariate in group j.

According to Bryk and Raudenbush (1992), the following statistic can be used to test (2.12):

$$T_4(\mathbf{y}, \mathbf{x}) = \sum_{j=1}^{J}(\hat{\beta}_{j,LS} - \hat{\beta})^2 / \frac{\hat{\sigma}^2}{\sum_{k=1}^{K_j}(x_{jk})^2}, \tag{2.16}$$

where $\hat{\beta}$ and $\hat{\sigma}^2$ can be estimated using the EM-algorithm (see Appendix B). Note that (2.16) is derived from the one presented by Bryk and Raudenbush (1992, p. 55). However the formula in (2.16) is less general than the one presented by Bryk and Raudenbush (1992), in the sense that it only applies to a random intercept model with one explanatory variable. The test statistic summarizes the distance between the least squares estimates of the slopes in each group and the maximum likelihood estimate of the slope, and weights each distance with an estimate of the sampling variance of $\hat{\beta}_{j,LS}$. Since (2.16) is less straightforwardly related to (2.12) than (2.13) and (2.14), and since the denominator of (2.16) is not the best estimate of the sampling variance of $\hat{\beta}_{j,LS}$ (the correct estimate is given in (2.15), see for example Wooldridge, 2000, p. 55) this statistic will only be used to illustrate that asymptotic approximations to distributions of test statistics can be inadequate. According to Bryk and Raudenbush (1992, p. 55), (2.16) is asymptotically distributed as a chi-square variable with $J - 1$ degrees of freedom. As will be shown in the sequel this asymptotic distribution is inadequate for finite sample sizes.

2.6 Design of the Study

2.6.1 Classical and Bayesian Frequency Properties

As indicated before, in this paper, three types of p-values are considered; classical p-value as in equation (2.2), posterior predictive p-value as in

equations (2.3) and (2.4), and plug-in p-value as in equations (2.5) and (2.6). In this section, each of these p-values will be denoted by the generic symbol $p_{\mathbf{y}}$ where the subscript $\mathbf{y}$ is added to denote the dependence on the data $\mathbf{y}$.

A p-value is said to have good frequency properties if p, considered as a random variable, has a uniform distribution on the interval [0,1] under the null hypothesis. That is,

$$P(p_{\mathbf{y}} \leq \alpha) = \int_{\mathbf{y}} I_{(p_{\mathbf{y}} \leq \alpha)} f(\mathbf{y} \mid \mathbf{x}, \theta) d\mathbf{y} = \alpha, \tag{2.17}$$

for all $\alpha \in [0, 1]$, where $I_{(p_{\mathbf{y}} \leq \alpha)}$ is an indicator variable defined to be one if $p_{\mathbf{y}} \leq \alpha$ and zero otherwise. Note that $f(\mathbf{y} \mid \mathbf{x}, \theta)$ is the conditional density of $\mathbf{y}$ given $\mathbf{x}$ and θ, and is proportional to the likelihood of a random intercept model with one explanatory variable (see Appendix A). As can be seen in (2.17), θ has to be known in order for the integral to be computed. This 'classical' frequency property will be investigated using a sample of 2000 data matrices from a fixed population. The population of interest will be described in the next subsection. For each of the 2000 data matrices asymptotic (where available), posterior predictive and plug-in p-values will be computed for each of the statistics and discrepancy measures introduced in Section 4 and Section 5. The classical frequency properties will be determined using constant priors for the model parameters. This ensures that the posterior distribution of the model parameters is proportional to the corresponding likelihood function (see Appendix A). An overview of the results will be provided in Section 7.

Bayesian frequency properties require a p-value to be uniform under the prior predictive distribution (Meng, 1994):

$$E_{\theta}[P(p_{\mathbf{y}} \leq \alpha)] = \int_{\theta} \int_{\mathbf{y}} I_{(p_{\mathbf{y}} \leq \alpha)} f(\mathbf{y} \mid \mathbf{x}, \theta) \Pi(\theta) d\mathbf{y} \, d\theta = \alpha, \tag{2.18}$$

where the expectation is taken over $\Pi(\theta)$, the prior distribution of the model parameters. Note that whereas the classical frequency properties require θ to be known, the Bayesian frequency properties require $\Pi(\theta)$ to be known. In truly Bayesian inference the user is required to formalize his prior knowledge with respect to the model parameters in prior distributions for these parameters. The consequence is that frequency properties can be evaluated with respect to these prior distributions instead of with respect to a fixed population. By sampling 2000 data matrices from their prior predictive distributions, Bayesian frequency properties can be determined. As will be explained below, increasingly informative prior distributions will be used. The better the distribution of any p-value computed for each of these 2000 data matrices resembles a uniform distribution, the better the Bayesian frequency properties. This evaluation of Bayesian frequency properties adds

to the results of Bayarri and Berger (2000a) who only consider 'classical' frequency properties, that is, sampling from a fixed population. An overview of the results will be given in Section 7.

2.6.2 Description of the Target Population

In the fixed population (subsequently referred to as Design 1), (2.1) adequately describes the relation between y_{jk} and x_{jk}. The parameter values chosen are σ^2=3.2, τ^2=4.8, μ=0.0, β=0.6 and $x_{jk} \sim N(0, 5)$. This corresponds to a population where, within each group, the correlation between the covariate and the response variable equals 0.6, and the intra-class correlation also equals 0.6. Each data matrix sampled from this population contains J=20 groups; with 5 groups containing 4 persons each, 5 groups containing 8 persons each, 5 groups containing 12 persons each and 5 groups containing 16 persons each.

For each data matrix the Gibbs sampler (see Appendix A) was used to sample $\theta_1, ..., \theta_R$, where $R = 20000$ after a burn-in period of 5000 iterations of the Gibbs sampler. Also, for each data matrix, $\hat{\theta}$ was estimated using the EM-algorithm (see Appendix B). Subsequently, for each data matrix, asymptotic p-values are computed for $T_1(\mathbf{y}, \mathbf{x})$ and $T_4(\mathbf{y}, \mathbf{x})$, and posterior predictive and plug-in p-values for $T_1(\mathbf{y}, \mathbf{x})$, $D_1(\mathbf{y}, \mathbf{x}, \theta)$, $T_2(\mathbf{y}, \mathbf{x})$, $D_2(\mathbf{y}, \mathbf{x}, \theta)$, $T_3(\mathbf{y}, \mathbf{x})$ and $D_3(\mathbf{y}, \mathbf{x}, \theta)$.

2.6.3 Description of Increasingly Informative Prior Distributions

To evaluate the Bayesian frequency properties, data matrices are sampled from Prior predictive distributions based on increasingly informative prior distributions for the model parameters as indicated in Table 2.1. This is done to investigate the effect of increasingly informative prior distributions on the Bayesian frequency properties of the various p-values considered in this paper.

Three (increasingly informative) designs (subsequently referred to as Design 2, 3 and 4, respectively) were used. For each design the parameters of the prior distribution can be found in Table 2.1 (note the similarity with the parameters of the fixed population). Note that
$\Pi(\sigma^2, \tau^2, \mu, \beta) = \text{Inv-}\chi^2(\upsilon_0, \sigma_0^2)\text{Inv-}\chi^2(\upsilon_1, \sigma_1^2)N(m_0, d_0^2)N(m_1, d_1^2)$, where Inv-$\chi^2$ denotes a scaled inverse chi-square distribution (see, Gelman, Carlin, Stern and Rubin, 1995, p. 474).

To sample a data matrix from its prior predictive distribution, first of all a parameter vector is sampled from $\Pi(\sigma^2, \tau^2, \mu, \beta)$. Subsequently, this parameter vector is used to simulate a data matrix according to (2.1) of J=20 groups; with 5 groups containing 4 persons each, 5 groups containing 8 persons each, 5 groups containing 12 persons each and 5 groups containing 16 persons each. Each simulated data matrix is then subjected to the

Table 2.1: Parameters of the Prior distributions.

Design No.	2	3	4
υ_0	50	100	150
σ_0^2	3.2	3.2	3.2
υ_1	5	10	15
σ_1^2	4.8	4.8	4.8
m_0	0.0	0.0	0.0
d_0^2	0.49	0.25	0.0625
m_1	0.6	0.6	0.6
d_1^2	0.0225	0.01	0.0025

computations described in the second paragraph of sub-section 2.6.2.

2.7 Results of the Study to Determine Frequency Properties

2.7.1 Frequency Properties of the Asymptotic P-values

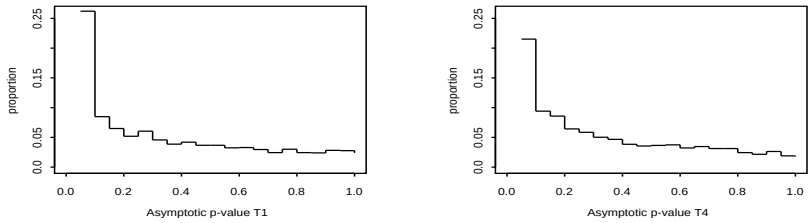

Figure 2.1: Design 1 null distributions of asymptotic p-values for $T_1(\mathbf{y}, \mathbf{x})$ and $T_1(\mathbf{y}, \mathbf{x})$.

In Figure 2.1 the null distributions of asymptotic p-values for $T_1(\mathbf{y}, \mathbf{x})$ and $T_4(\mathbf{y}, \mathbf{x})$ obtained for Design 1 (fixed population) are displayed. As can be seen the asymptotic approximations fail in finite samples with 20 groups and a total sample size of 200 persons (the requirement that $P(p_{\mathbf{y}} \leq \alpha) = \alpha$

is heavily violated). It is safe to conclude that for realistic sample sizes asymptotic p-values are rather uninformative when the goal is to detect violations of homogeneous residual variances or homogeneous regression slopes in the random intercept model. It is rather likely that this observation will not improve when the corresponding statistics are used to detect violations in more elaborate multilevel models.

2.7.2 Frequency Properties of p_{post} and p_{plug} for Statistics

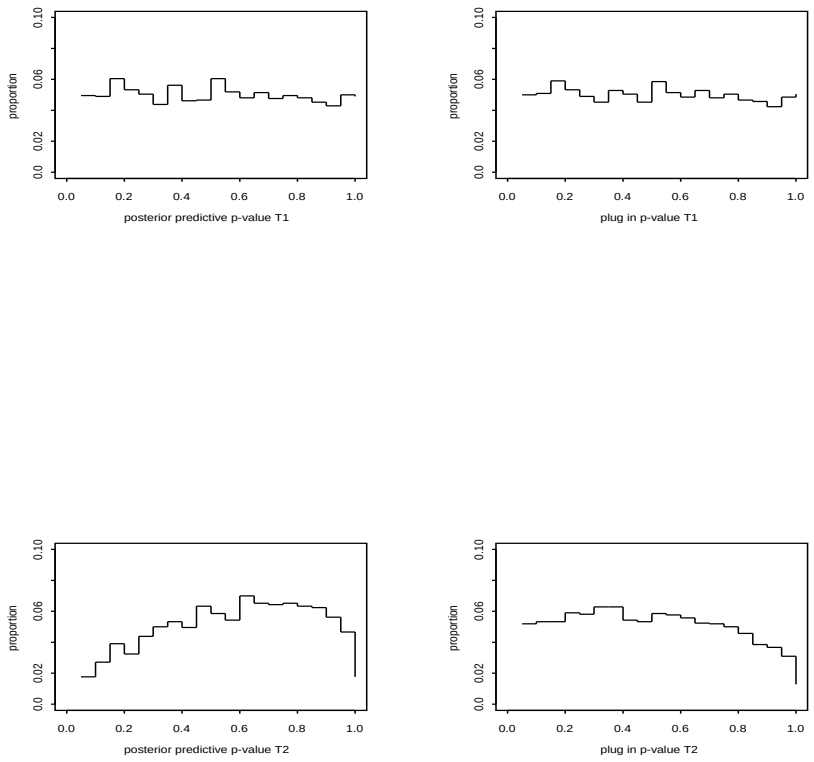

Figure 2.2: Design 1 null distributions of p_{post} and p_{plug} for $T_1(\mathbf{y}, \mathbf{x})$ and $T_2(\mathbf{y}, \mathbf{x})$.

In Figure 2.2 the null distributions of p_{post} and p_{plug} for $T_1(\mathbf{y}, \mathbf{x})$ and $T_2(\mathbf{y}, \mathbf{x})$ are displayed for Design 1 (fixed population). As can be seen, the null distribution of the almost pivotal statistic $T_1(\mathbf{y}, \mathbf{x})$ appear to be uniform both for the posterior predictive and the plug-in p-value. However, the null distribution of the 'non-pivot' $T_2(\mathbf{y}, \mathbf{x})$ is not uniform. The posterior predictive p-value is conservative ($P(p_{\mathbf{y}} \leq \alpha) < \alpha$ for small values of α), the plug-in p-value is liberal. Although the performance is substantially better than the performance of the asymptotic p-values, there is still room for improvement.

2.7.3 Frequency Properties of p_{post} and p_{plug} for Discrepancy Measures

In Figure 2.3 the null distributions of p_{post} and p_{plug} for $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ and $D_3(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ are displayed for Design 1 (fixed population). As can be seen, for both discrepancy measures and both p-values the null distributions are approximately uniform, that is $P(p_{\mathbf{y}} \leq \alpha) \approx \alpha$. Note, that the results for $D_1(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ where virtually the same as those for $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ and $D_3(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$. It appears that for fixed populations both homogeneity assumptions are best investigated using discrepancy measures. It also appears that the posterior predictive and the plug-in p-value perform equally good. It is rather likely that the same will hold when homogeneity of residual variances and homogeneity of regression slopes have to be investigated in more general multilevel models. The latter will be elaborated in the discussion. Note that this result is new, in the sense that it adds (limited to the context of multilevel models) to the results obtained by Bayarri and Berger (2000a), who consider p-values for statistics, but do not consider p-values for discrepancy measures.

2.7.4 Bayesian Frequency Properties

In Figure 2.4 the null distributions of p_{post} and p_{plug} for $T_2(\mathbf{y}, \mathbf{x})$ are displayed for Design 2, 3 and 4. As can be seen, the more informative the prior distributions and the prior family of populations, the more the null-distributions of both p-values approach uniformity. However, similar to what was noted in Section 7.2, there appears to be room for improvement.

In Figure 2.5 the null distributions of p_{post} and p_{plug} for $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ are displayed for Design 3 (the results for the other two designs and discrepancy measures were similar). As can be seen, $E_{\boldsymbol{\theta}}[P(p_{\mathbf{y}} \leq \alpha)] \approx \alpha$ for all values of α. It should be noted that in our simulation study the prior predictive distribution (see equation (2.18)) from which the data matrices were sampled and the posterior distribution of the model parameters (see equation (4.5) in Appendix A) are based on the same prior distribution of the model parameters. As indicated by a reviewer, the Bayesian frequency properties may not be so nice if the prior predictive distribution of the data

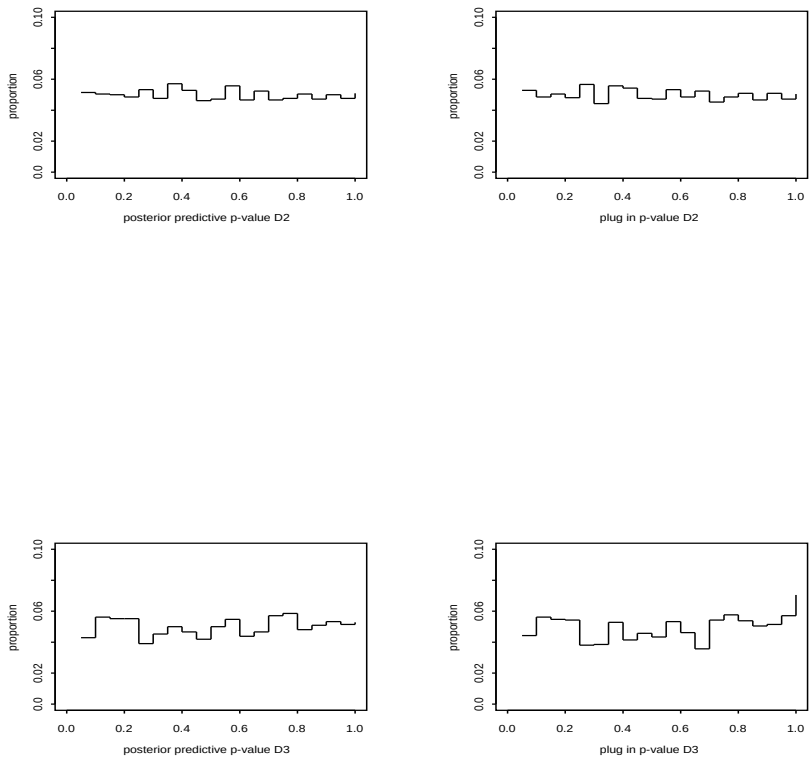

Figure 2.3: Design 1 null distributions of p_{post} and p_{plug} for $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}$ and $D_3(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$.

matrices and the posterior distribution of the model parameters are based on different prior distributions for the model parameters.

The discrepancy measures evaluated using either posterior predictive or plug-in p-values appear to have both excellent 'classical' and 'Bayesian' frequency properties. Stated otherwise, whenever homogeneity of residual variances and/or homogeneity of regression slopes have to be evaluated in multilevel models, the user is well advised to base his evaluation on discrepancy measures evaluated with either posterior predictive or plug-in p-values.

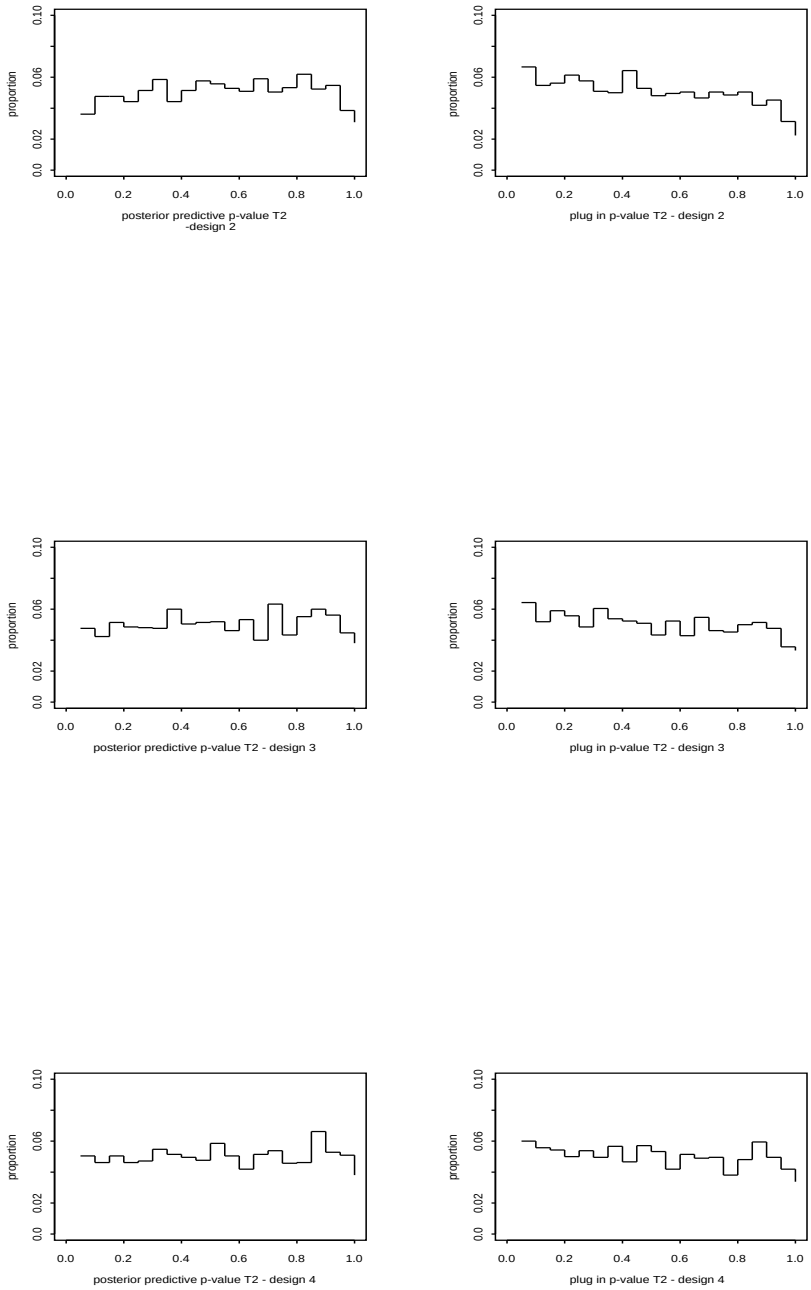

Figure 2.4: Design 2, 3 and 4 null distributions of p_{post} and p_{plug} for for $T_2(\mathbf{y}, \mathbf{x})$.

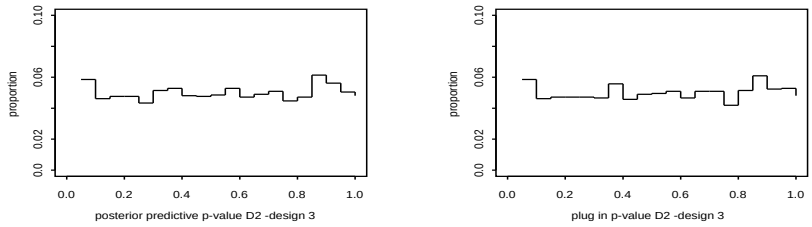

Figure 2.5: Design 3 null distributions of p_{post} and p_{plug} for $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$.

2.8 Discussion

The conclusion of this paper is that homogeneity assumptions in a random intercept model should be tested using discrepancy measures evaluated using posterior predictive or plug-in p-values. For the specific discrepancy measures $D_1(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$, $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ and $D_3(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ evaluated for a random intercept model, both the 'classical' and the Bayesian frequency properties are excellent. It may very well be that related discrepancy measures will also have good properties in more general multilevel models. However, more research is needed in order to be able to confirm this.

Testing homogeneity assumptions is an issue in multilevel modelling. Both Snijders and Bosker (1999, Chapter 6, Section 4) and Hox (2002, Chapter 4, Section 1) describe the construction of multilevel models using stepwise procedures. An inevitable step (see, for example, Hox, 2002, p. 52) is the decision which of a number of slopes should be random instead of fixed. Consider the following situation:

$$y_{jk} = \alpha_j + \beta_1 x_{1jk} + \beta_2 x_{2jk} + \beta_3 x_{3jk} + \ldots + \varepsilon_{jk}, \tag{2.19}$$

where β_1, β_2 and β_3 are fixed regression coefficients of level one variables x_{1jk}, x_{2jk} and x_{3jk}, and ... refers to level two variables (which in multilevel modelling are used to explain variation of the random intercept) and interactions between level two variables and level one variables (which in multilevel modelling are used to explain variation of the random slopes).

In this situation the following hypotheses have to be tested: $H_0 : \mathrm{Var}(\beta_1) = 0$, $H_0 : \mathrm{Var}(\beta_2) = 0$ and $H_0 : \mathrm{Var}(\beta_3) = 0$. This is easily done using a simplification of (2.14)

$$D_3^p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \frac{\sum_{j=1}^{J} (\hat{\beta}_{pj,LS} - \beta_p)^2}{J - 1}, \tag{2.20}$$

that is, an estimate of $\mathrm{Var}(\hat{\beta}_{p,LS})$ instead of an estimate of $\mathrm{Var}(\beta_p)$. In (2.20) $p = 1, \ldots, 3$ refers to the three level one predictors in (2.19). Note that, only the least-squares estimates of all regression coefficients in (2.19) are needed to be able to compute (2.20). Note also, that a direct generalization of (2.14) to general two-level models requires only a modification of (2.15).

Another step is the decision whether level one residual variances are homogeneous across level two units or not. See, for example, Snijders and Bosker (1999, pp. 126-128). For the general two-level model this can straightforwardly be done with $D_1(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$, but also with $D_2(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$ if the variance of the random intercept is replaced with the generalized variance of all random effects, that is, the determinant of the variance-covariance matrix of the random effects.

The main computational devices needed to be able to use the testing procedures described in this paper, are the Gibbs sampler and a procedure to find the maximum of the posterior distribution of multilevel models. The Gibbs sampler is implemented in MLwiN (http://www.ioe.ac.uk/mlwin/). The main other ingredients are least squares estimates of regression coefficients within each of the level two groups, and the corresponding residual variances. It should be only a matter of time that all these become available to the user of multilevel models.

Appendix A: Sampling from the Posterior Using the Gibbs Sampler

Let $\Pi(\sigma^2, \tau^2, \mu, \beta)$ be the prior distributions for σ^2, τ^2, μ and β, and $L(\mathbf{y} \mid \mathbf{x}, \sigma^2, \tau^2, \mu, \beta)$ denote the likelihood function. Assuming that the prior distributions of the parameters are independent, the posterior distribution of $\boldsymbol{\theta} = (\sigma^2, \tau^2, \mu, \beta)$ is:

$$g(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}) \propto \Pi(\sigma^2) \times \Pi(\tau^2) \times \Pi(\mu) \times \Pi(\beta) \times L(\mathbf{y} \mid \mathbf{x}, \sigma^2, \tau^2, \mu, \beta) \tag{2.21}$$

in which $L(\mathbf{y} \mid \mathbf{x}, \sigma^2, \tau^2, \mu, \beta)$ is given by

$$\prod_j \int_{\alpha_j} \{ \prod_k \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} \frac{(y_{jk} - \alpha_j - \beta x_{jk})^2}{\sigma^2}) \} h(\alpha_j \mid \mu, \tau^2) \, \partial\alpha_j,$$

where $h(\alpha_j \mid \mu, \tau^2)$ denotes the density of a normal distribution with expectation μ and variance τ^2.

The Gibbs sampler as presented by Zeger and Karim (1991) will be used to obtain a sample from $g(\cdot)$. The Gibbs sampler is an iterative procedure. Within each iteration $r = 1, \ldots, R$, ($r = 0$ denotes the initial values) each parameter is sampled (in a fixed sequence) from its posterior distribution conditional upon the current values of the other parameters.

Note that Gibbs sampling proceeds most smoothly by regarding the level 2 residuals $\alpha_1, ..., \alpha_J$ as latent variables to be sampled along with the model parameters σ^2, τ^2, μ and β. This is known as data augmentation (Tanner and Wong, 1987).

Each iteration of the Gibbs sampler consists of five steps:

1. Sample α_j for $j = 1, \ldots, J$ from its conditional posterior distribution

$$p(\alpha_j \mid \sigma^2_{r-1}, \tau^2_{r-1}, \mu_{r-1}, \beta_{r-1}, \mathbf{y}, \mathbf{x}),$$

which is a normal distribution with expectation

$$M_j = \frac{\frac{\mu_{r-1}}{\tau^2_{r-1}} + \frac{\sum_k (y_{jk} - \beta_{r-1} x_{jk})}{\sigma^2_{r-1}}}{\frac{1}{\tau^2_{r-1}} + \frac{K_j}{\sigma^2_{r-1}}} \tag{2.22}$$

and variance

$$V_j = \frac{1}{\frac{1}{\tau^2_{r-1}} + \frac{K_j}{\sigma^2_{r-1}}} \tag{2.23}$$

Note that the index $r - 1$ refers to values sampled in iteration $r - 1$.

2. Sample σ^2 from its conditional posterior distribution

$$p(\sigma^2 \mid \alpha_{1,r}, ..., \alpha_{J,r}, \beta_k, \mathbf{y}, \mathbf{x}),$$

which is an inverse chi-square distribution with degrees of freedom $\upsilon_0 + \sum_k K_j$ and scale

$$\frac{\upsilon_0 \sigma^2_0 + \sum_j \sum_k [y_{jk} - \alpha_{j,r} - \beta_{r-1} x_{jk}]^2}{\upsilon_0 + \sum_k K_j}$$

3. Sample τ^2 from its posterior conditional distribution

$$p(\tau^2 \mid \alpha_{1,r}, ..., \alpha_{J,r}, \mu_{r-1}),$$

which is an inverse chi-square distribution with degrees of freedom $\upsilon_1 + J$ and scale

$$\frac{\upsilon_1 \sigma^2_1 + \sum_j (\alpha_{j,r} - \mu_{r-1})^2}{\upsilon_1 + J}$$

4. Sample μ from its conditional distribution

$$p(\mu \mid \alpha_{1,r}, ..., \alpha_{J,r}, \tau_r^2),$$

which is a normal distribution with expectation

$$\frac{\frac{\sum_j \alpha_{j,r}}{\tau_r^2} + \frac{m_0}{d_0^2}}{\frac{1}{d_0^2} + \frac{J}{\tau_r^2}}.$$

and variance

$$\frac{1}{\frac{1}{d_0^2} + \frac{J}{\tau_r^2}}.$$

5. Sample β from its conditional posterior distribution

$$p(\beta \mid \alpha_{1,r}, ..., \alpha_{J,r}, \sigma_r^2, \mathbf{y}, \mathbf{x}),$$

which is a normal distribution with expectation

$$\frac{\frac{m_1}{d_1^2} + \frac{\sum_j \sum_k x_{jk}(y_{jk} - \alpha_{j,r})}{\sigma_r^2}}{\frac{1}{d_1^2} + \frac{\sum_j \sum_k x_{jk}^2}{\sigma_r^2}},$$

and variance

$$\frac{1}{\frac{1}{d_1^2} + \frac{\sum_j \sum_k x_{jk}^2}{\sigma_r^2}}$$

Appendix B: Using the EM Algorithm to obtain Posterior Modes

Maximization of (4.5) using the EM algorithm (Dempster, Laird and Rubin, 1977) renders the posterior mode of θ. In the E-step, for $j = 1, \ldots, J$, the expectation M_j in (2.22) and variance V_j in (2.23) of the posterior distribution of α_j given the current values σ_{r-1}^2, τ_{r-1}^2, μ_{r-1}, and β_{r-1} are computed. Note that here r refers to iterations of the EM algorithm, and not to iterations of the Gibbs sampler. In the M-step updated values of β, σ^2, μ and τ^2 are computed as follows:

1. Using (2.27), substitute β_{r-1} into (2.24), which renders an implicit equation as a function of σ^2), and iterate until σ_r^2 is very close to σ_{r-1}^2. This renders an updated value of σ^2 namely σ_r^2.

2. Using (2.26), substitute μ_{r-1} into (2.25), which renders an implicit equation as a function of τ^2, and iterate until τ_r^2 is very close to τ_{r-1}^2. This renders an updated value of τ^2 namely τ_r^2.

3. Substitute τ_r^2 and σ_r^2 into (2.26) and (2.27) respectively to obtain updated values of μ_r and β_r,

$$\sigma_r^2 = \frac{\upsilon_0\sigma_0^2 + \sum_j \sum_k [V_{j,r} + M_{j,r}^2 - 2(y_{jk} - \beta_{r-1}x_{jk})M_{j,r} + (y_{jk} - \beta_{r-1}x_{jk})^2]}{\upsilon_0 + 2 + \sum_j K_j} \tag{2.24}$$

$$\tau_r^2 = \frac{\upsilon_1\sigma_1^2 + \sum_j (V_{j,r} - 2\mu_{r-1}M_{j,r}^2 + \mu_{r-1}^2)}{\upsilon_1 + 2 + J}, \tag{2.25}$$

$$\mu_r = \frac{m_0\tau_r^2 + d_0^2(\sum jM_j)}{\tau_r^2 + Jd_0^2}, \tag{2.26}$$

$$\beta_r = \frac{m_1\sigma_r^2 + d_1^2 \sum_j \sum_k x_{jk}(y_{jk} - M_{j,r})}{\sigma_r^2 + d_1^2 \sum_j \sum_k x_{jk}^2}. \tag{2.27}$$

The EM-algorithm iterates between the E-step and the M-step until convergence of σ^2, τ^2, μ and β.

Chapter 3

Application of Multilevel Models to Structured Repeated Measurements

Abstract[1]

The linear mixed-effects model has been widely used for the analysis of continuous longitudinal data. This paper demonstrates that the linear mixed model can be adapted and used for the analysis of structured repeated measurements. A computational advantage of the proposed methodology is that there is no extra burden on the analyst since any software for linear mixed-effects models can be used to fit the proposed models. Two data sets from clinical psychology are used as motivating examples and to illustrate the methods.

[1]This chapter is accepted for publication in Quality and Quantity: International Journal of Methodology.

3.1 Introduction

In the medical, biological, behavioral and social sciences, studies are often designed to investigate changes in a specific variable which is measured repeatedly over time in the participating subjects. This allows one to model the process of change within individuals. This change could be some kind of growth for instance physical growth and learning achievement. Or it could be response to some treatment for instance reduction in blood pressure. These studies are in contrast to cross-sectional studies where the response of interest is measured only once for each individual. In general, longitudinal studies are most appropriate for the investigation of individual changes over time and for the study of factors likely to influence change.

Many approaches are available for the analysis of continuous longitudinal data. Over the last couple of decades, a lot of emphasis has been put on the hierarchical linear model (also called the linear mixed model), in which the repeated measurements are modelled using a linear regression model, with parameters which are allowed to vary over individuals, and which are therefore called subject-specific regression coefficients. See, for example Bryk et al. (1992), Goldstein (1995), Hox (2002), Longford (1993) and Snijders et al. (1999). In the literature most of the analyzes for continuous longitudinal data have been dealing with longitudinal data from designs in which measurements are taken repeatedly within one session which could last for a day, a month, a year or even a decade depending on the nature of the study at hand. Within this design, participating subjects may be randomized into a control group and one or more treatment groups. In this paper we will adapt and apply the hierarchical linear model to data obtained from designs where measurements taken repeatedly over time are structured in within-subjects conditions. In what follows these types of data will be called "Structured Repeated Measurements"(SRM) data.

The paper is organized as follows: Section 3.2 gives a brief overview of multilevel longitudinal models and then models which can be used to analyze SRM data are introduced and discussed. Section 3.3 describes two data sets from Clinical psychology. In Section 3.4, the data sets are analyzed and results presented. The paper is concluded with a discussion in Section 3.5.

3.2 Statistical background

3.2.1 General model

Data from individuals that are measured at a number of consecutive points in time can be understood as having a multilevel structure: individuals are considered as level-two units and the repeated measurements as level-one

units, so that the longitudinal measurements are nested within the individuals. For the purposes of this paper a two-level model will be considered. Considerable literature exists on the models for the analysis of two-level longitudinal data. See, for example, Catrien et al. (1998); Diggle et al. (1994); Laird et al. (1982); Snijders (1996); Verbeke et al. (1997, 2000) and Verbeke et al. (2001).

Let y_{jk} denote the response variable recorded on the jth subject at time point t_{jk} ($j = 1, \ldots, J$; $k = 1, \ldots, K_j$), where J denotes the total number of subjects and K_j is the number of repeated measurements available for the jth individual. Then a simple multilevel model for longitudinal data could be written as:

$$y_{jk} = \pi_{0j} + \pi_{1j}t_{jk} + \pi_{2j}x_{jk} + \varepsilon_{jk}, \tag{3.1}$$

at level one (measurements level), with

$$\begin{cases} \pi_{0j} &= \beta_{00} + \beta_{01}h_j + u_{0j}, \\ \pi_{1j} &= \beta_{10} + \beta_{11}h_j + u_{1j}, \\ \pi_{2j} &= \beta_{20} + \beta_{21}h_j + u_{2j}, \end{cases} \tag{3.2}$$

at level two (subjects level). This is a simple model in the sense that only linear effects have been included. More general models can be developed from this simple model by adding higher order terms e.g t^2. The π_{ij}'s, $i = 0, 1, 2$ in equation (3.1) are subject-specific regression coefficients, t is the time variable, x is a time varying covariate, h is a time invariant covariate such as gender and the treatment group, the β's are fixed effects, the u's are the random effects and ε is an error term.

The fixed effects are regression parameters which are assumed to be the same for all subjects, while the random effects differ for each subject. Equation (3.1) says that at level one, the profile of each subject can be described by fitting a function to the data of each subject separately, resulting in estimates for the subject-specific regression coefficients. Then at level two each of the subject-specific coefficients can be expressed as functions of variables (such as diagnostic groups, treatment groups, age,...) that influence subject's evolution over time. The random effects describe how subject-specific profiles deviate systematically from the average profiles. The random terms (u's and ε)in the model are assumed to be normally distributed:

$$(u_{0j}\ u_{1j}\ u_{2j})^T \sim N(0, \boldsymbol{V}),$$

and,

$$\varepsilon_{jk} \sim N(0, \sigma^2)$$

where $\mathbf{V}$ is a variance-covariance matrix and σ^2 denotes the variance of the error term. The u's are independent of $varepsilon$. Substitution of equation (3.2) into equation (3.1) renders one regression equation:

$$y_{jk} = \beta_{00} + \beta_{01}h_j + \beta_{10}t_{jk} + \beta_{11}h_jt_{jk} + \beta_{20}x_{jk} + \beta_{21}h_jx_{jk} + u_{0j} + u_{1j}t_{jk} + u_{2j}x_{jk} + \varepsilon_{jk}. \quad (3.3)$$

Taking

$$\begin{aligned} \mathbf{Q}_j &= (1 \;\; h_j \;\; t_{jk} \;\; h_jt_{jk} \;\; x_{jk} \;\; h_jx_{jk}), \\ \boldsymbol{\beta} &= (\beta_{00} \;\; \beta_{01} \;\; \beta_{10} \;\; \beta_{11} \;\; \beta_{20} \;\; \beta_{21})^T, \\ \mathbf{Z}_j &= (1 \;\; t_{jk} \;\; x_{jk}), \text{and} \\ \mathbf{u}_j &= (u_{0j} \;\; u_{1j} \;\; u_{2j})^T, \end{aligned}$$

equation (3.3) can be rewritten as a special case of the general linear mixed model (Laird and Ware, 1982; Verbeke and Molenberghs, 1997; Verbeke and Molenberghs, 2000):

$$\mathbf{y}_j = \mathbf{Q}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j, \quad (3.4)$$

where $\mathbf{y}_j$ is the K_j dimensional response vector for subject j, $\mathbf{Q}_j$ and $\mathbf{Z}_j$ are $(K_j \times p)$ and $(K_j \times q)$ dimensional matrices of known covariates , $\boldsymbol{\beta}$ is a p dimensional vector containing the fixed effects, $\mathbf{u}_j$ is a q dimensional vector containing the random effects, and $\boldsymbol{\varepsilon}_j$ is a K_j dimensional vector of residual components. Put another way, the theory of linear multilevel models is embodied in that of linear mixed models. It should be noted that the matrix of covariates $\mathbf{Z}_j$ will usually be a sub-matrix of the fixed effects matrix $\mathbf{Q}_j$.

As mentioned in the introduction, the prime advantage of a longitudinal study is its effectiveness for studying change. Longitudinal studies can be used to distinguish changes over time within subjects from differences among subjects in their baseline levels. The repeated measurements of subjects can be used to make inferences about the effects of explanatory variables and background characteristics of individuals on the response variable. Of primary interest is the estimation of changes over time and testing whether these changes depend on background characteristics of the subjects. Furthermore one can predict individual response trajectories and also make inferences about the variability between subjects. In fact in a longitudinal study it is possible to distinguish *within* subjects variability from *between* subjects variability. Within subjects variability arises because for most outcomes, there is variability across subjects due to the influence of characteristics such as genetic make-up, environmental exposures, personal habits and so on and so forth. Consequently even if, for example, one administers the same drug to different subjects, inherently

some subjects are high responders while others are low responders. On the other hand within subjects variability stems from the fact that the repeated measurements of a subject reflect the biochemical processes operating within that subject. This variation results in a correlation between pairs of measurements from the same subject. Consequently in a longitudinal study, when making inferences the naturally occurring differences among and within subjects have to be acknowledged.

3.2.2 Models for Structured Repeated Measurements

In this section, notation for SRM data is introduced and corresponding models for the analysis of such data are developed. In line with the two data sets to be analyzed in this paper, a two level framework will be considered. Consider a sample consisting of J subjects and $s = 1, \ldots, S$ within subject conditions such that within each condition $l = 1, \ldots, L_{js}$ measurements are obtained for each subject. Let y_{jsl} denote the value of the lth measurement ($l = 1, \ldots, L_{js}$) recorded for the jth subject in condition s. Then similar to Section 2.1, for each subject j there are K_j measurements but these K_j measurements are structured within the $s = 1, \ldots, S$ conditions. Put another way, $\sum_{s=1}^{S} L_{js} = K_j$.

In developing models for SRM data, the variables t, x and h as defined in Section 3.2.1 will be used but the index notation will change to take into account the fact that the measurements for the time varying covariates and response variable are structured in within-subject conditions. As mentioned in the introduction, a longitudinal study is one in which participants are followed through time, rather than observed at only one time. For SRM data, there are two possible scenarios for the time trend. In what follows, these two scenarios will be discussed and corresponding models will be presented. Assuming that the data of each subject j can be approximated by a quadratic function of time, an SRM model is given by:

$$y_{jsl} = \pi_{0js} + \pi_{1js} t_{jsl} + \pi_{2js} t_{jsl}^2 + \pi_{3js} x_{jsl} + \varepsilon_{jsl}, \tag{3.5}$$

where the π_{ijs} (i=0, 1, 2, 3) are subject-specific regression coefficients which may depend on the within subject condition s. So for each subject one needs to estimate an intercept, an x effect, a linear time and a quadratic time effect for each condition $s = 1, \ldots, S$. The term ε_{jsl} is an error term which is assumed to have a normal distribution with mean 0 and variance σ^2.

3.2.2.1 Case 1

The first scenario is the one consisting of one time trend across the within subject conditions. The Speech data introduced in Section 3.3.1 fall under this single time trend scenario.

Interest lies in finding how the within-subject conditions affect the evolution of subject's responses over time. Depending on specific research questions the π_{ijs} in (3.5) can be expressed as functions of within-subject conditions. For instance one can have:

$$\begin{cases} \pi_{0js} &= \beta_0 s + \beta_1 h_j + u_{0j}, \\ \pi_{1js} &= \gamma_0 s + \gamma_1 h_j + v_{0j}, \\ \pi_{2js} &= \alpha_0 s + \alpha_1 h_j + w_{0j}, \\ \pi_{3js} &= \lambda_0 s + \lambda_1 h_j + r_{0j}, \end{cases} \tag{3.6}$$

where, the β's, α's, γ's and λ's are fixed effects while the u's, v's, w's and r's are random effects. This yields equations (3.5) and (3.6) at level one (measurements level) and level two (subjects level) respectively. Note that (3.6) is valid only if the within-subject condition s is continous. If s is discrete, then (3.6) has to be written differently. An example of how to deal with the discrete case will be presented in Section 3.4.1 (Analysis of Speech data).

It should be noted that in the development of models for SRM data, subject specific parameters are expressed as functions of both between-subject conditions (e.g h in (3.6)) and within-subject conditions (e.g s in (3.6)). This is indeed the point of departure from the classical multilevel models where the subject specific parameters are expressed as functions of only between-subject conditions.

3.2.2.2 Case 2

On the other hand one can have multiple time trends which are such that intervals over which measurements are taken are the same within each condition. This would for instance occur in a situation where you have several 'treatment' sessions which are sequential. The researcher would then be interested in investigating research questions pertaining to the changes in specific variables as one proceeds from the first session to the last session. The Gilles de La Tourette's Syndrome data introduced in Section 3.3.2 fall under the multiple time trend scenario.

In order to answer interesting research questions the π_{ijs} in (3.5) can be re-written as linear functions of s. For instance expressing each of the π_{ijs} in (3.5) as linear functions of s yields

$$\begin{aligned} y_{jsl} = (A_{0j} + A_{1j}s) + (B_{0j} + B_{1j}s)t_{jsl} + (C_{0j} + C_{1j}s)t_{jsl}^2 \\ + (D_{0j} + D_{1j})x_{jsl} + \varepsilon_{jsl}. \end{aligned} \tag{3.7}$$

Consequently instead of estimating $4 \times S$ (given S within subject conditions) parameters for each subject, one estimates 8 parameters. Depending on specific research questions to be investigated the π_{ijs}'s can also be

expressed as functions of within-subject conditions and other plausible covariates. In equation (3.7), linear restrictions are imposed on the intercept, x term, linear time and quadratic time effects. For instance the term $A_{0j}+A_{1j}s$ in (3.7) stipulates that the average intercept π_{0js} from each session lie on a straight line with intercept A_{0j} and slope A_{1j}. Other types of restrictions are also possible. Note that in (3.7), the parameters A, B, C and D have the subscript j meaning that they differ for each subject. Analogous to section (2.2.1), one obtains equation (3.7) at level one with,

$$\begin{cases} A_{0j} &= \beta_{00} + \beta_{01}h_j + u_{0j}, \\ A_{1j} &= \beta_{10} + \beta_{11}h_j + u_{1j}, \\ B_{0j} &= \gamma_{00} + \gamma_{01}h_j + v_{0j}, \\ B_{1j} &= \gamma_{10} + \gamma_{11}h_j + v_{1j}, \\ C_{0j} &= \alpha_{00} + \alpha_{01}h_j + w_{0j}, \\ C_{1j} &= \alpha_{10} + \alpha_{11}h_j + w_{1j}, \\ D_{0j} &= \lambda_{00} + \lambda_{01}h_j + r_{0j}, \\ D_{1j} &= \lambda_{10} + \lambda_{11}h_j + r_{1j}, \end{cases} \tag{3.8}$$

at level two and as can be seen the subject specific parameters $\pi_{0js}, \ldots, \pi_{3js}$ are expressed as functions of both between-subject conditions and within-subject conditions. Just like in (3.6), the β's, α's, γ's and λ's in (3.8) are fixed effects while the u's, v's, w's and r's are random effects. Note the similarity between equations (3.1) and (3.5) and equations (3.2), (3.6) and (3.8) respectively.

3.2.2.3 Generalization

In this section it will be shown that the model developed for SRM data is also a special case of the general linear mixed model (3.4). Substitution of equation (3.8) into equation (3.7) and suppressing the subscripts of t, x and h renders one regression equation:

$$\begin{aligned} y_{jsl} = \beta_{00} + \beta_{01}h + +\beta_{10}s + \beta_{11}hs + \gamma_{00}t + \gamma_{01}ht + \gamma_{10}st + \\ \gamma_{11}hst + \alpha_{00}t^2 + \alpha_{01}ht^2 + \alpha_{10}st^2 + \alpha_{11}hst^2 + \lambda_{00}x + \\ \lambda_{01}hx + \lambda_{10}sx + \lambda_{11}hsx + u_{0j} + u_{1j}s + v_{0j}t + \\ v_{1j}st + w_{0j}t^2 + w_{1j}st^2 + r_{0j}x + r_{1j}sx. \end{aligned} \tag{3.9}$$

Writing

$$\begin{aligned} \mathbf{Q}_j &= (1\ h\ s\ hs\ t\ ht\ st\ hst\ t^2\ ht^2\ st^2\ hst^2\ x\ hx\ sx\ hsx), \\ \theta &= (\beta_{00}\ \beta_{01}\ \beta_{10}\ \beta_{11}\ \gamma_{00}\ \gamma_{01}\ \gamma_{10}\ \gamma_{11}\ \alpha_{00}\ \alpha_{01}\ \alpha_{10}\ \alpha_{11}\ \lambda_{00}\ \lambda_{01}\ \lambda_{10}\ \lambda_{11})^T, \\ \mathbf{Z}_j &= (1\ s\ t\ st\ x\ sx), \text{ and} \\ \delta_j &= (u_{0j}\ u_{1j}\ v_{0j}\ v_{1j}\ w_{0j}\ w_{1j})^T, \end{aligned}$$

renders

$$\mathbf{y}_j = \mathbf{Q}_j\theta + \mathbf{Z}_j\delta_j + \varepsilon_j, \tag{3.10}$$

which has the same form as (3.4). A similar result can be obtained if (3.6) is substituted into (3.5). Consequently the model parameters in a Structured Repeated Measurements model can be estimated with any software designed for fitting linear mixed models.

Structured Repeated Measurements can be used to answer research questions relating to the effect of within-subject conditions and between-subject conditions on the response variable. Note that the dependence of the response variable on the within-subject and between-subject conditions is investigated by estimating the regression coefficients in (3.5) and (3.6) for case 1 (see Section 3.2.2.1) or in (3.5), (3.7) and (3.8) if one considers case 2 (see Section 3.2.2.2).

3.2.2.4 Centering

In multilevel modelling continuous predictor variables are usually centered. The advantage of centering is two-fold. First it makes interpretation of the intercept term easier because then the intercept is the predicted value for a subject that has average values for each explanatory variable. Second, centering predictor variables can also reduce the chances of numerical errors in the estimation methods used for fitting multilevel models.

As will be seen in the sequel, in this paper all continuous predictors will be centered around their means (or medians).

3.2.2.5 Likelihood Ratio Test

A classical statistical test for the comparison of nested models is the likelihood ratio (LR) or Deviance test. In applications of multilevel models this test is mainly used for multi-parameter tests and for tests about the random effects. It works as follows. When parameters of a statistical model are estimated by the maximum likelihood method, the estimation also provides the likelihood which can then be used to obtain the *deviance* (equal to -2*log-likelihood). The deviance can be regarded as a measure of how well the model fits the data. However in most statistical models, the values of the deviance cannot be interpreted directly, only differences in deviance values for several nested models fitted on the same data set can be interpreted.

Suppose that two nested models M_1 and M_2 with n_1 and n_2 ($n_2 > n_1$) parameters respectively are fitted to one data set. That is, M_2 can be considered as an extension of M_1 with n_2 - n_1 extra parameters. Considering M_1 as the null model and M_2 as the model under the alternative hypothesis,

and indicating their deviances by D_1 and D_2, respectively, the difference D_1 - D_2 can be used as a test statistic. Under the null, this test statistic is asymptotically distributed as a Chi square distribution with n_2 - n_1 degrees of freedom. Then depending on the p-value obtained from the test, one can decide whether to reject or not to reject the null model.

Furthermore approximate t-statistics can be used for testing hypotheses about each of the fixed effects in the vectors β (in (3.4)) and θ (in (3.10)). For each single parameter, say, β_p , the hypothesis H_0: $\beta_p = 0$ can be tested by using the test statistic $T(\beta_p) = (\widehat{\beta_p})/S.E(\widehat{\beta_p})$. This test statistic can be used for one-sided as well as two-sided tests. According to Snijders and Bosker (1999, section 6.1), $T(\beta_p)$ has approximately a t-distribution with N-r-1 degrees of freedom where N and r are the total number of level-one units and total number of explanatory variables respectively. If N-r is large enough (say larger than 40), the t-distribution can be replaced by a standard normal distribution. The p-value obtained from the test can then be used to decide on whether to reject or not to reject the null hypothesis.

3.3 Data

3.3.1 The Speech data

In Clinical Psychology, many studies thus far have focused on the content and the organization of speech during exposure to fearful stimuli (see for example Amir et al., 1998; Foa et al., 1995; Pennebaker, 1993; Pennebaker et al., 1996; Pennebaker et al., 1988; Pennebaker et al., 1997; Van Minnen et al., 2002). The data introduced in this section are from a study to investigate the effect of fear on paralinguistic aspects of speech (content-filtered, non-semantic, non-grammatical aspects of the voice such as pitch and rate of speech). The question was whether activation of fear memories could be identified by specific physiological and behavioral responses as reflected in paralinguistic aspects of speech. Full details of this study are reported in Hagenaars et al.,(in press). When fearful memories are re-collected and recounted aloud, paralinguistic aspects of speech can reveal information about the emotional and physiological processes that take place during the re-collection of that memory. Consequently, speech could be a useful tool to provide information about emotional processing of fear during exposure to fearful stimuli. This is relevant for psychotherapy because speech is a central part of the treatment.

Twenty five people took part in the study all of whom met the DSM-IV criteria for Panic Disorder with Agoraphobia (APA, 1994). At the time of the study all of them were undergoing cognitive behavioral therapy at an outpatient clinic specialized in the treatment of anxiety disorders or had two intake sessions and were on the waiting list of that same clinic. In

order to avoid any possible differences in speech due to gender or cultural differences, all the participants were female and native speakers of dutch. Their age ranged from 21 to 63 years, with mean value equal to 38.12 years and a standard deviation of 11.32 years.

The experiment consisted of two modules namely Autobiographical Talking and Script Talking and each module consisted of two emotional conditions, Fearful and Happy. In the Autobiographical Talking module, each subject was asked to describe, in six minutes, two recent experiences (three minutes for each experience); one in which they had a serious panic attack (Autobiographical Talking Fearful) and another in which they had felt happy (Autobiographical Talking Happy). On the other hand, in the Script Talking module, each subject had to read aloud two scripts, one in which a fearful agoraphobic situation was described (Script Talking Fearful) and another script describing a happy situation (Script Talking Happy). The topics and grammar were kept the same except for a number of selected words. To familiarize the subjects with the tasks and to reduce baseline anticipatory tension, the participants were asked to read aloud a Neutral Script with the same length as the scripts used in Script Talking Fearful and Script Talking Happy. At the end of the experiment the same Neutral Script was read again by all subjects. In summary, each of the participants was subjected to six experimental conditions namely 'Script Talking Neutral', 'Autobiographical Talking Happy', 'Autobiographical Talking Fearful', 'Script Talking Reading Happy', 'Script Talking Fearful' and 'Script Talking Neutral'. In what follows these will be referred to as treatment conditions. The Autobiographical Talking conditions lasted 3 minutes and the Script Talking conditions lasted 2 minutes.

There were two variables of interest. The first one is subjective feelings of anxiety where a subject is required to note the degree of fear at a given moment. This is measured on the Subjective Unit of Distress (SUD) scale which is a visual analogue scale on which participants can place a number from 0 (indicating that the participant had no fear) to 10 (indicating that the participant was in total panic). The other variable is a speech measurement. From the speech signals, a computer program called PRAAT (Boersma and Weenick, 1996) is used to extract certain paralinguistic measures from digitally recorded spoken sound. The measurement used for the speech was the variability of the pitch (variability of the highness of the voice as we hear it) and was measured in Hertz (Hz). This variable will be denoted by $pitchvariability$. For each condition, the participants' subjective fear and pitch variability were recorded. Subjective fear was measured at the beginning and at the end of each condition as well as at one minute intervals during the condition. Pitch variability was measured every 0.1 seconds. 10 second intervals around the SUD measures (and 20 seconds after the first SUD and before the last SUD) were selected and the mean pitch variability for each of these intervals was calculated. Consequently each of the Script

Talking conditions renders 3 values of pitch variability and subjective fear respectively. On the other hand the Autobiographical Talking conditions render 4 values of pitch variability and subjective fear respectively. Thus on a 20 minute time trend, this yields 20 values of pitch variability and 20 values of subjective fear.

3.3.2 The Gilles de La Tourette's Syndrome Data

The data introduced in this section are from an exposure and response prevention therapy study on the treatment of Gilles de la Tourette's Syndrome (GTS). GTS is a tic disorder which occurs in 4-5 out of 10000 persons (APA, 1994). A tic is a sudden, rapid, recurrent, stereotyped motor movement or vocalization. Research had indicated that more than 90% of GTS patients experience uncomfortable sensations and urges which they try to get rid of by intentionally carrying out a tic (Leckman et al., 1993). The possible existence of a sequence of unpleasant sensations followed by a tic may mean that interventions can be developed, the purpose of which is to interrupt the sequence, thus preventing the tics. Various behavioral therapeutic strategies are used in the treatment of GTS. Recently a new treatment strategy involving exposure and response prevention (ER) has been developed. ER is based on the thought that if a subject is exposed for a prolonged period to the sensations, these sensations would eventually diminish, a process known as habituation.

Exposure therapy is a well known behavior intervention that has proven to be effective for the treatment of anxiety disorders, for instance obsessive compulsive disorder (OCD). There are some parallels between anxiety disorders and ticking behavior: in both disorders there is a stimulus that provokes a negative feeling (anxiety in OCD and sensory sensations in GTS) to which the patient responds with behavior that reduces the negative feeling (compulsive behavior in OCD and ticking in GTS). Typically in exposure therapy, the sequence between stimulus and response is interrupted by exposing the patient for a prolonged period of time to the stimulus (e.g sensory sensation in the case of GTS) and persuading them to avoid performing the response (ticking). As a result, habituation to the premonitory sensations may take place resulting in the amount of tics exhibited.

Previous studies (see for example Hoogduin et al., 1997) have shown that exposure therapy is also of value in the treatment of GTS. The purpose of the present study was to find out whether in GTS patients, habituation to the sensory sensations is a possible underlying mechanism of change in exposure therapy as it seems to be the case in anxiety disorders. To investigate the habituation hypothesis a total of 19 GTS patients were studied. The patients participated in a controlled comparative effect study between ER and habit reversal, another behavioral strategy used in the treatment of GTS (Verdellen et al, in press). For investigating the habituation hypothe-

sis, patients had to be familiar with sensory sensations. Of the participating subjects 15 were males and 4 were females. Nine patients were aged less than 18 years. The mean age of the patients was 23 with a standard deviation of 13.3.

The ER procedure consisted of 10 treatment sessions each consisting of 2 hours, during which the patients were encouraged not to tic whenever they felt some unpleasant sensations. Each treatment session was executed according to a treatment protocol for ER (Hoogduin et al., 1997; Verdellen et al., 1994). To measure any changes in the severity of the sensations, for each subject prior to treatment four sensations corresponding to the four most frequent tics were selected. Global tension was added as a fifth item for each subject. Subjects were then asked to rate the severity of these five items on a visual analogue scale from 0-4. These are used to compute a total sensation severity score ranging from 0-20. Before and during sessions, at 15 minute intervals, therapists asked the subjects to rate the severity of the five items as described above. For each session there were nine rating time points. Furthermore the therapists registered the incidence of tics that had not been suppressed. So the data consist for each of the patients, a total sensations severity score for each time point within each session and the total number of tics per time point exhibited in each session.

3.4 Data Analysis and Results

3.4.1 Analysis of Speech Data

The objective of the analysis is to investigate the effect of subjective anxiety on variation in pitch, and to test whether this effect depends on subject's age and within-subjects conditions. This is of importance for several reasons, both theoretical and practical. Firstly, recent fear theories in clinical psychology theories are mostly two dimensional. That is, they recognize two levels of information processing: a directly, automatically responding system (seated in the amygdala), and a somewhat slower, more cognitive, verbal, appraisal system (seated in the hippocampus amongst others) (Brewin et al., 1996; Brewin, 2001; Lang et al., 1998; Lang et al., 2000; LeDoux, 1996). Information about fear responses is often gained from self report, meaning the first, automatically responding system is not looked at. Speech could be an easy way to provide information about this system, because it reflects physiological (respiration) and behavioral (muscle tension) mechanisms. This is not only interesting theoretically (because it makes it possible to look at the interaction between the two response systems), but could also be useful as a tool in clinical practice. Patients with anxiety orders could easily be monitored in their therapy process, by looking at the second system (subjective measures), but also by looking at the first sys-

tem (measuring speech). Because there still is a large amount of relapse in patients with anxiety disorders, measuring both systems could provide more information about the fear processing stage they are in and therefore prevent relapse.

First, for these data the within-subjects conditions are the $neutral$, $happy$ and $fearful$ conditions while subject's age is a between subjects condition. Second, these data fall under case 1 (see Section 3.2.2.1) since we have a single time trend. Second, we have balanced data since all subjects have the same number of repeated measures, taken at time points which are also the same for all subjects. So the coding of $time$ does not depend on subject j and condition s, that is, $t_{jsl} = t_l$ for $l = 0, \ldots, 19$.

Let $Varpitch_{jl}$ and Sud_{jl} denote the variation in speech (variability of the pitch) and subjective anxiety respectively of the jth subject measured at t_l ($l = 0, \ldots, 19$) minutes. It should be emphasized that the 20 measurements of $Varpitch$ and sud respectively for each subject were recorded under 3 different conditions, that is, the measurements are structured within 3 conditions. Furthermore Age will denote a subject's age. For all the analyzes, the predictors Sud and $time$ across all within-subjects conditions will be centered at their median values which are 5 units and 10 minutes respectively.

The first model to be considered is a linear trend model consisting of random intercepts and random slopes for time and subjective anxiety (an initial analysis showed that the quadratic time trend was not significant):

$$Varpitch_{jl} = \pi_{0js} + \pi_{1js}(t_{jl} - 10) + \pi_{2js}(Sud_{jl} - 5) + \varepsilon_{jl} \tag{3.11}$$

with,

$$\begin{cases} \pi_{0js} &= \beta_{00} + u_{0j}, \\ \pi_{1js} &= \beta_{10} + u_{1j}, \\ \pi_{2js} &= \beta_{20} + u_{2j}, \end{cases} \tag{3.12}$$

At this point it is worthwhile to note that (3.11) is a simplified version of the general equation (3.5). This model (hereafter referred to as model 1) investigates how for each subject, the subjective anxiety affects variation in speech controlling for time. Note that this model does not take into account the fact that the measurements were recorded from several experimental conditions. It is expected that the higher the subjective anxiety, the lower the variation in pitch controlling for time. The deviance, parameter estimates and standard errors obtained from fitting this model are displayed in Table 3.1. In this paper all tests will be based on a 5 percent level of significance. As can be seen in Table 3.1, the parameter β_{00} is significant. Furthermore using a one sided test to test the hypothesis $H_0 : \beta_{20} = 0$ against the alternative $H_1 : \beta_{20} < 0$, renders a p-value of 0.021. Therefore the null hypothesis

is rejected in favor of the alternative hypothesis. Since β_{20} = -0.28, these results suggest that for each unit increase in subjective anxiety, the variation in speech decreases by 0.28 Hertz.

The next step is to investigate whether the effect of subjective anxiety on variation in pitch depends on subject's age and experimental condition of subject controlling for time. This renders a model given by equation (3.11) at measurements level and

$$\begin{cases} \pi_{0js} &= \beta_{00} + \beta_{01}Hapneu_j + \beta_{02}Fearneu_j + \beta_{03}Age_j + u_{0j}, \\ \pi_{1js} &= \beta_{10} + u_{1j}, \\ \pi_{2js} &= \beta_{20} + \beta_{21}Hapneu_j + \beta_{22}Fearneu_j + \beta_{23}Age_j + u_{2j} \end{cases} \tag{3.13}$$

at subjects level. $Hapne$ and $Fearneu$ are indicator variables defined to be one if the condition is a happy condition or a fearful condition respectively, and zero otherwise. Since we are dealing with discrete within-subjects conditions, indicator variables have been used to compare the within-subjects conditions. For instance in the first term of (3.13), the quantity

$$\beta_{00} + \beta_{01}Hapneu_j + \beta_{02}Fearneu_j$$

is used in place of the quantity β_{0s} in the first term of (3.6). Although the variable $time$ is not significant in model 1, it is still included in the extended model, because the aim is to investigate the effect of $subjective\ anxiety$, while controlling for the effect of $time$.

Fitting the extended model followed by an investigation of the estimates for the parameters in the model showed that no significant interaction seems to be present between subject's age and subjective anxiety (β_{23} was not significantly different from zero). There was also no significant interaction between experimental condition and subjective anxiety (both β_{21} and β_{22} were not significantly different from zero). Furthermore the variance of u_{2j} and its covariances with u_{0j} and u_{1j} respectively were all estimated to be zero. These non-significant terms were consequently removed from the model and a new model (hereafter referred to as model 2) fitted. The results of the model fit are shown in Table 3.1. As can be seen in Table 3.1, model 2 has 3 parameters more than model 1. The difference between their deviances is 44.531. On 3 degrees of freedom, this yields a p-value very close to zero implying that model 2 is a significant improvement to model 1. Second, the parameters β_{00}, β_{01}, β_{02} and β_{03} are significant. This implies that a subject's intercept (which equals the variability of the pitch at time = 10 and sud = 5) is influenced by experimental condition and age. Furthermore it can be seen from Table 3.1, that the parameter β_{10} is not significant. Using a one sided test to test the hypothesis $H_0 : \beta_{20} = 0$ against the alternative $H_1 : \beta_{20} < 0$, yields a p-value of 0.069 which is borderline. This suggests that inclusion of the within-subjects conditions; $Happy$, $Fearful$ and $Neutral$ into the model renders Sud less significant.

This is not unexpected since subjective anxiety and condition (happy, fearful or neutral) are correlated. Furthermore it may well be that in model 2, β_{20} is not significant because there is not enough power to reject the null hypothesis because of the small sample size (only 25 subjects).

Age has a strong main effect with a regression coefficient equal to 0.14: if a subject's age increases by one year then the variation in speech increases by 0.14 Hertz. Put another way the older a subject is the higher the variation in speech. Regression coefficients of β_{00} = 23.63, β_{01} = -1.854, and β_{02} = -3.749 for the other main effects imply that in the Neutral, Happy and Fearful conditions, the average variation in speech for a subject whose subjective fear of anxiety is at 5 (in the middle of the scale) are 23.63, 21.78 and 19.88 Hertz respectively. In other words a subject's variation in speech is lowest when subject is in fearful condition, higher in a Happy condition and highest in a Neutral condition.

To conclude, the above analyzes suggest that only main effects of age and condition are needed to explain variation in speech, that the variation in speech does not change over time and subjective anxiety does not affect one's variation in speech if condition is taken into account.

The results have important implications for clinical psychology. Speech analyses have proved to be a useful tool in measuring different experimental conditions, reflecting different emotional states. The automatically responding fear system proved to be reflected in speech, which opens the door for speech analyses as a means to measure the automatically occuring fear response in for example anxiety disordered patients. In other words, by measuring speech, information can be gained about the kind of memory that is activated. Improvement of patients with anxiety disorders in behavior therapy can be measured easily using the two dimensional model: subjective measures and speech measures provide a rather complete picture of the stage of fear processing that the patient is in.

Table 3.1: Speech data. Parameters estimates and standard errors for the models.

		Model 1	Model 2
Effect	Parameter	Estimate(S.E)	Estimate(S.E)
Intercepts:			
	β_{00}	26.823(0.909)	23.630(2.593)
Hapneu	β_{01}	-	-1.854(0.578)
Fearneu	β_{02}	-	-3.749(0.586)
Age	β_{03}	-	0.140(0.062)
Time effect:			
	β_{10}	-0.012(0.052)	-0.009(0.052)
Sud effects:			
	β_{20}	-0.280(0.162)	-0.172(0.159)
Hapneu*sud	β_{21}	-	-
Fearneu*sud	β_{22}	-	-
Age	β_{23}	-	-
Covariance of u_j:			
$Var(u_{0j})$		17.079(5.350)	13.750(4.274)
$Var(u_{1j})$		0.023(0.019)	0.026(0.019)
$Cov(u_{0j}, u_{1j})$		0.529(0.250)	0.461(0.222)
Residual variance:			
$Var(\varepsilon_{jsl})$	σ^2	29.379(1.958)	26.932(1.795)
Deviance		3176.215	3131.684

3.4.2 Analysis of Gilles de La Tourette's Data

First, one needs to realize that for these data the within-subject conditions are the *sessions* 1-10. Therefore the measurements for the "total sensation severity score" (TSSS) are structured within the 10 sessions. Second, these data fall under case 2 (see Section 2.2.2) since we have multiple time trends, that is for each s ($s = 1, \ldots, 10$), we have time points t_{sl} for $l = 1, \ldots, 9$.

Research findings in the treatment of anxiety disorders indicate that prolonged exposure to feared stimuli by means of ER has resulted in reductions in both subjective anxiety and heart rate (Marks, 1987). Based on these findings, therapists have hypothesized that the severity of sensations experienced by a subject within a session will decrease over time (within session habituation). Another hypothesis is that as one progresses from the first session through the last one, the severity of sensations experienced by a subject will decrease (across session habituation). The research project at hand was executed to investigate the following research questions: To find out whether

1. Within sessions the severity of sensations decreases over time (within session habituation).

2. The average sensations severity score depends on session, tics and the interaction between the two.

3. Across sessions, the severity of sensations decreases (across session habituation)

4. Within session habituation depends on session number and the number of tics exhibited within a session

Let sen_{jsl} denote the TSSS recorded for the jth subject at the lth time point in session s and $codetics_{js}$ be the variable representing the number of tics exhibited by subject j in session s ($j = 1, \ldots, 19$; $s = 1, \ldots, 10$; $l = 1, \ldots, 9$). The variable *codetics* has a scale of 1 - 5 where 1 means no tics exhibited, 2 means between 1 and 8 tics exhibited, 3 means between 9 and 16 tics exhibited, 4 means between 16 and 24 tics exhibited and lastly 5 means 25 or more tics exhibited. Assuming that within each session the data of each subject can be well approximated by a quadratic function of time, then sen_{jsl} satisfies the general equation,

$$sen_{jsl} = \pi_{0js} + \pi_{1js} t_{sl} + \pi_{2js} t_{sl}^2 + \varepsilon_{jsl}, \tag{3.14}$$

where the subject-specific regression coefficients π_{ijs}, for i=0,1,2 depend on the subject j and the session s. Just like equation (3.11) in Section 4.1, equation (3.14) is a simplified version of (3.5). It should be noted that the GTS data is also balanced since for each session, all subjects have the same number of repeated measures, taken at time points which are also the same

for all subjects. Consequently L_{js} (see first paragraph of Section 2.2) simplifies to L and therefore in (3.14), we use t_{sl} instead of the general term t_{jsl} in (3.5). Throughout the analysis the variables *session* and *codetics* are centered at their means (which are 5.5 and 3 respectively). Within each session, the variable time was converted to hours and then centered at its mean (which is 1). Expressing time in hours rather than minutes is done to avoid that the random slopes for the linear and quadratic time effects would show too little variability, which might lead to divergence of the numerical maximization routine (Verbeke and Molenberghs, 2000, Section 5.6).

The analysis will begin with a baseline model (model 1) in which the π_{ij}'s depend on subject j but not on session s. This renders,

$$sen_{jsl} = \pi_{0js} + \pi_{1js}(t_{sl} - 1) + \pi_{2js}(t_{sl} - 1)^2 + \varepsilon_{jsl}, \tag{3.15}$$

where,

$$\begin{cases} \pi_{0js} &= \beta_{00} + u_{0j}, \\ \pi_{1js} &= \gamma_{00} + v_{0j}, \\ \pi_{2js} &= \alpha_{00} + w_{0j}, \end{cases} \tag{3.16}$$

This model assumes that the profiles of the subjects are the same across sessions and it can be used to answer the first research question. The results of the model fit are displayed in Table 3.2. As can be seen in the table, the parameters β_{00} and γ_{00} are significant while α_{00} is not significant. Since γ_{00} is negative, it follows that within sessions the severity of sensations decreases over time.

Research questions 2 and 3 can be investigated by means of an extension of the baseline model. This extension is achieved by replacing π_{0js} in (3.16) with

$$\pi_{0js} = A_{0j} + A_{1j}(s - 5.5) + A_{2j}(codetics_{js} - 3) + A_{3j}(s - 5.5) * (codetics_{js} - 3) \tag{3.17}$$

where,

$$\begin{cases} A_{0j} = \beta_{00} + u_{0j}, \\ A_{1j} = \beta_{10} + u_{1j}, \\ A_{2j} = \beta_{20} + u_{2j}, \\ A_{3j} = \beta_{30}, \end{cases} \tag{3.18}$$

and leaving π_{1js} and π_{2js} intact. This renders model 2. The parameters β_{10}, β_{20} and β_{30} will be used to answer the research questions depending on whether they are significantly different from zero or not. For instance if

β_{10} is negative, then this suggests that the average severity of sensations decreases across sessions. Table 3.2 shows the results obtained from fitting model 2, and Table 3.3 displays the model fit summary. As can be seen in Table 3.3, comparing the deviances of models 1 and 2 shows that model 2 fits better than model 1.

An extension of model 2 will be used to investigate the fourth research question. The extension is obtained by replacing π_{1js} with

$$\pi_{1js} = B_{0j} + B_{1j}(s - 5.5) + B_{2j}(codetics_{js} - 3) \tag{3.19}$$

where

$$\begin{cases} B_{0j} = \gamma_{00} + \nu_{0j}, \\ B_{1j} = \gamma_{10} + \nu_{1j} \\ B_{2j} = \gamma_{20} + \nu_{2j} \end{cases} \tag{3.20}$$

producing model 3. In this model, the parameters γ_{10} and γ_{20} will be used to answer the fourth research question. At this point, it is worthwhile to note that equations (3.16), (3.18) and (3.20) are obtained by adapting the general equation (3.8) to the GTS data. In other words (3.16), (3.18) and (3.20) are versions of (3.8).

On fitting model 3, the algorithm failed to converge. The remedy to this was to reduce the number of random slopes in the model. Convergence was attained after removing the random terms ν_{1j} and ν_{2j} from the model. The results are displayed in Table 3.2. For brevity, for the random effects, only variances have been reported. Comparing the deviances of models 2 and 3 (see Table 3.3) indicates that model 3 is a significant improvement of model 2. Staying with model 3, it can be seen that comparing β_{20}, γ_{10} and α_{00} with their respective standard errors reveals that they are not significant. On the other hand β_{00}, β_{10}, β_{30}, γ_{00}, and γ_{20} are significant. All tests are based on a 5 percent level of significance. The effects corresponding to the non significant parameters γ_{10} and α_{00}, were therefore removed from the model and a new model was fitted. Although the average effect for codetics (β_{20}) is not significantly different from zero, it will not be removed from the model. This is because the interaction between session and codetics is significant and so the corresponding main effects session and codetics should also be in the model. The results of the new model fit are displayed in Table 3.2 under model 4. Model 3 has 7 parameters more than than model 4. The difference between their deviances is 64.025 and on 7 degrees of freedom, this yields a p-value very close to zero. Much as the parameter estimates of the effects which are considered in both models are pretty close to each other, the fact that the p-value is smaller than 0.05 suggests that model 3 should be preferred. Consequently in what follows, all interpretations are based on the parameter estimates from model 3.

First, the time effects will be considered. In model 3, γ_{00} = -0.69 implying that there is within session habituation. The within session habituation follows a linear trend since α_{00} is not significant. In addition, since γ_{20} is significant, then within session habituation depends on the number of tics exhibited within a session. The fact that γ_{20} is positive indicates that the larger the number of tics exhibited by a patient, the less the habituation. As an illustration consider two patients A and B who exhibit 5 tics (coded as 1 for the variable codetics) and 30 tics (coded as 5 for the variable codetics) respectively within a particular session. Note that the values 1 and 5 for codetics correspond to centered values of about -2 and 2 respectively. Then γ_{20}= 0.23 implies that within each session the severity of sensations decreases by π^{A}_{1js}= -0.69 + 0.23*-2 = -1.15 per unit time for patient A and decreases by π^{B}_{1js}= -0.69 + 0.23*2 = -0.23 per unit time for patient B. Since re-scaled time ranges from -1 to 1 (equivalent to two units of time) a decrease of 1.15 for patient A (0.23 for patient B) amounts to a within session habituation of 2.3 for patient A (0.46 for patient B).

Next, since β_{10} and β_{30} are significant, then the average number of sensations π_{0js} depends on session and the interaction between session and tics. β_{10}= -0.41 implies that the higher the session number, the smaller the average sensation severity score. So the average number of sensations decreases as patients progress from session 1 through session 10. For instance controlling for all other effects, the average number of sensations in session 6 will be about 2 units less than the average in session 1. Furthermore β_{30} = 0.05 indicates that with increasing session number, the difference in sensations between two persons exhibiting few and many tics respectively is increasing. In other words persons exhibiting few tics during the sessions, habituate better than persons exhibiting more tics.

Furthermore as can be seen in Table 3.2, the variance of the random effect u_{0j} is 11.861 and the residual variance is 3.495 which sums up to 15.356. This shows that at time = 1 (corresponding to an un-scaled time of 60 minutes), session = 5.5 (corresponding to the period after the 5th session and before the 6th session), patients who exhibit between 9 and 16 tics (equivalent to 3 on the codetics scale) experience on average 5 sensations (on a scale runing from 0-20) with a standard deviation of 3.92.

To conclude, the above results indicate that within each session the number of sensations felt by a patient decreases over time (within session habituation). Note that this decrease depends on the number of tics exhibited by a patient. Patients exhibiting more tics experience less reduction as compared to patients exhibiting fewer tics. Furthermore, on average the number of sensations felt by the patients decreases as patients progress from the first session through the last session (across session habituation).

The above results are of significant importance for a psychologist/behavior therapist. This is because they support the hypothesis that habituation is

an underlying mechanism of change in exposure therapy in Tourette's syndrome. This is evidenced by the following:

1. Subjective severity sensations ratings reduce within sessions (over time) and across sessions and over time.

2. Interupting exposure (more tics during sessions implies exposure) reduces the level of habituation (less reduction of subjective sensation severity scores).

Table 3.2: GTS data. Parameter (P) estimates and standard errors for the models.

		Model 1	Model 2	Model 3	Model 4
Effect	P	Estimate(S.E)	Estimate(S.E)	Estimate(S.E)	Estimate(S.E)
Average:					
	β_{00}	5.447(0.665)	5.222(0.797)	5.285(0.804)	5.304(0.767)
session	β_{10}	-	-0.417(0.099)	-0.411(0.099)	-0.416(0.100)
tics	β_{20}	-	0.038(0.238)	0.029(0.237)	0.078(0.242)
session*tics	β_{30}	-	0.053(0.023)	0.051(0.023)	0.052(0.024)
Time effects:					
	γ_{00}	-0.680(0.197)	-0.688(0.195)	-0.687(0.164)	-0.693(0.166)
session	γ_{10}	-	-	0.019(0.026)	-
tics	γ_{20}	-	-	0.227(0.058)	0.214(0.067)
Time squared effects:					
	α_{00}	0.083(0.283)	0.079(0.282)	0.075(0.283)	-
Random effects					
$Var(u_{0j})$		8.271(2.728)	11.667(3.916)	11.861(3.976)	10.749(3.626)
$Var(u_{1j})$		-	0.077(0.028)	0.077(0.028)	0.077(0.028)
$Var(u_{2j})$		-	0.538(0.226)	0.528(0.222)	0.547(0.234)
$Var(v_{0j})$		0.595(0.240)	0.623(0.234)	0.410(0.165)	0.420(0.170)
$Var(v_{1j})$		-	-	-	-
$Var(v_{2j})$		-	-	-	-
$Var(w_{0j})$		1.075(0.494)	1.209(0.492)	1.215(0.493)	-
Residual variance:					
$Var(\varepsilon_{jsl})$	σ^2	5.113(0.182)	3.499(0.126)	3.495(0.126)	3.669(0.131)
Deviance		7445.535	6925.565	6914.045	6978.070

Table 3.3: GTS data. Model fit summary.

Model	No. of parameters	Deviance difference	Reference model	df	p-value
1	10				
2	22	519.97	1	12	0.0000
3	24	11.52	2	2	0.0032

3.5 Discussion

This paper proposes methods which can be used to analyze data obtained from designs where measurements taken repeatedly over time are structured in within-subject conditions. The models for these data were obtained by adapting the linear mixed model and in Section 3.2, it was shown that the Structured Repeated Measurements models have the same general form as the linear mixed model.

The methods have been illustrated with an analysis of two data sets from clinical psychology; the first data set consisted of measurements taken over one time trend across the within-subject conditions while the second consisted of measurements taken from multiple time trends with the same time trend for each condition.

The models proposed in this paper are general and can be used to analyze Structured Repeated Measurements data from the medical, behavioral and social sciences. Furthermore since the models have the same general form as the linear mixed model, any commercially available package which allows fitting of linear mixed models (e.g., SAS, STATA and MLWIN) can be used.

Chapter 4

A Bayesian approach to Inequality Constrained Hierarchical Models: Estimation and Model selection

Abstract

Constrained parameter problems arise in a wide variety of applications. This paper deals with estimation and model selection in hierarchical linear models with inequality constraints on the parameters. It is shown that different theories can be translated into statistical models by putting constraints on the model parameters yielding a set of competing models. A new approach based on the principle of encompassing priors is proposed and used to compute Bayes factors and subsequently posterior model probabilities. Model selection is based on posterior model probabilities. The approach is illustrated using three examples.

4.1 Introduction

Order restricted parameter problems arise in a wide variety of applications (see for example the Journal of Statistical Planning and Inference, 2002, Volume 107, Issues 1-2 and, Robertson, Wright and Dykstra, 1988). This paper deals with hierarchical linear models with inequality constraints on the parameters. Constraints on parameters can be used to translate theories into statistical models. Most of the literature on order restricted parameter problems uses the frequentist approach. Few publications have focused on the Bayesian analysis of order restricted parameter problems especially with respect to hierarchical models. Furthermore even though a great deal of frequentist literature exists on order restricted parameter problems, most of the attention is focussed on estimation and hypothesis testing. This paper deals with estimation and model selection via the computation of posterior model probabilities.

Consider an experiment in medical science in which subjects have been randomized to either a control group (C) or one of two treatment groups. Suppose that treatment consists of a low (L) or a high (H) dose of a drug and that subjects in each of the three groups (C, L and H) are followed over time and measurements taken at several time points. Of primary interest is the estimation of changes over time and testing whether these changes are treatment dependent. A simple hierarchical model can be used to investigate this:

$$y_{jk} = (\beta_1 + u_{1j}) + (\beta_2 C_j + \beta_3 L_j + \beta_4 H_j + u_{2j}) t_{jk} + \varepsilon_{jk}, \tag{4.1}$$

$$\mathbf{u}_j = (u_{1j}, u_{2j})^T \sim N(\mathbf{0}, \mathbf{V}), \quad \varepsilon_{jk} \sim N(0, \sigma^2),$$

in which y_{jk} denotes the $k = 1, \ldots, K_j$-th measurement on the outcome variable for subject $j = 1, \ldots, J$ and t_{jk} the corresponding time at which the measurements are made. The covariance matrix of the random effects is $\mathbf{V}$ and σ^2 is the residual variance. Furthermore C_j, L_j, and H_j are indicator variables defined to be one if the subject belongs to the control, low-dose group or the high-dose group respectively, and zero otherwise. The parameter β_1 can be interpreted as the average response at the start of the treatment ($t_{jk} = 0$), whereas the parameters β_2, β_3, and β_4 represent the average slopes (averaged over persons) for the treatment groups C, L and H respectively. Interest lies in the comparison of these average slopes, as this directly measures the treatment effect on the outcome variable y. Assuming that treatment aims at increasing the values of the outcome variable, a researcher may have several theories about the experiment at hand; one plausible theory is that the higher the dose the higher the increase in y over time. This leads to a model with the inequality constraint, $\beta_2 < \beta_3 < \beta_4$. Another theory would be that both treatments have the same effect over time and that each treatment is better than no treatment. This yields a

model with the inequality constraint, $\beta_2 < \{\beta_3, \beta_4\}$. Alternatively the researcher may envisage the situation where the direction of the treatment effect is not known and therefore fit a model without any constraints on the βs. On the other hand, it may be that the low dose is better than than no treatment and high dose has adverse effects. This renders the constraint $\beta_2 < \beta_3 > \beta_4$. Lastly, another possibility is that there is no treatment effect, leading to a model in which the average slopes are constrained to be equal, that is $\beta_1 = \beta_2 = \beta_3$. Consequently, five competing models are obtained; one without constraints on the parameters, three with strict inequality constraints on the parameters and one with an equality constraint, that is

$$\begin{aligned}
M_1 &: \quad \beta_2, \beta_3, \beta_4; \\
M_2 &: \quad \beta_2 < \beta_3 < \beta_4; \\
M_3 &: \quad \beta_2 < \{\beta_3, \beta_4\}; \\
M_4 &: \quad \beta_2 < \beta_3 > \beta_4; \\
M_5 &: \quad \beta_2 = \beta_3 = \beta_4.
\end{aligned}$$

Note that the remainder of the parameters in (4.1) have no restrictions on them.

As a second example, consider a design where you have several *within* subject conditions. These conditions could for instance be treatment sessions, seasons, years etc. In this setting, a researcher could have some expectations about the effect of the treatment as one progresses from the first condition to the last. For instance, one possible theory is that the average response decreases as subjects progress from the first session to the last. Another theory could predict a single peaked trajectory from the first to the last session.

In this paper, a model in which all other models are nested will be called the *encompassing* model. For instance in the first example model M_1 is the encompassing model. Note that the other models do not have to be mutually nested. Again considering the first example it is clear that models M_2, M_3, M_4 and M_5 are not mutually nested. Further, the treatments (control, low dose or high dose) are *between* subject conditions whereas in the second example the treatment sessions are *within* subject conditions. It is the primary purpose of this paper to show that in a hierarchical data setting, several competing theories can be translated into different constrained hierarchical models which can then be analyzed by making use of Bayesian methodology. The competing theories can then be compared by looking at the posterior probabilities of the models corresponding to the different theories. Much as all the examples presented in this paper (see Section 4.3) are from a longitudinal data setting, the methodology can be adapted and applied to any multilevel data setting. In Section 4.2 the methodology for model estimation and model selection is introduced and explained. In Sec-

tion 4.3 the methodology is illustrated by applying it on three longitudinal data sets. The paper is concluded with a discussion in Section 4.4.

4.2 The Model

Longitudinal data consist of time sequences of measurements on several experimental or observational units. In this paper, the following notation is used: y_{jk} and x_{pjk} denote the $k = 1, \ldots, K_j$-th measurement on the outcome variable and $p = 1, \ldots, P$-th covariate respectively on the $j = 1, \ldots, J$-th unit and t_{jk} the corresponding time at which the measurements are made.

The standard hierarchical model for longitudinal data has the following structure:

$$y_{jk} = \boldsymbol{x}_{jk}^T \boldsymbol{\beta} + \boldsymbol{z}_{jk}^T \boldsymbol{u}_j + \varepsilon_{jk}, \tag{4.2}$$

where $\boldsymbol{x}_{jk}$ and $\boldsymbol{z}_{jk}$ of length P and Q, respectively are vectors of known covariates, $\boldsymbol{\beta}$ is a vector of length P containing the fixed effects parameters, $\boldsymbol{u}_j$ is a vector of length Q ($Q \leq P$) containing the random effects and ε_{jk} is an error term. The vector of covariates $\boldsymbol{z}_{jk}$ will usually be a subset or a linear combination of the fixed-effects covariates $\boldsymbol{x}_{jk}$. For instance, in (4.1), $\boldsymbol{x}_{jk}^T = (1,\ C_j t_{jk},\ L_j t_{jk},\ H_j t_{jk})$ and $\boldsymbol{z}_{jk}^T = (1,\ t_{jk})$ where $t_{jk} = C_j t_{jk} + L_j t_{jk} + H_j t_{jk}$. All random terms in the model are assumed to be normally distributed:

$$\boldsymbol{u}_j \sim N_Q(\boldsymbol{0}, \boldsymbol{V}),\quad \varepsilon_{jk} \sim N(0, \sigma^2),\quad \boldsymbol{u}_j \text{ and } \varepsilon_{jk} \text{ are independent,}$$

where $\boldsymbol{V}$ is the $(Q \times Q)$ covariance matrix of the random effects and σ^2 is the residual variance. Note that the standard hierarchical model can be adapted to deal with designs where one has *between* subject or *within* subject conditions or both. This will further be illustrated in Section 4.3.

4.2.1 Estimation

In the Bayesian framework the information with respect to the parameters of a model is summarized in the posterior distribution. The posterior distribution of the model parameters is based on the likelihood of the data given the model parameters and the prior distribution of the model parameters. As mentioned in Section 4.1, it is not uncommon that a researcher has a number of competing theories concerning some subject matter. These theories are then translated into statistical models by putting inequality constraints on the model parameters. For the models considered in this paper, the *encompassing* model is the one in which no constraints are put on the model parameters and all other models are nested in this model. Furthermore, the prior distribution of the parameters in the *encompassing*

model will be called the *encompassing* prior. Basically an encompassing prior distribution is one that is uninformative with respect to the parameters in the encompassing model. It follows from (4.2), that $\theta = (\boldsymbol{\beta}, \boldsymbol{V}, \sigma^2)$ is the vector of parameters in the encompassing model. To obtain a conjugate model specification normal priors will be used for the βs, an inverse Wishart prior for $\boldsymbol{V}$ and a scaled inverse chi square prior for σ^2. Using independent distributions for each model parameter, the prior distribution for the parameters in the encompassing model, which will be denoted by M_1 is:

$$\begin{aligned} g(\boldsymbol{\beta}, \boldsymbol{V}, \sigma^2|M_1) &= g(\boldsymbol{V}) \times g(\sigma^2) \times g(\boldsymbol{\beta}) \\ &= \text{Inv-W}(\boldsymbol{V}|\, \lambda, \boldsymbol{S}) \times \text{Inv-}\chi^2(\sigma^2|\, \upsilon_0, \sigma_0^2) \\ &\quad \times \prod_{p=1}^{P} N(\beta_p|\, \mu_p, \delta_p^2), \quad (4.3) \end{aligned}$$

where Inv-$\chi^2(\sigma^2|\, \upsilon_0, \sigma_0^2)$ denotes a scaled inverse Chi-square distribution with degrees of freedom υ_0 and scale σ_0, Inv-W$(\boldsymbol{V}|\, \lambda, \boldsymbol{S})$ denotes an inverse Wishart distribution with degrees of freedom λ and scale matrix $\boldsymbol{S}$ and $N(\beta_p|\, \mu_p, \delta_p^2)$ denotes a normal distribution with mean μ_p and standard deviation δ_p (see Gelman et al., 2000, pp. 474–475). In Section 2.2.2, more information will be given about the specification of the distributions in (4.3).

The prior distribution of any model M_r, $r = 2, \ldots, R$ that is nested in the encompassing model M_1 can be obtained from (4.3) by restricting the parameter space in accordance with the constraints imposed by a model M_r and is given by

$$g(\theta|M_r) = \frac{g(\theta|M_1) I_{M_r}(\theta)}{\int g(\theta|M_1) I_{M_r}(\theta) d\theta}. \quad (4.4)$$

In (4.4), $I_{M_r}(\theta)$ is the indicator function for model M_r, so that $I_{M_r}(\theta)$ equals 1 if the parameter values are in accordance with the restrictions imposed by model M_r, and equals 0 otherwise. In other words, for each model under consideration the corresponding constraints are accounted for in the prior distribution of the model. It should be noted that only the prior distribution of the parameters of the encompassing model has to be specified and that the prior distributions of the other models can be derived using (4.4).

Let $\{M_1, M_2, \ldots, M_R\}$ denote the finite set of all models under consideration, that is the set of competing models and $\boldsymbol{D} = (\boldsymbol{y}_{jk}, \boldsymbol{x}_{jk}, \boldsymbol{z}_{jk} : k = 1, \ldots, K_j;\ j = 1, \ldots, J)$ denotes the data. Then the posterior distribution of the parameters in model M_r, for $r = 1, \ldots, R$, is given by:

$$P(\theta \mid \boldsymbol{D}, M_r) \propto L(\boldsymbol{D} \mid \theta) g(\theta|M_r), \quad (4.5)$$

where L is the likelihood of the data and $g(.)$ is the prior distribution of the parameters θ in model M_r. The likelihood function, L is given by,

$$L(.\mid .) = \prod_{j=1}^{J} \int_{\mathbf{u}_j} \{ \prod_{k=1}^{K_j} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_{jk} - \mathbf{x}_{jk}^T\boldsymbol{\beta} - \mathbf{z}_{jk}^T\mathbf{u}_j)}{2\sigma^2}) \} h(\mathbf{u}_j \mid \mathbf{0}, \mathbf{V})\, d\mathbf{u}_j,$$

where $h(\mathbf{u}_j|\ \mathbf{0}, \mathbf{V})$ denotes a normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{V}$. Since conjugate priors are used to specify $g(\theta|M_r)$, it follows that for the encompassing (or unconstrained) model, the *full conditional* distributions of the parameters in θ are known distributions and are therefore available in closed form. In particular, the *full conditional* distribution of each β_p is a normal distribution while the *full conditional* distributions of $\mathbf{V}$ and σ^2 are inverse Wishart and scaled inverse chi square distributions respectively. Consequently the Gibbs sampler introduced by Geman and Geman (1984) in the context of image processing and now widely used as a general method for Bayesian calculations (see for example, Gelfand and Smith, 1990; Zeger and Karim, 1991) can be used to sample the model parameters from the posterior distribution in (4.5). It is worthwhile to note that Gibbs sampling proceeds most smoothly by regarding the random effects $\mathbf{u}_1, \ldots, \mathbf{u}_J$ as latent variables to be sampled along with the model parameters σ^2, $\mathbf{V}$, and β_p for $p = 1, \ldots, P$. The Gibbs sampling algorithm (see Appendix A for details) can be summarized as follows:

1. Sample $\mathbf{u}$ from the distribution of $\mathbf{u} \mid \mathbf{D}, \boldsymbol{\beta}, \sigma^2, \mathbf{V}$.

2. Sample σ^2 from the distribution of $\sigma^2 \mid \mathbf{D}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{V}$.

3. Sample $\mathbf{V}$ from the distribution of $\mathbf{V} \mid \mathbf{D}, \boldsymbol{\beta}, \mathbf{u}, \sigma^2$.

4. Sample each β_p ($p = 1, \ldots, P$) from the (possibly truncated) distribution of $\beta_p \mid \mathbf{D}, \mathbf{u}, \sigma^2, \mathbf{V}, \boldsymbol{\beta}_{-p}$, where $\boldsymbol{\beta}_{-p}$ denotes the vector $(\beta_1, \ldots, \beta_{p-1}, \beta_{p+1}, \ldots, \beta_P)$.

A similar algorithm is used in Browne (1998, Section 4) and Browne and Draper (2000) to sample model parameters from a random slopes regression model. The differences are that in the above mentioned papers, $\mathbf{z}_{jk}$ is either a sub-vector of $\mathbf{x}_{jk}$ or equal to $\mathbf{x}_{jk}$. Second, the model parameters are sampled without any constraints on them and consequently there the vector $\boldsymbol{\beta}$ is sampled at once from a multivariate normal distribution.

The total number of iterations for the Gibbs sampler is chosen large enough (see details in Section 4.3) to ensure that the sample converges to the target posterior distribution; the first few iterations serve as a 'burn in' and are usually discarded. The remaining sample can then be used to obtain all desired estimates. For instance, the values sampled for each model

parameter can used to obtain the expected a posteriori estimate (EAP), posterior standard deviation and central credibility interval for the respective parameter (see for example Hoijtink, 2000).

An extra step in the Gibbs sampler is that for all the constrained parameter models, each $\beta_p \in \boldsymbol{\beta}$ is sampled from a truncated normal distribution with some lower bound b and upper bound c. The value b is the largest $\beta_{p'}$ ($p \neq p'$) that must be smaller than β_p, and c is the smallest $\beta_{p'}$ ($p \neq p'$) that must be greater than β_p. To expound on this: suppose model M is specified using: $\beta_1 > \beta_2 > \beta_3 > \beta_4$. Then at each iteration of the Gibbs sampler, the sampled βs should fulfill the constraint imposed by the model. In this case, the parameters β_1, β_2, β_3 and β_4 would be sampled from normal distributions with lower bounds β_2, β_3, β_4 and $-\infty$ respectively and upper bounds ∞, β_1, β_2 and β_3 respectively. In general, for the constrained parameter models, sampling a β_p proceeds via sampling a deviate i from a uniform distribution on the interval (0,1) and then taking $\beta_p = \Phi^{-1}_{\beta_p}[\Phi_{\beta_p}(b) + i(\Phi_{\beta_p}(c) - \Phi_{\beta_p}(b))]$, where $\Phi_{\beta_p}(.)$ denotes the standard normal cumulative probability function of β_p evaluated at the argument (see Gelfand et al., 1992).

4.2.2 Model Selection

In order to select the best from a set of competing models a criterion for model selection is necessary. In this paper, posterior probabilities will be computed for each model under consideration. If $\{M_1, M_2, \ldots, M_R\}$ denotes the set of all models under consideration, then the posterior probability of model M_r is

$$P(M_r \mid \mathbf{D}) = \frac{P(\mathbf{D} \mid M_r)P(M_r)}{\sum_{r'=1}^{R} P(\mathbf{D} \mid M_{r'})P(M_{r'})}, \tag{4.6}$$

where $P(M_r)$ is the prior probability of model M_r and $P(\mathbf{D} \mid M_r)$ is the marginal likelihood of model M_r. As can be seen from (4.6), the magnitude of the posterior model probability of model M_r is influenced by the prior probability of the model. In the examples considered in this paper (see Section 4.3), prior probabilities are chosen to be the same for each model under consideration, that is each model is apriori equally likely. This ensures that any differences in posterior model probabilities are caused by differences in the extent to which the model conforms to the data.

The problem of calculating posterior probabilities for a collection of competing models continues to be a challenge for applied Bayesian statisticians. This is largely due to the complications involved in obtaining the marginal likelihood $P(\mathbf{D} \mid M_r)$. The marginal likelihood can be obtained as

$$P(\mathbf{D} \mid M_r) = \int_{\theta} L(\mathbf{D} \mid \theta) g(\theta|M_r) d\theta. \tag{4.7}$$

In the literature (see for example, Carlin and Chib, 1995; Kass and Raftery, 1995) various methods for the computation of (4.7) have been proposed. In the next section a new method which can be used to obtain posterior probabilities in the context of inequality constrained models will be presented.

4.2.2.1 Estimation of Posterior probabilities using Encompassing priors

In this section we propose a new method for the computation of posterior probabilities. The method is based on the principle of encompassing priors and it works as follows: consider two models, the encompassing model M_1, another model M_r (nested in M_1) and observed data $\mathbf{D}$. In general, for any model M_r, the marginal likelihood can be written as

$$P(\mathbf{D} \mid M_r) = \frac{L(\mathbf{D} \mid \theta)\, g(\theta|M_r)}{P(\theta \mid \mathbf{D}, M_r)},$$

where the numerator is a product of the likelihood function of the data and the prior distribution of θ under model M_r, while the denominator is the posterior density of θ under model M_r (see Chib, 1995, Section 2). Consequently the Bayes factor of M_1 to M_r is given by:

$$BF_{1r} = \frac{P(\mathbf{D}|M_1)}{P(\mathbf{D}|M_r)} = \frac{L(\mathbf{D} \mid \theta)g(\theta|M_1) \,/\, P(\theta \mid \mathbf{D}, M_1)}{L(\mathbf{D} \mid \theta)g(\theta|M_r) \,/\, P(\theta \mid \mathbf{D}, M_r)}. \tag{4.8}$$

Suppose $\theta = \theta^*$ is a parameter vector that is common between the models M_1 and M_r. Then substituting $\theta = \theta^*$ in (4.8) renders,

$$BF_{1r} = \frac{g(\theta^*|M_1)P(\theta^*|\, \mathbf{D},\, M_r)}{g(\theta^*|M_r)P(\theta^*|\, \mathbf{D},\, M_1)}.$$

Since M_r is nested in M_1, it follows that the densities $g(\theta^*|M_r)$ and $P(\theta^*|\, \mathbf{D}, M_r)$ can be rewritten as $B_r \times g(\theta^*|M_1)$ and $A_r \times P(\theta^*|\, \mathbf{D}, M_1)$ respectively, where A_r and B_r are constants. In other words, if model M_r is nested in model M_1 then the prior and posterior densities of M_r can be rewritten in terms of the prior and posterior densities of M_1. Effectively, $1/B_r$ is the proportion of the prior distribution of M_1 in agreement with M_r and $1/A_r$ is the proportion of the posterior distribution of M_1 in agreement with M_r. Assuming the two models are *a priori* equally probable (i.e., $P(M_1) = P(M_r) = 0.5$), we have that $BF_{1r} = P(M_1|\mathbf{D}) \,/\, P(M_r|\mathbf{D}) = A_r/B_r$. This procedure can be applied to any number of competing models as long as an encompassing model within which each of them is nested can be specified. A computational advantage of this method is that a researcher only needs to specify the prior distribution and subsequently the posterior distribution of the encompassing model and then sample from each of them to

determine the proportion of parameter vectors (i.e., $1/B_r$ and $1/A_r$ respectively) from each in agreement with any model nested in the encompassing model.

After obtaining $1/A_r$ and $1/B_r$ for each model in $\{M_2, M_3, \ldots, M_R\}$ the posterior probability of the encompassing model is estimated using

$$P(M_1|\mathbf{D}) = \left(\prod_{r=2}^{R} A_r\right) / \left[(\prod_{r=2}^{R} A_r) + \sum_{r=2}^{R}(B_r \prod_{s\neq r} A_s)\right], \tag{4.9}$$

and the posterior probability of model M_r, $r = 2, \ldots, R$, is estimated using

$$P(M_r|\mathbf{D}) = \left(B_r \prod_{s\neq r} A_s\right) / \left[(\prod_{r=2}^{R} A_r) + \sum_{r=2}^{R}(B_r \prod_{s\neq r} A_s)\right]. \tag{4.10}$$

4.2.2.2 Specification of the Encompassing prior

As explained in Section 4.2.1, only the prior distribution of the parameters of the encompassing model has to be specified. In the sequel we will distinguish between $\boldsymbol{\beta}$, which is the parameter vector of interest and $\mathbf{V}$ and σ^2 which are nuisance parameters. Specification of the encompassing prior is based on the following principles:

1. The encompassing prior should not favor the unconstrained or any of the constrained models. This is achieved by using similar and independent prior distributions for each element of the parameter vector $\boldsymbol{\beta}$, that is the prior distribution of β_p is $g(\beta_p) = g(\beta)$ for $p = 1, \ldots, P$. The consequence is that in (4.3), $\mu_p = \mu$ and $\delta_p^2 = \delta^2$ for $p = 1, \ldots, P$. In situations where $\boldsymbol{\beta}$ is subdivided into groups $w = 1, ..., W$ such that constraints apply only to parameters from the same group, then $g(\beta_p) = g^w(\beta)$ for $p \in \{w\}$. In other words, parameters from the same group would be assigned the same prior distributions.

2. The prior density $g(\beta)$ should be continuous on the real line or on a subinterval of the real line $\Re$. For discrete prior distributions, the properties derived in the remainder of this section do not hold.

3. The prior density $g(\beta)$ should be vague, that is, it should not exclude regions of the parameter space with substantial posterior probability. For instance, if at least 99% of the posterior distribution of each element of $\boldsymbol{\beta}$ is within the interval $[\alpha, \gamma]$, then each β_p could have a mean and variance computed in such a way that the prior distribution of β_p corresponds to a normal distribution with 0.5-th and 99.5-th percentiles α and γ respectively. The consequence is that the (encompassing) posterior is almost unaffected by the (encompassing) prior distribution.

4. The encompassing prior should be vague with respect to the nuisance parameters $\mathbf{V}$ and σ^2. This will be achieved by using scaled inverse Chi-square and inverse Wishart distributions with one degree of freedom and data based scale factors, namely the estimated residual variance and covariance matrix of the random effects respectively. Like in 3, this ensures that the (encompassing) posterior distribution is unaffected by the (encompassing) prior distribution.

The sensitivity of posterior probabilities with respect to the actual specification of the encompassing prior in (4.3) is an issue that deserves further attention. For a large class of models, posterior probabilities are virtually objective, that is, independent of the actual specification of a vague encompassing prior. In the sequel, this will be illustrated using $1/B_r$ and $1/A_r$, the proportion of the encompassing prior and posterior respectively, in agreement with a constrained model M_r (see Section 2.2.1). The sensitivity with respect to $g(\mathbf{V})$, $g(\sigma^2)$ and $g(\boldsymbol{\beta})$ will be discussed separately. From (4.3) and (4.4) it is immediately clear that $g(\mathbf{V})$ and $g(\sigma^2)$ do not influence $1/B_r$. On the other hand, from (4.5) it is also clear that $g(\mathbf{V})$ and $g(\sigma^2)$ do influence the posterior distribution, and accordingly $1/A_r$. However, it is well-known (see, for example, Gelman et al., 2000, p.101) that this influence is small if the sample size is large compared to the degrees of freedom of the scaled inverse Chi-square and inverse Wishart distributions. Similarly, if $g(\beta_p)$ is vague, its influence on the posterior distribution is negligible, and subsequently its influence on $1/A_r$ is negligible.

For a specific class of models the choice of a prior distribution $g(\beta)$ for the parameters β_p does not influence $1/B_r$. Let $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\}$. This class uses one or more constraints of the form

$$\sum_{p=1}^{P} \beta_p^{(1)} > \sum_{s=1}^{S} \beta_s^{(2)}, \tag{4.11}$$

with the restriction $P = S$. For encompassing priors it holds that $\beta_p^{(1)}, \beta_s^{(2)} \overset{\text{i.i.d.}}{\sim} g(\beta)$. The consequence is that $P(\sum_{p=1}^{P} \beta_p^{(1)} > \sum_{s=1}^{S} \beta_s^{(2)})$ is 0.5, and hence independent of the actual choice of $g(\beta)$. A similar result holds for combinations of constraints of the type (4.11). For example, $P(\beta_1 < \beta_2 < \beta_3) = 1/6$, which is independent of the choice of $g(\beta)$. This line of reasoning can be continued and applied to situations involving more than one group of parameters. As a further illustration consider, $\{\beta_1, \beta_2\}$ and $\{\beta_3, \beta_4\}$. Here $P(\{\beta_1 > \beta_2\} \bigcap \{\beta_3 > \beta_4\}) = P(\beta_1 > \beta_2) \times P(\beta_3 > \beta_4) = 0.5 \times 0.5 = 0.25$, which again is a result independent of the prior distribution for each group of parameters. Besides (4.11), there are other constraints for which model selection based on encompassing priors is virtually objective. However, these constraints are not used in the examples to be presented in this paper and will therefore not be discussed in this paper.

There are also constraints for which model selection based on encompassing priors is *not* objective. Consider, for example,

$$| \beta_1 - \beta_2 | \leq \xi. \tag{4.12}$$

If encompassing priors are used, this constraint represents the counterpart of traditional null-hypotheses like $H : \beta_1 = \beta_2$. The latter does not straightforwardly fit in the encompassing framework because both $1/A_r$ and $1/B_r$ are equal to zero. A possible solution is to use (4.12) with a small value for ξ. For a different context where this is used, see Berger and Delempady (1987, Section 2) where it is proposed that precise hypotheses are better represented as tests of, say,

$$H_0 : | \theta - \theta_0| \leq \xi \quad \text{versus} \quad H_1 : | \theta - \theta_0| \geq \xi,$$

where ξ is small. Like before, if $g(\beta)$ is vague then, $1/A_r$ is virtually independent of the choice of encompassing prior. However, this does not hold for $1/B_r$. Suppose $g(\beta) \sim U[-\tau, \tau]$, then $1/B_r \longrightarrow 0$ as $\tau \to \infty$. In other words, the larger τ becomes, the smaller the Bayes factor in favor of the unconstrained model. This phenomenon is similar to Lindley's paradox (see for example, Berger and Delempady, 1987 and Lee, 1997, pp. 130–131)) where the posterior probability of a model depends on the vagueness or variance of the prior distribution of the parameters involved.

To deal with constraints like (4.12), subjective choices with respect to the encompassing prior have to be made. That is, whether or not a model with the constraint $\beta_1 = \beta_2$ is $relevantly$ better than one where no constraints are imposed on the parameters can be determined by reformulating the former such that effect sizes are included. All a researcher has to do is to determine which difference between the parameters β_1 and β_2 is considered to be relevant.

To support our 'theorizing' with respect to sensitivity, in this paper in the examples in Section 3, sensitivity analysis will consist of two aspects. The first one is the sensitivity of results to the choice of prior distributions $g(\beta)$. Values for the parameters of our priors (the hyperparameters) will be varied and it will be illustrated how this affects the posterior model probabilities. In particular this will be accomplished by varying the prior variance δ^2. Note that the hyperparameters of the nuisance parameters $\boldsymbol{V}$ and σ^2 will not be varied. The second aspect of sensitivity involves investigating how the choice of threshold values ξ affects the posterior model probabilities. Note that the choice of threshold values is only relevant for models with constraints of the type $\beta_1 = \beta_2$ as the encompassing prior framework requires such constraints to be re-written in the form (4.12).

4.3 Examples

Building on the methodology developed in Section 4.2, the current section presents three illustrative examples. All the three examples are on longitudinal data. In Section 4.3.1, we present a study on milk protein content of cows randomised to three different diets. Section 4.3.2 is devoted to analyzing growth measurements of 16 boys and 11 girls. In Section 4.3.3, the example is about Sitka spruce trees grown under two different conditions, Ozone and Control over a two year period. The first two examples involve *between* subject conditions namely type of diet and gender respectively. The Sitka spruce tree example involves both *between* and *within* subjects conditions which are treatment and year, respectively.

4.3.1 The Milk Content Trial

These data are taken from Verbyla and Culis (1990) who obtained the data from a workshop at Adelaide University. In addition to the analyses in Verbyla and Culis (1990), other analyses of the available data can be found in Diggle (1990) and Diggle et al., (1994). The data consist of assayed protein content of milk samples taken from 79 Australian cows. The cows entered the experiment after calving and were randomly allocated to one of three diets: barley, lupins and a mixture of barley and lupins with 25, 27 and 27 cows in the three groups respectively. Measurements were taken weekly for 19 consecutive weeks and they consisted of the percentage protein content of milk samples. However, 42 cows had incomplete data largely towards the end of the experiment. For the analyses in this paper we will use the available data. For analyses dealing with the dropout process refer to Diggle and Kenward (1994, Chapter 11) and Verbeke and Molenberghs (2000, Chapter 24).

According to Verblya and Culis (1990), the first three weeks constituted a settling-in period and so in their analyses the data for the first three weeks were ignored. In the same spirit, in the analyses done in this paper the data for the first three weeks are also ignored. The main objective of the experiment was to describe the effect of diet on the milk protein content over time. Figure 4.1 shows the mean response profiles over the measurement period (ignoring the data for the first three weeks). The profiles suggest that the barley diet gives higher protein values than the barley-lupin diet, which in turn gives higher protein values than the lupins diet.

Let y_{jk} be the milk protein content measured for the jth cow in the kth week, for $k = 1, \ldots, K_j$ and $j = 1, \ldots, 79$. In addition, let bar_j, mix_j and lup_j be indicator variables defined to be 1 if the cow was assigned to the barley, mixed or lupins diet respectively, and 0 otherwise. Assuming that the average profiles for each group (barley, mixed and lupins) can be approximated by a quadratic function over time, the model to be used in the

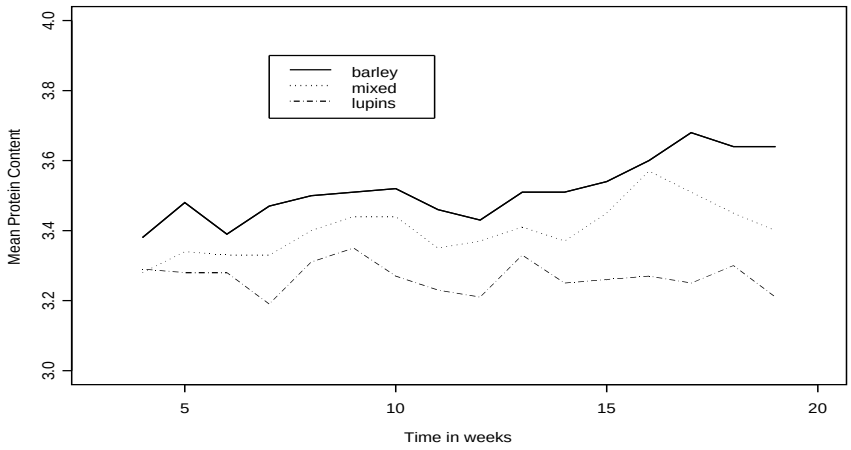

Figure 4.1: Milk Content Trial: Observed Mean Profiles.

analysis is:

$$y_{jk} = \pi_{1j} + \pi_{2j}(t_{jk} - 11.5) + \pi_{3j}(t_{jk} - 11.5)^2 + \varepsilon_{jk}, \tag{4.13}$$

where

$$\begin{cases} \pi_{1j} &= \beta_1 bar_j + \beta_2 mix_j + \beta_3 lup_j + u_{1j}, \\ \pi_{2j} &= \beta_4 bar_j + \beta_5 mix_j + \beta_6 lup_j + u_{2j}, \\ \pi_{3j} &= \beta_7 bar_j + \beta_8 mix_j + \beta_9 lup_j, \end{cases} \tag{4.14}$$

with

$$\mathbf{u} = (u_{1j}, u_{2j})^T \sim N(\mathbf{0}, \mathbf{V}), \quad \varepsilon_{jk} \sim N(0, \sigma^2).$$

and the time variable is centered at t = 11.5. The parameters β_1, β_2, and β_3 are the average intercepts for the cows on barley, mixed and lupins diets respectively at t=11.5 weeks. Further, the parameters β_4, β_5 and β_6 represent the average time effects (averaged over cows) for cows on barley, mixed and lupins diets respectively. Finally the parameters β_7, β_8, and β_9 represent the average time squared effects for cows on barley, mixed and lupins diets respectively. An initial analysis showed that there were hardly any quadratic time effects. Consequently in (4.13) and (4.14), we include random effects for only the intercepts and slopes for the linear time effect. We then seek to compare the following three models:

$$\begin{aligned} M_1 &: \quad \beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5,\ \beta_6,\ \beta_7,\ \beta_8,\ \beta_9 \\ M_2 &: \quad \{\beta_1 > \beta_2 > \beta_3\}, \{\beta_4 > \beta_5 > \beta_6\},\ \beta_7,\ \beta_8,\ \beta_9 \\ M_3 &: \quad \{\beta_1 \approx \beta_2 \approx \beta_3\}, \{\beta_4 \approx \beta_5 \approx \beta_6\}, \{\beta_7 \approx \beta_8 \approx \beta_9\} \end{aligned}$$

where M_1 is the encompassing model in which no constraints are imposed on the model parameters. Model M_2 investigates the following theory: at 11.5 weeks, the average protein content in the milk is highest, intermediate and lowest for the cows that received barley, mixed and lupins diets respectively. Further the average increase in percentage protein content per week is highest, intermediate and lowest for the cows that received barley, mixed and lupins diets respectively. This theory is translated into a statistical model by imposing order constraints on the intercept and slope parameters respectively. Note that the constraints in M_2 are of the type (4.11). The implication is that irrespective of the choice of encompassing prior, $1/B_r = 1/36$. Finally M_3 investigates the theory that there is no difference between the percentage protein content in the milk obtained from cows on each of the three diets. In other words, M_3 states that the type of diet does not affect the protein content of milk. To translate this theory into a model, the average, linear time and quadratic time parameters for the barley, mixed and lupins diets respectively are constrained to be about equal (see below for an elaboration).

Having specified the competing models, the next step is to specify the encompassing prior and subsequently compute the posterior model probabilities. To assess the sensitivity of our results to the encompassing prior specification for the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9)^T$, three different prior specifications will be used. These are displayed in Table 1. Note that parameters in the same group: intercepts, time slopes, and time squared slopes respectively have the same hyperparameters. This is in line with the first principle under specification of the encompassing prior (see Section 4.2.2.2). Further, note that the prior means μ are the same, that is the prior distributions vary only in their standard deviations δ; prior 1 is data based, while priors 2 and 3 are obtained by increasing the standard deviation of prior 1 by factors of 2 and 5 respectively. Turning to the prior distribution on $\boldsymbol{V}$, we take $\lambda = 1$ and $\boldsymbol{S} = \begin{pmatrix} 0.040 & 0.003 \\ 0.003 & 0.002 \end{pmatrix}$. Finally for the prior on σ^2 we set $\upsilon_0 = 1$ and $\sigma_0^2 = 0.04$. Note that these values of σ_0^2 and $\boldsymbol{S}$ are the estimates of σ^2 and $\boldsymbol{V}$ respectively obtained from the sample (see Section 4.2.2.2 on the specification of an encompassing prior).

A difference of $\xi = 0.05$ or less in the average percentage protein content and average increase in percentage protein content per week respectively for cows on each of the three diets is considered to be irrelevant and a good replacement for $\xi = 0$. Note that the observed values of protein content range from 2.45% to 4.44% percent. Consequently in order to obtain posterior probabilities for M_3, the constraints specifying M_3 will be replaced by constraints of the form (4.12) with $\xi = 0.05$. However, for purposes of a sensitivity analysis threshold values of $\xi = 0.08$ and $\xi = 0.10$ will also be considered.

Table 4.1: Prior specifications for β - Milk data.

		Prior 1	Prior 2	Prior 3
Parameter	μ	δ	δ	δ
β_1	3.3893	0.1876	0.3752	0.938
β_2	3.3893	0.1876	0.3752	0.938
β_3	3.3893	0.1876	0.3752	0.938
β_4	0.0065	0.011	0.022	0.055
β_5	0.0065	0.011	0.022	0.055
β_6	0.0065	0.011	0.022	0.055
β_7	-0.00013	0.00165	0.0033	0.00825
β_8	-0.00013	0.00165	0.0033	0.00825
β_9	-0.00013	0.00165	0.0033	0.00825

Table 4.2: Summary of results - Threshold = 0.05%.

	Prior 1		Prior 2		Prior 3	
Model	PP	BF_{1r}	PP	BF_{1r}	PP	BF_{1r}
M_1	0.0475		0.0464		0.0420	
M_2	0.9511	0.0499	0.9487	0.0489	0.8632	0.0486
M_3	0.0013	37.0291	0.0049	9.4416	0.0948	0.4429

Using the three different prior specifications with hyperparameters as shown in Table 4.1 in turn, coupled with the prior specifications for $\mathbf{V}$ and σ^2 as explained above, the Gibbs sampling algorithm (see Section 2.1) was used to obtain samples from the posterior distribution. For each prior specification the algorithm was run for 200 000 iterations after the initial 20 000 iterates were discarded as a burn-in. Further, 200 000 βs were drawn from each encompassing prior. Next the quantities $1/B_r$ and $1/A_r$ are estimated for models 2 and 3 using the proportion of the 200 000 βs sampled from prior and posterior distributions in agreement with constraints specified by the model at hand. Subsequently posterior probabilities are estimated using (4.9) and (4.10). The posterior model probabilities (PP) obtained for the different prior specifications and different thresholds are shown in Tables 4.2, 4.3, and 4.4. The Bayes factors (BF_{1r}) of M_1 to M_r for $r = 2$ and 3, respectively are also shown in the tables. Note that each model is assumed to have the same probability $a\ priori$, that is $P(M_i) = 1/3$ for $i = 1, 2, 3$.

As can be seen from the tables, the Bayes factor of M_1 to M_2 is virtually independent of the prior specifications. On the other hand, the Bayes factor of M_1 to M_3 does depend on the prior specification. These results lend support to our claim that if we use fairly vague (but proper) priors then for constraints of the type (4.11), the choice of encompassing prior does not influence $1/B_r$ while it does when it comes to constraints of the type

Table 4.3: Summary of results - Threshold = 0.08%.

	Prior 1		Prior 2		Prior 3	
Model	PP	BF_{1r}	PP	BF_{1r}	PP	BF_{1r}
M_1	0.0474		0.0462		0.0411	
M_2	0.9492	0.0499	0.9441	0.0489	0.8441	0.0486
M_3	0.0033	14.0782	0.0097	4.7719	0.1148	0.3576

Table 4.4: Summary of results - Threshold = 0.10%.

	Prior 1		Prior 2		Prior 3	
Model	PP	BF_{1r}	PP	BF_{1r}	PP	BF_{1r}
M_1	0.0473		0.0458		0.0391	
M_2	0.9458	0.0499	0.9349	0.0489	0.8039	0.0486
M_3	0.0069	6.8846	0.0194	2.3616	0.1570	0.2490

(4.12). Next, from the tables one can see that the posterior probability of model M_3 increases i.e., the model becomes more likely as the threshold value increases. This is not surprising. In Section 4.2.2.2, it was mentioned that in order to deal with equality constrained models, constraints of the form $\beta_p = \beta_{p'}$ would be replaced with $|\beta_p - \beta_{p'}| \leq \xi$ for some threshold value ξ. Increasing ξ leads to a decrease in the Bayes factor comparing the unconstrained model to the equality constrained model and equivalently an increase in the posterior model probability of the equality constrained model.

In summary, among the models presented, model M_2 is preferred. Note that irrespective of the choice of encompassing prior this model has the highest posterior probability. Table 4.5 shows the expected aposteriori (EAP) estimates, posterior standard deviations (PSD) and 95% central credibility intervals (CCI) for the parameters of M_2 obtained using constrained Gibbs sampling as explained in Section 4.2.1.

Model 2 predicts that the intercepts and slopes are ordered as $\beta_1 > \beta_2 > \beta_3$ and $\beta_4 > \beta_5 > \beta_6$, respectively. Inspection of the EAP estimates for these parameters in Table 4.5 reveals that β_1 and β_3 differ by 0.22 while β_4 and β_6 differ by 0.014. Note that the credibility intervals of β_1 and β_2 are not overlapping. The estimates of β_1, β_2 and β_3 indicate that after 11.5 weeks the average percentage protein contents in the milk samples are 3.50, 3.41 and 3.28 for the barley, mixed and lupins diet cows respectively. The estimate of β_4 is positive while those of β_5 and β_6 are negative. This suggests that over time, milk protein content is linearly increasing for the cows on barley diet and linearly decreasing for the cows on mixed and lupins diets. Further looking at the estimates of β_6, β_7 and β_8 one can conclude that for each of the three groups (barley, mixed and lupin) there is some

Table 4.5: Posterior summaries of the parameters in model M_2.

Parameter	EAP	PSD	CCI
β_1	3.5030	0.0358	(3.4346,3.5747)
β_2	3.4091	0.0339	(3.3420,3.4749)
β_3	3.2807	0.0355	(3.2108,3.3507)
β_4	0.0043	0.0052	(-0.0052,0.0150)
β_5	-0.0025	0.0045	(-0.0114,0.0065)
β_6	-0.0097	0.0051	(-0.0202,-0.0003)
β_7	-0.0019	0.00058	(-0.0031,-0.0008)
β_8	-0.0027	0.00057	(-0.0038,-0.0016)
β_9	-0.0020	0.00057	(-0.0031,-0.0009)
$\mathrm{Var}(u_{1j})$	0.0369	0.0068	(0.0257,0.0523)
$\mathrm{Cov}(u_{1j}, u_{2j})$	0.0027	0.0011	(0.0009,0.0050)
$\mathrm{Var}(u_{2j})$	0.0015	0.0003	(0.0010,0.0021)
σ^2	0.0403	0.0019	(0.0368,0.0442)

quadratic time effect pointing to a decrease in protein content over time. A similar phenomenon was also noticed by Verbeke and Molenberghs (2000, Chapter 24) when they performed a stratified analysis. However, as can be seen in Table 4.5, the EAP estimates of the quadratic time effects are very close to zero suggesting that the effects are quite small. In conclusion, diet affects protein content of milk, with the barley diet giving the highest protein content, followed by the mixed diet and then the lupins diet gives the lowest protein content.

4.3.2 Growth data

These data, taken from Potthoff and Roy (1964), consist of growth measurements for 16 boys and 11 girls and were collected by investigators at the University of North Carolina Dental School. For each subject, the distance (in millimeters) from the center of the pituitary to the pterygomaxillary fissure was recorded at the ages of 8, 10, 12 and 14 years, respectively. The research question is whether dental growth is related to gender. In the literature several authors have used these data for various analyses (see for example, Jennrich and Schluchter, 1986; Little and Rubin, 1987; Verbeke and Molenberghs, 2000). Figure 4.2 shows the observed mean profiles, one for each sex group. There seems to be a linear trend in both profiles and some indication that the profiles are not parallel, that is the profiles have different slopes. Let y_{jk} denote the response (distance) for the jth ($j = 1, \ldots, 27$) subject at age k= 8, 10, 12, 14. Also, let $\mathtt{Mal}_j$ and $\mathtt{Fem}_j$ be indicator variables defined to be 1 if the subject is a boy or girl respectively, and 0 otherwise. Assuming that the mean profile for each sex group can be represented by a

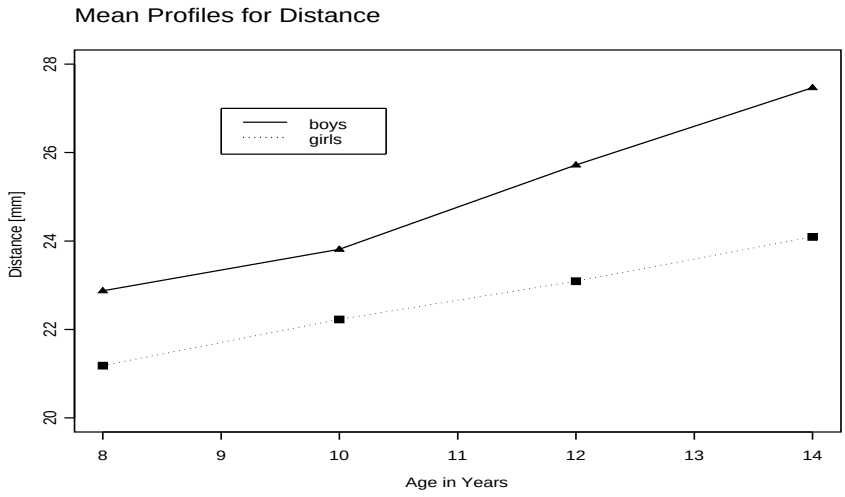

Figure 4.2: Growth Data: observed mean profiles.

linear function of time, the model to be used in the analysis is:

$$\begin{aligned} y_{jk} = \beta_{1m}(Mal_j)t_{1k} + \beta_{2m}(Mal_j)t_{2k} + \beta_{3m}(Mal_j)t_{3k} + \\ \beta_{4m}(Mal_j)t_{4k} + \beta_{1f}(Fem_j)t_{1k} + \beta_{2f}(Fem_j)t_{2k} + \\ \beta_{3f}(Fem_j)t_{3k} + \beta_{4f}(Fem_j)t_{4k} + u_{1j} + \varepsilon_{jk}, \end{aligned} \quad (4.15)$$

with $u_{1j} \sim N(0, V)$ and $\varepsilon_{jk} \sim N(0, \sigma^2)$. In (4.15), the variables t_{1k}, t_{2k}, t_{3k} and t_{4k} are defined to be 1 if the subject's age is 8, 10, 12 and 14 respectively, and 0 otherwise. This model renders a separate mean for each of the eight age $\times$ gender combinations. The parameters β_{1m}, β_{2m}, β_{3m} and β_{4m} represent the mean distance for the boys at age 8, 10, 12 and 14, respectively while the parameters β_{1f}, β_{2f}, β_{3f} and β_{4f} represent the corresponding mean distances for the girls. Further, the term u_{1j} for $j = 1, \ldots, 27$ represents the deviation of a subject's profile from the predicted mean profile.

The encompassing model for these data is one where no constraints are put on the βs:

$$M_1 : \beta_{1m},\ \beta_{2m},\ \beta_{3m},\ \beta_{4m},\ \beta_{1f},\ \beta_{2f},\ \beta_{3f},\ \beta_{4f}.$$

Of interest would be to look at the theory that boys grow faster than the girls. Three possible scenarios will be considered, leading to the the three models:

$$\begin{aligned} M_2 : &\quad \beta_{1m},\ \beta_{2m},\ \beta_{1f},\ \beta_{2f}, (\beta_{4m} - \beta_{3m}) > (\beta_{4f} - \beta_{3f}), \\ M_3 : &\quad \beta_{1m},\ \beta_{1f}, (\beta_{4m} - \beta_{3m}) > (\beta_{4f} - \beta_{3f}),\ (\beta_{3m} - \beta_{2m}) > (\beta_{3f} - \beta_{2f}), \end{aligned}$$

and

$$M_4: \quad (\beta_{4m} - \beta_{3m}) > (\beta_{4f} - \beta_{3f}),\ (\beta_{3m} - \beta_{2m}) > (\beta_{3f} - \beta_{2f}),\ (\beta_{2m} - \beta_{1m}) > (\beta_{2f} - \beta_{1f})$$

respectively. As can be seen M_2 states that from the age of 12 years, the slope of the profile for boys is bigger than for girls while model M_3 states that the profiles differ from the age of 10 years. Finally M_4 states that the profiles differ right from the age of 8 years. Note that models 2 to 4 all have constraints of the form (4.11).

In order to assess the sensitivity of results to the prior formulation for $\boldsymbol{\beta} = (\beta_{1m}, \ldots, \beta_{4m}, \beta_{1f}, \ldots, \beta_{4f})^T$, three different encompassing prior specifications will be used. The first encompassing prior is based on the data following principles 3 and 4 in Section 4.2.2.2. This renders a $N(24.206, 20.4)$ prior for each of the parameters in $\boldsymbol{\beta}$. Further, we assign V an inverse Wishart distribution with degrees of freedom 1 and scale parameter 4.23. Lastly, the parameter σ^2 is assigned a scaled inverse Chi square distribution with degree of freedom 1 and scale 1.44. The other two specifications for the encompassing prior are obtained by increasing the variance of the (common) normal prior for the parameters in $\boldsymbol{\beta}$ by factors of 4 and 9 respectively. Consequently, for the second and third encompassing prior each of the parameters in $\boldsymbol{\beta}$ are assigned $N(24.206, 81.6)$ and $N(24.206, 183.6)$ priors respectively. The prior specifications of the other parameters (V and σ^2) remain as they were specified under the first encompassing prior.

Subsequently 200 000 $\boldsymbol{\beta}$s (after a burn in of 20 000 iterates) were sampled from the encompassing prior and the corresponding posterior distributions for the three prior specifications respectively. For each prior and posterior, these samples were used to estimate the quantities $1/B_r$ and $1/A_r$, respectively which are then used to obtain Bayes factors and posterior model probabilities using the procedure presented in Section 4.2.2.1. In Tables 4.6 and 4.7, the results are displayed for the different prior specifications. It turns out that the Bayes factors are virtually independent of the choice of encompassing prior. The same holds for the posterior model probabilities. These results provide more support to our claim that for models with constraints of the form (4.11), the proportion of samples from the encompassing prior which are in agreement with a model at hand is virtually independent of the choice of the encompassing prior.

Further, the posterior probabilities show evidence in favor of model 4. However, although model 4 has the highest posterior probability, model 3 cannot be ruled out completely. These results favor the theory that boys grow faster than girls and that the differences between their dental growth rates are noticeable from the age of 10 years (models 3 and 4) and possibly 8 years on (model 4). In Table 4.8, the expected aposteriori (EAP) estimates, posterior standard deviations (PSD) and 95% central credibility intervals

Table 4.6: Summary of results - Bayes factors

	Prior 1	Prior 2	Prior 3
Bayes factor			
BF_{12}	0.5997	0.6012	0.6018
BF_{13}	0.2233	0.2242	0.2245
BF_{14}	0.1362	0.1384	0.1391

Table 4.7: Summary of results - Posterior probabilities (PP).

	Prior 1	Prior 2	Prior 3
Model	PP	PP	PP
M_1	0.0690	0.0697	0.0699
M_2	0.1151	0.1159	0.1162
M_3	0.3091	0.3109	0.3114
M_4	0.5068	0.5035	0.5025

(CCI) for the parameters (of model 4) obtained using constrained Gibbs sampling (see Section 4.2.1) are shown. From the table, the estimated mean vectors for boys and girls, respectively, suggest that on average, the dental growth of boys ranges from 22.69 to 27.47, while that of girls ranges from 21.53 to 24.10.

4.3.3 Sitka Spruce trees data

In this example data about growth of Sitka spruce trees are taken from Diggle et al. (1994). The data consist of measurements of log tree size (conventionally measured by the product of tree height and diameter squared) for 79 trees over two growing seasons (years 1988 and 1989). The objec-

Table 4.8: Posterior summaries of the parameters in model M_4.

Parameter	EAP	PSD	CCI
β_{1m}	22.689	0.559	(21.574,23.779)
β_{2m}	23.879	0.556	(22.772,24.967)
β_{3m}	25.681	0.560	(24.568,26.776)
β_{4m}	27.474	0.567	(26.355,28.589)
β_{1f}	21.530	0.662	(20.228,22.837)
β_{2f}	22.192	0.653	(20.907,23.479)
β_{3f}	23.185	0.659	(21.883,24.483)
β_{4f}	24.097	0.674	(22.765,25.420)
V	3.573	1.216	(1.837,6.522)
σ^2	2.015	0.334	(1.465,2.770)

tive of the study was to assess the effect of ozone pollution on tree growth. Of the 79 trees, 54 trees were grown with ozone exposure at 70 ppb and the remaining 25 trees were grown under control conditions. In this paper, interest lies in comparing the growth pattern of trees under the two treatments (ozone and control) as well as the two seasons (1988 and 1989). Figure 4.3 shows the observed profiles of the 79 trees over the two growing

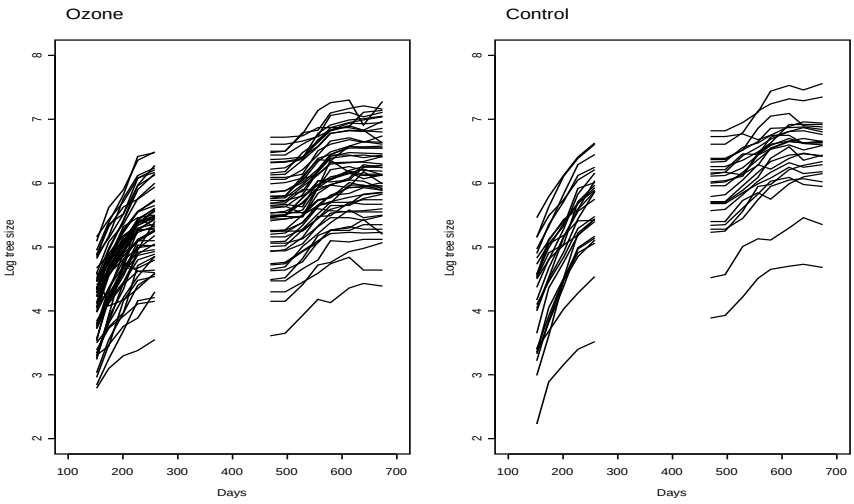

Figure 4.3: Sitka spruce trees. Growth curves of 79 trees over two growing seasons for ozone treated and control

seasons. The horizontal axis shows the time in days since the 1st of January 1988. In each panel (ozone and control), the first set of profiles corresponds to the year 1988 while the second set of profiles corresponds to the year 1989. The graphs make apparent some patterns. First, all trees are growing. Second, the general trend is that a profile tends to be high or low. That is, trees which have a high (log tree) size at the beginning of the observation period tend to be highest throughout. This suggests that the random effects structure is probably confined to the random intercept.

Let y_{jk} be the log tree size measured for the jth tree at time t_{jk}, for $k = 1, \ldots, K_j$ and $j = 1, \ldots, 79$ where t_{jk} is a scaled time variable defined as t_{jk} = (original time in days - 200)/100 and t_{jk} = (original time in days - 570)/100 for the years 1988 and 1989 respectively. This transformation (centering) of the original time implies that the average intercepts correspond to the log tree size after 200 and 570 days, respectively from 1st January 1988. The division by 100 is done to avoid small regression coefficients later in the estimation. Furthermore, let S_1 and S_2 be indicator variables defined to be 1 if a day belongs to 1988 or 1989 respectively, and 0 otherwise. And lastly, let C_j and E_j be indicator variables defined to be 1 if the tree was grown in

control or ozone conditions respectively, and 0 otherwise. Assuming that the average profiles for each treatment (ozone or control) by year (1988 or 1989) can be approximated by a linear function over time (which is reasonable given the profiles in figure 3), then the model to be used in the analysis is:

$$y_{jk} = \pi_{1j}S_{1k} + \pi_{2j}S_{2k} + u_j + (\pi_{3j}S_{1k} + \pi_{4j}S_{2k})t_{jk} + \varepsilon_{jk},$$

in which

$$\begin{aligned}
\pi_{1j} &= \beta_1 C_j + \beta_2 E_j, \\
\pi_{2j} &= \beta_3 C_j + \beta_4 E_j, \\
\pi_{3j} &= \beta_5 C_j + \beta_6 E_j, \\
\pi_{4j} &= \beta_7 C_j + \beta_8 E_j,
\end{aligned}$$

and

$$u_j \sim N(0, V), \quad \varepsilon_{jk} \sim N(0, \sigma^2).$$

For these data the following models will be compared:

$$\begin{aligned}
M_1 &: \quad \beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5,\ \beta_6,\ \beta_7,\ \beta_8, \\
M_2 &: \quad \{\beta_1 > \beta_2\}, \{\beta_3 > \beta_4\}\{\beta_5 > \beta_6\}, \{\beta_7 > \beta_8\}, \\
M_3 &: \quad \{\beta_1 \approx \beta_2\}, \{\beta_3 \approx \beta_4\}\{\beta_5 \approx \beta_6\}, \{\beta_7 \approx \beta_8\}.
\end{aligned}$$

M_1 is the encompassing model where no constraints are imposed on the model parameters, M_2 identifies with the theory that Ozone suppresses growth of trees: in this case within each year (1988 or 1989), the intercept and slope parameters for the ozone group are constrained to be smaller than their counterparts in the control group. The constraints in M_2 are of the type (4.11) and consequently irrespective of the choice of encompassing prior, $1/B_r = 1/16$. Lastly M_3, states that there is no treatment effect i.e growth is not affected by the treatment (ozone). So in this model the intercept and slope parameters relating to the same season are constrained to be about the same for ozone and control groups respectively.

In the same spirit as in Sections 4.3.1 and 4.3.2, a sensitivity analysis will be done. Three different encompassing priors will be used for the parameter vector $\boldsymbol{\beta}$. The hyperparameters of these priors are displayed in Table 4.9. To complete the encompassing prior specification: the variance parameter V will be assigned an inverse Wishart distribution with degree of freedom 1 and scale parameter 0.402 and the residual variance σ^2 assigned a scaled inverse chi square distribution with degree of freedom 1 and scale 0.196. Like in Sections 4.3.1 and 4.3.2, the values of μ and δ for Prior 1 (see Table 4.9) and the values of the scale parameters specified above are all data based. Priors 2 and 3 are obtained by increasing the standard deviations

Table 4.9: Prior specifications for β - Trees data.

		Prior 1	Prior 2	Prior 3
Parameter	μ	δ	δ	δ
β_1	4.885	0.306	0.612	0.918
β_2	4.885	0.306	0.612	0.918
β_3	6.095	0.379	0.758	1.137
β_4	6.095	0.379	0.758	1.137
β_5	1.323	0.186	0.372	0.558
β_6	1.323	0.186	0.372	0.558
β_7	0.371	0.042	0.084	0.126
β_8	0.371	0.042	0.084	0.126

Table 4.10: Summary of results - Threshold = 0.06.

	Prior 1		Prior 2		Prior 3	
Model	PP	BF_{1r}	PP	BF_{1r}	PP	BF_{1r}
M_1	0.1014		0.0984		0.0823	
M_2	0.8982	0.1129	0.8946	0.1100	0.7531	0.1093
M_3	0.0004	244	0.0070	14	0.1646	0.5

of the normals in Prior 1 by factors of 2 and 3 respectively. As before the hyperparameters for V and σ^2 are not varied.

A difference of $\xi = 0.06$ or less in the average log size of trees grown under control or ozone conditions is considered to be irrelevant and therefore a good replacement for $\xi = 0$. Note that the observed values of log size range between 2 and 8. Also, a difference of 0.06 or less in the average increase in log size (over a 100 day period) of trees grown under control or ozone conditions is considered to be irrelevant. Consequently in order to obtain posterior probabilities for M_3, constraints specifying M_3 will be replaced with constraints of the form (4.12), with $\xi = 0.06$. However, for purposes of a sensitivity analysis threshold values of 0.08 and 0.10 will also be considered.

As before, for each prior specification, the quantities $1/B_r$ and $1/A_r$ are estimated for models 2 and 3 using the proportion of samples from prior and posterior distributions in agreement with constraints specified by the model at hand. Subsequently posterior probabilities are estimated using (4.9) and (4.10).

The posterior model probabilities (PP) obtained from using the different priors and different thresholds are shown in Tables 4.10, 4.11 and 4.12 respectively. Further, the Bayes factors (BF_{1r}) of M_1 to M_r for $r = 2$ and 3, respectively are also shown in the tables. From the tables two things are evident. First, the Bayes factor of M_1 to M_2 does not depend on the prior.

Table 4.11: Summary of results - Threshold = 0.08.

	Prior 1		Prior 2		Prior 3	
Model	PP	BF_{1r}	PP	BF_{1r}	PP	BF_{1r}
M_1	0.1011		0.0960		0.0825	
M_2	0.8959	0.1129	0.8731	0.1100	0.7526	0.1093
M_3	0.0030	33.9	0.0309	3.1111	0.1649	0.1111

Table 4.12: Summary of results - Threshold = 0.10.

	Prior 1		Prior 2		Prior 3	
Model	PP	BF_{1r}	PP	BF_{1r}	PP	BF_{1r}
M_1	0.1000		0.0906		0.0638	
M_2	0.8859	0.1129	0.8235	0.1100	0.5822	0.1093
M_3	0.0141	7.0680	0.0860	1.0538	0.3539	0.0263

Second, the Bayes factor of M_1 to M_3 does depend on the prior. These findings are similar to those in Sections 4.3.1 and 4.3.2 and consequently strengthen our claim that if we use fairly vague (but proper) priors then for constraints of the type (4.11), the choice of encompassing prior does not influence $1/B_r$ while it does when it comes to constraints of the type (4.12). Further, from the tables one can conclude that the posterior probability of model M_3 increases i.e the model becomes more likely as the threshold value increases. A similar phenomenon was noticed in Section 4.3.1 and an explanation was given for it.

Among the models presented, model M_2 is preferred since it has the highest posterior probability. Table 4.13 shows the expected aposteriori (EAP) estimates , posterior standard deviations (PSD) and 95% central credibility intervals (CCI) for the parameters of model 2. The EAP estimates of

Table 4.13: Posterior summaries of the parameters in model M_2.

Parameter	EAP	PSD	CCI
β_1	4.948	0.102	(4.760,5.156)
β_2	4.753	0.074	(4.605,4.895)
β_3	6.232	0.102	(6.042,6.441)
β_4	5.890	0.074	(5.742,6.031)
β_5	1.410	0.045	(1.321,1.497)
β_6	1.204	0.031	(1.144,1.265)
β_7	0.380	0.015	(0.353,0.411)
β_8	0.359	0.012	(0.335,0.382)
V	0.384	0.063	(0.280,0.526)
σ^2	0.038	0.002	(0.035,0.042)

β_1 and β_2 indicate that 200 days since 1st January 1988, the average log sizes of trees grown in control and ozone conditions were 4.95 and 4.75 units respectively. Next, the estimates of β_5 and β_6 suggest that for every 100 extra days, the log size of trees increases by 1.4 and 1.2 units for trees grown in control and ozone conditions respectively. Turning to the year 1989, EAP estimates of β_3 and β_4 indicate that 570 days since 1st January 1988, the average log sizes of trees grown in control and ozone conditions were 6.23 and 5.89 units respectively. Finally the estimates of β_7 and β_8 suggest that for every 100 extra days, the log size of trees increases by 0.38 and 0.36 units for trees grown in control and ozone conditions respectively. Model 2 implies that Ozone suppresses the growth of trees. Looking at the EAP estimates obtained for model 2, it can also be concluded that the rate of growth (of the trees) is higher for the year 1988 than for 1989.

4.4 Discussion

In this paper, it has been shown that the Bayesian approach can be successfully applied to constrained hierarchical problems. We adopted a fully Bayesian approach to model selection and estimation that is flexible and allows users to incorporate information regarding expert opinion or some theory through a prior distribution. Using samples from the prior and (truncated) posterior distributions, it was possible to obtain the Bayesian estimates of model parameters, Bayes factors and posterior probabilities for each model. For a given data set, the posterior probabilities provide a natural way for researchers to arrive at a model or a smaller set of models that describe the data best out of a finite set of competing models. Much as all the examples in the paper consisted of longitudinal data, the methodology can also be applied on other multilevel data settings.

The advantage of the Bayesian approach is that more precise inferences are possible since all available information is incorporated into the analysis, that is practical model constraints are incorporated into the analysis via the prior distribution. Furthermore, incorporation of constraints on the model parameters offers no additional computational burden on the Gibbs sampler. The only difference between the constrained and the unconstrained (encompassing) model is that each regression coefficient is sampled from a truncated normal distribution instead of a normal distribution. In contrast, inferences for constrained problems in a non-Bayesian way often require specialized algorithms tailored for particular models and are therefore not easily generalizable. This has necessitated a vast literature on order-restricted estimators for different cases. Therefore, even for researchers more comfortable with classical methods, the Bayesian approach proposed in this paper should provide a useful machinery and hence an addition to their toolkit.

As noted by Gelfand et al. (1992, p. 526) and Chen and Shao (1998, p. 74), there is remarkably little literature on Bayesian approaches to constrained parameter problems let alone constrained hierarchical model problems. Mukerjee and Tu (1995) also note that the problem of order-restricted inference on regression coefficients has received relatively little attention. This paper is an addition to the available literature.

Appendix: Sampling from the Posterior distribution using the Gibbs Sampler

Starting with initial values of the model parameters, each iteration of the Gibbs sampler consists of four steps:

1. Sample $\mathbf{u}_j$ for $j = 1 \ldots, J$ from $N_2(\mathbf{\Phi}_j, \mathbf{\Sigma}_j)$ where

$$\mathbf{\Phi}_j = \left[\frac{\sum_{k=1}^{K_j} \mathbf{z}_{jk} \mathbf{z}_{jk}^T}{\sigma^2} + \mathbf{V}^{-1} \right]^{-1}$$

   and

$$\mathbf{\Sigma}_j = \frac{\mathbf{\Phi}_j}{\sigma^2} \sum_{k=1}^{K_j} \mathbf{z}_{jk}^T (y_{jk} - \mathbf{x}_{jk}\boldsymbol{\beta}).$$

2. Sample σ^2 from a scaled inverse chi square distribution with degrees of freedom $\upsilon_0 + \sum_{j=1}^{J} K_j$ and scale

$$\upsilon_0 \sigma_0^2 + \sum_{j=1}^{J} \sum_{k=1}^{K_j} (y_{jk} - \mathbf{x}_{jk}\boldsymbol{\beta} - \mathbf{z}_{jk}\mathbf{u}_j)^2.$$

3. Sample $\mathbf{V}$ from an Inverse Wishart distribution with degrees of freedom $\lambda + J$ and scale matrix

$$\sum_{j=1}^{J} \mathbf{u}_j^T \mathbf{u}_j + \mathbf{S}.$$

4. Sample β_p from a normal distribution with mean

$$\frac{\frac{\mu_p}{d_p^2} + \frac{\sum_{j=1}^{J} \sum_{k=1}^{k_j} \left[y_{jk} - \sum_{\substack{i=1 \\ i \neq p}}^{P} \beta_i x_{ijk} - \sum_{q=1}^{Q} u_{qj} z_{qjk} \right] x_{pjk}}{\sigma^2}}{\frac{1}{d_p^2} + \frac{\sum_{j=1}^{J} \sum_{k=1}^{K_j} x_{pjk}^2}{\sigma^2}}$$

and variance

$$\frac{1}{\frac{1}{d_p^2} + \frac{\sum_{j=1}^{J}\sum_{k=1}^{K_j} x_{pjk}^2}{\sigma^2}}.$$

Chapter 5

Posterior Predictive Model Selection in Inequality Constrained Hierarchical Models

Abstract

Model Selection is a fundamental activity in the analysis of data sets. This has become increasingly more important as computational advances enable the fitting of increasingly complex models. In this paper we explore predictive model selection in inequality constrained hierarchical models with random effects. Model selection is done by choosing the model that would best predict the observed data. The best prediction is defined via a criterion measuring the discrepancy between the observed outcome variable and the predictive distribution for its future counterpart. The potential of predictive model selection lies in its ability to handle non-nested models.

5.1 Introduction

Constrained parameter problems arise in a wide variety of applications (see for example, Barlow et al., 1972; Robertson et al., 1988; Journal of Statistical Planning and Inference (2002), Volume 107, Issues 1-2, pages 1-378). Like in many applications of statistics, the comparison of alternative scientific explanations via hypotheses is an important issue in the area of constrained parameter problems. The classical approach to hypothesis testing is one where a researcher states a null hypothesis H_0 and an alternative H_A, defines an appropriate test statistic and then computes a p-value. The decision to reject or not to reject the null hypothesis is then based on comparing the p-value with some prescribed level of significance, say 0.01 or 0.05. Existing approaches to hypothesis testing make use of the likelihood ratio statistic which has been discussed widely in the literature (especially for testing hypotheses where there are no constraints on the model parameters).

There is evidence (see for example Robertson et al., 1998, Chapter 1) that taking order restrictions into account can improve the efficiency of a statistical analysis by reducing the error or expected error of estimates or by increasing the power of the test procedures provided that the hypothesized order restriction actually holds. Testing hypotheses with constraints on the model parameters has also been discussed in books as well as in journal articles (see for example, Barlow et al., 1972; Robertson et al., 1998; Mukerjee and Tu, 1995 and Dunson and Neelon, 2003). As an example, consider the one way Anova problem with three groups. Let μ_1, μ_2 and μ_3 denote the mean in each group, respectively. Most of the (frequentist) literature on constrained parameter problems deals with comparing a model $M_1 : \mu_1 = \mu_2 = \mu_3$ with $M_2 : \mu_1 > \mu_2 > \mu_3$ and/or a model $M_2 : \mu_1 > \mu_2 > \mu_3$ with $M_3 : \mu_1, \mu_2, \mu_3$ via hypothesis testing.

The classical hypothesis testing procedure has some limitations. First, the approach can be applied only when the two models in question are nested, one within the other. However many practical situations involve a choice between two (or more) models that are not nested. Secondly, testing hypotheses with constraints on the parameters using frequentist methodology is often difficult or even impossible. To elaborate on this, in Anova situations, using the likelihood ratio principle, a test statistic $\bar{E}^2$, has been derived to test the null hypothesis that a set of independent normal populations have equal means and common but unknown variance against an ordered alternative hypothesis. The interested reader is referred to Robertson et al., (1998, Chapter 2). It turns out the null distribution of this statistic is a mixture of beta distributions in which the mixing coefficients depend on sample sizes. Even in simple problems when you have groups with unequal sample sizes, computation of the mixing coefficients can be

tedious. Several authors have developed and compared approaches for approximating the mixing coefficients (see for example, Wright and Tran, 1985; Futschik and Pflug, 1998; Sen and Silvapulle, 2002 and Dobler, 2002). These approximating procedures are subsequently used to derive (asymptotic) distributions of the test statistic. However this has been done for relatively simple models. Results for hierarchical models have not been published either because they are intractable or they have not been actively studied.

As an alternative to hypothesis testing, several information criteria for model selection have been proposed in the literature. Some examples are Akaike information criterion (AIC) (Akaike, 1974), Schwarz's information criterion (BIC) (Schwarz, 1978), CAIC (Bozdogan, 1987), Deviance information criterion (DIC) (Spiegelhalter et al., 1998) and the Bayes factor (Kass and Raftery, 1995). The first four are all penalised likelihood criteria while the Bayes factor is a ratio of marginal likelihoods. The criteria AIC, BIC, CAIC and DIC all consist of two terms, one representing the 'goodness of fit' and the other a penalty for increasing model complexity. The penalty term for the first three criteria (AIC, BIC and CAIC) is some function of the number of parameters in the model while that of DIC is some function of the deviance (see Carlin and Louis 2000, p. 221). For the models considered in this paper it is quite difficult to quantify the number of parameters in the models and consequently this precludes the use of AIC, BIC, CAIC as criterion for model selection. Anraku (1999) proposed an order-restricted information criterion (ORIC) for parameters under a simple order restriction. Still, computation of ORIC for the models to be considered in this paper is rather intractable. Potentially the DIC can also be computed for the models considered in this paper. However, we will work in predictive space and subsequently use posterior predictive quantities for model selection. Using posterior predictive model selection has several potential advantages. First, it enables the comparison of non-nested models. Secondly, there is no need to count the number of parameters in the models. Furthermore, it is possible to compute the variances of the criterion values obtained hence the ability to quantify the uncertainty in the criterion values. As far is known by the authors, this has not been done for other model selection criteria e.g., AIC and BIC. Another attractive aspect of posterior predictive model selection is that if a non informative prior distribution is used, then the likelihood would dominate the posterior distribution. As a result there would be no need to worry about the sensitivity of the results to the prior distribution. This would not be the case when Bayes factors are used for model selection.

This paper deals with posterior predictive model selection in hierarchical linear models with inequality constraints on the model parameters. In Section 5.2, the methodology is introduced and described. Examples to illustrate the method are presented in Section 5.3. In Section 5.4, results from

a simulation study to investigate the frequency properties of the method are presented. The paper is concluded with a discussion in Section 5.5.

5.2 Methodology

Hierarchical linear models are widely used to model continuous longitudinal data. Consider a set of J subjects followed over a time period. The jth subject provides a set of (possibly partly missing) longitudinal measurements y_{jk} ($k = 1, \ldots, K_j$) at times t_{jk} ($k = 1, \ldots, K_j$). Then the sequence of measurements $y_{j1}, \ldots, y_{jK_j}$ for the jth subject is modelled as

$$y_{jk} = \boldsymbol{x}_{jk}^T \boldsymbol{\beta} + \boldsymbol{z}_{jk}^T \boldsymbol{u}_j + \varepsilon_{jk}, \tag{5.1}$$

where the vectors $\boldsymbol{x}_{jk}$ and $\boldsymbol{\beta}$ represent possibly time-varying explanatory variables and their corresponding regression coefficients, respectively. The $\boldsymbol{u}_j$ are vectors of random effects corresponding to the explanatory variables $\boldsymbol{z}_{jk}$ (which are usually a subset or a linear combination of $\boldsymbol{x}_{jk}$) and are typically modelled as independent and identically distributed $N(\boldsymbol{0}, \boldsymbol{V})$ random variables where $\boldsymbol{V}$ is a covariance matrix. Lastly ε_{jk} is a sequence of mutually independent measurement errors. It follows from (5.1) that the likelihood function of the data is

$$\prod_{j=1}^{J} \int_{\boldsymbol{u}_j} \{ \prod_{k=1}^{K_j} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_{jk} - \boldsymbol{x}_{jk}^T \boldsymbol{\beta} - \boldsymbol{z}_{jk}^T \boldsymbol{u}_j)}{2\sigma^2}) \} N(\boldsymbol{u}_j \mid \boldsymbol{0}, \boldsymbol{V}) \, d\boldsymbol{u}_j. \tag{5.2}$$

Suppose that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ is a vector of length P. Our interest focuses on assessing evidence of a decreasing (or increasing) trend in (a subset of) the β_ps, for $p = 1, \ldots, P$. Often the goal of an analysis is not only to estimate the model parameters subject to the constraints, but also to assess evidence with respect to equalities in the βs versus some ordering. For example, we would want to compare the models $M_1 : \beta_1 = \beta_2 = \cdots = \beta_P$, $M_2 : \beta_1 \leq \beta_2 \leq \cdots \leq \beta_P$ (with at least one strict inequality) and the unrestricted model $M_3 : \beta_1, \beta_2, \ldots, \beta_P$. We exemplify this in the examples in Section 5.3.

As there can be several competing models, it is imperative to find a parsimonious model which fits the data best. To this end, we will evaluate the performance of a model by comparing what it predicts to what has been observed. Ideally we will work in predictive space, and a good model among those under consideration, should be one that makes predictions close to what has been observed for an identical experiment. In the spirit of Laud and Ibrahim (1995) and Gelfand and Ghosh (1998), given a set of competing models M_r ($r = 1, \ldots, R$), we will choose the model M_r that minimizes the posterior predictive discrepancy measure

$$W_r^2 = E\{(\boldsymbol{y}^{rep} - \boldsymbol{y}^{obs})^T (\boldsymbol{y}^{rep} - \boldsymbol{y}^{obs}) | \, \boldsymbol{y}^{obs}, \boldsymbol{x}, \boldsymbol{z}, M_r\}. \tag{5.3}$$

In (5.3), $\mathbf{y}^{obs} = \{y_{jk} : k = 1, \ldots, K_j \; ; \; j = 1, \ldots, J\}$ is the observed data and $\mathbf{y}^{rep}$ is the replicated data, the data we would get if the experiment were replicated with the same model and the same value of $\theta = (\boldsymbol{\beta}, \mathbf{V}, \sigma^2, \mathbf{u}_j)$ that produced $\mathbf{y}^{obs}$. Note that with respect to $\mathbf{y}^{obs}$ and $\mathbf{y}^{rep}$, the explanatory variables $\mathbf{x}_{jk}$ and $\mathbf{z}_{jk}$ are fixed, that is they are identical for both $\mathbf{y}^{obs}$ and $\mathbf{y}^{rep}$. The expectation in (5.3) is taken with respect to the posterior predictive distribution associated with $\mathbf{y}^{rep}$ under model M_r, i.e.,

$$p(\mathbf{y}^{rep}|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r) = \int p(\mathbf{y}^{rep}|\, \theta_r, \mathbf{x}, \mathbf{z}, M_r) p(\theta_r|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r) d\theta_r,$$

where θ_r is the vector of parameters for model M_r and $p(\theta_r|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r)$ is the posterior distribution of θ_r under model M_r. Computation of W_r^2 for each model M_r will be done using simulation-based model fitting:

1. Sample $\theta_r^1, \ldots, \theta_r^Q$ from the posterior distribution $p(\theta_r|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r)$. The Gibbs sampling algorithm in the Appendix will be used to sample from this posterior distribution.

2. Sample $\mathbf{y}_q^{rep}$ from $p(\mathbf{y}^{rep}|\, \theta_r^q, \mathbf{x}, \mathbf{z}, M_r)$ for $q = 1, \ldots, Q$. This renders a sample $\mathbf{y}_1^{rep}, \ldots, \mathbf{y}_Q^{rep}$ from $p(\mathbf{y}^{rep}|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r)$.

Subsequently an estimate of W_r^2 is

$$\frac{1}{Q} \sum_{q=1}^{Q} (\mathbf{y}_q^{rep} - \mathbf{y}^{obs})^T (\mathbf{y}_q^{rep} - \mathbf{y}^{obs}).$$

Note that $W_r = \sqrt{W_r^2}$ is a distance measure on the response axis, measured in the same units as the response variable $\mathbf{y}^{obs}$. So effectively W_r measures how 'close' the observed data $\mathbf{y}^{obs}$ is to the posterior predictive distribution for each model M_r. So, it follows that good models will have small values of W_r. To quantify the variability in the criterion value W_r, one can compute the standard deviation of W_r for each model M_r. For each model, estimates for the variance of W_r^2 can also be obtained as output via the simulation based procedure outlined above: that is, using each replicate $\mathbf{y}_q^{rep}$ and $\mathbf{y}^{obs}$, the quantity $W_{r,q}^2 = (\mathbf{y}_q^{rep} - \mathbf{y}^{obs})^T (\mathbf{y}_q^{rep} - \mathbf{y}^{obs})$ is computed for each $q = 1, \ldots, Q$. The variance of these Q estimates is an estimate of the variance of W_r^2. Subsequently, applying the delta method (Rice, 1995, Section 4.6) renders an estimate of the variance of W_r:

$$\mathrm{Var}(W_r) = \frac{\mathrm{Var}(W_r^2)}{4 \times W_r^2}.$$

Note that the estimate measures the variability in W_r given that model M_r is true.

The discrepancy measure W_r^2 can be decomposed into

$$[E(\mathbf{y}^{rep}) - \mathbf{y}^{obs}|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r]^2 + Var(\mathbf{y}^{rep}|\, \mathbf{y}^{obs}, \mathbf{x}, \mathbf{z}, M_r), \qquad (5.4)$$

(see Laud and Ibrahim (1995), p. 249) where the first term involves the means of the replicates while the other involves the variances. This implies that the model selection criterion consists of two terms, namely a goodness of fit measure $[E(\mathbf{y}^{rep}) - \mathbf{y}^{obs}]^2$ and a term $Var(\mathbf{y}^{rep})$ which penalizes for model complexity; in the sequel we notate these two terms as FIT and PEN respectively. Gelfand and Ghosh (1998) noted that for very simple models, the first term in (5.4) will be large. For both very simple and very complex models, the penalty term in (5.4) will also be large due to large predictive variances. However as models become more complex, the first term will decrease (since the models will conform to the data better) but the second term will begin to increase. The consequence is that model complexity is automatically penalised hopefully leading to selection of a parsimonious and well fitting model. Note that unlike in other model selection criteria, the penalty term in (5.4) does not necessarily increase with model complexity. Put another way, a simple model can have a higher penalty than a complex model.

For the analyses we will use proper priors, but with hyperparameters chosen so that the priors will have minimal impact relative to the data. Put another way, the priors will be chosen so that the Bayesian analysis mimics a corresponding likelihood analysis. Using conjugate prior specifications and assuming independence between the model parameters, the prior to be used in the analyses is

$$g(\boldsymbol{\beta}, \mathbf{V}, \sigma^2) = \text{Inv-W}(\mathbf{V}|\, \lambda, \mathbf{S}) \times \text{Inv-}\chi^2(\sigma^2|\, \upsilon_o, \sigma_o^2) \times \prod_{p=1}^{P} N(\beta_p|\, \mu_p, d_p^2). \qquad (5.5)$$

In (5.5), $\text{Inv-}\chi^2(\sigma^2|\, \upsilon_o, \sigma_o^2)$ denotes a scaled inverse Chi-square distribution with degrees of freedom υ_o and scale σ_o^2, $\text{Inv-W}(\mathbf{V}|\, \lambda, \mathbf{S})$ denotes an Inverse Wishart distribution with degrees of freedom λ and scale matrix $\mathbf{S}$ and $N(\beta_p|\, \mu_p, d_p^2)$ denotes a Normal distribution with mean μ_p and standard deviation d_p^2. For the models considered in this paper, the posterior distribution is obtained as the product of the expressions in (5.2) and (5.5). To obtain samples from the posterior distribution and compute the posterior predictive quantity of interest W_r^2 for each model M_r, an MCMC algorithm will be needed. The Gibbs sampler is well suited to the models discussed in the paper because the full conditional distributions for the model parameters are standard distributions. An outline of the Gibbs sampling procedure is presented in the Appendix. The following specifications for the hyperparameters of the prior distributions will be used: $\upsilon_o = -2$ and $\sigma_o^2 = 0$; $\lambda = -3$ and $\mathbf{S} = \mathbf{0}$; $\mu_p = 0$ and $d_p^2 = 1000$ for the model parameters σ^2, $\mathbf{V}$ and β_p, respectively. Further, for all computations we will

use MCMC sampling chains of 10 000 iterations, following a 5000 iteration 'burn-in' period.

5.3 Examples

Applying the methodology presented in Section 5.2, the current section presents two illustrative examples. In Section 5.3.1, we present a study on milk protein content of cows randomised to three different diets. Section 5.3.2 is devoted to analyzing growth measurements of male Wistar rats randomised to three groups.

5.3.1 The Milk Content Trial

These data are taken from Verbyla and Culis (1990) who obtained the data from a workshop at Adelaide University. In addition to the analyses in Verbyla and Culis (1990), other analyses of the available data can be found in Diggle (1990) and Diggle et al., (1994). The data consist of assayed protein content of milk samples taken from 79 Australian cows. The cows entered the experiment after calving and were randomly allocated to one of three diets: barley, lupins and a mixture of barley and lupins with 25, 27 and 27 cows in the three groups respectively. Measurements were taken weekly for 19 consecutive weeks and they consisted of the percentage protein content of milk samples. However, 42 cows had incomplete data largely towards the end of the experiment. For the analyses in this paper we will use the available data. For analyses dealing with the dropout process refer to Diggle and Kenward (1994, Chapter 11) and Verbeke and Molenberghs (2000, Chapter 24).

According to Verblya and Culis (1990), the first three weeks constituted a settling-in period and so in their analyses the data for the first three weeks were ignored. In the same spirit, in the analyses done in this paper the data for the first three weeks are also ignored. The main objective of the experiment was to describe the effect of diet on the milk protein content over time. Figure 5.1 shows the mean response profiles over the measurement period (ignoring the data for the first three weeks). The profiles suggest that the barley diet gives higher protein values than the barley-lupin diet, which in turn gives higher protein values than the lupins diet.

Let y_{jk} be the milk protein content measured for the jth cow in the kth week, for $k = 1, \ldots, K_j$ and $j = 1, \ldots, 79$. In addition, let bar_j, mix_j and lup_j be indicator variables defined to be 1 if the cow was assigned to the barley, mixed or lupins diet respectively, and 0 otherwise. Assuming that the average profiles for each group (barley, mixed and lupins) can be approximated by a quadratic function over time, the model to be used in the

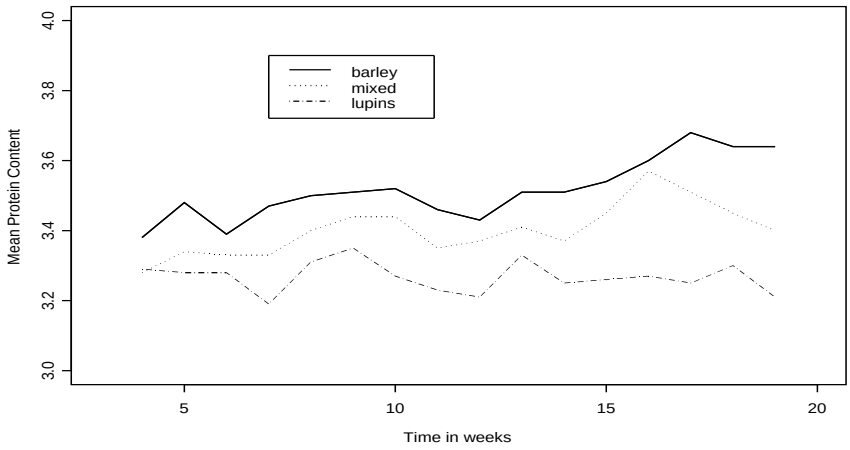

Figure 5.1: Milk Content Trial: Observed Mean Profiles.

analysis is:

$$y_{jk} = \pi_{1j} + \pi_{2j}(t_{jk} - 11.5) + \pi_{3j}(t_{jk} - 11.5)^2 + \varepsilon_{jk}, \tag{5.6}$$

where

$$\begin{aligned} \pi_{1j} &= \beta_1 bar_j + \beta_2 mix_j + \beta_3 lup_j + u_{1j}, \\ \pi_{2j} &= \beta_4 bar_j + \beta_5 mix_j + \beta_6 lup_j + u_{2j}, \\ \pi_{3j} &= \beta_7 bar_j + \beta_8 mix_j + \beta_9 lup_j, \end{aligned} \tag{5.7}$$

with

$$\mathbf{u} = (u_{1j}, u_{2j})^T \sim N(\mathbf{0}, \mathbf{V}), \quad \varepsilon_{jk} \sim N(0, \sigma^2).$$

and the time variable is centered at t = 11.5. The parameters β_1, β_2, and β_3 are the average intercepts for the cows on barley, mixed and lupins diets respectively at t=11.5 weeks. Further, the parameters β_4, β_5 and β_6 represent the average time effects (averaged over cows) for cows on barley, mixed and lupins diets respectively. Finally the parameters β_7, β_8, and β_9 represent the average time squared effects for cows on barley, mixed and lupins diets respectively. An initial analysis showed that there were hardly any quadratic (fixed and random) time effects. Consequently in (5.6) and (5.7), we include random effects for only the intercepts and slopes for the linear time effect. We then seek to compare the following three models:

$$\begin{aligned} M_1: &\quad \beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5,\ \beta_6,\ \beta_7,\ \beta_8,\ \beta_9 \\ M_2: &\quad \{\beta_1 > \beta_2 > \beta_3\}, \{\beta_4 > \beta_5 > \beta_6\},\ \beta_7,\ \beta_8,\ \beta_9 \\ M_3: &\quad \{\beta_1 = \beta_2 = \beta_3\}, \{\beta_4 = \beta_5 = \beta_6\}, \{\beta_7 = \beta_8 = \beta_9\} \end{aligned}$$

where M_1 is the model in which no constraints are imposed on the model parameters. Model M_2 investigates the following theory: at 11.5 weeks, the average protein content in the milk is highest, intermediate and lowest for the cows that received barley, mixed and lupins diets respectively. Further the average increase in percentage protein content per week is highest, intermediate and lowest for the cows that received barley, mixed and lupins diets respectively. This theory is translated into a statistical model by imposing order constraints on the intercept and slope parameters respectively. Finally M_3 investigates the theory that there is no difference between the percentage protein content in the milk obtained from cows on each of the three diets. In other words, M_3 states that the type of diet does not affect the protein content of milk. In this case the average, linear time and quadratic time parameters for the barley, mixed and lupins diets respectively are constrained to be equal.

For each model specified above, the Gibbs sampling algorithm was used to obtain samples from the posterior distribution of the model parameters. Subsequently the simulation based model fitting procedure presented in Section 2 was used to estimate W_r^2. For each model the estimates of W_r^2 are presented in Table 5.1, along with the partitioning of W_r^2 into the goodness of fit measure (FIT) and Penalty (PEN) components. For each model, estimates of W_r and the corresponding standard deviations are also given in the table. As can be seen in the table, M_2 renders the lowest value

Table 5.1: Posterior predictive measures for the Milk data.

Model	W_r^2	FIT	PEN	W_r	Stdev(W_r)
M_1	225.04	103.53	121.51	15.00	0.68
M_2	223.50	103.14	120.36	14.95	0.68
M_3	240.89	113.92	126.97	15.52	0.73

of W_r^2. This finding supports the theory that the milk protein content is highest, intermediate and lowest for the cows that received barley, mixed and lupins diets, respectively. However, it should be noted that the difference between W_1 and W_2 is 0.05 which is not that big given that the response variable percentage protein content ranges from 2.45 to 4.44. Further, given that the standard deviation of W_2 (the smallest criterion value) is 0.68 it follows that the difference between W_1 and W_2 is 0.074 of a standard deviation. So indeed the difference can be considered small. This suggests that for these data it is hard to distinguish between the unconstrained model (M_1) and the constrained model (M_2). In a situation such as this where two models have similar fits, we choose the simpler one. For these data, we would choose model M_2. Estimates of the parameters for model M_2 and associated posterior standard deviations can be obtained using the algorithm in the Appendix. Since estimation is not the main goal

of the paper we will not pursue that any further.

5.3.2 Rat Data

The data to be used in this section have been introduced and analysed by Verbeke and Lesaffre (1999). The data consist of craniofacial growth measurements of 50 male Wistar rats. The rats were randomised to either a control group or one of two treatment groups where the treatment consisted of a low or high dose of the drug Decapeptyl. This drug is an inhibitor for testosterone production in rats. The primary aim of the experiment was to investigate the effect of the inhibition of the production of testosterone on the craniofacial growth in male Wistar rats. The responses of interest are distances (in pixels) between well-defined points on X-ray pictures of the skull of each rat, taken after the rat had been anaesthetized.

In the same spirit as Verbeke and Lesaffre (1999), in this paper we will consider one of the measurements which can be used to characterize the height of the skull. The treatment started at the age of 45 days, and measurements taken every 10 days until the age of 110 days, with the first measurement taken at the age of 50 days. This would render 7 measurements for each rat. For each treatment group the individual profiles are shown in Figure 5.2. As can be seen in the figure not all rats have up to 7 measurements. This is because some rats did not survive the anaesthesia and therefore dropped out before the end of the study. In their analyses, Verbeke and Lesaffre (1999) use these data to investigate the effect of drop-out on the efficiency of longitudinal experiments. For the analyses in this paper, we will use the available data.

To analyse these data the following model proposed by Verbeke and Molenberghs (2000, Chapter 3) will be used:

$$y_{jk} = (\beta_1 + u_{1j}) + (\beta_2 C_j + \beta_3 L_j + \beta_4 H_j + u_{2j}) t_{jk} + \varepsilon_{jk}, \tag{5.8}$$

$$\mathbf{u}_j = (u_{1j}, u_{2j})^T \sim N(\mathbf{0}, \boldsymbol{V}), \quad \varepsilon_{jk} \sim N(0, \sigma^2),$$

in which y_{jk} denotes the $k = 1, \ldots, K_j$-th measurement for the $j = 1, \ldots, 50$-th rat and $t_{jk} = \ln[1 + (Age_{jk} - 45)/10)]$. In (5.8), C_j, L_j, and H_j are indicator variables defined to be 1 if the subject belongs to the control, low-dose group or the high-dose group respectively, and 0 otherwise. The transformation of the original time (age in days) implies that $t = 0$ corresponds to the start of the treatment. Note that the randomization in combination with this transformation of the original time scale allows one to assume that the subject-specific intercepts β_{1j} $(= \beta_1 + u_{1j})$ do not depend on treatment. Consequently the parameter β_1 represents the average response at the start of treatment. The parameters β_2, β_3 and β_4 represent the average slopes for the control, low dose and high dose groups respectively. In this

paper, interest lies in comparing several models relating to the treatment effect. Consequently these models will differ with respect to the constraints on the average slopes parameters since these directly measure the effect of treatment on the craniofacial growth. For these data, 5 models will be

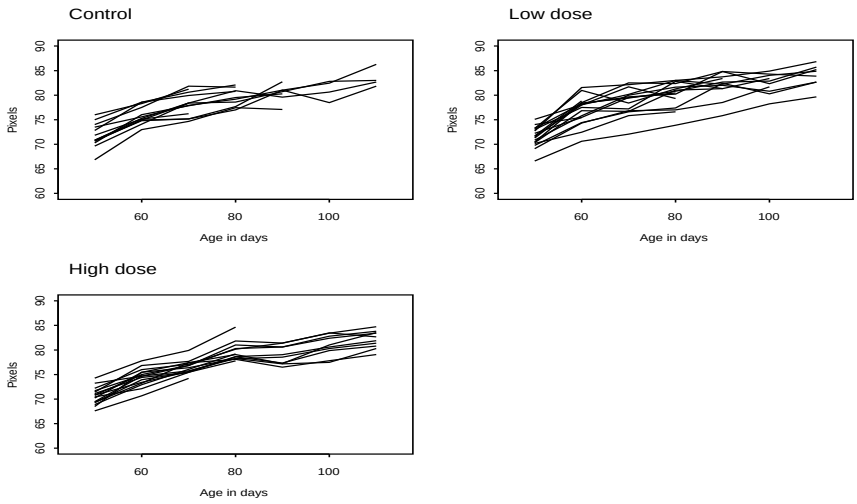

Figure 5.2: Profiles for each of the treatment groups

compared. These are

$$\begin{aligned}
M_1: &\quad \beta_1,\ \beta_2,\ \beta_3,\ \beta_4 \\
M_2: &\quad \beta_1,\ \beta_2 > \beta_3 > \beta_4 \\
M_3: &\quad \beta_1,\ \beta_2 > \{\beta_3, \beta_4\} \\
M_4: &\quad \beta_1,\ \beta_2 < \beta_3 < \beta_4 \\
M_5: &\quad \beta_1,\ \beta_2 = \beta_3 = \beta_4
\end{aligned}$$

Model 1 is the unconstrained model. Model 2 reflects the theory that the higher the dose, the lower the growth rate. In model 3, it is assumed that both treatments have the same effect over time and that each treatment is better than no treatment, hence the indicated constraint. Model 4 is just the reverse of model 2. Lastly model 5 investigates the theory that there is no treatment effect. In each of the above models, it is assumed that the random effects structure consists of both random intercepts and random slopes.

Estimates of W_r^2 and corresponding values for the goodness of fit measure (FIT) and penalty (PEN) terms for each of the models are reported in Table 5.2. The estimates of W_r and the corresponding standard deviations are also presented in the table. Model 1 emerges with the smallest W_r^2, hence we select it as the best model for the rat data. This result suggests that

Table 5.2: Posterior predictive measures for the Rat data.

Model	W_r^2	FIT	PEN	W_r	Stdev(W_r)
M_1	2690.99	1198.11	1492.88	51.87	4.74
M_2	2706.08	1220.50	1485.58	52.02	4.75
M_3	2719.77	1224.74	1495.03	52.15	4.79
M_4	2948.16	1328.27	1619.89	54.30	5.10
M_5	2852.40	1292.31	1560.09	53.41	4.93

there is some treatment effect but not strong enough to support the theories that inhibiting testosterone production in rats slows down their cranofacial growth (models 2 and 3). However, the difference between W_1 and W_2 is 0.15 which is quite small given that the response variable (distance in pixels) ranges from 66.6 to 86.3. Further, a standard deviation of 4.74 for W_1 (the smallest criterion value) implies that the difference between W_1 and W_2 is 0.032 of a standard deviation. So the difference is rather small. This suggests that for these data it is hard to distinguish between the unconstrained model (M_1) and the constrained model (M_2).

5.4 A Simulation Study

To investigate the performance of the predictive model selection criterion presented in Section 5.2, a simulation study was conducted and results are presented in this section. We shall use the rat data introduced in Section 5.3.2 as a reference point for the design in our simulations. The expected a posteriori (EAP) estimates and associated posterior standard deviations (PSD) for the parameters in the unconstrained model M_1 (see Section 5.3.2) are presented in Table 5.3, where $\mathbf{V}_{ab}$ is the element in the a-th row and b-th column of the covariance matrix $\mathbf{V}$. Later in this section these estimates

Table 5.3: EAP estimates for parameters in M_1.

Parameter	Estimate	PSD
β_1	68.610	0.313
β_2	7.334	0.300
β_3	7.505	0.240
β_4	6.888	0.245
$\mathbf{V}_{11}$	4.183	1.278
$\mathbf{V}_{12}$	-0.131	0.427
$\mathbf{V}_{22}$	0.213	0.180
σ^2	1.46	0.152

will be used as a point of reference for the populations to be used in the

simulation study.

Suppose that we have a designed experiment in which subjects are randomised to either a control group denoted by C or one of three treatment groups denoted by TG1, TG2 and TG3, respectively. Further, suppose there are 72 subjects (with 18 subjects in each group) and for each subject, measurements are taken every 10 days until 110 days. Note that this design has 4 groups (one group more than the number of groups in the empirical example of Section 5.3.2). The reason for this is to make the simulation design more elaborate than that of the rat data set. Like for the rat data set, we shall assume that treatment started at day 45. Let $y_j(t)$ denote the response variable taken for subject $j = 1, \ldots, 72$ at Age $= 50, \ldots, 110$. We then assume that $y_j(t)$ can be modelled as:

$$y_j(t) = \begin{cases} (\beta_1 + u_{1j}) + (\beta_2 + u_{2j})t + \varepsilon_j(t) & \text{if C} \\ (\beta_1 + u_{1j}) + (\beta_3 + u_{2j})t + \varepsilon_j(t) & \text{if TG1} \\ (\beta_1 + u_{1j}) + (\beta_4 + u_{2j})t + \varepsilon_j(t) & \text{if TG2} \\ (\beta_1 + u_{1j}) + (\beta_5 + u_{2j})t + \varepsilon_j(t) & \text{if TG3} \end{cases} \tag{5.9}$$

where $t = \ln\{1 + (Age - 45)/10\}$. The model assumes that on a logarithmic scale, the response variable evolves linearly over time. The parameter β_1 is the average intercept and is the same for each group while β_2, β_3, β_4 and β_5 are the average slopes for the subjects in the C, TG1, TG2 and TG3, respectively. Further every subject j is allowed to have its own intercept and slope, through the inclusion of random intercepts u_{1j} and random slopes u_{2j} which are assumed to jointly follow a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$. Lastly, all error components ε_j are assumed to be independent of each other as well as of the random effects u_{1j} and u_{2j}, and normally distributed with mean 0 and variance σ^2. Note that model (5.9) is a special case of (5.1).

The following set of models will be considered:

$$\begin{aligned} M_1 &: \quad \beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5 \\ M_2 &: \quad \beta_1,\ \beta_2 > \beta_3 > \beta_4 > \beta_5 \\ M_3 &: \quad \beta_1,\ \beta_2 > \{\beta_3, \beta_4, \beta_5\} \\ M_4 &: \quad \beta_1,\ \beta_2 < \beta_3 < \beta_4 < \beta_5 \\ M_5 &: \quad \beta_1,\ \beta_2 = \beta_3 = \beta_4 = \beta_5. \end{aligned}$$

The underlying model from which data was simulated was taken to be (5.9); 500 samples were generated from populations which conformed to each of the models $M_1, \ldots, M_5$, respectively. As can be seen in Table 5.3, the posterior standard deviations of the fixed effects parameters range from 0.24 to 0.31. Therefore, differences of at least 2 units in the average slopes would be relevant. Accordingly, for each of the models the following values were chosen for the parameter vector $\boldsymbol{\beta} = \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$:

M_1: (68, 5, 3, 8, 7)
M_2: (68, 10, 7, 4, 1)
M_3: (68, 8, 4, 7, 2)
M_4: (68, 1, 4, 7, 10)
M_5: (68, 7, 7, 7, 7)
In all cases the random effects $\mathbf{u}_j = (u_{1j}, u_{2j})$ were generated from a bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}=\begin{pmatrix} 4.0 & -0.15 \\ -0.15 & 0.50 \end{pmatrix}$. The error components were generated from a normal distribution with mean 0 and variance 1.5.

For each of the 500 data sets generated from each model $M_r\{r = 1, \ldots, 5\}$, the posterior predictive model selection procedure outlined in Section 5.2 was used to select the best of the 5 competing models. Table 5.4 shows the

Table 5.4: Frequencies of being selected as best of the 5 models given the true model.

	No of times model is selected				
True Model	M_1	M_2	M_3	M_4	M_5
M_1	500	0	0	0	0
M_2	110	220	170	0	0
M_3	180	0	320	0	0
M_4	210	0	0	290	0
M_5	110	40	70	110	170

frequencies of being selected as best of the 5 models based on 500 simulated data sets.

From the table, we see that whenever the unrestricted model M_1 is the true model, the procedure correctly identifies it, that is the procedure distinguishes M_1 from all the other models. Further, model M_2 can be distinguished from models M_4 and M_5, respectively and model M_3 can be distinguished from models M_2, M_4 and M_5, respectively. Lastly, model M_4 can be distinguished from models M_2, M_3 and M_5, respectively. Note that for about 34% of the data sets in which model M_2 is the true model, M_3 is selected as the best model. Nevertheless, this seems plausible since M_2 is closely related to M_3.

Further, the results suggest that it is not always possible to distinguish the constrained models M_2, M_3 and M_4, respectively from the unconstrained model M_1. For each of the three constrained models, a considerable amount of data sets are selected as coming from the unconstrained model. The same conclusion holds for the 'no treatment' effect model M_5 and the models $M_1, \ldots, M_4$ in which it is assumed there is a treatment effect. As can be seen in the table, when M_5 is the true model, out of 500 data sets simulated from the model, the procedure selects M_5 only 170 times.

Given the results from the simulation study, a re-evaluation of the results in Section 5.3 is in order. In Sections 5.3.1 and 5.3.2, models M_2 and M_1 were chosen as the best models, respectively. It could be that in reality these are the models which conform to the data best or it may well be that one of the other models was the best but the model selection procedure was not able to detect them.

5.5 Discussion

In this paper, we have used an approach to model choice which is based on the posterior predictive distribution of the replicated data. The approach is based on choosing a model that minimizes the mean squared distance between the observed data vector and a future observation vector.

The approach has a number of advantages. It avoids the problems associated with significance testing where one needs to derive (asymptotic) distributions for the likelihood ratio test under the null hypothesis. The method allows comparison of non nested models and there is no need to count the number of parameters in a model. Further, if a non informative prior distribution is used in the analysis, then there is no need to worry about the sensitivity of the results to the prior distribution. Lastly, in contrast to other model selection criteria, the uncertainty in the criterion values can be quantified via the computation of variances.

However, based on the results of the simulation study in Section 5.4, it turns out that the procedure is far from perfect. The simulation results show that given some empirical data it is not always possible to make a clear distinction as to which model (constrained or unconstrained) conforms to the data better.

More research is needed to investigate the performance of posterior predictive model selection. More simulations with different designs need to be done. Possible issues worth considering in the designs are varying the number of subjects and the number of measurements per subject.

The model selection criteria applied in the paper uses replicates which are obtained conditional on samples from the posterior distribution. For future investigation, a possible alternative is to compute the replicates conditional on the expected a posterior (EAP) estimates of the model parameters. It is worthwhile to note that a similar procedure can be used to obtain plug-in p values (Kato and Hoijtink, 2004, Section 3).

Furthermore, in the empirical examples that we analysed and the simulation study that was conducted all the models were assumed to have the same random effects structure. Future research would also consider models with different random effects structure e.g., random intercept models versus random coefficient models.

Appendix: Sampling from the Posterior distribution using the Gibbs Sampler

Starting with initial values of the model parameters, each iteration of the Gibbs sampler consists of four steps:

1. Sample $\mathbf{u}_j$ for $j = 1 \ldots, J$ from $N_2(\mathbf{\Phi}_j, \mathbf{\Sigma}_j)$ where

$$\mathbf{\Phi}_j = \left[\frac{\sum_{k=1}^{K_j} \mathbf{z}_{jk} \mathbf{z}_{jk}^T}{\sigma^2} + \mathbf{V}^{-1} \right]^{-1}$$

and

$$\mathbf{\Sigma}_j = \frac{\mathbf{\Phi}_j}{\sigma^2} \sum_{k=1}^{K_j} \mathbf{z}_{jk}^T (y_{jk} - \mathbf{x}_{jk}\boldsymbol{\beta}).$$

2. Sample σ^2 from a scaled inverse chi square distribution with degrees of freedom $\upsilon_o + \sum_{j=1}^{J} K_j$ and scale

$$\upsilon_o \sigma_o^2 + \sum_{j=1}^{J} \sum_{k=1}^{K_j} (y_{jk} - \mathbf{x}_{jk}\boldsymbol{\beta} - \mathbf{z}_{jk}\mathbf{u}_j)^2.$$

3. Sample $\mathbf{V}$ from an Inverse Wishart distribution with degrees of freedom $\lambda + J$ and scale matrix

$$\sum_{j=1}^{J} \mathbf{u}_j^T \mathbf{u}_j + \mathbf{S}.$$

4. Sample β_p from a normal distribution with mean

$$\frac{\frac{\mu_p}{d_p^2} + \frac{\sum_{j=1}^{J} \sum_{k=1}^{k_j} \left[y_{jk} - \sum_{\substack{i=1 \\ i \neq p}}^{P} \beta_i x_{ijk} - \sum_{q=1}^{Q} u_{qj} z_{qjk} \right] x_{pjk}}{\sigma^2}}{\frac{1}{d_p^2} + \frac{\sum_{j=1}^{J} \sum_{k=1}^{K_j} x_{pjk}^2}{\sigma^2}}$$

and variance

$$\frac{1}{\frac{1}{d_p^2} + \frac{\sum_{j=1}^{J} \sum_{k=1}^{K_j} x_{pjk}^2}{\sigma^2}}.$$

Note that for all the constrained parameter models, each $\beta_p \epsilon \boldsymbol{\beta}$ is sampled from a truncated normal distribution with some lower bound b and upper bound c. The value b is the largest $\beta_{p'}$ ($p \neq p'$) that must be smaller than β_p, and c is the smallest $\beta_{p'}$ ($p \neq p'$) that must be greater than β_p. To expound on this: suppose model M is specified using: $\beta_1 > \beta_2 > \beta_3 > \beta_4$. Then at each iteration of the Gibbs sampler, the sampled βs should fulfill the constraint imposed by the model. In this case, the parameters β_1, β_2, β_3 and β_4 would be sampled from normal distributions with lower bounds β_2, β_3, β_4 and $-\infty$ respectively and upper bounds ∞, β_1, β_2 and β_3 respectively. In general, for the constrained parameter models, sampling a β_p proceeds via sampling a deviate i from a uniform distribution on the interval (0,1) and then taking $\beta_p = \Phi^{-1}_{\beta_p}[\Phi_{\beta_p}(b) + i(\Phi_{\beta_p}(c) - \Phi_{\beta_p}(b))]$, where $\Phi_{\beta_p}(.)$ denotes the standard normal cumulative probability function of β_p evaluated at the argument (see Gelfand et al. (1992)).

References

Agresti, A. and Coull, B. A. (1996). Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics*, **52**, 1103–1111

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders, 4th edition*. Washington, D.C.

Amir N, Stafford J, Freshman M.S., Foa E.B. (1998). Relationship between trauma narratives and trauma pathology. *Journal of Traumatic Stress*, **11**, 385–392.

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika*, **86**, 141–152.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). Statistical Inference under Order Restrictions, John Wiley and Sons, London.

Bayarri, M. J. and Berger, J. O. (2000a), P Values for Composite Null Models. *Journal of the American Statistical Association*, **95**, 1127–1142.

Bayarri, M. J. and J.O. Berger (2000b), Rejoinder. *Journal of the American Statistical Association*, **95**, 1168–1170.

Bayarri, M. J. and J. O. Berger (2004), The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, **19**, 58-80.

Berkhof, J., H. Hoijtink and I. Van Mechelen (2000), Posterior predictive checks: Principles and discussion. *Computational Statistics*, **6**, 15, 337–354.

Berger, J.O., and Delempady, M. (1987). Testing precise hypotheses, *Statistical Science*, **2**, 317–335.

Boersma P, Weenink D. (1996) *Praat: A system for doing phonetics by computer, version 3.4, Report 132*. Institute of Phonetic Sciences, University of Amsterdam, The Netherlands.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Brewin C.R. (2001) A cognitive neuroscience account of posttraumatic stress disorder and its treatment. *Behaviour Research and Therapy*, **39**, 373–393.

Brewin CR, Dalgleish T, Joseph S. (1996) A dual representation theory of posttraumatic stress disorder. *Psychological Review*; **103**, 670–686.

Browne, W.J. (1998). *Applying MCMC Methods to Multi-level Models*. PhD dissertation, Department of Mathematical Sciences, University of Bath, United Kingdom.

Browne, W. and D. Draper (2000),Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, **15**, 391–420.

Bryk A.S. and Raudenbush S.W. (1992) *Hierarchical Linear Models: Application and Data Analysis Methods*. London: SAGE Publications.

Carlin, B.P., and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, B*, **57**, 473–484.

Carlin, B.P., and Louis, T.A. (2000). *Bayes and Empirical Bayes methods for data analysis*. Newyork: Chapman and Hall/CRC.

Catrien C.J.H.B., Van Der Kamp L.J.T., Mooijart A, Van Der Kloot W.A, Van Der Linden R, and Van Der Burg E. (1998) *Longitudinal Data Analysis: Designs, Models and Methods*. SAGE Publications: London.

Chen, M.H., and Shao, Q.M. (1998). Monte Carlo methods for Bayesian analysis of constrained parameter problems. *Biometrika*, **85**, 73–87.

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Cox, D.R. and D. V. Hinkley (1990), *Theoretical Statistics*. Chapman and Hall, London.

Dempster, A.P., N. M. Laird and D. B. Rubin (1977), Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Oxford: Oxford University Press.

Diggle, P.J. and Kenward, M.G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Applied Statistics*, **43**, 49–93

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Dobler, C.P. (2002). The one-way layout with ordered parameters: a survey of advances since 1988. *Journal of Statistical Planning and Inference*, **107**, 75–78.

Foa E.B., Molnar C, Cashman L. (1995). Change in rape narratives during exposure therapy for post-traumatic stress disorder. *Journal of Traumatic Stress*, **8**: 675–690.

Futschik, A. and Pflug, G.C. (1998). The likelihood ratio test for simple tree order: A useful asymptotic expansion. *Journal of Statistical Planning and Inference*. **70**, 57–68.

Gelfand, A.E and Ghosh, S.K. (1998). Model Choice: a minimum posterior predictive loss approach, *Biometrika*, **85**, 1–11.

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., Smith, A.F.M., and Lee, T.M. (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, **87**, 523–532.

Gelman, A., Carlin, J.B., Stern, H.S and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman and Hall, Newyork.

Gelman, A., Carlin, J.B., Stern, H.S and Rubin, D.B. (2000). *Bayesian Data Analysis*. Chapman and Hall, Newyork.

Geman, D., and Geman, S. (1984). Stochastic Relaxation, Gibbs distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Goldstein, H. (1995) *Multilevel Statistical Models*, 2nd edition. Edward Arnold: London.

Hoijtink, H. (2000). Posterior inference in the random intercept model based on samples obtained with Markov chain Monte Carlo methods, *Computational Statistics*, textbf15, 315–336.

Hoogduin K, Verdellen C, Cath D. (1997) Exposure and Response Prevention in the Treatment of Gilles de la Tourette's syndrome: Four Case Studies. *Clinical Psychology and Psychotherapy*; **4**, (2): 125-137.

Hox J. (2002) *Multilevel Analysis: Techniques and Applications*. Lawrence Erlbaum Associates, Inc.: New Jersey.

Jennrich, R.I., and Schluchter, M.D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices, *Biometrics*, **42**, 805–820.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kato, B.S., and Hoijtink, H. (2004). Testing Homogeneity in a random intercept model using asymptotic, posterior predictive and plug-in p-values. *Statistica Neerlandica*, **58**, 179–196.

Laird, N.M., and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lang P.J., Bradley M.M., Cuthbert B.N. (1998) Emotion, motivation and anxiety: Brain mechanisms and psychophysiology. *Society of Biological Psychiatry*; **44**, 1248–1263.

Lang P.J., Davis M. and hman A. (2000) Fear and anxiety: Animal models and human cognitive psychophysiology, *Journal of Affective Disorders*; **61**, 137–159.

Laud, P. and Ibrahim, J. (1995). Predictive model selection, *Journal of the Royal Statistical Society, Series B*, **57**, 247–262.

Lee, P.M. (1997), *Bayesian Statistics: An Introduction*. Arnold, London

LeDoux J. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster: New york

Longford N.T. (1993) *Random Coefficient Models*. Clarendon Press: Oxford.

Leckman J.F., Walker D.E. and Cohen D.J. (1993) Premonitory urges in Tourette's syndrome. *American Journal of Psychiatry*; **150**, 1, 98–102.

Marks I.M. (1987) *Fears, phobias and rituals. Panic, anxiety and their disorders*. Oxford University Press: Newyork.

Meng, X. L. (1994), Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142–1160.

Mukerjee, H. and Tu, R. (1995). Order-Restricted Inferences in Linear Regression, *Journal of the American Statistical Association*, **90**, 717–728.

Newton, M.A., and Raftery, A.E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap, *Journal of the Royal Statistical Society, B*, **56**, 3–48.

Pennebaker J.W. (1993) Putting stress into words: health, linguistic, and therapeutic implications. *Behavior Research and Therapy*; **31**, 539–548.

Pennebaker J.W. and Francis M.E. (1996) Cognitive, emotional and language processes in disclosure. *Cognition and Emotion*; **10**, 601–626.

Pennebaker J.W., Kiecolt-Glaser J.K. and Glaser R. (1998) Disclosure of traumas and immune function: Health implications for psychotherapy. *Journal of Consulting and Clinical Psychology*; **56**, 239–245.

Pennebaker J.W., Mayne T.J. and Francis, M.E. (1997) Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, **12**, 4, 863–871.

Potthoff, R.F., and Roy, S.N. (1964). A generalised multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, **51**, 313–326.

Rice, J.A. (1995). Mathematical Statistics and Data Analysis, 2nd ed., Carlifornia, Wadsworth Publishing Company.

Robertson, T., Wright F.T., and Dykstra, R.L. (1998) *Order Restricted Statistical Inference*. Chichester: John Wiley & Sons.

Robins, J. M., A. V. D. Van Der Vaart and V. Ventura (2000), Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association*, **95**, 1143–1156.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Sen, P.K. and Silvapulle, M.J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. *Journal of Statistical Inference and Planning*. **107**, 3–43.

Spiegelhalter, D.J., Best, N., and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Biostatistics, University of Minnesota.

Stram, D.O. and J. W. Lee (1994), Variance Components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.

Stram, D. O. and J. W. Lee (1995), Correction to: Variance Components testing in the longitudinal mixed effects model. *Biometrics*, **51**, 1196.

Snijders T.A.B. (1996) Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*; **30**, 405–426.

Snijders T.A.B. and Bosker R. (1999) *Multilevel Analysis*. Sage Publications: London.

Tanner, M. A. and W. H. Wong (1987), The calculation of posterior distributions by data augumentation. *Journal of the American Statistical Association*, **82**, 528–540.

Van Minnen A., Wessel I., Dijkstra T. and Roelofs K. (2002) Changes in narratives during prolonged exposure therapy for PTSD-patients: a replication and extension. *Journal of Traumatic Stress*; **15**, 255–258.

Verbeke, G. and Lesaffre, E. (1999), The effect of drop-out on the efficiency of longitudinal experiments, *Applied Statistics* **48**, 363-375.

Verbeke G. and Molenberghs G. (1997) *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Springer-Verlag: Newyork.

Verbeke G and Molenberghs G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer-Verlag: Newyork.

Verbeke G., Spiessens B. and Lesaffre E. (2001) *Conditional Linear Mixed Models*; *American Statistician*, **55**, 25–34

Verblya, A.P and Culis, B.R (1990). Modelling in Repeated Measures Experiments, *Applied Statistics*, **39**, 341–356

Verdellen C.W.J., Hoogduin C.A.L. and Plandsoen N. (1994) Exposure en responsepreventie bij de behandeling van het syndroom van Gilles de la Tourette: een gevalsbeschrijving. *Directieve Therapie*; **14**, 114–124.

Wooldridge, J. M. (2000), Introductory Economics: A Modern Approach. South-Western College Publishing, Thomas Learning.

Wright, F.T. and Tran, T. (1985). Approximating the level probabilities in order restricted inference: The simple tree ordering. *Biometrika*, **72**, 429–439.

Zeger, S.L., Karim, M.R., 1991. Generalised linear models with random effect; a Gibbs sampling. *Journal of the American Statistical Association*, **86**, 79–86.

Samenvatting

In een multilevel onderzoek is de gegevensstructuur in de populatie hirarchisch. De steekproef is een gelaagde steekproef uit de hirarchische populatie. Een voorbeeld uit onderzoek in het onderwijs: een populatie bestaat uit scholen en leerlingen binnen deze scholen. Andere voorbeelden van hirarchische gegevensstructuren vindt men in gezinsonderzoek, waarbij gezinsleden deel zijn van een gezin; bij longitudinaal onderzoek waarbij er per persoon verschillende waarnemingen zijn; bij bio-medisch onderzoek met nageslacht van dierenparen. De gegevens die dergelijk onderzoek opleveren staan bekend onder de overkoepelende naam "multilevel"-gegevens.

Elk statistiek onderzoek berust op kansmodellen, zelfs wanneer deze modellen niet gexpliciteerd zijn. Modellen berusten op parameters en statistische analyse houdt zich met name bezig met het 'kiezen' van waarden voor de parameters zodat het model op de gegevens past. Dit proefschrift behandelt enkele onderwerpen uit de analyse van multilevel-gegevens aan de hand van hirarchische lineaire modellen. De aandacht spitst zich toe op hirarchische lineaire modellen met restricties ten aanzien van de modelparameters.

In hoofdstuk 2 komen methoden aan de orde die kunnen worden gebruikt om homogeniteit te toesten in een "random intercept model". Het hoofdstuk begint met een overzicht van de definities van asymptotische, 'posterior predictive' en 'plug-in' p-waarden en het toont hoe deze p-waarden kunnen worden berekend. Eerst worden twee toetsingsgrootheden en twee discrepantie-maten gepresenteerd om de homogeniteit van residuele variantie te toetsen. Daarna worden een toetsingsgrootheid en een discrepantie-maat gepresenteerd om de homogeniteit van regressielijnen te toetsen. Vervolgens wordt elke toets en maat beoordeeld met behulp van asymptotische, 'posterior predictive' en' 'plug-in' p-waarden. Een simulatiestudie wordt gebruikt om de frequentie-eigenschappen van deze p-waarden te onderzoeken. Uit de resultaten van deze studie blijkt, dat, wanneer discrepantie-maten worden onderzocht met, hetzij posterior predictive, hetzij plug-in p-waarden, beide p-waarden goede 'klassieke' en Bayesiaanse frequentie-eigenschappen hebben. Dit wijst erop, dat wanneer de homogen-

iteit van residuele variantie en de homogeniteit van regressiehellingen in het random intercept model moeten worden beoordeeld, het aan te raden is discrepantie-maten te gebruiken met dan wel 'posterior predictive', dan wel 'plug in' p-waarden. Dit leidt tot de overtuiging dat homogeniteit in complexere hirarchische lineaire modellen kan worden onderzocht door uitbreidingen van de gepresenteerde methode en dat er niet noodzakelijkerwijs een heel nieuwe benadering nodig is.

Doorgaans hebben onderzoekers in de meeste onderzoeksvelden verschillende theorien of verwachtingen inzake de uitkomsten van hun onderzoek. Met andere woorden, voor elk willekeurig voor te leggen onderzoek kunnen verscheidene elkaar beconcurrerende theorien zich aandienen. Deze theorien kunnen afkomstig zijn van deskundigen op het betreffende gebied of voortkomen uit eerdere studies. Hoofdstuk 3 gebruikt klassieke statistiek om de houdbaarheid van de door deskundigen aangedragen theorien te onderzoeken. Hiertoe worden gegevens van twee projecten uit de klinische psychologie besproken en geanalyseerd; het eerste project is een onderzoek naar de invloed van angst op paralingustische aspecten van taal en het tweede project is een onderzoek naar de behandeling van het Gilles de la Tourette syndroom. De analyse van deze gegevens laat zien hoe het "linear mixed model" kan worden toegepast en gebruikt om herhaalde metingen te analyseren.

In de hoofdstukken 4 en 5 gaan over statistische modellering van theorien die een onderzoeker heeft met betrekking tot de uitkomst van het onderzoek. Hoofdstuk 4 behandelt verschillende onderwerpen. Allereerst wordt aangetoond hoe met restricties ten aanzien van de parameters van het model concurrerende theorien vertaald kunnen worden naar concurrerende modellen. Een volledig Bayesiaanse benadering wordt gebruikt om het 'beste' model te kiezen uit de serie concurrerende modellen. Voor elk model wordt de 'posterior' kans op het model berekend. Dit wordt bereikt door het gebruik van een nieuwe methode: 'method of encompassing priors'. Er wordt aangetoond dat voor ongelijkheidsgerestricteerde hirarchische lineaire modellen, posterior kansen berekend met deze methode vrijwel onafhankelijk zijn van de specificatie van de prior verdeling. Het model met de grootste posterior kans wordt gekozen als het model dat het best overeenstemt met de gegevens. In aanvulling op de selectie van een model, toont hoofdstuk 4 hoe de parameters van ongelijkheidsgerestricteerde hirarchische lineaire modellen kunnen worden geschat. Een voordeel van de Bayesiaanse benadering is dat schattingen van de parameters van het model makkelijk verkregen kunnen worden door gebruik te maken van de posterior verdeling. Bovendien voorziet deze benadering in de vergelijking van verschillende modellen of hypothesen: het aantal mogelijke modellen of hypothesen is onbeperkt en de modellen hoeven niet "genest" te zijn. In tegenstelling hiermee kan de klassieke wijze van het toetsen van hypothesen slechts twee hypothesen (nul en een alternatief) in

ogenschouw nemen, waarbij de nul-hypothese een restrictie is ten aanzien van de alternatieve hypothese.

Hoofdstuk 5 behandelt de "posterior predictive" modelselectie in de context van ongelijkheidsgerestricteerde hirarchische lineaire modellen. De selectie van een model wordt bereikt door de keus van een model dat de gemiddelde gekwadrateerde afstand tussen de waargenomen gegevens en de voorspelde waargenomen gegevens. Eigenlijk kiest deze benadering van modelselectie een model dat optimaal de waargenomen gegevens voorspelt. De methode wordt ingeleid en besproken aan de hand van twee voorbeelden. Het hoofdstuk besluit met een simulatiestudie om de prestaties van de methode te onderzoeken. Hieruit volgt dat posterior predictive modelselectie meerdere voordelen heeft boven de klassieke hypothese-toetsing.