UiO : Department of Mathematics
University of Oslo

# Focussed Model Selection for Longitudinal Data

**Tristan Curteis**

Master's Thesis, Spring 2018

This master's thesis is submitted under the master's programme *Modelling and Data Analysis*, with programme option *Statistics and Data Analysis*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group $E_8$, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

# Focussed Model Selection for Longitudinal Data

Tristan Curteis

Master's Thesis, Spring 2018

## Abstract

Longitudinal data arise when repeated measurements are taken on individuals over time. Commonly used models for such data are multivariate linear models, linear mixed effect models and generalised linear mixed models. This thesis begins by providing a detailed overview of these classes of models within the context of longitudinal data. Attention is then turned to model selection for such data. When selecting between models, one typically aims to come as close as one may to the underlying truth, without regard to the particular questions of interest. In contrast, the focussed information criterion (FIC) (Claeskens & Hjort 2003) approaches model selection with the goal of answering specific questions as accurately as possible. In this thesis, a multivariate slightly misspecified framework is put forward, within which the FIC is applicable as a covariate selector for multivariate linear models, linear mixed effect models, and generalised linear mixed models. Alternative approaches to focussed model selection for multivariate linear models and a selection of quantities of interest are also formulated.

## Acknowledgements

A list of acknowledgements can never make amends for the time and effort given, nor reach all of those one is indebted to. Nevertheless, I am grateful to friends Surya Tahora, and Dr. Snehal Shah for initial orientation and encouragement, even if the thesis took on a direction of its own. Thank you Matthew and Simon for your friendship, and for the opportunity to discuss statistical concepts. Thank you Kylie and Lewys for your time and patience. Lastly, and most importantly, I am grateful to my supervisor, Ørnulf Borgan, for granting me the right balance between guidance and independence. It has been fun working with you.

# Contents

# Chapter 1

# Introduction

## 1.1 Longitudinal and clustered data

Longitudinal data arise when measurements are taken repeatedly on the same subject or individual throughout time (Fitzmaurice et al. 2004*a*). For example, individuals randomised to different treatment groups may be observed on several occasions. Such a setting results in the measurements of any individual having temporal correlation, which must be accounted for in a statistical analysis (Laird & Ware 1982).

Similarly, clustered data are a more general form of longitudinal data. A commonly used example of non-longitudinal clustered data is exam results of students from different classes. The results of students within the same class are more likely, in comparison with students of different classes, to be similar. Thus, in this example, the class forms a cluster and the students form the units within each cluster. For longitudinal data, the repeated measurements of an individual form the units, and the individuals the clusters. In the school example, the within-cluster correlations are not temporal, but are still very much present.

In the regression setting, conceptually, independence means that, once having accounted for a principle set of variables, the connection or similarity between outcomes is considered weak or distant enough for any similarity between them to be attributable to chance. For longitudinal data, the measurements of different individuals, may (in general) be considered independent. For clustered data, the different clusters (i.e. the classes in the school example) may be considered independent. The within-cluster units, or within-individual measurements may not.

## 1.2 Model selection

For clustered or longitudinal data, multivariate linear regression, linear mixed effect models, generalised linear mixed models, and marginal models are commonly used classes of models. However, even within a given class, there is not just one model from which to draw conclusions. Generally, there is a list of competing models from which to choose, though this can often be shortened to some extent by considering what assumptions may be appropriate at the outset. The design of a study will also inform (or even determine) the form of a mean structure prior to model selection (Stroup 2012). Nevertheless, the statistician is regularly placed in a position of choice: is this model any better than the next? Combining the results from all models is also possible (Claeskens et al. 2008, Ch. 7), but how much should each model be relied upon?

A standard approach to model selection is to find the model which comes as close to the true data generating mechanism as possible. This is directly targeted by the Akaike information criterion (AIC), whose theoretical basis is in minimising the statistical distance from the true model (Claeskens et al. 2008, p.30). Selection of covariates via hypothesis tests (e.g. Wald or likelihood ratio tests) is also a commonly used means of model selection. The conditional AIC (cAIC) is an alternative information criterion for selection between mixed effect models (i.e. linear mixed effect models or generalised linear mixed models), whose goal is the same as that of the AIC but suitable when the focus of inference is at the level of a cluster, rather than the level of the population (Vaida & Blanchard 2005).

The focussed information criterion (FIC) (Claeskens & Hjort 2003, Claeskens et al. 2008) approaches model selection from a focussed point of view. Research in any field attempts to address specific questions which, in the sphere of parametric models, can be formulated as mathematical expressions in terms of parameters. The goal of the FIC is to estimate the parameters of interest as precisely as possible, thereby answering research questions as accurately as one may.

In longitudinal studies, there are specific questions to be addressed. Typically, variance-covariance or scale parameters are treated as nuisance parameters, and hypotheses addressing these questions can be formulated in terms of the regression coefficients (Fitzmaurice et al. 2004*a*). Questions such as, ‘Is there a positive trend with time?’, and ‘Are the mean slopes of these two treatment groups parallel?’ are frequently of interest (Fitzmaurice et al. 2004*a*). Thus, a focussed approach to model selection for longitudinal data is conceptually suitable, and is the main topic of this thesis.

The FIC, as formulated in Claeskens & Hjort (2003), is designed for independent data, though it is noted in Claeskens et al. (2008, p.259) that the extension to multivariate models is straightforward (and hence also for models of longitudinal data). FIC formula are also given in Cunen et al. (2017, 2018) for linear mixed effect (LME) models with an application to whale ecology. A quasi-FIC (QFIC) and associated model averaging schemes have also been introduced for selection of covariates in marginal models (which involve generalised estimating equations) for clustered data (Yang et al. 2017).

## 1.3 Structure of thesis

This thesis is structured as follows. In Chapter 2, a comprehensive overview of linear models for longitudinal data with and without random effects is given. That is, linear models (LMs) without random effects and linear mixed effect (LME) models are discussed in the context of longitudinal data, with a psychological clinical trial dataset used as an illustration. Chapter 3 gives an overview of generalised linear mixed models (GLMMs), with an application to binary longitudinal data. Chapter 4 begins with introducing the FIC for independent data as in Claeskens & Hjort (2003). Then, how the FIC is applicable to multivariate models of clustered (and in particular longitudinal) data within a slightly misspecified framework is made explicit. Examples are then given for multivariate linear regression and for a logistic GLMM. Lastly, in Chapter 5, alternative methods to arrive at focussed model selection formulae for linear models and a subset of quantities of interest are derived.

The focus of application in this thesis is on longitudinal data. All of the theory, however, is applicable to clustered data. Therefore in application, (since it is generally people that are studied over time) the term ‘individual’ will be used instead of the term ‘cluster’. With the exception of Chapter 2, which also lays out particular features of longitudinal data, the term ‘clusters’ will be used for the theoretical parts that are relevant to both longitudinal and clustered data.

# Chapter 2

# Linear Regression for Longitudinal Data

This chapter discusses linear regression for longitudinal data. The first part deals with linear regression without random effects and is based upon Chapters 2-7 of Fitzmaurice et al. (2004*a*). The second part discusses linear mixed effect (LME) models and is sourced from Bryk & Raudenbush (1992) and Galecki & Burzykowski (2013). Estimation and model diagnostics are then discussed for both classes of models simultaneously. Finally, a real data set is used to illustrate the LME model in action.

## 2.1 Linear regression for longitudinal data

Linear regression is one of the most commonly used statistical models. The principal assumption of ordinary linear regression is that the errors are independent of each other. This is reasonable in many situations. However, when the data exhibits a clustered structure, as is the case for longitudinal data where repeated measurements are clustered within individuals, this assumption is unacceptable. This is due to there being dependency between measurements: given an individual's measurement at one point in time, we have information on the same individual's measurement at another point in time.

Assuming independence between individuals (clusters) $i = 1, ..., N$, but not between the $n_i$ measurements (units) for a given individual (cluster), the linear model (LM) for longitudinal (or clustered) data is

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim N_{n_i}(\mathbf{0}, \Sigma_i(\boldsymbol{\theta})), \tag{2.1}$$

where $\boldsymbol{y}_i$ is the $n_i \times 1$ vector of continuous responses for the $i$th individual; $\boldsymbol{X}_i$ is an $n_i \times p$ design matrix for the $i$th individual; $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients; and $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of random errors for the $i$th individual, which is assumed to be normally distributed, mean zero, with a positive semi-definite, symmetric variance-covariance matrix $\Sigma_i(\boldsymbol{\theta})$, itself dependent upon parameters $\boldsymbol{\theta}$.

Typically, the first column of the design matrix will consist of a column of ones (the intercept term), and each remaining column of the design matrix will correspond to a particular covariate or interaction between covariates. Generally, it is the regression coefficients that are of primary interest, as many scientific hypotheses can be formulated in terms of the regression coefficients.

Note that linear regression for uncorrelated data and homogeneous variance (when $\Sigma_i = \sigma^2 \boldsymbol{I}$) is a special case of the LM (2.1). In fact, the variance-covariance matrix $\Sigma_i(\boldsymbol{\theta})$ specifies the correlation of within-individual measurements, an important feature of longitudinal modelling. This variance-covariance matrix can be modelled in different ways. The inclusion of random effects in the mean structure implicitly induces a structure on the covariance. Alternatively, certain structures can be explicitly imposed upon $\Sigma_i$, which usually exploit some pattern in the repeated measurements.

The variance-covariance matrix $\Sigma_i(\boldsymbol{\theta})$ need not vary between individuals when the study design is balanced and there is complete data. That is, the subscript $i$ may be dropped and it may be assumed that all individuals share the same variance-covariance. Furthermore, notice that the between-individual measurements are assumed to be independent. This is usually a sound assumption to make as the measurements of different individuals within a study do not usually influence one another. There are exceptions to this, however. For example, if two participants are living in the same household.

## 2.2 Modelling the mean

The mean trend of a response over time can be modelled in one of two different ways: treating time as discrete, or as continuous. With time continuous, the mean structure is a parametric or semi-parametric function of time (e.g. linear, or piece-wise linear), and relatively few parameters are required regardless of the number of measurement occasions.

### 2.2.1 Parametric trends

Consider the situation where we have two groups (e.g. girls and boys) measured for the same outcome on multiple occasions. If the change in mean response over time seems to be constant for both groups, though at possibly different rates, the following model for the mean response could be insightful:

$$\mathbb{E}[y_{ij}] = \beta_0 + \beta_1 \text{group}_i + \beta_2 t_{ij} + \beta_3 t_{ij} \text{group}_i,$$

for $i = 1, ..., N$, $j = 1, ..., n_i$; where $\text{group}_i$ is an indicator variable (taking values 0 or 1) for the group of individual $i$; $t_{ij}$ is the time at occasion $j$ (the subscript $i$ allows individuals to have different sets of times i.e. an unbalanced design); and the third term is an interaction between time and group. So, a $\beta_3$ significantly different from zero would indicate a significant difference in rate of linear change in response over time between the two groups.

The same model given in terms of the matrix formulation of (2.1) is

$$\mathbb{E}\left[\begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}\right] = \begin{pmatrix} 1 & \text{group}_i & t_{i1} & t_{i1}\text{group}_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \text{group}_i & t_{in_i} & t_{in_i}\text{group}_i \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Note that, above, the two groups are modelled to have different mean outcomes at baseline. That is, when $t = 0$, the reference group has mean baseline score $\beta_0$ and the non-reference group a mean score of $\beta_0 + \beta_1$. Such an assumption is reasonable at the outset in an observational study, where individuals are grouped according to naturally existing characteristics. But in studies where individuals are randomised to different groups after baseline measurement e.g. treatment and placebo, there is no reason to assume different baseline scores; the means can be assumed to coincide.

Another important assumption of the model is that there is linear change in the response variable over time for both groups. Such an assumption may not be appropriate if the mean profiles do not change at a constant rate. In such cases, a model with quadratic time may be appropriate. For example,

$$\mathbb{E}[y_{ij}] = \beta_0 + \beta_1\text{group}_i + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 t_{ij}\text{group}_i + \beta_5 t_{ij}^2\text{group}_i,$$

for $i = 1, ..., N$, $j = 1, ..., n_i$; where $\beta_3$ is now the mean change in response for the reference group for every unit change in time squared; and $\beta_5$ an interaction term between group and time squared. Such a model allows a convex or concave change in the mean response over time at different rates for both groups.

It should be noted that introducing a quadratic term (or higher order term) for time results in colinearity between predictors. Time and time squared will be almost perfectly correlated which can lead to computational problems. It is wise to centre the variable time to avoid such an issue. Choice of centering is not usually a problem in balanced designs: the mean time will suffice. But in unbalanced designs, the mean time may not have a clear interpretation for all individuals (as not all individuals may have been participating in the study at the mean time), and so a meaningful value for all individuals may be chosen.

One can easily imagine how these models generalise to multiple categorical or continuous covariates. Higher order polynomials in time and randomly-varying (with time) covariates are also possible, but the interpretation of regression coefficients becomes more challenging.

### 2.2.2 Semi-parametric trend

An alternative to assuming a smooth change in mean response over the whole study period is to assume piece-wise smoothness. That is, to break the study period up into sections, and to assume a parametric trend (usually defined by a polynomial) over each individual section.

For example, suppose there is a sharp change in behaviour in the mean response at time $u$. Then the model for the mean response could be

$$\mathbb{E}[y_{ij}] = \beta_0 + \beta_1 \text{group}_i + \beta_2 t_{ij} + \beta_3 (t_{ij} - u)_+ + \beta_4 t_{ij} \text{group}_i + \beta_5 (t_{ij} - u)_+ \text{group}_i,$$

for $i = 1, ..., N$, and $j = 1, ..., n_i$, where $(t_{ij} - u)_+$ equals $(t_{ij} - u)$ if $u \geq t_{ij}$ and zero otherwise. The first term captures the mean baseline score of the reference group; $\beta_0 + \beta_1$ the mean baseline score of the non-reference group; $\beta_2$ the change in response of the reference group induced by a unit increase in time prior to time $u$; $\beta_2 + \beta_3$ is the effect of a unit increase in time on the response of the reference group after time $u$; $\beta_4$ is the additional change in response for the non-reference group before time $u$ (in comparison with the reference group); and $\beta_4 + \beta_5$ is the additional change in response of the non-reference group after time $u$.

The above model assumes linear change in response before and after time $u$ for both groups, though the slopes of any group need not be the same before and after time $u$. The time $u$ in this model, where the joining of two differentiable curves meet, is known as a *knot*. There are methods to decide on the best locations for the knots of any model, but this will not be discussed in this thesis.

## 2.3 Modelling the covariance

In longitudinal studies, there are typically three sources of variability. The first is between-subject variability, which is simply that there will be a spread in response tendencies among participants. Some individuals will tend to have an above average response, some below average, and others somewhere in the middle. The second source is within-subject variability. This source accounts for the fact that the underlying process being measured for any individual (be it biological, psychological etc.) is constantly undergoing change. Because of this, there will be fluctuations in response over time for any given individual. The third source of variability is measurement error. Conceptually, this source of error is that even when two measurements are taken as close together in time as possible, such that one would expect identical results to be produced, the results are still not totally consistent due to the measurement instrument being used. When the instrument of measurement is a psychometric test, for example, this can be a substantial source of error.

These factors account in different ways for the general characteristics of correlation between repeated measures in longitudinal studies. The correlations between repeated measures for a given individual arising in such studies:

- are positive,
- generally decrease as time between measurements increases,
- rarely approach zero, no matter how much time has passed between measurements,
- rarely approach one, no matter how little time between measurements.

Why is it important to take into account correlation between measurements? If positive correlations between measurements are not taken into account, estimates of the variance (or variances if allowing for heterogeneity) will be inflated. These incorrectly estimated variances influence the standard errors of the regression coefficients and thereby affect statistical inference. So, failure to account for correlation between measurements results in faulty inference.

But the correlation between measurements is not a nuisance; it is the strength of a longitudinal study. Being able to account for the positive correlation results in more efficient estimates of the mean response at each occasion. In effect, the models taking into account correlation borrow information from all occasions to obtain more precise estimates of the mean response. This results in smaller standard errors for the regression parameters and thereby greater power to detect the effect of covariates on changes in the response over time. In this way, longitudinal studies capitalise on correlated measurements, a feature not possible in comparison of two separate cohorts in a cross-sectional study for example.

There are two options for modelling the covariance: to leave the matrix unstructured (but necessarily still symmetric and semi-positive definite) or to apply a structure. The application of a structure can either be done directly via variance-covariance pattern modelling, or indirectly through the introduction of random effects. First, variance-covariance pattern models will be discussed. Linear models with random effects will be discussed subsequently.

### 2.3.1 Examples of variance-covariance pattern models

Leaving the covariance matrix unstructured (aside from the symmetry and semi-positive definite requirements) may be suitable when the design is balanced and there are few measurement occasions. However, since $n$ variance parameters and $\frac{n(n-1)}{2}$ covariance parameters are required (where $n$ is the number of measurement occasions), various patterns are generally applied (especially when $n$ is large) to reduce the parameter burden. In the following, $\rho$ is assumed to be a parameter in the interval [0,1).

The *compound symmetric* structure is conceptually suited to studies where the ordering of within-cluster measurements does not matter (not the case for longitudinal studies) as the correlation between any pair of measurements is the same. The compound symmetric structure is

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & \cdots & \rho & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}, \tag{2.2}$$

where $\sigma^2$ is the variance which is assumed homogeneous across measurements. We will see this structure again later in this chapter, as it arises naturally when individual-specific random intercepts are introduced.

The *auto-regressive* structure assumes a Markov-type dependency on the errors: that the errors at one occasion depend upon the errors at the previous occasion, and thus correlations decay with time. The system of equations (e.g. Cressie & Wikle 2011, p.87) relating each error to the previous error is

$$\begin{aligned} \epsilon_{ij} &= \rho\epsilon_{ij-1} + \sigma w_{ij}, \qquad j = 2, ..., n_i \\ \epsilon_{i1} &\sim \; N(0, \sigma^2), \end{aligned}$$

where $w_{ij}$ are $N(0, (1-\rho^2))$ random variables. When this structure is imposed on the errors, their variance-covariance matrix becomes

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-1} & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-2} & \rho^{n-1} & \cdots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho & 1 \end{pmatrix},$$

which only requires two parameters, and where $n$ is the number of measurements.

The *Toeplitz* covariance pattern

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-2} & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-1} & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-2} & \rho_{n-1} & \cdots & 1 & \rho_1 \\ \rho_{n-1} & \rho_{n-2} & \cdots & \rho_1 & 1 \end{pmatrix}$$

assumes that pairwise measurements equally distant in time share the same pairwise correlations ($0 \leq \rho_1, ..., \rho_{n-1} < 1$).

The *exponential* covariance structure generalises the auto-regressive structure to settings where the time between measurement occasions are not necessarily evenly spaced. The covariance between the $j$th and $k$th response for subject $i$ is of the form

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma^2 \rho^{|t_{ij} - t_{ik}|},$$

which implies exponential decay in correlation between measurements with the progression of time.

In spatial statistics, a *nugget effect* is often included to account for discontinuities in the correlation function at the origin (Cressie & Wikle 2011, p.123). That is, whenever one moves the smallest of distances from one location to a new location, there will a drop in correlation between measurements at both locations. In this way, measurements at two different locations are modelled as never perfectly correlated. Although I have not seen this in longitudinal literature yet, the same principle can be applied to longitudinal data where the only 'spatial' dimension is time. When the difference in time between two measurements is zero (i.e. it is the same measurement), there is perfect correlation between two measurements. But when there is even the smallest of time lags between measurements, the correlation can no longer be perfect. In this manner, the nugget effect can account for the fact that no measurement instrument is totally reliable and/or that there is within-individual variability in the response.

For example, the nugget effect, $\kappa$, can be added to the exponential correlation pattern as

$$\text{Corr}(\epsilon_{ij}, \epsilon_{ik}) = \begin{cases} (1-\kappa)\rho^{|t_{ij}-t_{ik}|}, & \text{if} \quad |t_{ij}-t_{ik}| \geq 0, \\ 1, \quad \text{if} \quad |t_{ij}-t_{ik}| = 0, \end{cases}$$

with $0 < \kappa < 1$ small.

Note that a nugget effect is already implicitly implied for non-continuous correlation functions, such as the compound-symmetric and auto-regressive structures. So, including a nugget effect only offers a potential improvement for continuous correlation functions.

The above covariance patterns all assume homogeneity of variance. That is, the variance is assumed to be constant across time. This can be unrealistic in longitudinal studies where there are typically differing degrees in spread of the responses from baseline to the end of the study. There are different ways to to accommodate heterogeneity into these covariance patterns. For example (Galecki & Burzykowski 2013), assume no structure and allow $n$ different variance parameters to be estimated, or define the variance in terms of a function that depends upon unknown parameters e.g. $\text{Var}(\epsilon_{ij}) = |t_{ij}|^{\delta}$, or $\text{Var}(\epsilon_{ij}) = e^{|t_{ij}|\delta}$, with $\delta$ an unknown parameter.

Choosing an appropriate model for the variances and covariances is a matter of balance. One neither wants models that are too simple and fail to catch the intricacies of the covariance patterns, nor models that are too complex with too many parameters to be estimated. This balance of getting things just right is the case of a bias versus variance trade-off, and will be discussed in detail later.

## 2.4 Linear mixed effect models

Introducing random effects into the LM (2.1) is an alternative means to account for correlation between repeated measurements. Doing so creates a new class of models: linear mixed effect (LME) models. Since LME models are used in many different fields, they go by many different names. LME models are also known as multilevel models, random-effects models, linear mixed models, and random coefficient models. In particular, they are also termed hierarchical linear models which stresses the hierarchy of the data structure.

Typically, longitudinal data have a two-level hierarchy: there is the level of an individual's measurements (level-1), and a between-individual level (level-2). There are variables which relate to between-individual differences, e.g. gender, that do not change with time, and there are variables at the within-individual level which do change with time, e.g. time itself. One can also imagine higher levels to the hierarchy: the individuals may be grouped within schools or hospitals which have their own associated variables. The inclusion of a random effect at the between-individual level posits a diversity in the response accountable for by between-individual differences that have not been explained by the covariates. As such, linear mixed effects models offer a natural way to account for heterogeneity at different levels of the data-hierarchy.

Furthermore, the inclusion of random-effects partitions the variability in the data. The variability due to within-individual fluctuations and between-individual diversity can be separated, a feature not possible with the variance-covariance pattern models seen in Section 2.3.1. This separation of the variance facilitates inference at both the level of the individual and at the level of the population. Which means that, along with the population mean trend, individual-specific trajectories can be charted over time.

### 2.4.1 The general linear mixed effect model

As originally put forward by Laird & Ware (1982), the general LME model for individuals (clusters) $i = 1, ..., N$ of $j = n_1, ..., n_N$ measurements (units) respectively is

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{2.3}$$

where $\boldsymbol{y}_i$ is the $n_i \times 1$ continuous outcome vector; the $\boldsymbol{X}_i$ is the $n_i \times p$ fixed effect design matrix; $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects coefficients; $\boldsymbol{Z}_i$ and $\boldsymbol{b}_i$ are the $n_i \times k$ matrix of random effect covariates and the $k \times 1$ vector of random effect coefficients respectively; and $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ vector of random errors, which explain variability in the response of individual (cluster) $i$ not accounted for by the fixed effect, $\boldsymbol{X}_i\boldsymbol{\beta}$, or random effect, $\boldsymbol{Z}_i\boldsymbol{b}_i$ components of the mean structure. This is in contrast to the errors of (2.1) which account for variability unexplained by the marginal mean structure, $\boldsymbol{X}_i\boldsymbol{\beta}$, the fixed effect part alone.

In the general LME model it is assumed that

$$\boldsymbol{\epsilon}_i \sim N_{n_i}(\mathbf{0}, \boldsymbol{R}_i), \quad \boldsymbol{b}_i \sim N_k(\mathbf{0}, \boldsymbol{D}), \tag{2.4}$$

where $\boldsymbol{\epsilon}_i$ are independent of $\boldsymbol{b}_i$, and the $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are both themselves assumed independent for all $i = 1, ..., N$.

The fixed effects coefficients are common to all individuals (note the absence of subscript $i$). They represent the influence that the fixed effects covariates have on the population mean response. The random effect coefficients are individual-specific and describe the expected difference between individual $i$'s outcome and the population mean response (McNeish et al. 2017). In fact, it is the inclusion of the random effects term that differentiates the general LME model from the LM (2.1). Alternatively, one can view the LM (2.1) as a special case of the general LME model with the random effect coefficients set equal to zero.

The general LME model (as presented above) is the most general form of a 2-level LME model. Extension to a greater number of levels is straightforward and is discussed in, for example, Galecki & Burzykowski (2013).

### 2.4.2 The conditional distribution defined in the general LME model

The general LME model specifies the unconditional distribution of the random effects, $\boldsymbol{b}_i$, and the conditional distribution of the response given the random effects, $\boldsymbol{y}_i|\boldsymbol{b}_i$. Both are assumed to be multivariate normal. The random effects need not be considered normal, but doing so is mathematically and computationally simpler.[1] Therefore, random effects will be considered normal throughout. The distribution of $\boldsymbol{b}_i$ is given in (2.4); the expectation and variance of the conditional distribution of the response given the random effects are

$$\mathbb{E}[\boldsymbol{y}_i|\boldsymbol{b}_i] = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,$$

and

$$\text{Cov}(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{R}_i,$$

respectively, where $\text{Cov}(\cdot)$ denotes the variance-covariance matrix.

In theory, the only constraints on $\boldsymbol{R}_i$ and $\boldsymbol{D}$ are that they are symmetric and semi-positive definite. However, since these matrices are determined by parameters to be estimated, certain structures are often imposed to reduce the number of parameters, especially when sample sizes are small. In fact, it is the modeller who selects the specification of these structures, which model the within-individual dependency, and how the random effects of one covariate covary with another respectively (McNeish

[1] It is suggested in Stroup (2012, p.11) that non-normal random effects will be commonly used in the future. Just as how linear and generalised linear models would have been too advanced 30 years ago, but are now ubiquitous in statistics.

et al. 2017). If $\boldsymbol{R}_i$ is not diagonal, $\boldsymbol{\epsilon}_i$ no longer accounts purely for within-individual variability, a feature which may be desirable to retain. In addition, estimation of two general structures can be difficult due to identifiability reasons (Fitzmaurice et al. 2004*a*, p.195). Therefore, unless otherwise stated, $\boldsymbol{D}$ is considered to be a general semi-positive definite symmetric matrix, and $\boldsymbol{R}_i$ will be regarded as a diagonal matrix with homogeneous variances. That is, $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}$.

### 2.4.3 The marginal model implied by the general LME model

From the general LME model (2.3), it can be seen that, marginally,

$$\begin{aligned}\text{Cov}(\boldsymbol{y}_i) =& \text{Cov}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i)\\ =& \text{Cov}(\boldsymbol{Z}_i\boldsymbol{b}_i) + \text{Cov}(\boldsymbol{\epsilon}_i)\\ =& \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\intercal + \sigma^2\boldsymbol{I}_{n_i}. \end{aligned} \tag{2.5}$$

Thus, marginally, correlations between repeated-measurements are accounted for. In addition, it is the particular form of the random effect design matrix that determines the type of association between repeated-measurements. This is in contrast to the LM of Section 2.1 where the correlations were modelled explicitly. Furthermore, since $\mathbb{E}[\boldsymbol{y}_i] = \boldsymbol{X}_i\boldsymbol{\beta}$, the general LME model implies the existence of a marginal model, namely

$$\boldsymbol{y}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\intercal + \sigma^2\boldsymbol{I}_{n_i}). \tag{2.6}$$

Three classes of sub-models of the general LME model will now be discussed.

### 2.4.4 The random effects model

The random effects model, sometimes called the unconstrained or null-model, is the least complicated sub-model of the general LME model and corresponds to a one-way analysis of variance (ANOVA) with random effects. No covariates are taken into account at any level and the model is formulated as

$$y_{ij} = \beta_0 + b_i + \epsilon_{ij}, \tag{2.7}$$

where

$$\epsilon_{ij} \sim N(0, \sigma^2) \quad \text{and} \quad b_i \sim N(0, d_{11}) \tag{2.8}$$

are independent for all individuals $i = 1, ..., N$, and for all within individual measurements $j = n_1, ..., n_N$. That is, the response for each individual is predicted by the overall mean $\beta_0$; $b_i$ are the individual-specific deviations away from the overall mean; and $\epsilon_{ij}$ are the within-individual errors. The parameters $d_{11}$ and $\sigma^2$ describe the between and within-individual variances respectively. In other words, they provide a measure of the response dispersion due to differences between individuals and due to within-individual differences respectively.

The random effects model is unrealistically simple for most applications. Nevertheless, it is useful as an initial step in a statistical analysis of two-level hierarchical data, as it allows the statistician to gauge to what extent each level accounts for the total variation. Specifically, from the random effects model, we are able to extract the *intraclass correlation coefficient* (ICC) defined as

$$\frac{d_{11}}{\sigma^2 + d_{11}}, \tag{2.9}$$

which gives the proportion of the total variability of the response that is accounted for by the between-individual variation.

The random-effects model can also be formulated using a system of equations that clearly illustrate the two-level nature of the data. The level-1, within-individual, model is

$$y_{ij} = \gamma_{0i} + \epsilon_{ij}. \tag{2.10}$$

The individual-specific intercepts $\gamma_{0i}$ (note the subscript $i$) are allowed to vary for each individual and become the response in the level-2, between-individual, model. This is

$$\gamma_{0i} = \beta_0 + b_i, \tag{2.11}$$

which states that the only between-individual differences are in terms of intercepts. To see the equivalence with Equation (2.7), substitute Equation (2.11) back into Equation (2.10).

### 2.4.5 Random intercept models

Introducing the level-1 covariate time $t_{ij}$ as an explanation for within-individual differences, the level-1 model becomes

$$y_{ij} = \gamma_{0i} + \gamma_{1i}t_{ij} + \epsilon_{ij},$$

and the between-individual model can be formulated as

$$\begin{aligned} \gamma_{0i} &= \beta_0 + b_{0i}, \\ \gamma_{1i} &= \beta_1. \end{aligned} \tag{2.12}$$

Combining the above system of equations gives the combined model

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \epsilon_{ij}. \tag{2.13}$$

The above is an example of a random intercept model. Introducing a (level-2) covariate, e.g. treatment group, into Equation (2.12) allows slopes to vary for different values of the covariate, producing a non-randomly varying slopes model.

### Induced compound symmetric structure

Consider now a more general random intercept model allowing for the possibility of several covariates. In particular, let $\boldsymbol{X}_i$ be the $i$th cluster's fixed effects design matrix, and $\boldsymbol{\beta}$ be the vector of fixed effect regression coefficients, as in the general LME model (2.3).

For the random intercept model with independent and identically distributed within-individual error terms, the marginal variance-covariance matrix of the response is given by

$$\begin{aligned}\text{Cov}(\boldsymbol{y}_i) =& \text{Cov}(\boldsymbol{X}_i\boldsymbol{\beta} + \mathbb{1}b_i + \boldsymbol{\epsilon}_i)\\ =& \text{Cov}(\mathbb{1}b_i + \boldsymbol{\epsilon}_i)\\ =& \mathbb{1}\text{Cov}(b_i)\mathbb{1}^\intercal + \text{Cov}(\boldsymbol{\epsilon}_i)\\ =& \mathbb{1}d_{11}\mathbb{1}^\intercal + \sigma^2\boldsymbol{I}_{n_i},\end{aligned}$$

where $\mathbb{1}$ is a column vector of ones; $\boldsymbol{X}_i$ is the design matrix of fixed effect covariates; $\sigma^2$ describes the within-individual variation; and $d_{11}$ the between-individual variation. This can be written as

$$\text{Cov}(\boldsymbol{y}_i) = \begin{pmatrix} \sigma^2 + d_{11} & d_{11} & \cdots & d_{11} & d_{11} \\ d_{11} & \sigma^2 + d_{11} & \cdots & d_{11} & d_{11} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{11} & d_{11} & \cdots & \sigma^2 + d_{11} & d_{11} \\ d_{11} & d_{11} & \cdots & d_{11} & \sigma^2 + d_{11} \end{pmatrix}.$$

That is, the random intercept model induces a compound symmetric structure [see (2.2)] on the variance-covariance of the response. Note, however, that this is only the case when the within-individual errors are assumed independent and identically distributed, and does not hold in general (Hedeker & Gibbons 2006).

The essential assumption of both the random effects model and random intercept models is that individuals are allowed to differ randomly only in terms of their intercepts. That is, with the exception of non-randomly varying slope models, the fitted slopes of different individuals are the same. Indeed, the random effect model involves no slope at all. Such an assumption may be acceptable for some applications. However, empirical evidence and theoretical reasoning may be to the contrary, and allowing the slopes to vary randomly for each individual is often a valuable modelling approach (Bryk & Raudenbush 1992). Furthermore, inducing a compound symmetric structure on the mean marginal (or population) trend is unsuitable for longitudinal data, where correlations generally decay with time (Fitzmaurice et al. 2004*a*).

### 2.4.6 A random intercept and slope model

Introducing a random effect in Equation (2.12) allows the slope to vary randomly for each individual and results in the following random intercept and slope model:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}. \tag{2.14}$$

This corresponds to including both a column of ones, and a column for time, $t_{ij}$, in both the fixed effect design matrix, $\boldsymbol{X}_i$, and the random effect design matrix, $\boldsymbol{Z}_i$, of the general LME model (2.3). The model assumes that there is a between-individual variability both in terms of intercepts, and in terms of slopes. The combination $\beta_1 + b_{1i}$ has the interpretation as the individual-specific effect of a unit increase in time on the expected response.

Marginally, we have that

$$\begin{aligned}\text{Var}(y_{ij}) &= \text{Var}(\beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}) \\ &= \text{Var}(b_{0i}) + \text{Var}(b_{1i} t_{ij}) + 2\text{Cov}(b_{0i}, b_{1i} t_{ij}) + \text{Var}(\epsilon_{ij}) \\ &= d_{11} + d_{22} t_{ij}^2 + 2 d_{12} t_{ij} + \sigma^2,\end{aligned}$$

which means that the variance is a quadratic function in time, and whose coefficients are determined by the elements of the variance-covariance matrix of the random effects. This is similar for the marginal response covariances (Fitzmaurice et al. 2004*a*, p.197) and is a particular case of Equation (2.5). This means that the model (2.14) accounts for a quadratic time dependence in the marginal response.

## 2.5 Estimation

Two types of estimation procedures will be discussed here: maximum likelihood (ML), and restricted maximum likelihood (REML). Both of these procedures require an expression for the joint density of the data (or transformed data in the case of REML). Since random effects are unobserved, it is reasonable to produce parameter estimates (of the fixed effect and variance-covariance components) for the general LME (2.3) based on the implied marginal model (2.6) (Galecki & Burzykowski 2013). This model, (2.6), is a special case of the LM (2.1), so estimation for both the LM and the LME model will be discussed simultaneously. To this end, consider

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i \boldsymbol{\beta}, \sigma^2 \boldsymbol{\Sigma}_i^*(\boldsymbol{\phi})),$$

a re-parameterisation of (2.1), with $\sigma^2$ factorised such that $\sigma^2 \boldsymbol{\Sigma}_i^*(\boldsymbol{\phi}) = \Sigma(\boldsymbol{\theta})$ for later convenience. The parameters $\boldsymbol{\phi}$ are defined such that $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi}^\intercal)^\intercal$. For the special case of the marginal mixed effects model (2.6), we have $\sigma^2 \boldsymbol{\Sigma}_i^*(\boldsymbol{\phi}) = \sigma^2(\boldsymbol{Z}_i \boldsymbol{D}^*(\boldsymbol{\phi}) \boldsymbol{Z}_i^\intercal + \boldsymbol{I}_{n_i})$, with $\sigma^2 \boldsymbol{D}^*(\boldsymbol{\phi}) = \boldsymbol{D}(\boldsymbol{\phi})$, and where $\boldsymbol{\phi}$ are the parameters of random effect covariance matrix, $\boldsymbol{D}$.

### 2.5.1 Maximum likelihood

Maximum likelihood (ML) estimation is one of the most common methods used to obtain parameter estimates. The idea is this: given the data, which parameter values maximise the likelihood that the data was generated by the model under consideration?

Under model (2.5), the joint density of the response for individual (cluster) $i$ is

$$f(\boldsymbol{y}_i) = (2\pi\sigma^2)^{-\frac{n_i}{2}} |\boldsymbol{\Sigma}_i^*(\boldsymbol{\phi})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\intercal \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right),$$

where $|\cdot|$ denotes the matrix-determinant. The contribution to the log-likelihood from one individual (cluster) is thus

$$\begin{aligned}\ell_{\text{ML},i}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}; \boldsymbol{y_i}) = -\frac{1}{2}\Big[&n_i \log 2\pi\sigma^2 + \log|\boldsymbol{\Sigma}_i^*(\boldsymbol{\phi})| \\ &+ \frac{1}{\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\intercal \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\Big].\end{aligned}$$

The log-likelihood, assuming independence between individuals (clusters), is

$$\begin{aligned}\ell_{\text{ML},N}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}; \boldsymbol{y}) = -\frac{1}{2}\sum_{i=1}^{N}\Big[&n_i \log 2\pi\sigma^2 + \log|\boldsymbol{\Sigma}_i^*(\boldsymbol{\phi})| \\ &+ \frac{1}{\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\intercal \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\Big], \quad (2.15)\end{aligned}$$

where $\boldsymbol{y}$ is a stack of all the $\boldsymbol{y}_i$ vectors. This marginal log-likelihood can be maximised via profiling out $\boldsymbol{\beta}$ and respectively $\sigma^2$ as illustrated in Galecki & Burzykowski (2013). Alternatively, one may differentiate (2.15) with respect to $\boldsymbol{\beta}$, $\sigma^2$ and $\boldsymbol{\phi}$ directly, then equate to zero and re-arrange to form an iterative system of equations (Gumedze & Dunne 2011). For differentiating (2.15) see Section A.1 of Appendix A, where expressions are given for the derivatives of the log-determinant of the variance-covariance matrix and its inverse with respect to its parameters. We will return to these derivatives again in Section 4.3.

The regression parameters and the factor $\sigma^2$ at the $m$th step of the iteration procedure are given by

$$\hat{\boldsymbol{\beta}}_{(m)} = \Big(\sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Sigma}_i^{*-1}(\hat{\boldsymbol{\phi}}_{(m-1)})\boldsymbol{X}_i\Big)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Sigma}_i^{*-1}(\hat{\boldsymbol{\phi}}_{(m-1)})\boldsymbol{y}_i, \quad (2.16)$$

$$\hat{\sigma}^2_{(m)} = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{(m)})^\intercal \boldsymbol{\Sigma}_i^{*-1}(\hat{\boldsymbol{\phi}}_{(m-1)})(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{(m)}), \quad (2.17)$$

respectively. The variance-covariance parameters $\boldsymbol{\phi}$ at the $m$th step solve the $r$-dimensional system of equations (indexed by $j = 1, ..., r$, where $r$ is the dimension of $\boldsymbol{\phi}$)

$$\sum_{i=1}^{N} \mathrm{Tr}\left\{ \boldsymbol{\Sigma}_i^{*-1}(\hat{\boldsymbol{\phi}}_{(m)}) \frac{\partial \hat{\boldsymbol{\Sigma}}_i^*}{\partial \phi_j} \right\} = \\ \frac{1}{\hat{\sigma}^2_{(m-1)}} \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{(m-1)})^\intercal \hat{\boldsymbol{\Sigma}}_i^{*-1} \frac{\partial \hat{\boldsymbol{\Sigma}}_i^*}{\partial \phi_j} \hat{\boldsymbol{\Sigma}}_i^{*-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{(m-1)}), \quad (2.18)$$

and where $\hat{\boldsymbol{\Sigma}}_i^* = \boldsymbol{\Sigma}_i^*(\hat{\boldsymbol{\phi}}_{(m)})$ is the estimated variance-covariance matrix at the $m$th step.

The first step in the iteration procedure is to assign a starting value to $\boldsymbol{\phi}$, which is then used to obtain estimates of $\boldsymbol{\beta}$ and $\sigma^2$ via (2.16) and (2.17). Using (2.18), the values of $\boldsymbol{\beta}$ and $\sigma^2$ are then used to update the variance components $\boldsymbol{\phi}$, which in turn serve to update the estimates of $\boldsymbol{\beta}$ and $\sigma^2$. This procedure of repeatedly alternating between estimating both $\boldsymbol{\beta}$ and $\sigma^2$ and then $\boldsymbol{\phi}$ is continued until convergence is reached.

### 2.5.2 Restricted maximum likelihood

It is well known that estimates of variance components via ML are downwardly biased. This is because the ML estimators neglect the fact that the regression parameters are also being estimated, which results in a reduction in degrees of freedom. Restricted maximum likelihood estimation (REML) on the other hand, allows for unbiased estimates of the variance components. This can be accomplished by using a projection matrix that removes the regression parameters prior to constructing the likelihood (shown in Gumedze & Dunne 2011, p.1926).

By profiling out the regression coefficients, the relationship between the log-likelihood, and restricted-log-likelihood is given in Galecki & Burzykowski (2013, p.197) as

$$\ell_{\mathrm{REML},N}(\sigma^2, \boldsymbol{\phi}) = \ell_{\mathrm{ML},N}(\hat{\boldsymbol{\beta}}(\boldsymbol{\phi}), \sigma^2, \boldsymbol{\phi}) + \frac{p}{2} \log \sigma^2 - \frac{1}{2} \log \left| \sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi}) \boldsymbol{X}_i \right|, \quad (2.19)$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{\phi})$ is the generalised least squares (GLS) estimator

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\phi}) = \Big( \sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi}) \boldsymbol{X}_i \Big)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^\intercal \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi}) \boldsymbol{y}_i, \quad (2.20)$$

and where $p$ is the number of regression parameters.

Finding the maximiser of (2.19), either via profiling on $\sigma^2$ or via an iterative scheme, results in finding unbiased estimates of the variance-covariance components.

In particular, the REML estimator of $\sigma^2$, expressed here as a function of $\boldsymbol{\phi}$ is (Galecki & Burzykowski 2013)

$$\hat{\sigma}^2_{\text{REML}} = \frac{1}{\left(\sum_{i=1}^{N} n_i\right) - p} \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}(\boldsymbol{\phi}))^{\intercal} \boldsymbol{\Sigma}_i^{*-1}(\boldsymbol{\phi})(\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}(\boldsymbol{\phi})),$$

which, in contrast to (2.17), accounts for the fact that $\boldsymbol{\beta}$ is also being estimated. The estimated variance-covariance components may then be substituted into the expression (2.20) to find REML estimates for the regression parameters.

### 2.5.3 A note on model selection

The models for the mean and the covariance are inter-dependent (e.g. Fitzmaurice et al. 2004*a*, p.163). This is because the variance-covariance matrix is defined for the errors which are the response minus the mean trend. When the mean is misspecified, for example when linear growth is used with non-linearly behaving data, additional variance crops up in the covariance matrix of the errors. To avoid modelling a covariance matrix that is not representative of the actual covariation, but is a consequence of a misspecified mean, covariance model selection (or random effect selection for the case of LME models) should first be performed on a maximal mean structure (or maximal fixed effect structure) which has minimal bias. Model selection for the mean trend may then be looked into once a variance-covariance model has been chosen.

For balanced longitudinal data, the maximal mean structure may be formed by including time as a categorical variable since this imposes no specific time trend on the data (e.g. Fitzmaurice et al. 2004*a*, p.173). At this stage, since REML produces unbiased estimates of the variance-covariance parameters, REML is the preferred estimation procedure, and particularly so for smaller samples.[2] Comparison of different variance-covariance structures may then be carried out via, for example, the Akaike information criterion (AIC) based on the maximised restricted log-likelihood, $\hat{\ell}_{\text{REML},N}$ (e.g. Claeskens et al. 2008, p.271)

$$\text{AIC}_{\text{REML}} = 2(\hat{\ell}_{\text{REML},N} - s),$$

where $s$ is the number of variance-covariance parameters.

Once the covariance model (or random effect structure) has been chosen and different mean structures are to be compared, the REML log-likelihood is of no use. This is because models with different mean structures require a different transformation of the data to construct the restricted log-likelihood.[3] Hence, it is not possible

[2] The discrepancy between ML and ML diminishes as $N$ grows relative to the number of regression parameters (Fitzmaurice et al. 2004*a*, p.101).

[3] This aspect of constructing the restricted log-likelihood has not been demonstrated here, see Gumedze & Dunne (2011, p.1926) for more details.

to use the restricted log-likelihood for comparison of different mean structures (e.g. Galecki & Burzykowski 2013, p.87). Rather, one might use the ML based AIC,

$$\text{AIC}_{\text{ML}} = 2(\hat{\ell}_{\text{ML},N} - s),$$

as a selection criterion, where here, in contrast to $\text{AIC}_{\text{REML}}$, $s$ denotes the total number of model parameters.

One can view the AIC as a trade-off between fit and complexity. The former being captured by the (potentially restricted) log-likelihood component, the latter by the penalty in terms of number of parameters. In addition, the aim of the AIC criterion is to minimise the distance between the candidate models and the true underlying data mechanism (Claeskens et al. 2008, p.30). As formulated above, models with larger AIC are preferred.

For LME models, an alternative to the AIC when focus is at the level of the individual is the conditional AIC (cAIC) introduced by Vaida & Blanchard (2005). The cAIC maximises the conditional (given the random effects) log-likelihood and uses twice the effective number of model parameters as a penalty term. A corrected version of the cAIC which accounts for the fact that the variance-covariance parameters have to be estimated is given in Greven & Kneib (2010).

### 2.5.4 Predicting random effects

For the LME model, it is common to use

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^\intercal \hat{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) \tag{2.21}$$

as predictors of individual-specific deviations, $\boldsymbol{b}_i$, from the fixed-effect parameters. These are found by plugging in relevant estimators into the conditional means

$$\mathbb{E}[\boldsymbol{b_i}|\boldsymbol{y}_i] = \boldsymbol{D}\boldsymbol{Z}_i^\intercal \Sigma_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

where the dependence of $\Sigma_i$ and $\boldsymbol{D}$ on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively is implicit.

The uncertainty in the predictor (2.21) is[4]

$$\text{Cov}(\hat{\boldsymbol{b}}_i) = \boldsymbol{D}\boldsymbol{Z}_i^\intercal \Sigma_i^{-1}\left(\Sigma_i - \boldsymbol{X}_i\left(\sum_{i=1}^{N}\boldsymbol{X}_i\Sigma_i^{-1}\boldsymbol{X}_i^\intercal\right)^{-1}\boldsymbol{X}_i^\intercal\right)\Sigma_i^{-1}\boldsymbol{Z}_i\boldsymbol{D}. \tag{2.22}$$

However, to generate prediction intervals, the quantity

$$\text{Cov}(\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i) = \boldsymbol{D} - \text{Cov}(\hat{\boldsymbol{b}}_i), \tag{2.23}$$

which accounts for the variability of the random variable and for uncertainty in prediction (though not for uncertainty in estimation of variance-covariance parameters), is used (e.g. Fitzmaurice et al. 2004*a*, p.208).

[4] See the relevant part of Formula (9.30) in McCulloch & Searle (2001, p.256.).

## 2.6 Model diagnostics for linear models

For the general LME model, residuals pertaining to either the marginal (2.6) or the conditional model (2.3) may be defined (Galecki & Burzykowski 2013, p.265). For the LM (2.1), only marginal residuals can be defined.

The *marginal residuals* obtained for individual $i$ are

$$\boldsymbol{r}_{m,i} = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}},$$

where $\boldsymbol{y}_i$ are observed, and $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ is the marginal mean trend. Since the data are clustered, the residuals obtained for any given individual will be correlated. To 'de-correlate' and simultaneously standardise the residuals, the Cholesky transform may be applied to give the transformed residuals

$$\boldsymbol{r}^*_{m,i} = \boldsymbol{L}_i^{-1}\boldsymbol{r}_{m,i},$$

where $\boldsymbol{L}_i$ is the lower triangular matrix satisfying the Cholesky-decomposition $\boldsymbol{L}_i\boldsymbol{L}_i^\intercal = \hat{\Sigma}_i$. So, the particular form of $\boldsymbol{L}_i$ depends on the particular form of variance-covariance pattern or random effect structure. For the LM (2.1), these transformed residuals can be inspected in the same way as how residuals arising from ordinary linear regression are inspected e.g. graphical checks for homoskedasticity.

For the LME model, the *conditional residuals* for individual $i$ may also be defined as

$$\boldsymbol{r}_{c,i} = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i.$$

However, these residuals are 'contaminated' in that they may be confounded with the random effects $\boldsymbol{b}_i$. Therefore, for LME models, Santos Nobre and da Motta Singer (as cited in Galecki & Burzykowski 2013, p.266) suggest using the marginal residuals to assess propriety of the marginal mean trend. The conditional residuals may be reserved for detecting outlying observations, and homoskedasticity.

In the general LME model, assumptions are also made on the random effects, $\boldsymbol{b}_i$. According to Verbeke and Molenberghs (as cited in Galecki & Burzykowski 2013, p.265), the distribution of $\hat{\boldsymbol{b}}_i$ does not necessarily represent that of $\boldsymbol{b}_i$. Therefore, checking normality of $\hat{\boldsymbol{b}}_i$ via e.g. Q-Q plots is of little use. However, such plots may be useful for detecting outliers. It is also worth mentioning that both normality of residuals (for both the LM and the LME model) (e.g. Fitzmaurice et al. 2004*a*, p.61), and normality of the random effects (particularly if inference concerns only the $\beta$s) are not too critical assumptions (Verbeke and Molenberghs as cited in Galecki & Burzykowski 2013, p.265).

## 2.7 Dataset illustration

A dataset from a psychological clinical trial will be used here to illustrate the LME model in action. The dataset is explored at length in Hedeker & Gibbons (2006) and was obtained from its accompanying website (Hedeker 2006). The trial follows 66 patients with depression for a period of 5 weeks. All of whom received the same treatment consisting of a daily dosage of antidepressant Imipramine. The continuous response to be analysed is the Hamilton Depression Rating Scale (HAMD) score. The scale measures 17 variables each on either a 3 or 5 item scale (summed to produce a continuous summary statistic) and is administered by, preferably, two independent interviewers (Hamilton 1960). The HAMD scores were recorded at the beginning and end of a first placebo week, and then at the end of each of the following four treatment weeks (Hedeker & Gibbons 2006). Patients were diagnosed with either endogenous depression or non-endogenous depression, where endogenous depression is depression due to internal or biological causes, and non-endogenous depression is caused by external factors such as social or familial reasons (Hedeker & Gibbons 2006).

As is often the case in psychological research (McNeish et al. 2017), not all individuals were observed at each time point, resulting in missing data. Even though neither the LM nor the LME model require complete data (Fitzmaurice et al. 2004*a*), a complete case analysis involving $N = 46$ patients is performed here. The R package **plyr** (Wickham 2011) was used to help select the subset of patients with no missing data.

For simplicity of illustration, the specific drug-levels in the blood, measured as stochastic, time-varying covariates will be ignored. Instead, covariates of interest will be $t$, time in weeks from beginning of treatment treated as continuous, and binary covariate $ed$, taking value 1 if a patient has endogenous depression, and 0 if a patient has non-endogenous depression. Attention will also be given to the interaction term between $t$ and $ed$. This being said, the purpose of this illustration is to show LME models in action, not to draw definitive conclusions about the trial itself.

### 2.7.1 Exploratory analysis

Figures 2.1 and 2.2 show a scatterplot and box plot of the observed HDRS scores for the two types of diagnosis for the depressed patients respectively. The figures were put together using the **lattice** (Sarkar 2008) package.

The negative trend in both plots suggest that, in general, both groups of patients' depression scores are reduced over time. It also appears that non-endogenous patients start with lower depression scores at baseline. In addition, notice that the negative trend begins even before the treatment starts. An explanation of this could be regression to the mean, as patients recruited are likely to have more extreme depression than average at baseline. The slight decline during the placebo week could also be a case of what is described by Rosenthal & Rosnow (1991) as reactive observation, where the mere fact that patients are enrolled in a study and observed by therapists could induce a reduction in depression. The box plot shows increasing variability in the response as time increases. This suggests different rates of improvement for different individuals, implying that a model involving random slopes could be useful. It could also mean that the assumption of constant variances may not be appropriate, or that the stochastic time-varying drug-levels in the blood (ignored in our analysis) are influential.

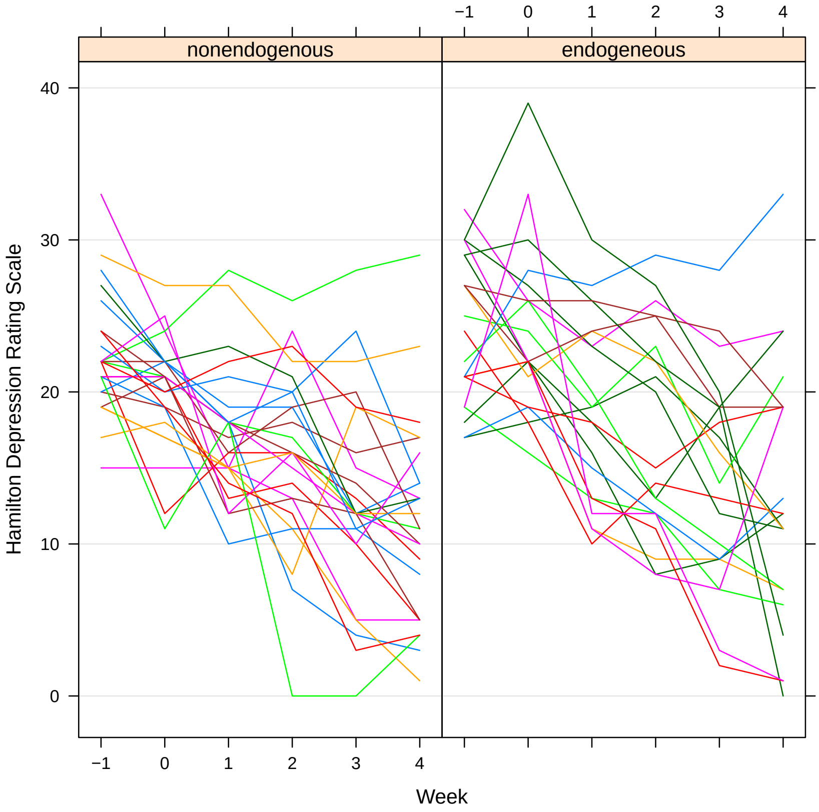


Figure 2.1: A spaghetti plot of observed scores against time for all patients with complete data. Patients are separated according to their diagnosis as endogenous (right) or not (left). Week -1 is the placebo week. Treatment begins from the beginning of week 0.

For the following, the placebo week is ignored and concentration is on fitting LME models during the period of treatment. With the treatment period defined as the beginning of week 0, baseline is considered to be the measurement at the end of week -1. All models are fit using the *lme* command from the **nlme** package (Pinheiro et al. 2016). The parameter estimates are presented in tables that were generated using the package **stargazer** (Hlavac 2015). Estimates relating to random effects have been added to the tables manually.

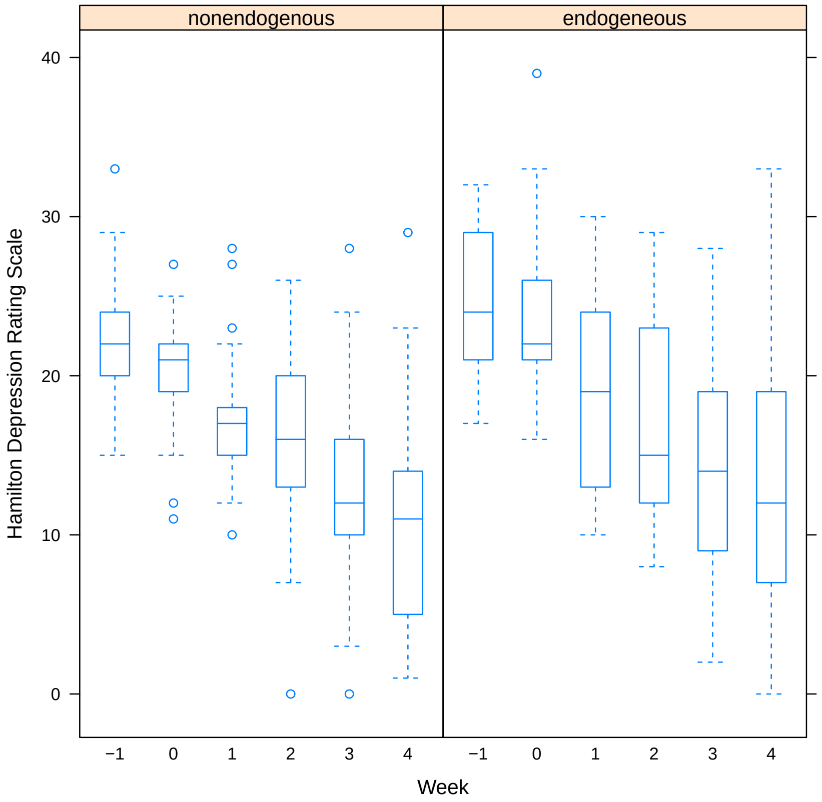


Figure 2.2: A box plot of scores against time for all patients with complete data, separated according to diagnosis. Week -1 is a placebo week. Treatment begins from the beginning of week 0.

### 2.7.2 Random effects model

As part of the data exploration, a random effects model was fitted. As explained in Section 2.4.4, the random effects model is used as a preliminary step in model fitting in order to extract the ICC, not as a realistic model of the data. The random effects model is

$$hd_{ij} = \beta_0 + b_i + \epsilon_{ij}, \tag{2.24}$$

for $i = 1, ..., 46$, $j = 0, ..., 4$, where $\beta_0$ is the overall mean HAMD score, and $b_i$ the individual deviations from the mean score.

Table 2.1 contains the REML parameter estimates obtained from this model. Extracting the between and within-individual variances enables calculation of the ICC. The ICC of the complete data set is 0.37, which suggests that the between-individual variation is accountable for over a third of the total variation. So, taking into account the effect of clustering is certainly worthwhile.

Table 2.1: Parameter estimates from the random effects model.

| | HDRS score |
|---|---|
| Fixed Effects: | |
| Intercept | $16.422^{***}$ |
| Random effects: | |
| $\sigma$ | 5.61 |
| $\sqrt{d_{11}}$ | 4.34 |

$^{*}$p < .05; $^{**}$p < .01; $^{***}$p < .001

### 2.7.3 Random intercept and slope model

Consider the following random intercept and slope model:

$$hd_{ij} = \beta_0 + \beta_1 ed_i + \beta_2 t_{ij} + \beta_3 t_{ij} ed_i + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij},$$

for all $i = 1, ..., 46$, $j = 0, ..., 4$.

This corresponds to having fixed effect and random effect design matrices

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & ed_i & t_{i0} & t_{i0}ed_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & ed_i & t_{i4} & t_{i4}ed_i \end{pmatrix} \quad \text{and} \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 & t_{i0} \\ \vdots & \vdots \\ 1 & t_{i4} \end{pmatrix}$$

in the general LME model respectively.

With this set up, the fixed effect parameters can be interpreted in the following way: $\beta_0$ represents the mean baseline HAMD score for the reference group, non-endogenous depressed patients; $\beta_1$ is the effect diagnosis as an endogenous depressed patient has on the mean baseline score; $\beta_2$ is the effect on the HAMD score of a non-endogenous patient due to a one week change in time; $\beta_3$ represents the additional change in score due to a one week increase in time for endogenous depressed patients. For patient $i$, $b_{i0}$ is the difference from the mean baseline score of the depression group to which patient $i$ belongs, and $b_{i1}$ is the difference in rate from the average rate of change of HAMD score of the depression group to which patient $i$ belongs.

The parameter estimates from the above model are presented in Table 2.2. Significance of covariates were tested using exact (under the model) t-tests. The effect of diagnosis as an endogenous depressed patient on the mean baseline score is a significant variable at the 0.05 level, suggesting differences between groups at baseline. The coefficient of time is negative and significant at the 0.05 level, suggesting a general improvement for the reference group over the course of the study. The coefficient of the interaction term was estimated to be -0.46, suggesting that the endogenous patients improve faster over time compared with non-endogenous patients, but this term was not statistically significant.

The variances of the random effect parameters can be interpreted as follows: provided the sample of patients are representative of their respective populations, $95\%$ of all non-endogenous patients have baseline HAMD scores ranging between $19.87 \pm Z_{0.975} \times \sqrt{11.32} = (12.45, 27.29)$; and over $90\%$ of all non-endogenous patients have negative slopes $(-2.18 + Z_{0.90} \times \sqrt{2.42} = -0.19 < 0)$.

This being said, the results should be taken lightly: the analysis was based only on patients with complete data, who could be a biased subset of the initial sample; the specific levels of antidepressant and its chemical transformation in the blood measured each week were ignored; and a linear time effect may be simplistic.

Table 2.2: Parameter estimates from the random intercept and slope model.

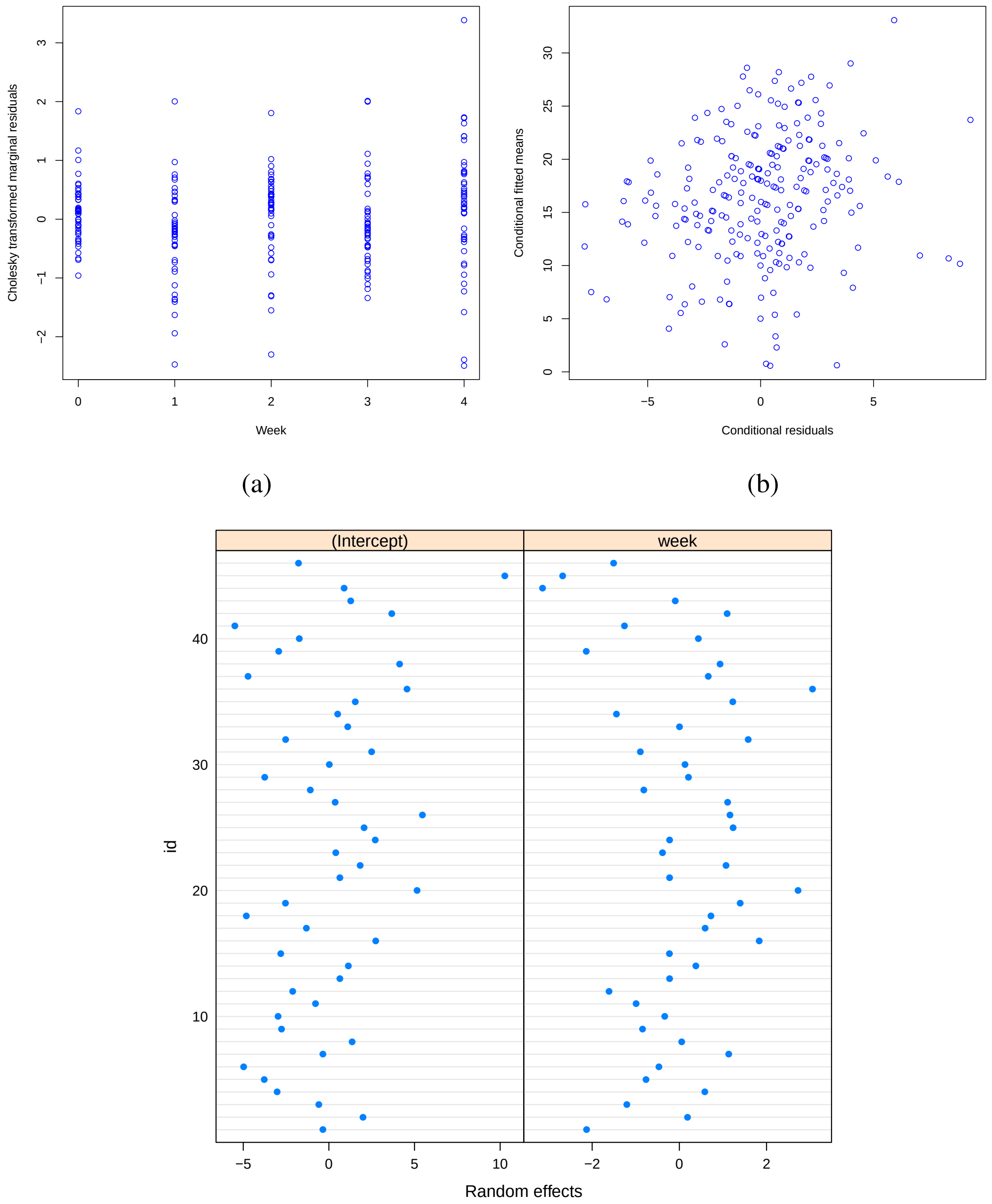
| | HDRS score |
|---|---|
| Fixed effects: | |
| Intercept | $19.87^{***}$ |
| Endo | $2.94^{*}$ |
| Time | $-2.18^{***}$ |
| Time:Endo | $-0.46$ |
| Random effects: | |
| $\sigma^2$ | 11.32 |
| $d_{11}$ | 14.33 |
| $d_{12}$ | -0.41 |
| $d_{22}$ | 2.42 |

$^{*}$p < .05; $^{**}$p < .01; $^{***}$p < .001

### 2.7.4 Model assumptions

It is difficult to detect any pattern in the scatterplot of marginal residuals versus the marginal mean. But, there appears to be slight variation between weeks (see Figure 2.3a) in whether the residuals are centered slightly above or below zero, perhaps suggesting linear time is too simplistic. With less variability at the more extreme values of the fitted conditional means, Figure 2.3b suggests that homoskedastic within-individual errors could be improved upon.

Figure 2.3c shows a scatterplot of the random effects for both intercept and time plotted against patient ID number. Individual 45 has an outlying HAMD score at baseline; individuals 20, 36, 44 and 45 have potentially outlying slopes.



(a)

(b)

(c)

Figure 2.3: Part (a) shows a scatterplot of the cholesky transformed marginal residuals versus time in weeks. Part (b) shows a scatterplot of the conditional residuals versus the fitted conditional means. Part (c) shows a scatterplot of the random effects for intercept (left) and time (right) against patient ID.

## 2.8 Concluding remarks

### 2.8.1 A short comparison of the general LME model and the LM

Modelling the variability via introducing random effects in the linear model accounts for the correlations between repeated measurements with relatively few parameters, regardless of the number of measurement occasions. This is in contrast to the variance-covariance pattern models of Section 2.3.1, whose simplicity (in terms of number of parameters) may depend upon the number of units within each cluster.

Furthermore, when the data is inherently hierarchical (longitudinal data are, indeed, always hierarchical), a hierarchical approach is natural. Introducing random effects takes into account the fact that the variability in the data can be separated into different levels of the hierarchy, meaning that variability due to within-individual fluctuations in the response, and due to between-individual differences can be separated. This facilitates inference at both the level of the individual, and at the level of the population.

This being said, the inclusion of random effects imposes additional assumptions [see (2.4)] that the linear model (2.1) avoids. Even prior to analysis and model selection, the choice of inclusion or exclusion of random effects can be made based on two factors. Firstly, the subject-matter or scientific knowledge may determine suitability of assumptions. How to determine whether an effect should be fixed or random is discussed in Stroup (2012, p.38) and depends largely upon whether the effect can be thought of as arising from a probability distribution. Secondly, the target of inference should also be considered: including random effects may be unnecessary if the goal of the study is inference at the level of the population. More details on this are given in McNeish et al. (2017).

### 2.8.2 Summary of chapter

This chapter began by describing linear regression without random effects for longitudinal data. It was shown how the mean trend can be modelled as a function of time, and how within-cluster dependency can be taken care of by a covariance pattern model. Random effects were then introduced as an alternative means to account for this dependency, and the general linear mixed effect model was presented. Random effect, random intercept, and random intercept and slope models were described. Subsequently, model diagnostics and estimation via ML and REML for LME models and the LM were explained. Finally, a clinical trial data set was used to demonstrate an application of LME models.

I hope to have distinguished differences between the LM and the LME model from a number of standpoints. And also, since Chapter 4 discusses model selection for multivariate models where the focus is on parameters, not random effects, shown how, marginally, the LME model is a special case of the LM.

# Chapter 3

# Generalised Linear Mixed Models

When the response variable is no longer continuous, is asymmetric or has heavy tails, a more general class of models is required than the linear models of Chapter 2. For example, when the response is a binary (e.g. success/failure) or count variable (e.g. the number of accidents), the linear models of Chapter 2 are of limited use.

The joint density of non-Normal multivariate data is not straightforward to specify. This is because specifying the joint distribution of a non-normal multivariate response without introducing random effects requires specifying more than just the pair-wise associations (the correlations in the linear model); higher order associations must also be specified, which typically entails a large number of parameters (Fitzmaurice et al. 2004*a*).

Thus, for non-Normal multivariate data, two methods are commonly used: marginal models (or population-averaged models) and generalised linear mixed models (GLMMs) (or subject-specific models). Marginal models avoid specification of the joint density of the data altogether (Liang & Zeger 1986), whereas GLMMs specify the joint density of a cluster via a conditional model making use of random effects.

GLMMs will be the focus of this chapter. They can be thought of as a generalisation of generalised linear models (GLMs) to clustered, or correlated data, and can also be considered a generalisation of the LME model to non-Normal or discrete data.

## 3.1 Formulating a GLMM

A GLMM requires four components:

- the conditional distribution of the response for the $i$th cluster and $j$th unit given the random effects, $y_{ij}|\boldsymbol{b}_i$,
- a linear predictor, $\boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\intercal}\boldsymbol{b}_i$,
- a link function, $g(\cdot)$,

- the distribution of the random effects, $\boldsymbol{b}_i$.

The *exponential family* of distributions includes a large number of commonly used distributions e.g. Poisson, binomial, gamma and so on. The conditional distribution of the response in a GLMM given the random effects must be a member of the exponential family. This means that, given the random effects, the conditional distribution of the response (of cluster $i$ and unit $j$) can be written in the form (McCulloch & Searle 2001, p.221)

$$f_{y_{ij}|\boldsymbol{b}_i}(y_{ij}|\eta_{ij},\phi) = \exp\left(\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} - c(y_{ij},\phi)\right) \tag{3.1}$$

with independence assumed for all $i = 1, ..., N$, and conditional independence given the random effects for all $j = 1, ..., n_i$. We have that $\phi$ is a scale parameter; $a(\cdot)$ is a function that determines the specific distribution, for example, $a(x) = e^x$ for the Poisson distribution, and $a(x) = \log(1 + e^x)$ for the Bernoulli distribution (Wand 2007); $c(y_{ij}, \phi)$ is a constant that makes the expression integrate to one; and the parameter $\eta_{ij}$ is known as the *canonical parameter*.

The link function $g(\cdot)$ relates the expected value of the $(i, j)$th response conditional on the random effects to the linear predictor $\boldsymbol{x}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\mathsf{T}\boldsymbol{b}_i$, where $\boldsymbol{x}_{ij}^\mathsf{T}$ and $\boldsymbol{z}_{ij}^\mathsf{T}$ are the $j$th rows of the $i$th fixed effect and random effect design matrices respectively. That is, we have

$$g(\mathbb{E}[y_{ij}|\boldsymbol{b}_i]) = \boldsymbol{x}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\mathsf{T}\boldsymbol{b}_i.$$

Note that, the linear predictor, $\boldsymbol{x}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\mathsf{T}\boldsymbol{b}_i$, is linear not necessarily in terms of the covariates, but in terms of the regression coefficients and random effects.

To complete the specification, we (typically) have that the $\boldsymbol{b}_i$ are mean zero, normally distributed, and with variance-covariance matrix $\boldsymbol{D}$.

At this point, it is worth noting why GLMMs get their name. The *models* are *linear* in the regression coefficients, *mixed* because the linear predictor includes fixed and random effects, and *generalised* because of the presence of a link function which need not be the identity.

The *canonical link*, which is unique to each distribution, is such that $g(\mathbb{E}[y_{ij}|\boldsymbol{b}_i]) = \eta_{ij}$ (De Jong et al. 2008, p.66). For example, as we will see in Section 3.2, the canonical link of the Poisson distribution is $\log(x)$, and the canonical link of the Bernoulli distribution is the logit link $\log(\frac{x}{1-x})$. Furthermore, for the canonical link, the canonical parameter becomes equal to the linear predictor. That is, we have $\eta_{ij} = \boldsymbol{x}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\mathsf{T}\boldsymbol{b}_i$.

### 3.1.1 Conditional moments of a GLMM

The conditional moments of model (3.1) can be found as illustrated in De Jong et al. (2008, p.37). In particular, we have

$$\mathbb{E}[y_{ij}|\boldsymbol{b}_i] = a'(\eta_{ij}),$$

where $a'$ is the derivative of $a$ with respect to $\eta_{ij}$. Furthermore, we have

$$\text{Var}(y_{ij}|\boldsymbol{b}_i) = \phi V(\mathbb{E}[y_{ij}|\boldsymbol{b}_i]),$$

where $V(\cdot) = a''(\cdot)$ is known as the *variance function* and relates the conditional variance to the conditional mean.

### 3.1.2 The marginal distribution derived from the GLMM

The marginal distribution of the $i$th cluster is

$$\begin{aligned} f_{\boldsymbol{y}_i}(\boldsymbol{y}_i) &= \int f_{\boldsymbol{y}_i,\boldsymbol{b}_i}(\boldsymbol{y_i}, \boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \int f_{\boldsymbol{y}_i|\boldsymbol{b}_i}(\boldsymbol{y}_i|\boldsymbol{b}_i) f_{\boldsymbol{b}_i}(\boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \int \prod_{j=1}^{n_i} f_{y_{ij}|\boldsymbol{b}_i}(y_{ij}|\boldsymbol{b}_i) f_{\boldsymbol{b}_i}(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned} \tag{3.2}$$

where the third equality follows from independence of the $(i, j)$th response given the random effects. Thus, by specifying the conditional distribution of the response and the distribution of random effects, an expression for the marginal density is obtainable. Using the rule of double expectations, we have that marginally (McCulloch & Searle 2001, p.222)

$$\mathbb{E}[y_{ij}] = \mathbb{E}[\mathbb{E}[y_{ij}|\boldsymbol{b}_i]] = \mathbb{E}[g^{-1}(\boldsymbol{x}_{ij}^\intercal \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\intercal \boldsymbol{b}_i)].$$

Thus, only for the identity link do we have $\mathbb{E}[y_{ij}] = \boldsymbol{x}_{ij}^\intercal \boldsymbol{\beta}$. This means that, for non-identity link functions, $\boldsymbol{x}_{ij}^\intercal \boldsymbol{\beta}$ does not have the interpretation as the marginal mean trend; the LME model (2.3) is indeed a special case. The regression coefficients of a model which does not have the identity link, can be interpreted in terms of the corresponding transform, $g(\cdot)$, of the expected response, or must be transformed via the inverse of the link function, $g^{-1}(\cdot)$ , to be interpretable on the same scale as the expected response.

In addition, the regression coefficients of a GLMM have an interpretation at the level of the cluster. This is because the regression coefficients must be interpreted while holding $\boldsymbol{b}_i$ fixed. For interpreting effects of continuous covariates, one should consider the conditional mean response of a given cluster with a specific $\boldsymbol{b}_i$. Whereas, for interpreting binary or categorical covariates, one should contrast two different clusters (perhaps of different covariate values) but that have the same random effect (Fitzmaurice et al. 2004*a*, p.361). This interpretation at the level of the cluster is a characteristic of GLMMs, and, as such, the target of inference of GLMMs is the level of the cluster. LME models are an exception in that both an interpretation at the level of the cluster and at the marginal level are available; for GLMMs without an identity link, this is not possible.

Lastly, it is worth mentioning that marginal models (or population averaged models), which posit no distributional assumptions and which specify only the marginal moments, are an alternative means for modelling non-Normal clustered or longitudinal data (Liang & Zeger 1986). The target of inference of such models is the level of the population (e.g. Fitzmaurice et al. 2004*a*, p.291).

## 3.2 Examples

### 3.2.1 A Bernoulli GLMM

Suppose that the response of interest is binary, taking values 0 and 1. Then, a Bernoulli GLMM with logit link can be formulated as

$$\begin{aligned} f_{y_{ij}|\boldsymbol{b}_i}(y_{ij}|\boldsymbol{b}_i) =& p_{ij}^{y_{ij}}(1-p_{ij})^{(1-y_{ij})}, \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) =& \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i, \\ \boldsymbol{b}_i \sim & \ N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{3.3}$$

where $p_{ij} = P(y_{ij} = 1|\boldsymbol{b}_i)$ is the probability of unit $j$ of cluster $i$ taking value one conditional on $\boldsymbol{b}_i$, and $1-p_{ij}$ is the probability of unit $j$ of cluster $i$ taking value zero. Such a model assumes a natural between-cluster diversity in the tendency to respond positively. The logit link function is the logarithm of the odds that $y_{ij}|\boldsymbol{b}_i$ takes value 1, where the odds are given by $\dfrac{p_{ij}}{1-p_{ij}}$. Thus, it is the odds that are log-linear in the regression coefficients.

To see how (3.3) is of the form (3.1) note that

$$\begin{aligned} p_{ij}^{y_{ij}}(1-p_{ij})^{(1-y_{ij})} &= \exp(y_{ij}\log(p_{ij}) + (1-y_{ij})\log(1-p_{ij})) \\ &= \exp\Big(y_{ij}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) + \log(1-p_{ij})\Big), \end{aligned} \tag{3.4}$$

from which it follows that the canonical parameter takes the form

$$\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right),$$

and, thus, the logit link is the canonical link for the Bernoulli distribution. In addition, inverting this relationship gives

$$p_{ij} = \frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}}. \tag{3.5}$$

Substitution of (3.5) into the second term in the exponent of (3.4) and forming a common denominator yields (3.1).

For the Bernoulli distribution then, we have that the scale parameter $\phi = 1$ and the constant $c(y_{ij}, \phi)$ is zero. Furthermore, it is well known that $\mathbb{E}[y_{ij}|\boldsymbol{b}_i] = p_{ij}$ and Var$(y_{ij}|\boldsymbol{b}_i) = p_{ij}(1 - p_{ij})$. So, the conditional variance of a Bernoulli random variable depends only on the conditional expectation. This is not necessarily the case for other distributions, whose variance may also depend upon the scale parameter (Stroup 2012, p.125).

Note that, in practice, for binary data and relatively few units within clusters, there is rarely enough variability in the data to estimate more random effect parameters than that of the variance of a random intercept (Fitzmaurice et al. 2004*a*, p.344).

### 3.2.2 A Poisson GLMM

When the response is a count variable (that is, taking values 0,1,...), the Poisson distribution may be appropriate. This is in contrast to binary data whose distribution is necessarily Bernoulli (Fitzmaurice et al. 2004*a*).

A Poisson GLMM with its canonical log-link can be expressed as

$$\begin{aligned} f_{y_{ij}|\boldsymbol{b}_i}(y_{ij}|\boldsymbol{b}_i) =& \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!} e^{-\lambda_{ij}}, \\ \log \lambda_{ij} =& \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i, \\ \boldsymbol{b}_i \sim& \ N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned}$$

where $\lambda_{ij} = \mathbb{E}[y_{ij}|\boldsymbol{b}_i]$ is the expected count of unit $j$ in cluster $i$. Re-writing the conditional distribution as

$$\exp(y_{ij} \log \lambda_{ij} - \lambda_{ij} - \log(y_{ij}!)),$$

it becomes visible that the log-link is indeed canonical since $\eta_{ij}$ takes the form $\log \lambda_{ij}$. We also have that $\phi = 1$, $c(y_{ij}, \phi) = \log(y_{ij}!)$ and $a(\eta_{ij}) = e^{\eta_{ij}}$.

For longitudinal data, the measurement occasions may not be equally separated. To account for this, an offset, $T_{ij}$, the length of the measurement interval from time $j - 1$ to time $j$, may be introduced into the linear predictor to give

$$\log \lambda_{ij} = \log T_{ij} + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i.$$

Re-arranging this gives

$$\log \left(\frac{\lambda_{ij}}{T_{ij}}\right) = \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i,$$

from which $\dfrac{\lambda_{ij}}{T_{ij}}$ gets the interpretation as the expected rate of counts at time $j$, a more easily interpretable measure across unevenly spread measurement occasions.

The Poisson model imposes the rather restrictive assumption that $\mathbb{E}[y_{ij}|\boldsymbol{b}_i] =$ Var$(y_{ij}|\boldsymbol{b}_i) = \lambda_{ij}$. *Overdispersion* is what occurs if Var$(y_{ij}|\boldsymbol{b}_i) > \mathbb{E}[y_{ij}|\boldsymbol{b}_i]$, and tends to be the case more often than not (Fitzmaurice et al. 2004*a*, p.297). So overdispersion has to, in general, be accounted for if the Poisson model is to be of any use in application.

# 3.3 Estimation

## 3.3.1 Maximum likelihood

Maximum likelihood is a commonly used method for estimation of parameters in a GLMM. Taking the logarithm of (3.2) gives

$$\log(f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)) = \log\left(\int \prod_{j=1}^{n_i} f_{y_{ij}|\boldsymbol{b}_i}(y_{ij}|\boldsymbol{b}_i) f_{\boldsymbol{b}_i}(\boldsymbol{b}_i) d\boldsymbol{b}_i\right),$$

which, summed over all clusters and inserting expressions for the densities $f_{y_{ij}|\boldsymbol{b}_i}(y_{ij}|\boldsymbol{b}_i)$ and $f_{\boldsymbol{b}_i}(\boldsymbol{b}_i)$, gives the marginal log-likelihood of model (3.1) as

$$\ell_N(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) =$$
$$\sum_{i=1}^{N} \log\left(\int \frac{1}{\sqrt{2\pi}|\boldsymbol{D}|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}\boldsymbol{b}_i^\intercal \boldsymbol{D}^{-1}\boldsymbol{b}_i + \sum_{j=1}^{n_i}\Big[\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} - c(y_{ij}, \phi)\Big]\Big) d\boldsymbol{b}_i\right), \tag{3.6}$$

where $\boldsymbol{y}$ is a stack of all clusters, and where $\boldsymbol{\theta}$ is a vector including scale parameter $\phi$, and the variance-covariance parameters of the random effects.

The integrals in (3.6) are analytically intractable. For the situations we are interested in (clustered/longitudinal data), it is suitable to maximise (3.6) numerically, which involves numerical integration (McCulloch & Searle 2001, p.226). For relatively low dimensions of random effects, Gauss-Hermite quadrature may be used (as in McCulloch & Searle 2001, p.270) to perform such integration.

In particular, for a univariate random effect, $b_i \sim N(0, \sigma_b^2)$, (3.6) is expressible as

$$\sum_{i=1}^{N} \log\left(\int h(b_i) \frac{1}{\sqrt{2\pi}\sigma_b} e^{\frac{-1}{2\sigma_b^2} b_i^2} db_i\right),$$

where

$$h(b_i) = \exp\left(\sum_{j=1}^{n_i}\Big[\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} - c(y_{ij}, \phi)\Big]\right),$$

and the dependency of $h$ on $b_i$ is through the canonical parameter $\eta_{ij}$. After a transformation of variable, such a sum of integrals becomes of the form

$$\sum_{i=1}^{N} \log\left(\int h(\sqrt{2}\sigma_b b_i) \frac{e^{-b_i^2}}{\sqrt{\pi}}\right), \tag{3.7}$$

and can be approximated via Gauss-Hermite quadrature (e.g. McCulloch & Searle 2001) as

$$\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} h(\sqrt{2}\sigma_b x_k) \frac{w_k}{\sqrt{\pi}}\right),$$

where $K$ is the number of evaluation points, $x_k$, and $w_k$ their corresponding weights. These can be obtained via, for example, using the **statmod** package (Smyth 2005) in R.

## 3.4 Data illustration

A randomised, double blind, clinical trial dataset comparing two oral treatments for toe nail infection is now considered. The dataset was obtained from the accompanying website (Fitzmaurice et al. 2004$b$) of the textbook Applied Longitudinal Analysis (Fitzmaurice et al. 2004$a$, p.355). The response of interest is the degree of onycholysis (extent of separation of the nail from the nail bed), which is measured as a binary variable (0 = none or mild, 1 = moderate or severe). 294 patients were measured on a maximum of 7 occasions during a period of 18.5 months. There is some spread in the exact timing of measurements, but measurements were obtained at baseline, and at (more or less) 1, 2, 3, 6, 9 and 12 months after baseline. Patients were randomised to two anti-fungal oral treatments (B=) Itraconazole and (A=) Terbinafine at baseline (Fitzmaurice et al. 2004$b$). Table 3.1 shows the number of severe or moderate cases for treatments A and B for all 7 visits.

Table 3.1: The number of severe or moderate cases by treatment group for each of the seven visits.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| A | 55 | 48 | 40 | 29 | 8 | 8 | 6 |
| B | 54 | 49 | 44 | 29 | 14 | 10 | 14 |

A large number of patients dropped out throughout the course of the study resulting in an incomplete data set. Nevertheless, an all-available data analysis will be conducted. For such an analysis to be valid, the reasons for missing data should be unconnected to either the observed or unobserved values, meaning that the observed values of the remainers (those that do not drop out) must be a random sample of the corresponding values in the population, which is likely an unrealistic assumption.

### 3.4.1 A plausible model

The *glmmML* command from the R package **glmmML** (Broström 2017) was used to fit model (3.3) with logit link and the following linear predictor:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}\text{group}_i + b_i, \tag{3.8}$$

where $p_{ij}$ is the probability of moderate or severe onycholysis at occasion $j$ for individual $i$; $t_{ij}$ is the exact time in months; and $\text{group}_i$ is the binary treatment group variable (reference group = Treatment B, non-reference group = Treatment A). Since the treatment groups were generated by randomisation (not by pre-existing distinguishing factors as in the depression data set (2.7)), it makes sense to assume that both groups have the same intercept at baseline. Furthermore, allowing separate slopes for both treatment groups facilitates comparison of both treatments. For obtaining ML estimates of the parameters, Gauss-Hermite quadrature with 30 quadrature points was used to approximate the integrals involved in maximising the log-likelihood.

### 3.4.2 Results

The parameter estimates obtained via ML are given in Table 3.2. Significance of estimates was judged using the asymptotic properties of ML estimators (with the inverse of the observed information matrix as the estimated covariance matrix of the model parameters). The negative (and significant) value for $\beta_1$ means that Treatment B is effective in reducing the probability/odds of a severe response at the level of the individual. However, since the interaction term in the linear model is also negative ($\beta_2 = -0.14$) and significant at the 0.05 level (p-value = 0.028), Treatment A can be concluded as more efficient than Treatment B at the level of the individual.

Table 3.2: Parameter estimates from model (3.8).

| | Binary outcome |
|---|---|
| Fixed Effects: | |
| Intercept | -1.70$^{***}$ |
| $t_{ij}$ | -0.39$^{***}$ |
| $t_{ij}\text{group}_i$ | -0.14$^{*}$ |
| Random effects: | |
| $\sigma_b$ | 4.01 |

$^{*}$p $<$ .05; $^{**}$p $<$ .01; $^{***}$p $<$ .001

On the scale of the linear predictor, the parameter estimates are not so easily interpretable. Instead, suppose that we are interested in understanding the change in odds of onycholysis with time for both treatments at the level of the individual. Since the variable month is modelled as linear, the effect on the odds of severe response by a unit increase in time is the same regardless of when the unit increase takes place. For any month $x$ up to one month before the end of study (unless a forecast beyond the study period is to be made), we have that for individual $i$ of treatment group B with random intercept $b_i$,

$$\begin{aligned}&\frac{\text{Estimated odds of severe onycholysis for individual i}|b_i,\text{group = B}, t = x+1}{\text{Estimated odds of severe onycholysis for individual i}|b_i,\text{group = B}, t = x}\\ =&\frac{e^{\hat{\beta}_0+\hat{\beta}_1(x+1)+b_i}}{e^{\hat{\beta}_0+\hat{\beta}_1x+b_i}}\\ =&e^{\hat{\beta}_1} \approx 0.68.\end{aligned}$$

That is, the odds of moderate or severe onycholysis for any individual in group B is estimated to be roughly $32\%$ smaller after one additional month of treatment. In this way, by defining odds ratios and arriving at a multiplicative effect, the parameter estimates are more readily interpretable.

Similarly, for an individual in group A we have that,

$$\begin{aligned}&\frac{\text{Estimated odds of severe onycholysis for individual i}|b_i,\text{group = A}, t = x+1}{\text{Estimated odds of severe onycholysis for individual i}|b_i,\text{group = A}, t = x}\\ =&\frac{e^{\hat{\beta}_0+\hat{\beta}_1(x+1)+\hat{\beta}_2(x+1)+b_i}}{e^{\hat{\beta}_0+\hat{\beta}_1x+\hat{\beta}_2x+b_i}}\\ =&e^{\hat{\beta}_1+\hat{\beta}_2} \approx 0.59.\end{aligned}$$

This means that the odds of moderate or severe onycholysis for any individual in group A is estimated to be roughly $41\%$ smaller after one additional month of treatment. So, the odds of moderate or severe onycholysis is reduced more by each additional month of Treatment A compared with Treatment B at the level of the individual.

Given the fixed effect parameter estimates and estimates of the realised values of the random effects, which together form estimates of the individual specific linear predictors, estimates of the individual specific probabilities can be obtained by application of the inverse logit transform, $\dfrac{e^x}{1+e^x}$. The estimated individual specific probabilities are plotted against time for treatment group A in Figure 3.1 and treatment group B in Figure 3.2.

From these plots it is visible that, by the end of the study, there is a selection of patients in group B that have slightly higher probabilities of a moderate or severe outcome in comparison with those in group A. It is also of note that a typical patient (i.e. with random effect set to its mean of zero) has a baseline probability of $0.15$ (for both groups). Thus, there are many probability trajectories hidden in the lower left corners of Figure 3.1 and Figure 3.2.

Finally, it may be of interest to know, as a measure of treatment efficacy for both groups, what fraction of individuals have probability greater than one half of moderate or severe onycholysis after 3 months of treatment. It turns out that $20\%$ of individuals in group A had a probability of severe onycholysis greater than one half after 3 months of treatment, whereas $24\%$ of individuals in group B had such a probability.

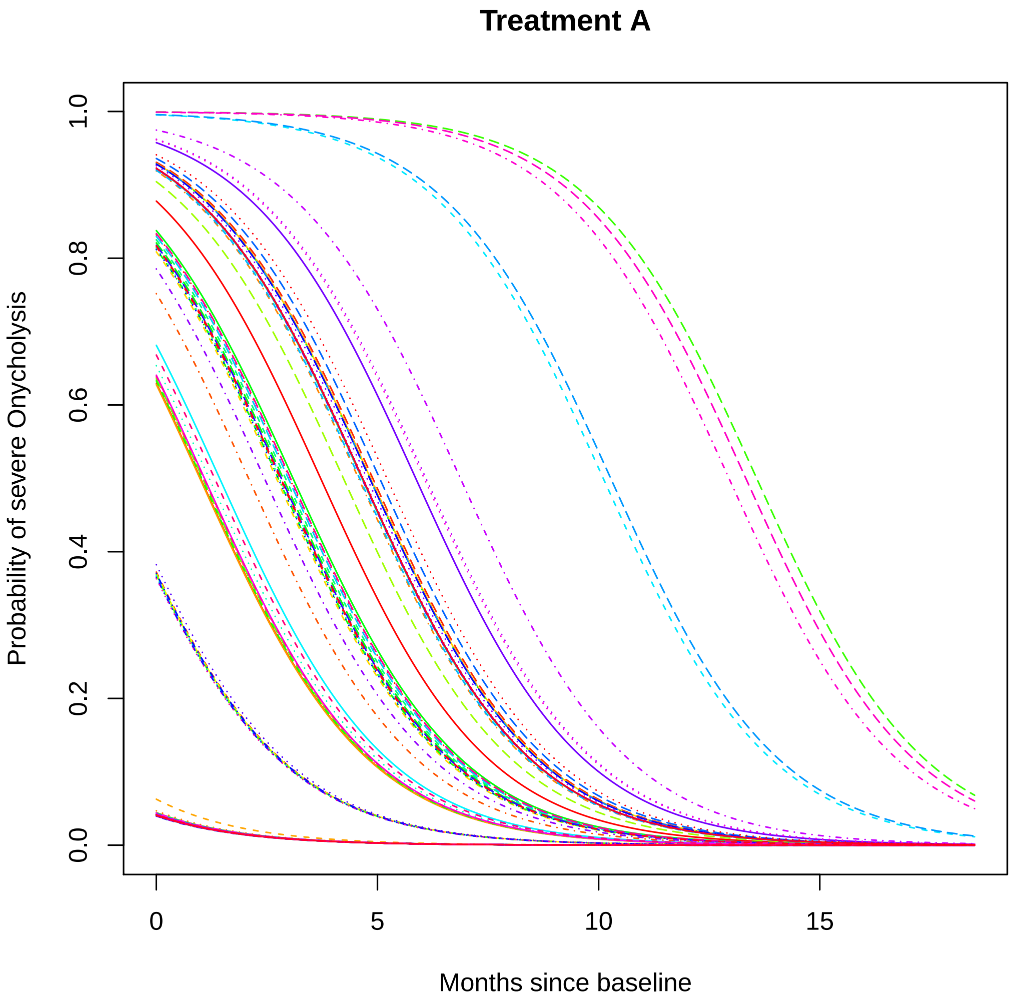


Figure 3.1: Estimated individual specific probabilities for treatment group A

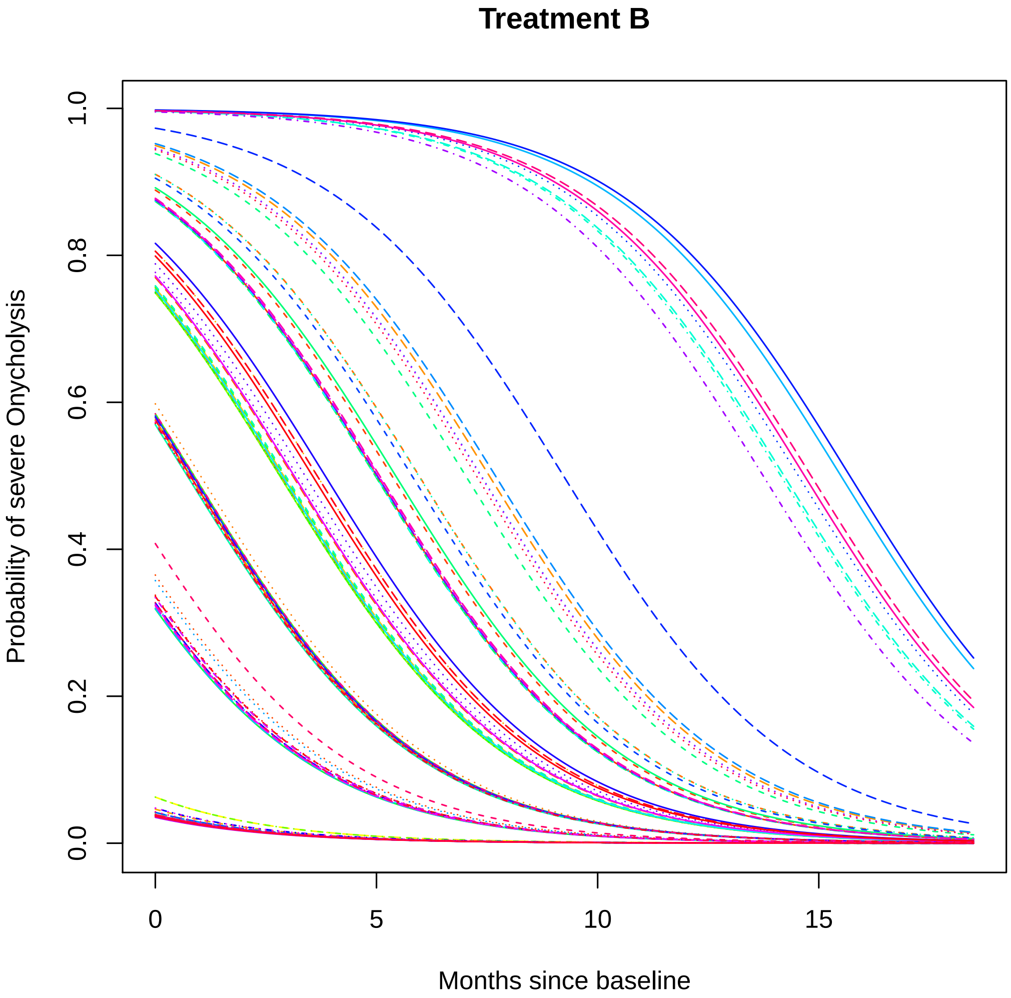


Figure 3.2: Estimated individual specific probabilities for treatment group B

## 3.5 Chapter summary

This chapter has presented essential theory of generalised linear mixed models. In particular, the exponential class of distributions was introduced, and the components of a general GLMM explained. The Bernoulli and Poisson models were given as examples of the general GLMM. Estimation via ML was then discussed along with Gauss-Hermite quadrature as a means for approximating the integrals of the likelihood. An application to a binary clinical trial dataset involving comparison of two treatments for toe-nail infection was then presented. Attention was given to interpretation of regression coefficients which, due to the presence of a non-identity link function, is not as straightforward as the Normal model.

# Chapter 4

# The Focussed Information Criterion for Clustered Data

In any field in which clustered or longitudinal data are collected (as with any field of research), there are specific questions to be answered. For the models of Chapter 2 and Chapter 3, these questions are typically formulated in terms of the regression coefficients. In general, model selection has traditionally been based around finding a model which fits the data well with as few parameters as possible, and without regard to the particular questions at hand. As discussed in Section 2.5.3, the Akaike information criterion (AIC), for example, facilitates such model selection. In this chapter, the focussed information criterion (FIC), a model selection criterion for targeting specific questions, as introduced by Claeskens & Hjort (2003), Claeskens et al. (2008), is presented. In addition, a multivariate framework is developed, within which the FIC becomes available as a covariate selector for multivariate LMs, LME models and GLMMs. This is followed by some theoretical results, a simulation study, and longitudinal data illustrations.

## 4.1 FIC for independent data

The focussed information criterion (FIC) is a model selection criterion that ranks models in terms of their appropriateness for a given purpose, or goal. The goal is to precisely estimate a parameter of primary interest, henceforth called the *focus parameter*. The focus parameter could be, for example, an interaction effect, an expected response, a rate ratio and so on.

The ability of a model to precisely estimate the focus is determined by the FIC scores which, for each model, are unbiased estimators of the mean squared error (MSE) of the limiting distribution of the focus. These estimators are derived using the large sample asymptotic theory of maximum likelihood (ML) estimators. In this section, the FIC scores for the regression setting with independent data are given. For which, the setting is laid out and the main steps regarding asymptotic results are presented. Initially, a multiple linear regression example will be used to familiarise the reader with the score vectors and information matrices. A natural extension of the FIC to multivariate models will then be presented in Section 4.2. Until then, keep in mind for this section that we are working with a sample of $n$ *independent* observations (no clustering).

### 4.1.1 Framework and goal

Under contention are a list of models ranging from the narrow model (of fewest parameters) to the widest model (of most parameters), all of which are nested within the wide model. The parameters $\boldsymbol{\nu}$, of dimension $p$, are included in every model. However, only the wide model includes the parameter vector $\boldsymbol{\gamma}$, of dimension $q$, in its entirety. The models in-between the wide and narrow models contain only some of the components of $\boldsymbol{\gamma}$. For this reason, $\boldsymbol{\nu}$ is said to be *protected*, whereas $\boldsymbol{\gamma}$ is said to be *unprotected*.

The candidate models are within a *locally misspecified framework*. That is, data $y_i$, which are assumed to be independent given covariates $\boldsymbol{x}_i$, are generated by the wide model

$$f_{\text{wide},n} = f(y_i|\boldsymbol{x}_i, \boldsymbol{\nu}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n}), \tag{4.1}$$

which is assumed to be true. Furthermore, $\boldsymbol{\nu}_0$ is the true value of $\boldsymbol{\nu}$, whereas $\boldsymbol{\gamma}_0$ is said to be the null value of $\boldsymbol{\gamma}$, and $\boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n}$ is the true value of $\boldsymbol{\gamma}$.

That is, we have a sequence of true wide models for each sample size $n$, and, as $n$ grows, the true value of $\boldsymbol{\gamma}$ approaches $\boldsymbol{\gamma}_0$. So, $f_{\text{wide},n}$, the true data generating mechanism, is assumed to be a distance $\boldsymbol{\delta}/\sqrt{n}$ away from the narrow model in terms of the unprotected $\boldsymbol{\gamma}$ parameters. When $\boldsymbol{\delta} = \sqrt{n}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) = 0$, we have that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, and we are back at the narrow model. In other words, $f_{\text{wide}}(\boldsymbol{\nu}, \boldsymbol{\gamma}_0) = f_{\text{narr}}(\boldsymbol{\nu})$. A model in-between the wide and narrow, model $S$ say, where $S$ is a subset of $\{1, ..., q\}$, contains the $\gamma_j$ with $j \in S$, and sets $\gamma_j = \gamma_{0,j}$ for those $j \notin S$.

Thus, the locally misspecified framework allows us to consider a range of models that are small perturbations away from a narrow model. Such a framework facilitates questions such as ‘How far from the narrow model is too far (in terms of uncertainty introduced by more parameters)?’, and ‘How close to the narrow model is too close (in terms of simplicity: too few parameters)?’. It may be that a true value that changes with sample size is conceptually bothering, but this is not how the framework is to be interpreted: the principal advantage of this framework (as we will see) is that it leads to attractive asymptotic results (Claeskens et al. 2008, p.128).

The focus parameter is defined as $\mu = \mu(\boldsymbol{\nu}, \boldsymbol{\gamma})$, a function of the model parameters. For example, the focus could be prediction of the mean response given a set of covariates, or a particular quantile of the true data generating mechanism. The aim of FIC is to estimate the MSE of the limiting distribution of the focus parameter for each of the models. The model with smallest estimated MSE of the limiting distribution of the focus provides the most precise estimate of the focus parameter and is therefore deemed the best by FIC. It is well-known that the MSE is expressible as the sum of a term for bias squared and a term for variance (e.g. Casella & Berger 2002, p.330). This means that for any estimator $\hat{\mu}$ of the true focus parameter value $\mu_0$, one may write its MSE in estimation as

$$\begin{aligned}\text{MSE}(\hat{\mu}) &= \mathbb{E}[(\hat{\mu} - \mu_0)^2] \\ &= (\mathbb{E}[\hat{\mu} - \mu_0])^2 + \mathbb{E}[\hat{\mu}^2] - \mathbb{E}[\hat{\mu}]^2 \\ &= \text{bias}^2(\hat{\mu}) + \text{Var}(\hat{\mu}).\end{aligned}$$

The FIC procedure involves estimating the mean squared error of the limiting distribution of the focus by finding separate estimators for the bias squared and variance of the limiting distribution and summing both.

### Example

As a simple example, consider a linear regression model with constant variance $\sigma^2$, intercept $\beta_0$, and covariate $x_1$ with effect $\beta_1$. Alternative models consist of including or excluding two more covariates $x_2$ and $x_3$. The goal is to find the best model for estimating the expected response when the covariates take on values $(x_1, x_2, x_3)$ = $(a, b, c)$, say. That is, the focus of interest is $\mu = \mu(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbb{E}[y_i|x_1 = a, x_2 = b, x_3 = c]$, and the wide model under consideration is

$$N(\beta_0 + \beta_1 x_{1,i} + \gamma_0 x_{2,i} + \gamma_1 x_{3,i}, \sigma^2).$$

The protected parameters are

$$\boldsymbol{\nu} = \begin{pmatrix} \sigma^2 \\ \beta_0 \\ \beta_1 \end{pmatrix},$$

and the unprotected parameters are

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix},$$

which have null value $\boldsymbol{\gamma}_0 = (0, 0)^\intercal$.

The narrow model is therefore,

$$N(\beta_0 + \beta_1 x_{1,i}, \sigma^2),$$

and the two models in-between the wide and narrow are

$$\begin{aligned}&N(\beta_0 + \beta_1 x_{1,i} + \gamma_0 x_{2,i}, \sigma^2),\\ &N(\beta_0 + \beta_1 x_{1,i} + \gamma_1 x_{3,i}, \sigma^2),\end{aligned}$$

which correspond to setting $\boldsymbol{\gamma} = (\gamma_0, 0)^\intercal$ and $\boldsymbol{\gamma} = (0, \gamma_1)^\intercal$ respectively.

Introducing two extra parameters may reduce the bias in estimation of the mean response, but doing so will also increase the number of parameters to be estimated, resulting in more uncertainty in the estimation of parameters. So here, the FIC will help the statistician find the best balance between a biased estimator with less uncertainty (variance) and a less biased, but more uncertain estimator of the mean response. This bias versus variance trade-off is called the *Goldilocks principle* in (Cressie & Wikle 2011, p.7).[1]

### 4.1.2 Score vectors and the expected information matrix

To be able to show the main steps in the derivation of the FIC, necessary quantities as in Claeskens & Hjort (2003) will first be introduced.

Provided sufficient smoothness in the model parameters about the null point $(\boldsymbol{\nu}_0, \gamma_0)$, each observation has an associated score vector

$$\begin{pmatrix} \boldsymbol{u}(y_i|\boldsymbol{x}_i) \\ \boldsymbol{v}(y_i|\boldsymbol{x}_i) \end{pmatrix} = \begin{pmatrix} \partial \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\nu}_0, \gamma_0)/\partial \boldsymbol{\nu} \\ \partial \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\nu}_0, \gamma_0)/\partial \boldsymbol{\gamma} \end{pmatrix}, \tag{4.2}$$

where $\boldsymbol{x}_i$ is the associated vector of covariates for observation $i$. For each observation, there is an associated expected information matrix, which evaluated at the null point is

$$\boldsymbol{J}_i = -\mathbb{E}\left[\begin{pmatrix} \frac{\partial^2 \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\nu}_0,\gamma_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^\intercal} & \frac{\partial^2 \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\nu}_0,\gamma_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\gamma}^\intercal} \\ \frac{\partial^2 \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\nu}_0,\gamma_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\nu}^\intercal} & \frac{\partial^2 \log f(y_i|\boldsymbol{x}_i,\boldsymbol{\nu}_0,\gamma_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\intercal} \end{pmatrix}\right],$$

and, under the wide model, is equal to the variance-covariance matrix of the score vector

$$\boldsymbol{J}_i = \text{Var}_0 \begin{pmatrix} \partial \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\nu}_0, \gamma_0)/\partial \boldsymbol{\nu} \\ \partial \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\nu}_0, \gamma_0)/\partial \boldsymbol{\gamma} \end{pmatrix}.$$

Averaged over all observations, this becomes

$$\boldsymbol{J}_{\text{full},n} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{J}_i = \begin{pmatrix} \boldsymbol{J}_{00,n} & \boldsymbol{J}_{01,n} \\ \boldsymbol{J}_{10,n} & \boldsymbol{J}_{11,n} \end{pmatrix},$$

where, averaged over all observations and evaluated at the null value, $\boldsymbol{J}_{00,n}$ is of dimension $p \times p$ and is the variance of the components of the score corresponding to the protected parameters $\boldsymbol{\nu}$; $\boldsymbol{J}_{11,n}$ is of dimension $q \times q$ and is the variance of the score of the unprotected parameters $\boldsymbol{\gamma}$; $\boldsymbol{J}_{01,n}$, of dimension $p \times q$, is the covariance between the $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ components of the score. Finally, being a symmetric matrix, $\boldsymbol{J}_{10,n}$ is the transpose of $\boldsymbol{J}_{01,n}$.

Under certain conditions, as $n$ tends to infinity, $\boldsymbol{J}_{\text{full},n}$ tends to the $(p+q)\times(p+q)$ limiting information matrix of the wide model (4.1)

$$\boldsymbol{J}_{\text{wide}} = \begin{pmatrix} \boldsymbol{J}_{00} & \boldsymbol{J}_{01} \\ \boldsymbol{J}_{10} & \boldsymbol{J}_{11} \end{pmatrix},$$

[1] In the children's story, Goldilocks is a little girl who wants things just right.

where the upper left block of $\boldsymbol{J}_{\text{wide}}$ is of dimension $p \times p$ and the lower right block is of dimension $q \times q$. This has inverse

$$\boldsymbol{J}_{\text{wide}}^{-1} = \begin{pmatrix} \boldsymbol{J}^{00} & \boldsymbol{J}^{01} \\ \boldsymbol{J}^{10} & \boldsymbol{J}^{11} \end{pmatrix} . \tag{4.3}$$

Let the size of set $S$ be denoted by $|S|$. Then, similarly the expected information matrix of dimension $(p + |S|) \times (p + |S|)$ for each model $S$, may be defined as

$$\boldsymbol{J}_{S,n} = \frac{1}{n} \sum_{i=1}^{n} \text{Var}_0 \begin{pmatrix} \boldsymbol{u}(y_i|\boldsymbol{x}_i) \\ \boldsymbol{v}_S(y_i|\boldsymbol{x}_i) \end{pmatrix} , \tag{4.4}$$

where $\boldsymbol{v}_S(y_i|\boldsymbol{x}_i) = \boldsymbol{\pi}_S \boldsymbol{v}(y_i|\boldsymbol{x}_i)$, with $\boldsymbol{\pi}_S$ a projection matrix of zeros and ones that maps any vector to the same vector but only containing its entries that belong to the set $S$. More formally, for any vector $\boldsymbol{m}$, $\boldsymbol{\pi}_S \boldsymbol{m} = \boldsymbol{m}_S$, where $\boldsymbol{m}_S$ is of dimension $|S|$ and contains only those entries of $\boldsymbol{m}$ that are in $S$. Similarly, for any matrix $\boldsymbol{M}$, $\boldsymbol{\pi}_S$ maps that matrix to the same matrix, but only includes the rows belonging to set $|S|$. That is, $\boldsymbol{\pi}_S \boldsymbol{M} = \boldsymbol{M}_S$, where the rows of $\boldsymbol{M}_S$ are those belonging to $\boldsymbol{M}$ that are in the set $S$.

Under certain conditions, $\boldsymbol{J}_{S,n}$ tends to the matrix

$$\boldsymbol{J}_S = \begin{pmatrix} \boldsymbol{J}_{00,S} & \boldsymbol{J}_{01,S} \\ \boldsymbol{J}_{10,S} & \boldsymbol{J}_{11,S} \end{pmatrix} = \begin{pmatrix} \boldsymbol{J}_{00} & \boldsymbol{J}_{01} \boldsymbol{\pi}_S^{\mathsf{T}} \\ \boldsymbol{\pi}_S \boldsymbol{J}_{10} & \boldsymbol{\pi}_S \boldsymbol{J}_{11} \boldsymbol{\pi}_S^{\mathsf{T}}, \end{pmatrix} \tag{4.5}$$

as $n$ grows, where the upper left block is of dimension $p \times p$ and the lower right block is of dimension $|S| \times |S|$. This has inverse

$$\boldsymbol{J}_S^{-1} = \begin{pmatrix} \boldsymbol{J}^{00,S} & \boldsymbol{J}^{01,S} \\ \boldsymbol{J}^{10,S} & \boldsymbol{J}^{11,S} \end{pmatrix} . \tag{4.6}$$

### Example

In the multiple linear regression example, the score vector for observation $i$, found by differentiating the logarithm of the density of the wide model with respect to the model parameters, is

$$\begin{pmatrix} \boldsymbol{u}(y_i|\boldsymbol{x}_i) \\ \boldsymbol{v}(y_i|\boldsymbol{x}_i) \end{pmatrix} = \frac{1}{\sigma} \begin{pmatrix} \epsilon_i^2 - 1 \\ \epsilon_i \\ x_{1,i}\epsilon_i \\ x_{2,i}\epsilon_i \\ x_{3,i}\epsilon_i \end{pmatrix} ,$$

where $\boldsymbol{x}_i = (1, x_{1,i}, x_{2,i}, x_{3,i})^{\mathsf{T}}$ and $\epsilon_i = (y_i - \beta_0 - \beta_1 x_{1,i} - \gamma_0 x_{2,i} - \gamma_1 x_{3,i})/\sigma$. The first three entries of the score are associated with the protected parameters $\boldsymbol{\nu} =$

$(\sigma, \beta_0, \beta_1)^\mathsf{T}$, and the last two entries of the score are associated with the unprotected parameters $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^\mathsf{T}$.

Differentiating again, taking minus the expected value and averaging over all observations gives

$$\boldsymbol{J}_{\text{full},n} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & x_{1,i} & x_{2,i} & x_{3,i} \\ 0 & x_{1,i} & x_{1,i}^2 & x_{1,i}x_{2,i} & x_{1,i}x_{3,i} \\ 0 & x_{2,i} & x_{1,i}x_{2,i} & x_{2,i}^2 & x_{2,i}x_{3,i} \\ 0 & x_{3,i} & x_{1,i}x_{2,i} & x_{2,i}x_{3,i} & x_{3,i}^2 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n\sigma^2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \end{pmatrix}. \tag{4.7}$$

Note that in order for $\boldsymbol{J}_{\text{full},n}$ to tend to $\boldsymbol{J}_{\text{wide}}$ a necessary condition is that $\frac{1}{n\sigma^2}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$ converges as $n$ grows.

In our example, $S$ can range from (the narrow) $S = \emptyset$ to (the wide) $S = \{1, 2\}$. When only $\gamma_0$ is included as an extra parameter, $S = \{1\}$ and the required projection matrix is thus $\boldsymbol{\pi}_s = (1, 0)$. The expected information matrix corresponding to this model (including $\gamma_0$ not $\gamma_1$) is

$$\boldsymbol{J}_{S,n} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & x_{1,i} & x_{2,i} \\ 0 & x_{1,i} & x_{1,i}^2 & x_{1,i}x_{2,i} \\ 0 & x_{2,i} & x_{1,i}x_{2,i} & x_{2,i}^2 \end{pmatrix},$$

where for example the lower right hand block (here only one entry), $\boldsymbol{J}_{11,S}$ [see (4.5)], which corresponds to the parameter $\gamma_0$ alone, is calculated as

$$\begin{aligned} \boldsymbol{J}_{11,S} =& \boldsymbol{\pi}_S \boldsymbol{J}_{11} \boldsymbol{\pi}_S^\mathsf{T} \\ =& \frac{1}{n\sigma^2} \sum_{i=1}^{n} \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} x_{2,i}^2 & x_{2,i}x_{3,i} \\ x_{2,i}x_{3,i} & x_{3,i}^2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} x_{2,i}^2. \end{aligned}$$

### 4.1.3 Limiting distributions

The score vectors and information matrices have now been introduced. The main steps required to derive the limiting distribution of the focus parameter will now be shown. For full details and in particular regularity conditions, the reader is referred to Section 3.1 and the appendix of Hjort & Claeskens (2003).

The averages of the score vectors defined for each observation in (4.2) are $\bar{\boldsymbol{u}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}(y_i|\boldsymbol{x}_i)$ and $\bar{\boldsymbol{v}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}(y_i|\boldsymbol{x}_i)$. Under the locally misspecified sequence of models (4.1), and due to independence of observations, it can be shown that the averaged score vector for model $S$ has the following limiting distribution as $n$ grows:

$$\begin{pmatrix} \sqrt{n}\bar{\boldsymbol{u}}_n \\ \sqrt{n}\bar{\boldsymbol{v}}_{S,n} \end{pmatrix} \xrightarrow{d} N_{p+|S|}\left( \begin{pmatrix} \boldsymbol{J}_{01}\boldsymbol{\delta} \\ \boldsymbol{\pi}_S \boldsymbol{J}_{11}\boldsymbol{\delta} \end{pmatrix}, \boldsymbol{J}_S \right). \tag{4.8}$$

The limiting distribution is not centered at zero because the score vectors are evaluated at the narrow model and thus away from the true wide model.

The maximum likelihood estimators of model $S$, $(\hat{\boldsymbol{\nu}}_S, \hat{\boldsymbol{\gamma}}_S)$, can be written in terms of the averaged score vector via a Taylor expansion, and thereby their limiting distribution is determined to be

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\nu}}_S - \boldsymbol{\nu}_0) \\ \sqrt{n}(\hat{\boldsymbol{\gamma}}_S - \boldsymbol{\gamma}_{0,S}) \end{pmatrix} \xrightarrow{d} N_{p+|S|}\left( \boldsymbol{J}_S^{-1} \begin{pmatrix} \boldsymbol{J}_{01}\boldsymbol{\delta} \\ \boldsymbol{\pi}_S \boldsymbol{J}_{11}\boldsymbol{\delta} \end{pmatrix}, \boldsymbol{J}_S^{-1} \right). \tag{4.9}$$

In particular, the limiting distribution of the wide model's unprotected parameters is

$$\boldsymbol{D}_n = \hat{\boldsymbol{\delta}}_{\text{wide}} = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0) \xrightarrow{d} \boldsymbol{D} \sim \ N(\boldsymbol{\delta}, \boldsymbol{Q}),$$

where $\boldsymbol{Q}$ is the lower right hand corner of $\boldsymbol{J}_{\text{wide}}^{-1}$, see (4.3). That is $\boldsymbol{Q} = \boldsymbol{J}^{11}$, and can be written (using a result on the inverse of block matrices, see e.g. Harville (1997, p.100)) as

$$\boldsymbol{Q} = (\boldsymbol{J}_{11} - \boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01})^{-1}. \tag{4.10}$$

For the following, define for submodel $S$ the bottom-right corner of $\boldsymbol{J}_S^{-1}$, i.e. the asymptotic variance covariance of $\boldsymbol{\gamma}_S$, as

$$\boldsymbol{Q}_S = \boldsymbol{J}^{11,S} = (\boldsymbol{J}_{11,S} - \boldsymbol{J}_{10,S}\boldsymbol{J}_{00,S}^{-1}\boldsymbol{J}_{01,S})^{-1} = (\boldsymbol{\pi}_S \boldsymbol{Q}^{-1}\boldsymbol{\pi}_S^{\mathsf{T}})^{-1},$$

of dimension $|S| \times |S|$, and

$$\boldsymbol{G}_S = \boldsymbol{\pi}_S^{\mathsf{T}} \boldsymbol{Q}_S \boldsymbol{\pi}_S \boldsymbol{Q}^{-1}$$

of dimension $q \times q$.

Then, the limiting distribution of the focus parameter can be found by applying the delta method to the limiting distribution of the maximum likelihood estimators, (4.9). Specifically, the limiting distribution of the focus as estimated by model $S$ can be written as

$$\sqrt{n}(\hat{\mu}_S - \mu_0) \xrightarrow{d} N(\boldsymbol{\omega}^{\mathsf{T}}(\boldsymbol{I} - \boldsymbol{G}_S)\boldsymbol{\delta}, \tau_0^2 + \boldsymbol{\omega}^{\mathsf{T}}\boldsymbol{G}_S\boldsymbol{Q}\boldsymbol{G}_S^{\mathsf{T}}\boldsymbol{\omega}), \tag{4.11}$$

where $\mu_0$ is the true value of the focus,

$$\boldsymbol{\omega} = \boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}\frac{\partial \mu}{\partial \boldsymbol{\nu}} - \frac{\partial \mu}{\partial \gamma} \tag{4.12}$$

is a column vector of dimension $q$;

$$\tau_0^2 = \left(\frac{\partial \mu}{\partial \boldsymbol{\nu}}\right)^{\mathsf{T}} \boldsymbol{J}_{00}^{-1} \frac{\partial \mu}{\partial \boldsymbol{\nu}} \tag{4.13}$$

is a scalar; and where partial derivatives in both $\boldsymbol{\omega}$ and $\boldsymbol{\tau}_0^2$ are evaluated at the null point $(\boldsymbol{\nu}_0, \gamma_0)$.

The quantity $\tau_0^2$ can be thought of as the minimal variance of the limiting distribution of the focus that persists in all models under consideration. In fact, it is the variance of the limiting distribution of the focus as estimated by the narrow model. Note that $\tau_0^2$ does not depend on $\boldsymbol{\gamma}$ and so is common to all models. In contrast, $\boldsymbol{\omega}$ depends on $\boldsymbol{\gamma}$ and so is different for each model.

The narrow and wide models are special cases of (4.11) with

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{\text{narr}} - \mu_0) &\xrightarrow{d} N(\boldsymbol{\omega}^{\intercal}\boldsymbol{\delta}, \tau_0^2),\\ \sqrt{n}(\hat{\mu}_{\text{wide}} - \mu_0) &\xrightarrow{d} N(0, \tau_0^2 + \boldsymbol{\omega}^{\intercal}\boldsymbol{Q}\boldsymbol{\omega}).\end{aligned}$$

Since the wide model is assumed true, the limiting distribution of its estimator of the focus is centered at zero, i.e. the bias disappears. The penalty the wide model pays for this is increased uncertainty due to more parameters having to be estimated. This is captured by the addition of $\boldsymbol{\omega}^{\intercal}\boldsymbol{Q}\boldsymbol{\omega}$ to the narrow variance. The bias and variance of those models in-between the wide and narrow will vary according to which unprotected parameters they involve.

Since the quantities introduced here ($\boldsymbol{G}_S, \boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\tau}^2, \boldsymbol{Q}_S$) all involve inverses of sums of matrices, it is no longer worth continuing with the linear model example. Exact expressions are difficult to obtain, and not required.

### 4.1.4 FIC scores

The FIC scores are calculated by estimating the MSE of the limiting distribution of the focus for each model. The MSE of the limiting distribution of the focus, (4.11), as estimated by model $S$ is

$$\text{MSE}_S = \boldsymbol{\omega}^{\intercal}(\boldsymbol{I} - \boldsymbol{G}_S)\boldsymbol{\delta}\boldsymbol{\delta}^{\intercal}(\boldsymbol{I} - \boldsymbol{G}_S)^{\intercal}\boldsymbol{\omega} + \tau_0^2 + \boldsymbol{\omega}^{\intercal}\boldsymbol{G}_S\boldsymbol{Q}\boldsymbol{G}_S^{\intercal}\boldsymbol{\omega},$$

with $\boldsymbol{I}$ the $q \times q$ identity matrix, and where the first term is for the bias squared, and the second and third combined are the variance.

The FIC score for model $S$ is constructed by estimating each term of $\text{MSE}_S$. This is straightforward for the variance as each element is simply replaced by its consistent estimator. However, the bias squared term requires a little more thought. In particular, $\boldsymbol{\delta}\boldsymbol{\delta}^{\intercal}$ can be estimated by $\boldsymbol{D}_n\boldsymbol{D}_n^{\intercal}$, but this estimator overshoots by the amount $\boldsymbol{Q}$. To see this, note that, for any multivariate random variable $\boldsymbol{X}$, we have $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\intercal}] = \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^{\intercal} + \text{Cov}(\boldsymbol{X})$. Since the expected value and variance-covariance of the limiting distribution of $\boldsymbol{D}_n$ are $\boldsymbol{\delta}$ and $\boldsymbol{Q}$ respectively [see (4.1.3)],

$$\mathbb{E}[\boldsymbol{D}_n\boldsymbol{D}_n^{\intercal}] = \mathbb{E}[\boldsymbol{D}_n]\mathbb{E}[\boldsymbol{D}_n]^{\intercal} + \text{Cov}(\boldsymbol{D}_n) \approx \boldsymbol{\delta}\boldsymbol{\delta}^{\intercal} + \boldsymbol{Q}.$$

So, a more appropriate, asymptotically unbiased, estimator of bias squared is $\boldsymbol{D}_n\boldsymbol{D}_n^{\intercal} - \hat{\boldsymbol{Q}}$. Furthermore, since this estimator may end up being negative, not desirable for a squared term, it is truncated at zero.

The FIC score for model $S$, an asymptotically unbiased estimator of $\text{MSE}_S$ adjusted to avoid being a negative bias squared term, is then

$$\text{FIC}_S = \max\Big(0, \hat{\boldsymbol{\omega}}^\mathsf{T}(\boldsymbol{I} - \hat{\boldsymbol{G}}_S)(\boldsymbol{D}_n\boldsymbol{D}_n^\mathsf{T} - \hat{\boldsymbol{Q}})(\boldsymbol{I} - \hat{\boldsymbol{G}}_S)^\mathsf{T}\hat{\boldsymbol{\omega}}\Big) + \hat{\tau}_0^2 + \hat{\boldsymbol{\omega}}^\mathsf{T}\hat{\boldsymbol{G}}_S\hat{\boldsymbol{Q}}\hat{\boldsymbol{G}}_S^\mathsf{T}\hat{\boldsymbol{\omega}}. \tag{4.14}$$

Special cases of the above for the narrow and wide model are

$$\begin{aligned}\text{FIC}_\text{narr} &= \max\Big(0, \hat{\boldsymbol{\omega}}^\mathsf{T}(\boldsymbol{D}_n\boldsymbol{D}_n^\mathsf{T} - \hat{\boldsymbol{Q}})\hat{\boldsymbol{\omega}}\Big) + \hat{\tau}_0^2,\\ \text{FIC}_\text{wide} &= \hat{\tau}_0^2 + \hat{\boldsymbol{\omega}}^\mathsf{T}\hat{\boldsymbol{Q}}\hat{\boldsymbol{\omega}}.\end{aligned}$$

The interpretation being that the narrow model has largest bias squared term, but smallest variance in its estimation of the focus. Whereas, the widest model has no bias (since assumed as the true model) but largest variance. The rest of the models will, in varying degrees, have less bias but more variance than the narrow model, and more bias but less variance than the wide model. The model with lowest FIC score strikes the best balance in this bias versus variance trade-off in the estimation of the focus parameter.

Since both $\boldsymbol{\omega}$ and $\tau_0^2$ depend on the focus, different focus parameters will lead to different MSEs of the focus limiting distributions, and thereby different FIC scores. This implies that different models may be be ranked differently by FIC when the purpose, the focus parameter, changes.

In Claeskens et al. (2008), $\tau_0^2$ is dropped from the FIC scores since it is common to all models and therefore does not affect their relative ranking. However, it is kept in here, since the FIC scores then retain their interpretation as unbiased estimates of the MSE of the limiting distribution of the estimated focus.

Finally, note that in practice (Claeskens et al. 2008, p.154), any consistent estimator of $\boldsymbol{J}_\text{wide}$ will do; either evaluating at the narrow $(\hat{\boldsymbol{\nu}}_\text{narr}, \gamma_0)$ or the wide $(\hat{\boldsymbol{\nu}}, \hat{\gamma})$ (or any model in-between) is acceptable. Furthermore, if an expression for the expected information is difficult to obtain, or difficult to arrive at numerically, the observed information matrix may serve as an approximation. For incomplete data this becomes especially relevant, as to arrive at the expected information one has to take into consideration the missing data mechanism (e.g. Gregoire et al. 2012, p.336).

## 4.2 FIC for clustered data

As we have seen, the FIC scores are based upon the limiting distributions of the focus parameter estimators (4.11), which are derived from the limiting distribution of ML estimators (4.9), and which are themselves derived from the limiting distributions of the averaged score vector (4.8). The derivation of the limiting distributions of the averaged score vector given in Hjort & Claeskens (2003) makes use of the independence of the data points conditional on the covariates. In fact, the limiting distributions are first derived in the independent and identically distributed case (i.e.

not the regression setting), and are then said to be generalisable to allow for covariates. In the regression setting discussed above, independence of observations given the covariates is assumed. This being said, a small adjustment can be made to derive the same FIC scores under the assumption of independence between clusters, rather than independence between observations. Such an adjustment, which allows for within-cluster dependency, will now be presented.

Since choice between variance-covariance or random effect structures would lead to null parameters on the border of their parameter space (for example, since variance parameters are constrained to be positive), and therefore asymptotic normality would become an issue, the variance components will be assumed protected. For the FIC to be presented, choice is only between covariates and the focus $\mu$ is allowed to be a function of both regression parameters and variance-covariance parameters.

### 4.2.1 Framework for using FIC for data with independent clusters

Assume that the multivariate response $\boldsymbol{y}_i$ for cluster $i$ is generated by the true wide model

$$f_{\text{wide},N} = f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\nu}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{N}). \tag{4.15}$$

Here, $N$ is the number of clusters with different clusters assumed independent; $\boldsymbol{X}_i$ is the matrix of covariates for cluster $i$; $\boldsymbol{\nu} = (\boldsymbol{\theta}, \boldsymbol{\beta})$ is the vector of protected parameters of length $p$ with true value $\boldsymbol{\nu}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$, where $\boldsymbol{\theta}$, of dimension $p_1$, is the vector of variance components, and $\boldsymbol{\beta}$ are the protected regression parameters of dimension $p_2$; and $\boldsymbol{\delta}/\sqrt{N}$ is the distance of the wide model from the null value $\boldsymbol{\gamma_0}$ of the unprotected regression parameters $\boldsymbol{\gamma}$ of dimension $q$. That is, a multivariate locally misspecified framework is assumed, with the number of units within each cluster, $n$, fixed, and the sequence of models indexed by the number of clusters $N$.

Assuming sufficient smoothness of the log-density about the null point $(\boldsymbol{\nu}_0, \boldsymbol{\gamma}_0)$, define the score vector for cluster $i$ as

$$\begin{pmatrix} \boldsymbol{u}(\boldsymbol{y}_i|\boldsymbol{X}_i) \\ \boldsymbol{v}(\boldsymbol{y}_i|\boldsymbol{X}_i) \end{pmatrix} = \begin{pmatrix} \partial \log f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\nu}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\nu} \\ \partial \log f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\nu}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma} \end{pmatrix}. \tag{4.16}$$

The expected information matrix obtained from one cluster, evaluated at $(\boldsymbol{\nu}_0, \boldsymbol{\gamma}_0)$ is thus

$$\boldsymbol{J}_i = -\mathbb{E}\left[\begin{pmatrix} \frac{\partial^2 \log f(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\nu}_0,\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\nu}^\intercal} & \frac{\partial^2 \log f(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\nu}_0,\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\nu}\partial\boldsymbol{\gamma}^\intercal} \\ \frac{\partial^2 \log f(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\nu}_0,\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\nu}^\intercal} & \frac{\partial^2 \log f(\boldsymbol{y}_i|\boldsymbol{X}_i,\boldsymbol{\nu}_0,\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\intercal} \end{pmatrix}\right],$$

which under the wide model is equal to

$$\text{Var}_0 \begin{pmatrix} \partial \log f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\nu}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\nu} \\ \partial \log f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\nu}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma} \end{pmatrix}.$$

Averaging over all clusters we get

$$\boldsymbol{J}_{\text{full},N} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{J}_i = \begin{pmatrix}\boldsymbol{J}_{00,N} & \boldsymbol{J}_{01,N}\\ \boldsymbol{J}_{10,N} & \boldsymbol{J}_{11,N}\end{pmatrix}, \tag{4.17}$$

where, averaged over all clusters and evaluated at $(\boldsymbol{\nu}_0, \gamma_0)$, the upper left block, $\boldsymbol{J}_{00,N}$ of dimension $p \times p$, corresponds to the variance of the score of the protected parameters; the lower right block, $\boldsymbol{J}_{11,N}$ of dimension $q \times q$, corresponds to the variance of the score of the unprotected parameters.

Under certain conditions, $\boldsymbol{J}_{\text{full},N}$ tends, as the number of clusters grows, to

$$\boldsymbol{J}_{\text{wide}} = \begin{pmatrix}\boldsymbol{J}_{00} & \boldsymbol{J}_{01}\\ \boldsymbol{J}_{10} & \boldsymbol{J}_{11}\end{pmatrix} \tag{4.18}$$

evaluated at the narrow. Similarly, definitions of the same form as (4.3), (4.4), (4.5) and (4.6) can be given for $\boldsymbol{J}_{\text{wide}}^{-1}$, $\boldsymbol{J}_{S,N}$, $\boldsymbol{J}_S$, and $\boldsymbol{J}_S^{-1}$. In particular, using the projection matrices $\boldsymbol{\pi}_S$, $\boldsymbol{J}_S$ can be defined as in (4.5), but in terms of the matrix (4.18).

Now, define the cluster averages of the scores (4.16) (but defined for model $S$) to be $\bar{\boldsymbol{u}}_N = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}(\boldsymbol{y}_i|\boldsymbol{X}_i)$ and $\bar{\boldsymbol{v}}_{S,N} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}_S(\boldsymbol{y}_i|\boldsymbol{X}_i)$. Due to independence between clusters, the limiting distribution of these averaged scores will be of the same form as (4.8), but in terms of the Fisher information matrix (4.18). This being so, the limiting distribution of the ML estimators are of the same form as (4.9), but in terms of (4.18). And similarly with the limiting distribution of the focus parameters estimators (4.11) where, again, the relevant Fisher information matrix upon which $\tau_0^2$, $\boldsymbol{\omega}$, $\boldsymbol{Q}$ and $\boldsymbol{G}_S$ are based is now (4.18). It is important to note that the limits considered here are no longer the growth of the total number of observations, but rather the number of clusters $N$. This means that the addition of information (in terms of data) comes from the addition of more and more clusters, and the number of units within each cluster remains fixed.

The consequence of these equivalent (cluster) limiting distributions is that FIC scores for models of clustered data are the same form as that of independent data. For model $S$ we have

$$\text{FIC}_S = \max\left(0, \hat{\boldsymbol{\omega}}^{\mathsf{T}}(\boldsymbol{I} - \hat{\boldsymbol{G}}_S)(\boldsymbol{D}_N\boldsymbol{D}_N^{\mathsf{T}} - \hat{\boldsymbol{Q}})(\boldsymbol{I} - \hat{\boldsymbol{G}}_S)^{\mathsf{T}}\hat{\boldsymbol{\omega}}\right) + \hat{\tau}_0^2 + \hat{\boldsymbol{\omega}}^{\mathsf{T}}\hat{\boldsymbol{G}}_S\hat{\boldsymbol{Q}}\hat{\boldsymbol{G}}_S^{\mathsf{T}}\hat{\boldsymbol{\omega}}, \tag{4.19}$$

with the difference from (4.14) being that the Fisher information matrix to be estimated for the estimates $\hat{\tau}_0^2$, $\hat{\boldsymbol{\omega}}$, $\hat{\boldsymbol{Q}}$ is that of (4.18), and with $\boldsymbol{D}_N$ now given by $\sqrt{N}(\gamma_{\text{wide}} - \gamma_0)$.

Since the multivariate LM (2.1), the general LME model (2.3), and GLMMs (3.2) all have an expression for the joint density, and thereby expected information matrices, formula (4.19) is applicable to these classes of models for selecting between covariates for a given variance-covariance or random effect structure. Note that for LME and GLMMs, the focus of interest should indeed be a parameter and not include a random effect. Application of (4.19) to these classes of models, along with finding expressions for information matrices and data illustrations will be the focus of the rest of this chapter.

## 4.3 FIC for multivariate linear regression, with and without random effects

Consider the situation where clustered data are generated by a wide model which is multivariate Normal of the following form:

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_{p,i}\boldsymbol{\beta} + \boldsymbol{X}_{u,i}\gamma, \Sigma_i(\boldsymbol{\theta})). \tag{4.20}$$

The matrix of covariates $\boldsymbol{X}_{p,i}$ appears in all models and is therefore said to be protected, whereas $\boldsymbol{X}_{u,i}$ is a matrix of covariates that is considered unprotected: not all of its columns appear in every model. Define the vector of parameters $\boldsymbol{\xi} = (\boldsymbol{\nu}, \gamma) = (\boldsymbol{\theta}, \boldsymbol{\beta}, \gamma)$, with $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ the protected variance-covariance and regression parameters respectively; the $\gamma$ parameters are unprotected regression parameters. That is, we are working with a fixed (chosen) variance-covariance structure, but looking to choose a mean structure.

The fixed, chosen, variance-covariance structure can indeed be induced by introducing random effects. As discussed in Chapter 2, marginally the LME model can be expressed as (4.20), but where $\Sigma_i(\boldsymbol{\theta})$ takes on the specific form $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\intercal} + \sigma^2\boldsymbol{I}_{n_i}$. Furthermore, since any focus parameter must necessarily be at the marginal level (At the level of the individual, a random variable, $\boldsymbol{b}_i$, which is not a parameter, would become involved.), we will consider the marginal LME model to be simply a special case of (4.20).

As we have seen, the FIC scores require estimates of the components of the expected information matrix. Although analytical expressions for these components are not strictly necessary, they may be useful. Therefore, expressions for the information matrices of model (4.20) will now be derived.

Formulae for the components of the expected information matrix in the case of correlated data for linear regression are given in Mardia & Marshall (1984). Clustered data are a special case of the the more general correlated data, which does not enforce independence between clusters (or even define clusters as such). The results given in Mardia & Marshall (1984) are used here but adjusted for the case where the correlation is solely within clusters, and where two sets of regression parameters are defined: those that are protected and those that are unprotected. Some details omitted in Mardia & Marshall (1984) are given here in Appendix A.

The full log-likelihood for model (4.20), assuming independence between clusters, is

$$\ell_N = -\sum_{i=1}^{N} \frac{1}{2}\Big[n_i \log(2\pi) + \log|\boldsymbol{\Sigma}_i(\boldsymbol{\theta})| \\ + (\boldsymbol{y}_i - \boldsymbol{X}_{p,i}\boldsymbol{\beta} - \boldsymbol{X}_{u,i}\boldsymbol{\gamma})^{\intercal}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{X}_{p,i}\boldsymbol{\beta} - \boldsymbol{X}_{u,i}\boldsymbol{\gamma})\Big].$$

Differentiating with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ gives

$$\frac{\partial \ell_N}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{X}_{p,i}^{\intercal}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{X}_{p,i}\boldsymbol{\beta} - \boldsymbol{X}_{u,i}\boldsymbol{\gamma})$$

and

$$\frac{\partial \ell_N}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{N} \boldsymbol{X}_{u,i}^{\intercal}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{X}_{p,i}\boldsymbol{\beta} - \boldsymbol{X}_{u,i}\boldsymbol{\gamma})$$

respectively. Using Equation (A.4) of Appendix A, the $k$th element of the derivative of $\ell_N$ with respect to $\boldsymbol{\theta}$ is given by

$$\left(\frac{\partial \ell_N}{\partial \boldsymbol{\theta}}\right)_k = \sum_{i=1}^{N} -\frac{1}{2}\left[\mathrm{Tr}\left\{\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial \theta_k}\right\} + \boldsymbol{\epsilon}_i^{\intercal}\frac{\partial \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_k}\boldsymbol{\epsilon}_i\right],$$

where $\boldsymbol{\epsilon}_i = (\boldsymbol{y}_i - \boldsymbol{X}_{p,i}\boldsymbol{\beta} - \boldsymbol{X}_{u,i}\boldsymbol{\gamma})$.

The observed information matrix can be written as

$$-\frac{1}{N}\frac{\partial^2 \ell_N}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^{\intercal}} = -\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \ell_{\boldsymbol{\theta\theta}} & \ell_{\boldsymbol{\beta\theta}}^{\intercal} & \ell_{\boldsymbol{\gamma\theta}}^{\intercal} \\ \ell_{\boldsymbol{\beta\theta}} & \ell_{\boldsymbol{\beta\beta}} & \ell_{\boldsymbol{\gamma\beta}}^{\intercal} \\ \ell_{\boldsymbol{\gamma\theta}} & \ell_{\boldsymbol{\gamma\beta}} & \ell_{\boldsymbol{\gamma\gamma}} \end{pmatrix},$$

where, for example, $\ell_{\beta\beta} = \dfrac{\partial^2 \ell_N}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\intercal}}$. For which, we have

$$\begin{aligned} \ell_{\beta\beta} &= -\boldsymbol{X}_{p,i}^{\intercal}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{p,i}, \\ \ell_{\gamma\gamma} &= -\boldsymbol{X}_{u,i}^{\intercal}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{u,i}, \\ \ell_{\beta\gamma} &= -\boldsymbol{X}_{p,i}^{\intercal}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{u,i}. \end{aligned}$$

In addition, the $k$th column of $\ell_{\boldsymbol{\beta\theta}}$ and $\ell_{\boldsymbol{\gamma\theta}}$ are

$$(\ell_{\boldsymbol{\beta\theta}})_k = \boldsymbol{X}_{p,i}^{\intercal}\frac{\partial \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_k}\boldsymbol{\epsilon}_i$$

and

$$(\ell_{\boldsymbol{\gamma\theta}})_k = \boldsymbol{X}_{u,i}^{\intercal}\frac{\partial \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_k}\boldsymbol{\epsilon}_i$$

respectively, which are both mean zero under the wide model. Using the product rule under the trace operation, the $(k,l)$th element of $\ell_{\boldsymbol{\theta\theta}}$ becomes

$$\begin{aligned}(\ell_{\boldsymbol{\theta\theta}})_{k,l} =& \frac{\partial^2 \ell_N}{\partial\theta_l \partial\theta_k} \\ =& -\frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{\Sigma}_i(\boldsymbol{\theta})^{-1}\frac{\partial^2\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k} + \frac{\partial\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\right\} - \frac{1}{2}\boldsymbol{\epsilon}_i^{\mathsf{T}}\frac{\partial^2\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}\boldsymbol{\epsilon}_i.\end{aligned} \tag{4.21}$$

Similarly, the expected information matrix of the wide model can be written as

$$\boldsymbol{J}_{\text{full},N} = -\mathbb{E}\left[\frac{1}{N}\frac{\partial^2\ell_N}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathsf{T}}}\right] = \begin{pmatrix}\boldsymbol{J}_{\boldsymbol{\theta\theta},N} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{J}_{\boldsymbol{\beta\beta},N} & \boldsymbol{J}^{\mathsf{T}}_{\boldsymbol{\gamma\beta},N} \\ \mathbf{0} & \boldsymbol{J}_{\boldsymbol{\gamma\beta},N} & \boldsymbol{J}_{\boldsymbol{\gamma\gamma},N}\end{pmatrix}, \tag{4.22}$$

where

$$\begin{aligned}\boldsymbol{J}_{\boldsymbol{\beta\beta},N} =& \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_{p,i}^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{p,i}, \\ \boldsymbol{J}_{\boldsymbol{\gamma\beta},N} =& \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_{u,i}^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{p,i}, \\ \boldsymbol{J}_{\boldsymbol{\gamma\gamma},N} =& \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_{u,i}^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_{u,i},\end{aligned} \tag{4.23}$$

and the $(k,l)$th element of $\boldsymbol{J}_{\boldsymbol{\theta\theta},N}$ is

$$(\boldsymbol{J}_{\boldsymbol{\theta\theta},N})_{k,l} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}\right\}. \tag{4.24}$$

Please see (A.3) in the Appendix for full details as to how (4.24) is arrived at. The block diagonality of (4.22) is thus apparent, and under mild regularity conditions $\boldsymbol{J}_{\text{full},N}$, (4.22), tends to the limit

$$\boldsymbol{J}_{\text{wide}} = \begin{pmatrix}\boldsymbol{J}_{\boldsymbol{\theta\theta}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{J}_{\beta\beta} & \boldsymbol{J}_{\beta\gamma} \\ \mathbf{0} & \boldsymbol{J}_{\gamma\beta} & \boldsymbol{J}_{\gamma\gamma}\end{pmatrix}, \tag{4.25}$$

as the number of clusters increases. Relating this back to the notation of (4.18), we have

$$\begin{aligned}\boldsymbol{J}_{00} =& \begin{pmatrix}\boldsymbol{J}_{\theta\theta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{J}_{\beta\beta}\end{pmatrix}, \\ \boldsymbol{J}_{10} =& \begin{pmatrix}\mathbf{0} & \boldsymbol{J}_{\gamma\beta}\end{pmatrix},\end{aligned} \tag{4.26}$$

and

$$\boldsymbol{J}_{11} = \boldsymbol{J}_{\gamma\gamma}.$$

Lastly, note that since these expressions do not depend on the regression parameters, evaluating at the narrow point makes no difference.

**The case of the marginal LME model**

Since the marginal variance-covariance takes the form $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\intercal + \sigma^2\boldsymbol{I}_{n_i}$ for the marginal LME model [see (2.5)], the variance-covariance component of its expected information matrix also has a particular form.

For the case of a general random effect variance-covariance matrix, $\boldsymbol{D}$, a compact expression (making use of elimination matrices, and the $\text{vec}(\cdot)$ operator) is given in Demidenko (2013, p.124). For the special case of constant variance and independence between random effect components, that is, when $\boldsymbol{\Sigma}_i = \sigma_b^2\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal + \sigma^2\boldsymbol{I}_{n_i}$, and $\boldsymbol{D} = \sigma_b^2\boldsymbol{I}_k$ is diagonal (e.g. independence between random intercepts and slopes), we have that, by (4.24), $\boldsymbol{J}_{\boldsymbol{\theta\theta}}$ is of the form

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\begin{pmatrix} \text{Tr}\left\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^{-1}\right\} & \text{Tr}\left\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i\boldsymbol{Z}_i^\intercal\boldsymbol{\Sigma}_i^{-1}\right\} \\ \text{Tr}\left\{\boldsymbol{Z}_i^\intercal\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i\right\} & \text{Tr}\left\{\boldsymbol{Z}_i^\intercal\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i(\boldsymbol{Z}_i^\intercal\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i)^\intercal\right\} \end{pmatrix},$$

where the upper left corner corresponds to the within-cluster error variance $\sigma^2$, the lower right corner corresponds to the variance of the random effects $\sigma_b^2$, and the off-diagonals correspond to the cross-terms. This corresponds to the lower right entry of Equation (6.62) in McCulloch & Searle (2001), but where there are only two levels of the data-hierarchy and hence only one vector of random effects.

### 4.3.1 The focus as solely a function of regression parameters

Consider the limiting expected information matrix of the Normal model, partitioned into blocks as in (4.25). In general, for discrete GLMMs (e.g. Binary, Poisson) the block diagonal nature does not hold. So the results in this section only (with some exceptions) apply to the Normal model.

When the focus parameter, $\mu$, is not a function of the variance-covariance parameters, the quantities $\tau_0^2$, $\boldsymbol{\omega}$ and $\boldsymbol{Q}$ drop their dependency upon the component of the expected information matrix corresponding to the variance-covariance parameters, $\boldsymbol{J}_{\boldsymbol{\theta\theta}}$. For instance, since in this case,

$$\frac{\partial\mu}{\partial\boldsymbol{\nu}} = \begin{pmatrix} \frac{\partial\mu}{\partial\boldsymbol{\theta}} \\ \frac{\partial\mu}{\partial\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \frac{\partial\mu}{\partial\boldsymbol{\beta}} \end{pmatrix},$$

we have that

$$\begin{aligned} \tau_0^2 &= \begin{pmatrix} \boldsymbol{0} & \left(\frac{\partial\mu}{\partial\boldsymbol{\beta}}\right)^\intercal \end{pmatrix} \begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\theta\theta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{\boldsymbol{\beta\beta}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{0} \\ \frac{\partial\mu}{\partial\boldsymbol{\beta}} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{0} & \left(\frac{\partial\mu}{\partial\boldsymbol{\beta}}\right)^\intercal \end{pmatrix} \begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\theta\theta}}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{\boldsymbol{\beta\beta}}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0} \\ \frac{\partial\mu}{\partial\boldsymbol{\beta}} \end{pmatrix} \\ &= \left(\frac{\partial\mu}{\partial\boldsymbol{\beta}}\right)^\intercal \boldsymbol{J}_{\boldsymbol{\beta\beta}}^{-1}\frac{\partial\mu}{\partial\boldsymbol{\beta}}. \end{aligned} \tag{4.27}$$

In a similar manner, the expressions for $\boldsymbol{\omega}$ and $\boldsymbol{Q}$ reduce to

$$\boldsymbol{\omega} = \boldsymbol{J}_{\gamma\beta}\boldsymbol{J}_{\beta\beta}^{-1}\frac{\partial\mu}{\partial\boldsymbol{\beta}} - \frac{\partial\mu}{\partial\gamma}, \tag{4.28}$$

and

$$\boldsymbol{Q} = (\boldsymbol{J}_{\gamma\gamma} - \boldsymbol{J}_{\gamma\beta}\boldsymbol{J}_{\beta\beta}^{-1}\boldsymbol{J}_{\gamma\beta}^{\intercal})^{-1}. \tag{4.29}$$

This result makes the FIC scores slightly simpler to arrive at for a focus that is purely a function of the regression parameters (using the explicit formula for $\boldsymbol{J}_{\boldsymbol{\theta\theta}}$, (4.24) can be a cumbersome activity). Furthermore, since $\boldsymbol{J}_{\boldsymbol{\theta\theta}}^{-1}$ is the covariance matrix of the limiting distribution of the variance-covariance parameters, increasing the number of $\theta$s in the model does not, in theory, increase the uncertainty, nor affect the bias of the estimator of a focus parameter that is only in terms of the regression parameters.

This means that, within an asymptotic framework and prior to estimation, provided the focus is a function of the $\beta$s alone, there may be little to gain in terms of mean squared error by attempting to restrict the number of variance-covariance parameters in the model: the uncertainty in the $\theta$s should not influence the uncertainty of the $\beta$s. Asymptotically, both sets of parameters draw their information from independent sources, albeit the same data. Thus, in terms of choice of variance-covariance matrix when the focus is on the regression parameters, there may be little advantage to a simpler model.

Since the expressions (4.23) have to be estimated (usually by plugging in the inverse of the estimated variance-covariance matrix), this could introduce some uncertainty into the focus that is purely a function of the $\beta$s. To see to what extent the above conclusion holds in practice (after estimation), a simulation study was considered. If this result does indeed hold true, there could be implications for how to go about choosing a variance-covariance matrix when the interest is only in the regression parameters.

### 4.3.2 Simulations

For foci that are only a function of the regression parameters, does the choice of covariance model (or random effect structure) affect the stability of the FIC scores? Instability in the FIC scores of a correct, but overly-specified variance-covariance model would indicate uncertainty in the $\theta$s influencing either (or both) the variance or bias of the focus estimators.

Thus, the specific question to be addressed is: given two correctly specified, covariance matrices $\Sigma_1$ and $\Sigma_2$, but where $\Sigma_2$ includes more parameters than necessary, is there any difference in terms of variability of FIC scores when the focus is only a function of $\boldsymbol{\beta}$? The answer is likely to depend on the number of clusters; the number of units within clusters; whether the dataset is balanced by design; whether there is missing data; the mean structures to be selected between; the particular focus of interest; and the different covariance structures in question. Addressing all of these variables at once is a tall order. A simplified scenario was thus constructed.

One hundred balanced longitudinal datasets were generated for each pair of $(N, n)$: the number of subjects $N$ taking values in the set {30, 50, 100, 200}, and the number of measurement occasions $n$ taking values {3, 4, 5, 6}. For each situation $(N, n)$, the true model was a LME model including a random intercept and a random slope for the effect of time. In particular, the true marginal model was

$$y_{ij} = \beta_0 + \beta_1 \text{group}_i + \beta_2 t_{ij} + \gamma_0 t_{ij}^2 + \gamma_1 t_{ij} \text{group}_i + \epsilon_{ij}, \tag{4.30}$$

with the true variance-covariance of the errors given by

$$\text{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^\intercal + \sigma^2 \boldsymbol{I}_n, \tag{4.31}$$

where $\boldsymbol{Z}_i$ is the true random effects design matrix, and $\boldsymbol{D}$ the variance-covariance matrix of the random effects, which are given by

$$\boldsymbol{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in} \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}.$$

The binary covariate group was generated using the command *rbinom* in R with probability 0.5. Covariate time was treated as continuous over the interval $[0, t^*]$, with $t^* = 12$, and centered about its mean, $(t = 6)$. For each situation of $(N, n)$, measurements were simulated for both endpoints, $t = 0, t = t^*$, and additional measurements divided the interval $[0, t^*]$ into equal length sub-intervals.

The true values of the parameters were set as $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (1, 1, 1), \boldsymbol{\gamma} = (\gamma_0, \gamma_1) = (-0.1, 1)$ which means that the true mean trend was increasing and slightly concave. And, in addition $\boldsymbol{\theta} = (\sigma^2, d_{11}, d_{12}, d_{22})^\intercal = (4, 4, 0.2, 1)$.

Four different mean structures were considered for each simulation and for each situation of $(N, n)$. The widest mean structure considered, M4, was

$$\mathbb{E}[y_{ij}] = \beta_0 + \beta_1 \text{group}_i + \beta_2 t_{ij} + \gamma_0 t_{ij}^2 + \gamma_1 t_{ij} \text{group}_i + \gamma_2 t_{ij}^2 \text{group}_i; \tag{4.32}$$

the narrowest model, M1, included only the protected covariates. The true model, M3, (4.30), was among the candidate models, and finally, the mean structure M2

$$\mathbb{E}[y_{ij}] = \beta_0 + \beta_1 \text{group}_i + \beta_2 t_{ij} + \gamma_2 t_{ij}^2 \text{group}_i, \tag{4.33}$$

which is the narrow model plus one interaction: the uninformative interaction between group and $t^2$, was also included.

The focus parameter of interest in every situation was the expected response of an individual belonging to the non-reference group at the last measurement occasion $n$, when $t = t^*$, that is,

$$\mu = \mathbb{E}[y_{in} | t = t^*, \text{group} = 1].$$

The covariance models under consideration were:

- $\mathbf{\Sigma}_0$, the marginal covariance matrix arising from a random intercept model, i.e. compound symmetric, see (2.2), with two parameters regardless of the number of measurement occasions: the variance of the random intercept and the variance of the errors.

- $\mathbf{\Sigma}_1$, the marginal covariance matrix arising from a random intercept and slope model, i.e. the true data generating model, with four unknown parameters $\boldsymbol{\theta} = (\sigma^2, d_{11}, d_{12}, d_{22})^\intercal$ regardless of the number of measurement occasions).

- $\mathbf{\Sigma}_2$, an unstructured covariance model with $n(n+1)/2$ parameters, ranging from 6 $(n = 3)$ to 21 $(n = 6)$ parameters.

- $\mathbf{\Sigma}_3$, the true covariance model with known parameter values $\boldsymbol{\theta} = (4, 4, 0.2, 1)^\intercal$.

The variance-covariance matrix $\mathbf{\Sigma}_0$ is thus misspecified, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_3$ are both true, but with unknown and known parameters respectively, and $\mathbf{\Sigma}_2$ is correctly specified, but involves too many parameters.

Given enough data, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are expected to converge to the same true covariance model. Since $\mathbf{\Sigma}_2$ involves more parameters (especially as $n$ grows), what are the consequences for the variability in the bias and variance of the focus?

### Results

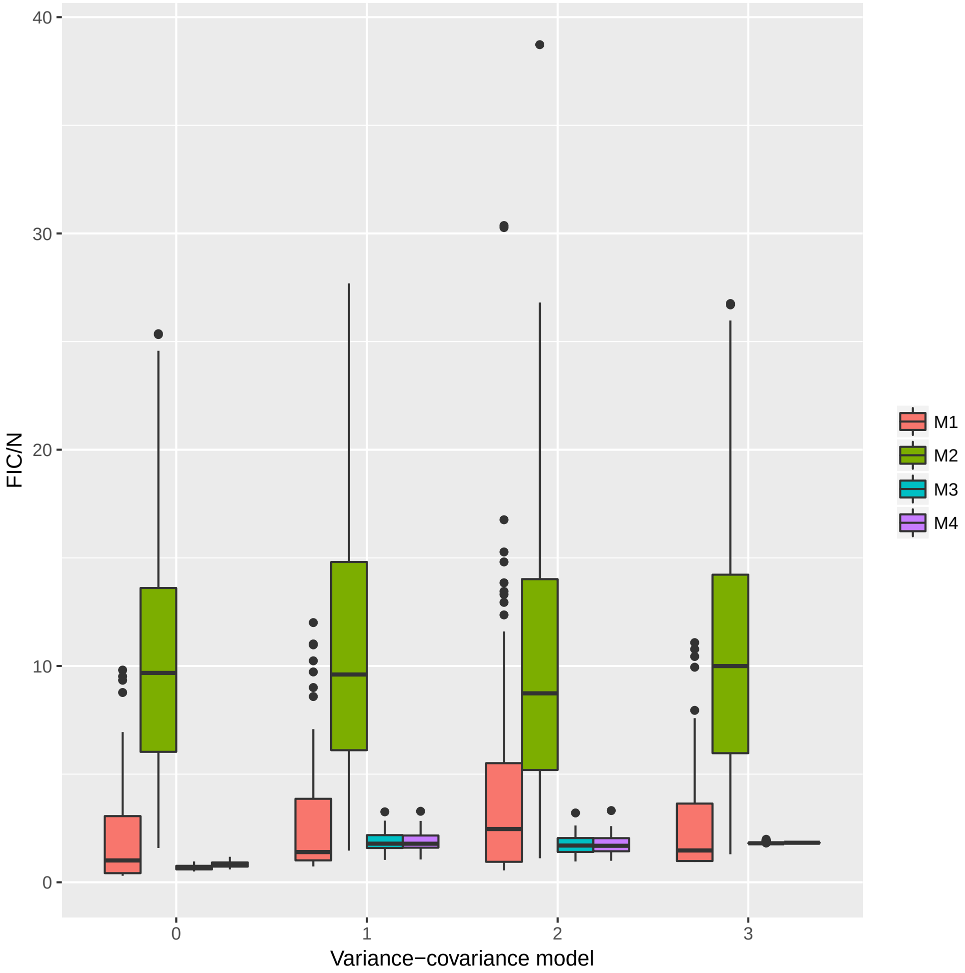
The package **simsalapar** (Hofert & Mächler 2016) was used to help run parallel simulations. Figures were produced using the package **ggplot2** (Wickham 2009). The random intercept model, $\mathbf{\Sigma}_0$, failed to converge on 5 out of the 100 simulations, but there were no other complications. Tables B.1 and B.2 in Section B.2 of Appendix B give the arithmetic mean over all 100 simulations of the FIC scores divided by $N$ (scaled to be interpretable as estimates of the MSEs) for each situation $(N, n, \text{M}, \mathbf{\Sigma})$.

In general, M3 and M4, were preferred over M1 and M2, as illustrated in Figure 4.1, which shows box plots (of all 100 simulations) of the FIC scores scaled down by a factor of $N$ for the particular situation $N = 50$ and $n = 6$, and with covariance matrix indexed along the x-axis. It is also visible that, for this case (and in fact more generally), that M3 and M4 displayed much less uncertainty in the FIC scores.

From inspection of the plots of FIC$/N$ for each situation $(N, n, \text{M}, \mathbf{\Sigma})$, it is clear that two cases arise: the cases of correctly and incorrectly specified mean structures need to be distinguished.

When the mean structure is misspecified, i.e. for M1 and M2, there is more variability (particularly for M1) in the FIC scores of the variance-covariance matrix $\mathbf{\Sigma}_2$ relative to that of $\mathbf{\Sigma}_1$. Figure 4.2 shows a grid of box plots in $4 \times 4 = 16$ cells. Each column of the grid represents a specific size for $n$, and each row a specific $N$. In each individual cell, FIC$/N$ for mean structure M1 is on the y-axis; the values of the x-axis $(0, 1, 2, 3)$ correspond to the indices of the different variance-covariance matrices. The red crosses mark the approximately true mean squared errors. That is,

Figure 4.1: Box plots of FIC$/N$ for the situation $N = 50$, $n = 6$ plotted for each covariance structure $\mathbf{\Sigma}$ (whose indices correspond with the values (0, 1, 2, 3) on the x-axis) and for each mean structure (M1, M2, M3 and M4) represented by different colours.



the red crosses show the sample means (over all 100 simulations) of the 'observed' squared errors: $(\hat{\mu}_j - \mu_{\text{true}})^2$. The box plots show more variability in the FIC scores (divided by $N$) of $\mathbf{\Sigma}_2$ consistently throughout the grid. The same set-up is given for mean structure M2 in Figure B.1 in Appendix B, where the differences still exist, but are much less severe.

Figures B.3 and B.4 along with Figures B.7 and B.8 in Appendix B, using the same set-up as described above, show box plots of the estimated bias squared and estimated variances of models M1 and M2. The red crosses mark the approximately true bias squareds $\left(\frac{1}{100}\sum_{j=1}^{100}(\hat{\mu}_j - \mu_{\text{true}})\right)^2$, and the sample variances in focus estimates over all 100 simulations for the bias squared (B.3 and B.4) and variance plots (B.7 and B.8) respectively. Inspection of these plots shows that the difference in spreads of FIC$/N$ between $\mathbf{\Sigma}_1$ compared with $\mathbf{\Sigma}_2$ is due to differences in bias squared, not variance (although the variances of $\mathbf{\Sigma}_2$ seem to be underestimated for mean structure M1).

The same set-up as Figure 4.2 is displayed in Figure 4.3 for the FIC$/N$ scores of correctly specified mean structure M3 (and for M4 in Figure B.2 in Appendix B). For M3 and M4, it is visible that there is very little difference in terms of variability of the FIC scores (divided by $N$) between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

Analysis of the relevant box plots (Figures B.9 and B.10) shows that for M3 and M4 the estimates of the variance term were similar for both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, in terms of spread. There is more variability in the bias squared estimates belonging to $\mathbf{\Sigma}_2$ for M3 (see Figure B.5), though since M3 is correctly specified, these differences do not contribute much to the FIC scores; the variance term dominates. Since M4 is the mean structure assumed correct by the FIC procedure, its bias squared estimates are effectively zero (see Figure B.6).

The conclusion is then that, provided the mean structure is correctly specified, the FIC scores will not suffer if too many parameters are used to correctly specify the variance-covariance model. For the correctly specified mean structures, the variance term tends to dominate, which appears to be less afflicted by lack of simplicity. However, the bias squared terms of misspecified models appear to show more spread in the estimates of over-parameterised variance-covariance models.

In general, the true specification of the mean structure is not known, so it is of interest to include potentially biased (misspecified) mean structures among the candidate models. Therefore, foci that are purely a function of the $\beta$s do require parsimonious variance-covariance models. This agrees then with the Goldilocks principle of traditional methods of choosing a variance-covariance model by balancing between fit and simplicity.

As a comment, note that the FIC$/N$ estimates of the misspecified $\mathbf{\Sigma}_0$ appear to be underestimated. This seems to be more in connection with the variance estimates rather than the bias squareds. For example, see Figure B.9. In my view, this is most likely attributable to the fact that the FIC procedure used model-based inference: $\mathbf{\Sigma}_0$ enforces compound-symmetry - an unrealistic structure. Model-robust inference

Figure 4.2: In this figure showing a grid of box plots in $4 \times 4 = 16$ cells, each column represents a specific size of $n$, and each row a specific $N$. In each individual cell, FIC$/N$ for the narrow mean structure M1 is plotted on the y-axis; the values of the x-axis $(0, 1, 2, 3)$ correspond to the indices of the different variance-covariance matrices, $\Sigma$.

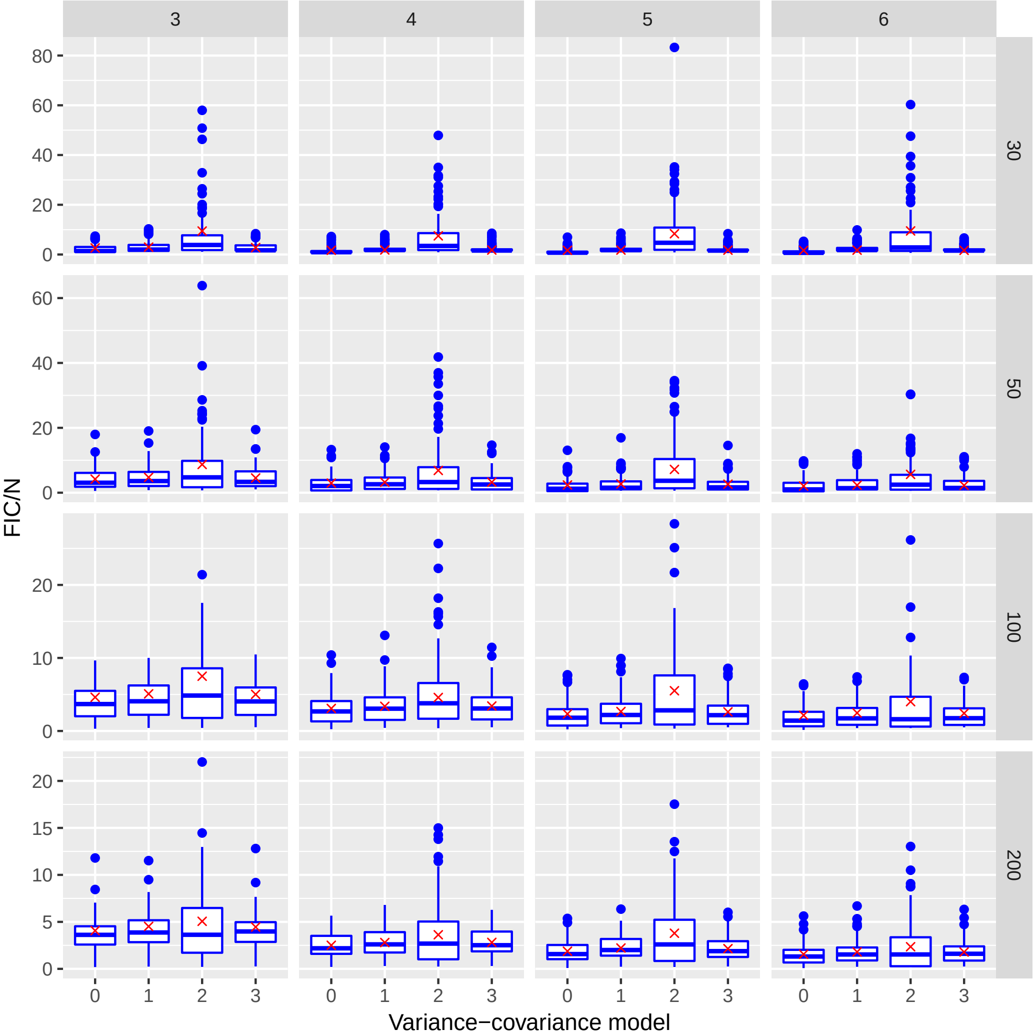
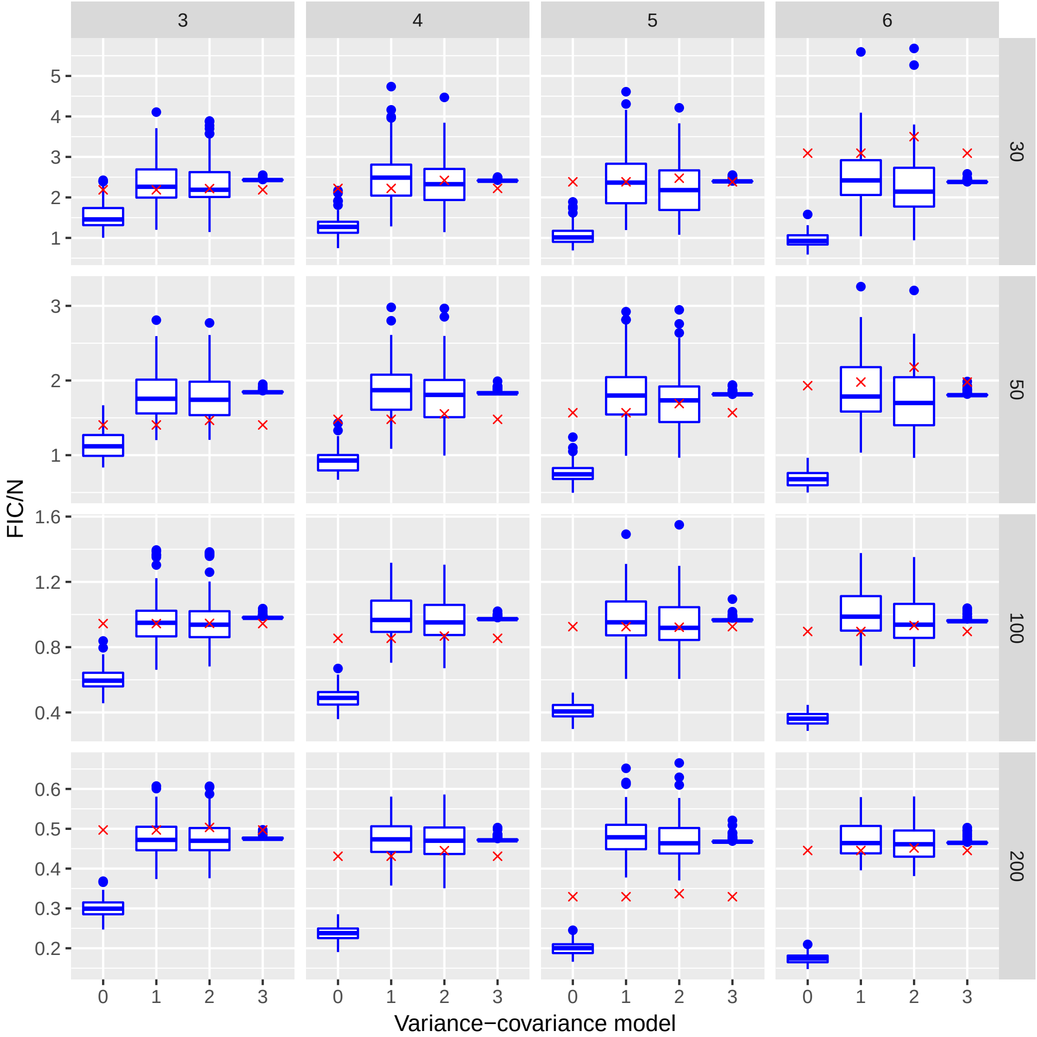


Figure 4.3: In this figure showing a grid of box plots in $4 \times 4 = 16$ cells, each column represents a specific size of $n$, and each row a specific $N$. In each individual cell, the y-axis gives the values of FIC$/N$ for the true mean structure M3; the values of the x-axis $(0, 1, 2, 3)$ correspond to the indices of the different variance-covariance matrices, $\mathbf{\Sigma}$.



would produce valid estimates for such misspecified covariance models.

Lastly, it is worth mentioning that the simulations were repeated but with less variable intercepts and slopes, $d_{11} = 2$, $d_{22} = 0.2$. The same conclusion was arrived at: the relative differences in variability of FIC scores between mean structures and covariance structures remained, only the scale of the variability had reduced.

### 4.3.3 Data illustration

The FIC, (4.19), will be used as a covariate selector for the data set introduced in Section 2.7, where 46 depressed patients are followed over a period of 5 weeks, and their HAMD depression scores are recorded. Two different foci will be considered. But, first of all, a variance-covariance model will be selected via maximising the REML based AIC.

#### Step 1: choosing a covariance model

A range of covariance models were considered with a maximal mean structure. Since a misspecified mean structure can result in the attempt to model covariance which is not truly there, imposing no assumptions on the maximal mean trend is best (e.g. Fitzmaurice et al. 2004*a*, p.173). And with balanced data, time was able to be treated as discrete (categorical), which forces no structure on the mean response. The maximal mean trend was thus taken as an interaction between discrete time and depression type, including the main effects of both.

The candidate covariance models include: M1, an unstructured variance-covariance matrix, the same for both groups; M2, an unstructured variance-covariance matrix but with separate sets of variance parameters for both groups;[2] M3, compound symmetric with the same unstructured variances for both groups; M4, M5 and M6 have exponentially decaying correlations and exponentially growing variances, the difference being that M5 and M6 allow different rates of exponential growth of variances for each group, and, in addition, only M6 includes a nugget effect; M7 is the covariance matrix arising marginally from a random intercept and slope model. Its fixed effect part was set as the same maximal mean structure so that, marginally, REML based AIC comparison was possible.

The results are presented in Table 4.1, where the column $p_1$ gives the number of variance-covariance parameters and $\hat{\ell}_{REML}$ the maximum restricted log-likelihood. Model M6 was deemed best by AIC.

[2] This model assumes that the correlations are the same for both groups. If there was more data, and R facilitated such an option, I would also specify separate sets of parameters for the unstructured correlations of both groups and check the AIC.

Table 4.1

| Model | $p_1$ | AIC | $\hat{\ell}_{REML}$ |
|---|---|---|---|
| M1 | 15 | -1367.84 | -658.92 |
| M2 | 20 | -1370.03 | -655.01 |
| M3 | 6 | -1475.97 | -721.98 |
| M4 | 3 | -1360.92 | -667.46 |
| M5 | 4 | -1360.28 | -666.14 |
| M6 | 5 | -1357.24 | -663.62 |
| M7 | 4 | -1359.22 | -665.61 |

### Step 2: covariate selection

### The focus as solely a function of regression parameters

Suppose that the focus is the expected response at the end of the study (the end of treatment week 4) of a non-endogeneous patient with an above average baseline response of 27 on the HAMD scale. That is,

$$\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbb{E}[hd_{i4}|base_i = 27, t_{i4} = 4, ed_i = 0]. \tag{4.34}$$

All models now under consideration were a subset of the wide model with the following mean structure

$$\mathbb{E}[hd_{ij}] = \beta_0 + \beta_1 basec_i + \beta_2 tc_{ij} + \gamma_0 ed_i + \gamma_1 tc_{ij}^2 + \gamma_2 tc_{ij} ed_i + \gamma_3 tc_{ij}^2 ed_i \tag{4.35}$$

for individual $i$ at time $j$, where the response is HAMD score, $hd_{ij}$; covariate $basec_i$ is the baseline score centered about the sample mean; $tc_{ij}$ is time, but centered around the middle of the treatment period; and $ed_i$ is an indicator variable for depression type, with non-endogeneous patients as the reference group; and where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ signify protected and unprotected regression coefficients respectively. Altogether, $2^4 = 16$ submodels were considered, corresponding to the different possible combinations of unprotected regression coefficients.

For the wide model, the matrix of protected covariates is

$$\boldsymbol{X}_{p,i} = \begin{pmatrix} 1 & basec_i & tc_{i0} \\ \vdots & \vdots & \vdots \\ 1 & basec_i & tc_{i4} \end{pmatrix}, \tag{4.36}$$

and the matrix of unprotected covariates is

$$\boldsymbol{X}_{u,i} = \begin{pmatrix} ed_i & tc_{i0} ed_i & tc_{i0}^2 & tc_{i0}^2 ed_i \\ \vdots & \vdots & \vdots & \vdots \\ ed_i & tc_{i4} ed_i & tc_{i4}^2 & tc_{i4}^2 ed_i \end{pmatrix}. \tag{4.37}$$

The covariance matrix (pattern M6) for endogenous, $\boldsymbol{\Sigma}_1$, and non-endogenous patients, $\boldsymbol{\Sigma}_2$, as estimated using REML by the wide model are

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 21.82 & 15.65 & 14.12 & 12.75 & 11.50 \\ 15.65 & 28.27 & 20.27 & 18.29 & 16.51 \\ 14.12 & 20.27 & 36.61 & 26.26 & 23.70 \\ 12.75 & 18.29 & 26.26 & 47.42 & 34.01 \\ 11.50 & 16.51 & 23.70 & 34.01 & 61.42 \end{pmatrix}$$

and

$$\hat{\boldsymbol{\Sigma}}_0 = \begin{pmatrix} 19.76 & 13.48 & 11.58 & 9.94 & 8.54 \\ 13.48 & 23.17 & 15.81 & 13.58 & 11.66 \\ 11.58 & 15.81 & 27.17 & 18.55 & 15.93 \\ 9.94 & 13.58 & 18.55 & 31.87 & 21.75 \\ 8.54 & 11.66 & 15.93 & 21.75 & 37.38 \end{pmatrix}$$

respectively. That the variances are different for each group is clear. When the covariances are standardised into correlations one can see that they are in fact the same for both groups. These matrices give estimates for the required components (4.23) of the expected information matrix of the wide model. These, along with partial derivatives of the focus evaluated at the null model, give estimates of the quantities (4.27), (4.28) and (4.29), which, in turn, combined with $\boldsymbol{D}_N$ and the relevant projection matrices, $\boldsymbol{\pi}_S$, give the FIC scores (4.19).

The results from the analysis of this focus, (4.34), are shown in Table 4.2. The first column states which unprotected covariates have been included. For example, an entry of 1100 corresponds to including $ed_i$ and $tc^2$ but excluding the interaction terms. Column $|S|$ is the number of unprotected parameters included in the model; column $\hat{\mu}$ contains the estimate of the focus parameter for each model; column bias$^2$ contains the estimates of the squared bias of the limiting distribution of the focus for each model; Var is the estimated variance of the limiting distribution of the focus for each model; the column AIC gives the (ML based) AIC scores of all models; rAIC and rFIC rank the AIC and FIC scores from best (=1) to worst (=16) respectively.

Model 1000 is the FIC favourite for this focus. With smallest FIC score its estimate appears furthest to the left on part (a) of Figure 4.4. That is, the model including only the main effect for group, $ed_i$, as an additional term produced the best estimate for this particular focus as judged by FIC. Table 4.2 shows that there is disagreement between AIC and FIC in terms of covariate selection. Part (a) of Figure 4.4, displaying two groups distinguished by similar sized confidence intervals and similar point estimates, shows that FIC is happy to accept some bias in exchange for smaller confidence intervals here.

When the focus is changed to the expected response at the end of treatment for an endogeneous patient with a below average baseline score of 18, the narrow model

Table 4.2: FIC results for the focus as an expected response at the end of the study for a non-endogeneous depressed patient with an above average baseline score of 27 on the HAMD.

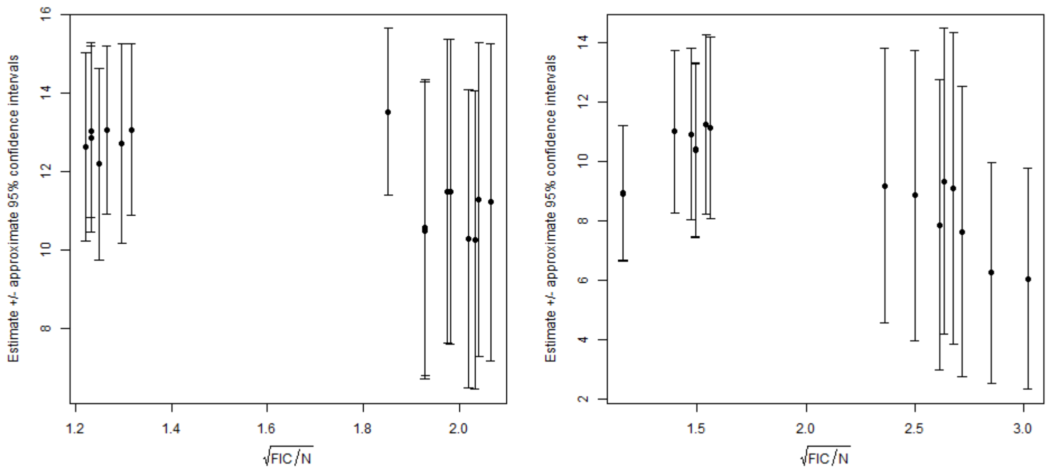
| Model | $\|S\|$ | $\hat{\mu}$ | bias$^2$ | Var | FIC | $\sqrt{\text{FIC}/N}$ | rFIC | AIC | rAIC |
|---|---|---|---|---|---|---|---|---|---|
| 1000 | 1 | 12.64 | 0 | 68.58 | 68.58 | 1.22 | 1 | -1363.04 | 14 |
| 0101 | 2 | 13.03 | 12.68 | 57.1 | 69.78 | 1.23 | 2 | -1360.09 | 4 |
| 1001 | 2 | 12.87 | 0.66 | 69.16 | 69.82 | 1.23 | 3 | -1360 | 3 |
| 1100 | 2 | 12.2 | 0 | 71.77 | 71.77 | 1.25 | 4 | -1362.31 | 10 |
| 0001 | 1 | 13.07 | 18.5 | 55.12 | 73.62 | 1.27 | 5 | -1358.14 | 1 |
| 1101 | 3 | 12.73 | 0 | 77.31 | 77.31 | 1.3 | 6 | -1361.87 | 8 |
| 0100 | 1 | 13.07 | 22.55 | 57.07 | 79.62 | 1.32 | 7 | -1362.8 | 12 |
| 0000 | 0 | 13.54 | 104.02 | 53.86 | 157.88 | 1.85 | 8 | -1365.77 | 16 |
| 0010 | 1 | 10.5 | 0 | 171.09 | 171.09 | 1.93 | 9 | -1362.09 | 9 |
| 1010 | 2 | 10.58 | 0 | 171.14 | 171.14 | 1.93 | 10 | -1363.14 | 15 |
| 0011 | 2 | 11.5 | 0 | 179.59 | 179.59 | 1.98 | 11 | -1359.19 | 2 |
| 1011 | 3 | 11.49 | 0 | 180.55 | 180.55 | 1.98 | 12 | -1361.17 | 6 |
| 1110 | 3 | 10.3 | 15 | 172.66 | 187.66 | 2.02 | 13 | -1362.63 | 11 |
| 0110 | 2 | 10.26 | 17.51 | 172.62 | 190.12 | 2.03 | 14 | -1361.66 | 7 |
| 0111 | 3 | 11.3 | 0 | 191.63 | 191.63 | 2.04 | 15 | -1361.03 | 5 |
| 1111 | 4 | 11.22 | 0 | 196.31 | 196.31 | 2.07 | 16 | -1362.96 | 13 |

is judged best by FIC. Table 4.3 shows the output for this focus.[3] Stark contrasts between AIC and FIC are again evident. The estimates of the focus for each model are plotted versus the square root of the FIC scores, scaled by a factor of $1/\sqrt{N}$ in part (b) of Figure 4.4. This particular transformation of the FIC scores has the interpretation of estimated root MSE, which is on the same scale as the standard errors of the focus estimates.

It should be noted that selecting only one of the models would result in a reduction in confidence of the confidence intervals of Figure 4.4. This is because simply choosing the best estimate would not be acknowledging the uncertainty involved in the model selection procedure. Taking the best model, as judged by FIC (or AIC for that matter), would be acting as if the model was known to be the best at the outset, which is not so; the uncertainty of the model selection procedure must be taken into account (Claeskens et al. 2008). Since the model selection was performed in two steps here, the uncertainty should be acknowledged as such. To my knowledge, there are no frequentist model averaging schemes which account for two model selection steps, though smoothed weights as in Section 7.2 of Claeskens et al. (2008) could well be explored for this purpose.

R code for this section is given in Appendix D. In particular, preparation of the

[3] The bias squared terms are mostly truncated to zero, in spite of varying estimates. This is because the correction for over-shooting produces a negative bias squared.



(a)

(b)

Figure 4.4: Figure (a) displays the estimated response at the end of treatment for a non-endogeneous patient of baseline score 27 along with 95% confidence intervals. Figure (b) displays the estimated response at the end of treatment for an endogeneous patient of baseline score 18 along with 95% confidence intervals.

depression dataset is given in Listing D.1, and code for the FIC procedure is given in Listing D.2.

### The focus now allowed to depend on the variance-covariance parameters

For the HAMD rating scale, Zimmerman et al. (2013) suggest severity ranges of $\leq 7$ for no depression, $(8 - 16)$ for mild depression, $(17 - 23)$ for moderate depression, and $\geq 24$ for severe depression.

Suppose that we now wish to rank models in terms of estimation of the probability of endogeneous patients, who have baseline HAMD scores ranging from $(14, 15, ..., 34)$, being depression free at the end of the study.[4] That is, rather than looking at a single focus determined by a single covariate value, as in the expectations of the previous section, a range of foci will be considered, each probabilities determined by a different baseline covariate value. The FIC scores will be calculated for each model and for each focus. The model with lowest FIC averaged over all foci is deemed the winner. In other words, the averaged FIC (AFIC) scheme of (Claeskens et al. 2008, Section 6.9) will be carried out with equal weighting for each focus.

Recall that the AIC favoured variance-covariance model was M6 (see Table 4.1). This, for occasions $t_j$ and $t_k$ in $\{1, 2, 3, 4, 5\}$ and for an endogeneously depressed

[4] I do not wish to extend too far beyond the sample baseline scores which range from 15 to 33.

Table 4.3: FIC results for the focus as the expected response at the end of the study of an endogeneous depressed patient with (a below average) baseline score of 18 on the HAMD.

| Model | $\|S\|$ | $\hat{\mu}$ | bias$^2$ | Var | FIC | $\sqrt{\text{FIC}/N}$ | rFIC | AIC | rAIC |
|---|---|---|---|---|---|---|---|---|---|
| 0000 | 0 | 8.94 | 0 | 61.72 | 61.72 | 1.16 | 1 | -1365.77 | 16 |
| 0100 | 1 | 8.96 | 0 | 61.76 | 61.76 | 1.16 | 2 | -1362.8 | 12 |
| 0001 | 1 | 11.02 | 0 | 89.67 | 89.67 | 1.4 | 3 | -1358.14 | 1 |
| 0101 | 2 | 10.93 | 0 | 99.62 | 99.62 | 1.47 | 4 | -1360.09 | 4 |
| 1000 | 1 | 10.38 | 0 | 102.29 | 102.29 | 1.49 | 5 | -1363.04 | 14 |
| 1100 | 2 | 10.42 | 0 | 102.33 | 102.33 | 1.49 | 6 | -1362.31 | 10 |
| 1001 | 2 | 11.25 | 0 | 109.07 | 109.07 | 1.54 | 7 | -1360 | 3 |
| 1101 | 3 | 11.16 | 0 | 112.16 | 112.16 | 1.56 | 8 | -1361.87 | 8 |
| 0011 | 2 | 9.19 | 0 | 256.37 | 256.37 | 2.36 | 9 | -1359.19 | 2 |
| 0111 | 3 | 8.87 | 0 | 287.34 | 287.34 | 2.5 | 10 | -1361.03 | 5 |
| 1110 | 3 | 7.87 | 27.03 | 286.8 | 313.83 | 2.61 | 11 | -1362.63 | 11 |
| 1011 | 3 | 9.33 | 0 | 318.67 | 318.67 | 2.63 | 12 | -1361.17 | 6 |
| 1111 | 4 | 9.11 | 0 | 329.42 | 329.42 | 2.68 | 13 | -1362.96 | 13 |
| 1010 | 2 | 7.64 | 53.63 | 285.89 | 339.52 | 2.72 | 14 | -1363.14 | 15 |
| 0110 | 2 | 6.25 | 208.09 | 166.05 | 374.14 | 2.85 | 15 | -1361.66 | 7 |
| 0010 | 1 | 6.05 | 253.67 | 165.5 | 419.17 | 3.02 | 16 | -1362.09 | 9 |

individual $i$, can be written as

$$\text{Cov}(y_{ij}, y_{ik}) = \begin{cases} \sigma^2 e^{2t_j\delta_1}, & \text{if} \quad j = k, \\ \sigma^2(1-\kappa)e^{2(t_j+t_k)\delta_1 - \frac{|t_j-t_k|}{R}}, & \text{if} \quad j \neq k, \end{cases} \tag{4.38}$$

in terms of parameter vector $\boldsymbol{\theta} = (\sigma, \delta_1, \delta_0, R, \kappa)^{\intercal}$, where $\sigma$ is a scaling parameter; $\delta_1$ regulates the exponential growth of the variances; $R$ regulates the exponential decay of correlations; and $\kappa$ is a nugget effect. The corresponding variance-covariance model of a non-endogeneous patient is found by replacing $\delta_1$ with rate $\delta_0$.

Recall also that the covariance matrix of an endogeneous patient was labelled $\boldsymbol{\Sigma}_1$, and that of a non-endogenous patient was labelled $\boldsymbol{\Sigma}_0$. Since the protected and unprotected design matrices, (4.36) and (4.37), are determined by baseline covariate score and depression type respectively, for a patient of baseline score $b$, denote the matrix of protected covariates as $\boldsymbol{X}_{p,b}$ and the unprotected covariates by $\boldsymbol{X}_{u,1}$ or $\boldsymbol{X}_{u,0}$ depending on the depression type ($1$ =endogenous, $0$ =non-endogenous). These quantities will determine the probability of a patient with a specific baseline score being depression free at the end of the study.

Thus, the foci of an endogeneous patient, indexed by baseline score $b$ in the set

$\{14, ..., 34\}$, are

$$\mu_b = P(y_{i4} \leq y^*) \tag{4.39}$$
$$= \int_{\Omega^*} (2\pi)^{-\frac{5}{2}} |\boldsymbol{\Sigma}_1(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\Big( - \frac{1}{2} \boldsymbol{w}^\mathsf{T} \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{w} \Big) d\boldsymbol{y}, \tag{4.40}$$

where

$$\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{X}_{p,b}\boldsymbol{\beta} - \boldsymbol{X}_{u,1}\boldsymbol{\gamma},$$

$\Omega^* = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, y^*)$, with $y^* = 7$ the no depression threshold value and $5$ the number of measurement occasions.

The chosen variance-covariance model, given by (4.38), has parameter vector $\boldsymbol{\theta} = (\sigma, \delta_1, \delta_0, R, \kappa)^\mathsf{T}$. In order to carry out AFIC for foci that are functions of these parameters, an estimate of the quantity $\boldsymbol{J_{\theta\theta}}$ will be needed. Therefore, making use of the explicit expression (4.24), the derivatives of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_0$ with respect to $\boldsymbol{\theta}$ are required. We have, for example, the $(i, j)$th element of the partial derivative of $\boldsymbol{\Sigma}_1$ with respect to $\sigma$ as

$$\Big(\frac{\partial \boldsymbol{\Sigma}_1(\boldsymbol{\theta})}{\partial \sigma}\Big)_{i,j} = \begin{cases} 2\sigma e^{2t_j\delta_1}, & \text{if} \quad i = j, \\ 2(1-\kappa)\sigma e^{|t_i - t_j|\delta_1 - \frac{|t_i - t_j|}{R}} & \text{otherwise.} \end{cases}$$

Similar calculations can be performed for the other variance-covariance parameters and likewise for $\boldsymbol{\Sigma}_0$ (see (A.4) in Appendix A), and using the estimates of the wide model, $\boldsymbol{J_{\theta\theta}}$ may be estimated.

In addition, for $\boldsymbol{\omega}$ and $\tau_0^2$, we require the partial derivatives of the focus evaluated at the narrow model. Differentiating under the integral sign of (4.40) with respect to the regression parameters, we get

$$\frac{\partial \mu_b}{\partial \boldsymbol{\beta}} = \int_{\Omega^*} \boldsymbol{X}_{p,b}^\mathsf{T} \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}(\boldsymbol{y}) f_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y}, \tag{4.41}$$
$$\frac{\partial \mu_b}{\partial \boldsymbol{\gamma}} = \int_{\Omega^*} \boldsymbol{X}_{u,1}^\mathsf{T} \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}(\boldsymbol{y}) f_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y},$$

where

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = (2\pi)^{-\frac{5}{2}} |\boldsymbol{\Sigma}_1(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\Big( - \frac{1}{2} \boldsymbol{\epsilon}^\mathsf{T}(\boldsymbol{y}) \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}(\boldsymbol{y}) \Big),$$

and $\boldsymbol{\epsilon}(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{X}_{p,b}\boldsymbol{\beta}$.

The partial derivatives of $\mu_b$ with respect to the $k$th element of $\boldsymbol{\theta}$ may be written as

$$\frac{\partial \mu_b}{\partial \theta_k} = \int_{\Omega^*} \frac{1}{2} \left( - \operatorname{Tr} \left\{ \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}_1}{\partial \theta_k} \right\} + \boldsymbol{\epsilon}^\mathsf{T}(\boldsymbol{y}) \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}_1}{\partial \theta_k} \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}(\boldsymbol{y}) \right) f_{\boldsymbol{y}}(\boldsymbol{y}),$$

where the derivative of the determinant was found by (A.3), and (A.1) was applied to the derivative of the inverse.

Table 4.4: For the set of foci (4.40), the AIC and AFIC scores are given for each model along with their relative ranking.

| model | AFIC | rAFIC | AIC | rAIC |
|---|---|---|---|---|
| 0000 | 0.131 | 1 | -1363.50 | 16 |
| 0100 | 0.135 | 2 | -1365.96 | 12 |
| 0001 | 0.168 | 3 | -1361.45 | 1 |
| 0101 | 0.169 | 4 | -1363.35 | 3 |
| 1000 | 0.18 | 5 | -1366.25 | 13 |
| 1100 | 0.188 | 6 | -1365.35 | 9 |
| 1001 | 0.192 | 7 | -1363.4 | 4 |
| 1101 | 0.193 | 8 | -1365.21 | 8 |
| 0011 | 0.236 | 9 | -1362.67 | 2 |
| 0111 | 0.241 | 10 | -1364.44 | 5 |
| 1110 | 0.279 | 11 | -1365.86 | 11 |
| 1011 | 0.286 | 12 | -1364.67 | 6 |
| 1111 | 0.286 | 13 | -1366.38 | 14 |
| 0110 | 0.286 | 14 | -1365.18 | 7 |
| 1010 | 0.346 | 15 | -1366.56 | 15 |
| 0010 | 0.447 | 16 | -1365.8 | 10 |

These three integrals, the first two of which have integrands of dimension $p_2 = 3$ and $q = 4$ respectively, were approximated via Monte Carlo integration. In order to do this, each integral was re-expressed as an expectation with respect to a multivariate normal distribution over the unrestricted domain of integration, $\Omega$. For example, it is possible to write (4.41) as

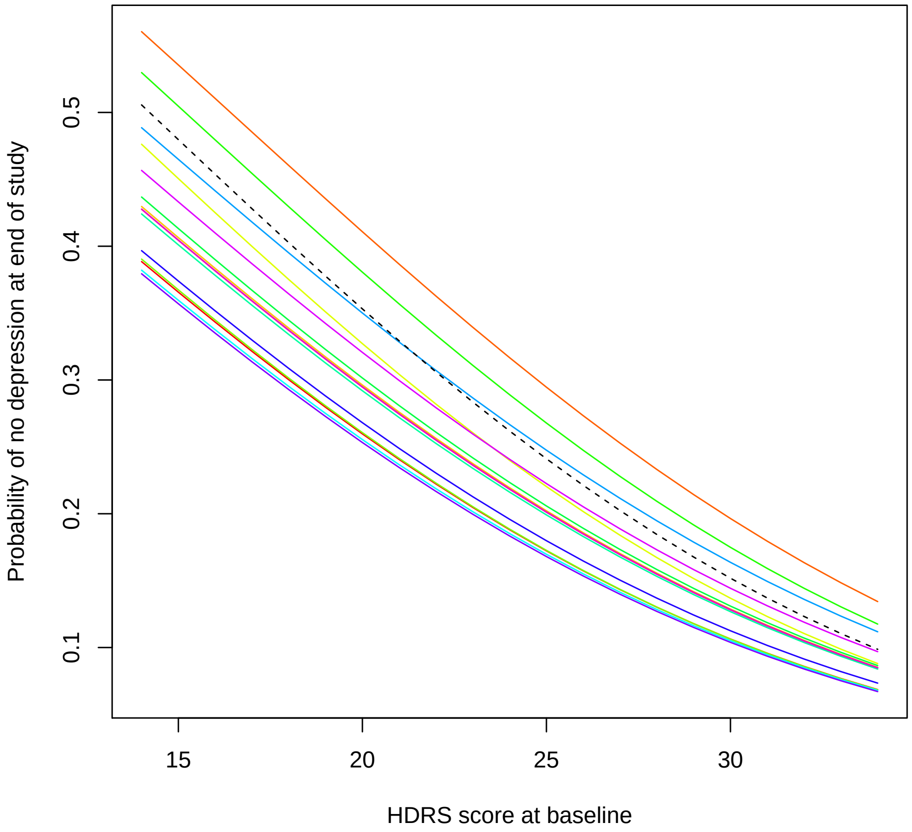
$$\begin{aligned}\frac{\partial \mu_b}{\partial \boldsymbol{\beta}} &= \int_\Omega \boldsymbol{X}_{p,b}^{\intercal} \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}(\boldsymbol{y}) \mathbf{1}_{\{y_{i4} \leq y^*\}} f_{\boldsymbol{y}}(\boldsymbol{y}) \\ &= \mathbb{E}_f[\boldsymbol{X}_{p,b}^{\intercal} \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}(\boldsymbol{y}) \mathbf{1}_{\{y_{i4} \leq y^*\}}]. \qquad (4.42)\end{aligned}$$

Furthermore, since the true $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and its derivatives are unknown, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Sigma}}$ etc. as estimated under the wide model via REML were used as direct substitutes.

Using the *rmvnorm* command of the **mvtnorm** package in R (Genz & Bretz 2009, Genz et al. 2017), 1000 samples were drawn from the appropriate multivariate normal density. The function $\boldsymbol{X}_{p,b}^{\intercal} \boldsymbol{\Sigma}_e^{-1} \boldsymbol{\epsilon}(\boldsymbol{y}) \mathbf{1}_{\{y_{i4} \leq y^*\}}$ was then evaluated at these samples and the mean of these evaluations was calculated to give an approximation to (4.41). This process was repeated for each of the required integrals, and, within a loop, for each of the foci.

The FIC scores for each foci and for each model were then calculated by plugging in the estimated and approximated quantities into (4.10), (4.12), and (4.13) (or rather, their cluster versions). Estimates of the focus, (4.40), were produced using the *pmvnorm* command of the **mvtnorm** package. These are displayed in Figure 4.5

Figure 4.5: The estimated probabilities for each model of being depression free by the end of study for endogeneous patients against baseline score. The estimates from the favourite AFIC model is shown with a dashed line.



for each of the baseline scores. The results are, of course, not continuous, but are displayed as such for the purpose of illustration.

The FIC scores averaged over all foci are displayed in Table 4.4 which shows, again, difference between the rankings of the ML based AIC and the AFIC scores. The model with smallest FIC averaged over all baseline situations is the narrow model and is shown in Figure 4.5 with a dashed line. This model estimates that all patients classified as severely depressed $(\text{HAMD} \geq 24)$ at baseline have a probability smaller than 0.26 of no depression by the end of study, and a probability between 0.43 and 0.26 for those moderately depressed $(17 \leq \text{HAMD} \leq 23)$ at baseline.

## 4.4 FIC for generalised linear mixed models

Since for GLMMs an expression for the marginal density of each cluster (3.2) is available, the FIC formula (4.19) for selection of covariates in GLMMs can be set to work with a chosen link function and random effect structure. As by Section 4.2, the vector of protected parameters $\boldsymbol{\nu}$ includes the protected regression parameters, the variance-covariance parameters of the random effects and any scale parameter. As with LME models, the focus parameter, $\mu$, should not include a random effect since random effects are random variables. By defining odds ratios, or rate ratios for example, one can arrive at sensible focus parameters.

The expected information matrix for the canonical class of GLMMs is given for general random effect structures with potentially numerous hierarchies in Wand (2007). Computing expected information matrices for GLMMs involves numerical integration, the complexity of which depends on the dimension of the random effects and the number of units within clusters. The observed information, which serves as an approximation, may be more readily available as output from software fitted models, and simpler to calculate, particularly so when there is missing data. For example, the observed information matrix of a logistic GLMM is readily available after calling *glmmML* in R, as the inverse of $N$ times the variance-covariance matrix of the model parameters. Nevertheless, it is interesting to see exact expressions for these matrices.

### 4.4.1 Derivations of information matrices of a logistic GLMM with a random intercept.

The logistic GLMM of Section 3.2.1, but with only a random intercept, will be considered here and expressions for its expected and observed information matrices are derived.

Define parameters $\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}$ as the vector of protected and unprotected regression coefficients. In addition, define the fixed effect design matrix as $\boldsymbol{X}_i = (\boldsymbol{X}_{p,i}\ \boldsymbol{X}_{u,i})$ to be the concatenation of protected and unprotected design matrices, with $j$th row $\boldsymbol{x}_{ij}^{\mathsf{T}}$. Then the wide model is

$$\begin{aligned} f_{y_{ij}|b_i}(y_{ij}|b_i) =& p_{ij}^{y_{ij}}(1-p_{ij})^{(1-y_{ij})}, && (4.43) \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) =& \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha} + b_i, && (4.44) \\ b_i \sim& \ N(0, \sigma_b^2), \end{aligned}$$

where $p_{ij} = P(y_{ij} = 1|b_i)$ is the probability of unit $j$ of cluster $i$ taking value one, and $1 - p_{ij}$ is the probability of unit $j$ of cluster $i$ taking value zero.

First of all, note that from (3.4) and (4.44) we have that (4.43) can be re-written as

$$f_{y_{ij}|b_i}(y_{ij}|b_i) = \exp\Big(y_{ij}(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha} + b_i) - \log(1 + e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i})\Big),$$

and by conditional independence we have

$$\begin{aligned} f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) &= \prod_{j=1}^{n_i} f_{y_{ij}|b_i}(y_{ij}|b_i) \\ &= \exp\left(\sum_{j=1}^{n_i} y_{ij}(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha} + b_i) - \log(1 + e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i})\right). \end{aligned} \tag{4.45}$$

The marginal log-likelihood of a single cluster is

$$\ell_i(\boldsymbol{\alpha}|\boldsymbol{y}) = \log\left(\int f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) f_{b_i}(b_i) db_i\right),$$

where $f_{b_i}(b_i)$ is the $N(0, \sigma_b^2)$ density of the random effects.

The score function of protected and unprotected regression coefficients for the $i$th cluster is thus

$$\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} = \frac{\int \frac{\partial f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i)}{\partial \boldsymbol{\alpha}} f_{b_i}(b_i) db_i}{\int f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) f_{b_i}(b_i) db_i}, \tag{4.46}$$

where

$$\frac{\partial f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i)}{\partial \boldsymbol{\alpha}} = f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) \sum_{j=1}^{n_i} \left[y_{ij} - \frac{e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}}{1 + e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}}\right] \boldsymbol{x}_{ij}^{\mathsf{T}}.$$

Similarly we have

$$\frac{\partial \ell_i}{\partial \sigma_b} = \frac{\int f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) \frac{\partial f_{b_i}(b_i)}{\partial \sigma_b} db_i}{\int f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) f_{b_i}(b_i) db_i}, \tag{4.47}$$

where

$$\frac{\partial f_{b_i}(b_i)}{\partial \sigma_b} = \frac{1}{\sqrt{2\pi\sigma_b^2}}\Big(\frac{b_i^2}{\sigma_b^2} - 1\Big) e^{-\frac{b_i^2}{2\sigma_b^2}} = v(b_i) f_{b_i}(b_i),$$

with

$$v(b_i) = \frac{1}{\sigma_b}\Big(\frac{b_i^2}{\sigma_b^2} - 1\Big). \tag{4.48}$$

### Expected information matrix

For the expected information, what we are interested in is the covariance of the scores. Since the scores are mean zero under the wide model (4.44), when averaged over all clusters, the covariance of the scores corresponding to the regression coefficients becomes (i.e. the component of the expected information corresponding to the regression coefficients is)

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial\ell_i}{\partial\boldsymbol{\alpha}}(\boldsymbol{y}_i)\right)^{\mathsf{T}}\frac{\partial\ell_i}{\partial\boldsymbol{\alpha}}(\boldsymbol{y}_i)\right]. \tag{4.49}$$

Similarly the component corresponding to the standard deviation parameter of the random effects is

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial\ell_i}{\partial\sigma_b}(\boldsymbol{y}_i)\right)^{2}\right], \tag{4.50}$$

and the cross-term is

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial\ell_i}{\partial\boldsymbol{\alpha}}(\boldsymbol{y}_i)\right)^{\mathsf{T}}\frac{\partial\ell_i}{\partial\sigma_b}(\boldsymbol{y}_i)\right], \tag{4.51}$$

which are averages of expectations with respect to $\boldsymbol{y}_i$. These expectations, which give expressions for the components of the expected information matrix of the logistic model, can be evaluated at the narrow $(\hat{\boldsymbol{\beta}}_{\text{narr}}, \mathbf{0})$ or at the wide estimate $(\hat{\boldsymbol{\beta}}, \hat{\gamma})$ (Claeskens et al. 2008, p.154) for use in the FIC framework.

In practice, there are integrals within integrals to be approximated here. Monte Carlo methods are typically more suited for the outer integrals (the expectations with respect to the data), since the number of units within each cluster may be relatively large. However, samples cannot be drawn directly from the marginal of $\boldsymbol{y}_i$. Instead, one may draw from the distribution of $b_i$, and then subsequently from the conditional distribution of $\boldsymbol{y}_i$ given the sampled $b_i$. The inner integrals consist of integrating out $b_i$ which, in this case, are univariate. Deterministic approaches such as Gauss-Hermite quadrature are therefore acceptable for the inner integrals. If this is the approach to be taken, integration with respect to the random effects has to be performed for each of the sampled $\boldsymbol{y}_i$.

### Observed information matrix

By the quotient rule, the component of the observed information corresponding to the regression parameters, found by taking the sum and derivative of (4.46) under the integral sign, is

$$-\frac{1}{N}\frac{\partial^2\ell_N}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathsf{T}}} = -\frac{1}{N}\frac{1}{(f_{\boldsymbol{y}_i}(\boldsymbol{y}_i))^2}\left(\frac{\partial h(\boldsymbol{y}_i)}{\partial\boldsymbol{\alpha}^{\mathsf{T}}}f_{\boldsymbol{y}_i}(\boldsymbol{y}_i) - \frac{\partial f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial\boldsymbol{\alpha}^{\mathsf{T}}}h(\boldsymbol{y}_i)\right), \tag{4.52}$$

where $\dfrac{\partial f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial \boldsymbol{\alpha}^{\mathsf{T}}}$ is the transpose of (4.46), $h(\boldsymbol{y}_i)$ is the numerator of (4.46), and

$$\frac{\partial h(\boldsymbol{y}_i)}{\partial \boldsymbol{\alpha}^{\mathsf{T}}} =$$

$$\int \left( \frac{\partial f(\boldsymbol{y}_i)}{\partial \boldsymbol{\alpha}^{\mathsf{T}}} \sum_{j=1}^{n_i} \left[ y_{ij} - \frac{e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}}{1+e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}} \right] \boldsymbol{x}_{ij}^{\mathsf{T}} - f(\boldsymbol{y}_i|b_i) \sum_{j=1}^{n_i} \left[ \frac{e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}}{(1+e^{\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\alpha}})^2} ]\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{T}} \right]\Bigg] \right) f_{b_i}(b_i) db_i.$$

Similarly, for the component corresponding to the standard deviation of the random effects, we have

$$-\frac{1}{N}\frac{\partial^2 \ell_N}{\partial \sigma_b^2} = -\frac{1}{N}\frac{1}{(f(\boldsymbol{y}_i))^2}\left( \frac{\partial^2 f(\boldsymbol{y}_i)}{\partial \sigma_b^2} f(\boldsymbol{y}_i) - \left( \frac{\partial f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial \sigma_b} \right)^2 \right), \qquad (4.53)$$

where

$$\frac{\partial^2 f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial \sigma_b^2} = \int f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) \left[ v(b_i) + \frac{\partial v(b_i)}{\partial \sigma_b} \right] f_{b_i}(b_i) db_i,$$

with $v(b_i)$ as in (4.48),

$$\frac{\partial v(b_i)}{\partial \sigma_b} = \frac{1}{\sigma_b^2}\Big(1 - \frac{3b_i}{\sigma_b^2}\Big),$$

and $\dfrac{\partial f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial \sigma_b}$ the score function given in (4.47). Finally, for the cross-term we have

$$-\frac{1}{N}\frac{\partial \ell_N}{\partial \sigma_b \partial \boldsymbol{\alpha}} = -\frac{1}{N}\frac{1}{(f_{\boldsymbol{y}_i}(\boldsymbol{y}_i))^2}\left( \frac{\partial^2 f(\boldsymbol{y}_i)}{\partial \sigma_b \partial \boldsymbol{\alpha}} f_{\boldsymbol{y}_i}(\boldsymbol{y}_i) - \frac{\partial f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial \boldsymbol{\alpha}} \frac{f_{\boldsymbol{y}_i}(\boldsymbol{y}_i)}{\partial \sigma_b} \right), \qquad (4.54)$$

where

$$\frac{\partial^2 f(\boldsymbol{y}_i)}{\partial \sigma_b \partial \boldsymbol{\alpha}} = \int f_{\boldsymbol{y}_i|b_i}(\boldsymbol{y}_i|b_i) h(\boldsymbol{y}_i) v(b_i) f_{b_i}(b_i) db_i.$$

The integrals in (4.52), (4.53) and (4.54) are all with respect to a univariate normal density. Thus, in application, Gauss-Hermite integration may be suitable for approximating these quantities.

### 4.4.2 Data illustration

Recall the binary dataset introduced in Section 3.4 of Chapter 3. Suppose that the focus of interest is the following odds ratio:

$$\frac{\text{Odds of onycholysis}|\text{group}_i = A, b_i, t = t_k}{\text{Odds of onycholysis}|\text{group}_i = B, b_i, t = t_k}. \qquad (4.55)$$

That is, the focus is an odds ratio comparing an individual of Treatment A to an individual of Treatment B at time $j$, both of which happen to share the same random effect, $b_i$. In other words, the same tendency (whatever strength that might be, but that persists throughout the study) to be observed as moderately or severely infected is shared by both individuals.

The widest model under consideration is the logistic cubic

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \gamma_0 t_{ij}^2 + \gamma_1 t_{ij}^3 \\ + \gamma_2 t_{ij}\mathrm{group}_i + \gamma_3 t_{ij}^2 \mathrm{group}_i + \gamma_4 t_{ij}^3 \mathrm{group}_i + b_i,$$

where time, $t$, was centered about its mean. The focus (4.55) can be written as

$$\frac{e^{\boldsymbol{x}_{A,ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}}{e^{\boldsymbol{x}_{B,ij}^{\mathsf{T}}\boldsymbol{\alpha}+b_i}},$$

where $\boldsymbol{x}_{A,ij}^{\mathsf{T}}$ is the $j$th row of the design matrix of individual $i$ from treatment group A, and similarly for $\boldsymbol{x}_{B,ij}^{\mathsf{T}}$. Since both individuals under comparison are assumed to have measurements observed at the same set of scheduled occasions (at 1, 2, 3, 6, 9 and 12 months), and also share the same random effect, $b_i$, this reduces to

$$e^{(\boldsymbol{x}_{A,ij}-\boldsymbol{x}_{B,ij})^{\mathsf{T}}\boldsymbol{\alpha}} = e^{\gamma_3 t_{ij}+\gamma_4 t_{ij}^2+\gamma_5 t_{ij}^3}. \tag{4.56}$$

For this set-up, there are four potential models (the narrow, 00000, and 10000, 01000, 11000), which all set $\gamma_3 = \gamma_4 = \gamma_5 = 0$ and thus estimate the focus to take value 1 with zero variance. Of these four, only the narrow is included as a candidate.

The derivatives of (4.56) with respect to the model parameters, evaluated at the narrow model, give estimates for the required partial derivatives in the cluster version of (4.13).

The observed information matrix [see formulas (4.52), (4.54), (4.53)] of the wide model was used as an approximation of the expected information matrix. Doing so is advantageous since this dataset contains missing data which has to be accounted for by the expected information matrix. 60 quadrature points were used in the Gauss-Hermite quadrature for approximating the integrals of the observed information matrix. The command *gauss.quad* from the R package **statmod** (Smyth 2005) was used to create the quadrature points. In addition, the observed information matrix was evaluated at the wide estimate (rather than the narrow), which builds in a certain model robustness (Claeskens et al. 2008, p.154).

This focus, (4.55), was estimated by each model for each of the scheduled time points, producing corresponding FIC scores. The AFIC was then calculated over all 7 time points. Table 4.5 presents the AFIC scores (divided by the number of patients $N = 294$) and the AIC scores of the top five models (out of 29) as rated by AFIC. Once again, there are disagreements between the rankings of AIC and AFIC.[5]

[5] Admittedly, the conditional AIC would be more appropriate than the AIC since the focus is at the level of the individual. However, implementation of this is a slow process in R package **cAIC4** (Saefken et al. 2018*a*,*b*) for Bernoulli models due to the number of model-refits required.

Table 4.5: The AFIC (divided by N) and AIC scores for the top 5 models (as judged by AFIC) for the onycholysis data set with focus given in (4.55).

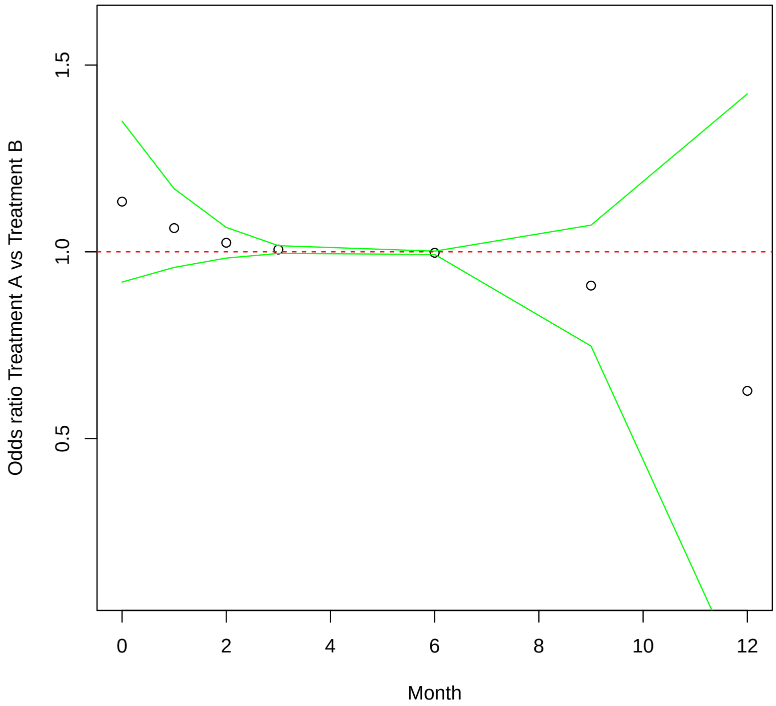
| model | AFIC/N | AIC |
|---|---|---|
| 10001 | 0.027 | -1243.64 |
| 00000 | 0.029 | -1261.84 |
| 01001 | 0.035 | -1244.58 |
| 11001 | 0.036 | -1241.7 |
| 00100 | 0.062 | -1260.75 |

Figure 4.6 shows the focus estimates of the favoured AFIC model, 10001, for each of the 7 time points along with 95% confidence intervals (which are optimistic as they neglect the uncertainty in the model selection procedure). Under the assumptions of model 10001 and assuming an all-available data analysis is appropriate, this figure suggests that, since the 95% confidence intervals include the no-difference between treatment line, there is no significant difference between the odds of onycholysis of two individuals from both treatments who happen to have the same random effect.

## 4.5 Chapter summary

In this chapter, the focussed information criterion for independent data as in in Claeskens & Hjort (2003) was introduced. How this generalises to clustered data within a multivariate misspecified framework was then made explicit. In particular, the framework formulated in Section 4.2.1 opens the door to application of the FIC for covariate selection in multivariate LMs, LME models and GLMMs. With regard to the Normal model, a simulation study was carried out to see how influential the uncertainty in over-parametrised variance-covariance models is upon the limiting distribution of focus parameters that are only a function of the regression coefficients. As examples of the framework in Section 4.2.1, the FIC was applied to select covariates in the multivariate Normal model and a logistic GLMM for two clinical trial datasets. Explicit formulae for the information matrices of the multivariate linear model and the logistic GLMM were also given.

Figure 4.6: The focus [see (4.55)] as estimated by model 10001 for each scheduled measurement occasion 1, 2, 3, 6, 9 and 12. The dashed red line is where no difference between treatments would be observed i.e. the odds of treatment A is a factor of one time the odds of Treatment B. The solid green lines give upper and lower 95% confidence intervals (albeit optmistic).



# Chapter 5

# Derivations of Mean Squared Error Formulae

For the Normal model, an alternative approach to constructing FIC scores is available for a subset of foci. The method, described in the appendix of Cunen et al. (2017), begins with finding the expected value and covariance of the generalised least squares (GLS) estimator for each candidate model, under the assumption that the wide model's first two moments are correctly specified, and, thereby, the bias squared and variance of the focus estimator of each model. This method was used in Cunen et al. (2017) to construct FIC formulas for linear mixed effect (LME) models, albeit only for the bias squared term. Since the focus of interest in Cunen et al. (2017) was a single regression coefficient, the formulas were given for that specific case, and noted to be readily available for more general functions of the regression coefficients.

In this chapter, the approach will be demonstrated for a general linear combination of regression coefficients and for a multivariate linear combination of regression coefficients. An extension, following the work of Kackar & Harville (1984), which also takes into account the uncertainty in the variance-covariance parameters is also presented. Lastly, a similar approach is demonstrated for ranking LME models in terms of the precision of predictors of cluster-specific trajectories.

## 5.1 MSE formula for a focus that is a general linear combination of regression parameters

For this chapter, the slightly misspecified framework of (4.15) is not needed. Rather, an alternative, but similar setup is required. It will be assumed that a wide model, with fullest mean structure is the true model. The form of its covariance matrix depends on whether we are considering a model with random effects (the marginal model), or no random effects at all. Then, independent vector responses $\boldsymbol{y}_i$ are generated by

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_{p,i}\boldsymbol{\beta} + \boldsymbol{X}_{u,i}\boldsymbol{\gamma}, \boldsymbol{\Sigma}_i(\boldsymbol{\theta})). \tag{5.1}$$

In the same way as the FIC setup, let model (say) $S$ be a function of the protected regression parameters $\boldsymbol{\beta}$ of dimension $p$ and the unprotected regression parameters $\boldsymbol{\gamma}_S$ of dimension $|S|$, a subset of the wide models $\boldsymbol{\gamma}$, which is of dimension $q$.

In terms of design matrices for each individual $i$, all models will include a matrix of protected covariates, $\boldsymbol{X}_{p,i}$, but only the wide model will additionally include all of the unprotected covariates, $\boldsymbol{X}_{u,i}$. So, the wide model has an $n_i \times (p+q)$ design matrix, $\boldsymbol{X}_i = (\boldsymbol{X}_{p,i}\ \boldsymbol{X}_{u,i})$, which is the concatenation of the protected and unprotected covariates. Candidate model $S$ has design matrix $\boldsymbol{X}_{S,i} = (\boldsymbol{X}_{p,i}\ \boldsymbol{X}_{u_S,i})$ where $\boldsymbol{X}_{u_S,i}$ is the matrix of columns of $\boldsymbol{X}_{u,i}$ whose indices are in the set $S$ ($S$ denotes a set; model $S$ refers to that model which includes the unprotected covariates whose indices appear in set $S$).

Let $\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}$ be the vector of all regression coefficients. For a focus $\mu$ that is a linear combination of the elements of $\boldsymbol{\alpha}$, whose weights are defined by the $(p+q)$ row vector $\boldsymbol{m}$, we may write $\mu = \boldsymbol{m}\boldsymbol{\alpha}$. Also required is the $p + |S|$ row vector $\boldsymbol{m}_S$ whose first $p$ entries correspond to the first $p$ entries of $\boldsymbol{m}$ (i.e. the protected regression weights) and the remaining $|S|$ entries are the weights of the unprotected $\gamma$s that appear in set $S$. It is assumed that the variance-covariance model, $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$, is the same across candidate models. Furthermore, its parameter values are considered known, with superscript $^o$ used to denote this. Mean squared error formulae will now be derived.

We have that the generalised least squares estimator of $\boldsymbol{\alpha}$ for model $S$ is

$$\hat{\boldsymbol{\alpha}}_S^o = \begin{pmatrix} \hat{\boldsymbol{\beta}}_S^o \\ \hat{\boldsymbol{\gamma}}_S^o \end{pmatrix} = \left( \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o) \boldsymbol{X}_{S,i} \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o) \boldsymbol{y}_i. \tag{5.2}$$

Then, the estimate of $\mu$ produced by model $S$ is $\hat{\mu}_S^o = \boldsymbol{m}_S \hat{\boldsymbol{\alpha}}_S^o$.

Since, under the wide model, $\mathbb{E}[\boldsymbol{y}_i] = \boldsymbol{X}_i\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the true parameter value, taking the expected value of (5.2) gives

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\alpha}}_S^o] = & \left( \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o) \boldsymbol{X}_{S,i} \right)^{-1} \left( \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o) \boldsymbol{X}_i \right) \boldsymbol{\alpha} \\ = & \boldsymbol{B}_S^{o-1} \boldsymbol{D}_S^o \boldsymbol{\alpha}, \end{aligned} \tag{5.3}$$

with

$$\boldsymbol{B}_S^o = \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o) \boldsymbol{X}_{S,i}, \quad \text{and} \quad \boldsymbol{D}_S^o = \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o) \boldsymbol{X}_i. \tag{5.4}$$

In addition, define

$$\boldsymbol{A}_S^o = \boldsymbol{B}_S^{o-1} \boldsymbol{D}_S^o. \tag{5.5}$$

For the bias squared term, we have that the bias in the estimator $\hat{\mu}_S^o$ is

$$\begin{aligned}\text{bias}_S = \mathbb{E}[\hat{\mu}_S^o - \mu] &= \mathbb{E}[\boldsymbol{m}_S \hat{\boldsymbol{\alpha}}_S^o - \boldsymbol{m}\boldsymbol{\alpha}] \\ =& \boldsymbol{m}_S \mathbb{E}[\hat{\boldsymbol{\alpha}}_S^o] - \boldsymbol{m}\boldsymbol{\alpha} = (\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})\boldsymbol{\alpha},\end{aligned} \tag{5.6}$$

where the fourth equality follows from (5.3) combined with (5.5).

A naive estimator of bias squared for model $S$ is then

$$\widehat{\text{bias}}_S^{o2} = ((\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})\hat{\boldsymbol{\alpha}}^o)^2,$$

where $\hat{\boldsymbol{\alpha}}^o$ is the estimator of the wide model. Since for any random variable say $b$, $\text{E}[b^2] = \text{E}[b]^2 - \text{Var}(b)$, this naive estimator overshoots by the amount $\text{Var}(\widehat{\text{bias}}_S^o)$. An improved estimator is therefore

$$(\boldsymbol{m}_S(\boldsymbol{A}_S^o - \boldsymbol{m})\hat{\boldsymbol{\alpha}}^o)^2 - \text{Var}(\widehat{\text{bias}}_S^o),$$

where, by (5.6),

$$\text{Var}(\widehat{\text{bias}}_S^o) = (\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})\text{Cov}(\hat{\boldsymbol{\alpha}}^o)(\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})^\intercal.$$

We have that

$$\text{Cov}(\hat{\boldsymbol{\alpha}}^o) = \left(\sum_{i=1}^N \boldsymbol{X}_i \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o)\boldsymbol{X}_i\right)^{-1} =: \boldsymbol{B}^{o-1},$$

and so, along with truncating at zero to avoid a negative bias squared, we get

$$\max\left[0, ((\boldsymbol{m}_S \;\boldsymbol{A}_S^o - \boldsymbol{m})\hat{\boldsymbol{\alpha}}^o)^2 - (\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})\boldsymbol{B}^{o-1}(\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})^\intercal\right] \tag{5.7}$$

as an improved estimator of bias squared.

As for the variance term, note that under the wide model, the variance-covariance of $\hat{\boldsymbol{\alpha}}_S^o$ is

$$\text{Cov}(\hat{\boldsymbol{\alpha}}_S^o) = \boldsymbol{B}_S^{o-1}. \tag{5.8}$$

Therefore, the variance in estimation of the true focus $\mu = \boldsymbol{m}\boldsymbol{\alpha}$ for model $S$ is

$$\begin{aligned}\text{Var}(\hat{\mu}_S^o) &= \boldsymbol{m}_S \text{Cov}(\hat{\boldsymbol{\alpha}}_S^o)\boldsymbol{m}_S^\intercal \\ &= \boldsymbol{m}_S \boldsymbol{B}_S^{o-1}\boldsymbol{m}_S^\intercal.\end{aligned} \tag{5.9}$$

Summing (5.9) and (5.7) together then gives an estimator for the MSE of model $S$ in estimation of $\mu = \boldsymbol{m}\boldsymbol{\alpha}$ as

$$\begin{aligned}\widehat{\text{MSE}}_S^o = \boldsymbol{m}_S \boldsymbol{B}_S^{o-1}\boldsymbol{m}_S^\intercal + \max\Big[0, &((\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})\hat{\boldsymbol{\alpha}}^o)^2 \\ &- (\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})\boldsymbol{B}^{o-1}(\boldsymbol{m}_S \boldsymbol{A}_S^o - \boldsymbol{m})^\intercal\Big],\end{aligned} \tag{5.10}$$

which, for known $\boldsymbol{\theta}$, the variance term is exact and the bias squared term is unbiased prior to truncation.

The same derivation when $\boldsymbol{\theta}$ has to be estimated does not hold: the uncertainty in the estimates of $\boldsymbol{\theta}$ needs to be taken into account. Therefore, for such a situation, a plausible option is to use the estimates of $\boldsymbol{\theta}$ based on the widest mean structure and simply plug them into the relevant quantities of formula (5.10). That is, to use the estimators

$$\hat{\boldsymbol{A}}_S = \hat{\boldsymbol{B}}_S^{-1} \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}) \boldsymbol{X}_i, \quad \hat{\boldsymbol{B}}_S^{-1} = \left( \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}) \boldsymbol{X}_{S,i} \right)^{-1}, \tag{5.11}$$

and

$$\hat{\boldsymbol{B}}^{-1} = \left( \sum_{i=1}^{N} \boldsymbol{X}_i^{\intercal} \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\theta}}) \boldsymbol{X}_i \right)^{-1}. \tag{5.12}$$

Plugging in estimates into (5.10) could, for example, rank models in terms of their estimation of the focus $\mu$, which is the expected marginal response given a vector of covariates $\boldsymbol{m}$, $\mathbb{E}[y_{ij} | \boldsymbol{x}_{ij}^{\intercal} = \boldsymbol{m}]$. Or, similar to Cunen et al. (2017), to estimate the $k$th regression coefficient, in which case $\boldsymbol{m} = \frac{\partial \boldsymbol{\beta}}{\partial \beta_k}$, i.e. a vector of zeroes with a single entry of one in the $k$th slot. Furthermore, although this method has been presented in the multivariate linear regression setting, equivalent formulas can equally be derived for univariate linear regression.

As a final comment, note that one could consider a similar situation for choosing between covariance models, which should be done based on a maximal mean (or fixed effect) structure. For such a situation, with the maximal mean structure assumed true, the bias squared term would be zero; the variance term is the only contributing factor. However, there would be no guarantee that the favoured models would be correctly specified, so no formula is given here.

### 5.1.1 Data illustration

As an illustration of formula (5.10) consider the depression data set introduced in Section 2.7. Consider also the first focus of Section 4.3.3, namely

$$\mu(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbb{E}[hd_{i4} | base_i = 27, tc_{i4} = 2, ed_i = 0],$$

the expected response of a non-endogeneous depressed $(ed_i = 0)$ individual with an above average depression score at baseline $(base_i = 27)$ by the end of the study $(tc_{i4} = 2)$. The same variance-covariance matrix M6 chosen in Section 4.3.3 was used here. The plug-in estimators in (5.11) and (5.12) were used to generate estimates of (5.10) for each of the candidate models that are submodels of the wide model which is given in (4.35). REML was used for estimation of the variance-covariance of the wide model.

Table 5.1 displays the results, ranging from the model with smallest MSE at the top, to the largest at the bottom. The columns bias$^2$ and Var give the estimated bias squared [see (5.7)] and variance [see (5.9)], MSE gives the estimated mean squared error [see (5.10)], which is in agreement with the FIC$/N$ scores to 12 decimal places. This is unsurprising since the FIC, being asymptotic, neglects the uncertainty in the variance components, and the asymptotic covariance of the regression coefficients under the wide model, (5.12), was used as an estimator in the MSE formula. This connection between FIC and exact MSE formulas under a Normal model is discussed in Claeskens et al. (2008, p.172).[1]

Table 5.1: Mean squared error estimates by formula (5.10) for a focus that is the expected response at the end of the study of a non-endogeneous patient with baseline score 27 on the HAMD scale.

| model | bias$^2$ | Var | MSE | FIC/N |
|---|---|---|---|---|
| 1000 | 0 | 1.4908 | 1.4908 | 1.4908 |
| 0101 | 0.2756 | 1.2414 | 1.517 | 1.517 |
| 1001 | 0.0144 | 1.5035 | 1.5179 | 1.5179 |
| 1100 | 0 | 1.5602 | 1.5602 | 1.5602 |
| 0001 | 0.4021 | 1.1983 | 1.6004 | 1.6004 |
| 1101 | 0 | 1.6806 | 1.6806 | 1.6806 |
| 0100 | 0.4902 | 1.2407 | 1.7308 | 1.7308 |
| 0000 | 2.2614 | 1.1708 | 3.4322 | 3.4322 |
| 0010 | 0 | 3.7194 | 3.7194 | 3.7194 |
| 1010 | 0 | 3.7205 | 3.7205 | 3.7205 |
| 0011 | 0 | 3.9041 | 3.9041 | 3.9041 |
| 1011 | 0 | 3.925 | 3.925 | 3.925 |
| 1110 | 0.3261 | 3.7535 | 4.0796 | 4.0796 |
| 0110 | 0.3806 | 3.7526 | 4.1331 | 4.1331 |
| 0111 | 0 | 4.1658 | 4.1658 | 4.1658 |
| 1111 | 0 | 4.2677 | 4.2677 | 4.2677 |

## 5.2 Extension to a multivariate focus

Often, the research question to be answered will be a hypothesis about the regression coefficients consisting of more than one constraint. In other words, a null hypothesis may be of the form $H_0\colon \boldsymbol{M\beta} = \boldsymbol{r}$, where $\boldsymbol{r}$ is a vector of constants and $\boldsymbol{M}$ is a matrix of weights for the regression coefficients which define the constraints of the

[1] A detail implicitly assumed in Claeskens et al. (2008, p.172) is that the variance parameter $\sigma^2$ should be known.

hypothesis. For example, $\beta_1 = \beta_2 = 0$ may be the null hypothesis that two non-reference group treatment effects coincide with that of the reference group. This can be expressed as

$$\begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Such hypotheses can be addressed via multivariate Wald tests, for example. One may also be interested in simultaneously predicting all entries of a multivariate response. In such situations, the focus is multivariate.

The MSE formula (5.10) can be extended to the situations just described, allowing the statistician to rank models in terms of mean squared error in estimation of a multivariate focus. Such a formula will be derived here, when working with a fixed, known, variance-covariance matrix.[2] Choice is between unprotected regression parameters. And, one should ensure that inclusion or exclusion of unprotected regression parameters does not affect the interpretation of those parameters of interest. In particular, interpreting the parameters of interest as by the wide model may be suitable.

Define the multivariate focus as $\boldsymbol{\mu} = \boldsymbol{M}\boldsymbol{\alpha}$, where $\boldsymbol{M}$ is a matrix of dimension $k \times (p+q)$ and whose entries are constants. As discussed, they could define the weights of the linear constraints of a hypothesis, or the values of covariates used to make a prediction of a vector response. In the former case, $k$ defines the number of constraints in a hypothesis. In the latter, it is the length of the response vector to be predicted.

When $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is considered fixed and known, the estimator of $\boldsymbol{\mu}$ by the model with mean structure $S$ is $\hat{\boldsymbol{\mu}}^o_S = \boldsymbol{M}_S\hat{\boldsymbol{\alpha}}^o_S$, where $\boldsymbol{M}_S$ is a $k \times (p + |S|)$ matrix whose first $p$ columns correspond to the weights of $\boldsymbol{\beta}$. If the aim is to test a hypothesis about $\boldsymbol{\beta}$ alone, the additional $|S|$ columns would be columns of zeros. If the aim is prediction of a multivariate response, then the additional $|S|$ columns correspond to the weights of $\boldsymbol{\gamma}_S$.

The mean squared error for estimator $\hat{\boldsymbol{\mu}}^o_S$ of the multivariate focus $\boldsymbol{\mu}$ by model with mean structure $S$ is

$$\text{MSE}(\hat{\boldsymbol{\mu}}^o_S) = \text{Tr}\{\text{Cov}(\hat{\boldsymbol{\mu}}^o_S)\} + (\mathbb{E}[\hat{\boldsymbol{\mu}}^o_S] - \boldsymbol{\mu})^\intercal(\mathbb{E}[\hat{\boldsymbol{\mu}}^o_S] - \boldsymbol{\mu}).$$

The first term can be thought of as the sum of the variances of each entry of the focus, and the second as the sum of the squared bias in each entry of the focus. Both terms can be estimated in a similar fashion to the univariate case. The bias, as in (5.6) but now a vector of length $k$, is

$$\begin{aligned} \textbf{bias}_S =& \mathbb{E}[\hat{\boldsymbol{\mu}}^o_S - \boldsymbol{\mu}] = \boldsymbol{M}_S\mathbb{E}[\hat{\boldsymbol{\alpha}}^o_S] - \boldsymbol{M}\boldsymbol{\alpha} \\ =& (\boldsymbol{M}_S\boldsymbol{A}^o_S - \boldsymbol{M})\boldsymbol{\alpha}, \end{aligned}$$

[2] The word ‘fixed’ is used in the sense of chosen, the same across candidate models, and could arise marginally from a random effect structure.

where $\boldsymbol{A}_S^o$ is as in (5.5). A naive estimator of the bias squared term is thus

$$\widehat{\textbf{bias}}_S^{o\intercal}\widehat{\textbf{bias}}_S^o = \hat{\boldsymbol{\alpha}}^{o\intercal}(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})^\intercal(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})\hat{\boldsymbol{\alpha}}^o.$$

This, akin to the univariate case, overshoots by

$$\mathrm{Tr}\{\mathrm{Cov}(\widehat{\textbf{bias}}_S^o)\} = \mathrm{Tr}\{(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})\mathrm{Cov}(\hat{\boldsymbol{\alpha}}^o)(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})^\intercal\}.$$

So, a better estimator of bias squared, truncated to avoid being negative, is

$$\max\left[0, \hat{\boldsymbol{\alpha}}^{o\intercal}(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})^\intercal(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})\hat{\boldsymbol{\alpha}}^o \right.$$
$$\left. - \mathrm{Tr}\left\{(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})\left(\sum_{i=1}^N \boldsymbol{X}_i^\intercal\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}^o)\boldsymbol{X}_i\right)^{-1}(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})^\intercal\right\}\right]. \quad (5.13)$$

The variance part of the MSE formula is also derived in a similar fashion to the univariate case. The covariance of $\hat{\boldsymbol{\mu}}_S^o$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\mu}}_S^o) = \boldsymbol{M}_S\boldsymbol{B}_S^{o-1}\boldsymbol{M}_S^\intercal \quad (5.14)$$

of which the diagonal elements are of interest, and where $\boldsymbol{B}_S^o$ is as in (5.4).

So, an estimator of the MSE in estimation of the multivariate focus for each model, found by taking the trace of (5.14) and summing with (5.13), is

$$\widehat{\mathrm{MSE}}_S^o = \mathrm{Tr}\left\{\boldsymbol{M}_S\boldsymbol{B}_S^{o-1}\boldsymbol{M}_S^\intercal\right\} + \max\left[0, \hat{\boldsymbol{\alpha}}^{o\intercal}(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})^\intercal(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})\hat{\boldsymbol{\alpha}}^o\right.$$
$$\left. - \mathrm{Tr}\left\{(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})\boldsymbol{B}^{o-1}(\boldsymbol{M}_S\boldsymbol{A}_S^o - \boldsymbol{M})^\intercal\right\}\right], \quad (5.15)$$

and when $k = 1$ we are back to formula (5.10). Thus, under the assumption that the wide model is true, candidate models may be ranked according to their MSE in estimation of the multivariate linear combination of protected regression parameters, with direct application to ranking models in terms of ability to reject/fail to reject a multivariate hypothesis, or for predicting a multivariate response.

## 5.3 Accounting for uncertainty in the variance components

A limitation of of the previously considered MSE formulae (5.10) and (5.15) is that uncertainty in the variance-covariance parameters is neglected. Following the work of Kackar & Harville (1984), approximate MSE formula for both of these but that account for uncertainty in the variance-covariance parameters may be constructed. The method will be shown here for adjusting formula (5.10). The true focus is $\mu =$

$\boldsymbol{m\alpha}$, and, from candidate model $S$, has estimators $\hat{\mu}_S^o$ and $\hat{\mu}_S$, which treat $\boldsymbol{\theta}$ as known and unknown respectively.

By Kackar & Harville (1984, p.854), we have that

$$\hat{\mu}_S - \mu = (\hat{\mu}_S^o - \mu) + (\hat{\mu}_S - \hat{\mu}_S^o), \tag{5.16}$$

and both terms on the right hand side are independently distributed. Thus, the actual mean squared error in estimation by model $S$ is

$$\mathbb{E}[(\hat{\mu}_S - \mu)^2] = \mathbb{E}[(\hat{\mu}_S^o - \mu)^2] + \mathbb{E}[(\hat{\mu}_S - \hat{\mu}_S^o)^2]. \tag{5.17}$$

Formula (5.10) provides an estimator for the first term on the right hand side of (5.17), which is, in fact, a lower bound for the actual MSE (Kackar & Harville 1984). The second term captures the additional MSE due to uncertainty in estimation of $\boldsymbol{\theta}$.

Using a second order Taylor expansion for $\hat{\mu}^2_{\text{diff},S} = (\hat{\mu}_S - \hat{\mu}_S^o)^2 = (\hat{\mu}_S(\hat{\boldsymbol{\theta}}) - \hat{\mu}_S(\boldsymbol{\theta}^o))^2$ as a function of $\hat{\boldsymbol{\theta}}$ about $\boldsymbol{\theta}^o$ as in Kackar & Harville (1984), and supposing $\hat{\boldsymbol{\theta}}$ is the REML estimator to ensure unbiasedness, an approximation to the second term on the right hand side of (5.17) is

$$\begin{aligned}
\mathbb{E}[\hat{\mu}^2_{\text{diff},S}] &\approx \frac{1}{2}\mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o)^\intercal \frac{\partial \hat{\mu}^2_{\text{diff},S}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}\bigg|_{\theta_0} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^o)] \\
&= \frac{1}{2}\operatorname{Tr}\left\{ \frac{\partial \hat{\mu}^2_{\text{diff},S}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}\bigg|_{\theta_0} \operatorname{Cov}(\hat{\boldsymbol{\theta}}) \right\} \\
&= \operatorname{Tr}\left\{ \left(\frac{\partial \hat{\mu}_S}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \hat{\mu}_S}{\partial \boldsymbol{\theta}}\right)^\intercal\bigg|_{\theta_0} \operatorname{Cov}(\hat{\boldsymbol{\theta}}) \right\},
\end{aligned} \tag{5.18}$$

where the first equality is a consequence of (A.5) and the second is explained in Section A.5 of Appendix A.

So, given estimates for the quantities in (5.18) an approximation of the amount by which (5.10) undershoots the actual MSE is available. The $\operatorname{Cov}(\hat{\boldsymbol{\theta}})$ may be estimated asymptotically as the inverse of the expected information corresponding to the variance-covariance parameters (Kackar & Harville 1984) (for which, an expression is given (4.24)). For the partial derivatives, note that as a function of the variance-covariance parameters, the focus estimator for model $S$ is expressible as

$$\hat{\mu}_S(\boldsymbol{\theta}) = \boldsymbol{m}_S \hat{\boldsymbol{\alpha}}_S(\boldsymbol{\theta}) = \boldsymbol{m}_S \boldsymbol{B}_S^{-1}(\boldsymbol{\theta}) \boldsymbol{C}_S(\boldsymbol{\theta}),$$

where

$$\boldsymbol{B}_S(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^\intercal \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}) \boldsymbol{X}_{S,i} \quad \text{and} \quad \boldsymbol{C}_S(\boldsymbol{\theta}) = \boldsymbol{C}_S(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^\intercal \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}) \boldsymbol{y}_i.$$

The required partial derivatives in (5.18) are then

$$\begin{aligned}
\frac{\hat{\mu}_S(\boldsymbol{\theta})}{\partial \theta_k} &= \boldsymbol{m}_S \frac{\partial \hat{\boldsymbol{\alpha}}_S(\boldsymbol{\theta})}{\partial \theta_k} \\
&= \boldsymbol{m}_S \left( \frac{\partial \boldsymbol{B}_S^{-1}(\boldsymbol{\theta})}{\partial \theta_k} \boldsymbol{C}_S(\boldsymbol{\theta}) + \boldsymbol{B}_S^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{C}_S(\boldsymbol{\theta})}{\partial \theta_k} \right),
\end{aligned}$$

where,

$$\frac{\partial \boldsymbol{B}_S^{-1}(\boldsymbol{\theta})}{\partial \theta_k} = -\boldsymbol{B}_S^{-2}(\boldsymbol{\theta}) \left( \sum_{i=1}^{N} \boldsymbol{X}_{S,i}^{\mathsf{T}} \frac{\partial \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_k} \boldsymbol{X}_{S,i} \right), \quad \frac{\partial \boldsymbol{C}_S(\boldsymbol{\theta})}{\partial \theta_k} = \sum_{i=1}^{N} \boldsymbol{X}_{s,i}^{\mathsf{T}} \frac{\partial \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_k} \boldsymbol{y}_i,$$

and $\dfrac{\partial \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial \theta_k}$ is given by (A.1) in Appendix A.

Therefore, given an estimate of the asymptotic variance-covariance of the covariance parameters and estimates of the partial derivatives of the variance-covariance matrix of the wide model, an approximate MSE formula for model selection which accounts for the uncertainty in the variance-covariance parameters is available. This is obtained by summing (5.10) with (5.18).

### 5.3.1 A simulation study

To see how useful formula (5.18) might be in practice, a small simulation study was carried out.

500 balanced longitudinal datasets were generated from a random intercept and slope model for a fixed number of measurement occasions $n = 5$, and for each of the number of individuals $N = 10, 20, 30, 100, 250$. Four models (M1-M4), each with the same random intercept and slope, were under consideration.

The mean structure of the widest model (M4) was

$$\mathbb{E}[y_{ij}] = \beta_0 + \gamma_0 t_{ij} + \gamma_1 t_{ij} \text{group}_i,$$

where covariate group was binary and generated by the command *rbinom* in R. Model M2 was the true data generating model, and given by

$$y_{ij} = \beta_0 + b_{i,0} + (\gamma_0 + b_{i,1}) t_{ij} + \epsilon_{ij},$$

with

$$\begin{pmatrix} \boldsymbol{b}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \sim N\left( \boldsymbol{0}, \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2 \boldsymbol{I}_n \end{pmatrix} \right).$$

The true values of the regression parameters were $(\beta_0, \gamma_0, \gamma_1) = (1, -0.1, 0)$, and the true values of the variance-covariance parameters were $(\sigma^2, d_{11}, d_{12}, d_{22}) = (4, 1, 0.2, 0.1)$.

The mean structure of model M3 was

$$\mathbb{E}[y_{ij}] = \beta_0 + \gamma_1 t_{ij} \text{group}_i,$$

and that of the narrowest model, M1, was

$$\mathbb{E}[y_{ij}] = \beta_0.$$

The focus under consideration was the marginal mean response of non-reference group individuals by the end of study. That is

$$\mu = \mathbb{E}[y_{i5} | \text{group}_i = 1] = \beta_0 + \gamma_0 t_5 + \gamma_1 t_5,$$

where $t_5 = 4$ is the time at the fifth measurement occasion. The focus had true value $1 + (-0.1 \times 4) = 0.6$.

### Results

For $N = 10$, 51 models out of 2000 failed to converge (4 mean structures times 500 datasets). No other complications arose.

It has been drawn to my attention (Cunen et al. 2018, p.14), that not truncating bias squared at zero (for both FIC and (5.10)) leads to nicer results computationally, in spite of the conceptual issue of a negative estimate of a squared term. This certainly proved true for these simulations. Therefore, it was thought sensible to compare the MSE estimates, (5.10), but whose bias squared had not been truncated at zero, with the same estimates (that is, also not truncating bias squared) but subsequently adjusted by formula (5.18). To see this, compare Tables 5.3, 5.2, and 5.4, which show the average MSE estimates (over all 500 datasets) of formula (5.10) without truncation, with truncation, and the true MSE values for each mean structure and for each value of $N$ respectively. In particular, for smaller $N$ and narrower mean structures formula (5.10) with truncation of bias squared, overestimates the true MSE.[3]

With regard to comparison with accounting for uncertainty in the variance components, note that the estimates of Table 5.3, especially as $N$ grows, are close to the true MSEs shown in Table 5.4. Thus, it is only really for smaller $N$ that there is potential room for improvement. However, as shown in Table 5.5, adjusting for uncertainty in $\boldsymbol{\theta}$ via (5.18) did not make any improvement. Except for model M1, the estimates are poor for small $N$, which could be due to the use of the asymptotic variance-covariance matrix of the variance components in formula (5.18).

Table 5.2: MSE error estimates produced from formula (5.10) truncating bias squared at zero.

| $N$ | M1 | M2 | M3 | M4 |
|---:|---:|---:|---:|---:|
| 10 | 0.4976 | 0.7610 | 0.6112 | 0.8042 |
| 20 | 0.2819 | 0.3351 | 0.2875 | 0.3382 |
| 30 | 0.2142 | 0.2237 | 0.1955 | 0.2247 |
| 100 | 0.1179 | 0.0666 | 0.0674 | 0.0671 |
| 250 | 0.0917 | 0.0266 | 0.0321 | 0.0267 |

[3] In practice, when working with a single dataset, it may still be desirable to truncate bias squared, but in the simulation setting this is less so.

Table 5.3: MSE error estimates produced from formula (5.10) without truncating bias squared at zero.

| $N$ | M1 | M2 | M3 | M4 |
|---:|---:|---:|---:|---:|
| 10 | 0.2233 | 0.6957 | 0.4065 | 0.8042 |
| 20 | 0.1948 | 0.3320 | 0.2379 | 0.3382 |
| 30 | 0.1644 | 0.2228 | 0.1650 | 0.2247 |
| 100 | 0.1100 | 0.0661 | 0.0603 | 0.0671 |
| 250 | 0.0906 | 0.0266 | 0.0307 | 0.0267 |

Table 5.4: The true MSE errors: averages (over 500 simulations) of the observed squared errors $(\hat{\mu}_{\mathrm{M},N,j} - \mu_{N,\mathrm{true}})^2$, for simulation $j$, mean structure M, and sample size $N$.

| $N$ | M1 | M2 | M3 | M4 |
|---:|---:|---:|---:|---:|
| 10 | 0.3588 | 0.6322 | 0.4178 | 0.7451 |
| 20 | 0.1948 | 0.3138 | 0.2304 | 0.3179 |
| 30 | 0.1635 | 0.2195 | 0.1685 | 0.2194 |
| 100 | 0.1104 | 0.0690 | 0.0634 | 0.0696 |
| 250 | 0.0899 | 0.0270 | 0.0307 | 0.0270 |

Table 5.5: The average estimates produced by the adjusted MSE which accounts for uncertainty in variance components by formula (5.18).

| $N$ | M1 | M2 | M3 | M4 |
|---:|---:|---:|---:|---:|
| 10 | 0.2523 | 5.7666 | 8.7159 | 33.2241 |
| 20 | 0.2042 | 0.7902 | 0.8867 | 2.2227 |
| 30 | 0.1693 | 0.3614 | 0.3911 | 0.7086 |
| 100 | 0.1109 | 0.0774 | 0.0721 | 0.0845 |
| 250 | 0.0909 | 0.0296 | 0.0319 | 0.0289 |

## 5.4 Mean squared prediction error formulae

In some situations, the focus may not be to estimate a true value, but to predict future behaviour of a random variable. Suppose that there are a list of LME models with the same random effect structure, but different fixed effect structures. Which model produces the best predictor of the multivariate response of single cluster, or for a collection of clusters from the current dataset?

To this end, define the focus as the multivariate response of cluster $i$,

$$\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i, \tag{5.19}$$

which arises from the LME model

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim N(\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i, \sigma^2\boldsymbol{I}_{n_i}). \tag{5.20}$$

Note that the focus is considered to be a random variable, and until realised, does not have a true value.

With variance-covariance parameters assumed known, the predictor from the widest model (which is assumed true) is

$$\hat{\boldsymbol{\mu}}_i^o = \boldsymbol{X}_i\hat{\boldsymbol{\alpha}}^o + \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i^o,$$

with $\hat{\boldsymbol{\alpha}}_S^o$ as in (5.2), and $\hat{\boldsymbol{b}}_i^o$ as in (2.21) but with known variance components. In addition, model $S$ produces predictor

$$\hat{\boldsymbol{\mu}}_{S,i}^o = \boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}_S^o + \boldsymbol{Z}_i\hat{\boldsymbol{b}}_{S,i}^o,$$

where

$$\hat{\boldsymbol{b}}_{S,i}^o = \boldsymbol{D}^o\boldsymbol{Z}_i^\intercal\boldsymbol{\Sigma}_i^{o-1}(\boldsymbol{y}_i - \boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}_S^o). \tag{5.21}$$

As a measure of precision in prediction of $\boldsymbol{\mu}_i$, the mean squared prediction error (MSPE) of model $S$ is given by

$$\text{MSPE}_{S,i} = \mathbb{E}[(\hat{\boldsymbol{\mu}}_{S,i}^o - \boldsymbol{\mu}_i)^\intercal(\hat{\boldsymbol{\mu}}_{S,i}^o - \boldsymbol{\mu}_i)]. \tag{5.22}$$

For which, an estimator of the MSPE in prediction of $\boldsymbol{\mu}_i$ by model $S$, assuming known variance components, is

$$\begin{aligned}\widehat{\text{MSPE}}_{S,i}^o =& \text{Tr}\{\boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}_{S,i}^\intercal + \boldsymbol{Z}_i\text{Cov}(\hat{\boldsymbol{b}}_{S,i}^o)\boldsymbol{Z}_i^\intercal\} + \hat{\boldsymbol{\alpha}}^{o\intercal}\boldsymbol{W}_{S,i}^{o\intercal}\boldsymbol{W}_{S,i}^o\hat{\boldsymbol{\alpha}}^o \\ &- 2\Big[\text{Tr}\{\boldsymbol{Z}_i\boldsymbol{V}_i^o\boldsymbol{Z}_i\boldsymbol{D}^o\boldsymbol{Z}_i^\intercal + (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\boldsymbol{V}_i^o)\boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}_{S,i}^\intercal\boldsymbol{V}_i^{o\intercal}\boldsymbol{Z}_i^\intercal\} \\ &+ \hat{\boldsymbol{\alpha}}^{o\intercal}\boldsymbol{W}_{S,i}^{o\intercal}\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}^o\Big] + \text{Tr}\{\boldsymbol{Z}_i\boldsymbol{D}^o\boldsymbol{Z}_i^\intercal\} + \hat{\boldsymbol{\alpha}}^{o\intercal}\boldsymbol{X}_i^\intercal\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}^o,\end{aligned} \tag{5.23}$$

where

$$\boldsymbol{W}_{S,i}^o = \boldsymbol{X}_{S,i}\boldsymbol{A}_S^o + \boldsymbol{Z}_i\boldsymbol{V}_i^o(\boldsymbol{X}_i - \boldsymbol{X}_{S,i}\boldsymbol{A}_S^o);$$

$\text{Cov}(\hat{\boldsymbol{b}}^{o}_{S,i})$ is as in (C.3); $\boldsymbol{\alpha}$ is estimated by the wide model; and for which the derivation, largely inspired by previous methods of this chapter, is given in Appendix C.

The above formula could, in theory, be used for ranking models in terms of prediction of the multivariate response of a given individual, or for ranking models in terms of predictions of the collection of responses of a group of individuals: the MSPE estimated for each of them, and then averaged accordingly. In practice, as with formulae (5.10) and (5.15), the variance components must be estimated, and so estimating $\boldsymbol{\theta}$ from the wide model may be, in general, acceptable.

## 5.5 Chapter summary

In this chapter, mean squared error formulas have been derived for foci that are either univariate or multivariate linear combinations of the regression parameters. This contrasts with Chapter 4, where linearity of the focus in the regression coefficients was not required. Note also that, for formulas (5.10), (5.15), and indeed (5.23) an identity link function is essential. The form of the wide model's first two moments must also be correctly specified. However, normality is not necessary. By derivation, the mean squared error formulas (5.10) and (5.15) are exact for the variance term and unbiased for the bias squared term (provided no truncating at zero takes place) when the variance components are known. However, in practice these are to be estimated, and in doing so from the wide model one arrives at the same estimates as produced by FIC. In Section 5.3, the uncertainty in the variance components was taken into consideration. A simulation study showed that the suggested formulas provided no gain when the asymptotic variance-covariance of the variance components was utilised. Lastly, a mean squared prediction error formula was introduced, with potential applications for model selection between LME models when interest is in predicting responses of clusters from the current dataset.

# Chapter 6

# Summary and Further Topics

## 6.1 Summary of thesis

This thesis began with a detailed overview of linear models for longitudinal data. That is, in Chapter 2, the well-established theory of LMs and LME models was presented in the context of longitudinal data. Chapter 3 explained the theory and principal ideas behind GLMMs. For both Chapter 2 and Chapter 3, clinical trial datasets were used as illustrative examples. In Chapter 4, the focussed information criterion as in (Claeskens & Hjort 2003, Claeskens et al. 2008) was introduced, and the main steps underlying the limiting distribution theory was exhibited. Building upon this theory, a multivariate slightly misspecified framework was put forward which permits application of the FIC for selection of covariates in the multivariate LM, LME models, and GLMMs. A simulation study was then conducted and showed that, even if in theory the limiting distribution of the regression coefficients is independent of the uncertainty in the variance components, in practice it is still worthwhile to acknowledge the uncertainty. The rest of Chapter 4 was dedicated to illustrations of this FIC setup for multivariate models. In particular, examples of FIC and AFIC were given for the multivariate LM and a logistic GLMM. In addition, expressions for the relevant information matrices were derived. Chapter 5 presented alternative formulae for estimating the MSE of foci that are a linear combination of the regression coefficients in the context of multivariate linear models. An extension to multivariate foci was also suggested. Then, even if perhaps of limited practical value, an approximation that accounts for uncertainty in the variance components, making use of Kackar & Harville (1984), was put forward. Lastly, a mean squared prediction error formula for selection of fixed effects in LME models was proposed.

## 6.2 Further topics

The framework introduced in Section 4.2.1 assumes that any variance-covariance parameters or scaling parameters are protected. However, this assumption could potentially be relaxed. Random effect structures are typically nested in the FIC sense, whereby setting a parameter (or parameters) equal to a null value gives a simplified structure (e.g. setting $d_{12} = d_{22} = 0$ in the variance-covariance matrix of a random intercept and slope gives a random intercept model). The covariance pattern models of Section 2.3.1 are not necessarily nested, though may be in some cases, for an example see Claeskens et al. (2008, p.259). However, since variance-covariance parameters and scale parameters are constrained to be non-negative, the issue of asymptotic normality about border parameters arises. For the Normal model, due to asymptotic independence of regression coefficients and variance-covariance parameters, as discussed in Section 4.3.1, asymptotic normality of variance-covariance parameters is not required if the focus is purely a function of the regression parameters. So treating variance-covariance parameters as unprotected is acceptable for the Normal model if the focus is purely a function of the regression parameters. However, in general, this is not so. In addition, if choice between variance-covariance is to be considered, model robust inference would be required, as the FIC may not favour models with correctly-specified variances.

The FIC of Cunen et al. (2018) for LME models is not within the slightly misspecified framework, and so avoids the issue of asymptotic normality of variance-covariance parameters about *null* values. This makes choice between covariates and different random effect structures possible even for foci that are functions of the variance-covariance parameters, provided that the *true* values of the variance-covariance parameters are not on the border of their parameter space. The asymptotic distribution of ML estimators with boundary restrictions is given in Claeskens et al. (2008, p.278) and could offer an avenue of future research for the FIC.

The multivariate MSE error formulas of Section 5.2 are not restricted to the Normal model. In particular, it is possible to assess the MSE for multivariate foci via the FIC, where the multivariate delta method would be required. This could have applications for multivariate hypothesis testing for both univariate and multivariate models, and for predicting multivariate responses of non-Normal multivariate data. In particular, after application of the multivariate delta method, the limiting distribution of a $k$-dimensional multivariate focus as estimated by model $S$ for independent data, can be written [similar to (4.11)] as:

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_0) \xrightarrow{d} N_k(\boldsymbol{\omega}^{\mathsf{T}}(\boldsymbol{I} - \boldsymbol{G}_S)\boldsymbol{\delta}, \boldsymbol{\tau}_0^2 + \boldsymbol{\omega}^{\mathsf{T}}\boldsymbol{G}_S\boldsymbol{Q}\boldsymbol{G}_S^{\mathsf{T}}\boldsymbol{\omega}),$$

where

$$\boldsymbol{\omega} = \boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}\Big(\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\nu}}\Big)^{\mathsf{T}} - \Big(\frac{\partial\boldsymbol{\mu}}{\partial\gamma}\Big)^{\mathsf{T}}$$

is of dimension $q \times k$, and similarly

$$\boldsymbol{\tau}_0^2 = \Big(\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\nu}}\Big)\boldsymbol{J}_{00}^{-1}\Big(\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\nu}}\Big)^{\mathsf{T}}$$

is now a matrix of dimension $k \times k$.

Model averaging is a topic that has not been covered in this thesis. So it is perhaps hypocritical to say that, in my view, the "quiet scandal" of statistics remains a scandal as long as remedies to resolve it do not become part of mainstream practice (Breiman 1992). Nevertheless, for statisticians to remain confident that the confidence in their confidence intervals is as confident as claimed, continued application of estimation post model selection is problematic (see Claeskens et al. (2008, p.199) for more details). Model averaging approaches are a method to account for loss of confidence after model selection (Claeskens et al. 2008, Ch. 7). With regard to this, there are as yet (to my knowledge) no smoothed frequentist model averaging weights that account for more than one stage of model selection (as illustrated in Section 4.3.3 for the LM, or say for separate stages of choice of link, random effect structure and covariates in a GLMM).

# Appendices

# Appendix A

## A.1 Results on derivatives of the determinant and inverse of a matrix

From Section 15.9 of Harville (1997), provided sufficient smoothness of a matrix $\Sigma(\boldsymbol{\theta})$ in the domain of its parameters $\boldsymbol{\theta}$, the first and second derivatives of the inverse of $\Sigma(\boldsymbol{\theta})$ are given by

$$\frac{\partial \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})}{\partial \theta_k} = -\Sigma^{-1}(\boldsymbol{\theta}) \frac{\partial \Sigma(\boldsymbol{\theta})}{\partial \theta_k} \Sigma^{-1}(\boldsymbol{\theta}), \tag{A.1}$$

and

$$\frac{\partial^2 \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})}{\partial \theta_l \partial \theta_k} = \\ \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \left( \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_l} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_k} + \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_k} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_l} - \frac{\partial^2 \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_l \partial \theta_k} \right) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \tag{A.2}$$

respectively. Furthermore, the first derivatives of the determinant and the logarithm of the determinant of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are given by

$$\frac{\partial |\boldsymbol{\Sigma}(\boldsymbol{\theta})|}{\partial \theta_k} = |\boldsymbol{\Sigma}(\boldsymbol{\theta})| \operatorname{Tr} \left\{ \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_k} \right\}, \tag{A.3}$$

and

$$\frac{\partial \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})|}{\partial \theta_k} = \operatorname{Tr} \left\{ \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_k} \right\} \tag{A.4}$$

respectively.

## A.2 Expectation of a quadratic form of mean-zero variables

Let $\boldsymbol{\epsilon}$ be a mean zero random vector with covariance matrix $\boldsymbol{\Sigma}$ and let $\boldsymbol{A}$ be a generic matrix, then (Searle 1971, p.56):

$$\mathbb{E}[\boldsymbol{\epsilon}^{\intercal} \boldsymbol{A} \boldsymbol{\epsilon}] = \operatorname{Tr}\{\boldsymbol{A}\boldsymbol{\Sigma}\}. \tag{A.5}$$

## A.3 Derivation of $\boldsymbol{J}_{\boldsymbol{\theta\theta},N}$

To find the expression for the $(k, l)$th element of $\boldsymbol{J}_{\boldsymbol{\theta\theta},N}$, (4.24), substitute equation (A.2) into the second term of (4.21). Then sum (4.21) over all clusters, multiply by minus one and take the expectation to get

$$\mathbb{E}\left[\sum_{i=1}^{N}\frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{\Sigma}_i(\boldsymbol{\theta})^{-1}\frac{\partial^2\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}+\frac{\partial\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\right\}+\frac{1}{2}\boldsymbol{\epsilon}_i^{\intercal}\boldsymbol{A}\boldsymbol{\epsilon}_i\right], \tag{A.6}$$

where

$$\boldsymbol{A}=\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\left(\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}+\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}-\frac{\partial^2\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}\right)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}).$$

Since that which is inside the trace of (A.6) is considered fixed (even if unknown), the trace can move outside of the expected value to give

$$\sum_{i=1}^{N}\frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial^2\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}+\frac{\partial\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\right\}+\mathbb{E}\left[\frac{1}{2}\boldsymbol{\epsilon}_i^{\intercal}\boldsymbol{A}\boldsymbol{\epsilon}_i\right]. \tag{A.7}$$

By application of Equation (A.5) the expected value of $\frac{1}{2}\boldsymbol{\epsilon}_i^{\intercal}\boldsymbol{A}\boldsymbol{\epsilon}_i$ is

$$\begin{aligned}\frac{1}{2}\operatorname{Tr}\Bigg\{&\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\\&+\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}-\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial^2\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}\Bigg\}.\end{aligned}$$

Applying Equation (A.1) to the first two terms of this gives

$$\mathbb{E}[\frac{1}{2}\boldsymbol{\epsilon}_i^{\intercal}\boldsymbol{A}\boldsymbol{\epsilon}_i]=\frac{1}{2}\operatorname{Tr}\left\{-\frac{\partial\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_k}-\frac{\partial\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}-\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})\frac{\partial^2\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}\right\}. \tag{A.8}$$

Trace being a linear operator, the first and third terms of (A.8) cancel with the trace of (A.7). What remains is

$$\sum_{i=1}^{N}\frac{1}{2}\operatorname{Tr}\left\{-\frac{\partial\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial\boldsymbol{\Sigma}_i(\boldsymbol{\theta})}{\partial\theta_l}\right\},$$

and by applying (A.1) once more, we get (4.24).

## A.4 Derivatives of $\boldsymbol{\Sigma}_1(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_0(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$

The $(i, j)th$ element of the partial derivative of $\boldsymbol{\Sigma}_1(\boldsymbol{\theta})$ with respect to the parameter that regulates the exponential growth of variances of the endogenous response, $\delta_1$; the parameter that regulates the exponential growth of variances of the non-endogenous response, $\delta_0$; the range of exponentially decaying correlations $R$; and the nugget effect $\kappa$ are

$$\Big(\frac{\partial \boldsymbol{\Sigma}_1(\boldsymbol{\theta})}{\partial \delta_1}\Big)_{i,j} = \begin{cases} 2\sigma^2 t_j \sigma^2 e^{2t_j\delta_1}, & \text{if} \quad i = j, \\ \sigma^2(1-\kappa)(t_i+t_j)e^{(t_i+t_j)\delta_1 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j, \end{cases}$$

$$\Big(\frac{\partial \boldsymbol{\Sigma}_1(\boldsymbol{\theta})}{\partial \delta_0}\Big)_{i,j} = 0 \qquad \text{for all } i \text{ and } j,$$

$$\Big(\frac{\partial \boldsymbol{\Sigma}_1(\boldsymbol{\theta})}{\partial R}\Big)_{i,j} = \begin{cases} 0, & \text{if} \quad i = j, \\ \frac{\sigma^2(1-\kappa)|t_j-t_i|}{R^2} e^{(t_i+t_j)\delta_1 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j, \end{cases}$$

$$\Big(\frac{\partial \boldsymbol{\Sigma}_1(\boldsymbol{\theta})}{\partial \kappa}\Big)_{i,j} = \begin{cases} 0, & \text{if} \quad i = j, \\ -\sigma^2 e^{(t_i+t_j)\delta_1 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j, \end{cases}$$

respectively. Similarly, for $\boldsymbol{\Sigma}_0$ we have

$$\Big(\frac{\partial \boldsymbol{\Sigma}_0(\boldsymbol{\theta})}{\partial \delta_1}\Big)_{i,j} = 0 \qquad \text{for all } i \text{ and } j;$$

$$\Big(\frac{\partial \boldsymbol{\Sigma}_0(\boldsymbol{\theta})}{\partial \delta_0}\Big)_{i,j} = \begin{cases} 2\sigma^2 t_j e^{2t_j\delta_0}, & \text{if} \quad i = j, \\ \sigma^2(1-\kappa)(t_i+t_j)e^{(t_i+t_j)\delta_0 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j; \end{cases}$$

$$\Big(\frac{\partial \boldsymbol{\Sigma}_0(\boldsymbol{\theta})}{\partial R}\Big)_{i,j} = \begin{cases} 0, & \text{if} \quad i = j, \\ \frac{\sigma^2(1-\kappa)|t_i-t_j|}{R^2} e^{(t_i+t_j)\delta_0 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j; \end{cases}$$

$$\Big(\frac{\partial \boldsymbol{\Sigma}_0(\boldsymbol{\theta})}{\partial \kappa}\Big)_{i,j} = \begin{cases} 0, & \text{if} \quad i = j, \\ -\sigma^2 e^{(t_i-t_j)\delta_0 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j; \end{cases}$$

and the derivative of $\boldsymbol{\Sigma}_0(\boldsymbol{\theta})$ with respect to the scaling parameter $\sigma$ is,

$$\Big(\frac{\partial \boldsymbol{\Sigma}_0(\boldsymbol{\theta})}{\partial \sigma}\Big)_{i,j} = \begin{cases} 2\sigma e^{2t_j\delta_0}, & \text{if} \quad i = j, \\ 2(1-\kappa)\sigma e^{(t_i+t_j)\delta_0 - \frac{|t_i-t_j|}{R}} & \text{if} \quad i \neq j. \end{cases}$$

## A.5 Result (5.18)

For the second equality in (5.18), note that, for $\hat{\mu}^2_{\text{diff},S}$ as a function of $\boldsymbol{\theta}$,

$$\frac{\partial \hat{\mu}^2_{\text{diff},S}(\boldsymbol{\theta})}{\partial \theta_k} = 2(\hat{\mu}_S(\boldsymbol{\theta}) - \hat{\mu}_S(\boldsymbol{\theta}^o))\frac{\partial \hat{\mu}_S(\boldsymbol{\theta})}{\partial \theta_k},$$

so,

$$\frac{\partial^2 \hat{\mu}^2_{\text{diff},S}(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_k} = 2\Big(\frac{\partial\hat{\mu}_S(\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial\hat{\mu}_S(\boldsymbol{\theta})}{\partial\theta_l} + (\hat{\mu}_S(\boldsymbol{\theta}) - \hat{\mu}_S(\boldsymbol{\theta}^o))\frac{\partial^2\hat{\mu}_S(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_k}\Big)$$

which equates to $2\frac{\partial\hat{\mu}_S(\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial\hat{\mu}_S(\boldsymbol{\theta})}{\partial\theta_l}\Big|_{\theta_0}$ when evaluated at $\boldsymbol{\theta}^o$. Hence, (5.18) follows.

# Appendix B

## B.1 Figures from the simulations of Section 4.3.2

Figure B.1: In this figure showing a grid of box plots in $4 \times 4 = 16$ cells, each column of the grid represents a specific size for $n$, and each row a specific $N$. In each individual cell, the y-axis gives the values of FIC$/N$ for the mean structure M2; the values of the x-axis $(0, 1, 2, 3)$ correspond to the indices of the different variance-covariance matrices, $\boldsymbol{\Sigma}$.

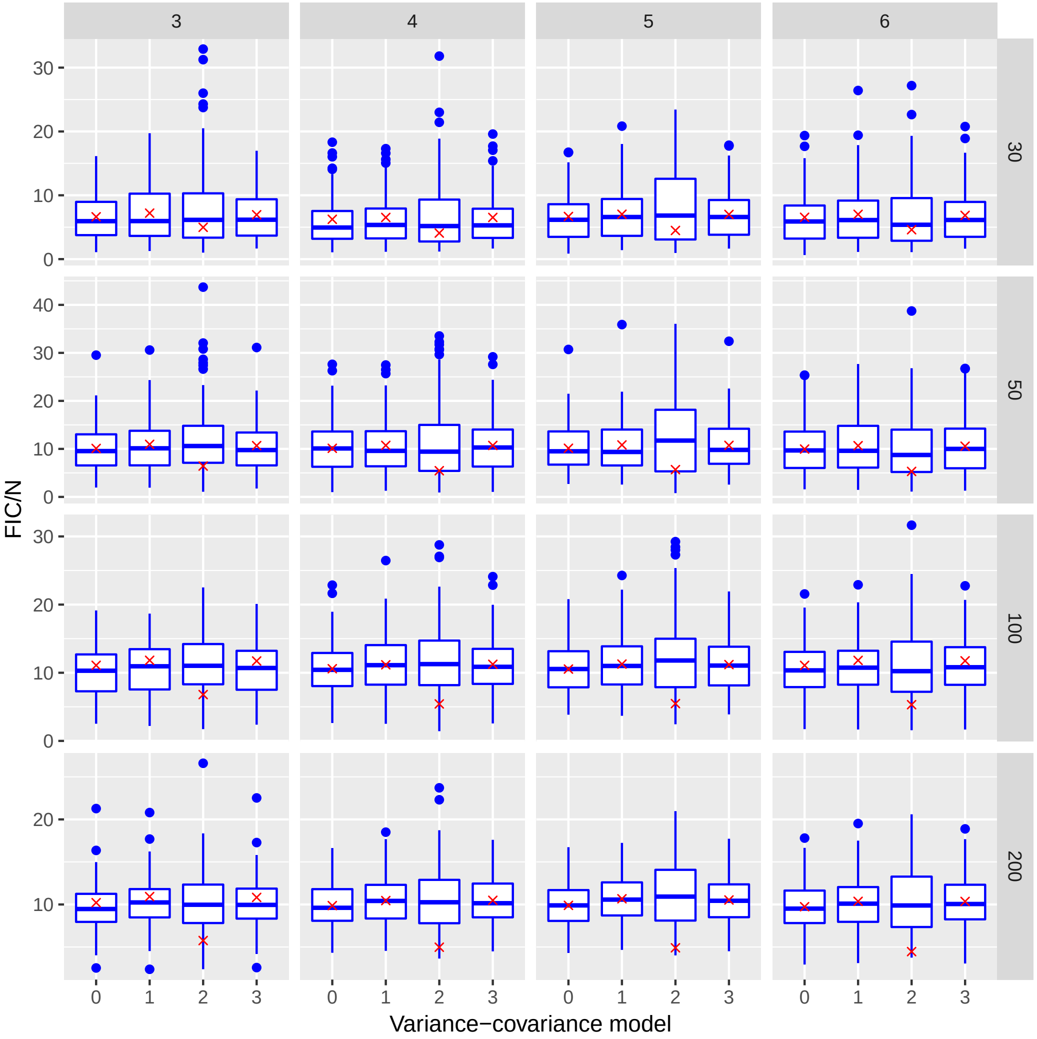


Figure B.2: The FIC$/N$ of mean structure M4 for each situation, plotted for different covariance matrices $\mathbf{\Sigma}$.

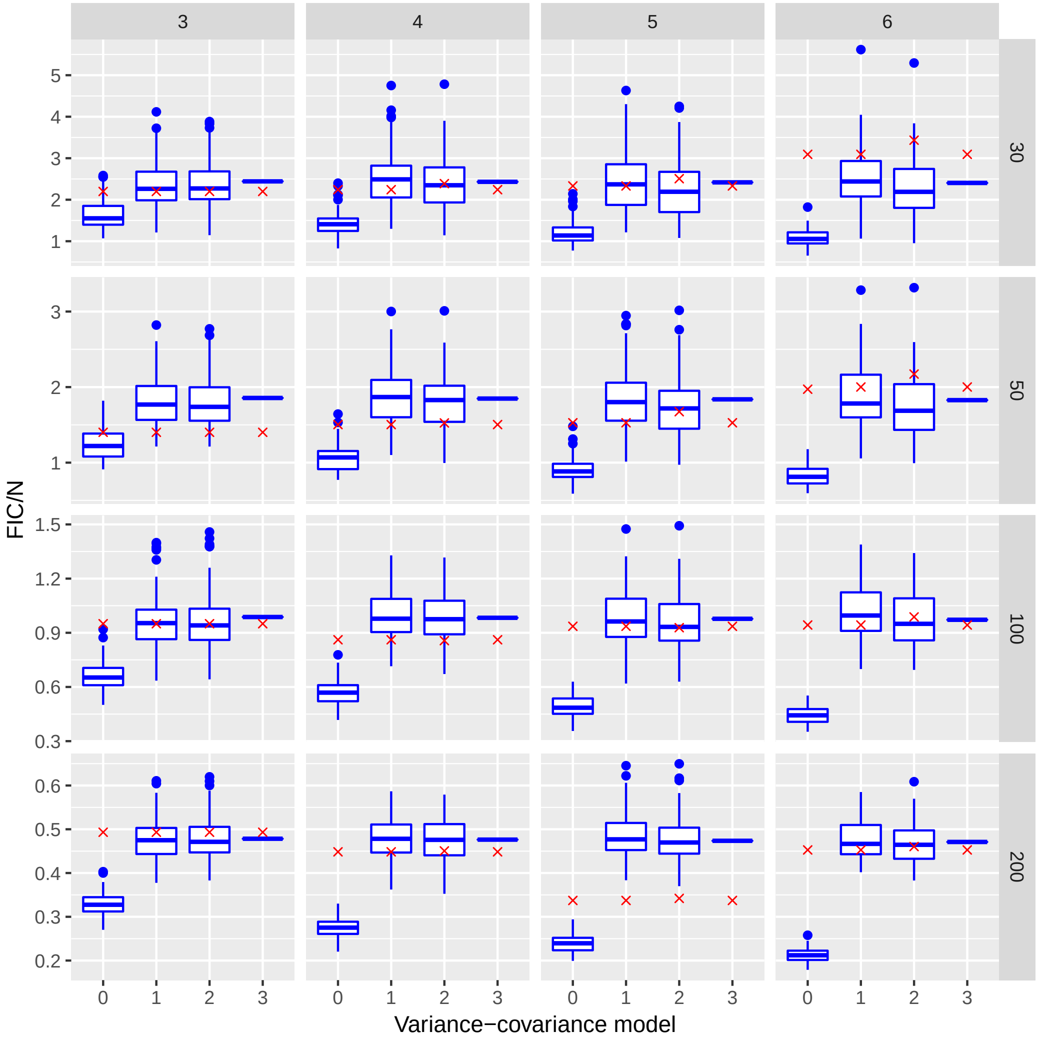


Figure B.3: In this figure showing a grid of box plots in $4 \times 4 = 16$ cells, each column represents a specific size of $n$, and each row a specific $N$. In each individual cell, the estimated bias squared term for the narrow mean structure M1 is plotted on the y-axis; the values of the x-axis $(0, 1, 2, 3)$ correspond to the indices of the different variance-covariance matrices, $\mathbf{\Sigma}$. The red crosses signify the approximately true bias squared in estimation.

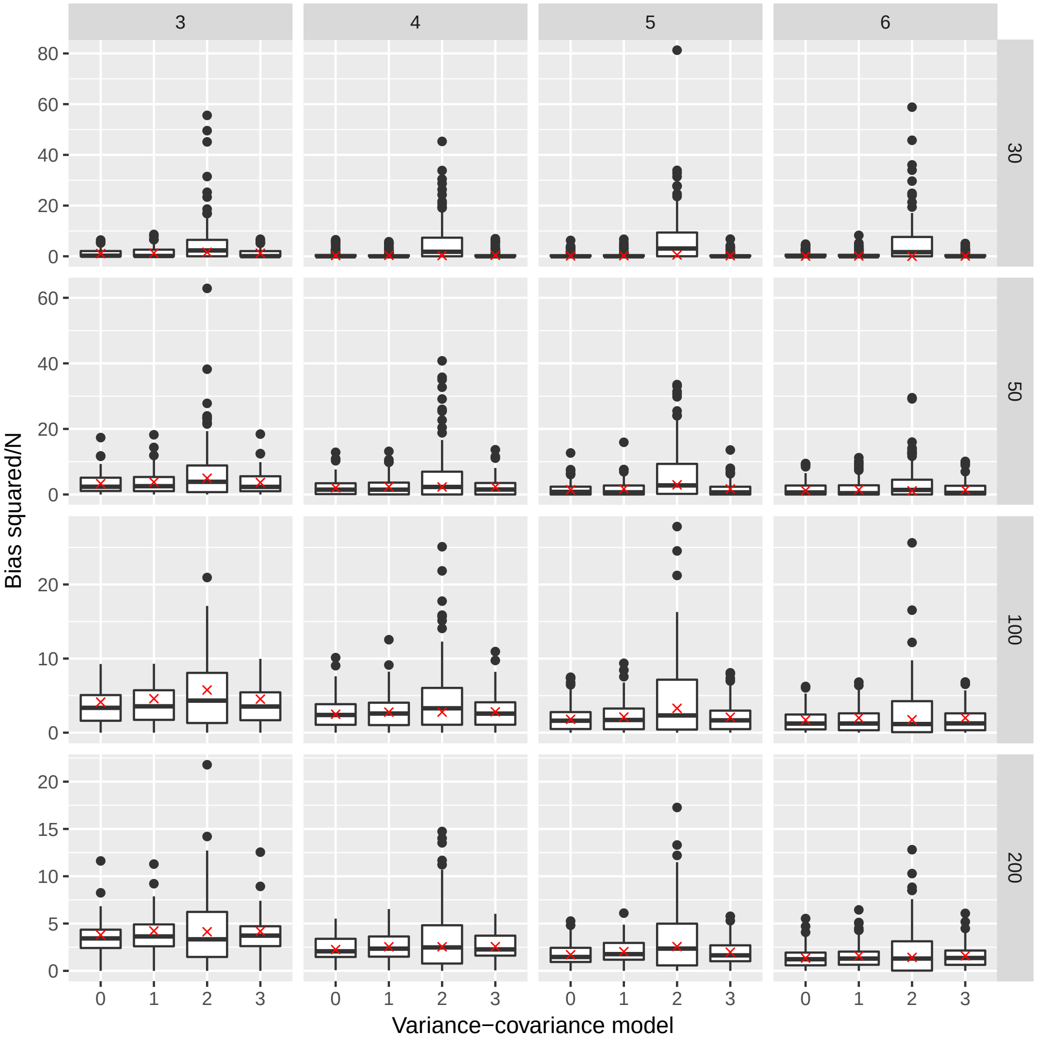


Figure B.4: Estimated bias squared term for the mean structure M2 are plotted against covariance matrices $\boldsymbol{\Sigma}$.

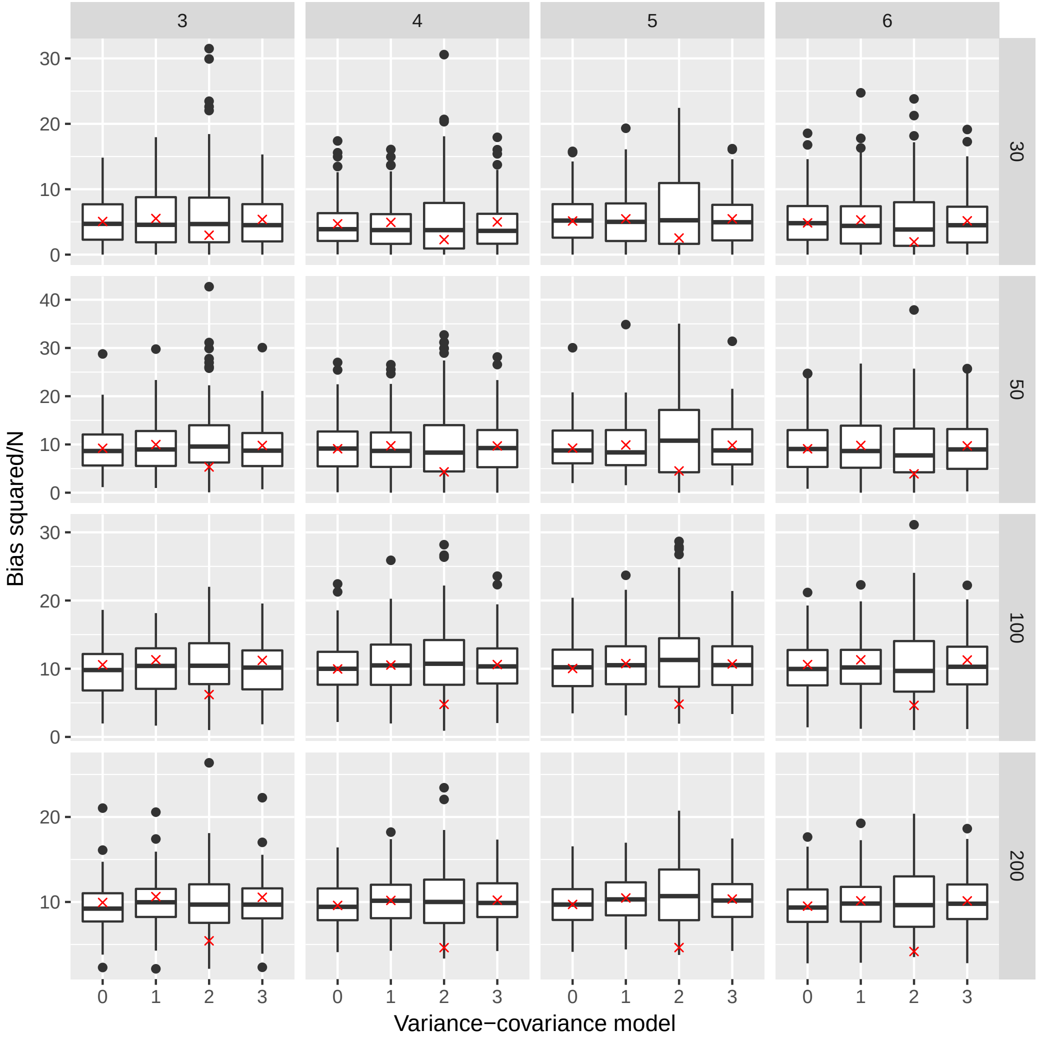


Figure B.5: Estimated bias squared term for the mean structure M3 plotted against covariance matrices $\Sigma$.

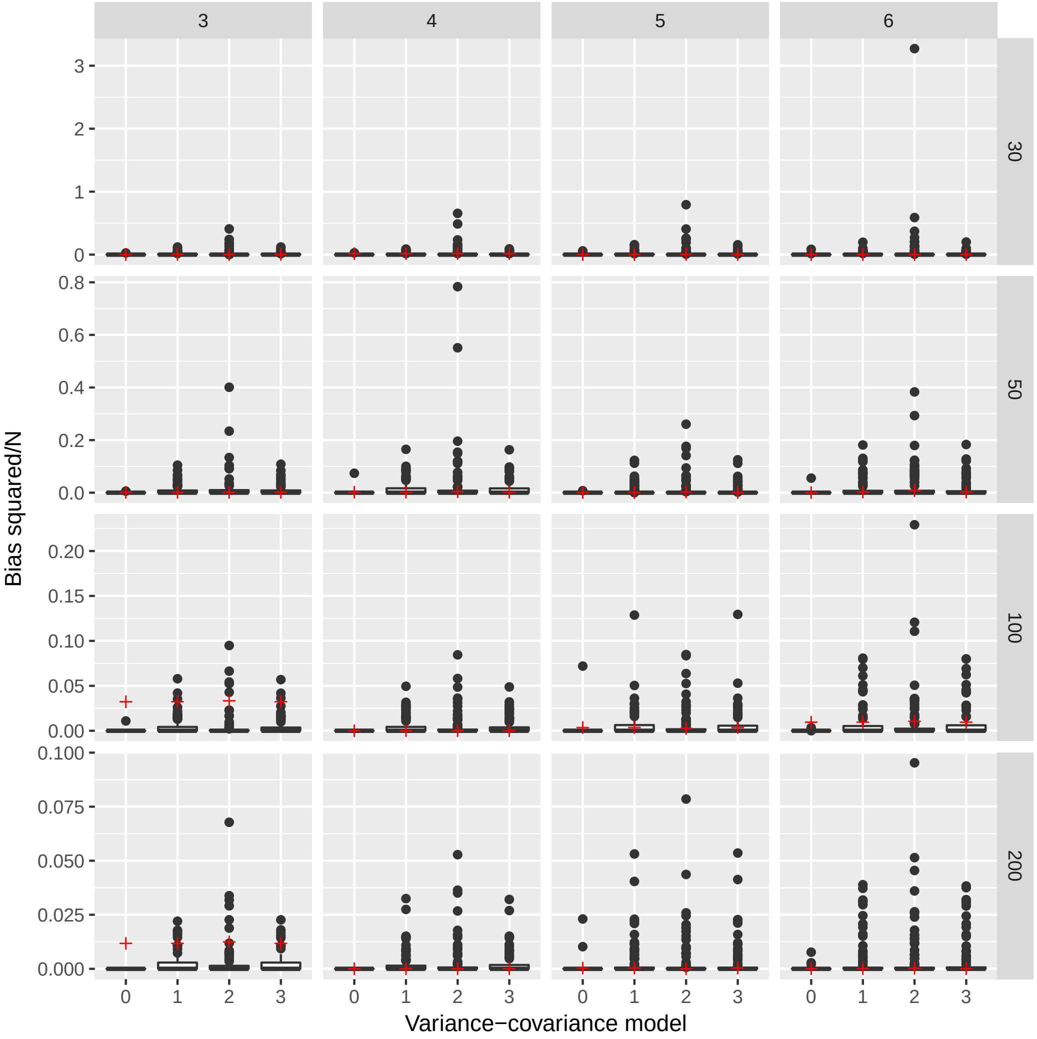


Figure B.6: Estimated bias squared term for the mean structure M4 plotted against covariance matrices $\Sigma$.

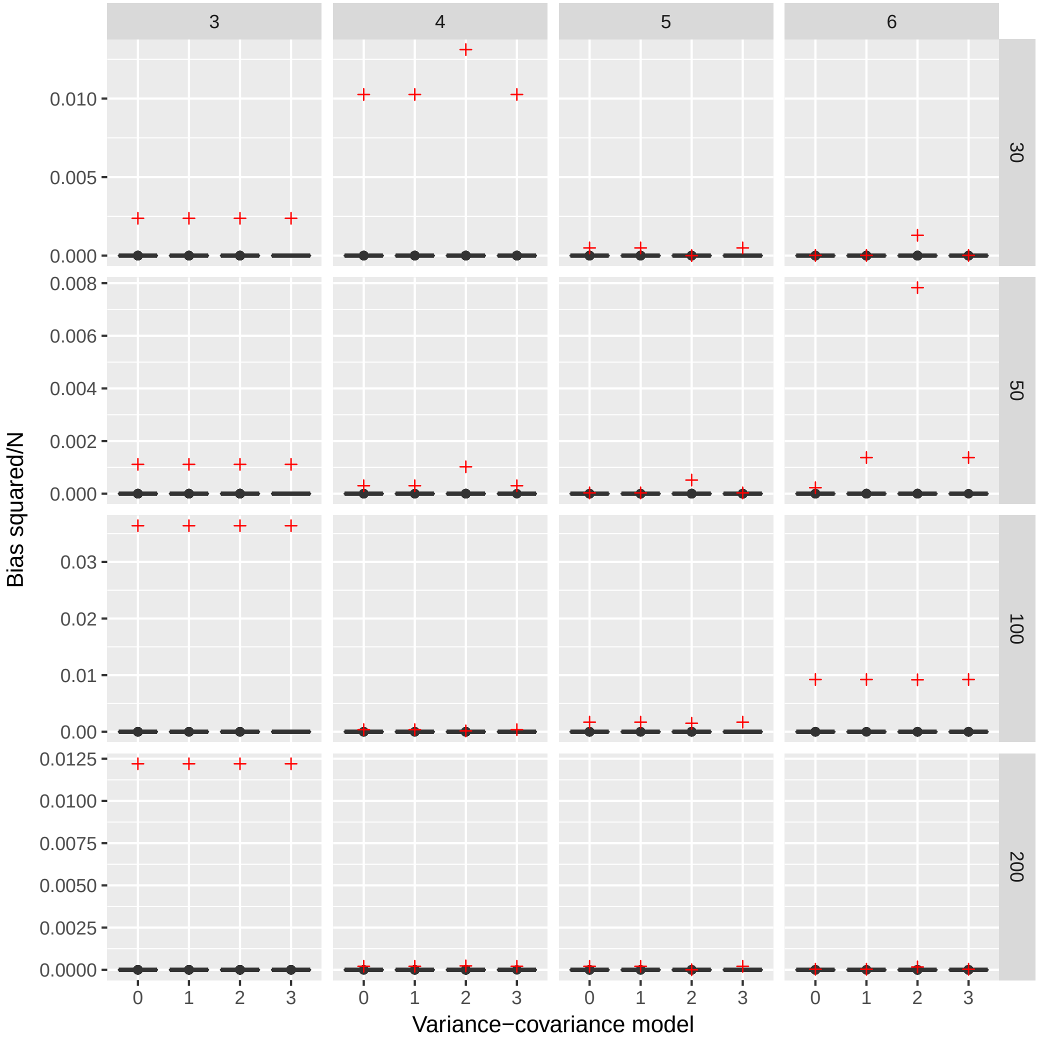


Figure B.7: In this figure showing a grid of box plots in $4 \times 4 = 16$ cells, each column represents a specific size for $n$, and each row a specific $N$. In each individual cell, the estimated variance term for the narrow mean structure M1 is plotted on the y-axis; the values of the x-axis $(0, 1, 2, 3)$ correspond to the indices of the different variance-covariance matrices, $\mathbf{\Sigma}$. The red crosses signify the approximately true variance in estimation.

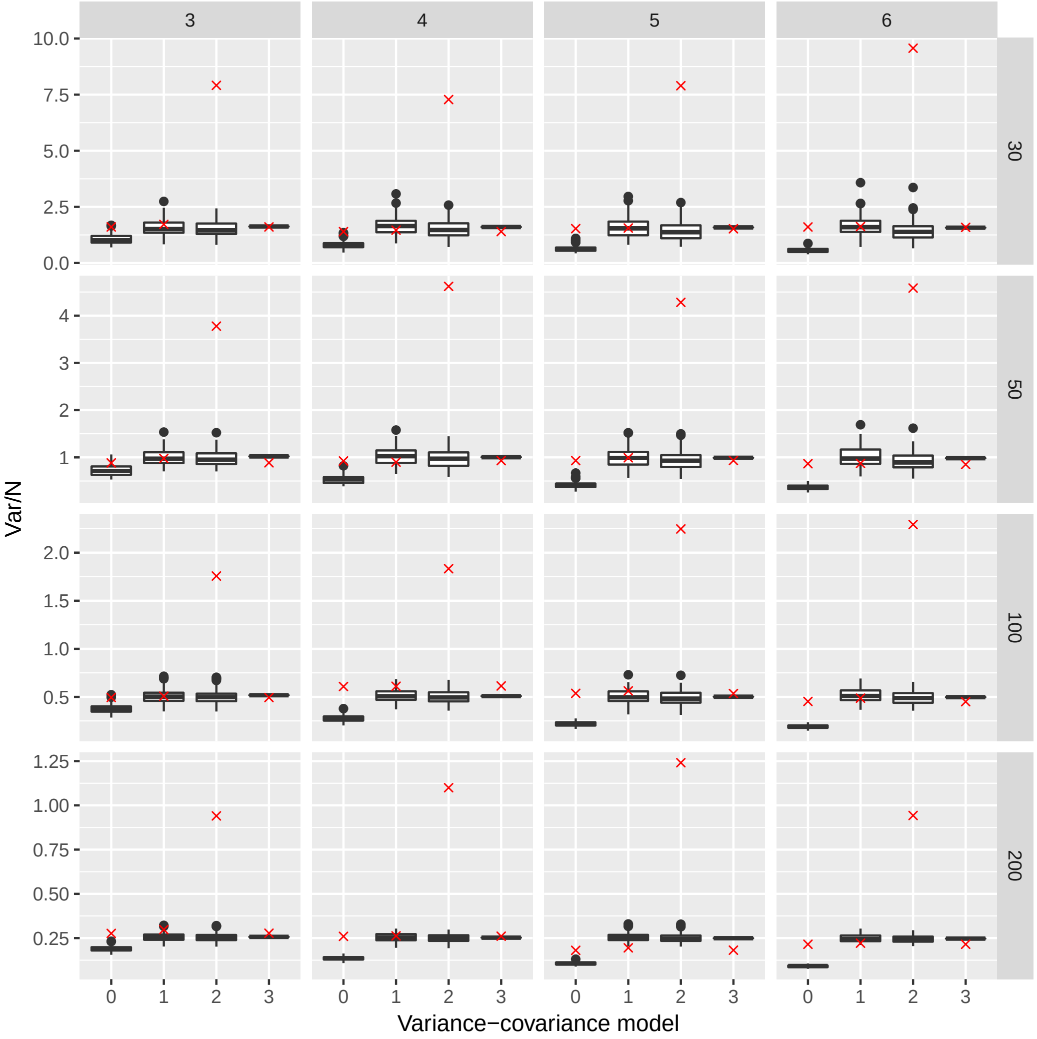


Figure B.8: Estimated variances for the mean structure M2 are plotted against covariance matrices $\Sigma$.

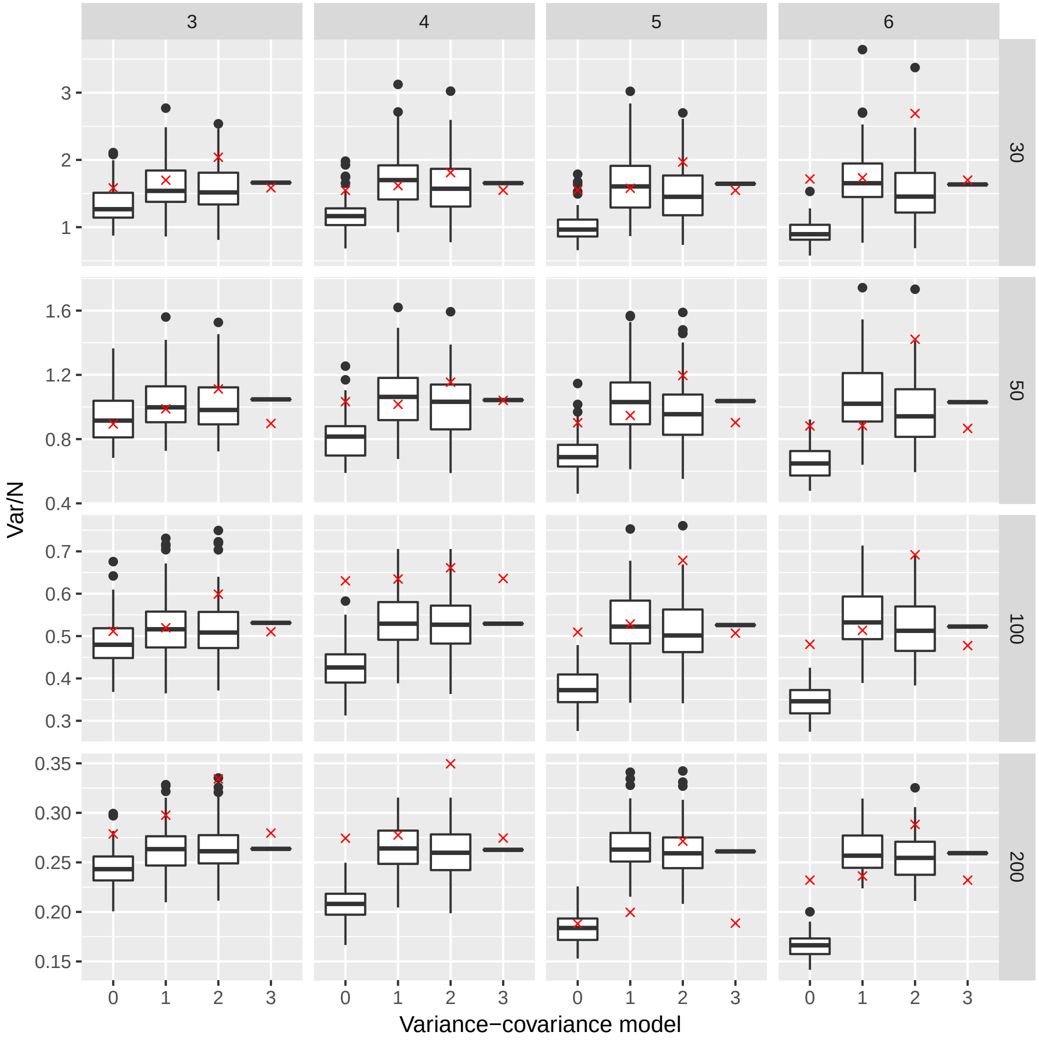


Figure B.9: Estimated variances for the mean structure M3 are plotted against covariance matrices $\Sigma$.

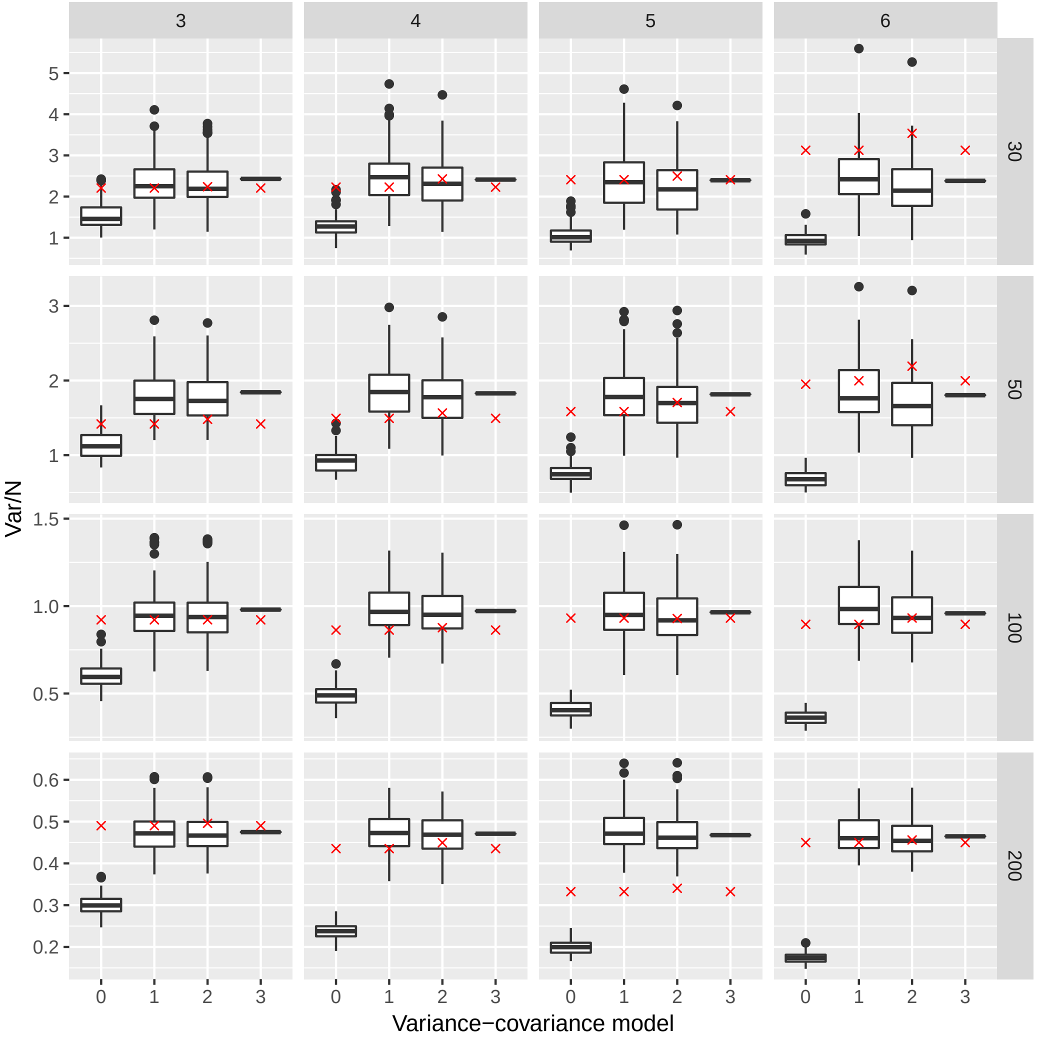
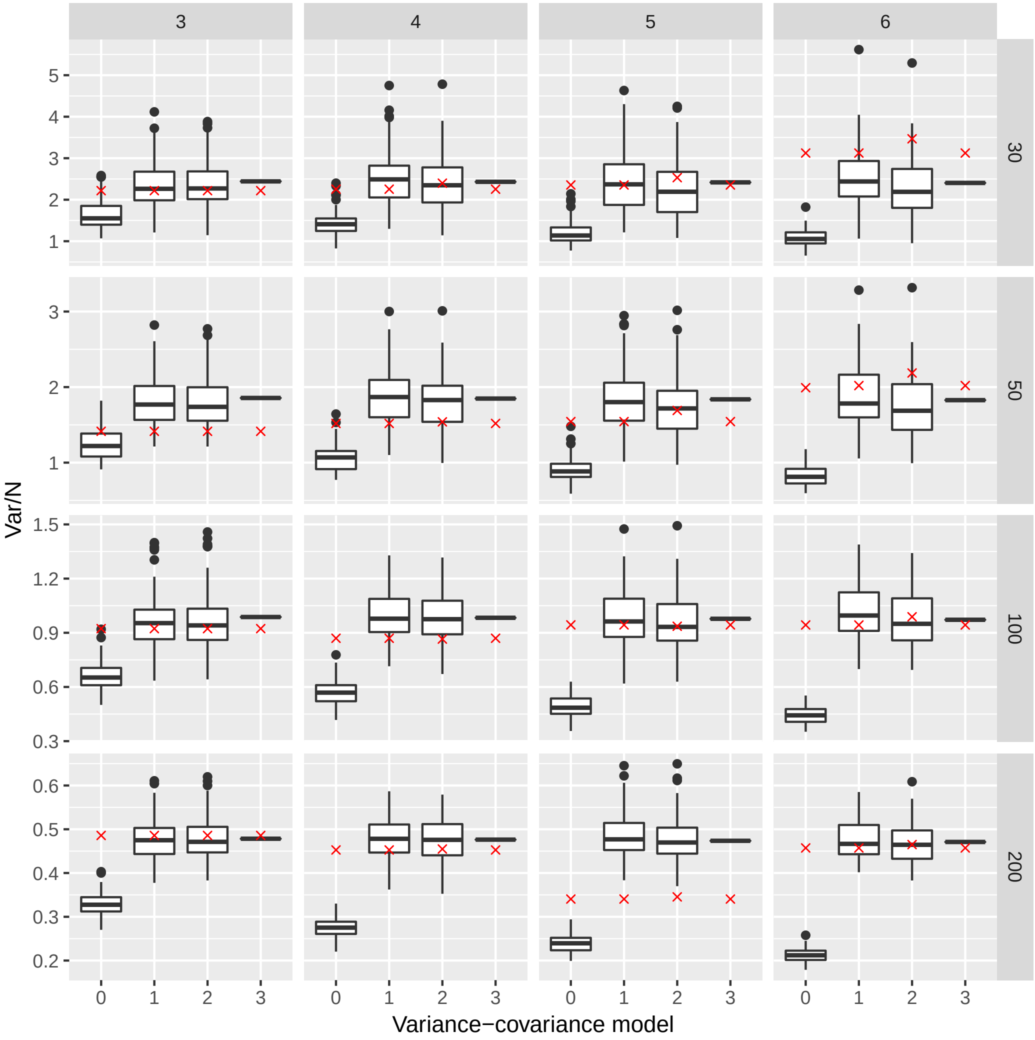


Figure B.10: Estimated variances for the mean structure M4 are plotted against covariance matrices $\Sigma$.



## B.2 Tables from the simulations of Section 4.3.2

| | $\mathbf{\Sigma}$ | 1 | | 2 | | 3 | | 4 | |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | $n$ \| M | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 30 | 3 | 2.3 (1.69) | 6.5 (3.63) | 3 (2.09) | 6.9 (4.19) | 7.4 (9.92) | 8.1 (6.30) | 2.8 (1.77) | 6.7 (3.81) |
| | 4 | 1.5 (1.41) | 6 (3.80) | 2.3 (1.36) | 6.2 (3.74) | 7.1 (8.60) | 6.9 (5.49) | 2.3 (1.49) | 6.3 (4.03) |
| | 5 | 1.1 (1.04) | 6.5 (3.68) | 2.1 (1.22) | 6.9 (3.94) | 8.7 (11.26) | 8.2 (5.90) | 2.1 (1.10) | 6.9 (3.90) |
| | 6 | 1.2 (1.12) | 6.4 (4.10) | 2.4 (1.46) | 7.2 (4.85) | 7.3 (10.24) | 7 (5.48) | 2.2 (1.14) | 6.9 (4.33) |
| 50 | 3 | 4 (3.12) | 10 (4.85) | 4.6 (3.51) | 11 (5.36) | 7.9 (9.45) | 12 (7.45) | 4.4 (3.33) | 10 (5.14) |
| | 4 | 2.8 (2.59) | 10 (5.19) | 3.4 (2.79) | 11 (5.60) | 6.8 (8.75) | 11 (8.25) | 3.3 (2.79) | 11 (5.53) |
| | 5 | 2.2 (2.32) | 10 (5.16) | 2.8 (2.68) | 11 (5.72) | 7.6 (8.75) | 13 (9.04) | 2.7 (2.48) | 11 (5.50) |
| | 6 | 2.1 (2.29) | 10 (5.56) | 2.8 (2.63) | 11 (6.24) | 4.6 (5.51) | 11 (7.29) | 2.7 (2.44) | 11 (5.98) |
| 100 | 3 | 3.9 (2.22) | 10 (3.59) | 4.3 (2.45) | 11 (3.88) | 6 (4.87) | 11 (4.68) | 4.2 (2.40) | 10 (3.83) |
| | 4 | 3 (2.10) | 11 (4.00) | 3.4 (2.31) | 11 (4.31) | 5 (4.81) | 12 (5.52) | 3.4 (2.28) | 11 (4.25) |
| | 5 | 2.3 (1.86) | 11 (3.86) | 2.7 (2.13) | 11 (4.23) | 5.1 (5.52) | 12 (6.01) | 2.7 (2.03) | 11 (4.11) |
| | 6 | 1.8 (1.48) | 11 (3.75) | 2.2 (1.66) | 11 (4.07) | 3.3 (4.06) | 11 (5.48) | 2.2 (1.62) | 11 (3.99) |
| 200 | 3 | 3.7 (1.80) | 9.6 (2.90) | 4.1 (1.94) | 10 (3.07) | 4.6 (3.75) | 10 (3.89) | 4 (1.95) | 10 (3.09) |
| | 4 | 2.5 (1.38) | 10 (2.63) | 2.9 (1.57) | 11 (2.94) | 3.7 (3.43) | 11 (3.85) | 2.9 (1.51) | 11 (2.80) |
| | 5 | 1.9 (1.15) | 10 (2.56) | 2.3 (1.26) | 11 (2.70) | 3.7 (3.62) | 11 (3.97) | 2.2 (1.27) | 11 (2.73) |
| | 6 | 1.5 (1.15) | 9.8 (2.87) | 1.8 (1.31) | 10 (3.11) | 2.3 (2.53) | 10 (3.72) | 1.8 (1.27) | 10 (3.05) |

Table B.1: The mean FIC scores (over all 100 simulations) divided by $N$ as estimates of the MSE for mean structures M1 and M2 and each situation $(N, n, \mathbf{\Sigma})$ are shown along with the corresponding standard deviation in brackets.

| | $\mathbf{\Sigma}$ | 1 | | 2 | | 3 | | 4 | |
|---|---|---|---|---|---|---|---|---|---|
| $N$ | $n$ \| M | 3 | 4 | 3 | 4 | 3 | 4 | 3 | 4 |
| 30 | 3 | 1.5 (0.33) | 1.6 (0.35) | 2.4 (0.61) | 2.4 (0.61) | 2.4 (0.62) | 2.4 (0.61) | 2.4 (0.02) | 2.4 (0.00) |
| | 4 | 1.3 (0.27) | 1.4 (0.30) | 2.5 (0.66) | 2.5 (0.66) | 2.4 (0.66) | 2.4 (0.67) | 2.4 (0.02) | 2.4 (0.00) |
| | 5 | 1.1 (0.23) | 1.2 (0.27) | 2.4 (0.71) | 2.4 (0.71) | 2.2 (0.69) | 2.3 (0.70) | 2.4 (0.03) | 2.4 (0.00) |
| | 6 | .96 (0.18) | 1.1 (0.21) | 2.5 (0.71) | 2.5 (0.71) | 2.3 (0.78) | 2.3 (0.71) | 2.4 (0.03) | 2.4 (0.00) |
| 50 | 3 | 1.1 (0.18) | 1.2 (0.20) | 1.8 (0.34) | 1.8 (0.34) | 1.8 (0.35) | 1.8 (0.35) | 1.9 (0.02) | 1.9 (0.00) |
| | 4 | .93 (0.15) | 1.1 (0.18) | 1.9 (0.36) | 1.9 (0.36) | 1.8 (0.38) | 1.8 (0.37) | 1.8 (0.03) | 1.8 (0.00) |
| | 5 | .76 (0.13) | .9 (0.15) | 1.8 (0.41) | 1.8 (0.41) | 1.7 (0.40) | 1.7 (0.41) | 1.8 (0.02) | 1.8 (0.00) |
| | 6 | .68 (0.11) | .83 (0.13) | 1.9 (0.42) | 1.9 (0.41) | 1.7 (0.41) | 1.8 (0.42) | 1.8 (0.03) | 1.8 (0.00) |
| 100 | 3 | .61 (0.07) | .67 (0.08) | .97 (0.15) | .97 (0.15) | .96 (0.15) | .97 (0.16) | .98 (0.01) | .99 (0.00) |
| | 4 | .49 (0.05) | .57 (0.06) | .99 (0.13) | .99 (0.13) | .97 (0.14) | .98 (0.14) | .98 (0.01) | .98 (0.00) |
| | 5 | .41 (0.04) | .49 (0.06) | .98 (0.15) | .99 (0.15) | .95 (0.15) | .96 (0.15) | .97 (0.02) | .98 (0.00) |
| | 6 | .36 (0.04) | .44 (0.05) | .99 (0.15) | 1 (0.14) | .96 (0.15) | .97 (0.15) | .97 (0.02) | .97 (0.00) |
| 200 | 3 | .3 (0.03) | .33 (0.03) | .48 (0.05) | .48 (0.05) | .48 (0.05) | .48 (0.05) | .48 (0.00) | .48 (0.00) |
| | 4 | .24 (0.02) | .28 (0.02) | .48 (0.05) | .48 (0.05) | .47 (0.05) | .47 (0.05) | .47 (0.01) | .48 (0.00) |
| | 5 | .2 (0.02) | .24 (0.02) | .48 (0.05) | .48 (0.05) | .47 (0.05) | .48 (0.05) | .47 (0.01) | .47 (0.00) |
| | 6 | .17 (0.01) | .21 (0.01) | .47 (0.04) | .48 (0.04) | .47 (0.04) | .47 (0.04) | .47 (0.01) | .47 (0.00) |

Table B.2: The mean FIC scores (over all 100 simulations) divided by $N$ as estimates of the MSE for M3 and M4 and each situation $(N, n, \mathbf{\Sigma})$ are shown along with the corresponding standard deviation in brackets.

# Appendix C

# Derivation of the MSPE formula (5.23)

The mean squared error in prediction (5.22), unlike the MSE, is not expressible as a sum of variances and squared biases. It can, however, be written as

$$\begin{aligned}
&\mathbb{E}[\hat{\boldsymbol{\mu}}^{o\intercal}_{S,i}\hat{\boldsymbol{\mu}}^{o}_{S,i} - 2\hat{\boldsymbol{\mu}}^{o\intercal}_{S,i}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^{\intercal}\boldsymbol{\mu}_i] \\
=&\mathbb{E}[\hat{\boldsymbol{\mu}}^{o\intercal}_{S,i}\hat{\boldsymbol{\mu}}^{o}_{S,i}] - \mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}]^{\intercal}\mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}] + \mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}]^{\intercal}\mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}] - 2\mathbb{E}[\hat{\boldsymbol{\mu}}^{o\intercal}_{S,i}\boldsymbol{\mu}_i] + \mathbb{E}[\boldsymbol{\mu}_i^{\intercal}\boldsymbol{\mu}_i] \\
=& \operatorname{Tr}\{\mathrm{Cov}(\hat{\boldsymbol{\mu}}^{o}_{S,i})\} + \mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}]^{\intercal}\mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}] - 2\mathbb{E}[\hat{\boldsymbol{\mu}}^{o\intercal}_{S,i}\boldsymbol{\mu}_i] + \mathbb{E}[\boldsymbol{\mu}_i^{\intercal}\boldsymbol{\mu}_i], \qquad \text{(C.1)}
\end{aligned}$$

where the first two terms on the right hand side of the first equality have been expressed as the sum of the uncertainties in each entry of the predictor. In the following, expressions will be derived for the above four terms. And, thereby, formula (5.23), for comparison of LME models (with different fixed effects) in terms of mean squared prediction error, is arrived at.

For the uncertainty in the predictor, we have

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\mu}}^{o}_{S,i}) = \mathrm{Cov}(\boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}^{o}_{S} + \boldsymbol{Z}_i\hat{\boldsymbol{b}}^{o}_{S,i}) &= \boldsymbol{X}_{S,i}\mathrm{Cov}(\hat{\boldsymbol{\alpha}}^{o}_{S})\boldsymbol{X}^{\intercal}_{S,i} + \boldsymbol{Z}_i\mathrm{Cov}(\hat{\boldsymbol{b}}^{o}_{S,i})\boldsymbol{Z}_i^{\intercal} \\
&= \boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^{\intercal}_{S,i} + \boldsymbol{Z}_i\mathrm{Cov}(\hat{\boldsymbol{b}}^{o}_{S,i})\boldsymbol{Z}_i^{\intercal}, \quad \text{(C.2)}
\end{aligned}$$

where the second equality follows since $\mathrm{Cov}(\hat{\boldsymbol{\alpha}}^{o}_{S}, \hat{\boldsymbol{b}}^{o}_{S,i}) = 0$ (Henderson 1975); the third equality by (5.8); and where, as in (2.22),

$$\mathrm{Cov}(\hat{\boldsymbol{b}}^{o}_{S,i}) = \boldsymbol{V}_i^{o}\left(\boldsymbol{\Sigma}_i^{o} - \boldsymbol{X}_{S,i}\left(\sum_{i=1}^{N}\boldsymbol{X}_{S,i}\boldsymbol{\Sigma}_i^{o-1}\boldsymbol{X}^{\intercal}_{S,i}\right)^{-1}\boldsymbol{X}^{\intercal}_{S,i}\right)\boldsymbol{V}_i^{o\intercal}, \qquad \text{(C.3)}$$

where $\boldsymbol{V}_i^{o} = \boldsymbol{D}^{o}\boldsymbol{Z}_i^{\intercal}\boldsymbol{\Sigma}_i^{o-1}$.

For the second term of (C.1), we will require $\mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}]$, which, by (5.3), (5.21), and assuming (5.20) is the true model, is equal to

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}^{o}_{S} + \boldsymbol{Z}_i\hat{\boldsymbol{b}}^{o}_{S,i}] &= \boldsymbol{X}_{S,i}\boldsymbol{A}^{o}_{S}\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{D}^{o}\boldsymbol{Z}_i^{\intercal}\boldsymbol{\Sigma}_i^{o-1}\mathbb{E}[\boldsymbol{y}_i - \boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}^{o}_{S}] \\
&= [\boldsymbol{X}_{S,i}\boldsymbol{A}^{o}_{S} + \boldsymbol{Z}_i\boldsymbol{V}_i^{o}(\boldsymbol{X}_i - \boldsymbol{X}_{S,i}\boldsymbol{A}^{o}_{S})]\boldsymbol{\alpha} =: \boldsymbol{W}^{o}_{S,i}\boldsymbol{\alpha}, \qquad \text{(C.4)}
\end{aligned}$$

where $\boldsymbol{A}^{o}_{S}$ is as in (5.5).

It is fruitful to express the cross-term in (C.1) as

$$\mathbb{E}[\hat{\boldsymbol{\mu}}^{o\intercal}_{S,i}\boldsymbol{\mu}_i] = \operatorname{Tr}\{\mathrm{Cov}(\hat{\boldsymbol{\mu}}^{o}_{S,i}, \boldsymbol{\mu}_i)\} + \mathbb{E}[\hat{\boldsymbol{\mu}}^{o}_{S,i}]^{\intercal}\mathbb{E}[\boldsymbol{\mu}_i], \qquad \text{(C.5)}$$

since expressions for $\text{Cov}(\hat{\boldsymbol{\mu}}^o_{S,i}, \boldsymbol{\mu}_i)$, $\mathbb{E}[\hat{\boldsymbol{\mu}}^o_{S,i}]$ and $\mathbb{E}[\boldsymbol{\mu}_i]$ are more easily obtained. In particular, we have that

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\mu}}^o_{S,i}, \boldsymbol{\mu}_i) &= \text{Cov}(\boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}^o_S + \boldsymbol{Z}_i\hat{\boldsymbol{b}}^o_{S,i}, \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i) \\ &= \text{Cov}(\boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}^o_S + \boldsymbol{Z}_i\hat{\boldsymbol{b}}^o_{S,i}, \boldsymbol{Z}_i\boldsymbol{b}_i) \\ &= \boldsymbol{X}_{S,i}\text{Cov}(\hat{\boldsymbol{\alpha}}^o_S, \boldsymbol{b}_i)\boldsymbol{Z}_i^\top + \boldsymbol{Z}_i\text{Cov}(\hat{\boldsymbol{b}}^o_{S,i}, \boldsymbol{b}_i)\boldsymbol{Z}_i^\top. \end{aligned} \tag{C.6}$$

For which, we have

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\alpha}}^o_S, \boldsymbol{b}_i) =& \text{Cov}\left(\boldsymbol{B}_S^{o-1}\sum_{j=1}^{N}\boldsymbol{X}^\top_{S,j}\boldsymbol{\Sigma}_j^{o-1}\boldsymbol{y}_j, \boldsymbol{b}_i\right) \\ =& \text{Cov}(\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{\Sigma}_i^{o-1}\boldsymbol{y}_i, \boldsymbol{b}_i) \\ =& \boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{\Sigma}_i^{o-1}\text{Cov}(\boldsymbol{y}_i, \boldsymbol{b}_i) \\ =& \boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{\Sigma}_i^{o-1}\text{Cov}(\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \boldsymbol{b}_i) \\ =& \boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{\Sigma}_i^{o-1}\boldsymbol{Z}_i\text{Cov}(\boldsymbol{b}_i) \\ =& \boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{\Sigma}_i^{o-1}\boldsymbol{Z}_i\boldsymbol{D}^o \\ =& \boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}, \end{aligned} \tag{C.7}$$

where the second equality follows from independence of $\boldsymbol{y}_j$ and $\boldsymbol{b}_i$ for $i \neq j$; and the fourth from the assumed true model (5.20), with $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$. Furthermore,

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{b}}^o_{S,i}, \boldsymbol{b}_i) =& \text{Cov}(\boldsymbol{V}_i^o(\boldsymbol{y}_i - \boldsymbol{X}_{S,i}\hat{\boldsymbol{\alpha}}^o_S), \boldsymbol{b}_i) \\ =& \boldsymbol{V}_i^o[\text{Cov}(\boldsymbol{y}_i, \boldsymbol{b}_i) - \boldsymbol{X}_{S,i}\text{Cov}(\hat{\boldsymbol{\alpha}}^o_S, \boldsymbol{b}_i)] \\ =& \boldsymbol{V}_i^o[\text{Cov}(\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \boldsymbol{b}_i) - \boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}] \\ =& \boldsymbol{V}_i^o[\boldsymbol{Z}_i\boldsymbol{D}^o - \boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}]. \end{aligned} \tag{C.8}$$

Therefore, by combining (C.6) with (C.7) and (C.8), we have

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\mu}}^o_{S,i}, \boldsymbol{\mu}_i) &= \boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}\boldsymbol{Z}_i^\top + \boldsymbol{Z}_i\boldsymbol{V}_i^o[\boldsymbol{Z}_i\boldsymbol{D}^o - \boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}]\boldsymbol{Z}_i^\top \\ &= \boldsymbol{Z}_i\boldsymbol{V}_i^o\boldsymbol{Z}_i\boldsymbol{D}^o\boldsymbol{Z}_i^\top + (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\boldsymbol{V}_i^o)\boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}\boldsymbol{Z}_i^\top. \end{aligned} \tag{C.9}$$

In addition,

$$\mathbb{E}[\boldsymbol{\mu}_i] = \mathbb{E}[\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i] = \boldsymbol{X}_i\boldsymbol{\alpha}. \tag{C.10}$$

So, by (C.5), (C.4), (C.10) and (C.9), the third term in (C.1) is expressible as minus twice

$$\text{Tr}\{\boldsymbol{Z}_i\boldsymbol{V}_i^o\boldsymbol{Z}_i\boldsymbol{D}^o\boldsymbol{Z}_i^\top + (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\boldsymbol{V}_i^o)\boldsymbol{X}_{S,i}\boldsymbol{B}_S^{o-1}\boldsymbol{X}^\top_{S,i}\boldsymbol{V}_i^{o\top}\boldsymbol{Z}_i^\top\} + \boldsymbol{\alpha}^\top\boldsymbol{W}^{o\top}_{S,i}\boldsymbol{X}_i\boldsymbol{\alpha}. \tag{C.11}$$

The final term in (C.1) is $\mathbb{E}[\boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i]$, and is equal to

$$\text{Tr}\{\text{Cov}(\boldsymbol{\mu}_i)\} + \mathbb{E}[\boldsymbol{\mu}_i]^\top\mathbb{E}[\boldsymbol{\mu}_i] = \text{Tr}\{\boldsymbol{Z}_i\boldsymbol{D}^o\boldsymbol{Z}_i^\top\} + \boldsymbol{\alpha}^\top\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{\alpha}. \tag{C.12}$$

Finally, taking the trace of (C.2) and summing with (C.12), (C.4) multiplied by its transpose, and minus twice (C.11) gives an expression for the true MSPE under the first two moments of the wide model, and is estimated by (5.23).

# Appendix D

## D.1 R code

The code for the data illustrations in this thesis, written using the software R (R Development Core Team 2008), were all of a similar form. The narrow and wide models are typically fit to the data first. The wide model provides estimates of the necessary quantities for the FIC. The narrow is dealt with separately, then all other models are fit within a loop and the FIC scores are simultaneously calculated. An example is given in Listing D.2. For the AFIC this procedure is repeated within a bigger loop for the different foci. Listing D.1 gives the data preparation for the Riesby depression dataset.

Listing D.1: Prepare Riesby depression dataset.

```
riesbynow <- read.table("http://hedeker.people.uic.edu/RIESBY.DAT.
    txt", nrows = 396, na.strings = ".")

colnames(riesbynow) <- c("id", "hd", "intcpt", "week", "endo", "
    interaction") #0='NonEndog' 1='Endog'
library(plyr)

searchcols <- c('hd') #delete individuals with missing data
riesby <-ddply(riesbynow, "id", function(x)if(any(is.na(x[,
    searchcols]))) NULL else x)
riesby$id2 <-rep(1:46,each=6)

#set baseline as covariate
base <- ddply(riesby, "id2", function(x){ rep(x[1,2], length(x[,2])
    )})
colnames(base)<- c("id2","y1", "y2", "y3", "y4", "y5", "y6")
base.now <- reshape(base, direction ="long", varying = c("y1", "y2"
    , "y3", "y4", "y5", "y6"), sep = "", idvar = "id2")
base.now <- base.now[order(base.now$id2),]
riesby$base <- base.now[,3]

riesby <- subset(riesby, week>0) #only data from treatment period

#c("nonendo", "endo") = c(0,1)
riesby$endo.f <- factor(riesby$endo , labels = c("Non-endo","Endo")
    )
riesby$week.f <- factor(riesby$week , labels = 0:4)

riesby$time <- riesby$week #center time
riesby$timec <- riesby$time - mean(riesby$time)
riesby$timesq <- riesby$timec^2
```

```
riesby$basec <- riesby$base - mean(riesby$base) #center baseline
```

Listing D.2: FIC for depression data set predictions.

```
library(nlme)

b0 <- 27- mean(riesby$base) #high scorer at baseline
t0 <- 2 #end of week 4 #centred time is 2
e0 <- 0 #non-endogeneous


#needed for estimate of focus
xu0xp0 <- c(1,b0, t0, e0, e0*t0, t0^2, e0*t0^2)

#narrow
narr.f <- formula(hd~ 1 + basec + timec)
narrow <- gls(narr.f,  correlation = corExp(form=~time|id2, nugget
   = T),
                    weights = varExp(form=~time|endo), #AIC
                        favoured
              method = "ML", data = riesby)
#wide
wide.f <- formula(hd~ 1 + basec + timec + endo  +  timesq + timec:
    endo + timesq:endo)
wide <- gls(wide.f,  correlation = corExp(form=~time|id2, nugget =
    T), weights = varExp(form=~time|endo), #AIC favoured
         data = riesby)   #By REML

Sigma.non <- getVarCov(wide, individual = "2") #non-endo
Sigma.endo <- getVarCov(wide, individual = "3") #endo

#protected and unprotected design matrices
XP <- cbind(1, riesby$basec, riesby$timec)
XU <- cbind(riesby$endo, riesby$timesq, riesby$timec*riesby$endo,
    riesby$timesq*riesby$endo)
XX <- cbind(XP, XU)  # wide design

NN <- 46      # number of individuals
nn <- 5       #each of 5 measurements
p1 <- 5       # 1 sigma, 2 variance, 1 corr, 1 nugget
p2 <- 3       #protected regression parameters
qq <- 4       #unprotected regression parameters
pp <- p1+p2
MM <- 2^qq    #number of models


#get estimate of Jhat
Jhat <- matrix(rep(NA, (pp+qq)*(pp+qq)), nrow = (pp+qq), ncol = (pp
    +qq))
Jreg <- NULL
sumnow <- list()
```

```
for(i in 1:NN){
  XXnow <- subset(XX, riesby$id2 == i)
  Sigmanow <- getVarCov(wide, individual = i)
  sumnow[[i]] <- t(XXnow) %*% solve(Sigmanow) %*% XXnow
}
Jreg <- (1/NN)*Reduce("+", sumnow)
Jhat[-(1:p1),-(1:p1)] <- Jreg

JBB <- Jhat[(p1+1):pp, (p1+1):pp]
JGB <- Jhat[(pp+1):(pp+qq), (p1+1):pp]
JGG <- Jhat[(pp+1):(pp+qq), (pp+1):(pp+qq)]
JBG <- t(JGB)

#partial derivatives
dmudbeta <- c(1, b0, t0)
dmudgamma <- c(e0,t0*e0,t0^2, (t0^2)*e0)

#FIC quantities
tau0sq <- t(dmudbeta) %*% solve(JBB) %*% dmudbeta
tau0 <- sqrt(tau0sq)
omega <- (JGB %*% solve(JBB) %*% dmudbeta) - dmudgamma
QQ <- solve(JGG - JGB %*% solve(JBB) %*% JBG)
Dn <- sqrt(NN)*(coef(wide)[(p2+1):(p2+qq)])
I <- diag(qq)

combinations = function(n) #as from lectures
{ comb = NULL
  {for (i in 1:n)
    comb = rbind(cbind(0,comb),cbind(1,comb))
    return(comb)} }
subsets <- combinations(qq)

#storage space
modout <- list() ; FIC <- NULL; AIC <- NULL bias <- NULL ;
bias.sq <- NULL ; bias.sq.t <- NULL ; var <- NULL ;
muhat <- NULL ; num.param <- NULL ; num.qq <- NULL

#narrow
bias[1]<- t(omega) %*% Dn
bias.sq[1] <- t(omega)%*% ( Dn %*% t(Dn) - QQ ) %*% omega
bias.sq.t[1] <- max(0, bias.sq[1])
var[1]<- tau0^2
FIC[1] <- bias.sq.t[1]+var[1]
AIC[1] <- -(AIC(narrow))
num.param[1] <- pp
num.qq[1] <- 0
muhat[1] <- t(xu0xp0)%*% c(coef(narrow), rep(0,qq))

for(k in 2:nrow(subsets)){ #for rest of models

  #define subset of unprotected params
  where <- (1:qq)[subsets[k, ] == 1]
  dims <- length(where)

  #fit models
  prot <- c("1", "basec", "timec")
  unprot <- c("endo", "timesq", "timec:endo", "timesq:endo")
  vec <- c(prot, unprot[subsets[k, ] == 1])

  meannow <- paste(vec, collapse = "+")
  formula <- as.formula(paste("hd ~", meannow, sep = ""))

  modout[[k]] <- gls(formula,
                     correlation = corExp(form=~time|id2, nugget = T),
                     weights = varExp(form=~time|endo), #AIC
                        favourite covariance
                     method = "ML", data = riesby)

  #FIC
  piS <- I[where, ] #projection matrices
  dim(piS) <- c(dims,qq)
  QS <- solve(piS %*% solve(QQ) %*% t(piS))
  GS <- t(piS) %*% QS %*% piS %*% solve(QQ)
  #
  bias.sq[k] <- t(omega) %*% (I - GS) %*% ( Dn %*% t(Dn) - QQ ) %*%
      t(I - GS) %*% omega
  bias.sq.t[k] <- max(0,bias.sq[k])
  bias[k] <- t(omega) %*% (I - GS) %*% Dn
  var[k] <- tau0^2 + t(omega) %*% GS %*% QQ %*% t(GS) %*% omega
  #
  FIC[k] <- bias.sq.t[k] + var[[k]]

  #focus estimate
  MLest <- 0*(1:(p2+qq))
  MLnow <- coef(modout[[k]])
  MLest[1:p2] <- MLnow[1:p2]
  MLest[p2+where] <- MLnow[(p2+1):length(MLnow)]
  muhat[k] <- t(xu0xp0) %*% MLest
  #
  AIC[k] <- -AIC(modout[[k]])
  num.param[k] <- pp + dims
  num.qq[k] <- dims
}
```

# Bibliography

Breiman, L. (1992), 'The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error', *Journal of the American Statistical Association* **87**(419), 738–754.

Broström, G. (2017), *glmmML: Generalized Linear Models with Clustering*. R package version 1.0.2.
**URL:** *https://CRAN.R-project.org/package=glmmML*

Bryk, A. S. & Raudenbush, S. W. (1992), *Hierarchical linear models: applications and data analysis methods*, Sage Publications.

Casella, G. & Berger, R. L. (2002), *Statistical inference*, Vol. 2, Duxbury Pacific Grove, CA.

Claeskens, G. & Hjort, N. L. (2003), 'The focused information criterion', *Journal of the American Statistical Association* **98**(464), 900–916.

Claeskens, G., Hjort, N. L. et al. (2008), *Model selection and model averaging*, Vol. 330, Cambridge University Press Cambridge.

Cressie, N. & Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, John Wiley & Sons.

Cunen, C., Walloe, L. & Hjort, N. L. (2017), 'Decline in energy storage in antarctic minke whales during the jarpa period: Assessment via the focussed information criterion (fic)'.

Cunen, C., Walloe, L. & Hjort, N. L. (2018), 'Focused model selection for linear mixed models, with an application to whale ecology'.

De Jong, P., Heller, G. Z. et al. (2008), *Generalized linear models for insurance data*, Vol. 10, Cambridge University Press Cambridge.

Demidenko, E. (2013), *Mixed models: theory and applications with R*, John Wiley & Sons.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004*a*), *Applied longitudinal analysis*, John Wiley & Sons.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004*b*), 'Applied longitudinal analysis toenail dataset'.
**URL:** *https://content.sph.harvard.edu/fitzmaur/ala2e/*

Galecki, A. & Burzykowski, T. (2013), *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*, Springer Science & Business Media.

Genz, A. & Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics, Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. & Hothorn, T. (2017), *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-6.
**URL:** *https://CRAN.R-project.org/package=mvtnorm*

Gregoire, T. G., Brillinger, D. R., Diggle, P., Russek-Cohen, E., Warren, W. G. & Wolfinger, R. D. (2012), *Modelling longitudinal and spatially correlated data*, Vol. 122, Springer Science & Business Media.

Greven, S. & Kneib, T. (2010), 'On the behaviour of marginal and conditional aic in linear mixed models', *Biometrika* **97**(4), 773–789.

Gumedze, F. & Dunne, T. (2011), 'Parameter estimation and inference in the linear mixed model', *Linear Algebra and its Applications* **435**(8), 1920 – 1944.

Hamilton, M. (1960), 'A Rating Scale for Depression', *Journal of Neurology, Neurosurgery & Psychiatry* **23**(1), 56–62.

Harville, D. A. (1997), *Matrix algebra from a statistician's perspective*, Vol. 1, Springer.

Hedeker, D. (2006), 'Riesby depression dataset'.
**URL:** *http://hedeker.people.uic.edu/long.html*

Hedeker, D. & Gibbons, R. D. (2006), *Longitudinal Data Analysis*, John Wiley & Sons.

Henderson, C. R. (1975), 'Best linear unbiased estimation and prediction under a selection model', *Biometrics* pp. 423–447.

Hjort, N. L. & Claeskens, G. (2003), 'Frequentist model average estimators', *Journal of the American Statistical Association* **98**(464), 879–899.

Hlavac, M. (2015), *stargazer: Well-Formatted Regression and Summary Statistics Tables*, Harvard University, Cambridge, USA. R package version 5.2.
**URL:** *http://CRAN.R-project.org/package=stargazer*

Hofert, M. & Mächler, M. (2016), 'Parallel and other simulations in R made easy: An end-to-end study', *Journal of Statistical Software* **69**(4), 1–44.

Kackar, R. N. & Harville, D. A. (1984), 'Approximations for standard errors of estimators of fixed and random effects in mixed linear models', *Journal of the American Statistical Association* **79**(388), 853–862.

Laird, N. M. & Ware, J. H. (1982), 'Random-Effects Models for Longitudinal Data', *Biometrics* **38**(4), 963–974.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Mardia, K. V. & Marshall, R. J. (1984), 'Maximum likelihood estimation of models for residual covariance in spatial regression', *Biometrika* **71**(1), 135–146.

McCulloch, C. E. & Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, Wiley Series in Probability and Statistics.

McNeish, D., Stapleton, L. M. & Silverman, R. D. (2017), 'On the unnecessary ubiquity of hierarchical linear modeling', *Psychological Methods* **22**(1), 114–140.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2016), *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-128.
**URL:** *http://CRAN.R-project.org/package=nlme*

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *http://www.R-project.org*

Rosenthal, R. & Rosnow, R. L. (1991), *Essentials of Behavioral Research: Methods and Data Analysis*, McGraw-Hill.

Saefken, B., Ruegamer, D., Kneib, T. & Greven, S. (2018*a*), *cAIC4: Conditional Akaike information criterion for lme4*.

Saefken, B., Ruegamer, D., Kneib, T. & Greven, S. (2018*b*), 'Conditional model selection in mixed-effects models with caic4', *ArXiv e-prints* .

Sarkar, D. (2008), *Lattice: Multivariate Data Visualization with R*, Springer, New York. ISBN 978-0-387-75968-5.
**URL:** *http://lmdvr.r-forge.r-project.org*

Searle, S. (1971), *Linear Models*, number v. 1 *in* 'Wiley Publication in Mathematical Statistics', John Wiley & Sons.

Smyth, G. K. (2005), 'Numerical integration', *Encyclopedia of Biostatistics* pp. 3088–3095.

Stroup, W. W. (2012), *Generalized linear mixed models: modern concepts, methods and applications*, CRC press.

Vaida, F. & Blanchard, S. (2005), 'Conditional akaike information for mixed-effects models', *Biometrika* **92**(2), 351–370.

Wand, M. (2007), 'Fisher information for generalised linear mixed models', *Journal of Multivariate Analysis* **98**(7), 1412–1416.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
**URL:** *http://ggplot2.org*

Wickham, H. (2011), 'The split-apply-combine strategy for data analysis', *Journal of Statistical Software* **40**(1), 1–29.
**URL:** *http://www.jstatsoft.org/v40/i01/*

Yang, H., Lin, P., Zou, G. & Liang, H. (2017), 'Variable selection and model averaging for longitudinal data incorporating gee approach', *Statistica Sinica* **27**(1), 389–413.

Zimmerman, M., Martinez, J. H., Young, D., Chelminski, I. & Dalrymple, K. (2013), 'Severity classification on the hamilton depression rating scale', *Journal of affective disorders* **150**(2), 384–388.