NTNU – Trondheim
Norwegian University of
Science and Technology

Statistical Modelling and Inference for Long Gene Expression Time Series

Karen Sofie Sollie Holt

Master of Science in Physics and Mathematics
Submission date: March 2014
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This thesis concludes my Master of Science in Applied Physics and Mathematics, with specialisation in Industrial Mathematics, at the Norwegian University of Science and Technology (NTNU). The work was carried out at the Department of Mathematical Sciences.

I would like to thank Astrid Lægreid and Torunn Bruland, for this opportunity to be part of their work.

I would also like to thank my supervisor Associate Professor Mette Langaas for excellent supervision and guidance. She has been a brilliant motivator and source of inspiration. I feel fortunate to have had the opportunity to work with such a wonderful person.

A special thank to my brother Jarl Christian for proofreading this thesis, and to Patrick, the rest of my family and close friends for support and encouragement.

Trondheim, March 2014
Karen Sofie Sollie Holt

Sammendrag

Denne oppgaven tar for seg statistisk modellering og analyse av lange genuttrykkstidsrekker ved bruk av lineære blandede (linear mixed effects) modeller. Denne type regresjonsmodell er mye brukt innen fagfelt som biologi, økologi og medisin.

Analysene er gjort på et datasett fra en mikromatrisestudie. Studien er gjennomført ved Institutt for kreftforskning og molekylær medisin (IKM) ved NTNU i 2009. Forskerne målte genuttrykk til to cellelinjer ved 12 ulike tidspunkt. Den ene cellelinjen ble stimulert med hormonet gastrin, mens den andre fungerer som en ustimulert kontroll. Forsøket ble gjort på to biologiske replikater. Datasettet består altså av to parvise tidsrekker av genuttrykk, med to biologiske replikater for hver av tidsrekkene.

Den lineære blandede modellen kan tilpasses hvert av genene i datasettet. Vi har i denne oppgaven undersøkt om arealet under den tilpassede kurven for tidsrekkene kan brukes som et mål på styrken av aktivert genuttrykk over tid. Dersom arealet kan brukes som et mål på denne effekten over tid, kan det brukes som en metode for å rangere gener i en genuttrykksstudie på. Ved bruk av hypotesetester kan signifikansen av reguleringen i genuttrykkene vurderes. Vi har i denne oppgaven foreslått å bruke en hypotesetest basert på arealet, fremfor de faste effektene (fixed effects) i modellen, for å gjøre signifikansvurderinger.

Vi har brukt både en parametrisk og en ikke-parametrisk tilnærming i denne oppgaven. Nye testobservatorer knyttet til hypotesetestene våre er foreslått. Vi beskriver en permuteringstest for å generere data under nullhypotesen om at arealet er lik null. Gjennom et lite simuleringseksperiment sjekker vi permuteringsalgoritmen vår. Vi gjør multippel testing av hypoteser på flere gener i datasettet vårt. Resultatene fra multippel testing ved bruk av både parametrisk og ikke-parametrisk metode blir så sammenlignet og evaluert.

Abstract

The main objective of this thesis is to model and analyse long gene expression time series from a microarray study using the linear mixed effects model. This regression model is widely used in the fields of biology, ecology and medicine. The linear mixed effects model combines both fixed and random effects on a linear scale.

We will use data from a microarray study conducted by Astrid Lægreid & Torunn Bruland and collaborators at Department of Cancer Research and Molecular Medicine (IKM) at Norwegian University of Science and Technology (NTNU) in 2009. The data set consists of paired time series, one gastrin stimulated treatment and one unstimulated control, for 8956 genes. The response value is a logarithmic measure of gene expression, and is measured for two biological replicates.

The linear mixed effects model can be fitted for each of the genes in the data set. We have examined if the area under the estimated time series curve may be used as a measure of strength of the gene expression activation over time, and if this area can be used to rank the genes with respect to effect size over time. Significant activation can be assessed with the aid of hypothesis tests. With the area as a measure of strength of the gene expression activation over time, we have suggested a hypothesis test for assessing gene significance.

Analyses will be performed based on parametric assumptions and on permutation. Test statistics related to the analyses are suggested. Our permutation strategy is validated through a small scale simulation study. Multiple testing of hypotheses are conducted. The parametric and permutation approach will be compared and evaluated using statistical inference.

Contents

Chapter 1

Introduction

Gene expression data from long time series can be analysed with the aid of linear mixed effects models, with a polynomial effect in time. The linear mixed effects model is a popular regression model for biological and medical analysis, due to the presence of repeated measures. The method combines, on a linear scale, both fixed and random effects.

In a microarray study in the order of tens of thousands of genes are studied. The biologists are looking for a way to rank the genes with respect to effect size of activation over time The aim of this thesis is to find such a ranking method and assess significance, based on parametric assumptions and on permutation, and to compare and evaluate the strategies using statistical inference.

1.1 Biological problem and the data set

Gene expression can be thought of as the result of activity of a genotype in an organism, and is measured by the mRNA production of a gene in a cell. Analysis of gene expression data is the analysis of this activity. In this thesis we will look at how the gastric hormone gastrin may affect the mRNA production of several genes.

Gastrin is a peptide hormone which is responsible for stimulating the secretion of gastric acid (HCl) and thus maintaining the appropriate acidic pH level in the stomach, Fjeldbo (2012). It is also important in the regulation of function and growth in the gastric muchosa. Prolonged elevated plasma gastrin levels (hypergastrinemia) have been associated with cancer in the gastrointestinal system.

The data used as basis for this thesis comes from a study conducted by Astrid Lægreid & Torunn Bruland and collaborators at IKM (NTNU) in 2009, Page (2012). The data is preprocessed by Arnar Flatberg (The Norwegian Microarray Consortium).

The data set consists of time series measurements of gene expression for 8956 genes. It is a collection of 12 observations for each gene, taken over a time frame of 14 hours, with two biological replicates for pairs of gastrin stimulated treatment (Gm) and unstimulated control (Un). In this thesis we will only study the difference time series, which we will refer to as GmUn. By the biologists, the two biological replicates are named BR2 and BR3.

1.2 Organisation of the report

In Chapter 2, the linear mixed effects model is defined. The model can be fitted for each gene. We will look at the area under the estimated time series curve as a measure of strength of the gene expression activation over time. This is presented in Chapter 3, with results for selected genes. Analyses will be performed based on parametric assumptions. As an alternative approach to the parametric assumptions, the method of permutation is presented in Chapter 4. A short simulation experiment is done for a better understanding of our permutation algorithm. An introduction to the multiple testing problem is found in Chapter 5. Chapter 6 contains analysis of a random sample of approximate 1000 genes. Results from a permutation test on these 1000 genes are presented, as well as the multiple testing problem for all genes. Discussion and conclusion are found in Chapter 7.

1.3 Statistical software

All statistical analysis in this thesis have been done using the statistical software `R`, R Core Team (2012). The programming code is found in Appendix A.

Chapter 2

Method

This chapter will present the Linear Mixed Effects (LME) model. This type of model can be used to describe the relationship between a response variable and explanatory variables or factors for longitudinal, clustered, and repeated measures data, which are often found in the fields of biology, ecology and medicine.

Data where the response variable is measured more than once for each subject, is of type repeated measures. The gene expression time series is an example of this. For each gene, the response value (gene expression) is measured for two or more biological replicas. The data is thus correlated within each biological replica, but the observations from different biological replicas are independent. On a linear scale, the LME model combines both fixed and random effects, and can be considered as an extension of the ordinary linear models, which require all data to be independent and identically distributed. For the LME model in a gene expression study, the fixed effects are associated with the gene, while the random effects are associated with each biological replica.

This presentation is based on Chapter 5 in Zuur et al. (2009) and Chapter 2 in Østgård (2011).

2.1 The Linear Mixed Effects Model

For each biological replica we consider the following model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $i = 1, ..., b$ represents the b different biological replicas.

The vector of continuous responses, in our data this is gene expressions, from the ith biological replica is defined as

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{1i} \\ \mathbf{Y}_{2i} \\ \vdots \\ \mathbf{Y}_{ni} \end{bmatrix},$$

where n is the total number of observations for each biological replica i.

The fixed effects design matrix $\mathbf{X}_i$ is an $n \times (k+1)$ matrix representing covariates. We will use as covariates a polynomial in time including the kth order. In the design matrix, the first column is equal to 1 for all observations as to include an intercept in the model. The design matrix is defined as

$$\mathbf{X}_i = \begin{bmatrix} 1 & x_{1i}^{(1)} & x_{1i}^{(2)} & \cdots & x_{1i}^{(k)} \\ 1 & x_{2i}^{(1)} & x_{2i}^{(2)} & \cdots & x_{2i}^{(k)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{ni}^{(1)} & x_{ni}^{(2)} & \cdots & x_{ni}^{(k)} \end{bmatrix},$$

where for example $x_{1i}^{(1)} = t_{1i}$ and $x_{1i}^{(2)} = t_{1i}^2$ for repeated measure number one.

The parameter vector $\boldsymbol{\beta}$ is the fixed effects vector, consisting of $(k+1)$ regression coefficients, one for each covariate, and one for the intercept. This is common to all replicas. The vector is defined as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

The random effects design matrix $\mathbf{Z}_i$ is an $n \times (q+1)$ matrix defined as

$$\mathbf{Z}_i = \begin{bmatrix} 1 & z_{1i}^{(1)} & z_{1i}^{(2)} & \cdots & z_{1i}^{(q)} \\ 1 & z_{2i}^{(1)} & z_{2i}^{(2)} & \cdots & z_{2i}^{(q)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_{ni}^{(1)} & z_{ni}^{(2)} & \cdots & z_{ni}^{(q)} \end{bmatrix}.$$

The random effects vector $\mathbf{u}_i$ is defined as

$$\mathbf{u}_i = \begin{bmatrix} u_{0i} \\ u_{1i} \\ \vdots \\ u_{qi} \end{bmatrix},$$

where the random effects $\mathbf{u}_i$ are independent between replicas and assumed to follow a multivariate normal distribution $N_{(q+1)}(\mathbf{0}, \mathbf{D})$, and $\mathbf{D}$ is a positive definite covariance matrix.

The simplest model for our problem has only a random intercept for each biological replica, hence the random effects design matrix $\mathbf{Z}_i$, the random effects vector $\mathbf{u}_i$ and the covariance matrix $\mathbf{D}$ will be reduced to

$$\mathbf{Z}_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{0i} \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \sigma_B^2.$$

For a slightly more complicated model with intercept and slope in time, the $\mathbf{Z}_i$, $\mathbf{u}_i$ and $\mathbf{D}$ will be reduced to

$$\mathbf{Z}_i = \begin{bmatrix} 1 & t_{1i} \\ \vdots & \\ 1 & t_{ni} \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

The vector of errors $\boldsymbol{\varepsilon}_i$ is defined as

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{ni} \end{bmatrix},$$

where each element of $\boldsymbol{\varepsilon}_i$ is the error associated with each response for the ith biological replica. The errors are independent between replicas and assumed to follow a multivariate normal distribution $N_n(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ might have a general structure. However, the most common structure is $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n\times n}$, where $\mathbf{I}_{n\times n}$ is the identity matrix, so that errors also within replicas are assumed independent. We will only use this model further. The two random vectors $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ are assumed independent of each other.

2.2 The Marginal Model

The linear mixed effects model (2.1) include two random terms, and to ease the interpretation we look at the marginal model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*, \tag{2.2}$$

where $\boldsymbol{\varepsilon}_i^* = \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$. The marginal model (2.2) can be used to study the variance structure of the response $\mathbf{Y}_i$.

Since the sum of two independent normally distributed random variables is also normally distributed, we get that $\boldsymbol{\varepsilon}_i^*$ is normally distributed with expected value

$$\begin{aligned} \mathrm{E}(\boldsymbol{\varepsilon}_i^*) &= \mathrm{E}(\mathbf{Z}_i\mathbf{u}_i) + \mathrm{E}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathrm{E}(\mathbf{u}_i) + \mathrm{E}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathbf{0} + \mathbf{0} \\ &= \mathbf{0}, \end{aligned}$$

and covariance matrix

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\varepsilon}_i^*) &= \mathrm{Cov}(\mathbf{Z}_i\mathbf{u}_i) + \mathrm{Cov}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathrm{Cov}(\mathbf{u}_i)\mathbf{Z}_i^T + \mathrm{Cov}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \sigma^2\mathbf{I}. \end{aligned}$$

We define the marginal variance-covariance matrix as

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}, \tag{2.3}$$

and

$$\boldsymbol{\varepsilon}_i^* \sim N_n(\mathbf{0}, \mathbf{V}_i),$$

and thus the marginal distribution of $\mathbf{Y}_i$ is defined as a multivariate normal distribution

$$\mathbf{Y}_i \sim N_n(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{V}_i).$$

The covariance structure between observations Y_{1i} and Y_{2i} from the same biological replica will be explained by the element $V_{[1,2]}$ in the marginal variance-covariance matrix (2.3). The correlation between the two observations is

$$\mathrm{Corr}(Y_{i1}, Y_{i2}) = \frac{V_{[1,2]}}{\sqrt{V_{[1,1]}}\sqrt{V_{[2,2]}}}.$$

Recall that the covariance between two observations from different biological replica is equal to zero by definition, since the biological replicas are independent.

2.3 Compound Symmetry

For models where the random part is just a random intercept, we have $\mathbf{D} = \sigma_B^2$, and we then get the following expression for the marginal variance-covariance matrix

$$\begin{aligned}
\mathbf{V}_i &= \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [\sigma_B^2] \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} + \sigma^2 \mathbf{I} \\
&= \begin{bmatrix} \sigma^2 + \sigma_B^2 & \sigma_B^2 & \cdots & \sigma_B^2 \\ \sigma_B^2 & \ddots & & \vdots \\ \vdots & & \ddots & \sigma_B^2 \\ \sigma_B^2 & \cdots & \sigma_B^2 & \sigma^2 + \sigma_B^2 \end{bmatrix}.
\end{aligned}$$

This structure, where the elements on the diagonal are the same and the elements on the off-diagonal are the same, is called compound symmetry.

Theorem 2 in Dobbin & Simon (2005) states that if one has a matrix on the form $\mathbf{I}_n \alpha^2 + \mathbf{J}_{n,n} \beta$, the inverse of this matrix, if it exists, is $\mathbf{I}_n \frac{1}{\alpha^2} - \mathbf{J}_{n,n} \frac{\beta}{\alpha^2(\alpha^2 + n\beta)}$. Here $\mathbf{I}_n$ is the $n \times n$ identity matrix, and $\mathbf{J}_{n,n}$ is an $n \times n$ matrix of ones.

For our model with only a random intercept for each biological replicate, where $\alpha^2 = \sigma^2$ and $\beta = \sigma_B^2$, we can write

$$\mathbf{V}_i = \mathbf{I}_n \sigma^2 + \mathbf{J}_{n,n} \sigma_B^2.$$

The expression for the inverse of the marginal variance-covariance matrix is thus

$$\begin{aligned}
\mathbf{V}_i^{-1} &= \mathbf{I}_n \frac{1}{\sigma^2} - \mathbf{J}_{n,n} \frac{\sigma_B^2}{\sigma^2(\sigma^2 + n\sigma_B^2)} \\
&= \frac{1}{\sigma^2(\sigma^2 + n\sigma_B^2)} \begin{bmatrix} \sigma^2 + (n-1)\sigma_B^2 & -\sigma_B^2 & \cdots & -\sigma_B^2 \\ -\sigma_B^2 & \ddots & & \vdots \\ \vdots & & \ddots & -\sigma_B^2 \\ -\sigma_B^2 & \cdots & -\sigma_B^2 & \sigma^2 + (n-1)\sigma_B^2 \end{bmatrix}. \qquad (2.4)
\end{aligned}$$

We will use this result when looking at the expression for the predicted random effects.

The intraclass correlation coefficient, defined as the correlation between two observations from the same biological replicate, can be calculated from $\mathbf{V}_i$ and is given as

$$\rho = \frac{\sigma_B^2}{\sigma^2 + \sigma_B^2}, \qquad (2.5)$$

and can be used to describe an effective sample size when the observations are correlated within each biological replicate. Because the standard error depends on the sample size, and we want a small standard error, a large sample size might help achieve this. When observations in each biological replicate are highly correlated, we cannot treat them as independent observations. In problems like this we can calculate the design effect, defined as

$$\text{design effect} = 1 + (n-1)\rho,$$

where n is the number of observations in each biological replicate and ρ is the intraclass correlation coefficient. If $\rho = 0$, then the design effect is 1, and the effective sample size is n. If design effect > 1, then the effective sample size is lower than the total number of observations, $N_{\text{eff}} < n$. The adjusted sample size, or effective sample size is given by

$$N_{\text{effective}} = \frac{bn}{\text{design effect}}, \qquad (2.6)$$

where n is the number of observations within each biological replicate and b is the total number of biological replicates, as before.

To describe the correlation between the repeated observations, we may use the effective sample size as an alternative to using the intraclass correlation ρ directly.

2.4 The Full Model

The linear mixed effects model for each biological replicate $i = 1, ..., b$ is defined in (2.1). A full model is a model where all the different biological replicates are put together into one model, given as

$$\underbrace{\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\\ \vdots\\ \mathbf{Y}_b\end{bmatrix}}_{(nb\times 1)} = \underbrace{\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_b\end{bmatrix}}_{(nb\times(k+1))} \underbrace{\begin{bmatrix}\boldsymbol{\beta}\end{bmatrix}}_{((k+1)\times 1)} + \underbrace{\begin{bmatrix}\mathbf{Z}_1 & 0 & \cdots & 0\\ 0 & \mathbf{Z}_2 & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \mathbf{Z}_b\end{bmatrix}}_{(nb\times b(q+1))} \underbrace{\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \vdots\\ \mathbf{u}_b\end{bmatrix}}_{(b(q+1)\times 1)} + \underbrace{\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}}_{(nb\times 1)}.$$

The full model can also be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}.$$

2.5 Maximum and Restricted Maximum Likelihood Estimation

The method of Maximum Likelihood (ML) estimation is used to estimate the parameters $\boldsymbol{\beta}$, and the variances $\mathbf{D} = \sigma_B^2$ and σ^2. ML estimation involves constructing the likelihood function for the observed data. Since the random effects $\mathbf{u}_i$ are not observed, we will use the marginal distribution of $\mathbf{Y}_i$ instead of the distribution given in (2.1) to estimate the parameters, as suggested in Galecki & Burzykowski (2013).

Finding a good estimator for the fixed effects $\boldsymbol{\beta}$ is not straightforward. First we will find the likelihood function assuming that $\mathbf{V}_i$, and hence σ_B^2 and σ^2, is known. This will result in the Best Linear Unbiased Estimator (BLUE) for $\boldsymbol{\beta}$. Because $\mathbf{V}_i$ is in fact not known, we use the expression for the BLUE to find a new likelihood function, where the unknown parameters are σ_B^2 and σ^2. By the use of this new likelihood function, we get an expression for $\hat{\mathbf{V}}_i$. With this expression we can find the Empirical Best Linear Unbiased Estimator (EBLUE) for $\boldsymbol{\beta}$, which is our main goal with ML estimation.

The marginal distribution of $\mathbf{Y}_i$ is multivariate normal distributed, $\mathbf{Y}_i \sim N_n(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$. From Johnson & Wichern (2007), we have the following expression for the probability density function of the multivariate normal distribution

$$f(\mathbf{Y}_i|\boldsymbol{\beta}, \sigma_B^2, \sigma^2) = (2\pi)^{-\frac{n}{2}}|\mathbf{V}_i|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}.$$

The likelihood function contribution for the ith biological replicate, given the observed data $\mathbf{Y}_i = \mathbf{y}_i$, is

$$L_i(\boldsymbol{\beta}, \sigma_B^2, \sigma^2) = (2\pi)^{-\frac{n}{2}}|\mathbf{V}_i|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}.$$

The joint likelihood function is then the product of these b likelihood functions, given as

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma_B^2, \sigma^2) &= \prod_{i=1}^{b} L_i(\boldsymbol{\beta}, \sigma_B^2, \sigma^2) \\ &= \prod_{1=1}^{b} (2\pi)^{-\frac{n}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}. \end{aligned}$$

The log-likelihood function is thus

$$l(\boldsymbol{\beta}, \sigma_B^2, \sigma^2) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2}\sum_{i=1}^{b} \ln(|\mathbf{V}_i|) - \frac{1}{2}\sum_{i=1}^{b} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \qquad (2.7)$$

By assuming that $\mathbf{V}_i$, and hence σ_B^2 and σ^2, is known, we can find the Best Linear Unbiased Estimator (BLUE) for $\boldsymbol{\beta}$. An estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is the best linear unbiased estimator if it can be written as $b^T Y$. This means that $E[b^T Y] = \boldsymbol{\beta}$ (unbiased), and that it has the smallest variance among the unbiased linear estimators.

Treating $\mathbf{V}_i$ as known, the log-likelihood function becomes a function of $\boldsymbol{\beta}$ only, and the maximization of this function is equal to minimization of the last term in (2.7),

$$q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{b} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

To minimize $q(\boldsymbol{\beta})$, we use the method of generalized least squares, which states that we differentiate with respect to $\boldsymbol{\beta}$, set this equal to zero and solve for $\boldsymbol{\beta}$. By doing this we will find the BLUE for $\boldsymbol{\beta}$.

$$\frac{\partial q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{1}{2}\sum_{i=1}^{b} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right) = 0$$

$$\begin{aligned} &\Rightarrow \frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{b} \left(\mathbf{y}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i - \mathbf{y}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i + \boldsymbol{\beta}^T \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\boldsymbol{\beta}\right) = 0 \\ &\Rightarrow \sum_{i=1}^{b} \left(0 - \frac{1}{2}\mathbf{y}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i - \frac{1}{2}\mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i + \frac{1}{2}(\mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i + \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i)\boldsymbol{\beta}\right) = 0 \\ &\Rightarrow \sum_{i=1}^{b} \left(-\mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i + \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\boldsymbol{\beta}\right) = 0 \\ &\Rightarrow \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{b} \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{b} \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i \qquad (2.8) \end{aligned}$$

To obtain the ML estimation for the covariance parameters σ_B^2 and σ^2, we construct a log-likelihood function, $l_{ML}(\sigma_B^2, \sigma^2)$, by replacing the fixed effects $\boldsymbol{\beta}$ by the BLUE for $\boldsymbol{\beta}$, (2.9).

$$l_{ML}(\sigma_B^2, \sigma^2) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2}\sum_{i=1}^{b} \ln(|\mathbf{V}_i|) - \frac{1}{2}\sum_{i=1}^{b} \left(\mathbf{r}_i^T \mathbf{V}_i^{-1}\mathbf{r}_i\right), \qquad (2.9)$$

where

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} = \mathbf{y}_i - \mathbf{X}_i\left((\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}_i \right). \tag{2.10}$$

There are no closed form solutions to the ML estimates of the covariance parameters. Statistical software programs can solve the equations numerically by the use of methods like Newton-Raphson.

With the estimated covariance parameters σ_B^2 and $\hat{\sigma}^2$ found numerically by `R`, we can now calculate the EBLUE for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$. By inserting these estimated values into $\mathbf{V}_i$, we get the following expression for the estimated marginal variance-covariance matrix

$$\hat{\mathbf{V}}_i = \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T + \hat{\sigma}^2\mathbf{I}.$$

By replacing $\mathbf{V}_i$ by $\hat{\mathbf{V}}_i$ in the log-likelihood function (2.7), we find that the maximization of the log-likelihood function is equivalent to minimization of the last term, as before. We obtain the EBLUE for $\boldsymbol{\beta}$ by using the method of weighted least squares,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{b}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{y}_i \tag{2.11}$$

By the method of ML estimation, the parameter estimates for σ_B^2 and σ^2 will be biased. This can be solved by a REstricted Maximum Likelihood (REML) estimation, which produces unbiased estimates for σ_B^2 and σ^2. The estimated parameters $\hat{\boldsymbol{\beta}}$ with the REML-method will not be identical to the $\hat{\boldsymbol{\beta}}$ using the ML-method.

The REML estimation is preferred over ML estimation as it produces unbiased estimates of the variance-covariance parameters σ_B^2 and σ^2. REML estimation takes into account the loss of degrees of freedom resulting from estimating the linear effects in $\boldsymbol{\beta}$. The log-likelihood function for the REML-method is given by

$$\begin{aligned} l_{\text{REML}}(\sigma_B^2, \sigma^2) = & -\frac{1}{2}(n-p)\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{b}\ln(|\mathbf{V}_i|) \\ & -\frac{1}{2}\sum_{i=1}^{b}\left(\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i\right) - \frac{1}{2}\sum_{i=1}^{b}\ln(|\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i|), \end{aligned} \tag{2.12}$$

where $\mathbf{r}_i$ is given by Equation (2.10), and p is the number of estimated parameters.

Comparing this expression with the log-likelihood function for the ML-method, we see that the REML subtracts $n-p$ instead of n in the first term, and it subtracts an extra term at the end.

By the use of the REML log-likelihood, we obtain unbiased estimates of the covariance parameters σ_B^2 and σ^2. With the estimated variance-covariance matrix $\hat{\mathbf{V}}_i$, we can compute the REML estimates of the fixed effects parameters by using the expression (2.11) for the EBLUE for $\boldsymbol{\beta}$ from the ML estimation. The fixed effects estimates will differ by the use of the two methods, and we see that the reason for this is the difference in how the two methods estimate the variance-covariance matrix $\mathbf{V}_i$.

2.6 Asymptotic distribution of fixed effects parameters

We want to examine the properties of the asymptotic distribution of the estimated fixed effects parameters. The expression for the EBLUE for $\boldsymbol{\beta}$ in (2.11) implies that $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, since the estimator is unbiased. Here we assume that $\mathbf{V}_i$ is known (implicit) by using the estimated value, $\hat{\mathbf{V}}_i$.

We look at $\hat{\boldsymbol{\beta}}$ as a linear combination of $\mathbf{y}_i$'s,

$$\hat{\boldsymbol{\beta}} = \mathbf{C}_i\mathbf{y}_i = \mathbf{C}_i^T\mathbf{X}_i\boldsymbol{\beta} + \mathbf{C}_i^T\boldsymbol{\varepsilon}^*,$$

where $\mathbf{C}$ is a $(k+1) \times n$-matrix of constant elements. With $\mathrm{E}(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*$, the mean of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}\left((\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}_i\right)\\
&= (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathrm{E}(\mathbf{y}_i)\\
&= (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}(\mathbf{X}_i\boldsymbol{\beta}+0)\\
&= (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}(\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)\boldsymbol{\beta}\\
&= \boldsymbol{\beta},
\end{aligned}$$

as expected.

To express the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, we need to find an expression for the covariance of $\hat{\boldsymbol{\beta}}$. As $\hat{\boldsymbol{\beta}} = \mathbf{C}_i\mathbf{y}_i$, we know that $\mathbf{C}_i = (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}$, hence $\mathbf{C}_i^T = \sum_{i=1}^{b}(\mathbf{V}_i^{-1})^T\mathbf{X}_i[(\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}]^T$. Remembering that the covariance for $\mathbf{y}_i$ is equal to $\mathbf{V}_i$, and some basic rules for matrix algebra, we find the following expression for the covariance of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}(\mathbf{C}_i\mathbf{y}_i)\\
&= \mathbf{C}_i\mathrm{Cov}(\mathbf{y}_i)\mathbf{C}_i^T\\
&= (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{V}_i(\mathbf{V}_i^{-1})^T\mathbf{X}_i[(\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}]^T\\
&= (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i[(\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}]^T\\
&= (\sum_{i=1}^{b}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}. \qquad (2.13)
\end{aligned}$$

With the results above, the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}, (\sum_{i=1}^{b} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1}\right). \tag{2.14}$$

The expression for the covariance of $\hat{\boldsymbol{\beta}}$ can be written as $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. We will use this notation throughout this thesis. The asymptotic distribution using this notation is then $\hat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$.

2.7 The Top-Down Strategy

The Top-Down strategy for fitting a linear mixed effects model from Zuur et al. (2009) will be used in this study. By the law of parsimony and the Top-Down strategy, the goal is to find the simplest model that fits the data best.

The Top-Down strategy can be described in four steps, where the first step is to start with a beyond optimal model. This is a model that includes all the possible fixed effects. The next step is then to find an optimal structure of the random effects. This is done by performing REML-based likelihood ratio tests for the associated covariance parameters $\mathbf{D}$ and σ^2. When the optimal structure of the random effects is found, the strategy states we find the optimal structure of the fixed effects. This can be done using ML-based *ANOVA* to compare models with different number of fixed effects parameters. If the p-value of the test is significant on an $\alpha = 0.05$ significance level, that is, if $p < 0.05$, the larger model in the test is indicated to have the better fit for the data.

The Akaike Information Criteria, *AIC*, can be used to choose the best fitted model for the data, and is defined by

$$AIC = 2p - 2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}_B^2, \hat{\sigma}^2),$$

where p is the total number of parameters estimated in the model, both fixed and random, and $l(\hat{\boldsymbol{\beta}}, \hat{\sigma}_B^2, \hat{\sigma}^2)$ is either the ML- or REML log-likelihood function. When comparing the *AIC* value for different models, the model with the lowest *AIC* value is assumed to be the best fit for the data, relative to the other fitted models.

The last step of the strategy is model verification or diagnostics. This step is described in detail in Chapter 2.8.

2.8 Diagnostics

When a linear mixed effects model is fitted, we must check and verify that the underlying assumptions for the random effects and the error terms are valid for the data. We will now list different plots for diagnostic purposes in the LME framework, suggested by Nobre & da Motta Singer (2007). As we will see in Chapter 3, some of the plots described are not relevant for our data set.

To check for normality in the errors, a normal probability (QQ) plot of the conditional standardised residuals is used. The quantiles should fall on an approximate straight line for the assumption of normality to be valid. To check for homoscedasticity (constant variance) of errors, a plot of the conditional standardised residuals versus the fitted values is used. A plot of the marginal standardised residuals versus the fitted values can serve as a check for linearity in the fixed effects. To check if the random effects are normally distributed, a weighted QQ plot of standardised linear combinations of the random effects can be used.

The raw marginal residuals, $\mathbf{r}_{m,i}$, are defined as

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

and the raw conditional residuals, $\mathbf{r}_{c,i}$, are defined by

$$\mathbf{r}_{c,i} = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{u}}_i = \mathbf{r}_{m,i} - \mathbf{Z}_i\hat{\mathbf{u}}_i.$$

Ideally, we would use the standardized residuals, defined as the raw residuals divided by their true standard deviations, (West et al. 2007, p. 42). However, the true standard deviation is rarely known, but the estimated standard deviations can be found using `R`. The Pearson residuals are standardized residuals using the estimated standard deviation, and is considered appropriate to use when the variability of the estimated fixed effects $\hat{\boldsymbol{\beta}}$ can be ignored. The Pearson residuals are defined as

$$\mathbf{r}_{m,i,\text{Pearson}} = \frac{\mathbf{r}_{m,i}}{\sqrt{\widehat{\text{Var}}(\mathbf{Y}_i)}},$$

and

$$\mathbf{r}_{c,i,\text{Pearson}} = \frac{\mathbf{r}_{c,i}}{\sqrt{\widehat{\text{Var}}(\mathbf{Y}_i|\mathbf{u}_i)}},$$

where m and c are abbreviations for marginal and conditional residuals, respectively, and i represents the biological replicates, as before.

2.9 Predicting the values of the random effects

After fitting a linear mixed effects model, we want to look at the predicted value of the random effects, $\hat{\mathbf{u}}_i$. As written in West et al. (2007), the values $\mathbf{u}_i$ are random variables that are assumed to follow a multivariate normal distribution, and therefore we predict the values of the random effects rather than estimate them.

According to West et al. (2007), the random effects are predicted conditional on the given observed response value. The expression for $\hat{\mathbf{u}}_i$ is defined as

$$\hat{\mathbf{u}}_i = \mathbf{E}(\mathbf{u}_i|\mathbf{Y}_i = \mathbf{y_i}) = \hat{\mathbf{D}}\hat{\mathbf{Z}}^T\hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}). \tag{2.15}$$

For the simple model with only a random intercept for each biological replicate, the covariance matrix $\mathbf{D}$ is equal to σ_B^2, and the random effects matrix $\mathbf{Z}_i$ is a vector of n ones. The marginal variance-covariance matrix $\mathbf{V}_i$ has compound symmetry, and the expression for its inverse is given in (2.5). The expression for the Best Linear Unbiased Predictor (BLUP) for the random effects for our model, for the ith biological replicate, can then be written as

$$\begin{aligned}\hat{\mathbf{u}}_i &= \hat{\mathbf{D}}\hat{\mathbf{Z}}^T\hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\ &= \begin{bmatrix} \frac{\sigma_B^2}{\sigma^2 + n\sigma_B^2} & \cdots & \frac{\sigma_B^2}{\sigma^2 + n\sigma_B^2} \end{bmatrix}_{(1\times n)} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\ &= \frac{\sigma_B^2}{\sigma_B^2 + \frac{\sigma^2}{n}}(\bar{\mathbf{y}} - \bar{\mathbf{X}}\boldsymbol{\beta}), \end{aligned} \tag{2.16}$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^n \mathbf{y}_j$ and $\bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^n \mathbf{X}_j$. This expression for the BLUP is equal to the one defined in McCulloch & Neuhaus (2011).

In this thesis we will use the statistical software R to predict these values.

2.10 Hypothesis test

From Casella & Berger (2002) we get the following information about hypothesis test.

A hypothesis is a statement about a population parameter. The goal with a hypothesis test is to decide which of the two complementary hypotheses is true. These two complementary hypotheses are called the null hypothesis, denoted H_0, and the alternative hypothesis, denoted H_1.

When testing hypotheses, two types of error can occur, namely type I error and type II error. A type I error is an incorrect rejection of a true null hypothesis, and is also called false positives. The probability of type I error is,

$$P(\textit{type I error}) = \alpha.$$

A type II error is the failure to reject a false null hypothesis, and is also called false negatives. The probability of type II error is,,

$$P(\textit{type II error}) = \beta.$$

A summary of the two types of errors in a single hypothesis test is found in Table 2.1.

Table 2.1: Summary table for single hypothesis testing

	Accept H_0	Reject H_0
True H_0	Correct decision	Type I error
Non-true H_0	Type II error	Correct decision

2.11 P-values

We will look at p-values when evaluating the result of a hypothesis test, since a p-value gives more information than just "accept H_0" or "reject H_0". The following definition of a p-value is found in Casella & Berger (2002).

Let $\mathbf{X} = (X_1, ...X_n)$ be independent and identically distributed random variables. A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha,$$

where α is the significance level.

A p-value gives information about the probability, assuming that the null hypothesis is true, of observing a test statistic at least as extreme as the one observed. The null hypothesis will be rejected when the p-value is small; more exact when the p-value is less than or equal to the significance level α.

If $P_\theta(p(\mathbf{X}) \leq \alpha) = \alpha$, the p-value is called an exact p-value. The probability distribution of an exact p-value is the uniform distribution.

2.12 Conditional tests for fixed effects parameters

The conditional tests for fixed effects parameters are conditional on the covariance parameters σ_B^2 and σ^2. According to Pinheiro & Bates (2000, p. 91), the conditional tests are recommended instead of likelihood ratio tests for assessing the significance of terms in the fixed effects.

The conditional t-test

The conditional t-tests for fixed effects parameters are used in situations where we want to test hypotheses on the form

$$H_0: \quad \beta = 0 \qquad vs. \qquad H_1: \quad \beta \neq 0.$$

The t-statistic associated with the test is defined as

$$t = \frac{\hat{\beta}}{\sqrt{\Sigma_{\boldsymbol{\beta}}}}.$$

This t statistic follow an approximate t-distribution, with degrees of freedom equal to the denominator degrees of freedom, see Chapter 2.13.

When using the `lme` function in `R`, the conditional t-tests are implemented in the `summary` method. The marginal significance of each fixed effect coefficient when all other fixed effects are present in the model, is tested. Thus the method is also conditional on all other fixed effects coefficients in the model. This is called a type III strategy.

The conditional F-test

The conditional F-tests for fixed effects parameters are used to test p hypotheses on the form

$$H_0: \quad \mathbf{C}\boldsymbol{\beta} = 0 \qquad vs. \qquad H_1: \quad \mathbf{C}\boldsymbol{\beta} \neq 0,$$

that is, linear combinations of the fixed effects where $\mathbf{C}$ is a known matrix. The F-statistic associated with the test is defined as

$$F = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{C}^T (\mathbf{C}\Sigma_{\boldsymbol{\beta}}\mathbf{C}^T)^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}}{rank(\mathbf{C})}. \tag{2.17}$$

This statistic follows an approximate F-distribution, where the numerator degrees of freedom is equal to the rank of $\mathbf{C}$ and the denominator degrees of freedom is defined below.

2.13 Denominator degrees of freedom

The conditional t- and F-tests use denominator degrees of freedom. As we have a balanced data structure, the expression for the denominator degrees of freedom is thus

$$denDF = N - (b + k),$$

where N is the total number of observations for all b biological replicates and k is the number of estimated fixed effects, disregarding the intercept. The formula is found in Pinheiro & Bates (2000) and is simplified for our model.

When using the `lme` function, `R` uses a method called containment to calculate the denominator degrees of freedom. In the newer function `lmer`, the degrees of freedom is not calculated. There is an ongoing discussion regarding the denominator degrees of freedom and how to calculate this. Some popular statistical tools such as `SPSS` and `SAS` use the Satterthwaite and Kenward-Roger method, respectively, while `Stata` does not calculate the denominator degrees of freedom. We have balanced data, and therefore choose to follow the guidelines in Pinheiro & Bates (2000) by using the `lme` function.

Chapter 3

Area as a measure of consistent activity

In Chapter 2 we have seen that the linear mixed effects model can be used to model long gene expression time series. We will now study the area under the fitted curves, from this point on simply called the area. This measure is not to be confused with Area Under the Curve (AUC), which is often used for a Receiver Operating Characteristic (ROC) curve.

The area under the gene expression curve relative to the gene expression at the starting point in time can be seen as a measure of consistent activity over time. The area is, mathematically speaking, a fairly easy thing to calculate. The computational costs are low, so it is possible to calculate for large amounts of data. This combination makes the area, seen from a biologist' view, an attractive method for ranking genes with respect to effect over time.

Another motivation for looking at the area is the following, in Chapter 6 we will see that fitting the LME model to the gene expression time series of all genes will claim that about 60% of the genes are significant. This is unlikely. One possible explanation for this can be that the wrong hypothesis has been tested. Hypothesis tests about the area might be a better way to asses consistent activation of genes. Such hypothesis tests will be presented in this chapter and Chapter 5. Another explanation might be that the data does not meet the criteria for the linear mixed effects model. Other models could be fitted, but this will not be done in this thesis. Instead, permutation tests will be introduced in Chapter 4 and applied to the long time series gene expression data set in Chapter 6.

In this chapter, we will present how to calculate the area and how to test hypotheses about the area. Finally, the linear mixed effects model is fitted and hypothesis tests concerning the area is performed, on a selection of genes.

3.1 Description of the data

The original data consists of, for each of the 8956 genes, two paired time series, one Gastrin (G) stimulated treatment and one UNstimulated (UN) control. The two paired time series have two biological replicates each, so $i = 1, 2$. The response $\mathbf{Y}_i$ is a logarithmic measure of gene expression. We have created a difference time series for each gene, of the difference between the gene expression for the stimulated and the unstimulated time series, for each biological replicate. It is these difference time series we will study further in this thesis.

For each time series and biological replicate there are $n = 12$ observations in time, given in minutes, where the first observation is $t_0^* = 0$ minutes, and the last observation is $t_{11}^* = 840$ minutes (14 hours). We will use scaled time points for our analysis. We have chosen to use polynomials in time, where the scaled times are defines as

$$\begin{aligned} t &= \frac{t^* - \big(\min(t^*) + \frac{1}{2}(\max(t^*) - \min(t^*))\big)}{\frac{1}{2}\big(\max(t^*) - \min(t^*)\big)} \\ &= \frac{t^*}{420} - 1 \end{aligned}$$

so that the range of the times are between -1 and 1. The scaling is done to stabilize the use of polynomials. A table with the original time points and the scaled time points is shown in Table 3.1.

Table 3.1: Table of the scaled times (t) and the corresponding original times (t^*).

t	Scaled time	t^*	Original time
t_0	-1.0000000	t_0^*	0
t_1	-0.9642857	t_1^*	15
t_2	-0.9285714	t_2^*	30
t_3	-0.8571429	t_3^*	60
t_4	-0.7857143	t_4^*	90
t_5	-0.7142857	t_5^*	120
t_6	-0.4285714	t_6^*	240
t_7	-0.1428571	t_7^*	360
t_8	0.1428571	t_8^*	480
t_9	0.4285714	t_9^*	600
t_{10}	0.7142857	t_{10}^*	720
t_{11}	1.0000000	t_{11}^*	840

3.2 Fitting Linear Mixed Effects models

We have fitted a linear mixed effects model for the difference *G-Un* time series for all genes. This has been done by following the Top-Down Strategy described in Chapter 2. The `R` code for fitting LME models is found in Appendix A. It should be noted that when working with a data set containing about 9000 genes, part of the Top-Down Strategy is not possible to perform. For instance, looking at several different residual plots for each gene, would be too time consuming of a task. Instead we have decided to trust the choice of model based on the criteria of the lowest AIC value. However, in Sections 3.6-3.8 we present residual plots for selected genes.

Step 1

The LME model is given in equation (2.1). We have chosen to focus on models with polynomials including fourth order, for two reasons. First, the biologists are not interested in modelling erratic activity over time, which would have been the case with a higher order polynomial. If data cannot be fitted using, at most, a fourth order polynomial, the data consists most likely of noise or other effects not related to the true effect of gastrin treatment over time. Second, we do not wish to estimate too many parameters compared to the number of observations.

The beyond optimal fixed effects vector is then given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \tag{3.1}$$

and the fixed effects design matrix for the ith biological replicate is given as

$$\mathbf{X}_i = \begin{bmatrix} 1 & t_{1i} & t_{1i}^2 & t_{1i}^3 & t_{1i}^4 \\ 1 & t_{2i} & t_{2i}^2 & t_{2i}^3 & t_{2i}^4 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_{12i} & t_{12i}^2 & t_{12i}^3 & t_{12i}^4 \end{bmatrix}. \tag{3.2}$$

As the beyond optimal random effect model we have chosen the random intercept model, giving a compound symmetry variance-covariance matrix $\mathbf{V}_i$. The random effects design matrix $\mathbf{Z}_i$ is then reduced to a vector of n_i ones, where $n_i = 12$ observations for each biological replicate, the random effects vector $\mathbf{u}_i = u_{0i} \sim N(\mathbf{0}, \mathbf{D} = \sigma_B^2)$ with σ_B^2 scalar, and the vector of errors $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_i)$.

Step 2

With the beyond optimal model, we want to find the optimal structure of the random component. As explained in Chapter 2.1, we decided on the simplest model with only a random intercept and will not consider more complex random structure in this presentation.

Step 3

We now want to find the optimal structure of the fixed effects, and start with four potential models we want to compare and find the best fitted for our data:

$$\begin{array}{ll}\text{Model 1 :} & y_i = \beta_0 + \beta_1 t + u_i + \boldsymbol{\varepsilon}_i \\ \text{Model 2 :} & y_i = \beta_0 + \beta_1 t + \beta_2 t^2 + u_i + \boldsymbol{\varepsilon}_i \\ \text{Model 3 :} & y_i = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + u_i + \boldsymbol{\varepsilon}_i \\ \text{Model 4 :} & y_i = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + u_i + \boldsymbol{\varepsilon}_i.\end{array}$$

We use the AIC value to choose between the four potential models. The AIC value does not say anything about whether a model is a good fit in an absolute sense, but by comparing this value for different models, we can say something about the relative fit compared to other models. Thus we choose the models with the lowest AIC values as our final model.

Step 4

The final models are refitted using the REML-method to get unbiased estimators for σ_B^2 and σ^2, which we will use at a later stage. Due to the size of the data set, it is impossible to investigate different types of residual plots for each gene. Model verification and diagnostics is therefore left out at this stage. With the results from the fitted LME models, we will in Chapters 3.6-3.8 look closer at some of the genes, and at this point we will do some model verification.

3.3 Area

We want to look at the area under the gene expression curves relative to the response value at the initial time point $t_0 = -1$. An expression for the theoretical area can be written as

$$A = \int_{-1}^{1} \Big(f(t) - y_0 \Big) dt,$$

where $f(t)$ is the true gene expression in time and y_0 is the true gene expression value at $t = -1$. The four different models we have fitted for the time series of the different genes will give different initial gene expression values y_0. The different models and the corresponding initial response values are found in Table 3.2.

The area for model 4, A_4, can then be expressed as

$$\begin{aligned} A_4 &= \int_{-1}^{1} \big(\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 - \beta_0 + \beta_1 - \beta_2 + \beta_3 - \beta_4 \big)\, dt \\ &= \left[\frac{1}{2}\beta_1 t^2 + \frac{1}{3}\beta_2 t^3 + \frac{1}{4}\beta_3 t^4 + \frac{1}{5}\beta_4 t^5 + \beta_1 t - \beta_2 t + \beta_3 t - \beta_4 t \right]_{-1}^{1} \\ &= 2\beta_1 - \frac{4}{3}\beta_2 + 2\beta_3 - \frac{8}{5}\beta_4. \end{aligned}$$

Table 3.2: Table of the true models, $f(t)$, and the corresponding true initial response value, y_0.

Model	$f(t)$	y_0
1	$\beta_0 + \beta_1 t$	$\beta_0 - \beta_1$
2	$\beta_0 + \beta_1 t + \beta_2 t^2$	$\beta_0 - \beta_1 + \beta_2$
3	$\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$	$\beta_0 - \beta_1 + \beta_2 - \beta_3$
4	$\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$	$\beta_0 - \beta_1 + \beta_2 - \beta_3 + \beta_4$

The true area for the other models can be found in a similar manner. The expressions for the four different areas are given below.

$$\begin{aligned}
A_1 &= 2\beta_1 \\
A_2 &= 2\beta_1 - \frac{4}{3}\beta_2 \\
A_3 &= 2\beta_1 - \frac{4}{3}\beta_2 + 2\beta_3 \\
A_4 &= 2\beta_1 - \frac{4}{3}\beta_2 + 2\beta_3 - \frac{8}{5}\beta_4.
\end{aligned}$$

We have fitted the curve $f(t)$ with linear mixed effects models and call this $\hat{f}(t)$. The response value at t_0 is equal to -1. When fitting the LME models, we use a fitted curve combining the two biological replicates into one curve, instead of individually fitted curves for each biological replicate. An estimate for y_0 might then be this fitted curve at $t = -1$, that is $\hat{y}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}$ at $t = -1$. We define the estimated area under the fitted gene expression curve as

$$\hat{A} = \int_{-1}^{1} \left(\hat{f}(t) - \hat{y}_0\right) dt,$$

with $\hat{f}(t) = \mathbf{X}\hat{\boldsymbol{\beta}}$, and $\hat{y}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}$ at $t = -1$. The expressions for the estimated areas, $\hat{A}$, are then similar to those for the true area, except we now use the estimated values for the fixed effects parameters, resulting in the following expressions for the estimated areas

$$\begin{aligned}
\hat{A}_1 &= 2\hat{\beta}_1 \\
\hat{A}_2 &= 2\hat{\beta}_1 - \frac{4}{3}\hat{\beta}_2 \\
\hat{A}_3 &= 2\hat{\beta}_1 - \frac{4}{3}\hat{\beta}_2 + 2\hat{\beta}_3 \\
\hat{A}_4 &= 2\hat{\beta}_1 - \frac{4}{3}\hat{\beta}_2 + 2\hat{\beta}_3 - \frac{8}{5}\hat{\beta}_4.
\end{aligned}$$

3.4 Asymptotic distribution of the estimated area

The area under the theoretical fitted curve, A, can be written as a function of the fixed effects parameters

$$A = g(\boldsymbol{\beta}).$$

This function $g(\boldsymbol{\beta})$ can be seen as a linear combination of the fixed effects parameters

$$g(\boldsymbol{\beta}) = \mathbf{c}^T\boldsymbol{\beta},$$

where $\mathbf{c}$ is a vector of constants unique for each of the possible linear models, and $\boldsymbol{\beta}$ is the vector of the fixed effects parameters.

We may estimate the area under the fitted curves by

$$g(\hat{\boldsymbol{\beta}}) = \mathbf{c}^T\hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}$ is the estimator for the fixed effects parameters. This estimator has the asymptotic distribution $\mathrm{N}_{k+1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the covariance of $\hat{\boldsymbol{\beta}}$ and is equal to $(\sum_{i=1}^{2} \mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}$.

When assuming the variance-covariance matrix $\mathbf{V}_i$ known, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution, and the distribution of the estimated area $\hat{A}$ will then also be multivariate normal, with mean equal to $\mathbf{c}^T\boldsymbol{\beta}$ and variance

$$\begin{aligned} \mathrm{Var}(g(\hat{\boldsymbol{\beta}})) &= \mathrm{Var}(\mathbf{c}^T\hat{\boldsymbol{\beta}}) \\ &= \mathbf{c}^T\mathrm{Cov}(\hat{\boldsymbol{\beta}})\mathbf{c} \\ &= \mathbf{c}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{c}. \end{aligned}$$

3.5 Hypothesis tests

In order to test whether the area is significantly different from zero or not, we perform hypothesis tests. We know that in the situation where all the fixed effects parameters are zero, $\boldsymbol{\beta} = 0$, the area must also be equal to zero. The reverse is not necessarily true, an area equal to zero is a possibility even if the fixed effects parameters are not zero. This is because the total area can be zero, but the absolute value of the area might not be zero, thus the fixed effects parameters are not equal to zero. The situation where the fixed effects parameters are equal to zero is thus a subset of the situation where the area is equal to zero. This can be summarized in Figure 3.1. We will therefore perform two hypothesis tests, one for the fixed effects parameters, and one for the area.

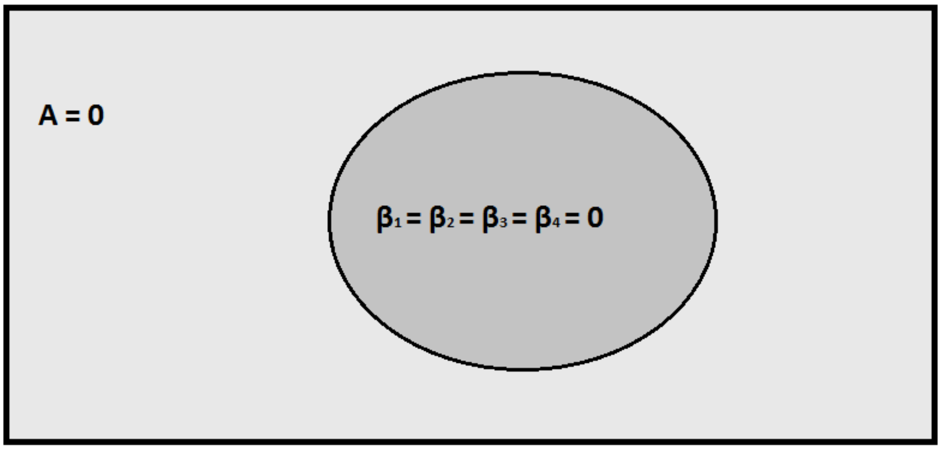

Figure 3.1: The data under H_0^1 (dark grey), where the fixed effects parameters are equal to zero, is a subset of H_0^2 (light grey), where the area is equal to zero.

We now consider the model with $k = 4$, as this is our most complex model. For the first test,

$$\mathrm{H}_0^1 : \beta_1 = \cdots = \beta_4 = 0 \qquad vs. \qquad \mathrm{H}_1^1 : \text{At least one } \beta \text{ different from } 0$$

we will calculate a test statistic with approximate F-distribution. Let

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

for $k = 4$. The test statistic is defined as

$$T_1^* = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{C}^T (\mathbf{C}\hat{\boldsymbol{\Sigma}}_\beta \mathbf{C}^T)^{-1} \mathbf{C}\hat{\boldsymbol{\beta}}}{\nu_2} \sim f_{\nu_1,\nu_2},$$

where ν_1 is the nominator degrees of freedom, equal to k, and ν_2 is the denominator degrees of freedom, as presented in 2.12.

The second test

$$\mathrm{H}_0^2 : A = 0 \qquad vs. \qquad \mathrm{H}_1^2 : A \neq 0$$

is the equivalent of testing

$$\mathrm{H}_0^2 : g(\boldsymbol{\beta}) = 0 \qquad vs. \qquad \mathrm{H}_1^2 : g(\boldsymbol{\beta}) \neq 0.$$

The true asymptotic distribution of $g(\boldsymbol{\beta})$ is multivariate normal

$$g(\boldsymbol{\beta}) \sim N(\mathbf{c}^T\boldsymbol{\beta}, \mathbf{c}^T\boldsymbol{\Sigma}_\beta\mathbf{c}),$$

while the asymptotic distribution under the null hypothesis H_0^2 is

$$g(\boldsymbol{\beta}) \sim N(0, \mathbf{c}^T\boldsymbol{\Sigma}_\beta\mathbf{c}).$$

The corresponding test statistic is defined as

$$T_2^* = \frac{\hat{A}}{\sqrt{\mathbf{c}^T\hat{\boldsymbol{\Sigma}}_\beta\mathbf{c}}} \sim t_{\nu_2},$$

where ν_2 is the denominator degrees of freedom. Notice that ν_2 is equal for both hypothesis tests, since $\mathbf{\Sigma}_\beta$ is used in the expression for both test statistics.

3.6 Analysis of gene with largest estimated area

The gene with the largest estimated area in our data set, calculated after fitting the LME model to the gene expression time series and analysing the results, is named "LOC286960". In this chapter we will present the results and analysis of this time series, using the Top-Down strategy and other methods presented in Chapter 2 and 3. A plot of the raw data is found in Figure 3.2.

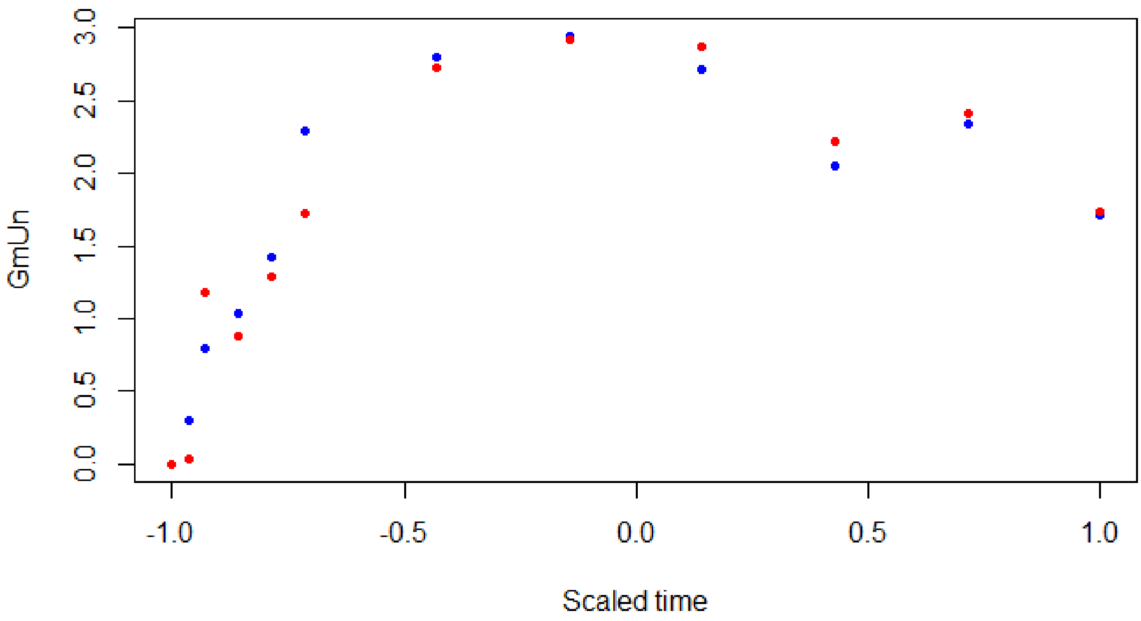

Figure 3.2: Plot of the raw data for gene "LOC286960", which has the largest estimated area in our data set. The blue dots are gene expression measurements from biological replicate BR2, and the red dots are from BR3.

The four possible models were fitted, and the following model is, by the criteria of the lowest AIC value, the best fitted

$$y_i = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + u_i + \varepsilon_i,$$

for the two biological replicates $i = 1, 2$. The estimated values of the fixed and random effects parameters are found in Table 3.3.

A normal probability plot, or QQ-plot, of the residuals, conditioned on biological replicate,

Table 3.3: Estimated values of the fixed and random effects parameters for gene ”LOC286960”.

Parameter	Estimated value
β_0	2.8189860
β_1	-0.6180437
β_2	-1.1550846
β_3	1.5145633
β_4	-0.7736399
u_1	$1.048461 \cdot 10^{-11}$
u_2	$-1.048461 \cdot 10^{-11}$

is found in Figure 3.3. From the plot, we see that the residuals seem to be approximately normally distributed for both biological replicates. This is confirmed by an Anderson-Darling test on the residuals. With p-value $= 0.63$, the null hypothesis of the errors being normally distributed is not rejected on an $\alpha = 0.05$ significance level.

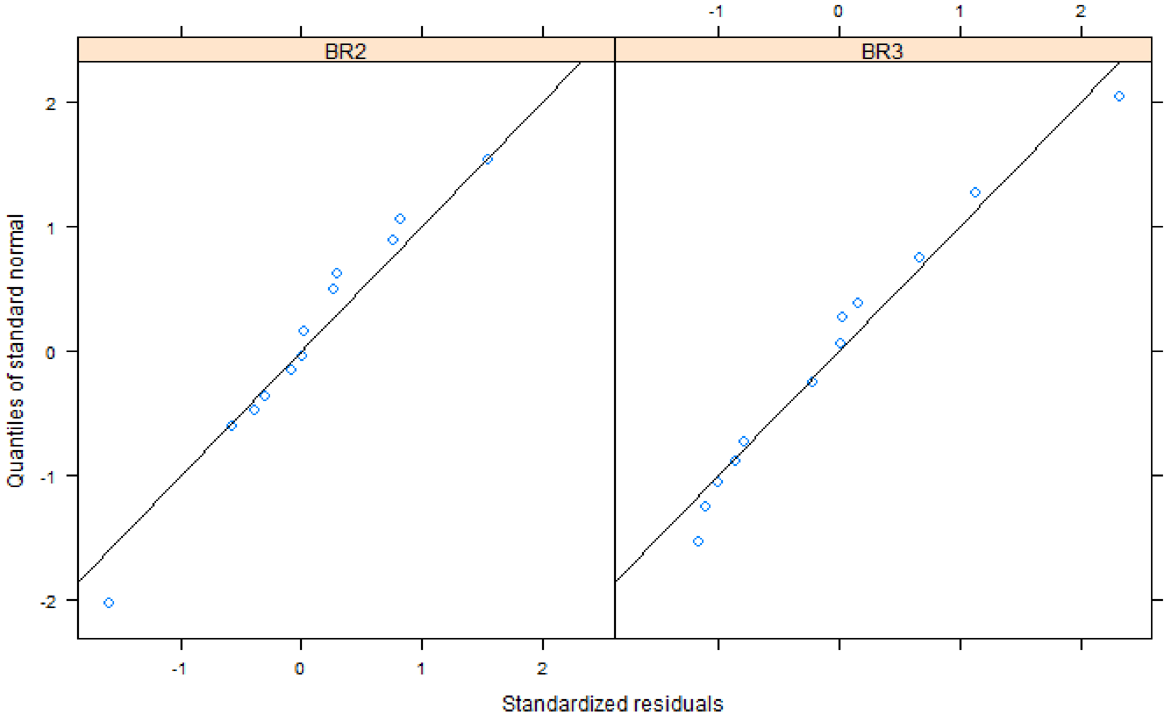

Figure 3.3: A normal probability (QQ) plot of the residuals, conditioned on biological replicate, for gene ”LOC286960”.

To check if the assumption of constant variance in the errors holds, we plot the conditional residuals against the fitted values in Figure 3.4. When plotting the marginal residuals versus the fitted values, we get an identical plot. This is due to the low values of the random effects, which is discussed in Chapter 3.9. The points seems to be randomly scattered, although some of the points deviates from the zero line. We conclude that the variance of the errors is nearly constant, and that the fixed effects are approximate normally distributed.

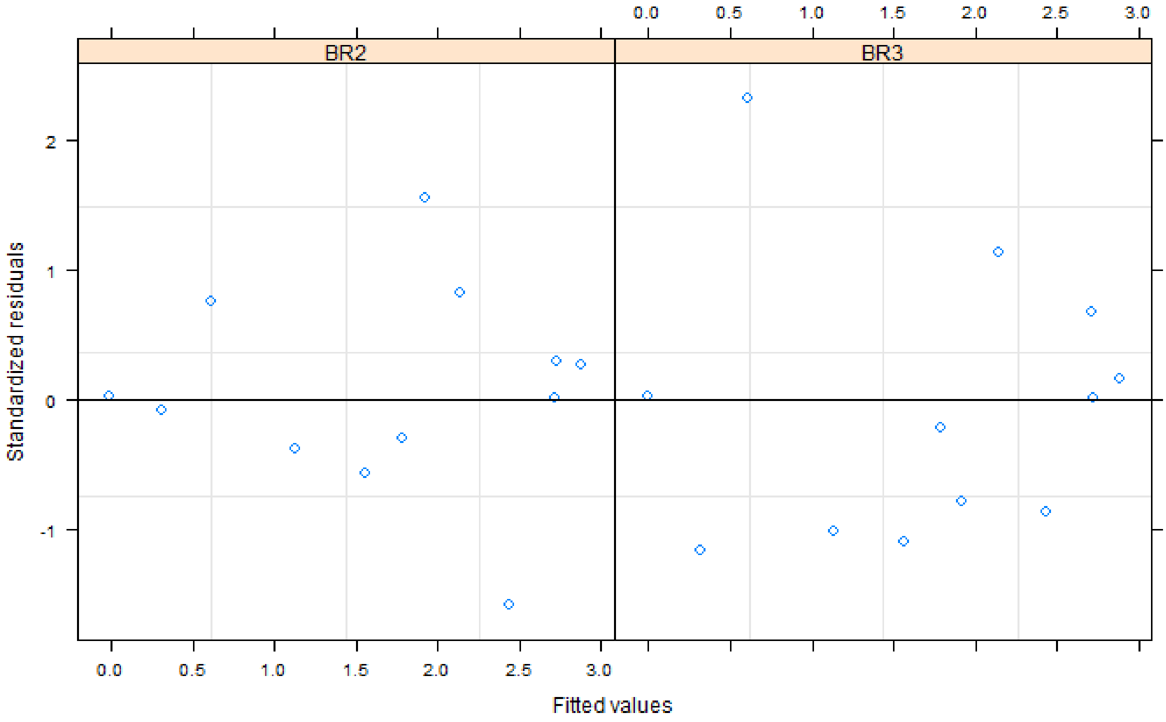

Figure 3.4: Plot of the residuals versus the fitted values, conditioned on biological replicate, for gene ”LOC286960”.

To summarise the diagnostics, we believe that the fitted LME model is a good fit for our data. A plot of the raw data and the predictions for gene ”LOC286960” is found in Figure 3.5. From calculations using the estimated values of the fixed effects parameters, the area under the fitted gene expression curve for gene ”LOC286960” is $\hat{A} = 4.57$. This is the largest estimated area for our data set. From Figure 3.5, we can see that clearly neither the fixed effects parameters nor the area is equal to zero. The results from the two hypothesis tests confirm this. The null hypothesis $H_0^1 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, provides a $p-$value $= 1.36 \cdot 10^{-11}$. The null hypothesis is thus rejected on an $\alpha = 0.05$ significance level. The null hypothesis $H_0^2 : A = 0$ provides a p-value $= 1.93 \cdot 10^{-12}$, and is also rejected on the same significance level.

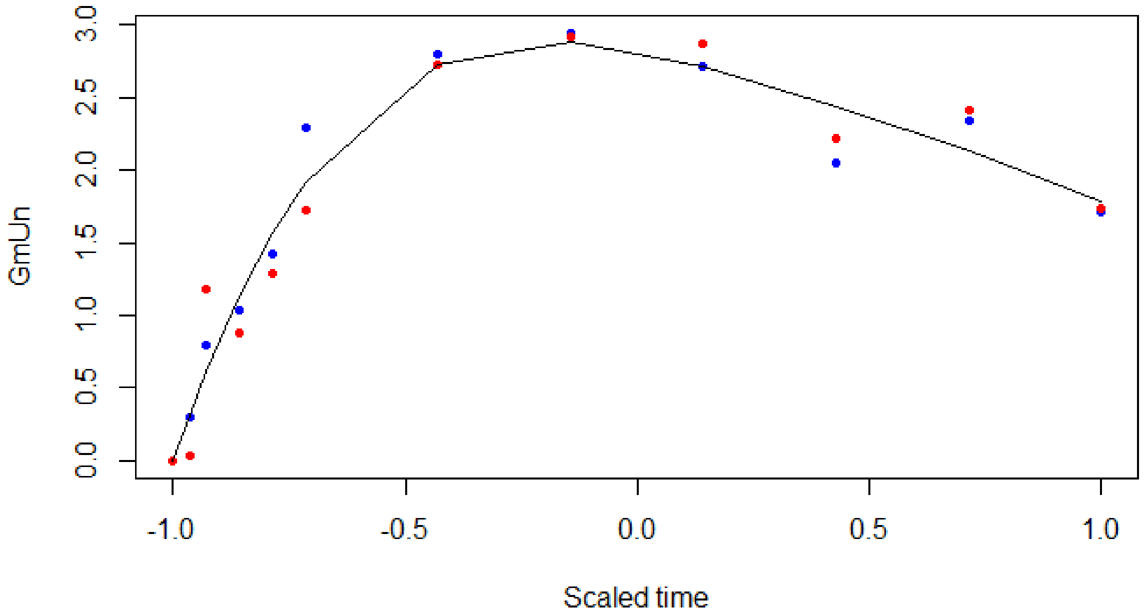

Figure 3.5: Plot of the raw data, seen as dots, and the fitted model, seen as a black line, for gene "LOC286960". The blue dots represents biological replicate BR2, and the red represents BR3.

3.7 Analysis of gene with poorest normal approximation of residuals

We will present results for and analysis of the gene in our data set which has the poorest normal approximation of the residuals after fitting LME models. This result is based on p-values obtained from Anderson-Darling tests on all fitted genes in our data set. This gene is called "Mob4", and a plot of the raw data is found in Figure 3.6. From the plot we see that there is less consistency between the two biological replicates BR2 and BR3, compared to the data for the gene studied in Chapter 3.6. The scale of the y-axis tells us that the effect of treatment versus control for this gene is somewhat limited. We expect the area to be close to zero, and the gene is thus probably not interesting with respect to consistent activity over time for the biologists.

The Top-Down strategy resulted in model 3 as the final, best fitted model among the four possible models,

$$y_i = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + u_i + \varepsilon_i,$$

where the estimated values for the fixed and random effects parameters are found in Table 3.4.

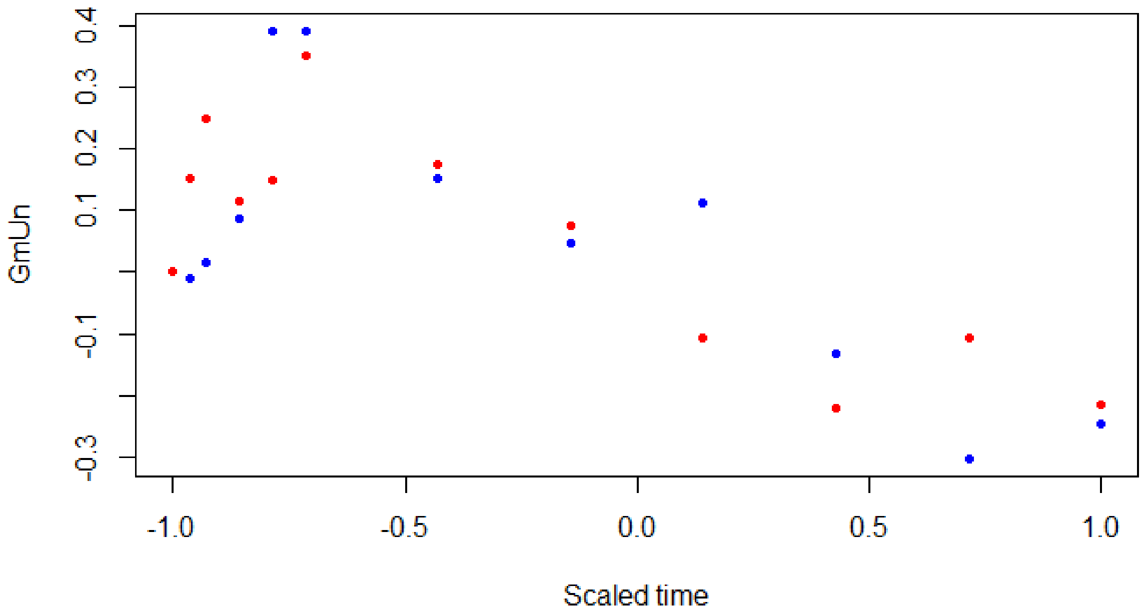

Figure 3.6: Plot of the raw data for gene ”Mob4”, which has the poorest normal approximation of the residuals in our data set. The blue dots are gene expression measurements from biological replicate BR2, and the red dots are from BR3.

Table 3.4: Estimated values of the fixed and random effects parameters for gene ”Mob4”.

Parameter	Estimated value
β_0	0.05551842
β_1	-0.55945147
β_2	-0.13160157
β_3	0.43286250
u_1	$-1.152803 \cdot 10^{-11}$
u_2	$1.152803 \cdot 10^{-11}$

Figure 3.7 shows a normal probability plot of the residuals for the fitted model for gene ”Mob4”. The errors seems not to be normally distributed, as several quantiles depart from the straight line representing the desired normal approximation. This is the case for both biological replicates. An Anderson-Darling test of the residuals provides a $p-$ value $= 3.182793 \cdot 10^{-5}$, rejecting the null hypothesis of the errors being normally distributed, as expected from the plot.

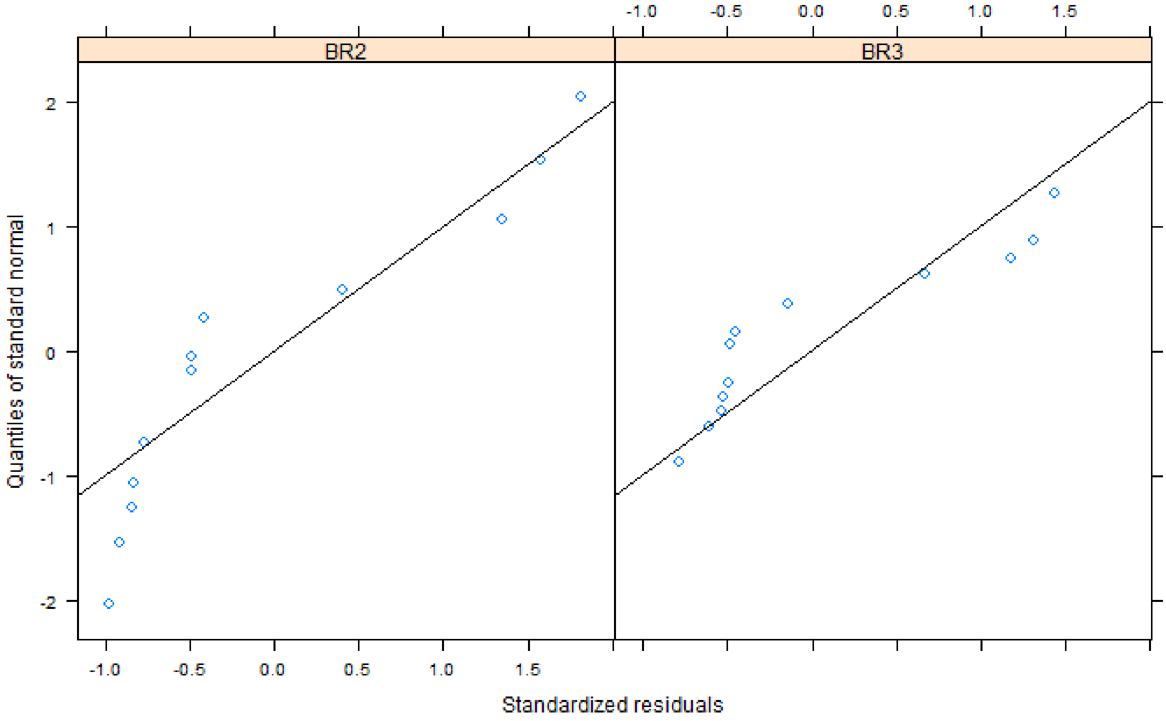

Figure 3.7: A normal probability (QQ) plot of the residuals, conditioned on biological replicate, for gene "Mob4".

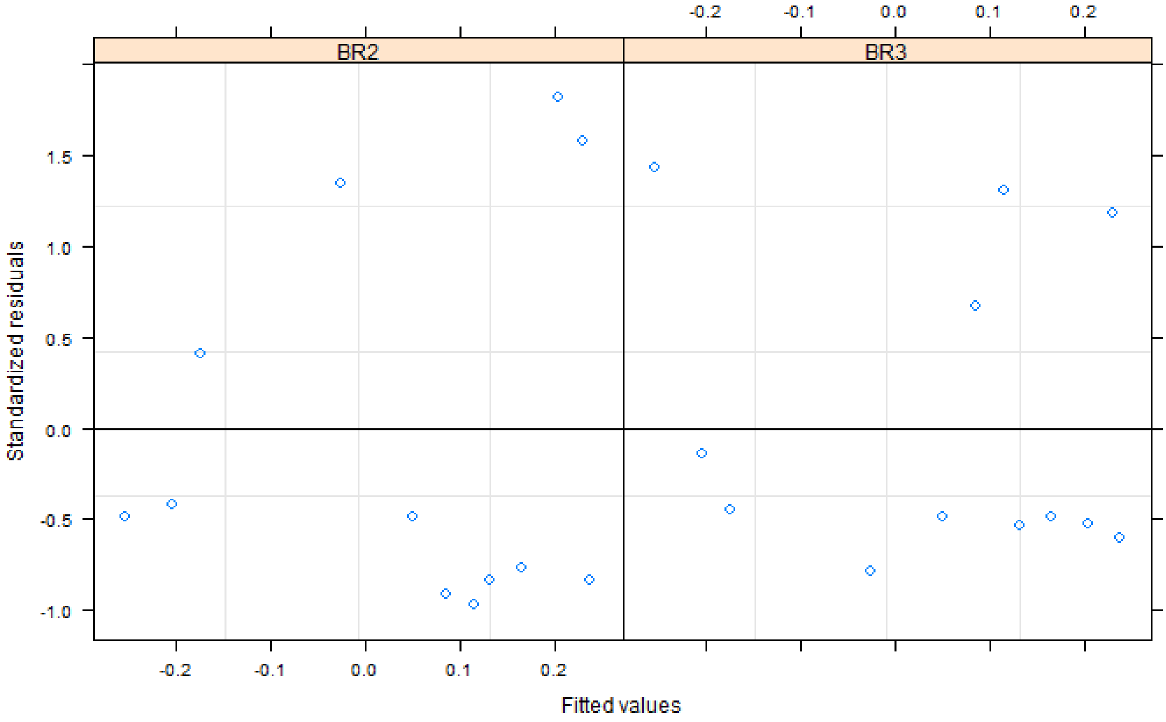

Figure 3.8: Plot of the residuals versus the fitted values, conditioned on biological replicate, for gene "Mob4".

Figure 3.8 shows a plot of the residuals versus the fitted values for "Mob4". The residuals lies far from zero. The assumption of constant variance in the errors is incorrect and/or the normality assumptions for the fixed effects does not hold. Diagnostics using this type of plot for our fitted data, where marginal and conditional residuals provide the same plot, is problematic. In accordance with linear models, a plot of the residuals versus the fitted values is often used to check for homoscedasticity in the errors. The conclusion must be that either one or both criteria are not met. The model seems not to fit the data well. From Figure 3.9, showing the predicted fitted line without random effects, we can also see that the fit is poor for the data of this gene. Note that this plot is identical to a predictive plot including the random effects, due to low values of the random effects.

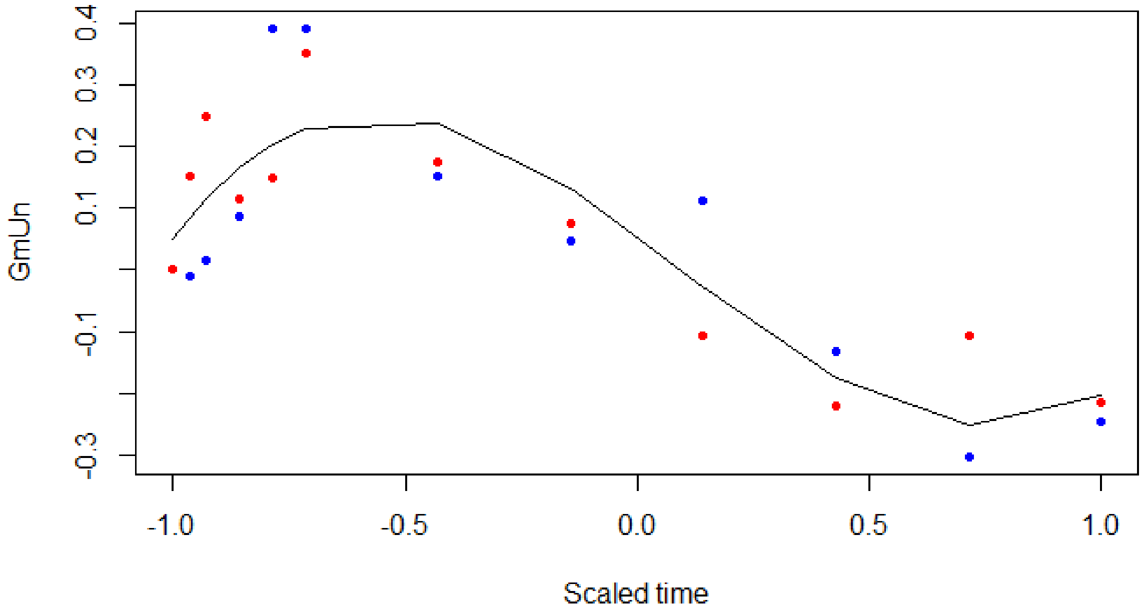

Figure 3.9: Plot of the raw data, seen as dots, and the fitted model, seen as a black line, for gene "Mob4". The blue dots represents biological replicate BR2, and the red represents BR3.

From the biologists' view, this gene is not considered interesting; with little consistent activity over time. This is confirmed by the estimated area, calculated as $\hat{A} = -0.078$. The first null hypothesis tested, H_0^1, provides a test statistic $T_1 = 11.33$ and a p-value of $3.26 \cdot 10^{-6}$. The null hypothesis of the fixed effects parameters being equal to zero is rejected on an $\alpha = 0.05$ significance level. The second null hypothesis, H_0^2, gives a test statistic $T_2 = -0.73$ and a p-value $= 0.45$. This null hypothesis of the estimated area being equal to zero is not rejected, using the same significance level as the first hypothesis. This result is as expected, since the area is in fact close to zero.

3.8 Analysis of gene with discrepancy in results between the two hypothesis tests

The last example of detailed analysis of a gene we will present in this thesis, is for the gene called ”Higd2al1”. From analysis, we found that this gene has the largest discrepancy in results between the two hypothesis tests. In Figure 3.10, the raw data for this gene is plotted. We see that the points lie both above and under zero, so a fitted model will most likely include several fixed effects parameters not equal to zero. By a rough visual estimate, the estimated area seems to be close to zero. These two observations about the fixed effects parameters and the estimated area might explain the discrepancy.

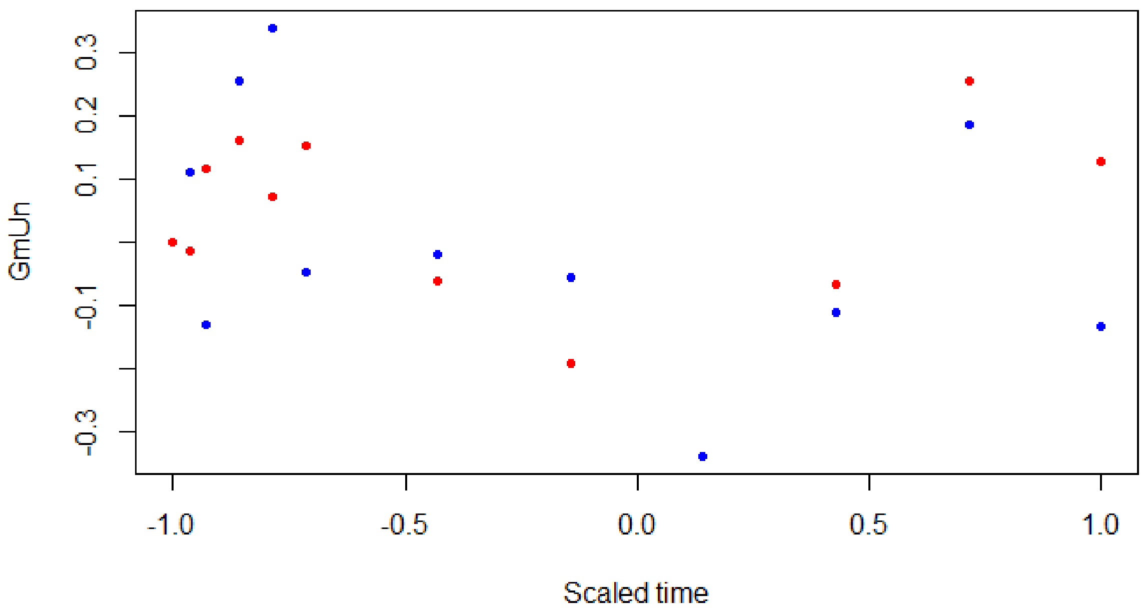

Figure 3.10: Plot of the raw data for gene ”Higd2al1”, which, in our data set has the largest difference in p-values for the two null hypothesis tested. The blue dots are gene expression measurements from biological replicate BR2, and the red dots are from BR3.

The best fitted LME model for this gene is model 4

$$y_i = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + u_i + \varepsilon_i,$$

where the estimated fixed and random effects parameters are found in Table 3.5.

Table 3.5: Estimated values of the fixed and random effects parameters for gene "Higd2al1".

Parameter	Estimated value
β_0	-0.28378673
β_1	-0.03273886
β_2	1.46949119
β_3	0.05858136
β_4	-1.19224910
u_1	$-8.006525 \cdot 10^{-12}$
u_2	$8.006525 \cdot 10^{-12}$

Figure 3.11 shows a normal probability plot of the residuals for the final model, conditioned on each biological replicate. As the points lie on or close to the line, the errors seems to be approximately normally distributed for both biological replicates. An Anderson-Darling test on the errors provides a p-value of 0.53, supporting the assumption of normality in the errors.

The residuals versus the fitted values for gene "Higd2al1" are plotted in Figure 3.12. We see that the residuals might be randomly distributed in the plot, it is difficult to say. The assumption of constant variance in the errors might hold and/or the assumption of normality in the fixed effects might hold.

The area is estimated to be $\hat{A} = -3.80 \cdot 10^{-5}$, very close to zero. The results from testing the null hypothesis of the estimated area being equal to zero, H_0^2, provides a p-value $= 0.9998$. The null hypothesis is thus not rejected on a $\alpha = 0.05$ level. The test result for testing H_0^1, provides a p-value $= 0.00046$, and the null hypothesis of the fixed effects parameters being equal to zero is rejected with the same significance level as for H_0^2. These results are as expected when looking at the plot of the raw data and fitted model in Figure 3.13.

It is worth mentioning that this gene is almost unique in our data set, with such difference in results for the two hypotheses. Other genes, such as the "Mob4" also have a difference, but not as extreme as "Higd2al1".

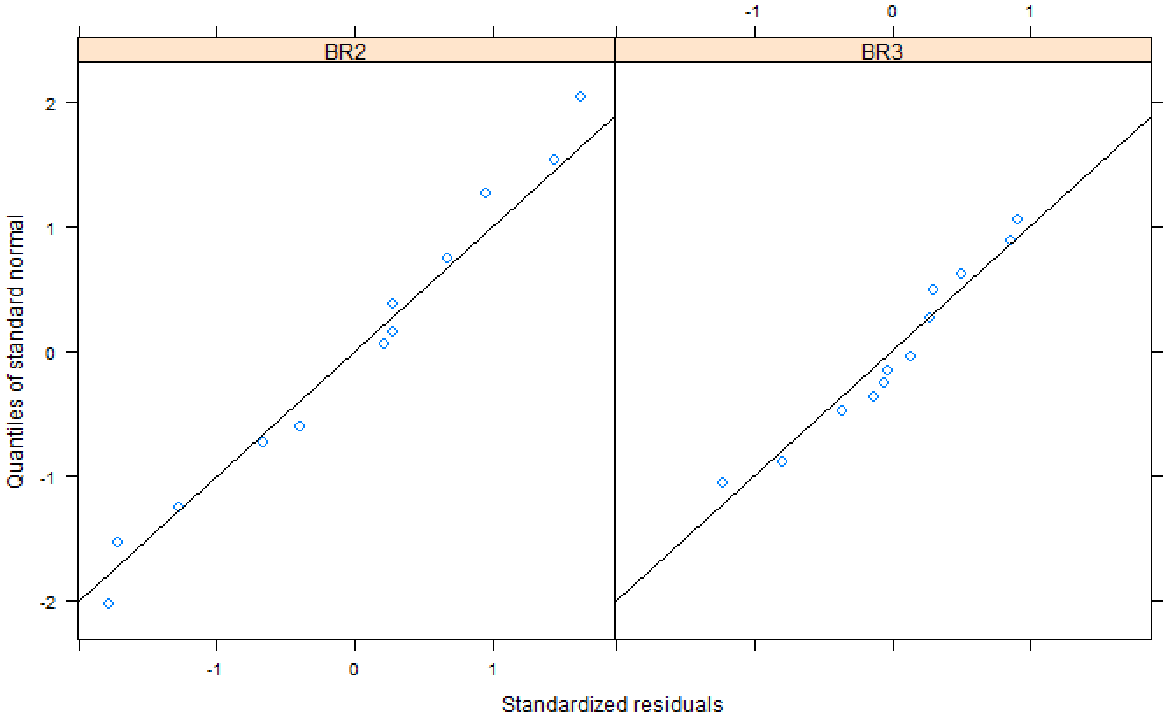

Figure 3.11: A normal probability (QQ) plot of the residuals, conditioned on biological replicate, for gene ”Higd2al1”.

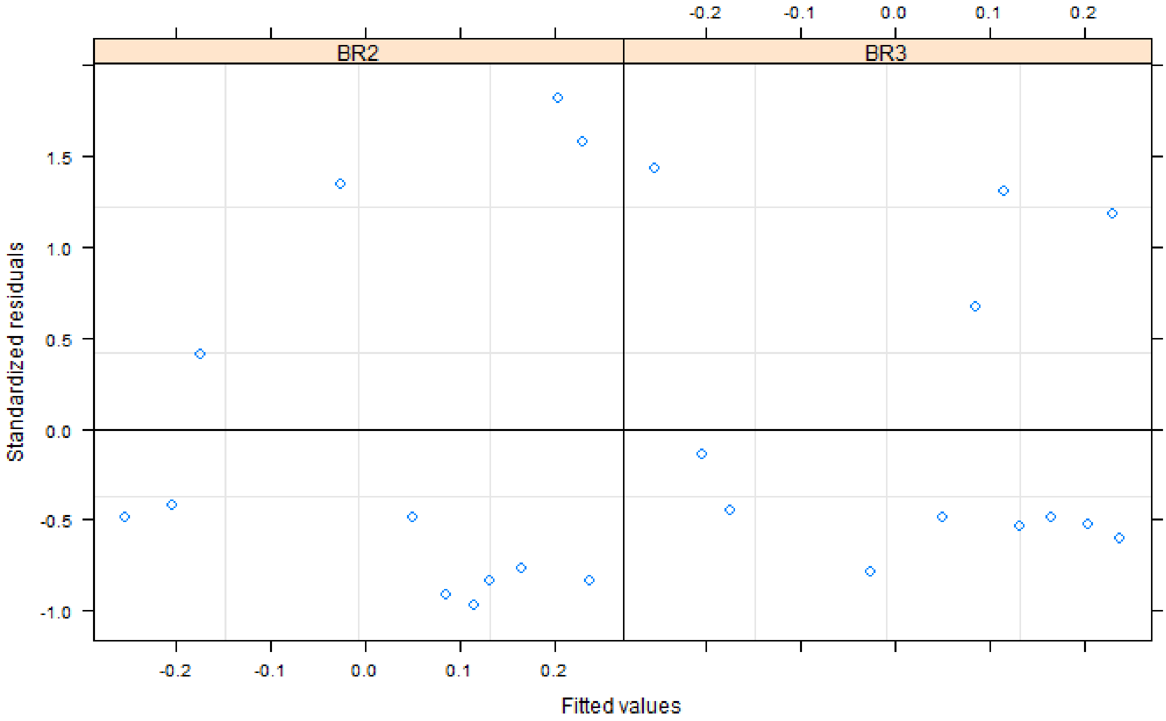

Figure 3.12: Plot of the residuals versus the fitted values, conditioned on biological replicate, for gene ”Higd2al1”.

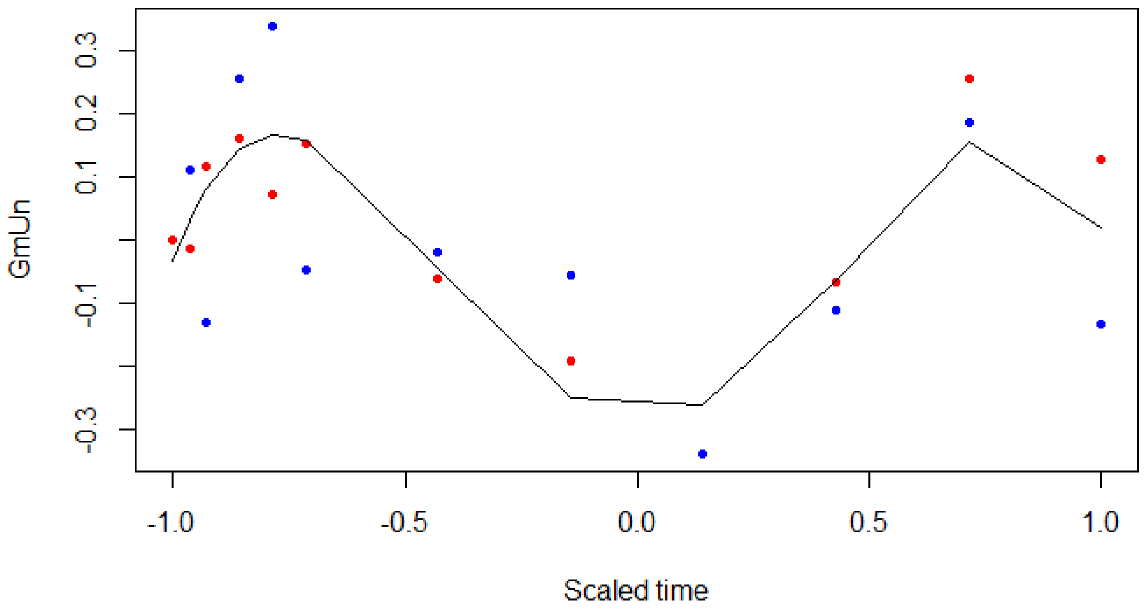

Figure 3.13: Plot of the raw data, seen as dots, and the fitted model, seen as a black line, for gene "Higd2al1". The blue dots represents biological replicate BR2, and the red represents BR3.

3.9 Remarks

Throughout this chapter, we have presented the estimated area under the fitted curves as a possible measure of consistent activity over time for gene expression time series. Some examples have been presented, with analysis of the fitted model using diagnostic plots and hypothesis testing based on the results of the parametric distribution of the estimated area.

In Chapter 2.8, we have presented different diagnostic plots for the LME model. In our analysis, however, we have only looked at two of these, the normal probability plot of the residuals, and the plot of Pearson residuals versus fitted values. We have not specified if the residuals in the latter are marginal or conditional. Due to the low values of the random effects, the plots of marginal and conditional residuals are equal. It is thus difficult to say whether a bad fit in the plots of residuals versus fitted values are caused by lack of linearity in the fixed effects, or by lack of homoscedasticity in the errors. From the framework of linear models, the plot of residuals versus fitted values is used as a check for homoscedasticity in the errors. We may interpret our plots as verification or non-verification of this assumption.

As we have chosen a model with only two values for the random effects, one for each biological replicate, a QQ plot of the random effects will have no purpose for our difference gene expression data set.

Generally, for most genes in our data set, the estimated variances of random effects and of errors are small. For the random effects, the estimated values of σ_B^2 are of the order $10^{-10} - 10^{-12}$. For the errors, the estimated values of σ^2 are of order $10^{-1} - 10^{-2}$. As a result, the estimated values of the random effects are small, in the order of 10^{-11}. When calculating the intraclass correlation for all genes in our data set using (2.5), we find the values to be in the order of 10^{-10}. The correlation between two observations from the same biological replicate is small, almost negligible. Thus, the effective sample sizes, calculated using (2.6), are $N_{\text{effective}} = 24$ for all genes. The effective sample sizes are equal to the the actual number of observations for each gene, $bn = 24$, and the observations can thus be treated as independent observations.

In this thesis we have only studied the difference gene expression data, i.e. the difference between the gastrin stimulated treatment and unstimulated control time series. For each of the stimulated and unstimulated gene expression data, the random effects are found to be greater than for our difference gene expression data, with values in the order of 10^{-2}. The choice of using the LME framework was made to handle gene expression data in general.

The three genes presented in the prior chapters are not necessarily representative for our difference data set. We have included analysis of these genes to illustrate some of the extreme cases. We have used two hypothesis tests, H_0^1 and H_0^2. The null hypothesis H_0^1 have been used by the biologists in our project to assess significance of differentially expressed genes. Both "Mob4" and "Higd2al1" are considered significant findings under H_0^1. However, under our suggested null hypothesis H_0^2, the two genes are not considered significant findings. The light grey shaded area in Figure 3.1 illustrates these situations. Introducing the area for significance assessment through H_0^2 can be an alternative method to the use of H_0^1. The next chapters of this thesis will explore this further, with the use of simulation and permutation.

Chapter 4

Permutation Test

A permutation test is a type of non-parametric test. It is a method for making inference without assuming a specific form for the distribution of the chosen test statistics under the null hypothesis.

If one need not consider the cost of data analysis, the permutation test would be conducted on all genes in our gene expression data set. However, this is not possible within the time frame of a master thesis. We will describe the test and how it can be performed. Analysis and results using the method of permutation on our gene expression data set is described in Chapter 6. Our procedure is summarized in the following permutation algorithm.

4.1 The permutation algorithm

The algorithm can be divided into three steps. We will look at the difference time series only, and we will look at the two biological replicates separately.

The two different null hypotheses we wish to test, are the same as presented in Chapter 3.5;

$$\begin{aligned} H_0^1 &: \beta_1 = \cdots = \beta_4 = 0 \quad vs. \quad H_1^1 : \text{ At least one } \beta \text{ different from } 0, \\ H_0^2 &: A = 0 \quad vs. \quad H_1^2 : A \neq 0. \end{aligned}$$

In advance of the permutation test, we calculate the two test statistics for the observed data. The expressions for the test statistics are found below.

Step 1: Generate permutations

Considering the two biological replicates separately, with $n = 12$ observations for each time series, we have 12! possible permutations if we randomly shuffle the time points. As 12! is approximately equal to 4.79×10^8, taking all permutations might be costly in terms of computer resources. This will depend on how the permutation algorithm is implemented, and on the calculations needed for the test statistics. We can choose between a sampled permutation test, taking a sample of the 12! possible permutations, and an exact or complete permutation test, which involves all 12! permutations. We denote the number of permutations performed B.

Step 2: Fit LME and calculate test statistics

For each permutation, we will fit the data to an LME model and calculate test statistics under the two null hypothesis defined in Chapters 3.5 and 4.1. We can predefine the same model for all the permutations done on each gene, or we can fit the four different models described in Chapter 3.2 for each permutation, and then use the best fitted model for calculating the test statistics. The two different scenarios will now be described in detail.

Scenario 1: Predefined fixed model

In the case where a predefined fixed model is used to fit the data, there is no issue with degrees of freedom, since this will be the same for all permutations made for a gene. We will use model 4 as the fixed model. This decision is based on two observations; model 4 is often the best fit for the original data, and by using the most complex model, the data will not be underfitted.

The two test statistics to be calculated under each of the two null hypothesis, H_0^1 and H_0^2, are named T_1^* and T_2^*, respectively. The test statistics have been defined in 3.5, but we will include the expressions here for consistency. The first test statistic is

$$T_1^* = F \sim F_{\nu_1,\nu_2}, \tag{4.1}$$

where ν_1 is the numerator degrees of freedom, and ν_2 is the denominator degrees of freedom. The expression for F is found in Equation (2.17). The second test statistic is

$$T_2^* = \frac{\hat{A}}{\sqrt{\mathbf{c}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \mathbf{c}}} \sim t_{\nu_2}. \tag{4.2}$$

Note that the numerator degrees of freedom, ν_1, will be equal to both k, the number of fixed effects parameters in the model, and the model number chosen. Note that the observed data is also fitted to the fourth LME model. The corresponding test statistics for the observed data are denoted $T_{1,\text{obs}}^*$ and $T_{2,\text{obs}}^*$.

Scenario 2: Model selection using AIC

In this case, all four different possible models will be fitted to the data, and the model with the lowest AIC value will be chosen. This is done for each permutation. Since each permutation can take four different models, the degrees of freedom will vary with the model choice. The test statistics need to be comparable for all models, and this is done by computing adjusted test statistics. This method is, in terms of computational costs, more expensive than using a predefined fixed model.

The first adjusted test statistic is defined as

$$T_1 = \frac{F - \mathrm{E}(F)}{\sqrt{\mathrm{Var}(F)}}, \tag{4.3}$$

where

$$\begin{aligned} \mathrm{E}(F) &= \frac{\nu_2}{\nu_2 - 2} && \text{if } \nu_2 \geq 3, \\ \mathrm{Var}(F) &= \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} && \text{if } \nu_2 \geq 3, \end{aligned}$$

and where ν_1 and ν_2 are the numerator and denominator degrees of freedom.

The second adjusted test statistic is defined as

$$T_2 = \frac{T_2^*}{\sqrt{\mathrm{Var}(T_2^*)}}, \tag{4.4}$$

where $\mathrm{Var}(T_2^*) = \dfrac{\nu_2}{\nu_2 - 2}$ for $\nu_2 \leq 3$. The distributions of the adjusted test statistics T_1 and T_2 are not known. The two test statistics calculated for the observed data are denoted $T_{1,\text{obs}}$ and $T_{2,\text{obs}}$.

Step 3: Produce p-value

We will now calculate permutation p-values for each of the test statistics. Let B be the number of permutations, and B^* be the number of test statistics as extreme as, or more extreme than, the observed test statistics. Then

$$p\text{-value} = \frac{B^* + 1}{B + 1}. \tag{4.5}$$

4.2 Simulation study

A small scale simulation study is done prior to the use of the permutation algorithm. A simulation study can give us a better understanding of how the algorithm works, and it can serve as a quality check of the algorithm. By simulating data under the null hypothesis H_0^1 for one randomly selected gene, we can examine the various test statistics and p-values, and we can check if the random shuffling of time points is a good method for creating permutations. We have chosen to simulate data under H_0^1, and not under H_0^2, since this will require the fixed effects parameters to be zero, which is easy to simulate. The `R` codes used for the simulation study are found in Appendix A.

We will simulate $K = 100$ new data sets under the null hypothesis

$$H_0^1: \quad \mathbf{y}_i = \mu + u_i + \boldsymbol{\varepsilon},$$

with $\mu = \beta_0$. We sample a random gene from our original data set and fit an LME model 4 to the original data. From this model, the estimated values of β_0, σ_B^2 and σ^2 are obtained. The estimated value of β_0 is fixed and will be used for all K new data sets, and is expressed as μ. Remember that $\mathbf{u} \sim N(0, \sigma_B^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ and independent of each other. The values u_1 and u_2, for the two biological replicates BR2 and BR3 respectively, and $\boldsymbol{\varepsilon}$ are simulated from the normal distribution with mean 0, and variance equal to the estimated values from the LME fit of the original data. These values are simulated for each new data set. The response values $\mathbf{y}_i$ for each new data set is obtained using the formula $\mathbf{y}_i = \mu + u_i + \boldsymbol{\varepsilon}$. Test statistics are calculated. For each of the new data sets, the two scenarios of the algorithm described in Chapter 4.1 are performed with $B = 10000$ permutations, and test statistics and p-values are calculated.

The sampled random gene is called ”Kcng2”, with identification symbol ”ILMN_1353916”.

4.2.1 Simulation study using a predefined fixed model

For this simulation, we run the version of the permutation algorithm where the data is fitted to a predefined fixed model for all B permutations. This predefined model is model 4. The degrees of freedom are equal for all permutations, so the test statistics T_1^* (4.1) and T_2^* (4.2) are used.

For each of the K new data sets, a total of K parametric test statistics $\mathbf{T}^*_{1,\text{par}}$ and $\mathbf{T}^*_{2,\text{par}}$, and K associated parametric p-values, $\mathbf{p}_{\mathbf{1},\text{par}}$ and $\mathbf{p}_{\mathbf{2},\text{par}}$, are calculated. A total of B permutations are done for each new data set, resulting in K permutation p-values, $\mathbf{p}_{\mathbf{1},\text{perm}}$ and $\mathbf{p}_{\mathbf{2},\text{perm}}$, associated with the two test statistics. If the permutation algorithm works well, that is, if the shuffling of time points to generate permutations is a valid method for generating data under the null hypothesis, the parametric test statistics from the simulation should be F- and t-distributed. The p-values under the null hypotheses, both parametric and permutation, should be uniformly distributed, as stated in Chapter 2.11.

Results: Estimated parameters

The original data from a randomly sampled gene is fitted to an LME model 4. The estimated values for the parameters β_0, σ and σ_B are found in Table 4.1. These values are used for simulating the K new data sets.

Table 4.1: Estimated values of parameters

Parameter	Estimated value
$\hat{\beta}_0$	0.1941745
$\hat{\sigma}$	0.4046577
$\hat{\sigma}_B$	$5.849867 \cdot 10^{-6}$

Results: Test statistics

We start by looking at the test statistics calculated for each of the simulated new data sets, the parametric test statistics $\mathbf{T}^*_{1,\text{par}}$ and $\mathbf{T}^*_{2,\text{par}}$. Since the LME model 4 is fitted for all new data sets, the degrees of freedom are equal for all K data sets. A plot of the parametric test statistics $\mathbf{T}^*_{1,\text{par}}$ and the theoretical F-distribution with numerator degrees of freedom $\nu_1 = 4$ and denominator degrees of freedom $\nu_2 = 18$ is found in Figure 4.1. We can see that the the density of the test statistic (black line) is approximately F-distributed (red line). Some deviation from the F-distribution can be seen in the tails, but considering that the result comes from a small simulation study with $K = 100$ simulations, this is as expected.

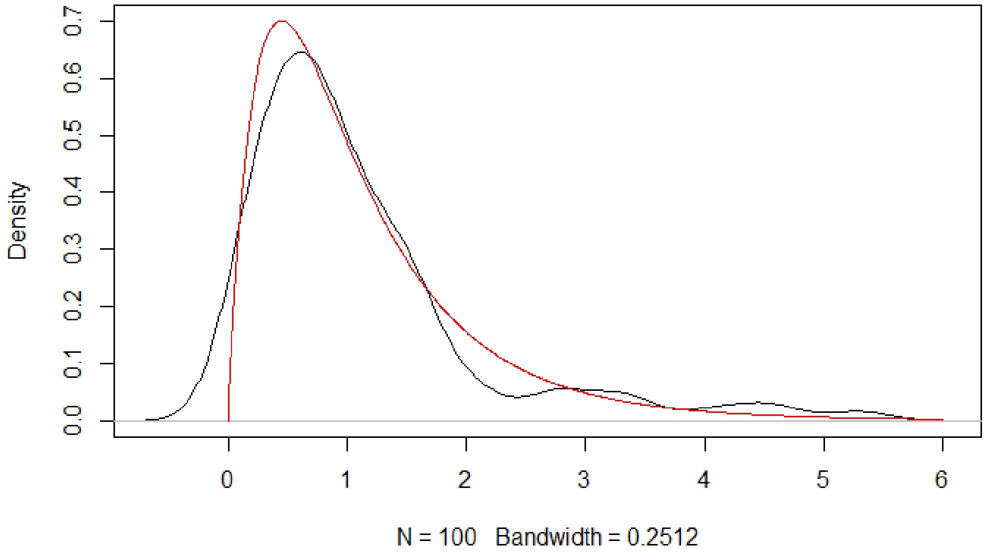

Figure 4.1: The density of the parametric test statistics $T^*_{1,\text{par}}$ from the 100 simulated data sets is shown in black. The theoretical F-distribution with $\nu_1 = 4$ and $\nu_2 = 18$ is shown in red.

A plot of the parametric test statistics $\mathbf{T}^*_{2,\text{par}}$ and the theoretical t-distribution with $\nu_2 = 18$ degrees of freedom is found in Figure 4.2. From the plot we can see that the mean of the density of the test statistics is somewhat shifted to the left, and some deviations are found in the tails, especially in the lower tail. Still, the density of the test statistics seems to be approximately t-distributed. We conclude that both parametric test statistics are approximately F- and t-distributed, and hence the permutation algorithm is a valid method for generating data under both null hypotheses H_0^1 and H_0^2.

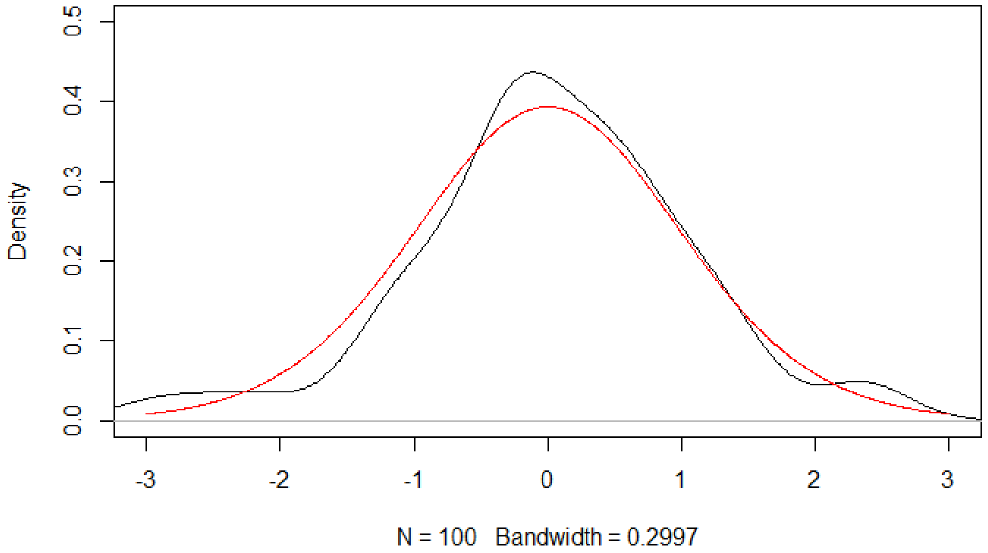

Figure 4.2: The density of the parametric test statistics $T^*_{2,\text{par}}$ from the 100 simulated data sets is shown in black. The theoretical t-distribution with $\nu_2 = 18$ degrees of freedom is shown in red.

Results: p-values

We will now look at the various p-values generated from this simulation study. Associated with the parametric test statistics, we have $\mathbf{p}_{\mathbf{1},\text{par}}$ and $\mathbf{p}_{\mathbf{2},\text{par}}$, and associated with the permutation test statistics, we have $\mathbf{p}_{\mathbf{1},\text{perm}}$ and $\mathbf{p}_{\mathbf{2},\text{perm}}$, all of which should be uniformly distributed. Empirical Cumulative Distribution Function (ECDF) plots for test statistics number one (left panel) and two (right panel) are found in Figure 4.3.

In the left panel of Figure 4.3, we see that the p-values for both the parametric and the permutation test statistics seems approximate uniformly distributed. However, the parametric p-values seems to be a better fit for the uniform distribution than the permutation p-values. Both the parametric and the permutation p-values lie mostly under the theoretical uniform distribution, indicating that the number of significant p-values are underestimated. Kolmogorov-Smirnov tests using the function `ks.test()` in `R` are used to test whether the two data sets $\mathbf{p}_{\mathbf{1},\text{par}}$ and $\mathbf{p}_{\mathbf{1},\text{perm}}$ differ significantly from each other. The same test is also conducted to test whether each of $\mathbf{p}_{\mathbf{1},\text{par}}$ and $\mathbf{p}_{\mathbf{1},\text{perm}}$ differs significantly from the uniform distribution. On a $\alpha = 0.05$ significance level, all tests conclude that both the parametric and the permutation p-values are from the same distribution, and that both follow the uniform distribution. The results from the Kolmogorov-Smirnov tests are found in Table 4.2.

In the right panel of Figure 4.3, we see that the p-values associated with the second test statistic have a greater distance between the uniform distribution compared to the p-values associated with the first test statistic. Both the parametric and the permutation p-values deviates somewhat from the uniform distribution, but there is not evidence that the two are not approximate uniformly distributed. This conclusion is also supported by Kolmogorov-Smirnov tests, which states that both densities of p-values follow the uniform distribution. The results from the Kolmogorov-Smirnov tests are found in Table 4.2.

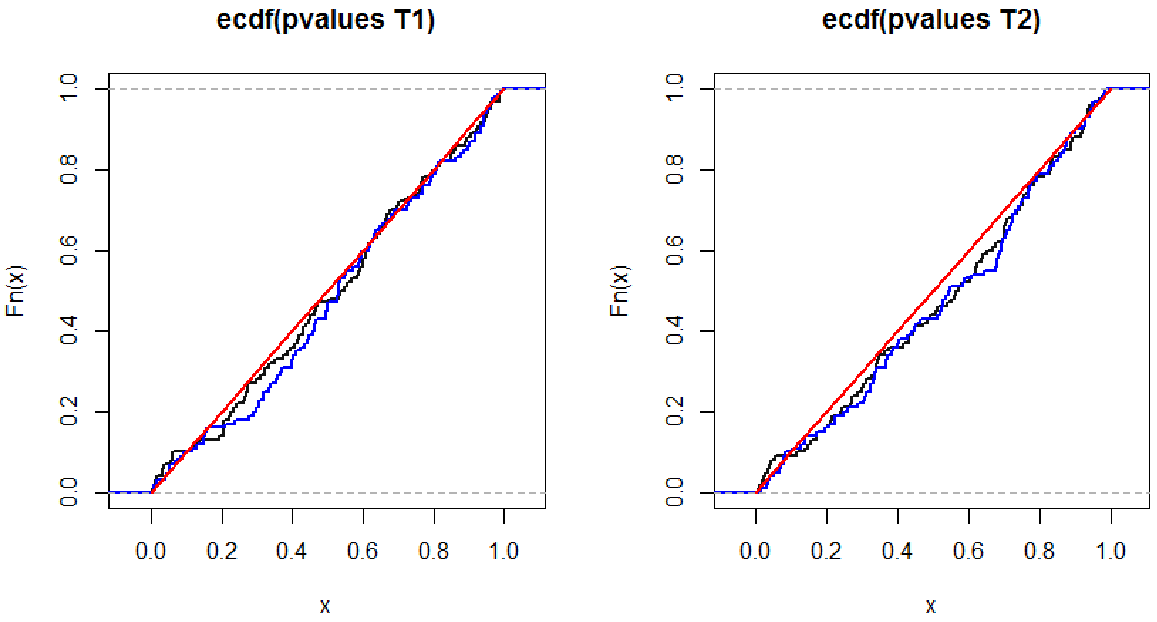

Figure 4.3: ECDF plots of the p-values associated with the test statistics for the method of a predefined fixed model. Left panel: p-values associated with test statistic T_1. Right panel: p-values associated with test statistic T_2. The black line represents results from parametric calculations, the blue line represents results from permutation. The red line is the theoretical uniform distribution.

Table 4.2: Kolmogorov-Smirnov test results for simulation by predefined fixed model.

Test data	p-value	Decision on $\alpha = 0.05$ level
$\mathbf{p_1}_{,\text{par}}$ versus uniform	0.9938	Fail to reject H_0
$\mathbf{p_2}_{,\text{par}}$ versus uniform	0.4676	Fail to reject H_0
$\mathbf{p_1}_{,\text{perm}}$ versus uniform	0.3667	Fail to reject H_0
$\mathbf{p_2}_{,\text{perm}}$ versus uniform	0.3667	Fail to reject H_0
$\mathbf{p_1}_{,\text{par}}$ versus $\mathbf{p_1}_{,\text{perm}}$	0.8127	Fail to reject H_0
$\mathbf{p_2}_{,\text{par}}$ versus $\mathbf{p_2}_{,\text{perm}}$	0.9938	Fail to reject H_0

Summary

From this simulation study using a predefined fixed model, we see that the two methods, parametric and permutation, provide comparable results. The test statistics follow approximate F- and t-distribution, and the p-values are uniformly distributed. In addition, we have seen that test statistic T_2^* is a suitable choice also for data simulated under H_0^1. Thus we have verified our permutation strategy and code, and we may proceed with the next simulation study.

4.2.2 Analysis based on model selection using AIC

The simulation based on model selection using AIC values follows the same procedure as the one in the previous chapter, except for an additional step when fitting the LME model for each simulation and permutation. The estimated values of the parameters used for simulating $K = 100$ new data sets are the same as for the predefined fixed model scenario. These estimated values are found in Table 4.1.

For each of the $K = 100$ new data sets, all four possible LME models are fitted. By comparing the AIC values, the best fitted model is chosen and adjusted parametric test statistics are calculated. We use adjusted test statistics T_1 and T_2, since the degrees of freedom will vary among the fitted models. For each of the new data sets, $B = 10000$ permutations are done. For each permutation, all four possible LME models are fitted, and the best fitted with respect to the lowest AIC value is chosen. Adjusted permutation test statistics are calculated for each permutation. When all B permutations for each new data set are done, permutation p-values associated with the test statistics are calculated.

We will now have a closer look at the results from this simulation study.

Results: p-values

Figure 4.4 shows ECDF plots of the p-values associated with the parametric and permutation test statistics. The left panel shows the results for adjusted test statistic T_1, while the right panel shows the results for adjusted test statistic T_2. The black lines represents results for parametric p-values, the blue lines from permutation p-values. The red line is the theoretical uniform distribution.

Looking at the left panel of Figure 4.4, we see that the p-values lie mostly above the uniform distribution. That is, we observe more small p-values than should be expected under the null hypothesis. This is also the case for the p-values associated with the second test statistic, found in the right panel. This is the opposite of what we see in Figure 4.3 for the case of the predefined fixed model, where the p-values lie mostly below the uniform distribution.

For both panels in Figure 4.4, we see that the permutation p-values (blue lines) lie closer to the theoretical uniform distribution than the parametric p-values (black line). The parametric p-values associated with test statistic T_1 (left panel) seems not to follow an approximate uniform distribution. This might also be the case for the parametric p-values associated with test statistic T_2 (right panel). In addition, the parametric and permutation p-values associated with T_1 have a greater distance than the p-values associated with T_2. This observation indicates that the parametric and permutation p-values for T_1 does not come from the same distribution.

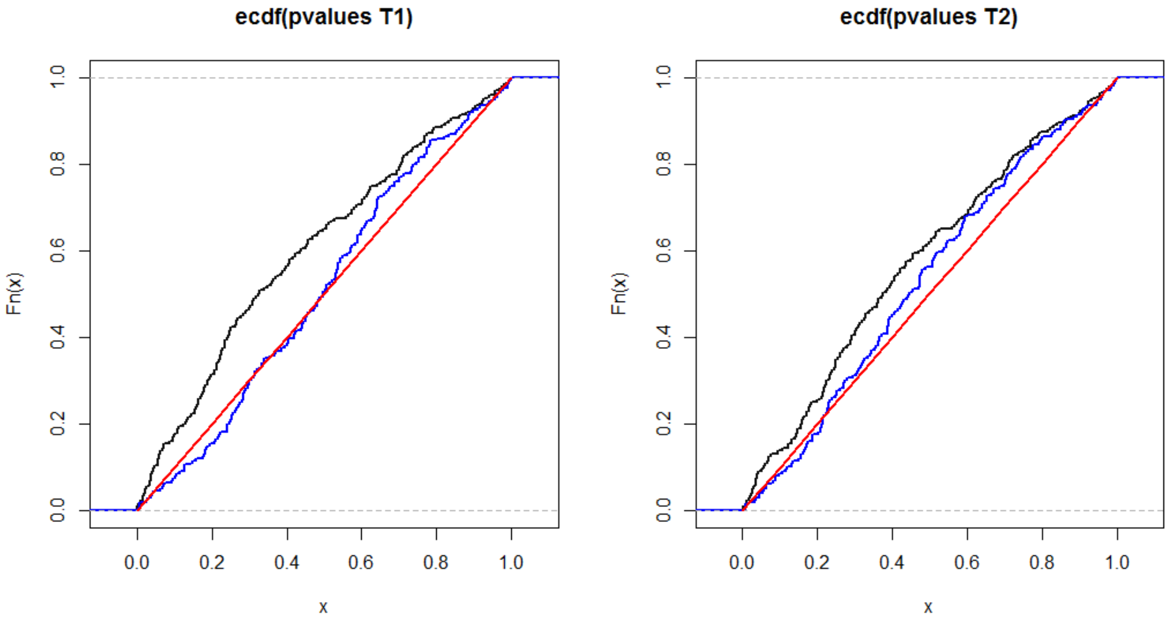

Figure 4.4: ECDF plots of the p-values associated with the test statistics for the method of model selection using AIC. Left panel: p-values associated with test statistic T_1. Right panel: p-values associated with test statistic T_2. The black line represents results from parametric calculations, the blue line represents results from permutation. The red line is the theoretical uniform distribution.

Various Kolmogorov-Smirnov tests are conducted for looking closer at the distribution of the p-values, and the results are found in Table 4.3. The results supports our observations from the ECDF plots, the parametric p-values for T_1 are not uniformly distributed, and hence not from the same distribution as the permutation p-values for T_1. The results from the Kolmogorov-Smirnov tests confirm that, in general, the permutation p-values for both test statistics are better approximations to the uniform distribution under the null hypotheses than the parametric p-values.

The non-uniform distribution of the parametric p-values associated with T_1 must be caused by the model selection step. This will in turn result in many low p-values for the parametric method in real data.

Table 4.3: Kolmogorov-Smirnov test results for simulation by model selection using AIC.

Test data	p-value	Decision on $\alpha = 0.05$ level
$\mathbf{p_1}_{,\text{par}}$ versus uniform	0.04242	Reject H_0
$\mathbf{p_2}_{,\text{par}}$ versus uniform	0.2099	Fail to reject H_0
$\mathbf{p_1}_{,\text{perm}}$ versus uniform	0.9084	Fail to reject H_0
$\mathbf{p_2}_{,\text{perm}}$ versus uniform	0.6807	Fail to reject H_0
$\mathbf{p_1}_{,\text{par}}$ versus $\mathbf{p_1}_{,\text{perm}}$	0.0001387	Reject H_0
$\mathbf{p_2}_{,\text{par}}$ versus $\mathbf{p_2}_{,\text{perm}}$	0.1913	Fail to reject H_0

Summary

To sum up, we have seen that the permutation p-values are approximate uniform distributed. The parametric p-values for T_1 are not uniform distributed, and different from the permutation p-values. The parametric p-values for T_2 are, by the Kolmogorov-Smirnov tests using a significance level of $\alpha = 0.05$, uniform distributed and equal to the permutation p-values.

We have seen that the model selection step is the cause of the non-uniform distribution of parametric p-values associated with T_1. The non-uniformity will thus cause many small p-values in real data, and this might impact the control of the type I errors in real data.

4.3 Remarks

In this present chapter, we have presented our permutation algorithm, and performed two small simulation studies using the algorithm. From the analysis found in Chapter 4.2.1, we have seen that the use of test statistic T_2^* is valid for data generated under H_0^1. The two strategies, predefined fixed model and model selection using AIC, have been compared through the two simulation studies. We found that the step of model selection using AIC produces smaller p-values than the method of using a predefined fixed model. The permutation p-values are found to be approximate uniform distributed, and thus validating the programming code and strategy. The distribution of parametric p-values seems problematic when using model selection, but not for a predefined fixed model.

The next step of this thesis is to apply our permutation algorithm on real data for several genes. In the next chapter, we will present methods used for comparing multiple hypothesis tests. Analysis on several genes using our permutation algorithm and method concerning multiple comparisons are found in Chapter 6.

Chapter 5

Multiple testing

When analysing experimental data such as the gene expression data described in Chapter 3, we might wish to perform hypothesis tests for all genes. An individual test is performed for each null hypothesis. Often the significance level is set to $\alpha = 0.05$, meaning that the probability of making a type I error is controlled at 5% for each hypothesis.

Multiple testing, also called multiple comparisons, is used when we want to test two or more hypothesis simultaneously. This is very much the case with gene expression data, as we have many genes, and thus many null hypothesis, to evaluate. The type I error may be generalised to involve more than one test. In this present chapter, we will look at two possible generalisations called the Family Wise Error Rate (FWER) and the False Discovery Rate (FDR), and present two procedures for control. The methods presented will in Chapter 6 be used for analysis of the entire gene expression data set.

5.1 Type I error rates

From Benjamini & Hochberg (1995) we can summarize the multiple testing problem in Table 5.1. Here V represents the number of type I errors, or false discoveries, and T represents the number of type II errors. The total number of hypotheses is m, and m_0 is the number of true null hypotheses. The number of non-true null hypotheses is m_1, and is equal to $m - m_0$. The total number of not rejected null hypotheses is W, and this number can also be written as $m - R$. The total number of rejected null hypotheses is R. The only known variables are R and m. The goal is to minimize V and T, thus controlling the type I and type II error rates.

Table 5.1: Summary table for the multiple testing problem

	Accepted H_0	Rejected H_0	Total
True H_0	U	V	m_0
Non-true H_0	T	S	m_1
Total	W	R	m

From Table 5.1, we take V the be the number of type I errors among all m hypotheses. By Ge et al. (2003), the family wise error rate is defined as the probability of at least one type I error,

$$\text{FWER} = \Pr(V > 0).$$

A strong control of the FWER is to control the FWER at level α under any combination of true, m_0, and false, m_1, null hypotheses.

The false discovery rate can be defined as the expected proportion of type I errors among the rejected hypotheses, $E(V/R)$. There are different ways to define the expression for the FDR, depending on how they handle the case of zero rejected hypotheses, when $R = 0$. By using the indicator function, we can write

$$\text{FDR} = E\left(\frac{V}{R} \cdot I(R > 0)\right).$$

Note that under the complete null hypothesis, that is, when $m_0 = m$, FDR is equal to FWER. Methods that control the FDR also control the FWER in a weak sense.

5.2 Controlling type I error rate

There have been developed many techniques to control type I error rates in the multiple testing problem. We will look at two methods, the Bonferroni correction and the Benjamini-Hochberg (BH) step-up procedure. The two methods seek to control the FWER and the FDR, respectively. The presentation of this topic is based on Ge et al. (2003) and Benjamini & Hochberg (1995).

Bonferroni correction

The Bonferroni correction is a method of strong level α FWER control, as it seek to reduce the probability of even a single type I error, or false discovery. The procedure is considered the simplest and most conservative method to control the FWER.

The raw p-values from any statistical test are denoted p_i, where $i = 1, ..., m$, and m is the total number of hypotheses to be tested. The Bonferroni correction method will reject any hypothesis H_i with p-value less than or equal to $\frac{\alpha}{m}$. This will control the FWER to be less than or equal to α.

The Bonferroni single-step method will generate adjusted p-values, $\tilde{p}_i$, given by

$$\tilde{p}_i = \min(mp_i, 1).$$

In practice, these adjusted p-values can be found by using the function `p.adjust()` in `R`.

Benjamini-Hochberg procedure

The Benjamini-Hochberg step-up procedure is a method for strong control of the FDR. It is designed to control the expected proportion of false discoveries.

Let $p_{(k)}$ for $k = 1, ..., m$ be the ordered raw p-values, so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. Then do the step-up order, starting from $k = m$, then $k = m - 1$, until $k = 1$. We then define k^* to be the first integer k such that $p_{(k)} \leq \frac{k}{m}\alpha$. No null hypothesis is rejected if k^* is not defined. Otherwise the hypotheses H_k are rejected for $k = 1, ..., k^*$. The corresponding adjusted p-values for the BH step-up procedure are

$$\tilde{p}_k = \min_{l=k,...,m} \{\min(\frac{m}{l} p_{(l)}, 1)\}.$$

In detail, we start by ordering the raw p-values from the smallest to the largest, with indices $k = 1, ..., m$. The smallest raw p-value is $p_{(1)}$ and the largest among all hypotheses is $p_{(m)}$. The threshold value $\dfrac{k}{m}\alpha$ depends on the index k, so this will most likely be unique for all m hypotheses. Starting at $k = m$ (step-up), calculating the threshold value, we compare this with the corresponding p-value $p_{(k)}$. If $p_{(k)} \leq \dfrac{k}{m}\alpha$, then we reject all $H_{(1)}, ..., H_{(k)}$ hypotheses, and stop the procedure. The remaining $H_{(k+1)}, ..., H_{(m)}$ hypotheses are not rejected. If $p_{(k)} > \dfrac{k}{m}\alpha$, we continue with comparing the next p-value $p_{(k-1)}$ to the corresponding threshold value $\dfrac{(k-1)}{m}\alpha$, and evaluate in the same manner. The procedure will always stop after we have found the first p-value that is smaller than or equal to the corresponding threshold value. All hypotheses with index smaller and equal to the index at which the procedure is stopped, will be rejected.

The adjusted p-values using the BH step-up procedure can be found using the function `p.adjust()` in `R`.

Chapter 6

Statistical analysis

In this chapter we will use the methods described in Chapters 3.5, 4 and 5 to analyse the long time series gene expression data set. First we perform parametric tests, as described in Chapters 2 and 3, on the full gene expression data set. Then we perform permutation tests for a randomly selected subset of 1000 genes, where methods described in Chapters 4 and 5 are used. Compared to the analysis from Chapter 3, where a few genes are studies in detail, we will in this chapter look at many genes collectively. The results will provide information about our method of testing the null hypotheses H_0^1 and H_0^2 when applied to a large amount of data.

Chapter 5 describes two methods for controlling type I error rates when testing multiple hypotheses. The Bonferroni correction, controlling the family wise error rate, provides a strong level α control. However, the method most widely used and recognised for gene expression data sets controls the false discovery rate. We will use the Benjamini-Hochberg step-up procedure for controlling the FDR.

6.1 Multiple testing on all genes

We will first test the two hypothesis H_0^1 and H_0^2 on each of the $m = 9856$ genes in our gene expression data set, using the parametric methods of Chapters 2 and 3. The two hypotheses to be tested are specified in Chapters 3.5 and 4.1. The test statistics T_1^* and T_2^*, for each of the hypothesis tests, are calculated by equations (4.1) and (4.2) for each gene after fitting LME models. The models are fitted using two scenarios, one with a predefined fixed model (model 4), and one with model selection using AIC. The results for both the Bonferroni correction and the Benjamini-Hochberg procedure, using $\alpha = 0.05$, are found in Table 6.1.

The percentage of significant genes for the Bonferroni correction method is calculated with level α. Any hypothesis with p-value less than or equal to $\dfrac{\alpha}{m} \approx 5.07 \cdot 10^{-6}$ is rejected. We see that controlling the FWER compared to the FDR gives fewer significant genes.

Control of the FDR is the preferred method for controlling the type I error rates for gene expression data. We see from the results of testing the first hypothesis H_0^1 that 60% of the genes are declared significantly differentially expressed, for the method of model selection using AIC. The opinion of the biologists in this project is that 60% differentially expressed genes is too high. Testing the second hypothesis H_0^2 of the estimated area, however, the percentage of significant genes is nearly cut in half, to 31.6%. This is a major improvement.

We believe that model selection using AIC is the preferred method over a predefined fixed model for fitting LME models. This is due to the fact that the variation in the gene expression data for different genes is large, thus forcing on a "one model fits all" philosophy does not make sense. Instead, various models should be fitted, and by the AIC value criteria, the best fitted model is chosen for each gene expression time series. Comparing the results using the two methods, we see that the method of model selection using AIC provides a higher percentage of significant genes. This is explained by the model selection step for choosing the best fitted models in advance of calculating test statistics. It is thus expected that the number of significantly differentially expressed genes is higher when using model selection compared to using a fixed model.

Table 6.1: Multiple testing of the two hypotheses H_0^1 and H_0^2 on all 9856 genes, using the Bonferroni correction and the Benjamini-Hochberg procedure, with $\alpha = 0.05$. The results are presented as a percentage of all 9856 genes.

	$H_0^1: \quad \boldsymbol{\beta} = 0$	$H_0^2: \quad A = 0$
Bonferroni (FWER):		
p-value, fixed model 4	5.4 %	1.8 %
p-value, model selection using AIC	6.2 %	2.9 %
BH procedure (FDR):		
p-value, fixed model 4	58 %	25.2 %
p-value, model selection using AIC	60 %	31.6 %

6.2 Multiple testing on a randomly selected subset

In a similar manner as Chapter 6.1, the two hypothesis tests are carried out on 967 randomly selected genes from our data set. We have performed analysis based on both parametric and permutation methods. We have used both schemes for fitting LME models, the predefined fixed model (model 4) and model selection using AIC.

Ideally, we would study all 9856 genes. However, analysis show that when using the scheme of model selection with AIC, it takes about 24 hours to fit models and calculate test statistics for 100 genes with $B = 10000$ permutations each, on a Xeon 2.67 GHz (Intel CPU) running Linux (Ubuntu 10.4). The time is reduced to about 7 hours when using a predefined fixed model. Sampling all genes would take approximate 90 days to process. Therefore, we sampled 500 genes twice (with different seed for the randomisation), running the calculations on two different servers. The result is 967 unique genes randomly selected from our data set.

The `R` code for the analysis is found in Appendix A. A summary in pseudo code is found below.

```
Make B = 10000 permutations by random shuffle of time points

For (id in 1:967)
{
     Fit LME model to original data using
     a) predefined fixed model 4, and
     b) model selection using AIC
     Calculate parametric test statistics
     Calculate parametric p-values

     For (b in B)
     {
          Construct a new data set using the permutations
          Fit LME model using a) and b)
          Calculate permutation test statistics for each permutation
     }

     Calculate permutation p-values
}
```

Controlling the family wise error rate

We would like to compare the results on the original data in Table 6.1, with the results from the $m = 967$ randomly selected genes. This is not sensible to do for adjusted permutation p-values using the Bonferroni correction, since the result is 0% significant genes with level $\alpha = 0.05$. This is explained by the threshold value $\dfrac{\alpha}{m} \approx 5.17 \cdot 10^{-5}$, and the lowest possible p-value from the permutations; $\dfrac{1}{10001} \approx 10^{-4}$. Thus, none of the hypotheses can be rejected based on permutation p-value.

Instead we have chosen to report the number of genes with a p-value below 0.05, just to provide a comparison between the parametric and permutation methods. The results, in percentages of all 967 genes, are found in Table 6.2. We see that in general, the numbers from the permutation p-values are lower compared to the parametric numbers.

Table 6.2: Results from testing the two hypotheses H_0^1 and H_0^2 on a sample of 967 genes, without using the methods of multiple testing. The numbers are the percentage of all 967 genes with p-values below $\alpha = 0.05$.

	$H_0^1: \quad \boldsymbol{\beta} = 0$	$H_0^2: \quad A = 0$
Parametric p-values:		
Predefined fixed model 4	66.7%	39.9%
Model selection using AIC	64.5%	43.0%
permutation p-values:		
Predefined fixed model 4	37.0%	23.5%
Model selection using AIC	31.0%	20.3%

Controlling the false discovery rate

The results from the BH procedure for controlling the FDR are found in Table 6.3. Comparing the values from the parametric p-values with the p-values from Table 6.1, we see that the results are very similar. With this, we will assume that the random sampling of 967 genes is a good representative, reflecting the original data set well.

Table 6.3: Multiple testing of the two hypotheses H_0^1 and H_0^2 on a sample of 967 genes, using the Benjamini-Hochberg procedure for controlling the FDR with $\alpha = 0.05$. The results are presented as a percentage of all 967 genes.

	$H_0^1: \quad \boldsymbol{\beta} = 0$	$H_0^2: \quad A = 0$
Parametric p-values:		
Predefined fixed model 4	60.1%	26.8%
Model selection using AIC	61.5%	33.5%
Permutation p-values:		
Predefined fixed model 4	11.7%	2.6%
Model selection using AIC	7.7%	2.7%

For both original and parametric results, the method of using a predefined fixed model gives a lower number of significant genes compared to the method of model selection using AIC. This case is, surprisingly, the opposite of what the results for the permutation p-values show, comparing 11.7% with 7.7% under H_0^1. The results for permutation p-values under H_0^2, 2.6% and 2.7%, shows that the two methods provides approximate equal percentages.

ECDF plots for the p-values from model selection using AIC are found in Figure 6.1. The left panel shows the p-values associated with test statistic T_1. We see that for both parametric (black line) and permutation (blue line) methods, the p-values lie above the theoretical uniform distribution. Thus, both methods produce small p-values. Although not as extreme as for test statistic T_1, we see that the p-values associated with test statistic T_2 in the right panel of Figure 6.1, also lie above the theoretical uniform distribution. This is a result of using the BH step-up procedure for controlling the FDR.

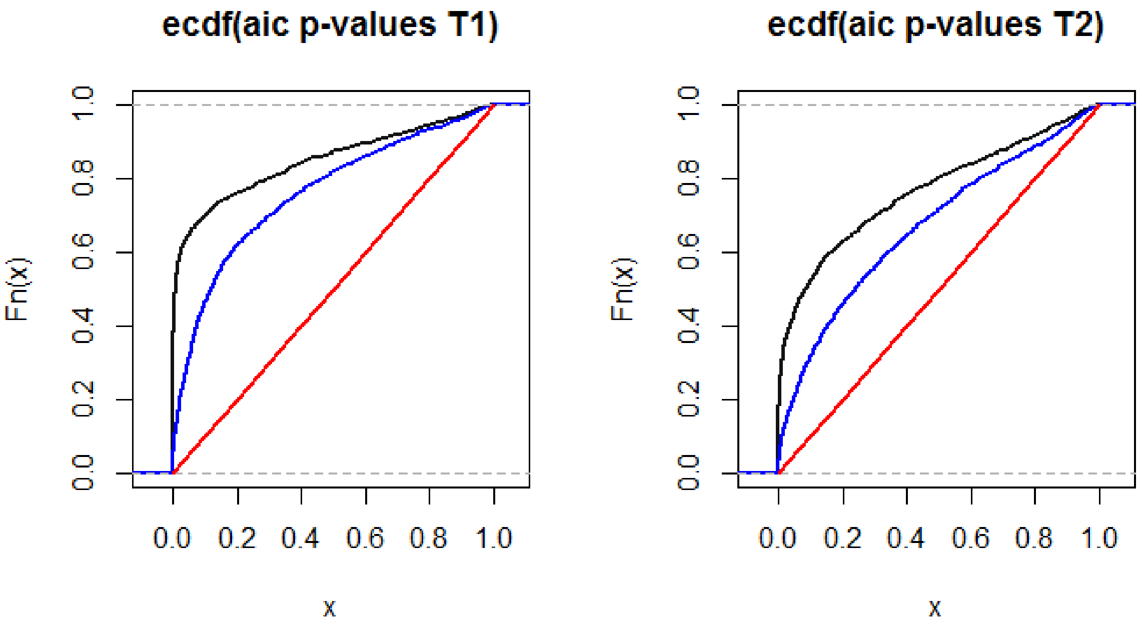

Figure 6.1: ECDF plot of the p-values using the Benjamini-Hochberg procedure for controlling the FDR, with $\alpha = 0.05$, for the model selection set-up. Left panel: p-values associated with test statistic T_1. Right panel: p-values associated with test statistic T_2. The black line represents parametric results, the blue line represents permutation results. The red line is the theoretical uniform distribution.

It would be interesting to analyse the 2.7% genes in detail, but this is a topic for further research and is discussed in Chapter 7. However, we find that the permutation results of 2.6% and 2.7% under H_0^2, are subsets of the corresponding parametric results of 26.8% and 33.5%. This is also the case under H_0^1, where the percentage of significant genes using permutation is a subset of the corresponding parametric percentages.

We also find the results for H_0^2 to be a subset of the corresponding results for H_0^1, as presented in Table 6.4. From the biologists in this project, we know that the number of 60% significant genes is an overestimation. By testing H_0^2, the number of significant genes are reduced. For parametric tests, the number is almost cut in half. For permutation tests, the reduction is even more drastic. From the results presented in Table 6.4, we see that the significant genes under H_0^2 also also marked as significant under H_0^1. We may assume that the significant genes under H_0^2, is a better estimate of which genes are actually found significant by our biologists, using other detailed analysis than presented in this thesis. Thus, we find that the use of H_0^2 is an appropriate method for assessing significance of differentially expressed genes in our gene expression study.

Table 6.4: Subset analysis of the results from the Benjamini-Hochberg procedure for controlling the FDR on 967 randomly selected genes.

Analysed subset
36.8% is a subset of 60.1%
22.3% is a subset of 61.5%
2.6% is a subset of 11.7%
2.7% is a subset of 7.7%

Using p-values as test statistics

Besides evaluating the results from multiple testing, we want to check whether the adjusted test statistics T_1 and T_2 are appropriate to use instead of the well-known F and t statistics T_1^* and T_2^*, respectively. We thus calculate new test statistics T_1^{new} and T_2^{new} to be the negative of the parametric p-values for the F and t statistics, and use these in the permutation algorithm. Using T_1^{new} and T_2^{new} give very similar results as using T_1 and T_2. We may conclude that T_1 and T_2 are appropriate to use in our model selection setup.

6.3 Remarks

We have done analysis of all 8956 genes using the Bonferroni correction method and the BH step-up procedure for controlling the FWER and FDR, respectively. We randomly selected 967 genes to represent the whole gene expression data set. None of the hypotheses tested could be rejected by controlling the FWER using the Bonferroni method. For control of the FDR using the BH procedure, we found the results from parametric and permutation to be different.

Initially, we believed that the model selection strategy would be a better strategy than the fixed model, even though the results using model selection gives a higher percentage of significantly differentially expressed genes. Before the analysis using permutation, we also assumed this method to provide better estimates of significantly expressed genes than the parametric approach.

The biologists in our project know that the result of 60% significantly expressed genes under H_0^1 is an overestimation. This finding is one of the main reasons why we have suggested the alternative test of H_0^2 to assess significance. For our data set, the biologists believe that the number of significantly differentially expressed genes is in the range of $1500-3300$ genes. In percentages, this is approximately $16.7-36.8\%$. These findings are in accordance with research on similar gene expression problems.

When comparing our findings with the expectation from the biologists, we see that the parametric test of H_0^2 seems realistic. Our initial thought of the model selection step being the better strategy than using a fixed model, does not seem to be correct. The fixed model strategy provides less significantly expressed genes than the model selection step. The results from permutation seems to be an underestimation, when comparing our result with the actual findings of the biologists. The size of B might influence the results from permutation. This is discussed in Chapter 7.

Chapter 7

Discussion and conclusion

In this thesis we have suggested an alternative statistical hypothesis test for assessing significant activation of genes over time, using the area under the estimated curves as a measure of consistent activity over time. We have used the framework of linear mixed effects models for our gene expression data set. Together with assumptions concerning the parametric distribution of the area, a hypothesis test is suggested. Detailed analysis of model fit and hypothesis tests for some genes are presented, and multiple hypotheses tests on several genes are conducted. Methods of simulation and permutation are used as verification. In addition, adjusted test statistics for the hypothesis tests are suggested and analysed.

In the present chapter, we will discuss some of the methods and results. A conclusion based on our observations throughout this thesis is made.

7.1 The LME model for gene expression data

We have used the LME model for our data set. This model is widely used in various biostatistical contexts. Compared to other linear models, the random effects of the LME model seems appropriate for data sets with repeated measurements. As we have seen in the detailed analysis of data for some genes in Chapter 3, the random effects are quite small, in the order of 10^{-11}. The contributions from the random effects are thus very small in the fitted LME models. The choice of LME models were based on the whole gene expression data set, not only the difference data set which we have studied in this thesis. For the gastrin treatment gene expression time series and the unstimulated controls, the random effects have been seen to take on values in the order of 10^{-3} to 10^{-1}.

In Chapter 3, we present detailed analysis of model fit for data from three genes. This is done to illustrate some of the extremities, such as the largest estimated area and the poorest normal approximation of the residuals we found in our data set. From detailed analysis of the gene expression data and the fitted LME models for several genes, we find the LME model to be a good fit for our gene expression data.

7.2 Choice of B in permutations

In this thesis, we have used the non-parametric method of permutation. Permutations are made by randomly shuffling time points to create data under the null hypotheses we have tested. Ideally we would do a complete permutation test, including all $12! = 4.79 \times 10^8$ possible permutations for our gene expression time series consisting of 12 time points. The computational costs for a complete permutation test is high, and in practice it is not a real possibility to complete. We have thus chosen to use sampled permutation tests with $B = 10000$ permutations. There are some limitations using this few permutations in a multiple testing setting, as we will now discuss .

Controlling the FWER for permutations based on the Bonferroni correction method

With the chosen $B = 10000$ for the randomly selected subset of $m = 967$ genes, we have seen that the smallest possible p-value from the permutations will never be smaller than $\frac{\alpha}{m}$. None of the hypotheses tested are rejected. We find the minimum number of permutations to be $B = 20000$ for the p-values to be just small enough, with level $\alpha = 0.05$ control of the FWER.

How many permutations are needed for a permutation study of approximate $m = 9000$ genes? If we would like to use a Bonferroni cut-off, we have $\frac{\alpha}{m} \approx 5.56 \cdot 10^{-6}$, so the minimum number of permutations needed for the p-values to be just small enough, is $B = 180000$. The time needed for conducting 10000 permutations on 100 genes was estimated to be approximate 24 hours. A permutation test with $B = 180000$, or preferably more, on 9000 genes is thus regarded as impossible in reality.

A control of the FWER is more strict than a control of the FDR. In the case of permutation tests on large amount of data, we deem the control of FWER impossible. Controlling the FDR is thus regarded as a better method for multiple tests on gene expression data.

Controlling the FDR for permutations based on the BH step-up procedure

In Chapter 6.2, we saw that the result from permutation on a sample of genes, for H_0^2 with model selection using AIC, gave a percentage of 2.7 significant genes. When using a fixed model, the result is 2.6. Comparing these results with the corresponding parametric results of 33.5% and 26.8%, respectively, we see that the permutation gives a fairly large reduction. The biologists believe that the 2.7% and 2.6% is an underestimation.

The question is then, what if the number of permutations increased to, say, 100000 or one million? Would the amount of significant genes then increase, decrease or remain the same? We saw that for the Bonferroni correction method, the number B is important, and we would assume that it would also influence the result in a BH procedure. As mentioned before, looking at the results in Table 6.2, the trend is that the number of significantly differentially expressed genes are higher when using the parametric approach compared to using the permutation approach. This result suggests that the number of significant genes associated with test statistic T_2 for permutations, should be lower than for the corresponding parametric p-values. Answering these questions could be an interesting topic for further research.

7.3 A predefined fixed model versus model selection using AIC

During the work of this thesis, we have used two different strategies for fitting LME models in parallel. In Chapter 4.2, we did a simulation study based on data from one gene, where the only difference in the two strategies was the introduction of model selection using AIC in Chapter 4.2.2. From results and plot in Chapter 4.2.2, we saw that the method of model selection using AIC produces small parametric p-values. The effect of model selection using AIC is thus large, even larger than we anticipated in advance. Would the use of model selection generate many type I errors in real data? A simulation experiment using data not only from one gene, but from several, or all, genes in our data set might give us an answer to this question. Due to the time frame of this thesis, this is a suggestion for further study.

Our first impression is that the model selection method is the preferred one over a predefined fixed model. Our gene expression data should be fitted to a model which is "best fitted", not just fitted to the same model regardless of how the data behaves. From the results presented in Table 6.3, we have seen that the percentages of significant genes from permutation under H_0^2 are very similar for the two methods, equal to 2.6% and 2.7% for a fixed model and model selection, respectively. This suggests that the preferred scheme for permutations is using a predefined fixed model, since this is a less time consuming and computationally intensive approach than the model selection scheme.

When comparing the two schemes for the parametric results, we see a greater difference than for the permutation results. Under H_0^1, the results from a predefined model and model selection are also very similar. A greater difference is seen for the corresponding results under H_0^2. The predefined model, with the lowest number of significantly differentially expressed genes, seems to be the better of the two schemes when performing parametric calculations.

7.4 Communication with biologists

We have looked at the estimated area under the fitted gene expression curves as a measure of consistent activity over time. A parametric approach have been presented in Chapters 2 and 3. The calculation of the estimated area after fitting LME models to the gene expression data is fairly simple, and can be done on large amount of gene expression data with little cost in computation time. The estimated area may be used as a measurement for ranking genes with respect to effect size of activation over time.

Besides using a parametric approach to the area, we have used the non-parametric approach of permutation as presented in Chapters 4 and 5. In addition to the widely used hypothesis test H_0^1 concerning the fixed effects, we have suggested a second hypothesis test H_0^2, concerning the area. Based on our observations throughout this thesis, we recommend testing the area as in H_0^2, in addition to, or rather than testing H_0^1. The permutation approach under both H_0^1 and H_0^2 seems to give underestimations. The parametric approach when testing the area seems to provide a better result with respect to the number of significantly differentially expressed genes, compared to testing the fixed effects as in H_0^1.

The biologists have findings suggesting that the percentage of significantly differentially expressed genes in our data set should be in the range of $16.7 - 36.8\%$. From the results in Chapter 6, we may conclude that using a parametric approach when testing the area under H_0^2 is in accordance with the findings of the biologists.

7.5 Conclusions

The aim of this thesis have been to find a ranking method for genes in a gene expression study, and to assess significance of differentially expressed genes. For this purpose, we have looked at the estimated area under the fitted gene expression curves for our difference gene expression time series.

We have introduced adjusted test statistics T_1 and T_2, which have the advantage of being comparable for fitted models with different degrees of freedom, unlike the F and t statistics. Based on the analysis in Chapter 6.2, we have concluded that the adjusted test statistics T_1 and T_2 are appropriate to use in our model selection set-up.

We have presented the estimated area under the fitted difference gene expression curves as a measure of consistent activity over time. This area provides an easy, fast and cheap way of ranking genes in a gene expression study with respect to effect size of activation over time.

With the use of the area, we have constructed a hypothesis test $H_0^2 : A = 0$. The test may be carried out using both a parametric and a permutation approach, although the parametric approach seem to give more realistic results than the permutation. When we in addition to this take computational costs into account, which we have seen to be approximate four times greater for the model selection step compared to the fixed model, our recommendation is thus to use a predefined fixed model to perform parametric hypothesis tests H_0^2 on observed gene expression data.

Bibliography

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methododical)* **57**(1), 289–300.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, 2nd edn, Duxbury Thomson Learning.

Dobbin, K. & Simon, R. (2005), 'Sample size determination in microarray experiments for class comparison and prognostic classification', *Biostatistics* **6**(1), 27–38.

Fjeldbo, C. S. (2012), Gastrin-mediated regulation of gene expression; a systems biology approach, PhD thesis, Norwegian University of Science and Technology.

Galecki, A. T. & Burzykowski, T. (2013), *Linear Mixed-Effects Models using R, A Step-by-Step Approach*, Springer Science+Business Media.

Ge, Y., Dudoit, S. & Speed, T. P. (2003), 'Resampling-based multiple testing for microarray data analysis', *Sociedad de Estadística e Investigacion Operativa Test* **12**(1), 1–77.

Johnson, R. A. & Wichern, D. W. (2007), *Applied Multivariate Statistical Ananlysis*, 6th edn, Pearson Prentice Hall.

McCulloch, C. E. & Neuhaus, J. M. (2011), 'Prediction of random effects in linear and generalized linear models under model misspecification', *Biometrics* **67**, 270–279.

Nobre, J. S. & da Motta Singer, J. (2007), 'Residual analysis for linear mixed models', *Biometric Journal* **49**(6), 863–875.

Page, C. M. (2012), Estimating time-continuous gene expression profiles using the linear mixed effects framework, Master's thesis, Norwegian University of Science and Technology.

Pinheiro, J. C. & Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, Springer.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Østgård, E. T. (2011), Statistical modeling and analysis of repeated measures, using the linear mixed effects model, Master's thesis, Norwegian University of Science and Technology.

West, B. T., Welch, K. B. & Galecki, A. T. (2007), *Linear Mixed Models, A Practical Guide Using Statistical Software*, Chapman & Hall/CRC.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. & Smith, G. M. (2009), *Mixed Effects Models and Extensions in Ecology with R*, Springer Science+Business Media.

Appendix A

R code

A.1 The data set

The data set used in this thesis is the called `GmUn[i,]`, a large matrix representing the difference time series of gene expression data. Here $i = 1, ..., 8956$ is the total number of genes. The polynomial scaling in time is called `pbasis12`, and `biol12` is a vector containing the two possible biological replicates BR” and BR3. A printout from `R` of `pbasis12`, `biol12` and an example of `GmUn[i,]` is found below.

```
> GmUn[i=2772,]
     BR2_Un_0min_1   BR2_G17_15min_3    BR2_G17_30min_5   BR2_G17_60min_7
       0.000000000      -0.173953742        0.327978438      -0.007784667
   BR2_G17_90min_9 BR2_G17_120min_11  BR2_G17_240min_13 BR2_G17_360min_15
      -0.362210346      -0.227695180       -0.003444469      -0.223049983
 BR2_G17_480min_17 BR2_G17_600min_19  BR2_G17_720min_21 BR2_G17_840min_23
       0.093410069      -0.138779868        0.015321052      -0.034654529
     BR3_Un_0min_1   BR3_G17_15min_3    BR3_G17_30min_5   BR3_G17_60min_7
       0.000000000      -0.025587692       -0.105694925      -0.265250216
   BR3_G17_90min_9 BR3_G17_120min_11  BR3_G17_240min_13 BR3_G17_360min_15
      -0.168058293      -0.069448713       -0.252837199      -0.078891159
 BR3_G17_480min_17 BR3_G17_600min_19  BR3_G17_720min_21 BR3_G17_840min_23
       0.131747867       0.011070331       -0.070914042      -0.304416634

> biol12
 [1] "BR2" "BR2" "BR2" "BR2" "BR2" "BR2" "BR2" "BR2" "BR2"
[10] "BR2" "BR2" "BR2" "BR3" "BR3" "BR3" "BR3" "BR3" "BR3"
[19] "BR3" "BR3" "BR3" "BR3" "BR3" "BR3"
```

```
> pbasis12
           1          2            3            4
1  -1.0000000 1.00000000 -1.000000000 1.0000000000
2  -0.9642857 0.92984694 -0.896638120 0.8646153296
3  -0.9285714 0.86224490 -0.800655977 0.7434662641
4  -0.8571429 0.73469388 -0.629737609 0.5397750937
5  -0.7857143 0.61734694 -0.485058309 0.3811172428
6  -0.7142857 0.51020408 -0.364431487 0.2603082049
7  -0.4285714 0.18367347 -0.078717201 0.0337359434
8  -0.1428571 0.02040816 -0.002915452 0.0004164931
9   0.1428571 0.02040816  0.002915452 0.0004164931
10  0.4285714 0.18367347  0.078717201 0.0337359434
11  0.7142857 0.51020408  0.364431487 0.2603082049
12  1.0000000 1.00000000  1.000000000 1.0000000000
13 -1.0000000 1.00000000 -1.000000000 1.0000000000
14 -0.9642857 0.92984694 -0.896638120 0.8646153296
15 -0.9285714 0.86224490 -0.800655977 0.7434662641
16 -0.8571429 0.73469388 -0.629737609 0.5397750937
17 -0.7857143 0.61734694 -0.485058309 0.3811172428
18 -0.7142857 0.51020408 -0.364431487 0.2603082049
19 -0.4285714 0.18367347 -0.078717201 0.0337359434
20 -0.1428571 0.02040816 -0.002915452 0.0004164931
21  0.1428571 0.02040816  0.002915452 0.0004164931
22  0.4285714 0.18367347  0.078717201 0.0337359434
23  0.7142857 0.51020408  0.364431487 0.2603082049
24  1.0000000 1.00000000  1.000000000 1.0000000000
```

A.2 Fitting an LME model

The R code for fitting LME models using the function nlme as suggested by Pinheiro & Bates (2000) is found below. The methods of ML and REML can be specified, and are used as described in Chapter 2.7.

```
library(nlme)
orgyvalue <- GmUn[i,]

lme1 <- lme(orgyvalue~pbasis12[,1], random=~1|biol12,
                method = "REML", na.action = na.omit)
lme2 <- update(lme1, orgyvalue~pbasis12[,1:2])
lme3 <- update(lme1, orgyvalue~pbasis12[,1:3])
lme4 <- update(lme1, orgyvalue~pbasis12[,1:4])

AICvecks<- c(AIC(lme1),AIC(lme2),AIC(lme3),AIC(lme4))
mod <- which.min(AICvecks)

thislme <- get(paste("lme",mod,sep=""))
```

Plotting raw data

Here we present the `R` code for plotting the raw data, scaled time versus gene expression:

```
plot(pbasis12[1:ntp],GmUn[i,1:ntp],pch=20,col="blue",
                      xlab="Scaled time", ylab="GmUn")
points(pbasis12[1:ntp],GmUn[i,(ntp+1):(2*ntp)],pch=20,col="red")
```

Diagnostics

When the LME models are fitted, we continue with the best fitted model according to the criteria of lowest AIC value, and do diagnostics. Residual plots are made by the commands `plot()` and `qqnorm()`, taking an LME object as argument. The types of residuals (marginal, conditional, pearson, etc) can be specified. Predictions are made using `predict()`.

A function to calculate the IntraClass Correlation (ICC) is made. The function takes as argument an LME object and returns the ICC value of this object.

```
ICC <- function(lmeobj)
{
  res <- as.numeric(VarCorr(lmeobj))
  return(res[1]/(res[1]+res[2]))
}
```

Anderson-Darling tests are used as support for other primary diagnostic tools such as residual plots. The test is in general used on a given sample of data, to test if the sample is drawn from a given probability distribution. We have used the test on residuals to check for normality. The null hypothesis tested is then ”The given sample is drawn from the normal probability distribution”. We use a significant level of 0.05 for decisions regarding the p-values. The test is applied to a sample by using the command `ad.test()`, found in the package ”Nortest”.

A.3 Calculating test statistics and p-values

The framework of the function `calctestobsWP` is used for multiple purposes, including permutation tests, simulation and calculating test statistics and p-values. Several functions are made, for each purpose, and all of them are stored in the file called ”calctestobs.R”. We will give an example of the function which is used for permutation and simulation with model selection using AIC. The function returns test statistics, p-values, model number and the estimated area of the fitted model.

```
calctestobsWP <- function(yval, pbas, bio, Acoeffs)
{
  # Fit LME
  plme1 <- lme(yval~pbas[,1],random=~1|bio,method="REML",na.action=na.omit)
  plme2 <- update(plme1, yval~pbas[,1:2])
  plme3 <- update(plme1, yval~pbas[,1:3])
  plme4 <- update(plme1, yval~pbas[,1:4])
  AICvecperm <- c(AIC(plme1),AIC(plme2),AIC(plme3),AIC(plme4))
  mod <- which.min(AICvecperm)
  plme <- get(paste("plme",mod,sep=""))

  # t1: F-test (hyp H_0^1)
  numDF<- anova(plme)$"numDF"[1]
  denDF <- anova(plme)$"denDF"[2]
  fF <- anova(plme)$"F-value"[2]
  peF <- denDF/(denDF-2)
  varF <- (2*denDF^2*(numDF+denDF-2))/(numDF*(denDF-2)^2*(denDF-4))
  t1 <- (fF-peF)/(sqrt(varF))

  # t2: t-test (hyp H_0^2)
  permcoeffs <- c(plme$coefficients$fixed,rep(0,4-mod))
  estAreal <- sum(permcoeffs*Acoeffs)
  pcc <- Acoeffs[(1:(mod+1)),,drop=FALSE]
  sigmamatrix <- plme$varFix
  pvarg <- as.vector(t(pcc)%*%sigmamatrix%*%pcc)
  t0 <- (estAreal)/(sqrt(pvarg))
  vart0 <- denDF/(denDF-2)
  t2 <- t0/(sqrt(vart0))

  # Other p-values for original data:
  pvalAt <- 2*pt(abs(estAreal)/sqrt(pvarg),denDF,lower.tail=FALSE)
  pvalAnorm <- 2*pnorm(2*abs(estAreal),0,sqrt(pvarg),lower.tail=FALSE)
  pvalANOVA <- anova(plme)$"p-value"[2]

  return(list(t1=t1,t2=t2,mod=mod,t1p=pvalANOVA,t2p=pvalAt,area=estAreal))
}
```

A.4 Simulation study

This is the R code used in Chapter 4. The two methods for simulation use the same framework, but with some differences. We only print the code for the simulation described in Chapter 4.2.2.

As support for other diagnostic tools such as the ECDF plots, we have used Kolmogorov-Smirnov (K-S) tests ks.test(). K-S is a test for equality of probability distributions. By using a two-sample K-S test, we compare two samples. The null hypothesis of the K-S test is "Samples are drawn from the same distribution", and we use significance level 0.05 to make decisions.

```
load("lmeImage")       # Data set
source("calctestobs.R")
library(nlme)

K <- 1e2
B <- 1e4
SUB <- 1

Acoeffs <- matrix(c(0, 2, -4/3, 2, -8/5), ncol=1)
permuts <- matrix(data = NA, nrow = B, ncol = 12)

Ogeneset <- setdiff(1:ngenes, 5044)
set.seed(123)
geneset <- sample(Ogeneset, SUB, replace=FALSE)

set.seed(233)
for(b in 1:B)
{
  permuts[b,] <- sample(1:12, 12, rep = FALSE)
}

# could have been loop here
id <- geneset[1]
orgyval <- GmUn[id,]

# parameters used to make model, use model 4 - partly
plme <- lme(orgyval~pbasis12[,1:4], random=~1|biol12,
              + method =  "REML", na.action = na.omit)
mod <- 4
beta0 <- plme$coeff$fixed[1]
sigma <- plme$sigma
sigmaB <- as.numeric(VarCorr(plme)[1,2])
orgtestobs <- calctestobs4WP(orgyval, pbasis12, biol12, Acoeffs)
```

```
# now loop over K data sets,
newmu <- rep(beta0, 24)

newtestobs <- matrix(nrow=K, ncol=5)
permpvals <- matrix(ncol=2, nrow=K)
set.seed(9876)
for (k in 1:K)
   {
    print(k)
    u1 <- rep(rnorm(1, 0, sigmaB), 12)
    u2 <- rep(rnorm(1, 0, sigmaB), 12)
    eps <- rnorm(24, 0, sigma)
    newy <- newmu + c(u1,u2) + eps
    # now need y=0 for first obs for each series
    newyJ <- newy
    newyJ[1:12] <- newy[1:12] - newy[1]
    newyJ[13:24] <- newy[13:24] - newy[13]
    newy <- newyJ
    newtestobs[k,] <-unlist(calctestobsWP(newy,pbasis12,biol12,Acoeffs))
    resmat <- matrix(ncol=3, nrow=B)

    for (b in 1:B)
        {
          permyvalue <- newy[c(permuts[b,], permuts[b,]+12)]
          permtestobs <- calctestobs(permyvalue, pbasis12, biol12,Acoeffs)
          resmat[b,1] <- permtestobs$t1
          resmat[b,2] <- permtestobs$t2
          resmat[b,3] <- permtestobs$mod
        }

    permt1p <- (sum(resmat[,1 ]>= newtestobs[k,1])+1)/(B+1)
    permt2p <- (sum(abs(resmat[,2]) >= abs(newtestobs[k,2]))+1)/(B+1)
    permpvals[k,] <- c(permt1p, permt2p)
    cat(permpvals[k,], "\n", file="rescheckAIC", append=T)
    }

dput(list(newtestobs,permpvals),paste("simcheckAIC",id,"res.dd",sep=""))
```