NTNU

Norwegian University of
Science and Technology

Statistical Methods for early Prediction of Cerebral Palsy based on Data from Computer-based Video Analysis

Martina Hall

Master of Science in Physics and Mathematics

Submission date: June 2017

Supervisor: Mette Langaas, IMF

Co-supervisor: Turid Follestad, Enhet for anvendt klinisk forskning, NTNU
Lars Adde, Institutt for laboratoriemedisin, barne- og kvinnesykdommer

Norwegian University of Science and Technology
Department of Mathematical Sciences

Summary

In this thesis we have investigated the association between summary variables from the General Movement Toolbox (GMT), and cases of abnormal fidgety movements (FMs) and cerebral palsy (CP). The GMT-software calculates summary variables from video recordings of infants movements, based on changes of pixel-values between video frames. In previous studies, low values of the variable for variations in the centroid of motion (C_{sd}) have shown to predict both normal FMs and no CP. However, these results were carried out for small datasets, consisting of only Norwegian infants.

Here, we have used data from 693 infants with a total of 798 video recordings from Norway, USA and India. We have use both a frequentist approach with the glmer()-function from the lme4-package and a Bayesian approach with INLA-package, for prediction of FMs in R. Due to repeated measurements, we used a mixed effects logistic regression model with random intercepts, with the C_{sd} variable as covariate. We have also used the same variable in a logistic regression model for prediction of CP. For both models we found the same association as in previous studies, but the effect of C_{sd} on the occurrence of normal FMs varied between countries. To investigate the stability and the uncertainty of the frequentist FM-model for different number of repeated measurements, a simulation study was performed. The results showed that having many observations without repeated measurements could cause unstable results with large confidence intervals for the estimated coefficients. However, for only two or more repeated measurements the estimated coefficient values were much more stable and the size of the confidence intervals were reduced considerably.

In the search for a better model for predictions of CP, we included several GMT-variables and other available variables, and used the Lasso method for variable selection. The results here showed that it was in fact the y-direction of the C_{sd} variable that was associated with the occurrence of CP, but also the mean value in the y-direction of the centroid of motion, mean and standard deviation variables of the area of motion and the standard deviation of the quantity of motion. Inclusion of other available variables increased the model fit a bit. The gender and an indication variable for extreme preterm infants were selected in the model. In addition, the length of the video recordings were accounted for. However, statistical inference, in the form of bootstrapping and the multi sample-splitting method, showed that only the mean value of the centroid of motion in y-direction had a statistically significant association with the occurrence of CP.

Sammendrag

I denne oppgaven har vi undersøkt assosiasjonen mellom oppsummerende variabler fra the General Movement Toolbox (GMT) programvaren, og tilfeller av unormale "fidgety" bevegelser (FMs) og cerebral parese (CP). GMT-programvaren beregner oppsummerende variabler fra videoopptak av bevegelsene til spedbarn, basert på endringer i pixelverdier mellom bildene i videoopptaket. I tidligere studier har lave verdier av variabelen for variasjonen i massesentret av bevegelsene (C_{sd}) vist seg å predikere både normale FMs og ingen CP. Disse studiene ble imidlertid utført på små datasett, kun bestående av norske spedbarn.

Her har vi brukt data fra 693 spedbarn med totalt 798 video opptak, fra Norge, USA og India. Vi har brukt både en frekventistisk metode med glmer()-funksjonen i lme4-pakken og en Bayesiansk metode med INLA-pakken, for prediksjon av FMs i R. På grunn av repeterte målinger brukte vi en blandet effekts logistisk regresjonsmodell med et tilfeldig skjæringspunkt, med C_{sd} som kovariat. Vi brukte også samme variabel i en logistisk regresjonsmodell for prediksjon av CP. Vi fant samme assosiasjon som i de tidligere studiene, men effekten av C_{sd} på forekomsten av normale FMs varierte mellom landene. For å undersøke stabiliteten og usikkerheten for den frekventistiske FM-modellen for ulikt antall kvadraturpunkt, utførte vi en simuleringsstudie. Resultatene her viste at mange observasjoner uten repeterte målinger kunnne føre til usabile resultat og store konfidensintervall for de estimerte koeffisientene. For kun to eller flere repeterte målinger, ble imidlertid de estimerte verdiene for koeffisientene mye mer stabile og størrelsen på konfidensintervallene ble betraktelig redusert.

For å finne en bedre modell for prediksjon av CP, inkluderte vi mange av GMT-variablene og andre tilgjengelige variabler, og brukte Lasso-metoden for variabelseleksjon. Resultatene her viste at det var y-retningen av C_{sd} variabelen som var assossiert med CP, men også gjennomsnittsverdien i y-retning av massesenteret for bevegelse, gjennomsnitts- og standardavviks-variabelene for arealet av bevegelsene, samt standardavviket for mengden bevegelse. Inkludering av andre tilgjengelige variabler ga en litt bedre tilpasning av dataene enn ved kun GMT-variablene. De inkluderte variablene her var variabelen for kjønn og en indikasjonsvariabel for ekstremt tidligfødte. I tillegg ble det justert for lengden av videoopptakene. Imidlertid viste statistisk inferens, i form av bootstrapping og multi sample-splitting, at bare gjennomsnittsverdien av massesentet av bevegelsene i y-retning hadde en statistisk signifikant assosiasjon med forekomsten av CP.

Preface

This thesis is written as a Masters degree in Industrial Mathematics at the Norwegian University of Science and Technology. Some previous analysis has been carried out in a project thesis in the autumn semester 2016, while the Master thesis has been carried out in the spring semester 2017, both at the Department of Mathematical Science.

I would like to thank both my supervisors, Associate Professor Turid Follestad at the Unit of Applied Clinical Research, NTNU, and Professor Mette Langaas at the Department of Mathematical Science, NTNU, for excellent guidance and advising. I would also like to thank my co-supervisor Researcher Lars Adde at the Department of Laboratory Medicine, Children's and Women's Health, NTNU, for great help in understanding the the medical background and the computer-based method, and for letting me work on such an exciting project. In addition, warm though go to my boyfriend and friends, which lighten up my everyday and make me think clear again when I am down.

Table of Contents

List of Tables

List of Figures

Abbreviations

FMs	=	Fidgety Movements
CP	=	Cerebral Palsy
GMT	=	General Movement Toolbox
GMA	=	General Movement Assessment
ROC-curve	=	Receiver Operating characteristic Curve
AUC	=	Area under the ROC-curve
INLA	=	Integrated Nested Laplace Approximation
C	=	Centroid of motion
A	=	Area of motion
Q	=	Quantity of motion
H	=	Hight of motion
W	=	Width of motion

Chapter 1

Introduction

Cerebral palsy (CP) is one of the most common causes of childhood physical disability, and occur in 2-2.5 per 1000 children (Yarnell and O'Reilly (2013), Oskoui et al. (2013)). An early detection can have a positive effect on the motor development of the child with CP (Blauw-Hospers and Hadders-Algra, 2005), and may reduce later daily life problems. A method for assessing the young nervous system, called the General Movement Assessment (GMA) has shown good results in predicting cerebral palsy at an early stage. Especially the absence of fidgety movements (FMs) have shown to predict CP with high sensitivity and specificity (Prechtl et al., 1997). As there are few trained clinicians to perform the GMA-analysis, computer-based methods can be applied. In this thesis, we consider a computer-based method for assessing the young nervous system, called the General Movement Toolbox (GMT). This program analyses video recordings of infants by their movements, based on changes of pixel-values between frames. The toolbox returns several summary measures, where each summary measure is one value per child (Adde et al., 2010).

In this thesis, we use GMT-summary variables from 798 video recordings of 693 infants, to predict normal FMs and CP. Since some of the infants have repeated measurements, we use a mixed effects logistic regression model with random intercepts to predict the FMs. For these data, we consider both a frequentist approach with the glmer()-function from the lme4-package in R (Bates et al., 2015), and a Bayesian approach, using the INLA-package in R (Rue et al., 2009). In addition, we perform a simulation study for the frequentist approach, to investigate whether small numbers of repeated measurements per infant could cause uncertain estimates. To predict CP, we remove the repeated measurements, and use a logistic regression model for the 693 infants. Then, we include several of the GMT-variables and other available variables in the model, and use the Lasso method for variable selection.

Since this thesis is a continuation of a project from autumn 2016, parts of the theory in the thesis are based or inspired by the work from the project. These are Section 2.1 and 2.2 in Chapter 2, and Section 3.1, 3.2 and 3.4.1 in Chapter 3.

We start this thesis by explaining details about the diagnosis of cerebral palsy and the

framework for the GMA- and GMT-methods in Chapter two. In this chapter, we also describe the data used for the analysis in the thesis, before we present the aim of the thesis. Then, in Chapter three, we go into details of the statistical methods used to model the data. Here, we first consider the methods for prediction of cerebral palsy. We start by presenting the well known logistic regression model, and methods for model evaluations. Then we consider the Lasso method for variable selection in the logistic regression model. In the final part of this chapter, we present the mixed effects logistic regression model with random intercepts, and look at both a frequentist and a Bayesian approach for estimating the model.

In Chapter four, we present the results from the model fitting, starting with the models for the FMs. Here, we also introduce the simulation study with results. Then, in the final part of Chapter four, we present the results from the CP-models. Here, we first present the results from the model which is similar to the ones for the FM-responses, before we look at the results from the variable selection using the Lasso method, with and without the other available variables. Next, in Chapter five, we conclude and discuss the results, and compare them to previous studies. Finally, we point out improvements for further work in Chapter 6.

Chapter 2

Background

Cerebral palsy (CP) describes a group of permanent disorders of the development of movements that occur in the developing fetal or infant brain (Rosenbaum et al., 2007). The damage can occur during pregnancy, delivery, the first month, or less commonly in early childhood (Yarnell and O'Reilly, 2013). The abnormal gross and fine motor functioning can lead to difficulties with walking, eating, coordinated eye movements, articulation of speech and other musculoskeletal functions (Rosenbaum et al., 2007). Being born preterm (born $<$37 weeks' of gestation) or with a very low birth weight (weighing $<$1500 g/$<$32 weeks' of gestation) or extreme low birth weight ($<$ 1000 g/$<$28 weeks' of gestation) is associated with significant motor impairment (de Kieviet et al., 2009), and as many as 5-15% of infants with a very low birth weight develop CP, (Veelken and Just (2013), Sellier et al. (2016), Platt et al. (2007), Oskoui et al. (2013)). Extreme preterm infants (born before 28 weeks' gestation) are born during a period of active brain development and maturation, placing them at an extremely high risk of brain injury (Stephens and Vohr, 2009).

We start this chapter by describing methods for diagnosing CP, for which we will focus on the computer-based method called *The General Movement Toolbox*. Then the dataset used for analysis in this thesis is described, before the aim of the thesis is introduced.

2.1 Diagnosing cerebral palsy

Before the age of 36 months, the motor capacity is not easily assessed, as it is not fully developed. The diagnose of CP before this age might therefore be difficult and misleading. Most of the false positive tests are done before the age of 18 months due to confusion with other neurodevelopmental disorders (Bosanquet et al., 2013).

Even though the CP diagnosis is permanent, an early detection can give earlier and closer follow up of the child, and can give relief to parents of children unlikely to develop CP. The brain‘s ability to adapt and change it‘s structure and functions is called the plasticity of the brain (McLellan et al., 2011). The plasticity of the brain is at its highest during the first two years, and decrease gradually thereafter (de Graaf-Peters and Hadders-Algra, 2006). It has been shown that intervention may be most efficient when the plasticity of the

brain is high (Heineman and Hadders-Algra (2008) with references), and an early detection of brain impairment is therefore crucial. An earlier follow up and training program can have a positive effect of the motor development of the child with CP (Blauw-Hospers and Hadders-Algra, 2005), in particular through prevention of limb contractions (Lindström and Bremberg, 1997), and might make a difference in the child's ability to handle everyday challenges. In addition, an early detection of CP gives the parents more time for adjustment and preparation.

"It has been shown that spontaneous motility is an excellent marker for neural dysfunction caused by brain impairment" (Einspieler and Prechtl, 2005), which normally would not become evident and clinically manifested for years (Darsaklis et al., 2011). A method called "General Movements Assessment" developed by Heinz F. R. Prechtl is a known diagnostic tool for the functional assessment of the young nervous system, and has shown good results in predicting CP at an early stage (Prechtl et al., 1997).

2.1.1 General Movements Assessment

General movements (GMs) are gross movements which involve the entire body. They are recognized by the variable sequence of arm, leg, neck and trunk movements which varies in speed, intensity and force with a gradual beginning and end. They include rotations along the axis of the limbs and slight changes in the direction of the movement. The fluent and elegant movements give the impression of complexity and variability (Prechtl, 1990). GMs have turned out to be an effective measure for the functional assessment of the young nervous system. They are complex, occur frequently, and last long enough to be observed properly (Einspieler and Prechtl, 2005).

Already at nine to twelve weeks postmenstrual age, the nervous system of a human fetus develops a large variety of movement patterns. Movements such as GMs, stretching, yawning and breathing are included, and they do not change their form after birth, independently of when the birth occurs. In this way, one can easily assess the functionality of the young nervous system through observations of GMs. The GMs are separated into different phases from early fetal life until the first half year post term, as shown in Figure 2.1, where each phase has it's own characteristics (Einspieler and Prechtl, 2005).

Fetal and Preterm GMs

From week nine up to term age, the GMs are referred to as Fetal and Preterm General Movements. These GMs may have large amplitudes and are often of fast speed (Hopkins and Prechtl, 1984).

Writhing movements

During term age and up to two months post term, the GMs are characterized by small-to-moderate amplitude and speed, typically in an elliptic form. This phase of GMs is referred to as writhing movements and may appear "awkward and ungrateful" (Hopkins and Prechtl, 1984).

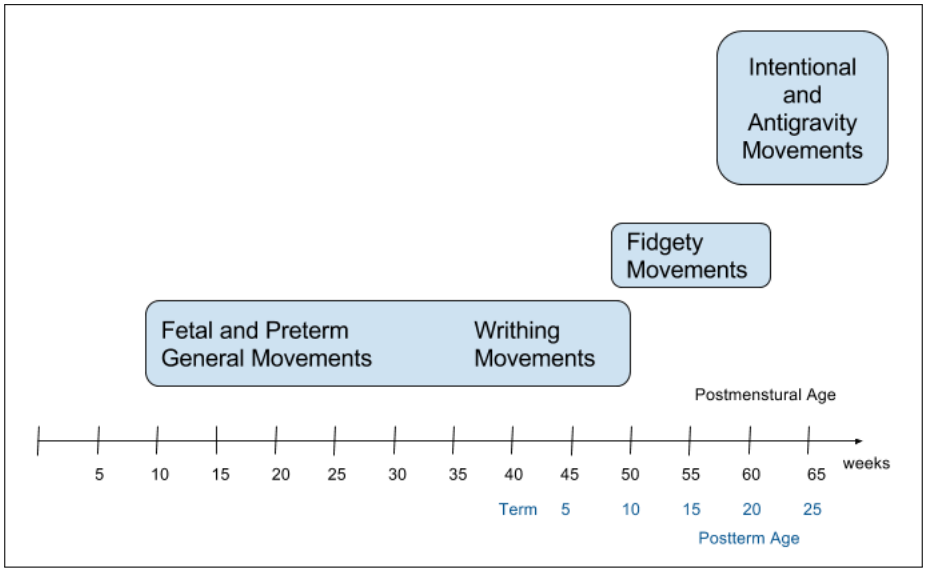


Figure 2.1: Developmental course of general movements, with inspiration from Einspieler and Prechtl (2005).

Fidgety movements

The next phase of GMs appears at six to nine weeks post term and disappears at 18 to 20 weeks post term age. These movements are referred to as Fidgety Movements (FMs) and differs from writhing movements by their rounded and elegant movements of the entire body. ”Fidgety movements are small movements of moderate speed and variable acceleration, of neck, trunk and limbs, in all directions, continual in the awake infant, except during fussing and crying” (Einspieler et al., 2004). They appear smooth, as arms and fingers move smoothly with full flex and extent, and their wrists rotates (Hopkins and Prechtl, 1984). At the end of the first half year, FMs gradually disappear and intentional and antigravity movements starts to dominate (Einspieler and Prechtl, 2005).

During the FMs period, one can classify the movements as normal or abnormal. There exists two methods for classification of abnormal and normal FMs with slightly different classification categories and terminology; the Prechtl’s approach and the Hadders-Algra approach (Adde, 2010). We will focus on Prechtl’s approach, since the GMA observers in the studies from which we have our data, are trained and certified in this approach.

The FMs are classified as normal if FMs are present. According to Prechtl’s approach, the presence of FMs are categorized in three groups. Continual FMs (++) when the FMs occur frequently with only short pauses, intermittent FMs (+) when the FMs occur often, but with longer pauses than for continual FMs, and sporadic FMs (+-) when there are some occurrence of FMs, but only sporadic.

The classification of abnormal FMs, includes both abnormal and absent FMs. Abnormal FMs (Exagg.) are defined as present FMs, but the movements are greatly or moder-

	Normal FMs	Abnormal FMs
Old	$++, +, +-$	$-$, Exagg
New	$++, +$	$+-, -$, Exagg

Table 2.1: New and old Prechtl's approach for classification of normal and abnormal FMs, where the FMs are categorized as continual, $(++)$, intermittent, $(+)$, sporadic, $(+-)$, absent, $(-)$, and exaggerated (Exagg).

ately exaggerated with respect to amplitude and speed. If FMs are not observed at all in the period 9 to 20 weeks post-term, they are classified as absent (-).

We call the above method Prechtl's "old" approach. The "new" Prechtl's approach separates from the old by including sporadic FMs in the category of abnormal FMs. The old and new Prechlt's approach are shown in Table 2.1 for different categories of FMs.

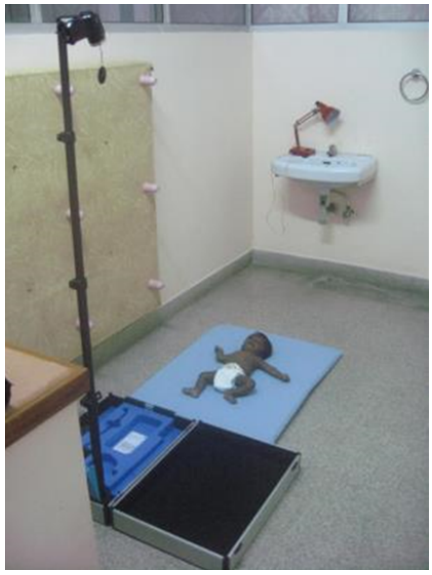
It has been shown that abnormal and absent FMs increase the risk of development of neurological impairment. Particularly, the absence of FMs has been shown to be highly predictive of CP, while normal FMs are associated with normal neurological outcome (Einspieler and Prechtl, 2005).

A previous study (Spittle et al., 2009) showed that both GMA and magnetic resonance imaging (MRI) had a sensitivity of 100 % when predicting the development of CP by the age of 12 months in preterm infants. A MRI is expensive, require highly skilled personnel and is not available for everyone. GMA however, is non-expensive, requires only a video camera and a trained person to analyze the video. A limitation of GMA is that there are few trained clinicians to analyze the video. Unexperienced clinicians, as well as experienced clinicians, working alone have a risk of drifting away from the GMA standards over time (Adde et al., 2009). Because of this, computer-based methods can be very useful tools for the clinician, and can perhaps be used without trained personnel to give earlier identification of infants unlikely or likely to develop CP. It only requires someone to perform the video recording within the requirements, and then computer-based methods can analyze the video for the child's movements. In this way, the clinicians get an objective second opinion, which will hopefully contribute to a higher accuracy for identifying infants with or without CP.

2.1.2 Computer-based video analysis

There exist several computer-based methods for assessing movements from infants at high risk of neurological and motor impairments (Marcroft et al. (2015) and references therein). The computer-based methods can be separated into two categories; i) using motion capturing systems; and ii) traditional color cameras. Using the motion capturing system, the limb movements can be tracked indirectly by the 3D system or directly through body-worn sensors. The body-worn sensors seem to be a promising application for prediction of abnormal movement patterns, but it has not yet been applied to a sample size large enough to do sensitivity and specificity analysis.

The 3D motion capturing system has shown good results in separating healthy infants from high risk infants. However, the method is both computationally and cost expensive, and is more adaptable to a research environment than a clinical environment. For tradi-



(a)

(b)

Figure 2.2: Setup for the video recording (b) and a snapshot (a) of cropped a video recording. (Lars Adde, St.Olavs Hospital/NTNU, Trondheim, with approval)

tional color cameras analysis, the price is considerably smaller than for a motion capturing system, and the set up is easier. An easier set up allows for usage of the method outside the clinical or research environment, and the method can be applied in a more natural environment for the infant, for example at home. In this project, we will use data from a video-based method, called the General Movement Toolbox.

General Movement Toolbox

This section is mostly based on the article by Adde et al. (2009). The General Movement Toolbox (GMT) is a software solution using video recordings of young infants to study their general movements. It has been customized from the open source software "The Musical Gesture Toolbox" (MGT), developed by Jensenius et al. in 2004 for studying music-related movements.

The infants are placed on a standard mattress and video recorded with a stationary digital video camera placed above them for typically 3-10 minutes. The GMT-software processes the video file and analyses the movements of the infant. The infants, wearing a diaper and a body, must be awake, active and in a comfortable state (non-crying and no pacifier) and lying on their back, for the analysis to work properly. The videos recordings are trimmed to typically 3-5 minutes length to ensure the correct state of the infant. All videos are also cropped so that only the image only consist of the infant on the mattress. Figure 2.2a shows the setup for video recording, while Figure 2.2b shows a snapshot of the cropped video recording.

Each second of the video recording typically consists of 25 images. One cropped image typically consists of $M = 320$ times $N = 240$ pixels, and each pixel has a value between 0 and 255 (8 bits) that represents the intensity (Adde, 2010). A motion image is calculated from the change in pixel values between two following frames, as shown in Figure 2.3. In the motion image the pixels have values 0 or 1, where 0 is black and represents no movements between the images, and 1 is white and represents movement. Hence, the white pixels are the active pixels. In this way, all the movements in the video are calculated from the motion image. The GMT uses this to calculate plots called motiongrams and several summary variables that summarizes the movements into one value for each child. Analyzing the motiongrams and the summary variables, one can see the amount, variation and location of the movements. The motiongrams and three of the summary variables are described below.

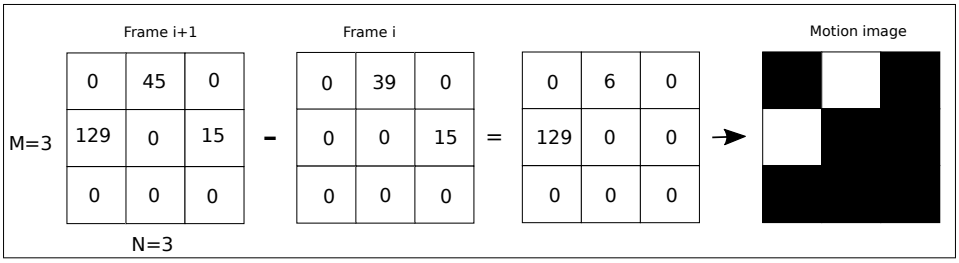
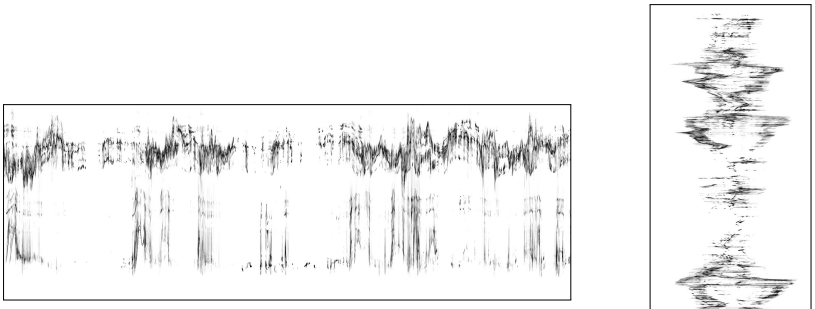


Figure 2.3: Visualization of calculation of the motion image. Each square represents a pixel in the frame that consists of 3x3 pixels. No change between frames is represented as black in the motion image, while change is displayed as white.

Motiongram

A motiongram is a plot representing the motion image over time. Using an x-y coordinate system on the motion image, one can take an average pixel value in both x and y-directions. A horizontal motiongram calculates the averages of pixel values in the x-direction, such that one gets a matrix with dimension M$\times$1 where M is the number of pixels in the y-direction and the entries in the matrix are the corresponding average of the pixel values in the x-direction. Plotting these matrices against time, movements of the upper body can be seen in the upper part of the y-axis and movements from the lower part of the body can be seen in the bottom. A vertical motiongram uses the average pixel value in the y-direction and shows the movements on the left and right side of the body.

Examples of horizontal and vertical motiongrams are shown in Figure 2.4. Using both motiongrams, one gets an indication of both the amount and the location of the movements over time. From the horizontal motiongram in Figure 2.4a one can see that there are more movements in the upper part of the body than in the lower part. The vertical motiongram in Figure 2.4b shows that the infants' movements on the left and right side seem to be mostly symmetric.



(a)

(b)

Figure 2.4: Examples of (a) a horizontal motiongram where time is running along the x-axis and vertical movements along the y-axis, and (b) a vertical motiongram where time is running along the y-axis and horizontal movements along the x-axis. (Lars Adde, St.Olavs Hospital/NTNU, Trondheim, with approval).

Quantity of motion, Q

The quantity of motion, Q, is defined as the sum of all active pixels in the motion image divided by the total number of pixels in the image ($n = M \times N$). Hence,

$$Q = \frac{\sum_i^n p_i}{n}, \quad \text{where } p_i = \begin{cases} 1 & \text{if pixel } i \text{ is white} \\ 0 & \text{if pixel } i \text{ is black.} \end{cases}$$

Plotting this variable against time gives an indication of amount of movements over time. To get one measure per child, the mean value Q_{mean}, the maximum value Q_{max} and the standard deviation Q_{sd} are calculated and used as summary variables.

Centroid of motion, C

The centroid of motion, C, measures the centre of all movements in the motion image for each frame. It is the spatial centre for the active pixels in the motion image and can be thought of as the centre point for the movements of the infant. This variable is given in the x- and y-direction (C_x, C_y), so we have the Euclidian distance $C = \sqrt{C_x^2 + C_y^2}$. To get just one summary value from this, one could calculate at the mean from all motion images in the x- and y-direction (C_{xmean}, C_{ymean}) and the Euclidian distance between them, $C_{mean} = \sqrt{C_{xmean}^2 + C_{ymean}^2}$. The standard deviations, C_{sd}, C_{xsd}, C_{ysd}, can also be calculated, where C_{xsd}, C_{ysd} are the standard deviation for their corresponding vector C_x and C_y, and C_{sd} is the Euclidean metric between them, $C_{sd} = \sqrt{C_{xsd}^2 + C_{ysd}^2}$.

Variable	Description
Q	**Quantity of motion**
Q_{mean}	mean
Q_{sd}	standard deviation
C	**Centroid of motion**
C_{xmean}	mean in x-direction
C_{ymean}	mean in y-direction
C_{xsd}	standard deviation in x-direction
C_{ysd}	standard deviation in y-direction
C_{sd}	standard deviation
A	**Area of motion**
A_{mean}	mean
A_{sd}	standard deviation
H_{mean}	mean height of motion
W_{mean}	mean width of motion
H_{sd}	standard deviation of height of motion
W_{sd}	standard deviation of width of motion

Table 2.2: Important summary variables given by the GMT-toolbox.

Area of motion (A)

The area of motion A is a measure of the area of which the infant is moving. The height H is the difference between the largest and smallest y-value for the active pixels in the motion image, and the width W is the corresponding value in x-direction. The area, A is then the height times the width. This measure is calculated for each motion image, and A_{mean} is the mean value of the area from all the motion images, while A_{sd} is the standard deviation. Mean and standard deviation for the height and width are also calculated and are denoted as H_{mean}, W_{mean}, H_{sd} and W_{sd}.

Table 2.2 gives an overview of the mentioned GMT-variables which will be used in the next chapters.

2.1.3 Previous studies

It has been shown in previous studies (Adde et al., 2009) that among all the obtained variables from the GMT-analysis, it is the variability of the centre of motion, C_{sd}, that is the most precise at predicting FMs, with both high sensitivity and high specificity. A low C_{sd} imply a stable spatial centre in the motion image and seem to correlate with both the presence of FMs and non-development of CP (Adde et al., 2010). The interpretation of this result can be explained in the following way: A stable centroid of motion may reflect the ongoing stream of small movements in the whole body as a system, described in the GMA methodology as FMs, and a stable centroid of motion gives a low value for C_{sd}.

The previous studies using the GMT-analysis have been performed on small samples from St.Olavs University Hospital in Norway. This goes for all computer-based methods presented in the article by Marcroft et al. (2015). They have been tested on small sample

sizes with mostly too few infants with neurologic brain impairment, to be reliable. Even though many of the methods indicate promising results, they have yet to be tested on larger datasets.

2.2 Data

In this project, we use data based on projects at St.Olavs University Hospital Trondheim/NTNU, Norway. The data, consisting of infants at a high risk of developing CP, have been collected from three different countries at different times. Parts of the data have been analyzed in previous studies (Adde et al. (2009), Adde et al. (2010), Adde et al. (2016)), but only when considering infants from the same country. The results from these studies conclude that low values of the C_{sd} variable from the GMT-software are associated with normal FMs and no CP, and that normal FMs are associated with a normal motor development.

2.2.1 Design

Seven hospitals from three different countries; Norway, USA and India, received a suitcase including a standard mattress, a video camera and standard equipment to place the video camera in the correct height above the infant. Infants with high risk of developing CP (described in the article by Adde et al. (2007)) were video filmed at ten to 18 weeks post term age, when FMs should be present. All the infants had one recording taken, some had two or more, and the recordings were sent to Trondheim, Norway for further analysis. In Trondheim, the GMT-software was used on the video recordings to analyze the movements of the infants. In addition, three physiotherapists worked in pairs to do a GMA analysis on all the recordings. The GMT analysis were done in the period 2012-2013, and in February 2017 all participating infants had CP status registered between 18 months to four years of age.

A total of 879 video recordings of 754 infants from seven hospitals have been taken in the period from 2009 to 2013. In Norway there were 155 participants from four different hospitals: University hospital of North Norway (Tromsø) ($n = 12$), St.Olavs University Hospital (Trondheim) ($n = 90$), University Hospital in Oslo (Rikshospitalet) ($n = 45$) and Levanger Hospital (Levanger) ($n = 8$). From USA there were 276 participants from two hospitals in Chicago: Lurie Children's Hospital of Chicago (LCH) ($n = 150$) and Chicago University Hospital (UOC) ($n = 122$). In India, there were 327 participants from Christian Medical College, Vellore, Tamil Nadu.

2.2.2 Participants

From the total of 754 infants two moved out of their country, two died at age six and eight months, two dropped out of the studies and one was unavailable for follow up. In addition, the GMT analysis was not performed on all video recordings, in fact 73 recordings are registered without GMT analysis. When removing those without the GMT analysis, the remaining number of infants is 693 with a total of 798 video recordings. Figure 2.5 illustrates the remaining number of participants from the different countries. The infants

were video recorded at mean twelve weeks post term age (sd=1.54) and the mean length of the recordings are four minutes and 19 seconds (sd=one minute, four seconds). Most infants have one video recording taken, some infants from Norway and USA have two or more.

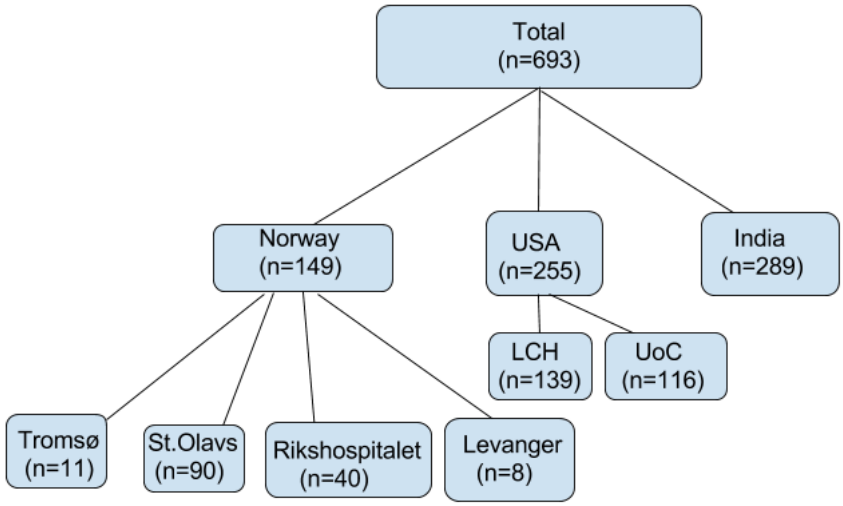


Figure 2.5: Number of participating infants from different hospitals.

	Norway	USA	India	Total
N *n(%)*	149 (21)	255 (37)	289 (42)	693
Gender				
Male n(%)	91 (61)	132 (52)	147 (51)	370 (53)
Female n(%)	58 (39)	123 (48)	142 (49)	323 (47)
Birth weight *mean(sd)*	2026 (1353)	1724 (1154)	1277 (186)	1603 (992)
Gestational age *mean(sd)*	33 (6.38)	31 (5.96)	32 (2.34)	32 (4.95)
Neurologic outcome				
CP n(%)	25 (16.78)	18 (7.06)	3 (1.04)	46 (6.64)
Non-CP n(%)	124 (83.22)	237 (92.94)	286 (98.96)	647 (93.36)

Table 2.3: Background variables and neurological outcome for the participants in each country. Percentage for gender and neurological outcome are given within the countries.

Table 2.3 shows some of the background information for the remaining participants from each country. The number of video recordings done for each infant in each hospital are shown in Table 2.4. Both tables have excluded the participants with missing CP diagnosis and GMT analysis. Table 2.3 shows that among the 693 participating infants, 46 (6.6%) have developed CP. In Norway, there are 25 out of the 149 (16.8%) participants

with CP, in USA, there are 18 out of 256 (7%) participants with CP and in India there are only three out of 289 (1%) participants with CP.

	Number of video recordings					
	1	2	3	4	Total	Number of infants
Tromso	11	0	0	0	11	11
St.Olavs	28	62	0	0	152	90
Rikshospitalet	15	25	0	0	65	40
Levanger	1	2	3	2	22	8
Norge	**55**	**89**	**3**	**2**	**250**	**149**
LCH	139	0	0	0	139	139
UOC	112	4	0	0	120	116
USA	**251**	**4**	**0**	**0**	**259**	**255**
India	**289**	**0**	**0**	**0**	**289**	**289**
Total	**595**	**93**	**3**	**2**	**798**	**693**

Table 2.4: Number of video recordings taken per infant, total number of recordings and number of infants in each city and summed up in each country.

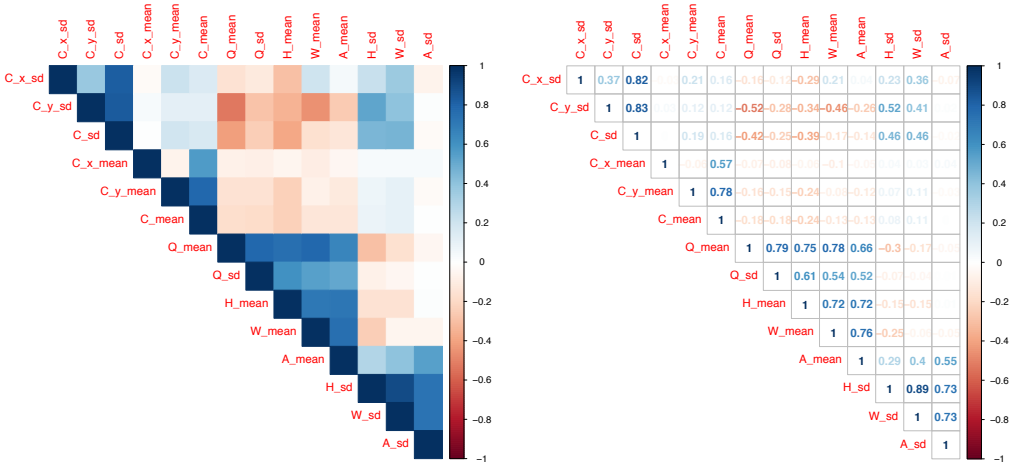
We see from Table 2.4 that only in Levanger there have been more than two video recordings per infant. In Levanger, the infants were recorded at two dates, and at each date, two video recordings should have been performed. The recordings are separated by lifting the infant and laying it down only seconds after. Hence, there should have been four recordings for all participants in Levanger, but we see that some are not in our data for different reasons that we will not look into. Summarizing all recordings done in Norway, we see that most infants have two recordings, but many have only one. In USA there are only four infants with two recordings while 251 infants have only one recording. In India all infants have been recorded only once. In total, we see that 595 infants have one recording, 93 infants have two, three infants have three and only two infants have four recordings.

2.2.3 Fidgety movements and General Movements Toolbox-variables

The pairwise correlations of the GMT-summary variables are visualized in Figure 2.6. The figure shows that some of the variables are highly correlated. In addition to the expected correlations between C_{xmean}, C_{ymean} and their Euclidean distance C_{mean}, and between C_{xsd} and C_{ysd} and their Euclidean distance C_{sd}, also the the height and width variables, H, W, are correlated with each other and the area variables, A. In addition, Q_{mean} is correlated with Q_{sd} and the mean variables for the height, the width and the area.

The data for FMs and the GMT-summary variables are shown in Table 2.5 for both cases of the neurological outcome, separated by countries. The table shows that only one of the infants with normal FMs developed CP, and only a few of those with intermittent or sporadic FMs developed CP. Among the 89 recordings of infants with absent FMs, 43 of the recordings were of infants that developed CP, which corresponds to 48.3%.

When looking at the GMT-summary variables it is important to notice that in India there are only three recordings of infants with CP. Looking at the values in Table 2.5 and



(a) **(b)**

Figure 2.6: Pairwise correlation plot of the GMT-summary variables visualized by a) colors and b) numbers.

	Norway		USA		India		Total	
	CP (n= 42)	No CP (n= 208)	CP (n= 19)	No CP (n= 240)	CP (n= 3)	No CP (n= 286)	CP (n= 64)	No CP (n= 734)
FMs								
Exagg *n(%)*	0 (0)	4 (100)	0 (0)	5 (100)	0 (0)	5 (100)	0 (0)	14 (100)
- *n(%)*	28 (68.3)	13 (31.7)	14 (37.8)	23 (62.2)	1 (9.10)	10 (90.9)	43 (48.3)	46 (51.7)
-+ *n(%)*	2 (10.5)	17 (89.5)	1 (4.55)	21 (95.5)	1 (3.57)	27 (96.4)	4 (5.80)	65 (94.2)
+ *n(%)*	11 (6.96)	147 (93.0)	4 (2.50)	158 (97.5)	1 (0.541)	184 (99.5)	16 (3.17)	489 (96.8)
++ *n(%)*	1 (3.57)	27 (96.4)	0 (0)	33 (100)	0 (0)	60 (100)	1 (0.826)	120 (99.2)
Quantity of motion								
Q_{mean} ($\times 10^3$) *mean(sd)*	7.72 (5.38)	8.13 (4.58)	9.00 (5.32)	10.1 (5.13)	2.42 (1.58)	8.03 (4.58)	7.85 (5.37)	8.72 (4.86)
Q_{sd} ($\times 10^3$) *mean(sd)*	13.4 (6.75)	10.7 (6.02)	10.5 (4.07)	11.3 (3.98)	4.14 (1.17)	9.18 (3.19)	12.1 (6.27)	10.3 (4.50)
Area of motion								
A_{mean} ($\times 10$) *mean(sd)*	1.82 (0.738)	2.00 (1.25)	2.02 (0.965)	2.27 (0.780)	0.862 (0.432)	1.83 (0.792)	1.84 (0.826)	2.02 (0.958)
A_{sd} ($\times 10$) *mean(sd)*	1.33 (0.387)	8.33 (7.25)	1.28 (0.291)	1.50 (0.376)	1.00 (0.095)	1.33 (0.30)	1.30 (0.356)	3.37 (38.6)
H_{mean} ($\times 10$) *mean(sd)*	4.27 (1.01)	4.46 (0.938)	4.08 (1.31)	4.29 (0.927)	2.34 (1.21)	3.72 (1.04)	4.12 (1.17)	4.12 (1.03)
H_{sd} ($\times 10$) *mean(sd)*	2.39 (0.416)	2.09 (0.901)	2.02 (0.395)	2.09 (0.392)	2.15 (0.042)	2.08 (0.337)	2.27 (0.432)	2.09 (0.569)
W_{mean} ($\times 10$) *mean(sd)*	3.38 (1.06)	3.62 (1.05)	3.88 (1.40)	4.40 (1.07)	2.04 (1.10)	3.79 (1.18)	3.46 (1.22)	3.94 (1.16)
W_{sd} ($\times 10$) *mean(sd)*	1.93 (0.553)	1.77 (1.02)	2.02 (0.395)	2.09 (0.392)	2.15 (0.042)	2.08 (0.337)	1.97 (0.497)	2.00 (0.640)
Centroid of motion								
C_{xmean} ($\times 10$) *mean(sd)*	4.75 (0.603)	5.05 (0.533)	4.81 (0.554)	4.82 (0.610)	5.55 (0.476)	5.11 (0.632)	4.80 (0.600)	5.00 (0.611)
C_{ymean} ($\times 10$) *mean(sd)*	5.95 (0.736)	5.48 (0.682)	5.98 (0.577)	5.75 (0.657)	6.26 (0.050)	5.61 (0.751)	5.97 (0.673)	5.62 (0.709)
C_{mean} ($\times 10$) *mean(sd)*	7.63 (0.720)	7.48 (0.575)	7.69 (0.566)	7.53 (0.636)	8.37 (0.313)	7.62 (0.691)	7.69 (0.676)	7.55 (0.644)
C_{xsd} ($\times 10$) *mean(sd)*	1.04 (0.265)	0.919 (0.260)	1.13 (0.189)	1.18 (0.329)	1.12 (0.161)	1.24 (0.303)	1.07 (0.241)	1.13 (0.329)
C_{ysd} ($\times 10$) *mean(sd)*	1.43 (0.329)	1.25 (0.265)	1.36 (0.320)	1.25 (0.274)	1.37 (0.116)	1.31 (0.294)	1.41 (0.32)	1.27 (0.280)
C_{sd} ($\times 10$) *mean(sd)*	1.79 (0.349)	1.57 (0.298)	1.78 (0.338)	1.73 (0.374)	1.78 (0.165)	1.81 (0.356)	1.78 (0.336)	1.72 (0.360)

Table 2.5: The number of cases and percentage within the countries for FMs, and mean and standard deviations of the GMT-variables of the 798 video recordings. FMs are categorized in absent ($-$), sporadic ($-+$), intermittent ($+$), continual ($++$) and exaggerated (Exagg).

the histograms for the GMT-variables in Figure 2.7 and Figure 2.8, it is not easy determine which of the variables that stands out, as most of the histograms seem to have the same shape for both CP and no CP, and the mean values in Table 2.5 seems quite similar between groups. The mean values for the variables Q_{mean}, A_{mean}, A_{sd}, H_{mean} and C_{xmean} in the table are lower for the CP group than for the no CP group. However, only the histograms

for Q_{mean}, A_{mean} and W_{mean} shows some small differences between the groups, while the histograms for the other mentioned variables are very similar between groups.

For the variables H_{sd}, C_{ymean}, C_{mean}, C_{ysd} and C_{sd}, the mean values in the table are higher for the CP group than for the no CP group. The histograms for all these except H_{sd} shows that there are more recordings with higher values of these variables in the CP group than in the no CP group.

Looking at the histograms for the area variables, we see that there are some outliers in the no CP group. These are most easily seen in the histograms for A_{mean}, A_{sd}, H_{sd} and W_{sd}. These outliers correspond to two video recordings of different infants. The outliers are most extreme for the A_{sd} variables, where they take the values 69.7 and 78.4. The outlier values have been noticed and checked in previous studies, but no clear answer to why these values differ that much from the other values has been found.

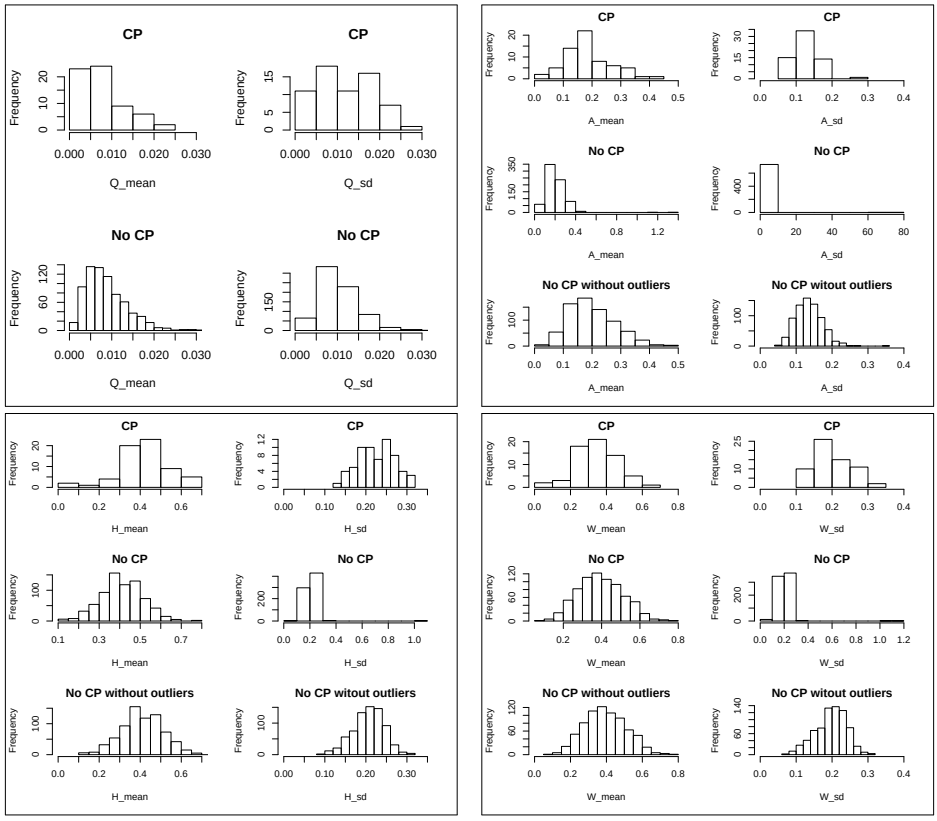


Figure 2.7: Histograms of the mean and standard deviation values for the quantity of motion (Q), area of motion (A), height of motion (H) and width of motion (W). The two outliers for the area-variables has been removed in the third row in the corresponding histograms.

As all the GMT-variables are calculated from the number and locations of the pixel changes in the motion image, one would expect that the infants‘ area would have an effect on these variables. The trunk area of the infants in the recordings are also registered

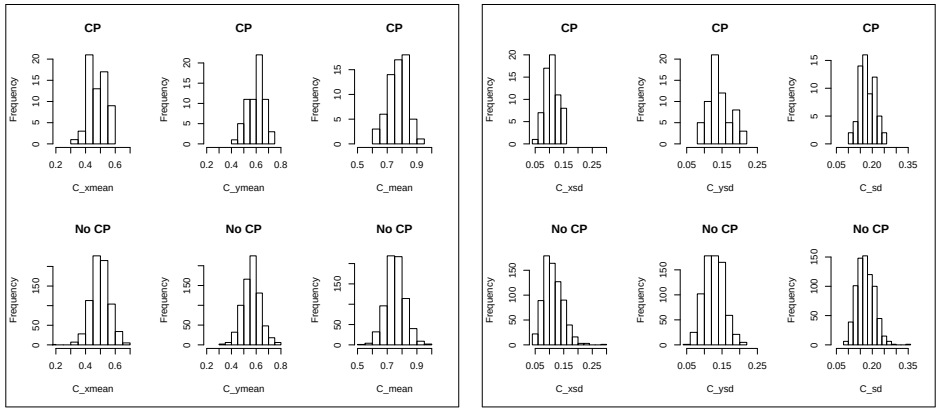


Figure 2.8: Histograms of the mean and standard deviation values for the centroid of motion variables.

through the GMT-software, with mean value 400 cm^2 (sd= 71.6 cm^2). Looking at Figure 2.9, there seems to be some small dependency between some of the GMT-variables, mostly the area variables, and the trunk area. However, in general there are not much dependency between the variables and the trunk area, so we consider the GMT-variabels without normalizing for trunk area, throughout the thesis.

2.3 Aim of the thesis

In this thesis, the main goal is to develop statistical methods for prediction of CP for high risk infants, using one or several GMT-variables. Since FMs has been used as a surrogate measure for CP, we start by predicting normal FMs using a mixed effects logistic regression model and the data described above with 798 observations. We refer to these data as the *FM-data*. Since the FM-statuses are classified based on human judgement, it might include wrong classifications. In this thesis, however, we assume that the FM-statuses given in the dataset are the correct ones. A previous study showed that low variations in the centroid of motion were associated with normal FMs (Adde et al., 2009), on a dataset consisting of 82 Norwegian infants with a total of 132 recordings. We will investigate if a dataset consisting of infants from three different countries, Norway, USA and India, give the same association between C_{sd} and FMs. In addition, we are interested in differences between the countries, and will see if the effect of C_{sd} on the occurrence of normal FMs varies between countries.

In the project in the autumn 2016, we found that the frequentist method using the lme4-package in R gave strange results for these data. We present these results with the frequentist approach, but we also consider a Bayesian approach for modeling these data, to see if this method give more reasonable results. In addition, we will perform a simulation study to investigate if increasing the number of repeated measurements give more stable results with less uncertainties for the frequentist approach.

Then, we turn to the prediction of CP. Using CP as outcome, we remove repeated

recordings such that each infant only have one recording, and we can model the data using logistic regression. We refer to these data as the *CP-data*. A previous studies by Adde et al. (2010) showed that low values of C_{sd} are associated with not having CP, but this was found on a dataset consisting of only 30 Norwegian infants. Here, we look for the same association using data of infants from the three different countries, and we will investigate if there is a different effect of C_{sd} on the occurrence of CP for the different countries.

Finally, we consider several GMT-variables and other available variables to see if we can find a model that is better at predicting CP. In this part of the analysis, we use the Lasso method to investigate which of the variables that are associated with having CP and give the best model for prediction of CP. For statistically inference of the Lasso estimates, we consider both bootstrapping and the multi sample-splitting method for calculations of p-values and confidence intervals for the variables.

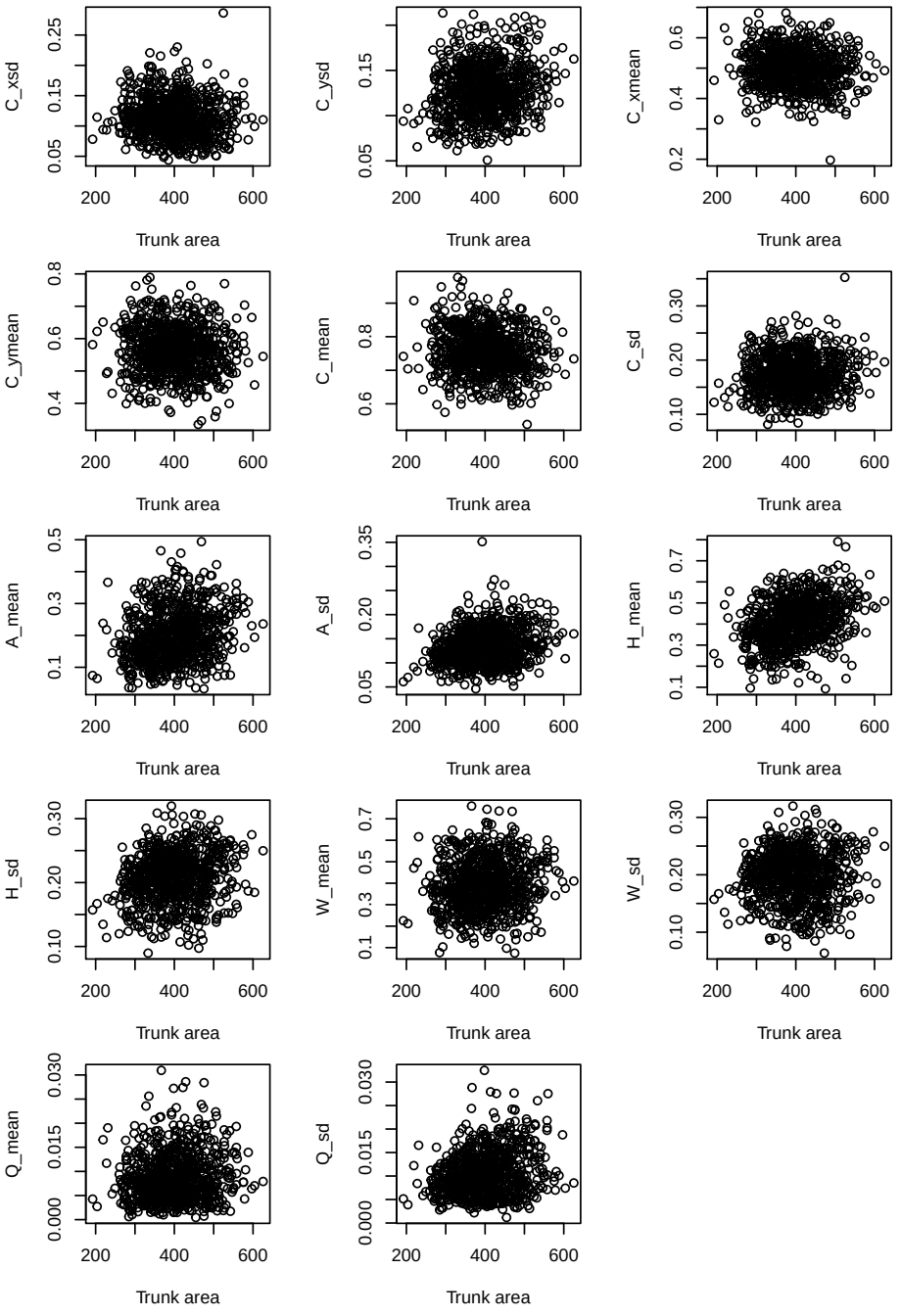


Figure 2.9: Scatter plots of the GMT-variables against trunk area of the infants. For the area of motion variables (A, H, W), the outliers have been removed.

Chapter 3

Statistical methods

In this chapter we present statistical methods for modeling and model evaluation of binary data. We start by presenting some well known theory for logistic regression, by presenting the logit model and estimation, and the likelihood ratio test for the regression coefficients. Then we look into methods for model evaluation and diagnostic tests. These methods will be applied also for the models presented later in the chapter. Then, we look into methods for variable selection, with focus on the Lasso method for logistic regression models.

In the final section of the chapter, we turn to mixed effects logistic regression models. We start the section by looking into the method for estimating the regression coefficients using a frequentist approach. Then, we introduce Bayesian theory and look into the Bayesian approach for estimation of the regression coefficients, using the the Integrated Nested Laplace Approximation (INLA).

3.1 Logistic regression

This section is mostly based on the Lecture Notes on Generalized Linear Models (Rodríguez, 2007). In a generalized linear model with binary outcome, the response takes only two values, $y_i = \{0, 1\}$. This type of model is also called a logistic regression model. Assuming independent variables, the model takes the form

$$y_i \sim \text{Bernoulli}(\pi_i) \quad \text{for } i = 1, 2, ..., n,$$

where all observations $y_1, ..., y_n$ are independent and each take value 1 with probability π_i. The Bernoulli distribution has a density on the form

$$f(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \quad \text{for} \quad y_i = 0, 1,$$

with expectation $E(y_i) = \pi_i$ and variance $Var(y_i) = \pi_i(1-\pi_i)$. The parameter, π_i, is the probability of y_i being 1 and $(1-\pi_i)$ is the probability of y_i being 0. Since the probability π_i should depend on the covariate vector $\boldsymbol{x_i}$ and be within the interval $[0, 1]$, the logit link

function is used on π_i to make it a linear function of the covariates. The link function is a one-to-one continuous differentiable function and the logit link is given by

$$logit(\pi_i) = \beta_0 + \mathbf{x_i}^T\boldsymbol{\beta},$$

where $\mathbf{x_i}$ is a vector of covariates, β_0 is the intercept and $\boldsymbol{\beta}$ is a vector of regression coefficients. The quantity $\eta_i = \beta_0 + \mathbf{x_i}^T\boldsymbol{\beta}$ is often referred to as the linear predictor. The logit function is given by

$$logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$$

and hence,

$$\frac{\pi_i}{1-\pi_i} = exp(\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta}). \tag{3.1}$$

Solving for π_i gives the form

$$\pi_i = \frac{exp(\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta})}{1 + exp(\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta})}. \tag{3.2}$$

Equation (3.1) is the probability of the observation y_i being 1, divided by the probability of y_i being 0, and is called the odds for observation y_i. In a logistic regression model, the regression coefficient β_j represents the change in the logit of π_i for each unit change for the covariate x_{ij}. By changing the j'th covariate by one unit, the odds will be multiplied by $\exp(\beta_j)$. In this way, this is a multiplicative model, and the coefficient $\exp(\beta_j)$ represents the odds ratio for the j'th covariate. The interpretation of the odds ratio is that the odds for $y_i = 1$ is $exp(\beta_j)$ times larger for an observation $x_{ij} + 1$ than the odds for an observation x_{ij}.

3.1.1 Estimation

For simplicity, we denote the regression coefficients $\{\beta_0, \boldsymbol{\beta}\}$ for $\boldsymbol{\beta}$. To estimate the regression coefficients, maximum likelihood estimation is used. The likelihood, L, for n independent Bernoulli observations is the product of the density functions. Hence,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

where π_i is given in Equation (3.2). The log likelihood is then,

$$\begin{aligned}\log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \Big(y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\Big) \\ &= \sum_{i=1}^{n} \Big(y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i})\Big).\end{aligned} \tag{3.3}$$

The regression coefficients are found by maximizing the likelihood for $\boldsymbol{\beta}$. There are no analytic solution when solving for $\boldsymbol{\beta}$, and the expression must be solved numerically. In the glm-function in R, the regression coefficients are found by the iterative method called Iterative Re-weighted Least Squares (IRLS) (R Core Team, 2014). This method works in the following way.

1. Given the current estimate $\hat{\boldsymbol{\beta}}$, the linear predictor $\hat{\eta}_i = \mathbf{x_i^T}\hat{\boldsymbol{\beta}}$ and the fitted values $\hat{\mu}_i = \text{logit}^{-1}(\hat{\eta}_i)$ are calculated.

2. These values are used to calculate the dependent variable $\mathbf{z}$ with elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(1-\hat{\mu}_i)},$$

   and the diagonal matrix $\mathbf{W}$ with weights

$$w_{ii} = \hat{\mu}_i(1-\hat{\mu}_i).$$

3. Then, the weighted least square estimates are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T W X})^{-1}\mathbf{X^T W z}$$

   where $\mathbf{X}$ is the matrix of covariates with the first row consisting of ones.

4. Go back to 1.

This algorithm runs until the difference between the current and previous estimate, $\hat{\boldsymbol{\beta}}_{new} - \hat{\boldsymbol{\beta}}_{old}$, is small enough, and gives the maximum likelihood estimator.

The estimated covariance matrix for the regression coefficients is given by

$$cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X^T W X})^{-\mathbf{1}},$$

where $\mathbf{W}$ is the diagonal matrix with weights evaluated in the last iteration. The diagonals of the covariance matrix are the variances of the coefficients, $\hat{\sigma}^2_{\hat{\beta}_j}, j = 0, 1, 2, ...$

For large sample sizes, n, the law of large numbers says that $\hat{\beta}_j$ is approximately normally distributed with mean β_j and variance $\hat{\sigma}^2_{\hat{\beta}_j}$, $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \hat{\sigma}^2_{\hat{\beta}_j})$. Hence, to test the significance of the estimated coefficients, a Z-test can be used. A Z-test tests the hypothesis

$$\text{H}_0 : \beta_j = 0 \quad \text{vs} \quad \text{H}_1 : \beta_j \neq 0,$$

where β_j is the coefficient for covariate j. Assuming H_0 is true, the Z-statistic is given by

$$Z = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}},$$

and is standard normally distributed with mean 0 and variance 1. A p-value for Z less than the chosen significance level, often 0.05, leads to rejection of the hypothesis and implies that the coefficient β_j is significant and that covariate x_j has a significant effect on the response.

3.1.2 Confidence intervals

The estimated probabilities, $\hat{\boldsymbol{\pi}}$, are found by replacing the coefficients in Equation (3.2) with the estimated coefficients. In order to estimate the confidence intervals of the probabilities, the standard errors of the linear predictor, $\hat{\boldsymbol{\eta}} = logit(\hat{\boldsymbol{\pi}})$ must be calculated. As the estimated linear predictor is a linear combination of the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ that are normally distribution, also the estimated linear predictor is normally distributed. The confidence interval of the estimated probabilities is then given by

$$\frac{exp(\hat{\boldsymbol{\eta}} \pm z_{\alpha/2}\widehat{SE}(\hat{\boldsymbol{\eta}}))}{1 + exp(\hat{\boldsymbol{\eta}} \pm z_{\alpha/2}\widehat{SE}(\hat{\boldsymbol{\eta}}))}$$

where

$$\widehat{SE}(\hat{\boldsymbol{\eta}}) = \sqrt{Var(\mathbf{X}\hat{\boldsymbol{\beta}})} = \sqrt{\mathbf{X}^T Var(\hat{\boldsymbol{\beta}})\mathbf{X}}$$

which can be written on the form

$$\widehat{SE}(\hat{\eta}_i) = \sqrt{Var(\hat{\beta}_0) + \sum_{k=1}^{N}(x_{ik}^2 Var(\hat{\beta}_k) + 2x_{ik}Cov(\hat{\beta}_0, \hat{\beta}_k)) + 2\sum_{k\neq l}^{N} x_{ik}x_{il}Cov(\hat{\beta}_k, \hat{\beta}_l)}.$$

The confidence interval of the odds ratio is found through the confidence interval of the covariate. For large sample sizes, a 95% confidence interval for β_j is given by $\hat{\beta}_j \pm 1.96 \cdot \hat{\sigma}_{\hat{\beta}_j}$. Then, the confidence interval for the odds ratio for β_j is given by $(exp(\hat{\beta}_j \pm 1.96 \cdot \hat{\sigma}_{\hat{\beta}_j}))$.

3.1.3 Likelihood ratio test

To investigate if one or more of the covariates are significant effects in the model, a likelihood ratio test (LRT) can be used. A LRT compare two nested models, based on the difference between their deviances, and test the hypothesis

$$\mathrm{H}_0 : \boldsymbol{\beta'} = \mathbf{0} \quad \text{vs.} \quad \mathbf{H_1} : \boldsymbol{\beta'} \neq \mathbf{0},$$

where $\boldsymbol{\beta'}$ is a vector containing a subset of the regression coefficients which we want to test. The null hypothesis states that all the regression coefficients contained in $\boldsymbol{\beta'}$ are equal to zero, while the alternative hypothesis states that one or more of the regression coefficients in $\boldsymbol{\beta'}$ are unequal to zero. The likelihood ratio is given by

$$\lambda = \frac{L(\hat{\boldsymbol{\beta}}_{\boldsymbol{small}})}{L(\hat{\boldsymbol{\beta}}_{\boldsymbol{large}})},$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{small}}$ denote the vector of maximum likelihood estimators for the model where all the regression coefficients in $\boldsymbol{\beta'}$ are equal to zero, and $\hat{\boldsymbol{\beta}}_{\boldsymbol{large}}$ is vector for the maximum likelihood estimators for the model where all the regression coefficients in $\boldsymbol{\beta'}$ are unequal to zero. For large sample sizes, $-2\log\lambda$ is approximately χ^2-distributed with degrees of freedom equal to the difference in number of regression coefficients between the two models.

The deviance for a model with estimated coefficients $\hat{\boldsymbol{\beta}}$ is given as

$$D(\hat{\boldsymbol{\beta}}) = -2\log\frac{L(\hat{\boldsymbol{\beta}})}{L(saturatedmodel)},$$

where $L(saturatedmodel)$ is the likelihood for the saturated model. When comparing the deviances for the larger and smaller model, the likelihoods for the saturated model cancels and we have that

$$\chi^2 = -2\log\lambda = D(\hat{\boldsymbol{\beta}}_{\boldsymbol{small}}) - D(\hat{\boldsymbol{\beta}}_{\boldsymbol{large}}) = -2\big(\log L(\hat{\boldsymbol{\beta}}_{\boldsymbol{small}}) - \log L(\hat{\boldsymbol{\beta}}_{\boldsymbol{large}})\big)$$

which for large sample sizes n is approximately χ^2-distributed with the difference in number of parameters as degrees of freedom. If the p-value for χ^2 is less than the significance level, the null hypothesis is rejected and the LRT indicate that the coefficient(s) should be included in the model.

3.2 Model evaluation and diagnostic tests

After having chosen and fitted the model, one is often interested in the accuracy of the model. Here, we consider methods for validation of the model and some measures for the model accuracy, in terms of diagnostic test, Brier score and ROC-curves. We start by describing a diagnostic test for binary data.

A diagnostic test can be used to investigate how well the model predicts the true data (Lydersen, 2012). We call the results from the model fit for the "test results" and the true outcome for the individuals for "disease status". For a binary test with two possible disease statuses, there are four possible outcomes. For a positive test, the test can either be true positive, if the individual is diseased, or false positive, if the individual is not diseased. The same applies for a negative test, which has the two options; true negative and false negative. Table 3.1 shows the possible combinations for the test results and the disease status.

	Test results		
Disease Status	Positive (T=1)	Negative (T=0)	Total
Diseased (D=1)	True positive (a)	False negative (b)	a+b
Non-diseased (D=0)	False positive (c)	True negative (d)	c+d
Total	a+c	b+d	a+b+c+d

Table 3.1: Possible outcomes of a diagnostic binary test with binary disease status.

The sensitivity and specificity of a test are the probabilities that the test results will give the true disease status of the individual. Hence they are direct properties of the test,

$$\text{Sensitivity} = P(\text{Positive test}|\text{Diseased}) = P(T = 1|D = 1),$$

and

$$\text{Specificity} = P(\text{Negative test}|\text{Non-diseased}) = P(T = 0|D = 0).$$

Using Table 3.1, the sensitivity and specificity of the test can be estimated by

$$\text{Sensitivity} = \frac{a}{a+b},$$

and

$$\text{Specificity} = \frac{d}{c+d}.$$

Receiver Operating characteristic Curve and Area Under Curve

As the estimated outcome of a logistic regression is a probability for $y_i = 1$, one need to determine a cutoff value for which the probability for $y_i = 1$ indicate a positive test. If the estimated probability is below the cutoff value, the predicted value is 0, while an estimate above or equal to the cutoff value, the predicted value is 1. One can compute the sensitivity and specificity for every possible cutoff value. A plot of these sensitivities as a function of $1-$ specificity is called a receiver operating characteristic curve (ROC). The area under the ROC-curve (AUC) is a measure of the test's ability to distinguish between the diseased and the not diseased. If the AUC is equal to 1, the ROC curve would go as straight lines from (0,0) up to (0,1) and further to (1,1), and we would have a perfect diagnostic test. Table 3.2 shows a general rule for strength of discrimination for different AUC values, which can be used to determine how well the test, and hence model, performs and to compare tests.

AUC	Strength of discrimination
0.5	No discrimination
$0.7 - 0.79$	Acceptable
$0.8 - 0.89$	Excellent
$0.9 - 1.0$	Outstanding

Table 3.2: General rule for strength of discrimination for different AUC values, (Lydersen, 2012).

Briers score

The Brier score is a popular tool in medical research for assessment and comparison of binary predictions (Rufibach, 2010). It is defined as,

$$B = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2,$$

where y_i is the true observed outcome and $\hat{\pi}_i$ is the estimated probability for $y_i = 1$. The Brier score is hence a mean square difference between the observed outcome and the predicted probability for the outcome. Comparing two binomial models in terms of Brier scores, the best model is the one with the lowest Brier score.

Cross validation

An internal validation of the estimated model is found by first fitting the model with a dataset, and then use the model to predict the outcomes in the same dataset. In this way, the same dataset is used for model fitting and model evaluation.

To have an external validation of the estimated model, cross validation can be applied. The dataset is divided into a set of test sets and training sets. A model is estimated based on the training sets and this model is used to predict the outcomes of the test sets. In this way, the model is validated on an external dataset. For a logistic regression model, the probabilities for $y_i = 1$ of the test set is predicted by the model based on the training set, and the predicted probabilities of the test set can be compared to the true outcomes.

In our analysis, k-fold cross validation and leave-one-out cross validation will be considered. A k-fold cross validation randomly divide the dataset into k folds. A model is estimated using the $k-1$ sets as training sets, and external validation is performed on the remaining k'th set. This is done repeatedly for all k, such that all the sets are used in the external validation exactly once. This is an advantage for small sample sizes.

A leave-one-out cross validation is similar to the k-fold cross validation, dividing the data into n folds. This method estimates a model using $n-1$ of the n observations, and test the model on the remaining observation. This is repeated until all observations have been predicted. In this way, all the data are externally validated and the model is estimated with a large amount of the available data, which is also an advantage for small sample sizes.

To compare models, the external validated data can be used for calculations of sensitivity and specificity, ROC-curves with corresponding AUC-values and Brier scores. The values for each of the evaluation methods can be used for model comparison in order to select the best model.

3.3 The Least Absolute Shrinkage and Selector Operator (Lasso)

Next, we look into methods for variable selection. We start by shorty presenting some methods for regression and variable selection, before we look into the method of the Least Absolute Shrinkage and Selector Operator (the Lasso). For simplicity, we start by describing the methods for linear models, before we look into details of the Lasso method for logistic regression models. At the end of the section, we consider methods for model evaluation for the Lasso model.

3.3.1 Overview of methods for variable selection and model estimation

Suppose that the data $y_1, ..., y_n$ are independent realizations from a normal distribution, $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\eta}, \sigma^2\mathbf{I})$, with mean $\boldsymbol{\eta}$ and covariance matrix $\sigma^2\mathbf{I}$. Suppose further that we have data on p predictors $\mathbf{x_1}, ..., \mathbf{x_p}$ which takes values x_{i1}, ..., x_{ip} for the i'th unit. Assuming that the expected responses are linear functions of these predictors, the linear predictor

takes the form

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{i=1,...,n}$$

for some unknown regression coefficients $\beta_0, ..., \beta_p$. When observing the data $(\mathbf{x_1}, y_1)$, ..., $(\mathbf{x_n}, y_n)$, the most popular estimation method for the regression coefficients is the *least squares* method (Hastie et al., 2001), where one minimizes the residual sum of squares,

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \eta_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

Assuming that $\mathbf{X}$ is of full rank, the least square estimates that minimizes the RSS are given by the unique solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}.$$

If $\mathbf{X}$ is not of full rank, then $\mathbf{X}^T\mathbf{X}$ is singular, and the least square estimates $\hat{\boldsymbol{\beta}}$ are no longer unique.

When $n \gg p$ the least square estimates tend to have low variance and bias, and give a good fit to the model, provided that the true relationship between the response and the predictors is approximately linear (James et al., 2013). However, when n is not much larger than p, there can be large variability in the least square fit, and the fitted model might prone to overfitting and perform poorly when predicting future observations. Also, when $p > n$, there are no longer unique least square estimates, the model will almost surely overfit the data, the variance is infinite and the method can not be used.

Another problem with the least square estimates are the complex interpretation of the model, if it includes many variables. Often, many of the variables are not associated with the response, and are irrelevant in the model. The least square estimates are unlikely to estimate some coefficients to zero, and hence, all variables are included in the model. To exclude the irrelevant or nearly irrelevant variables, their coefficients can be set equal to zero.

Fortunately, there are methods that overcome these problems of least square estimates. Below we shortly describe three popular methods that avoids either problems with model fitting when $p > n$ and/or inclusion of irrelevant variables in the model.

- Best-Subset Selection: For each $k \in \{1, ..., p\}$, the method finds the subset of size k with the smallest residual sum of squares. Among these p chosen subsets, the method chooses the one with the best tradeoff between bias and variance. AIC, adjusted R^2 and deviance are popular selection criteria. One problem with the best subset selection is the that there are 2^p possible subsets. Then, when p is large there will be an enormous amount of possible sets, which will slow down the algorithm. Take for example $p = 20$. This will give over one million possibilities. Having $p = 30$ there will be more than one billion possibilities. Another problem is that the larger the search space, the higher is the chance of finding models that look good in the test sets, but not for future datasets.

- Stepwise selection: A less computationally demanding method than best subset is the method of the stepwise selection. This method explore a far more restricted set of models. In forward stepwise selection, one starts with only the intercept in the model and sequently add variables that improves the fit of the model. In this way, the k variables included in the k'th model will also be included in the $k+l$'th model for all $l \geq 1$. Hence, the search space is reduced for each added variable. Backward subset selection is similar to the forward method, but starts with all p variables, and sequently remove the variables that have less impact on the fit. Both forward and backward subset selection have $\sum_{k=0}^{p-1}(p-k)$ possible sets, so with $k = 20$ there are only 221 subsets. In both methods, least squares are used to fit the subsets, so for high dimensions where $p > n$, only the forward selection method is possible, but then the estimates are not unique.

  Even though stepwise selection method is computationally more efficient than best subset, the stepwise methods have some disadvantages. The first problem is that the solutions are not unique for $p > n$ for the forward selection, while they do not exist for backward selection. In addition, the p-values for the coefficients do not take into account the multiple testing, so they can not be fully trusted (Finos et al., 2010). Another problem with the methods is that it might not give the best model. One could have that the variables added early in the algorithm might not be that important when other variables are included. Since the stepwise selection is a discrete process, it often suffer from large variances of the estimates and hence doesn't reduce the prediction error in the full model. Shrinkage methods, however, are continuous processes, and doesn't suffer that much from high variability (James et al., 2013).

- Shrinkage: Shrinkage methods uses all p variables to fit the model, but a penalty term is added to the expression that is to be minimized. By adding the penalty term, the model forces some coefficients towards zero, which has the effect of reducing the variance of the predicted values. For $p > n$, the least square estimates does not yield a unique solution, whereas the shrinkage methods performs well by trading off a small increase in bias for a large decrease in variance (Hastie et al., 2001). There are two popular shrinkage methods: the Ridge regression and the Lasso regression, where the Ridge regression estimates can be shrunken towards zero and the Lasso regression estimates can be shrunken to exactly zero. Hence, the Lasso regression also performs variable selection.

In the following section we describe the Lasso regression method in detail. Since we are interested in finding out which GMT-variables that are important in the prediction of cerebral palsy, the Lasso is a suitable method, as the variables that are not important are shrunken to zero and makes the important variables stand out.

3.3.2 The Lasso

In the Lasso regression, a shrinkage penalty is added to the residual sum of squares, imposing a penalty on the size of the parameters. The Lasso coefficients $\boldsymbol{\beta}_\lambda^L$ are the ones that

minimizes the quantity

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \frac{1}{2n}RSS + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{3.4}$$

where $\sum_{j=1}^{p}|\beta_j|$ is the ℓ_1 penalty and $\lambda \geq 0$ is the tuning parameter, to be determined separately (Hastie et al., 2015). The tuning parameter controls the impact of the two terms in Equation (3.4), where the RSS-term is small when the model fits the data well, and the shrinkage penalty $\lambda\sum_{j=1}^{p}|\beta_j|$ is small when the coefficients are close to zero. The Lasso regression differs from the Ridge regression by the ℓ_1 penalty, where Ridge uses the $\ell_2 = \sum_{j=1}^{p}\beta_j^2$ penalty. For sufficiently large values of λ, the ℓ_1 penalty shrinks some coefficients to be exactly zero, while for Ridge regression they are shrunken towards zero, but never exactly to zero. In this way, the Lasso regression also performs variable selection.

An equivalent formulation of Equation (3.4) is

$$\underset{\beta}{\text{minimize}}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s. \tag{3.5}$$

For Ridge regression, the constraint is now $\sum_{j=1}^{p}\beta_j^2 \leq s^2$. Figure 3.1 illustrates the differences between the Ridge and the Lasso regression for $p = 2$. For only two parameters, the constraints are $|\beta_1| + |\beta_2| < s$ for the Lasso coefficients and $\beta_1^2 + \beta_2^2 < s^2$ for the Ridge coefficients. These form a diamond and circle in the figure, while the ellipsis are constant values for the RSS. As the ellipsis expand from the least square estimates $\hat{\boldsymbol{\beta}}$, the RSS increases. Equation (3.5) indicate that the Lasso and Ridge coefficients estimates are given by the first point the ellipsis meets the constraint. Due to the corners of the diamond at the axis, the ellipse will often intersect with the axis, which corresponds to a coefficient being exactly equal to zero. This is not the case for the circle, where the estimated coefficients will be exclusively non-zero. For larger values of p, the dimension of the diamond and amount of corners increase, but the methodology holds. For p larger than 2, the intersection can take place in several corners.

When $\lambda \to \infty$, the impact of the shrinkage penalty grows, and all the Lasso coefficients will be exactly zero. When $\lambda = 0$, the Lasso regression model includes all the p predictors and the estimated coefficients are identical to the least square estimates (Hastie et al., 2015). In order to determine the value for the tuning parameter, one can apply k-fold cross validation. For a range of λ values, a model is fitted in the training set to obtain the Lasso coefficient estimates $\hat{\boldsymbol{\beta}}_\lambda^L$, and the fitted model is used for model validation in the test set. This results in k model validations for each value of λ in the given range. For each value of λ, one can calculate the mean prediction accuracy and standard errors from the mean (Hastie et al., 2015). We use the notation λ_{min} for the value of the tuning parameter giving the model with the "best" model validation, i.e. largest prediction accuracy, and λ_{1se} as the smallest value of the tuning parameter yielding the model which is within one standard error of the best model. The latter could be useful when the two models appears to be almost equally good, and one could choose the simplest model.

For models with highly correlated variables, the Lasso tend to pick only one of the correlated variables, and shrink the coefficients of the other correlated variables to zero

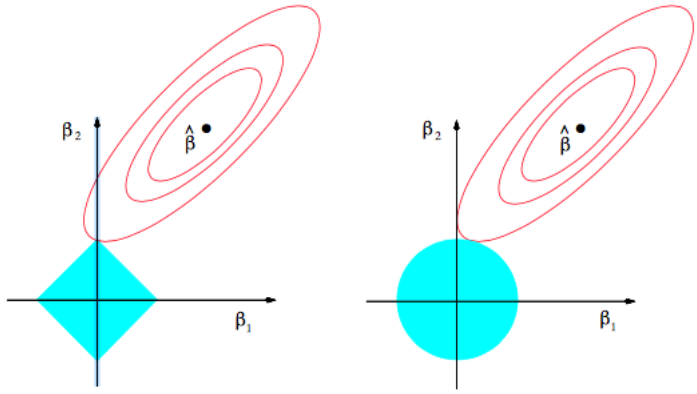


Figure 3.1: Estimation for the Lasso regression (left) and Ridge regression (right) for $p = 2$. The solid blue lines represent the constraint regions $|\beta_1| + |\beta_2| < s$ and $\beta_1^2 + \beta_2^2 < s^2$ respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point $\hat{\boldsymbol{\beta}}$ describes the least-squares estimate. The figure is copied with approval from Hastie et al. (2001).

(Friedman et al., 2010). The Ridge however, is known to shrink the coefficients of the correlated variables toward each other. Hence, correlated variables are allowed to borrow strength from each other in Ridge, while only one of the variables are chosen in the Lasso.

In the setting where the number of parameters are larger than the number of observations, $p > N$, the Lasso solution can only include N nonzero coefficients. Hence, if there are one million variables in a model, but only 100 observations, the Lasso solution can not include more than 100 nonzero coefficients.

Estimation of the Lasso coefficients

Let's first consider the case where we only have one predictor $\mathbf{x}_1$ and samples $\{(x_{1i}, y_i)\}_{i=1}^n$. We consider the standardized predictor z_i, which is standardized such that $\frac{1}{n}\sum_{i=1}^n x_i = 0$ and $\frac{1}{n}\sum_{i=1}^n x_i^2 = 1$. Then, the Lasso coefficient estimate is the one that solves

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - z_i\beta)^2 + \lambda|\beta| \right\}.$$

Simple differentiation with respect to β and putting the expression equal to zero, gives the Lasso coefficient estimate,

$$\hat{\beta}^L = \begin{cases} \frac{1}{n}\langle \mathbf{z}, \mathbf{y} \rangle - \lambda & \text{if } \frac{1}{n}\langle \mathbf{z}, \mathbf{y} \rangle > \lambda \\ \frac{1}{n}\langle \mathbf{z}, \mathbf{y} \rangle + \lambda & \text{if } \frac{1}{n}\langle \mathbf{z}, \mathbf{y} \rangle < -\lambda \\ 0 & \text{if } \frac{1}{n}|\langle \mathbf{z}, \mathbf{y} \rangle| \leq \lambda \end{cases} \tag{3.6}$$

where $\langle \mathbf{z}, \mathbf{y} \rangle = \mathbf{z}^T\mathbf{y} = \sum_{i=1}^n z_i y_i$. The notation in Equation (3.6) can be written as

$$\hat{\beta}^L = \mathcal{S}_\lambda(\frac{1}{n}\langle \mathbf{z}, \mathbf{y} \rangle).$$

The operator

$$\mathcal{S}_\lambda(x) = sign(x)(|x| - \lambda)_+$$

is called the soft-tresholding operator, and shrinks its argument x towards zero by the amount λ and is set to zero if $|x| \leq \lambda$ (Hastie et al., 2015).

Next, we consider a method for solving the Lasso problem for multiple predictors, called Cyclic Coordinate Descent. For this method, one repeatedly cycle through the predictors $j = 1, ..., p$ in a arbitrary order, but with the same order in each cycle. At the j'th step, the coefficient β_j is updated by minimizing the objective function

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j|,$$

while holding all other coefficients $\hat{\beta}_{k\neq j}$ fixed at their current value. Letting the partial residuals take the form $r_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k$, the outcomes from the current fit are removed for all predictors except the j'th. Then, the update for each $\hat{\beta}_j$ can be written as

$$\hat{\beta}_j = \mathcal{S}_\lambda(\frac{1}{n}\langle \mathbf{x}_j, \mathbf{r}^{(j)}\rangle).$$

Equivalently, letting the full residuals take the form $r_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$, the update can be written as

$$\hat{\beta}_j = \mathcal{S}_\lambda(\hat{\beta}_j + \frac{1}{n}\langle \mathbf{x}_j, \mathbf{r}\rangle).$$

This is done repeatedly in a cycling manner, updating each of the predictor in the model in every cycle (Hastie et al., 2015).

3.3.3 The Lasso for logistic regression models

Until this point, we have only been considering linear regression models for the Lasso. Now, we will consider the Lasso method for generalized linear models, with focus on logistic regression. For a generalized linear model, the Lasso solution is found by minimizing the negative log likelihood along with the shrinkage penalty,

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}}\Big\{ -\frac{1}{n}\mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) + \lambda||\boldsymbol{\beta}||_1\Big\},$$

where $\mathcal{L}$ is the log likelihood function of the generalized linear model with outcomes $\mathbf{y}$, coefficients $\boldsymbol{\beta}$ and covariates $\mathbf{X}$. For a logistic regression model, the log likelihood is given in Equation (3.3), and the Lasso coefficients are the ones that solves,

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}}\Big\{ -\frac{1}{n}\sum_{i=1}^{n}\Big(y_i(\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta}) - \log(1 + e^{\beta_0 + \mathbf{x_i}^T\boldsymbol{\beta}})\Big) + \lambda||\boldsymbol{\beta}||_1\Big\}. \tag{3.7}$$

The glmnet-package in R uses cyclic coordinate descent also when solving generalized linear models. One sufficient condition is that the function being minimized is continuously differentiable and strictly convex in each coordinate (Hastie et al., 2015). For

the Lasso solution of a generalized linear regression, this is fulfilled, as the negative log likelihood function is continuous, differentiable and convex and the shrinkage penalty is convex. The algorithm for finding the Lasso solutions consist of three loops, where the outer loop decrement λ. Since the log likelihood in Equation (3.7) is a concave function of the parameters, a iterative reweighted least squares can be used to maximize it. This leads to a quadratic objective function of the form,

$$\ell_Q(\beta_0, \beta) = -\frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \beta_0 - \mathbf{x_i}^T\boldsymbol{\beta})^2 + C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}),$$

where $z_i = \tilde{\beta}_0 + \mathbf{x_i}^T\tilde{\boldsymbol{\beta}} + \frac{y_i - \tilde{p}_i(\mathbf{x_i})}{\tilde{p}_i(\mathbf{x_i})(1-\tilde{p}_i(\mathbf{x_i}))}$ is the current observation, $\tilde{p}_i(\mathbf{x_i})$ is the current estimate for $P(Y = 1|\mathbf{X} = \mathbf{x_i})$, $w_i = \tilde{p}_i(\mathbf{x_i})(1 - \tilde{p}_i(\mathbf{x_i}))$ and C is a constant. With the current value of λ with corresponding current parameters $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$, the quadratic approximation ℓ_Q is computed in the middle loop. Then, in the inner loop, the coordinate descent is used to solve the problem

$$\underset{\beta_0, \boldsymbol{\beta}}{\text{minimize}} \Big\{ -\ell_Q(\beta_0, \boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1 \Big\}.$$

Using a warm start up on a fine grid values for λ, generally makes the quadratic approximation very accurate, and few iterations are required for convergence.

For a logistic regression model, one can consider several decision rules for selecting the value of λ that gives the best model fit from the cross validations. In this thesis, we focus on three decision rules: binomial deviance, misclassification error and AUC-values. The binomial deviance for the current λ value with the corresponding Lasso estimates $\hat{\boldsymbol{\beta}}_\lambda^L$ is given by

$$D_\lambda(\hat{\boldsymbol{\beta}}_\lambda^L) = -2\log\frac{L(\hat{\boldsymbol{\beta}}_\lambda^L)}{L(\text{saturated model})}.$$

Using the binomial deviance as decision rule, we seek the tuning parameter with corresponding Lasso estimates that minimizes the binomial deviance. This value for the tuning parameter is denoted λ_{min}.

Another option is to choose the best model based on the misclassification error. The fitted model is used to predict values for the outcomes y in the test set, and the number of misclassifications are counted based on a 0.5 cutoff. The best model is the one with the lowest misclassification error, and the value of the tuning parameter that gives the best model is denoted λ_{min}.

Also the AUC-value can be used as a decision rule for the cross validation of the model. Using this, one is interested in the model that gives the largest AUC-value for the test set, and we denote λ_{max} for it's corresponding tuning parameter.

3.3.4 Validation of the Lasso model

In addition to the predictive ability of the selected model, one are often interested determining the statistical strength of the included variables. Based on the adaptive nature of the estimation process, one can not easily assess p-values and confidence intervals for the variables (Hastie et al., 2015). Here, we describe two methods for making statistical inference of the variables: *Bootstrapping* and *Multi sample-splitting.*

Bootstrapping

Nonparametric bootstrapping is a method that can be used on the Lasso solution to asses the sampling distribution of the Lasso estimate $\hat{\boldsymbol{\beta}}_\lambda^L$ (Hastie et al., 2015). Assuming that the n observations $\{(\mathbf{x_i}, y_i)\}_{i=1}^n$ are independent and identically distributed by the cumulative distribution function F, a bootstrap sample is defined to be a random sample of size n from the empirical distribution function $\hat{F}_n$ that puts mass $1/n$ to each data point $(\mathbf{x_i}, y_i)$. A bootstrap sample is hence obtained by drawing n samples with replacement from the data. For each bootstrap sample, one can calculate the Lasso estimate $\hat{\boldsymbol{\beta}}_\lambda^L$, and repeating this process B times, one can use the B bootstrap samples to make inference about the parameters. In this thesis, we use the bootstrap samples to evaluate the number of times each coefficient for the variable is estimated to a non-zero value, and we look at the boxplots of their estimated values.

Multi sample-splitting

Multi sample-splitting is a method for constructing hypothesis tests and confidence intervals for regression parameters in a high dimensional setting, and are described in Dezeure et al. (2015). Here, we will not go into the details about the method, but only give a short description of the method.

The idea is to split the sample into two equal halves, I_1 and I_2, and use I_1 for model selection. The selected variables are gathered in $\hat{S}(I_1)$ and the covariates for the selected variables from the second half, $\mathbf{X}_{I_2}^{(\hat{S}(I_1))}$, is used for constructing p-values. In this thesis, we use the Lasso for variable selection. When testing $H_{0,j} : \beta_j^0 = 0$ for $j \in \hat{S}(I_1)$, we obtain the p-values, $P_{t-test,j}$, from the t-test when assuming Gaussian errors. Then, the raw p-value for the j'th variable is defined by,

$$P_{raw,j} = \begin{cases} P_{t-test,j} \quad \text{based on } Y_{I_2}, \mathbf{X}_{I_2}^{(\hat{S}(I_1))}, \text{ if } j \in \hat{S}(I_1) \\ 1, \quad \text{if } j \notin \hat{S}(I_1). \end{cases}$$

To correct the p-values for multiple testing, a Bonferroni corrected p-value for $H_{0,j}$ is given by

$$P_{corr,j} = \min(P_{raw,j} \cdot |\hat{S}(I_1)|, 1).$$

These adjusted p-values control the familywise error rate in multiple testing.

The procedure above is referred to as the single sample-splitting procedure. A problem with this approach is that the choice of sample splits can lead to wildly different p-values. To overcome this problem, the above procedure can be run B times, to obtain a collection of p-values for the j'th hypothesis $H_{0,j}$,

$$P_{corr,j}^{[1]}, ..., P_{corr,j}^{[B]} \quad \text{for } j = 1, ..., p$$

An aggregation to a single p-value P_j is obtained by defining an empirical γ-quantile with $0 < \gamma < 1$,

$$Q_j(\gamma) = \min(\text{emp. } \gamma\text{-quantile}\{P_{corr,j}^{[b]}/\gamma; b = 1, ..., B\}, 1).$$

Then, the aggregated p-value is obtained by choosing a properly scaled γ-quantile in the range $(\gamma_{min}, 1)$. Hence, the aggregated p-values are given by,

$$P_j = \min\Big((1 - \log(\gamma_{min}) \inf_{\gamma \in (\gamma_{min}, 1)} Q_j(\gamma), 1\Big) \quad \text{for } j = 1, ..., p. \tag{3.8}$$

The multi sample-splitting can easily be applied for generalized linear models, using the Lasso method for generalized linear models in the variable selection and constructing the p-values from the asymptotically distribution of the maximum likelihood estimator. The rest of the procedure is similar as described above.

Confidence intervals from the multi sample-split method for linear models are constructed based on the duality with the p-values from Equation (3.8). This method is explained in detail in Dezeure et al. (2015). Here, we only focus on confidence intervals for generalized linear models. The calculations of confidence intervals for these models are not described in Dezeure et al. (2015), so we have found the details by reading the R-code for the multi sample-split function. The R-code showed that for all other than linear models, the confidence intervals are calculated in a simpler manner. First, all the B confidence intervals are set to $(-\infty, \infty)$. Then for each loop, variable selection and model fitting is performed as described above. For a logistic regression model, the model with the selected variables is fitted as in Section 3.1.1 and the confidence intervals for the variables in the model is calculated as in Section 3.1.2. Selecting a significance level of 0.95, the multi sample splitting returns confidence intervals with lower and upper limit 1.3% and 98.8% respectively. For the variables that were not chosen in the given loop, the confidence interval for this loop is still $(-\infty, \infty)$. Then, based on all B loops, the median value of the confidence intervals for each variable is returned. In this way, for generalized linear models, the returned p-values from the function are adjusted for multiple testing, while the returned confidence intervals are not.

3.4 Mixed effects logistic regression with random intercepts

Now, we turn to statistical methods for modeling binary data with repeated measurements. In our data, we have repeated measurements for some of the infants, which we can model using mixed effects logistic regression with random intercepts. We start by describing the model, before we consider the frequentist approach for estimating and testing the regression coefficients in the model, as well as predictions and odds ratios for the mixed effect logistic regression model. Then, we look at some Bayesian theory for estimating the regression coefficients, before we describe the method for the Integrated Nested Laplace Approximation (INLA) in details. At the end of the chapter, we look at the Bayesian approach using INLA for predictions and odds ratios for mixed effect logistic regression models.

A mixed effects logistic regression model is similar to a logistic regression model, but it includes random effects. The repeated measurements within a cluster, like hospitals, countries or individuals, might be more equal to one another than two measurements for two different clusters. Due to this, one can not assume that all observations are independent, as for a logistic regression (Rabe-Hesketh and Skrondal, 2012). However, when

adding random intercepts to the model, the observations for individual j in different occasions i are independent, given the covariates and the random intercept. Adding random intercepts to the model gives the following conditional logit probability

$$logit(\pi_{ij}) = \beta_0 + \mathbf{x_{ij}^T}\boldsymbol{\beta} + \zeta_j$$

with

$$\pi_{ij} = P(y_{ij} = 1|\zeta_j) = \frac{exp(\beta_0 + \mathbf{x_{ij}^T}\boldsymbol{\beta} + \zeta_j)}{1 + exp(\beta_0 + \mathbf{x_{ij}^T}\boldsymbol{\beta} + \zeta_j)} \tag{3.9}$$

where i is the i'th measurement within individual j and ζ_j is the random intercept for individual j. The parameter π_{ij} is the probability of observation y_{ij} being equal to 1, and $\boldsymbol{x_{ij}}$ is the vector of covariates. The fixed intercept, β_0, is the mean intercept for all individuals, and the random intercept ζ_j is particular for the individual j. The random intercepts $\zeta_j \sim \mathcal{N}(0, \psi^2)$ are assumed to be independent and identically distributed across individuals and independent of the covariates $\boldsymbol{x_{ij}}$. Given ζ_j and $\boldsymbol{x_{ij}}$, the responses $y_{ij}, i = 1, ..., n_j$ and $j = 1, ..., n$ for individual j with observation i, are independently Bernoulli distributed. Hence, a mixed effects logistic regression model with random intercepts takes the form

$$\begin{aligned} y_{ij}|\pi_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\ logit(\pi_{ij}) &= \beta_0 + \mathbf{x_{ij}^T}\boldsymbol{\beta} + \zeta_j \\ \zeta_j &\sim \mathcal{N}(0, \psi^2). \end{aligned}$$

3.4.1 Frequentist approach

The marginal likelihood is the joint probability of all responses for all individuals given the covariates,

$$L(\boldsymbol{\beta}, \psi) = \prod_{j=1}^{N} P(y_{1j}, ..., y_{n_j j}|\boldsymbol{x_{ij}}, \boldsymbol{\beta}, \psi), \tag{3.10}$$

where $\boldsymbol{\beta}$ includes the intercept β_0. When the model includes random intercepts $\boldsymbol{\zeta}$, the responses are conditionally independent given the random intercept ζ_j and the covariates $\boldsymbol{x_{ij}}$. Hence, the joint probability for the responses y_{ij} for individual j, given the random intercept and covariates, is the product of conditional probabilities for the individual responses,

$$P(y_{1j}, ..., y_{n_j j}|\boldsymbol{x_{ij}}, \zeta_j) = \prod_{i=1}^{n_j} P(y_{ij}|\boldsymbol{x_{ij}}, \zeta_j) = \prod_{i=1}^{n_j} \pi_i^{y_{ij}}(1 - \pi_i)^{1-y_{ij}}.$$

Using the expression for π_{ij} given in Equation (3.9), we find after some calculation that

$$P(y_{1j}, ..., y_{n_j j}|\boldsymbol{x_{ij}}, \zeta_j) = \prod_{i=1}^{n_j} \frac{[exp(\beta_0 + \boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \zeta_j)]^{y_{ij}}}{1 + exp(\beta_0 + \boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \zeta_j)}.$$

To obtain the marginal joint probability for the responses, not conditioning of the random intercept, the random intercept can be integrated out,

$$P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}}) = \int_{-\infty}^{\infty} P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},\zeta_j)\phi(\zeta_j;0,\psi^2)d\zeta_j,$$

where $\phi(\zeta_j;0,\psi^2)$ is the normal density for the random intercept, with mean 0 and variance ψ^2. This integral doesn't have a closed form solution, and must be approximated numerically. In the glmer()-function in the lme4-package in R (Bates et al., 2015), the approximation is performed using an adaptive Gauss-Hermite approximation. Details about the Gauss-Hermite quadrature points and locations are described in Süli and Mayers (2003). In this thesis we focus on why we can approximate the integral using adaptive Gauss-Hermite approximation, based on Rabe-Hesketh et al. (2005).

Transforming the integration limits with $v_j = \zeta_j/\psi$, one get the standard normal density for v_j and the integral takes the form

$$P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{v_j^2}{2}} \prod_{i=1}^{n_j} \frac{[exp(\beta_0 + \boldsymbol{x_{ij}}^T\boldsymbol{\beta} + v_j\psi)]^{y_{ij}}}{1+exp(\beta_0 + \boldsymbol{x_{ij}}^T\boldsymbol{\beta} + v_j\psi)} dv_j.$$

Now, since this integral can be expressed on the form $\int exp(-x^2)f(x)dx$ it can be approximated numerically using Gauss-Hermite quadrature,

$$\int_{-\infty}^{\infty} \phi(v_j)P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},v_j)dv_j \approx \sum_{r=1}^{R} w_r P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},v_r = e_r), \quad (3.11)$$

where $\sqrt{\pi}w_r$ and $e_r/\sqrt{2}$ are the weights and locations of the r'th point of the Gaussian quadrature, and R is the number of quadrature points.The method is exact whenever $f(x)$ is a polynomial of degree less than $2R-1$.

A problem can occur when the function being integrated has sharp peaks. Using adaptive Gauss-Hermite quadrature can be an improved approximation in this situation. In the adaptive method, the properties of the integrand $\phi(v_j)P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},v_j)$ in (3.11) are taken into account. Using that the integrand is a product of a prior distribution for v_j, $\phi(v_j)$, and the joint probability of the responses given v_j, $P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},v_j)$, which can be thought of as a likelihood, the integrand can be used as a posterior distribution. Assuming large cluster sizes, n_j, and that the prior and the likelihood are positive and twice differentiable, the Bayesian central limit theorem states that the posterior density can be approximated by a normal distribution (pages 142-143 in Carlin and Louis (1996)).

If μ_j and τ_j^2 are the mean and variance of the posterior density, one would therefore expect that the ratio $\phi(v_j)P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},v_j)/g(v_j;\mu_j,\tau_j^2)$ would be well approximated by a low degree polynomial, where $g(v_j;\mu_j,\tau_j^2)$ is a normal distribution with mean μ_j and variance τ_j^2. Then, letting the integral take the form

$$\int_{-\infty}^{\infty} g(v_j;\mu_j,\tau_j^2)\Big(\frac{\phi(v_j)P(y_{1j},...,y_{n_jj}|\boldsymbol{x_{ij}},v_j)}{g(v_j;\mu_j,\tau_j^2)}\Big)dv_j$$

and changing the integration limits to $z_j = (v_j - \mu_j)/\tau_j$, it can be approximated by

$$\sum_{r=1}^{R} q_{jr} P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}}, \alpha_{jr}),$$

using Gauss-Hermite quadrature. The quadrature weights and locations are now given by

$$\begin{aligned} q_{jr} &= \sqrt{2\pi}\tau_j \exp(e_r^2/2)\phi(\mu_j + \tau_j e_r)p_r. \\ \alpha_{jr} &= \mu_j + \tau_j e_r \end{aligned} \tag{3.12}$$

The quadrature weights and locations are found through an iterative approach where solving Equation (3.12) with starting values $\mu_j^{(0)} = 0$ and $\tau_j^{(0)} = 1$. The k'th iteration is given by

$$\begin{aligned} P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}})^{(k)} &= \sum_{r=1}^{R} q_{jr}^{(k-1)} P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}}, \alpha_{jr}^{(k-1)}), \\ \mu_j^{(k)} &= \frac{\sum_{r=1}^{R} \alpha_{jr}^{(k-1)} q_{jr}^{(k-1)} P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}}, \alpha_{jr}^{(k-1)})}{P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}})^{(k)}} \\ \tau_j^{(k)} &= \sqrt{\frac{\sum_{r=1}^{R} (\alpha_{jr}^{(k-1)})^2 q_{jr}^{(k-1)} P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}}, \alpha_{jr}^{(k-1)})}{P(y_{1j}, ..., y_{n_j j} | \boldsymbol{x_{ij}})^{(k)}} - (\mu_j^{(k)})^2}, \end{aligned}$$

and the procedure runs until convergence.

When the integrals are approximated for all individuals j, the expressions can be inserted in Equation (3.10) to find the likelihood. The log likelihood can be maximized numerically with respect to $\boldsymbol{\beta}$ and ψ to find the estimated coefficients of the model.

The advantage of adaptive quadrature against ordinary quadrature is illustrated in Figure 3.2. Using adaptive quadrature, the weights are located directly under the integrand, as opposed to ordinary quadrature, where they are located under the prior density. The drawback of using quadrature rules to approximate the integral is that the algorithm might be very slow and can fail to converge when not using sufficiently number of quadrature points. Also, the approximation or a normal distribution for the posterior distribution might not be correct when n_j is too small.

Likelihood ratio test for random intercepts

The likelihood ratio test (LRT) can be used to investigate if random intercepts should be included in the model. Testing the random intercepts $\boldsymbol{\zeta}$, the hypothesis is the following,

$$\mathrm{H}_0 : \psi^2 = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \psi^2 > 0,$$

where ψ^2 is the variance for the random intercept. This test is equivalent to testing the hypothesis $H_0 : \zeta_j = 0$ for all j. The difficulty here is that the test statistic $-2\log\lambda$ is not χ^2-distributed with one degree of freedom under H_0, because the null hypothesis is on the

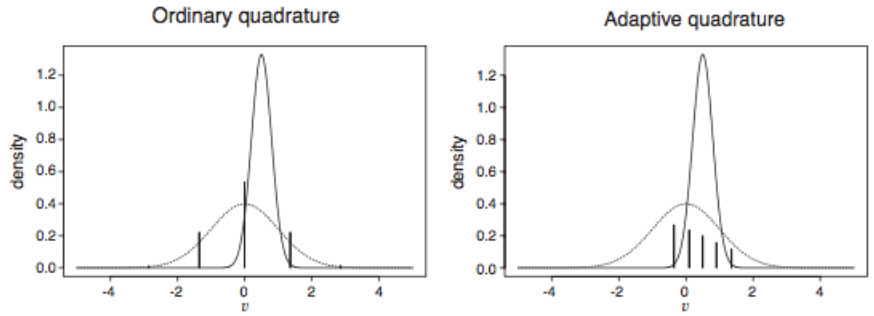


Figure 3.2: Prior (dotted curve) and posterior (solid curve) densities together with the quadrature weights (bars) for ordinary and adaptive quadrature, from Rabe-Hesketh et al. (2005) with approval.

boundary of the parameter space since $\psi^2 \geq 0$. One can prove that the correct asymptotic sampling distribution is a 50:50 mixture of a χ^2_0-distribution, which is a spike at 0, and a χ^2_1-distribution (Self and Liang, 1987). Due to this, the correct p-value for the test can be obtained by dividing the p-value obtained by the LRT with one degree of freedom by two. A p-value less than the significance level leads to the rejection of the hypothesis, and the random intercepts should be included in the model.

Prediction

The probability π_{ij} for $y_{ij} = 1$ given in Equation (3.9), is the conditional or *subject specific* probability. When doing predictions for π_{ij}, given the covariates, one can either look at the predicted subject specific probability or the predicted *population averaged* probability (Rabe-Hesketh and Skrondal, 2012). For predictions of the outcome for new data, we look at the population averaged predicted probabilities. Using the obtained estimated coefficients from the fitted model and the covariates for the new data, the population averaged predicted probabilities are found by averaging out the unknown random intercept for the new data with numerical integration,

$$\begin{aligned}\hat{\pi}_{ij} &= \int \hat{P}(y_{ij} = 1|\mathbf{x_{ij}}, \zeta_j)\phi(\zeta_j; 0, \hat{\psi}^2)d\zeta_j \\ &= \int \frac{exp(\mathbf{x_{ij}^T}\hat{\boldsymbol{\beta}} + \zeta_j)}{1 + exp(\mathbf{x_{ij}^T}\hat{\boldsymbol{\beta}} + \zeta_j)}\phi(\zeta_j; 0, \hat{\psi}^2)d\zeta_j.\end{aligned}$$

Hence, the predicted probabilities for the new observations are based on the population average. Note that these are not equal to to the predicted probabilities when setting the random intercept to zero.

$$\hat{\pi}_{ij} \neq \frac{exp(\mathbf{x_{ij}^T}\hat{\boldsymbol{\beta}} + 0)}{1 + exp(\mathbf{x_{ij}^T}\hat{\boldsymbol{\beta}} + 0)}. \tag{3.13}$$

When the true outcome is known or if the new observation is part of a cluster with estimated values, one can consider the subject specific probabilities in the prediction. For

these data, the random intercept has been estimated in the model fit and are considered known. Then, the predicted probabilities are found though simple calculations using the estimated coefficient values and the estimated random intercept, together with the covariates,

$$\hat{\pi}_{ij} = \frac{exp(\mathbf{x_{ij}^T}\hat{\boldsymbol{\beta}} + \hat{\zeta}_j)}{1 + exp(\mathbf{x_{ij}^T}\hat{\boldsymbol{\beta}} + \hat{\zeta}_j)}.$$

The subject specific probabilities describe the probability of $y_{ij} = 1$ within the cluster j. Examples can be participants within hospitals, where the probabilities describe the the probability for participant i having a disease within hospital j. On a population level, we would interpret the population averaged probabilities as the probability that participant i have the disease, regardless of the hospital. For our analysis, the clusters are the participants, and there are repeated measurements within the participants. Since we are interested in making predictions for recordings from new participants, we will in this thesis only consider the population averaged probabilities.

Using the predict function in the lme4-package in R, one can choose between the predicted population averaged and the predicted subject specific probabilities. However, it turns out that for population averaged probabilities, the function does not integrate out the random intercept (Pavlou et al., 2015), it just let the random intercept be equal to zero. Hence, it returns the probabilities in Equation (3.13). In lack of better prediction methods, we will show these predicted probabilities in the results, but we can not interpret them as population averaged probabilities.

Odds ratio for the mixed effects logistic regression model

When interpreting the odds ratios for a mixed effects logistic regression model with random intercepts, one should note that these are subject specific odds ratios, while the odds ratios for a logistic regression model are population averaged odds ratios. To see this, we look at the odds for observation y_{ij},

$$\frac{\pi_{ij}}{1 - \pi_{ij}} = exp(\beta_0 + \mathbf{x_{ij}}^T\boldsymbol{\beta} + \zeta_j).$$

Now, given that the random intercepts is constant and increasing covariate x_{ijk} by one unit, we have that

$$\begin{aligned}\frac{\pi_{ij}}{1 - \pi_{ij}} &= exp(\beta_0 + \mathbf{x_{ij}}^T\boldsymbol{\beta} + \beta_k + \zeta_j) \\ &= exp(\beta_0 + \mathbf{x_{ij}}^T\boldsymbol{\beta} + \zeta_j)exp(\beta_k).\end{aligned}$$

Given that the random intercept is constant, the odds increase by $exp(\beta_k)$ as for a logistic regression model. Hence, the subject specific odds ratio is $exp(\beta_k)$ for a unit increase in x_{ijk}.

The confidence interval for the odds ratio must also be interpreted as subject specific, but the calculations are similar. The estimated coefficients for the fixed effects are normally distributed for large cluster sizes n_j, also in the mixed effect logistic regression model, and the fixed effects are independent of the random intercept. Hence, the confidence interval for the subject specific odds ratio is calculated in the same way as for a logistic regression model.

3.4.2 Bayesian approach

Until this point, we have been looking at the frequentist approach for doing inference. Now, we turn to the Bayesian approach. The Bayesian approach differs from the frequentist approach by treating the parameters of interest as random variables instead of fixed quantities. In this way, one consider the parameter distribution, not just it's value (Carlin and Louis, 1996). In a frequentist approach, one starts with a hypothesis, say $\theta = 0$ and calculates the probability of observing the outcome of the data when the hypothesis is true. In a Bayesian setting, one specifies a prior distribution of the parameter of interest, say $p(\theta)$ and uses the observed data to update the prior, yielding the posterior distribution $p(\theta|\mathbf{y})$; the distribution of the parameter, given the observed data. Most of the following theory in this section is based on the book by Blangiardo and Cameletti (2015).

Bayes theorem

The well known Bayes theorem was introduced already in the 18th century by Thomas Bayes,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

More generically, if we let $B_1, ..., B_K$ be a set of mutually exclusive and exhaustive events, meaning that one of the event must occur and several events cannot occur simultaneously, we have that

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{K} P(A|B_i)P(B_i)}.$$

The application of Bayes theorem to observable events is uncontroversial and well established. Having observable events where the probabilities are frequencies of observed events, one can easily use Bayes theorem to calculate a value for the probability of interest. When applying it in Bayesian inference, we do calculations on probability distributions.

Bayesian inference

In this setting, we are interested in some parameters $\boldsymbol{\theta}$. Without observing the data, we have some prior beliefs about the parameters, and assume that they follow the the distribution $p(\boldsymbol{\theta})$, which is called the prior distribution. The observed data $\mathbf{y}$ are assembled in the likelihood function

$$L(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}),$$

which specifies the distribution of the data $\mathbf{y}$ given the parameters $\boldsymbol{\theta}$. Using Bayes theorem, we define the posterior distribution for parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

where $p(\mathbf{y})$ is the marginal distribution of the observed data $\mathbf{y}$. Since the marginal distribution of the observed data does not depend on $\boldsymbol{\theta}$, the posterior distribution is proportional

to the product of the likelihood and the prior,

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Hence, the posterior distribution contains information about the unknown parameter when having observed the data $\mathbf{y}$.

Hierarchical models

In some cases, one could have that the distribution of the parameters $\boldsymbol{\theta}$ depends on some hyperparameters $\boldsymbol{\phi}$. These hyperparameters can follow a distribution $p(\boldsymbol{\phi})$. Having different levels of the data, one could specify the model as a hierarchical model with three levels as follows:

- Level 1: $\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi} \sim p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})$
- Level 2: $\boldsymbol{\theta}|\boldsymbol{\phi} \sim p(\boldsymbol{\theta}|\boldsymbol{\phi})$
- Level 3: $\boldsymbol{\phi} \sim p(\boldsymbol{\phi})$.

Applying Bayes theorem on a three-level hierarchical model, the posterior distribution is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}).$$

Point estimation and credible intervals

For simplicity, lets go back to considering the two-level model. Having the posterior density $p(\theta|\mathbf{y})$, one can make statistical inference about the parameters of interest. Location estimates such as the the posterior mean

$$E(\theta|\mathbf{y}) = \int \theta f(\theta|\mathbf{y}) d\theta,$$

the posterior mode

$$\text{Mod}(\theta|\mathbf{y}) = \arg\max_{\theta} f(\theta|\mathbf{y}),$$

and the posterior median

$$\theta_M \Rightarrow \int_{-\infty}^{\boldsymbol{\theta_M}} f(\theta|\mathbf{y}) d\theta = \int_{\theta_M}^{\infty} f(\theta|\mathbf{y}) d\theta = 0.5,$$

can be calculated. Also, the credible interval

$$(\theta_L, \theta_U) \Rightarrow \int_{\boldsymbol{\theta_L}}^{\boldsymbol{\theta_U}} f(\theta|\mathbf{y}) d\theta = 1 - \alpha,$$

can be calculated. Note that the interpretation of a Bayesian credible interval is different from a frequentist confidence interval. A credible interval indicate that the parameter θ lies within the interval with probability $1 - \alpha$. A confidence interval, however, suggests that, if we do the same experiment a large number of times, the true parameter value θ would fall out of the interval $\alpha\%$ of the time.

Choice of prior

To be able to make inference and calculate these location estimates, a nice expression for the posterior is preferable. As the posterior is proportional to the likelihood and the prior, the choice of prior plays an important role in the distribution of the posterior.

According to Blangiardo and Cameletti (2015), there are two important aspects when selecting a prior. First, the type of distribution and second, the amount of information available provided through the hyperparameters. Knowing some quantities of the parameteres, e.g. if it is a fraction, positive or symmetric, there is often a "natural" candidate for the type of distribution. When choosing a prior such that the posterior distribution belong to the same family as the prior, the prior is called a *conjugate* prior. Having conjugate priors is convenient, as the posterior distribution and it's hyperparameters are known. When knowing the distribution, the summary statistics and other convenient quantities of interest are easy to calculate. Table 3.3 shows a list of the likelihood, the posterior and the prior for some conjugate models.

Likelihood	**Conjugate prior**	**Posterior distribution**
$y\|p \sim \text{Bin}(n,p)$	$p \sim \text{Be}(\alpha,\beta)$	$p\|y \sim \text{Be}(\alpha+y, \beta+n-y)$
$y\|\lambda \sim \text{Po}(e\lambda)$	$\lambda \sim \text{Ga}(\alpha,\beta)$	$\lambda\|y \sim \text{Ga}(\alpha+y, \beta+e)$
$y\|\lambda \sim \text{Exp}(\lambda)$	$\lambda \sim \text{Ga}(\alpha,\beta)$	$\lambda\|y \sim \text{Ga}(\alpha+1, \beta+y)$
$y\|\mu \sim \mathcal{N}(\mu,\sigma^2)^*$	$\mu \sim \mathcal{N}(\nu,\tau^2)$	$\mu\|y \sim \mathcal{N}((A)^{-1}(\frac{x}{\sigma^2}+\frac{\nu}{\tau^2}),(A)^{-1})$
$y\|\sigma^2 \sim \mathcal{N}(\mu,\sigma^2)^{**}$	$\sigma^2 \sim \text{IGa}(\alpha,\beta)$	$\sigma^2\|y \sim \text{IGa}(\alpha+\frac{1}{2}, \beta+\frac{1}{2}(x-\mu)^2)$

* : μ known.
** : σ^2 known.
$A = \frac{1}{\sigma^2}+\frac{1}{\tau^2}$

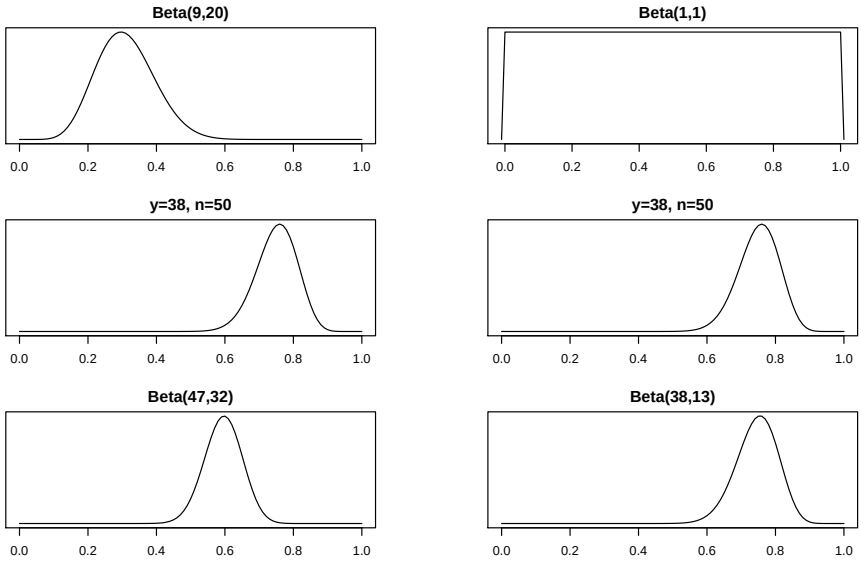
Table 3.3: Likelihood, prior and posterior distributions for some conjugate models.

Having chosen the distribution of the prior, the hyperparameters should be specified such that the distribution gives the desired amount of information. Depending on the amount of information, the prior could be informative or non-informative, relative to the likelihood (Box and Tiao, 1992). Non-informative priors are appealing in models where there is little prior information on the values of the parameter. They contain little information about the parameters besides its' range, such that the posterior distribution is decided mostly by the data. Non-informative priors are also called *flat* priors. Let's consider an example. Assume that we have binomial data and wants to assig a prior to the probability parameter p. A non-informative prior here would be $\pi(p) \sim Unif(0,1)$, which assign probability 1 to all values in the interval [0,1].

The major drawback of a non-informative prior is that it is not invariant for transformations of the parameters (Blangiardo and Cameletti, 2015). When transforming the parameters, the prior would not be flat any more. There are several rules proposed to make the prior not affected by the transformation, but a small problem with these priors is that most of them are improper, i.e. does not integrate to 1, which could lead to improper posteriors. To avoid these problems for non-informative priors, one could approximate a non-informative prior using a prior which is not non-informative on the entire parameter space. These type of priors are called *vague* priors and avoid the problems with impropri-

ety. Examples of vague priors are $\mathcal{N}(0, 10^6)$ and Gamma(0.01,0.01), which are often used as priors for regression parameters and the inverse of the variance.

An informative prior, however, assign different probabilities for different values in the range. Typically, these priors are based on previous findings or expert opinions. Using very informative priors, the posterior distribution will typically be closer to the prior distribution than to the likelihood. For a non-informative prior, the posterior would typically be closer to the likelihood, as it is mostly influenced by the data. Figure 3.3 shows examples of the posterior distribution when using an informative prior (a) and a non-informative prior (b). For the informative prior, with the mean has been centered between 0.2 and 0.4, one can see that the posterior distribution is affected by the prior, while in the non-informative case, the posterior is almost equal to the likelihood.



(a) Informative prior **(b)** Non-informative prior

Figure 3.3: Prior (top panel), likelihood (middle panel) and posterior distribution (bottom panel) for an informative prior and a non-informative prior.

We have seen how the posterior change with the prior, but the posterior is also affected by the amount of information in the data. This is easily seen on log scale where,

$$\log(p(\boldsymbol{\theta}|\mathbf{y})) = c + \log(L(\boldsymbol{\theta})) + \log(p(\boldsymbol{\theta})),$$

where c is a constant. Hence, if there is much information in the data, the posterior will be affected mostly by the likelihood. If there is little information in the data, the posterior would be more affected by the prior. To check how much information our data gives us, one could compare the posterior and prior distributions for different priors.

Challenges with the Bayesian approach

The use of priors is the most debated aspect of Bayesian statistics, (Carlin and Louis (1996), Blangiardo and Cameletti (2015), Box and Tiao (1992)). Inclusion of current knowledge is of course popular, as it would give strength to the result. On the other hand, it is feared that inclusion of other data besides the observed ones leads to biased results. Due to this, it is recommended that the posterior distribution is presented for different priors. In a medical setting, Adamina et al. (2009) recommended that the priors presented include skeptical, neutral and optimistic priors. An example of an optimistic prior is when testing the treatment effect of a disease. An optimistic prior for the probability of the treatment effect would be centered around a value for the probability corresponding to a high treatment effect.

We have seen that using conjugate priors give nice expressions for the posterior, as it can easily be calculated. However, conjugate priors are often not available. Some likelihoods, like for generalized linear regression models, does not have conjugate priors. For these models and other models where there are no conjugate priors, one needs to rely on simulation or approximation methods to do Bayesian inference about the parameters. Markov Chain Monte Carlo (MCMC) is a popular simulation method for making Bayesian inference (Blangiardo and Cameletti, 2015), but if the involved distributions are complex and there are large amount of data, the simulations can become computationally intensive and time consuming. In this thesis, we consider an approach by Rue et al. (2009) which has turned out to be both accurate and time efficient.

3.4.3 Bayesian inference using the Integrated Nested Laplace Approximation (INLA)

The INLA-algorithm is a deterministic algorithm designed for doing Bayesian inference on latent Gaussian models. A latent Gaussian model is a subclass of structured additive regression models, where the response y_i is assumed to come from an exponential family, and the mean μ_i is linked to a structural additive predictor η_i through the link function $g(\cdot)$, such that $g(\mu_i) = \eta_i$. The linear predictor account for effects of various covariates for the response on the form

$$\eta_i = g(\mu_i) = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ji} + \sum_{\gamma}^{n_f} f_\gamma(u_{\gamma,i}) + \epsilon_i \quad \text{for } i = 1, ..., n.$$

Here, the $\boldsymbol{\beta}$'s represents the effect of covariates $\mathbf{z}$, the $f_\gamma(\cdot)$'s are unknown functions for covariates $\mathbf{u}$, and the $\boldsymbol{\epsilon}$'s are the unstructured error terms. The applications for this type of model are very flexible due to the many different forms of the unknown functions $f_\gamma(\cdot)$, such as smooth and non-linear effects of the covariates, time trends, seasonal effects, random effects and temporal and spatial effects (Blangiardo and Cameletti, 2015). One of the most useful features of this type of model is that the effects can easily be added or removed from the model, while the model framework and computations stay the same.

All the latent, non-observable components can be collected in the *latent field* $\boldsymbol{\theta} = \{\boldsymbol{\eta}, \alpha, \boldsymbol{\beta}, f_1, f_2, ...\}$ with dimension n, which is typically very large ($10^2 - 10^5$) (Rue et al., 2016). Letting the latent field be controlled by the K hyperparameters, $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$,

and assuming that the responses are conditional independent, given the latent field and the hyperparameters, a hierarchical model can be applied here. Level 1 is the likelihood

$$\text{Level 1:} \quad \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi} \sim \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\phi}),$$

where each data point y_i is connected to only one element, θ_i of the latent field. Assuming Gaussian priors for all the components in the latent field, the joint distribution of the latent field is also Gaussian and hence, $\boldsymbol{\theta}$ is a latent Gaussian field. Assuming a latent Gaussian field with mean $\mathbf{0}$ and precision matrix $\mathbf{Q}(\boldsymbol{\phi})$, level 2 takes the form

$$\text{Level 2:} \quad \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q^{-1}}(\boldsymbol{\phi})) = (2\pi)^{-n/2}|\mathbf{Q}(\boldsymbol{\phi})|^{1/2}\text{exp}\Big(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{Q}(\boldsymbol{\phi})\boldsymbol{\theta}\Big),$$

where $|\cdot|$ denotes the matrix determinant. The hyperparameters in level 3 account for the variability and strength of dependence of the parameters in the latent field, and can take any distribution,

$$\text{Level 3:} \quad \boldsymbol{\phi} \sim p(\boldsymbol{\phi}).$$

Now, the joint posterior density of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is given by

$$\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) &\propto p(\boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \\
&= p(\boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi}) \prod_{i=1}^{n} p(y_i|\theta_i, \boldsymbol{\phi}) \\
&= p(\boldsymbol{\phi})|\mathbf{Q}(\boldsymbol{\phi})|^{1/2}\text{exp}\Big(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{Q}(\boldsymbol{\phi})\boldsymbol{\theta}\Big) \prod_{i=1}^{n} \text{exp}(\log(p(y_i|\theta_i, \boldsymbol{\phi}))) \\
&= p(\boldsymbol{\phi})|\mathbf{Q}(\boldsymbol{\phi})|^{1/2}\text{exp}\Big(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{Q}(\boldsymbol{\phi})\boldsymbol{\theta} + \sum_{i=1}^{n} \log(p(y_i|\theta_i, \boldsymbol{\phi}))\Big). \quad (3.14)
\end{aligned}$$

One main assumption when using INLA is that the latent field $\boldsymbol{\theta}$ is a Gaussian Markov Random Field (GMRF), meaning that in addition to being Gaussian, the latent field must also be conditionally independent. Then, θ_i and θ_j are conditionally independent given the remaining elements $\boldsymbol{\theta}_{-ij}$. A very useful consequence of this is that the precision matrix of the latent field gets very sparse, and calculations with sparse matrices can be very fast.

The idea in Bayesian inference is to use the posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})$$

to approximate the posterior marginals, $p(\theta_i|\mathbf{y})$ and $p(\phi_j|\mathbf{y})$ directly. The marginals are given by

$$\begin{aligned}
p(\theta_i|\mathbf{y}) &= \int p(\theta_i, \boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi} = \int p(\theta_i|\boldsymbol{\phi}, \mathbf{y})p(\boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi}, \quad \text{and} \\
p(\phi_j|\mathbf{y}) &= \int p(\boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi}_{-j}.
\end{aligned}$$

Thus, we need to perform two tasks: i) Compute $p(\boldsymbol{\phi}|\mathbf{y})$ from which all the relevant marginals $p(\phi_j|\mathbf{y})$ can be obtained, and ii) Compute $p(\theta_i|\boldsymbol{\phi}, \mathbf{y})$ which is needed to compute the marginal posteriors $p(\theta_i|\mathbf{y})$. In MCMC sampling, these marginals are computed through simulations, which can be a time-consuming process when the expressions are not that nice. The INLA-algorithm uses Laplace approximations to construct nested approximations,

$$\begin{aligned}\tilde{p}(\phi_j|\mathbf{y}) &= \int \tilde{p}(\boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi}_{-j}, \quad \text{and} \\ \tilde{p}(\theta_i|\mathbf{y}) &= \int \tilde{p}(\theta_i|\boldsymbol{\phi}, \mathbf{y})\tilde{p}(\boldsymbol{\phi}|\mathbf{y})d\boldsymbol{\phi}, \end{aligned} \tag{3.15}$$

where $\tilde{p}(\boldsymbol{\phi}|\mathbf{y})$ is the the approximated posterior distribution for the hyperparameters and $\tilde{p}(\theta_i|\boldsymbol{\phi}, \mathbf{y})$ is the approximated posterior marginals for the latent field. These integrations are solved using numerical integration, which is possible when the dimension of $\boldsymbol{\phi}$ is small (typically ≤ 6) (Rue et al., 2009). Hence, there are two main assumptions when using INLA: i) The latent field must be a GMRF, and ii) The number of hyperparameters must be small.

Laplace approximation

Before stating the method and steps of INLA, we pause to look at the Laplace approximation of a integral over the function $f(x)$. Let's assume that we are interested in computing the integral on the form

$$\int f(x)dx = \int \exp(\log f(x))dx.$$

Using a second order Taylor expansion centered around the mode x^*, we have that

$$\begin{aligned}\int f(x)dx &\approx \int \exp\Big(\log f(x^*) + \frac{(x-x^*)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}\Big)dx \\ &= \exp(\log f(x^*)) \int \exp(-\frac{(x-x^*)^2}{2\sigma_*^2}))dx,\end{aligned}$$

where $\sigma_*^2 = \Big[-\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}\Big]^{-1}$. Hence, the integrand is normally distributed with mean x^* and variance σ_*^2, so evaluating the integral $\int f(x)dx$ on the interval (α, β) gives approximately

$$\int_\alpha^\beta f(x)dx \approx f(x^*)\sqrt{2\pi\sigma_*^2}(\Phi(\beta) - \Phi(\alpha)),$$

where $\Phi(\cdot)$ is the cumulative density function for a Normal(x^*, σ_*^2).

Computation of $p(\phi_j|\mathbf{y})$

Returning to the approximations of the marginals, we begin with task i); computing $p(\phi_j|\mathbf{y})$. Applying Bayes theorem on the posterior, we have that

$$\begin{aligned} p(\boldsymbol{\phi}|\mathbf{y}) &= \frac{p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y})}{p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}, \boldsymbol{\phi})}{p(\mathbf{y})} \frac{1}{p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{y})} \\ &\propto \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})}{p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{y})}. \end{aligned}$$

Now, the numerator consist of only known terms. The denominator however, is unknown. For the INLA-method, if the denominator is not Gaussian, a Laplace approximation is used to approximate it by a Gaussian distribution. Using the expression for the posterior given in Equation (3.14), we have that

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\phi}) &\propto \exp\Big(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{Q}(\boldsymbol{\phi})\boldsymbol{\theta} + \sum_{i=1}^{n} \log(p(y_i|\theta_i, \boldsymbol{\phi}))\Big) \\ &\approx (2\pi)^{-n/2}|\mathbf{P}(\boldsymbol{\phi})|^{1/2}\exp\Big(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}(\boldsymbol{\phi}))^T \mathbf{P}(\boldsymbol{\phi})(\boldsymbol{\theta} - \boldsymbol{\mu}(\boldsymbol{\phi}))\Big), \end{aligned} \tag{3.16}$$

where $\mathbf{P}(\boldsymbol{\phi}) = \mathbf{Q}(\boldsymbol{\phi}) + \text{diag}(\mathbf{c}(\boldsymbol{\phi}))$ and $\boldsymbol{\mu}(\boldsymbol{\phi})$ is the location of the mode (Rue et al., 2016). The vector $\mathbf{c}(\boldsymbol{\phi}))$ is a vector of the i second derivatives of the negative log likelihoods with respect to θ_i, evaluated at the mode. This approximation turns out to be accurate since $p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{y})$ appears to be almost Gaussian as $\boldsymbol{\theta}$ is a GMRF (Blangiardo and Cameletti, 2015). Hence, the approximated posterior distribution for the hyperparameters becomes

$$\tilde{p}(\boldsymbol{\phi}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})}{\tilde{p}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\phi})}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\phi})}, \tag{3.17}$$

where $\boldsymbol{\theta}^*(\boldsymbol{\phi})$ is the mode for a given $\boldsymbol{\phi}$ and $\tilde{p}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\phi})$ is the Gaussian approximation in Equation (3.16). Now, since the dimension of $\boldsymbol{\phi}$ is low, the marginals $\phi_i|\mathbf{y}$ can be derived directly from Equation (3.17).

Computation of $p(\theta_i|\boldsymbol{y})$

The next task is to approximate the posterior marginals for the latent field. This task is more complicated because of the usually large dimension of the latent field $\boldsymbol{\theta}$. There are two challenges when approximating the posterior marginals for the latent field in Equation (3.15) (Rue et al., 2016). First, when the dimension of $\boldsymbol{\phi}$ is less than 2, classical numerical integration is applied, without too much computational cost. However, then the dimension is large, a standard numerical integration over $\boldsymbol{\phi}$ has a computational cost which is exponential in the dimension of $\boldsymbol{\phi}$. To perform the integration without a large computational cost, integration points on a sphere around the center is used. Details are explained in Rue et al. (2016). This is the default approach in INLA, and is shown to balance the computational cost and accuracy well. Other methods can be chosen to increase the accuracy, but at the expense of computational costs.

The second challenge is to approximate $p(\theta_i|\boldsymbol{\phi}, \mathbf{y})$ for a subset of all $i = 1, ..., n$. Due to the large dimension of $\boldsymbol{\theta}$, a standard application of the Laplace approximation will be too

demanding. The default approach in INLA is called the *simplified Laplace approximation*. This method computes a Taylor expansion around the mode of the Laplace approximation,

$$\log(p(\theta_i|\mathbf{y}, \boldsymbol{\phi})) \approx -\frac{1}{2}\theta_i^2 + b_i(\boldsymbol{\phi})\theta_i + \frac{1}{6}c_i(\boldsymbol{\phi})\theta_i^3. \tag{3.18}$$

In this way, a linear and a cubic correction term is provided to the Gaussian approximation. Matching a skew-Normal distribution to Equation (3.18), the linear term provides a correction term for the mean, and the cubic term provides a correction for the skewness. Hence, the posterior marginal for the latent field in Equation (3.15) is approximated by a mixture of skew-Normal distributions.

In this thesis, we use the INLA approach for fitting a mixed effect logistic regression model. For the analysis, we use the INLA-package in R, which uses the method described above to calculate posterior distributions for the regression coefficients. It has been shown that the INLA method is fast and accurate when fitting mixed logistic regression models with random intercepts, but the results depend on the choice of prior distribution, in particular in small samples (Grilli et al., 2015).

Prediction in the Bayesian approach

Once the model parameters are specified and derived, one are often interested in the model's ability to predict the outcomes $\mathbf{y}$. Suppose that a model has been fitted using the observed outcomes $\mathbf{y}$. Then, assuming that both a new occurrence y^* and the previous observations $\mathbf{y}$ are realizations from the distribution of Y, which is governed by the parameters in the latent field $\boldsymbol{\theta}$, the predictive distribution $p(y^*|\mathbf{y})$ is given by

$$\begin{aligned}
p(y^*|\mathbf{y}) &= \frac{p(\mathbf{y}, y^*)}{p(\mathbf{y})} \\
&= \frac{\int p(y^*|\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta})}{p(\mathbf{y})} \quad \text{by exchangeability} \\
&= \frac{\int p(y^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})d\boldsymbol{\theta}}{p(\mathbf{y})} \quad \text{applying Bayes theorem} \\
&= \int p(y^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.
\end{aligned}$$

Hence, the observed responses are used to update the uncertainty of the model into the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, which is used to do inference about the new occurrence y^* (Blangiardo and Cameletti, 2015).

In the INLA-package, the predictions can be computed together with the model fit. Letting the data consist of the observations $\mathbf{y}$ and missing values for the new occurrences, the INLA-function uses the previous observations to predict posterior distributions for the new outcomes. In order to perform a leave-one-out cross validation, one can in each loop set the observed value in the test set to missing, and let the model predict its posterior distribution.

Odds ratio for the Bayesian approach with binary outcomes

When using the Bayesian approach for fitting a model with repeated binary outcomes, one could also be interested in the odds ratio for the variables in the model. In the frequentist approach, the subject specific odds ratio is calculated from the estimated value for the coefficients. For the Bayesian approach, we consider posterior distributions of the coefficients, not just a single value. However, one can obtain the posterior distribution for the subject specific odds ratio by transformation. Then, one could find the median or mode value and the credible interval for the posterior distribution of the subject specific odds ratio.

Chapter 4

Results

In this chapter, we present the results when fitting models for both the CP-data and the FM-data. Since FMs has been used as a surrogate measure of the CP-status before one is able to diagnose CP, we start with the results for prediction of FMs. As prediction of CP is the main goal of this thesis, we will only consider the model with C_{sd} as covariate, since it has shown good results for prediction of FMs in a previous study (Adde et al., 2009).

First, we look at the results from the frequentist approach, using the glmer()-function from the lme4-package in R. Then, we show the results from a simulation study where we investigate if adding more repeated measurement give more stable results than for the original data. Finally, we look at the results from the Bayesian approach, using the INLA-package in R to fit the model.

Then we turn to the prediction of CP in Section 4.2. In this section, we first look at the results when fitting a model using the C_{sd} variable for prediction of CP. Then, we consider the Lasso method for variable selection among the GMT-variables, and first present the results when fitting the CP-data using the Lasso and then from a Lasso model including other available variables as well. Finally, we evaluate the uncertainty of the Lasso estimates, using the bootstrap and the multi sample-splitting method. The R-code used for the simulation study, the INLA-analysis, the bootstrapping and the multi sample-splitting are shown in Appendix C- F.

4.1 Prediction of fidgety movements

First, we consider the fidgety movements (FM) data. The FMs are categorized into a binary response on the form

$$\text{FM} = \begin{cases} 1 & \text{if the FMs are normal} \\ 0 & \text{if the FMs are abnormal.} \end{cases}$$

In this thesis, we have only considered the old Prechtl's classification approach for having normal or abnormal FMs, as only the group of absent FMs in Table 2.5 included many

cases of CP. As mentioned in Section 2.1.1, the old classification approach is given by

$$\text{FMs} = \begin{cases} 1 & \text{if the FMs are ++,+ or +-} \\ 0 & \text{if the FMs are - or exaggerated.} \end{cases}$$

Again, we remind the reader that we assume that the classifications in the dataset are the correct ones, even though they are based on human judgement. Previous studies (Adde et al., 2009) has shown that the C_{sd} variable is associated with having normal FMs. Here, we investigate this association on a larger dataset including infants from three different countries, when differences between countries are taken into account.

Statistical model for the fidgety movements data

As the response is binary, a logistic regression model can be applied to fit the data. However, among the 693 participating infants, 98 infants have two, three or four repeated measurements. For an infant with repeated measurements, both the FMs outcome and the C_{sd} value can vary, but one would expect that these values are more alike within each participant than they are between participants. Hence, one can not assume that all the observations are independent of each other, as in a logistic regression model. Using instead a mixed effects logistic regression model with random intercepts, the observations for participant j are independent, given the covariates and the random intercept.

When including the participants' countries in the model, one could argue that also the variable for country should be a random effect in the model, as the observations within each country is more alike than between countries. However, only three countries are represented in our data and adding country as a random effect could cause an uncertain estimation of the variance between countries.

Letting country be a factor variable in the model with dummy variables $Norway$, USA and $India$, the model with Norway as the reference takes the form

$$\begin{aligned} \text{FM}_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\ logit(\pi_{ij}) &= \beta_0 + \zeta_j + \beta_1 C_{sd,ij} + \beta_2 USA_{ij} + \beta_3 India_{ij} \\ &\quad + \beta_4 C_{sd,ij} USA_{ij} + \beta_5 C_{sd,ij} India_{ij} \\ \zeta_j &\sim \mathcal{N}(0, \psi^2), \end{aligned} \tag{4.1}$$

where ij denotes the observation $i = 1, 2, 3$ or 4 for participant j, and π_{ij} is the probability that observation i for participant j is classified as normal FMs. The dummy variables USA_{ij} and $India_{ij}$ are equal to zero if the observation is from a Norwegian infant, $USA_{ij} = 1, India_{ij} = 0$ if the observation is from an American infant and $USA_{ij} = 0, India_{ij} = 1$ if the observation is from an Indian infant. Adding interactions between C_{sd} and the countries allows the effect of C_{sd} on the occurrence of normal FMs to vary between countries.

Figure 4.1 gives an interpretation of the coefficients in the model. For this model, all observations have a common intercept, β_0, and a subject specific intercept, ζ_j. If the observation is from a Norwegian infant, the logit probability for having normal FMs will be $\beta_0 + \zeta_j$ when C_{sd} is zero and increase linearly by β_1 for increasing C_{sd} values. For

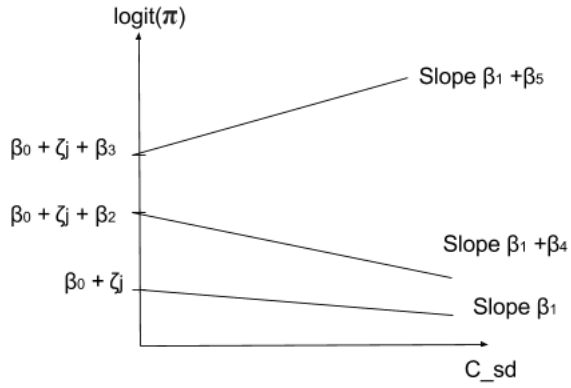


Figure 4.1: Logit probability for having normal FMs against C_{sd} with randomly chosen values for the coefficients.

American infants, β_2 is added to the intercept, and the logit probability for having normal FMs increase linearly by $\beta_1 + \beta_4$ for increasing C_{sd} values. Similarly for an Indian infant, the intercept is $\beta_0 + \zeta_j + \beta_3$ and the logit probability for having normal FMs increase linearly by $\beta_1 + \beta_5$ for increasing C_{sd} values.

To be certain that we don't have complete separation of our data, we look at the spread of the normal and abnormal FMs. Table 4.1 shows that even though there are few cases of abnormal FMs compared to normal FMs. However, all the countries have observations of both cases, and troubles due to complete separation of data should not be an issue here.

	Norway	USA	India	Total
FM= 1	205	217	273	695
FM= 0	45	42	16	103

Table 4.1: Frequency of normal and abnormal FMs for different countries.

4.1.1 Frequentist result for the fidgety movements data

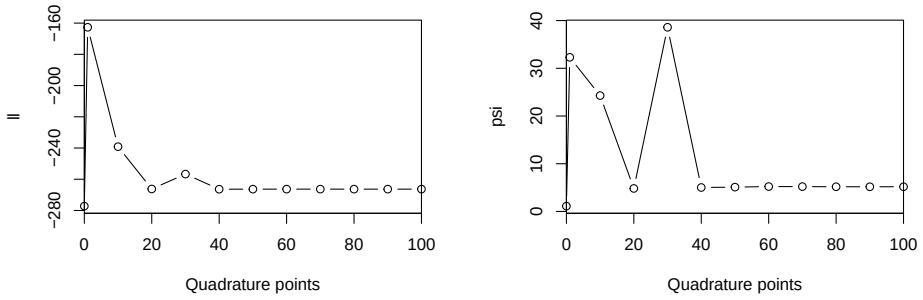
When fitting the logistic mixed effects model, the glmer()-function in R is used. The number of quadrature points used to approximate the integral in Equation (3.11) can be chosen manually in the specifications of the function. The default value is 1, which corresponds to the Laplace approximation. Values larger than 1 corresponds to the adaptive Gauss Hermite approximation, and larger values are supposed to produce greater accuracy in the approximation at the expense of speed. The value 0 corresponds to a simpler and faster, but less exact form for the parameter estimation

Running the glmer()-function with the model from Equation (4.1), the FM-dataset and number of quadrature points $= 0, 1, 10, 20, ..., 100$ gives the following warnings at 10 and 30 quadrature points:

- For 10 quadrature points: Warning messages:
  1: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
  unable to evaluate scaled gradient
  2: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
  Model failed to converge: degenerate Hessian with 1 negative eigenvalues

- For 30 quadrature points:Warning message:
  In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
  Model failed to converge with $\max|\text{grad}| = 0.0292726$ (tol $= 0.001$, component 1)

Hence, the method seems to not converge to stable estimates when using 10 and 30 quadrature points. Figure 4.2 shows the estimated log likelihoods and standard deviation for the random intercept, $\hat{\psi}$, when using glmer(), plotted against the chosen number of quadrature points. One can see that the estimated standard deviation for the random intercept, $\hat{\psi}$,



(a) Log likelihood **(b)** Standard deviation for the random intercept

Figure 4.2: Estimated log likelihood and standard deviation for the random intercept, ψ, for different number of quadrature points using the glmer()-function in R.

varies between 0 and 40 for different number of quadrature points. Especially for 1,10 and 30 quadrature points, the estimated value differs a lot from estimated values for different number of quadrature points. Using 40 quadrature points or more, the method seems to find stable estimated results as both the estimated log likelihood and $\hat{\psi}$ seems to have the same estimated value for increasing number of quadrature points.

To investigate if the random intercept is needed in the model, a LRT for the random intercept has been performed. The test statistic $-2\log\lambda$ was calculated to 28.5, corresponding to a p-value less than 0.001. Hence, the random intercept is statistically significant at a 5% significance level and should be included in the model.

Table 4.2 shows the hierarchical ANOVA-table for the variables in the FM-model. For the calculation of the ANOVA-table, a model without the interaction was fitted using 60 quadrature points, after doing the same analysis of quadrature points as for the full model (figure not shown). The same is done for the model without the country and interaction variables, and here the quadrature points analysis showed that 80 quadrature points were

needed (figure not shown). Table 4.2 shows that both the country variable and the interaction between C_{sd} and country are statistically significant at a 5% level, and should hence be included in the model. Hence, there are both differences between the logit probability of having normal FMs between countries, and there are differences in the effect of C_{sd} on the occurrence of normal FMs for the different countries.

Coefficient	Log likelihood	Resid. Df	χ^2	Df	p-value
C_{sd}	-283.8	795			
Country	-271.5	793	24.7	2	< 0.001
C_{sd}:Country	-266.3	791	10.3	2	0.006

Table 4.2: ANOVA-table with LRT for the mixed effects logistic regression model.

The results when running the glmer()-function with 50 quadrature points are shown in Table 4.3, together with the results when fitting the data with ordinary logistic regression, ignoring the within-subject correlation. Norway is used as the reference country. We see

Fixed effects					**Logistic regression**	
	Estimate	Std. Error	p-value	95% CI	Estimate	95% CI
Intercept	11.9	3.61	< 0.001	(4.86, 19.0)	6.01	(3.92, 8.11)
C_{sd}	-42.8	17.7	0.015	(-77.5, -8.22)	-26.7	(-38.5, -14.9)
USA	-4.79	3.81	0.209	(-12.3, 2.68)	-3.77	(-6.39, -1.15)
$India$	-10.3	4.78	0.031	(-19.7, -0.946)	-5.92	(-9.28, -2.56)
C_{sd}:USA	32.4	21.9	0.139	(-10.5, 75.4)	23.3	(8.58, 37.9)
C_{sd}:$India$	83.0	30.6	0.007	(22.9, 143)	42.6	(23.0, 62.1)
Random effects:						
	Variance	St.Dev.				
Intercept	25.9	5.09				

Table 4.3: Left: Estimated coefficients, standard error, p-values form the Z-test and confidence intervals when fitting the mixed logistic regression model to the data using the glmer()-function with 50 quadrature points. Right: Estimated coefficients and confidence intervals when fitting a logistic regression model to the data.

that the estimated coefficients for the fixed effects in the FM-model are similar to the estimates from the logistic regression model, but are more extreme. In addition, the estimates seem quite uncertain with very large confidence intervals, compared to the estimates and confidence intervals of the logistic regression model.

From Table 4.3, we see that there is no statistically significant difference between American and Norwegian infants for having normal FMs. However, there is a statistically significant difference between Indian and Norwegian infants. For Norwegian infants, the effect of C_{sd} on the occurrence of normal FMs is statistically significant at a 5% level, where increased C_{sd} values corresponds to decreased logit probability for having normal FMs. The effect of C_{sd} on the occurrence of normal FMs are not statistically significant different between American and Norwegian infants, but it is for Indian infants compared to Norwegian infants. To investigate the differences in the effect of C_{sd} on the occurrence of normal FMs for American and Indian infants, and to test the effects for each country

without comparing to Norwegian infants, we formulate three hypothesis,

$$\begin{aligned} &1) \quad H_0 : \beta_1 + \beta_4 = 0, \quad vs. \quad H_1 : \beta_1 + \beta_4 \neq 0 \\ &2) \quad H_0 : \beta_1 + \beta_5 = 0, \quad vs. \quad H_1 : \beta_1 + \beta_5 \neq 0 \\ &3) \quad H_0 : \beta_4 - \beta_5 = 0, \quad vs. \quad H_1 : \beta_4 - \beta_5 \neq 0. \end{aligned}$$

The first one test if there is an effect of C_{sd} for the occurrence of normal FMs for American infants. The second tests the same for Indian infants, while the third tests if there is a difference in effects of C_{sd} on the occurrence of normal FMs for American and Indian infants.

Letting L be the matrix on the form,

$$L = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

and letting $\hat{\boldsymbol{\beta}}$ be a vector of the estimated coefficients, $(\hat{\beta}_0, ..., \hat{\beta}_5)^T$, we have that

$$L\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 + \hat{\beta}_4 \\ \hat{\beta}_1 + \hat{\beta}_5 \\ \hat{\beta}_4 - \hat{\beta}_5 \end{pmatrix} \quad \text{and } Cov(L\hat{\boldsymbol{\beta}}) = LCov(\hat{\boldsymbol{\beta}})L^T$$

Plugging in our estimated values, we have that

$$L\hat{\boldsymbol{\beta}} = \begin{pmatrix} -10.4 \\ 40.1 \\ -50.6 \end{pmatrix} \quad Cov(L\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 192 & -32.2 & 225 \\ -32.2 & 522 & -554 \\ 225 & -554 & 779 \end{pmatrix}$$

This shows that the logit probability for having normal FMs decrease with increasing C_{sd} values for American infants, while it increase for Indian infants. In addition, we see that there is a quite large difference between the effect of C_{sd} on the occurrence of normal FMs between American and Indian infants.

Denoting $\boldsymbol{a} = L\hat{\boldsymbol{\beta}}$ and $\boldsymbol{b} = diag(Cov(L\hat{\boldsymbol{\beta}}))$, we can use the Z-test to test the three hypothesis formulated above,

$$Z_i = \frac{a_i}{\sqrt{b_i}} \quad \text{for } i = 1, 2, 3.$$

Using the estimated values, we find that all three hypothesis are rejected at a 5% significance level. Hence, the effect of C_{sd} on the occurrence of normal FMs is not statistically significant for American infants or for the Indian infants. In addition, the third test shows that even if there is a large difference between the effect of C_{sd} on the occurrence of normal FMs between American and Indian infants, the effect is not statistically significant at a 5% significance level.

Transformation of these values into subject specific odds ratios are shown in Table 4.4 for a 0.1 increase of C_{sd}.

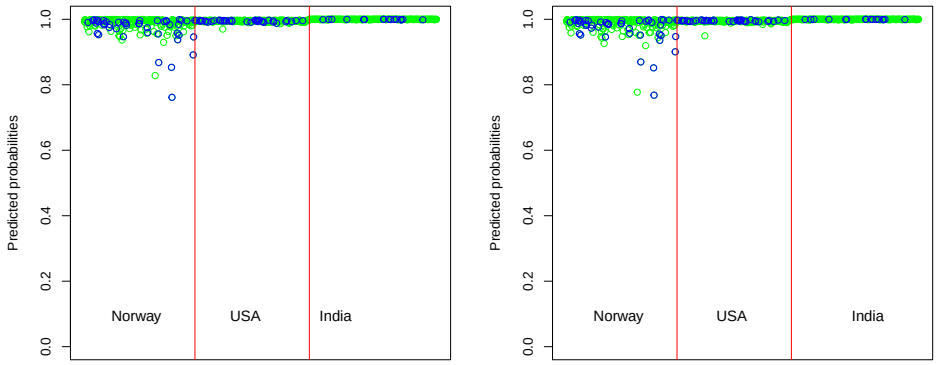
	OR	95%CI
Norway	0.014	(0.0004, 0.440)
USA	0.353	(0.023, 5.36)
India	55.4	(0.629, 4881)

Table 4.4: Subject specific odds ratios and confidence intervals for the effect of C_{sd} on the occurrence of normal FMs for a 0.1 increase in C_{sd} for all three countries.

Validation of the model

As mentioned in Section 3.4.1, in this thesis, we only consider the population averaged probabilities. However, the predict function in the lme4-package does not integrate out the random intercept before it returns the predicted probabilities, it only sets the random intercepts equal to zero. In lack of better methods for calculating the correct population average probabilities, we have chosen to present the ones returned from the predict function in the lme4-package, but we will be careful when interpreting them.

The predicted probabilities for having normal FMs are shown in Figure 4.3 for both the internal and the external validation of the model. A leave-one-out cross validation has been performed for the external validation. The green points represents infants with normal FMs while the blue points represents infants with abnormal FMs. The figure shows that there are small differences between the internal and external validation. The predicted probabilities for having normal FMs, $\hat{\pi}_{ij}$, are in general quite high, and there are no clear separation between the values of $\hat{\pi}_{ij}$ for the recordings of infants with normal FMs and the recordings of the ones with abnormal FMs. Only for Norwegian infants, there are some recordings that have values of $\hat{\pi}_{ij}$ less than 0.9, and it seems that most of these are in fact the ones recording infants with abnormal FMs.



(a) **(b)**

Figure 4.3: Predicted probabilities for having normal FMs for recordings of infants with normal FMs (green) and recordings of infants with abnormal FMs (blue) from the internal (a) and the external (b) validation of the mixed effects logistic regression model with random intercepts.

The ROC-curves and the AUC-values for both the external and internal validation are shown in Figure 4.4. The internal AUC-value is 0.701, while the external AUC-values is 0.666. The external AUC-value corresponds to a strength of discrimination which is below acceptable, according to Lydersen (2012). Calculation of the Brier score for the external

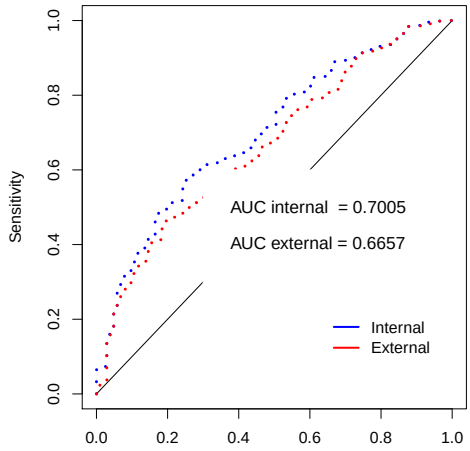


Figure 4.4: ROC-curves and AUC-values for the internal and external validation of the mixed effects logistic regression model with random intercepts.

validated data for this model, gives a value of 0.125. Again, we remind the reader that these values are calculations not from the population average probabilities, but from the probabilities where the random intercepts have been to zero for all measurements.

When performing sensitivity and specificity, one must determine a cutoff for which the predicted probabilities for having normal FMs lager than this value corresponds to having normal FMs. Normally, 0.5 is a reasonable value for the cutoff, but for this model most of the predicted probabilities for having normal FMs are almost equal to one. Hence, deciding a reasonable cutoff for this model is difficult and the values for sensitivity and specificity could be misleading. Therefore, we will not consider the sensitivity and specificity for this model or for any of the other models presented in this chapter.

4.1.2 Simulation study for fidgety movements data

In Figure 4.2, we saw that there were unstable results for the estimated coefficients, using different number of quadrature points in the glmer()-function. In addition, Table 4.3 showed that there large confidence intervals for the estimated coefficients. To investigate if increasing the number of repeated measurements could lead to more stable and certain estimates, we perform a simulation study.

The theory from Section 3.4.1 tells us that the normal approximation of the posterior density is only valid for large cluster sizes, i.e. many observations within each participant. In our data, most of the participants have only been observed once. Only 93 participants have two observations and only five participants have three or four observations. Hence,

the approximation of a normal distribution might not be valid in for our data. In addition, estimating a variance for the random intercepts might be difficult when there are many infants with only one repeated measurement. To investigate the uncertainty of few repeated measurements, the glmer()-function has been run for four different datasets, where the number of repeated observations varies. At the most, there are four repeated observations, which might not correspond to large cluster size, but we will investigate if the results gets better when increasing the cluster size.

In this simulation study we consider four cases of different datasets used to fit the mixed effect logistic regression model from Equation (4.1). All case datasets are based on the FM-dataset, where C_{sd} and the countries are the explanatory variables, and FMs is the binary outcome. The datasets differ by the number of repeated observations for the participants. When the case data includes more repeated measurements than in the original dataset, new C_{sd} values have been supplemented by creating new values for these participants. How these values has been created are described below the specifications of the cases.

- In case 1, all the participants have either one or two recordings. The third and fourth recording from the original data have been removed. Hence, 595 participants have a cluster size $n_j = 1$ and 98 participants have a cluster size $n_j = 2$.

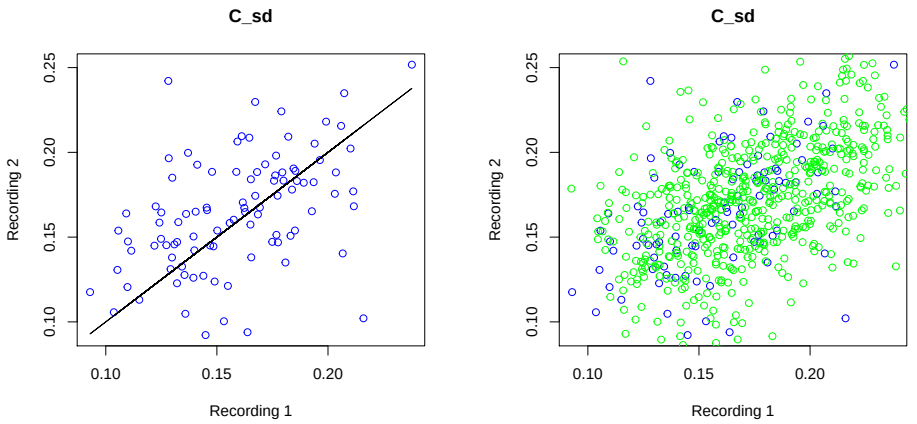
- In case 2, all the participants have two recordings. This dataset is similar to the case 1 dataset, but a second recording has been created for those with only one recording in case 1. Hence, all 693 participants have cluster size $n_j = 2$.

- In case 3, all the participants have three recordings. The dataset is similar to the case 2 dataset, but the third recording for participants with three or four recordings have been included from the original dataset. All other third recordings are created. Here, all 693 participants have cluster size $n_j = 3$.

- In case 4, all the participants have four recordings. The dataset is similar to the case 3 dataset, but the fourth recording for participants with four recordings have been included from the original dataset. All other fourth recordings are created. Here, all 693 participants have cluster size $n_j = 4$.

In case 2 there were drawn 595 values of C_{sd}, such that every participant has two values of C_{sd}. Figure 4.5a shows the association between the first and second recording from the original dataset. The figure indicate that there is a linear association between the two repeated values of C_{sd} for a participant. Due to this relationship, a second value of C_{sd} was created based on the linear regression model on the form $C_{sd2} = \beta_0 + \beta_1 C_{sd1} + \epsilon$, where ϵ is normally distributed with mean zero and variance σ^2_{csd}. The parameters in the model were estimated based on the 98 participants with two observed values of C_{sd}. Then, for the participants with only one recording, the second observation of C_{sd} was created by adding noise to the predicted values, $\hat{C}_{sd2,i} = \hat{\beta}_0 + \hat{\beta}_1 C_{sd1,i} + \hat{e}_i$, where $\hat{e}_i$ are drawn for a normal distribution with mean 0 and variance $\hat{\sigma}^2_{csd}$.

The dataset in case 3 was based on the data from case 2. In addition, the observed third values of C_{sd} and 688 created values of C_{sd} were added to the data. The created values were obtained through a linear model of the first and second repeated measurements, where



(a) **(b)**

Figure 4.5: Figure (a) shows the association between values of C_{sd} from the first and second recording for 98 participants with two observed values of C_{sd}. Figure (b) shows the predicted second values (green) plotted against the true first vales, together with the original values for the 98 participants (blue).

the second repeated measurements were used as covariates to predict the third 688 ones. The dataset in case 4 was obtained in the same way, adding the two observed forth values of C_{sd} and creating 691 new observations. Also here, the linear model for the first and second repeated measurements were used, and the third observations were used as covariates for prediction of the 691 fourth values of C_{sd}.

For each case, 1000 iterations were performed. In each iteration, the following algorithm was run.

1. Find π_{jj} from Equation (4.1), using $\mathbf{x_{ij}}$ from the case dataset, the estimated regression coefficients, $\hat{\boldsymbol{\beta}}$, from the logistic regression model (Table 4.3) and the random intercept ζ_j drawn from $\mathcal{N}(0, \psi^2 = 5^2)$.
2. Draw n values of $y_{ij} \sim$ Bernoulli(π_{ij}) where n is the number of observations in the case.
3. Estimate the regression coefficients in the model (4.1) using glmer() with 0,1,10,20,30,40,50,60,80 and 100 quadrature points.

As one would expect that the estimated coefficients in the mixed effects model are not that different from the coefficients of the ordinary logistic regression model, the regression coefficients $\hat{\boldsymbol{\beta}}$'s in step 1 are set to be equal to the estimated coefficients when fitting the original data with a logistic regression model. These values are shown in Table 4.3. Note

that these regression coefficients are used in all four cases. The choice of the standard deviation for the random intercept, $\psi = 5$, is based on Figure 4.2, where the estimated values for ψ are approximately 5 for each of the models using more than 30 quadrature points.

The plots of the 1000 estimated coefficients, log likelihoods, number of warnings and number of errors for different number of quadrature points when using glmer() to estimate the FM-model, are shown for each case in Appendix A. In all the cases, there are on average simulated 66% normal FMs, and 34% abnormal FMs (step 2). Comparing the 10 000 (1000 simulations $\times$ 10 different values for the number of quadrature points) simulated log likelihoods to the log likelihood of the models without the random intercept, the null hypothesis that the random intercept was not needed in the model was rejected by the LRT for all the 10 000 simulations, in all four cases.

Figure 4.6 shows the number of warnings and errors for each number of quadrature points for all cases. It turns out that only in case 1, there are errors when estimating the coefficients. The number of warnings are highest for case 1, for all number of quadrature points, and are in fact highest using 10 quadrature points. There are some warnings in all the cases, but most of them happen when the number of quadrature points used is equal to 1, which is the Laplace approximation. The figure shows that when using the Laplace approximation, the number of warnings are reduced with increasing number of repeated measurements. However, using more than 1 quadrature points, the number of warnings reduces considerably. An interesting point for case 1, is that for increasing number of quadrature points, the number of warnings reduces, but the number of errors increases. In fact, when using 100 quadrature points in case 1, there were four errors in the 1000 simulations.

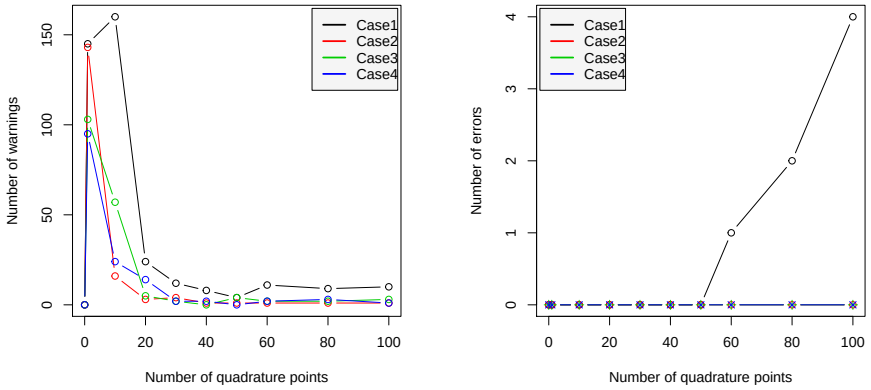


Figure 4.6: The number of warnings and errors in the 1000 simulations plotted against the number of quadrature points for all cases.

In general for all the estimated variables in case 1, the median and the 25th and 75th percentile seems to stay at the same values for more than 20 quadrature points. However, there are many outliers from the whiskers, where many of them take extreme values com-

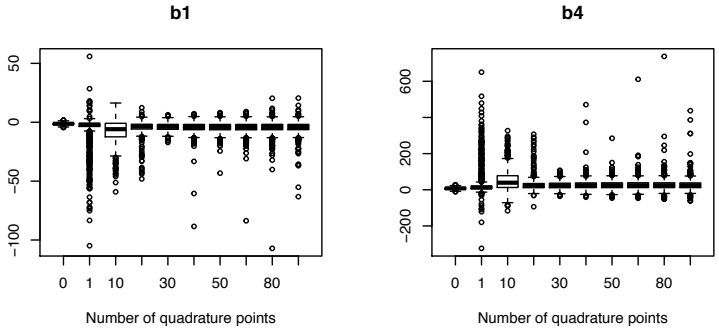


Figure 4.7: Estimated coefficients from case 1 for all the chosen number of quadrature points for the C_{sd} variable and the variable for the interaction between C_{sd} and USA.

pared to the median. Figure 4.7 shows the values for the estimated coefficients for C_{sd} and the estimated coefficients for the interaction between C_{sd} and USA. The values for the estimated coefficient for C_{sd}, $\hat{\beta}_1$, vary between -100 and 50. The values for the estimated coefficient for the interaction between C_{sd} and USA, $\hat{\beta}_4$, varies between -300 and 700. In addition, the estimated standard deviation for the random intercept, ψ, have large outliers. Figure 4.8 shows the estimated values for ψ for all the cases. In case 1, the value of the outliers for $\hat{\psi}$ seems to increase for increasing number of quadrature points. The largest value is 120 while the median value is 5, when considering at the estimated values for more than 20 quadrature points.

For all other cases, the estimated values for all the variables seems stable for more than 20 quadrature points, and the plots are similar to the one for the standard deviation of the random intercept in Figure 4.8 Even though these cases consist of more data, there is a considerably difference in the outliers values for the first case and case 2-4. Only when using 1 quadrature point (the Laplace approximation) there are large variations and many outliers. However, for case 2, 3 and 4, the estimated values in each iteration are almost identical for more than 20 quadrature points with much less extreme outliers than for case 1.

Table 4.5 shows that the median values and 25th and 75th percentiles are quite equal for all cases when using 50 quadrature points. The intervals between the 25th and the 75th percentile are large, which means that there are large variations in the estimated values. In addition, the table shows that the size of the intervals are reduced with increased number of repeated measurements.

As the median values seem to be quit stable for more than 20 quadrature points, most of the estimated values can be trusted for all coefficients. However, due to the extreme outliers in case 1, when having only one or two repeated observations, one could be unfortunate and get an estimated value that is very different from the ”true” value. When increasing the cluster sizes such that all participants have two repeated observations, the outliers disappear for more than 20 quadrature point, and these results are less uncertain. Hence, using more than 20 quadrature points to estimate the model gives trustable results

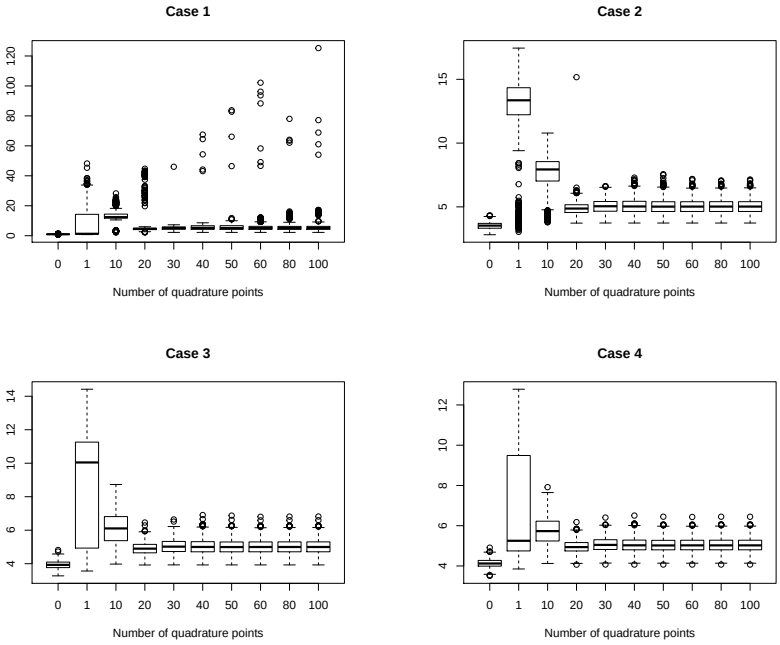


Figure 4.8: Estimated values for the standard deviation of the random intercept, ψ, from the simulations shown for all the four cases.

in most cases, but for small cluster sizes one should be aware that the estimated values can be far off the "true" value.

4.1.3 Bayesian approach for fitting the fidgety movements data

Now, we consider a Bayesian approach for fitting the FM-model in Equation (4.1), using the INLA-package in R. The latent field is given by $\boldsymbol{\theta} = \{\boldsymbol{\eta}, \boldsymbol{\beta}, f(\boldsymbol{\zeta})\}$, where $\boldsymbol{\eta}$ is a vector of the linear predictors, $\boldsymbol{\beta}$ is a vector of the fixed regression coefficients and $f(\boldsymbol{\zeta}) = \boldsymbol{\zeta}$ is the function for the random intercept. We assign Gaussian priors to all the coefficients in the latent field and assume that they are conditionally independent. In a hierarchical form, our model with $j = 1, ..., 693$ participants and $i = 1, .., n_j$ repeated measurements, $n_j \in \{1, 2, 3, 4\}$, takes the form

$$\begin{aligned}
\text{Level 1: } \mathbf{y}|\boldsymbol{\theta}, \phi &\sim \prod_{j=1}^{N} \Big(\prod_{i=1}^{n_j} \frac{[\exp(\beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \zeta_j)]^{y_{ij}}}{1 + \exp(\beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \zeta_j)} \Big), \\
\text{Level 2: } \boldsymbol{\theta}|\phi &\sim \mathcal{N}(\mathbf{0}, Q(\psi)), \\
\text{Level 3: } \phi &\sim p(\phi),
\end{aligned}$$

where $\boldsymbol{\theta}$ is the latent Gaussian field, $\mathbf{Q}(\phi)$ is the precision matrix for the latent field, and ϕ is the hyperparameter for the precision of the random intercept, $\phi = 1/\psi^2$. Hence,

	Case 1	Case 2	Case 3	Case 4
Log Likelihood	-475	-708	-939	-1146
	(-483, -468)	(-720, -693)	(-958, -918)	(-1170, -1123)
ψ	5.1	5.0	5.0	5.0
	(4.2, 6.7)	(4.6, 5.4)	(4.7, 5.3)	(4.8, 5.3)
Intercept	6.2	6.0	6.0	6.0
	(4.6, 8.4)	(4.5, 7.6)	(5.0, 7.1)	(5.2, 7.0)
C_{sd}	-4.1	-3.8	-3.8	-3.8
	(-6.4,-1.8)	(-5.7, -1.9)	(-5.0, -2.6)	(-4.9, -2.8)
USA	-6.2	-5.8	-6.0	-5.9
	(-8.7, -3.9)	(-7.8, -4.1)	(-7.3, -4.7)	(-7.0, -4.8)
$India$	-28	-27	-27	-27
	(-39, -18)	(-36, -17)	(-33, -21)	(-32, -22)
C_{sd}:USA	25	23	23	23
	(12, 39)	(12, 35)	(16, 30)	(17, 30)
C_{sd}:$India$	45	42	43	42
	(32,60)	(32, 54)	(35, 51)	(36, 49)

Table 4.5: Median and 25th and 75th percentiles for estimated coefficient values for all cases when using 50 quadrature points.

we need to specify the distribution of the hyperparameter and the precision matrix for the latent field. For the fixed effects $\boldsymbol{\beta}$, we assign a non-informative prior for the intercept and vague priors for the fixed effects, with mean and precision as follows:

- Intercept: $\mathcal{N}(0, 0)$
- $\beta_i : \mathcal{N}(0, 0.1)$ for $i = 2, ..., 6$.

For the specification of the hyperparameter, we use three different priors which has been proposed in the paper by Grilli et al. (2015). These are vague Gamma priors;

- a popular with distribution Ga(0.001,0.001),
- Fongs prior with distribution Ga(0.5, 0.0164), and
- the default prior for the INLA-package with distribution Ga(1, 0.0005).

Here, the first parameter is the shape and the second is the rate for the Gamma probability density function. The first prior is a popular choice in Bayesian analysis (Grilli et al., 2015) and is the default in the BUGS-software. For simplicity, we call this the "Popular" prior. The second, which we have chosen to call "Fongs" prior, is a prior proposed by Fong et al. (2010). The last one is the default prior of INLA, and we refer to it as "Default". Figure 4.9 shows the density of these three priors. One can see that the default prior is very flat, the popular prior is also flat, but with a peak close to zero, and that the prior proposed by Fong et al. (2010) is a weakly informative prior, placing more weight near zero.

Using this model, both criteria for using INLA are met; i) The latent field is a GMRF and ii) the number of hyperparameters is small, in fact, there is just one hyperparameter.

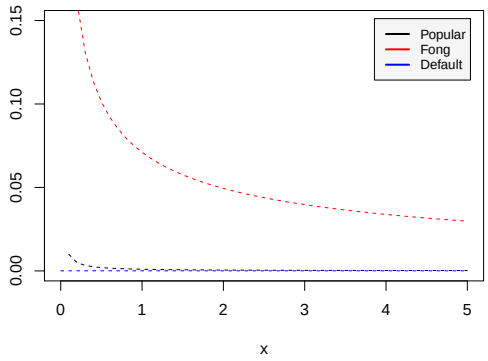


Figure 4.9: Density of the three different priors for the precision of the random intercept.

The output after having run INLA consists of marginal distributions for the parameters and hyperparameters. Using this, one could calculate summary variables such as the posterior mean, median and standard deviations, and credible intervals.

Figure 4.10 shows the density of the posterior marginals for the regression coefficients for the fixed effects when using default priors for the fixed effects and the three different priors on the precision of the random effect. The figure shows that there are small differences in the posterior marginals for the fixed effects when applying different priors for the precision of the random intercept. The mode of the posterior marginal distributions seems to be equal for different priors of the precision of the random intercept, but the distribution seems to be most narrow for the default, a bit wider for the prior proposed by Fong et al. (2010) and the widest for the popular prior.

Looking at the posterior distribution for the precision of the random intercept in Figure 4.11, one can see that the choice of prior has a greater effect on the posterior marginals for the precision of the random intercept than for the fixed effects. The popular prior and the prior proposed by Fong et al. (2010) seems to give quite similar posterior distributions for the precision of the random effect, but the popular prior gives a steeper posterior distribution. Both have their posterior mode close to 0.5. The default prior gives a posterior distribution which is close to zero everywhere, with a small peak near zero, but the mode is in fact near 0.5. One can see that the posterior when using the popular prior and the prior proposed by Fong et al. (2010) differs much from the prior distributions, while the posterior when using the default prior is close to the prior distribution. Hence, the posterior is highly affected by the prior, which tells us that there is little information about this parameter in the data. We see that the popular prior and Fongs prior, which are vague priors, gives quite similar posteriors that differs from the priors. For the default prior, which is a flat prior, the posterior is almost similar to the prior.

Based on the similar mode values and the similar posteriors for the fixed effects, it seems that the choice of prior for the precision of the random intercept doesn't matter that much. Hence, we can choose any of them without too much effect on other parts of the

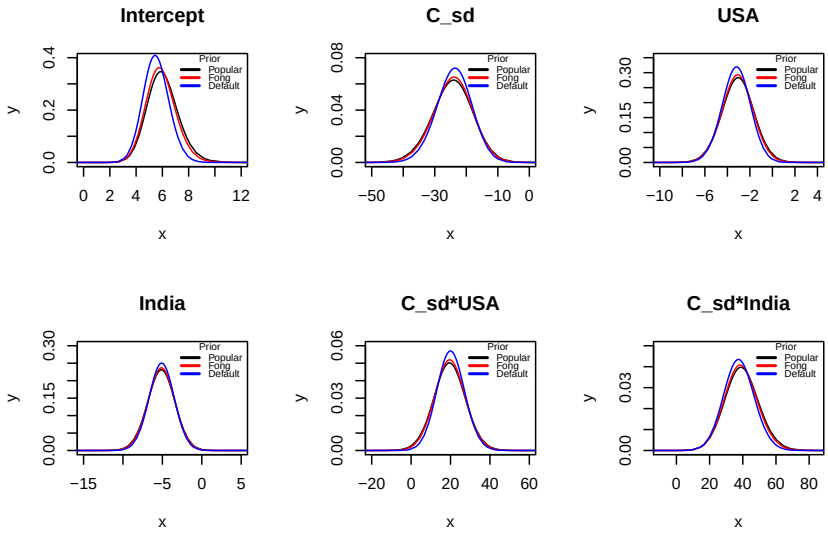
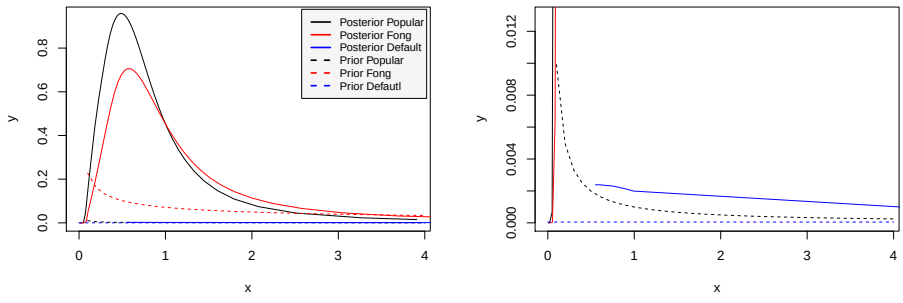


Figure 4.10: Density for the posterior marginals for the fixed effects. The fixed effects have default priors while the precision of the random intercept have three different priors; Fongs prior, a popular choice and the default.



(a) **(b)**

Figure 4.11: Posterior marginal distributions for the precision of the random effect plotted together with their prior distribution for all three priors. Figure (b) is a zoomed version with smaller values of the y-axis from figure (a).

model, so we use the popular prior Ga(0.001, 0.001) as the prior for the precision of the random intercept in the following analysis.

Next, we look at the posterior marginals for the fixed effects plotted against their priors. Figure 4.12 shows that for most of the posterior marginals, the likelihood updates the prior information, such that the posterior marginals differs from the prior distribution. However, for the interaction variables, C_{sd}·USA and C_{sd}·India, the shape of the posteriors are very similar to the shape of the priors, but we see that the likelihoods have updated the mode for the posterior distributions. Hence, there are much information in the likelihood of the data for most of the variables, but for the interaction variables, the prior information seems to affect the posteriors the most.

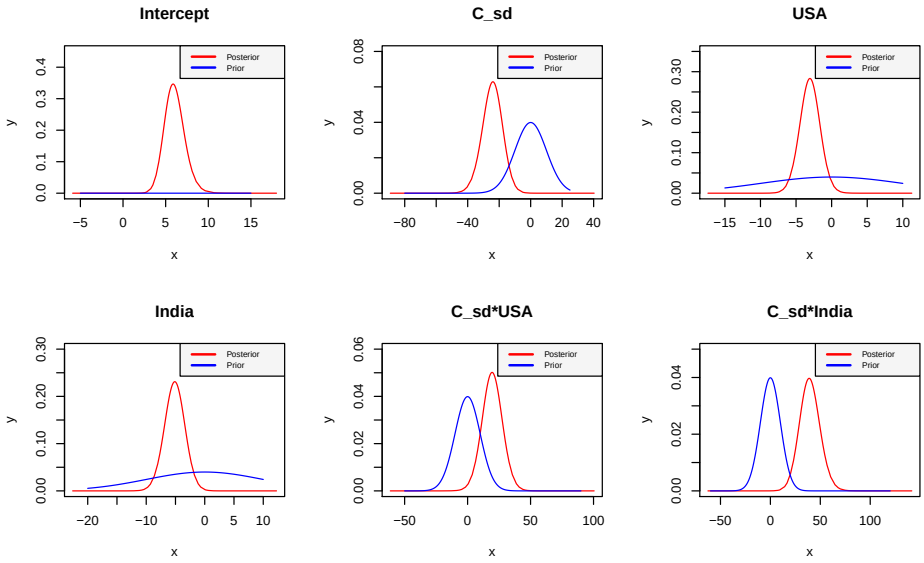


Figure 4.12: Posterior marginals for the fixed effects plotted together with their prior distributions.

The results from the Bayesian model are summarized in Table 4.6. Comparing this table to Table 4.3, showing the results from the frequentist approach using the glmer()-function, we see that the estimated coefficients for this model are quite different. The values here are of the similar magnitude, but they are less extreme than for the ones in Table 4.3. The credible intervals for this model are narrower than the confidence intervals in the model where the glmer()-function has been used, which corresponds to more certain estimates. In fact, the estimated values for the fixed effects and credible intervals are more equal to the ones from the logistic regression model than to the estimates from the glmer()-model. To get an indication of the differences in values for the precision of the random intercept in both analysis, one of the values must be inverted. Inverting the mode from the INLA-analysis, we find that the variance is $1/0.49 = 2.04$, which is quite different from the variance for the random intercept in the glmer()-model which is 25.9.

Transforming the posterior mode values into subject specific odds ratios are not that

Fixed effects			
	Mode	Mean	95% CredInt
Intercept	5.87	6.05	(3.88, 8.58)
C_{sd}	-24.0	-24.4	(-37.6, -12.07)
USA	-3.05	-3.06	(-5.88, -0.242)
$India$	-5.12	-5.15	(-8.61, -1.76)
C_{sd}:USA	19.5	19.6	(3.68, 35.5)
C_{sd}:$India$	39.5	38.8	(20.2, 60.4)
Random effects:			
	Mode	Mean	95% CredInt
Precision for Intercept	0.49	10.5	(0.179, 65.2)

Table 4.6: Estimated mode, mean and 95% credible intervals for the posterior marginals when using INLA to fit the model with the popular prior for the precision of the random intercept and default priors for the fixed effect.

straight forward as in the frequentist analysis. The effects of countries can not be added as easily here, since the values here are not estimates, but summary measures from the posterior marginal distributions. To find the subject specific odds ratio for the C_{sd} variable for American and Indian infants, the model has to be run again with these countries as references, and we look at the value for the mode of the exponential-transformed C_{sd} posterior. Doing this, the subject specific odds ratios for the occurrence of normal FMs by a 0.1 increase in C_{sd} are shown in Table 4.7 together with their credible intervals.

	Norway	USA	India
Mode C_{sd}	-24.0	-4.29	12.0
OR	0.09	0.65	3.3
95% CredInt	(0.023, 0.295)	(0.236, 1.70)	(0.820, 15.8)

Table 4.7: Mode, subject specific odds ratios of the mode and credible intervals for a 0.1 increase in C_{sd} on the occurrence of normal FMs from the Bayesian model.

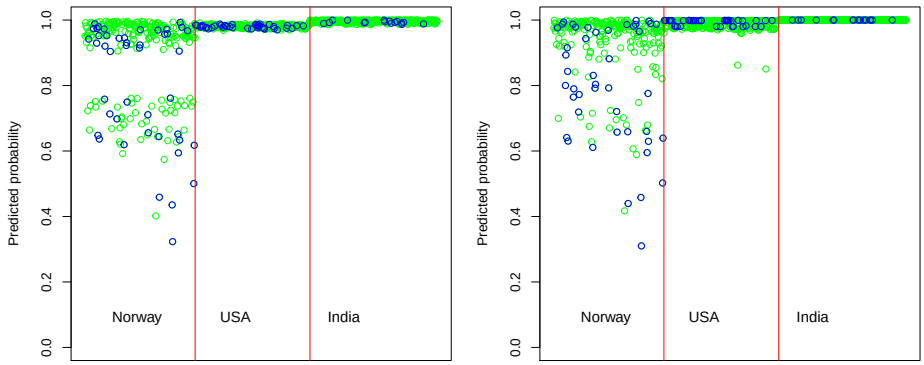
The table shows that also for the Bayesian approach, the effects of C_{sd} on having normal FMs differs between countries. Again, for Norwegian and American infants, the subject specific odds ratio are less than 1, and increased C_{sd} values gives reduced odds for having normal FMs, given that the random intercept is constant. For Indian infants, increased C_{sd} values gives an increase in the odds for having normal FMs, given that the random intercept is constant. Even though these values seems more reasonable than the ones from the glmer()-analysis, the fact that Indian infants have increased odds or normal FMs for increased C_{sd} values, while Norwegian and American infants have decreased odds, does not sound reasonable in a clinical setting. The credible intervals here are also quite large, even though they are not as large as in the frequentist analysis. Again, we have that the credible intervals for the American and Indian infants covers values of the subject specific odds ratio which corresponds to both increased and decreased odds for having normal FMs for increased C_{sd} values, given constant random intercepts. Hence, for these data, there are no significant effect of C_{sd} on the occurrence of normal FMs for American

and Indian infants.

Validation of the INLA-model

Now, we will see how well this model performs at predicting normal FMs. An internal validation and a leave-one-out cross validation, both at the population level, have been performed. Figure 4.13 shows the predicted posterior mode of the population averaged probabilities, $\hat{\pi}_{ij}$ for having normal FMs from both the internal and the external validation. There seems to be small differences between the internal and the external validation, as these figures seems quite equal. We see that for the Norwegian infants there are a spread of the values of $\hat{\pi}_{ij}$, but there are no clear separation between those that are classified with normal FMs from those who are classified with abnormal FMs. For the American and Indian infants, there are very small differences between the ones classified with normal FMs and the ones classified with abnormal FMs, and almost all the $\hat{\pi}_{ij}$ have values close to one, regardless of their true outcome.

Comparing these plots to the ones from the frequentist approach, we see these values are in general higher than the $\hat{\pi}_{ij}$ values from the frequentist approach, and have more spread in the values for the Norwegian infants. However, non of the methods seem to make any clear separation between those classified with normal FMs from those classified with abnormal FMs.



(a) **(b)**

Figure 4.13: Estimated population averaged probabilities for having normal FMs for those classified with normal FMs (green) and those classified with abnormal FMs (blue) for the internal (a) and external (b) validation using the posterior mode for each predicted population averaged probability.

Figure 4.14 shows the ROC-curve with corresponding AUC-values for the internal and external validation of the posterior mode of the predicted population averaged probabilities for having normal FMs. The curves are quite similar, but the internal validated curve are mostly higher than the external, and the AUC-values are thereafter with values 0.698 and 0.630 for the internal and external validation, respectively. Comparing these values to the

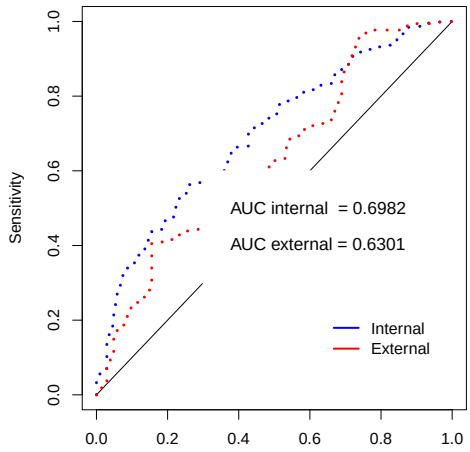


Figure 4.14: ROC-curve for internal and external validation with INLA model, using the posterior mode for each predicted population averaged probability for having normal FMs.

ones from the frequentist approach, we find that these values are smaller and are in fact below the acceptable range of discrimination, according to Lydersen (2012).

The Brier score for the Bayesian model is also calculated based on the posterior mode values for the predicted population averaged probabilities for having CP from the external validation. It has a value of 0.113, which is a decrease compare to the frequentist model and corresponds to a better model fit than for the frequentist model. Again, we remind the reader that the predicted probabilities for having normal FMs for frequentist approach were not population averaged, but calculated by setting the random intercepts equal to zero. Hence, we trust the results from the validation of the Bayesian approach more then the ones from the frequentist approach, when it comes to assessing the population averaged predicted probabilities.

4.2 Prediction of cerebral palsy

Now, we turn to the prediction of cerebral palsy (CP). First, we consider a model similar to the one with FMs as response, where we see how well the C_{sd} variable is at predicting cases of CP. Since there are several GMT-variables and other available variables, we search for a better models for prediction of CP. We start by using the Lasso method for all the GMT-variables, and see how well the selected model performs at predicting CP. Then, we also include some available variables in the Lasso analysis, and see if adding these variables will increase the model fit. Finally, we evaluate the estimated coefficient values for the selected model including both GMT- and the available variables, in terms of bootstrapping and multi sample-splitting for calculations of p-values and confidence intervals.

Since the CP-diagnose is only made once, there are not different outcomes correspond-

ing to each repeated measurement of the GMT, as for the FM-data. Due of this, we must select one of the GMT-recordings per infant for the CP-analysis. To avoid problems with selection bias, the chosen recordings were selected at random.

However, the first dataset that were selected at random included one of the outliers for the A_{sd}, (see Section 2.2). It turned out that this one outlier had some effect in the Lasso-analysis. Figure 4.15 shows the amount of times each of the variables were included in the model when running the bootstrap algorithm for the dataset including one outlier of A_{sd} and for a dataset not including any of the outliers. When including the outlier, the A_{sd} variable was selected in less than 40% of the 1000 bootstrap replicates. With no outlier, however, the variable was selected in more than 70% of the replicates. The remaining variables seem to stay at similar values, regardless of the outlier. Hence, it seems that the one outlier brings large uncertainty to the model, so to be on the safe side, we only consider the dataset where the recodings have been selected at random, but without any of the outliers of the A_{sd} variable.

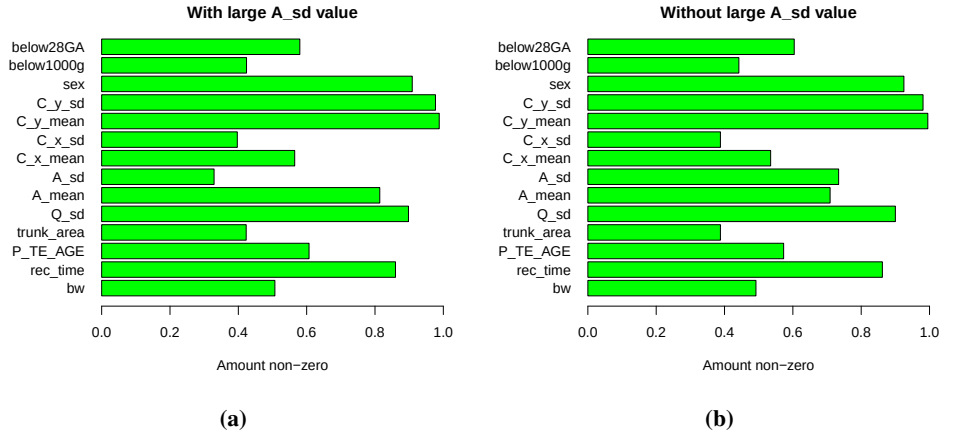


Figure 4.15: Bootstrap samples with (a) and without (b) one of the large outliers for the A_{sd} variable.

4.2.1 Prediction of cerebral palsy by the standard deviation of the centroid of motion

First, we look at the association between the C_{sd} variable and the CP-status. To investigate the predictive power of C_{sd} on the occurrence of CP, a logistic regression model is used to fit the data. To adjust for the infants' countries, both the country variable and the interaction of C_{sd} and country are included in the model. Country is a factor variable, with three categories; Norway, USA and India. Based on the same reasoning as for the FM-model, country will be treated as a fixed factor variable in the model. Our model takes the form

$$logit(\pi_i) = \beta_0 + \beta_1 C_{sd} + \beta_2 USA + \beta_4 India + \beta_5 C_{sd} USA + \beta_6 C_{sd} India,$$

where π_i is the probability of infant $i = 1, ..., 693$ having CP.

	Resid. Df	Resid. Dev	Df	χ^2	p-value
C_{sd}	691	338.2	1	0.2	0.672
Country	689	294.3	2	44	< 0.001
C_{sd}:Country	687	291.3	2	3.0	0.225

Table 4.8: Hierarchical ANOVA-table for the logistic regression model with CP as response and C_{sd}, Country and their interaction as covariates.

The ANOVA-analysis in Table 4.8 shows that, according to the LRT, the interaction between C_{sd} and country is not statistically significant at a 5% level, and it is hence omitted from the model. Then the new model takes the form

$$logit(\pi_i) = \beta_0 + \beta_1 C_{sd} + \beta_2 USA + \beta_3 India \quad \text{for i} = 1, ..., 693$$

For the new model without the interaction terms, the results for the fitted model are shown in Table 4.9 and Figure 4.16. We see that for all three countries, increased values

	Estimate	Std. Error	p-value	95% CI	OR	95% CI
Intercept	-3.08	0.774	< 0.001	(-4,60, -1.56)		
C_{sd}	9.09	4.49	0.043	(0.295, 17.9)	2.48	(1.03, 5.99)
USA	-1.13	0.341	< 0.001	(-1.80, -0.458)	0.324	(0.166, 0.633)
$India$	-3.18	0.633	< 0.001	(-4.42, -1.94)	0.042	(0.012, 0.144)

Table 4.9: Estimated coefficients with standard error, p-values from the Z-test, confidence intervals and odds ratios with confidence intervals for the logistic regression model with CP as response and C_{sd} and Country as covariates. The odds ratio for the C_{sd} is calculated with a 0.1 increase in C_{sd}.

of C_{sd} correspond to increased logit probability for having CP. An increase of 0.1 in C_{sd} gives an increase in odds of a factor 2.48 for having CP, and the odds for having CP is larger for Norwegian infants, compared to American and Indian infants.

Validation of the model

For an external validation, a leave-one-out cross validation was performed on the model with CP as response and C_{sd} and Country as covariates. The predicted probabilities for having CP, $\hat{\pi}_i$, from both the internal and external validation are shown in Figure 4.17, separated for each country. The values of $\hat{\pi}_i$ from the internal and external validation seems very similar, but looking closer one can see that the values of $\hat{\pi}_i$ from the internal validation is a bit higher than the ones form the external validation. In general, there are quite low values of $\hat{\pi}_i$. For Norwegian infants, most of the participants have values of $\hat{\pi}_i$ of about 0.1 and 0.2, and the model seems to not distinguish much between those that have CP and those that doesn't. The values of $\hat{\pi}_i$ are lower for American infants and even lower for Indian infants. Most of the values for $\hat{\pi}_i$ stay at the same level, and the model doesn't seem to catch the ones that do have CP.

Figure 4.18 shows the ROC-curves for both the internal and external validation for this model together with their AUC-values. As observed in Figure 4.17, the values of the $\hat{\pi}_i$'s are a bit higher for the internal validation than for the external validation. This is also

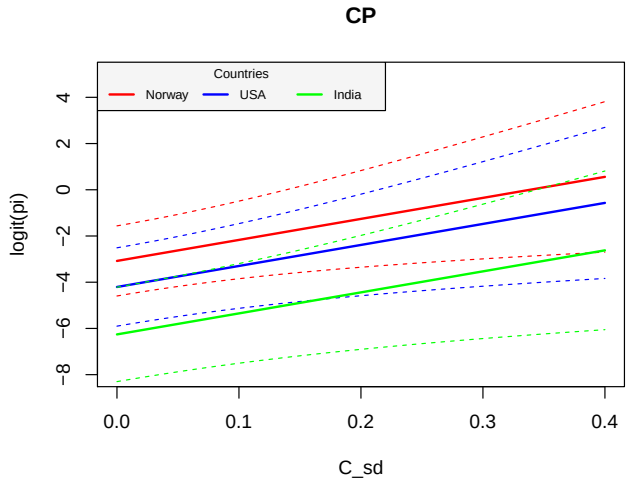


Figure 4.16: Estimated logit probability for having CP for different values of C_{sd} and different countries, with 95% confidence intervals.

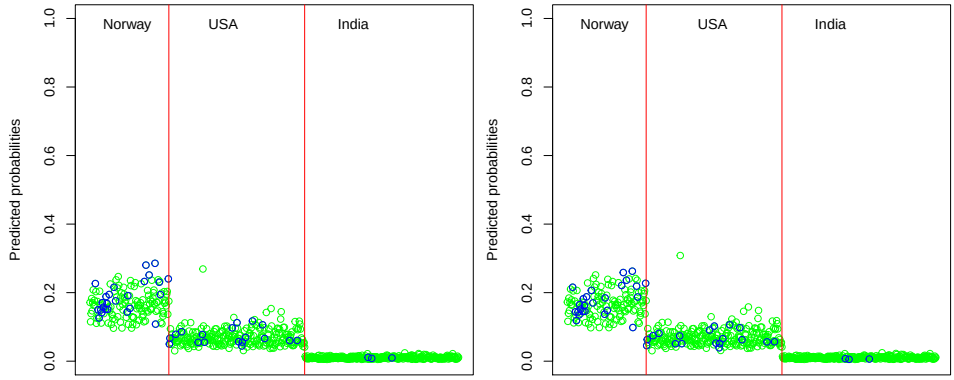


Figure 4.17: Predicted probabilities for having CP for the infants diagnosed with CP (blue) and the infants diagnosed without CP (green) from the internal (a) and external (b) validation.

reflected in the ROC-curves, where the internal ROC-curve is in general higher than the external ROC-curve. According Lydersen (2012), the external AUC-value corresponds to an acceptable strength of discrimination. The Brier score has been calculated for the predicted probabilities from the external validation. For this model, the Brier score is 0.058.

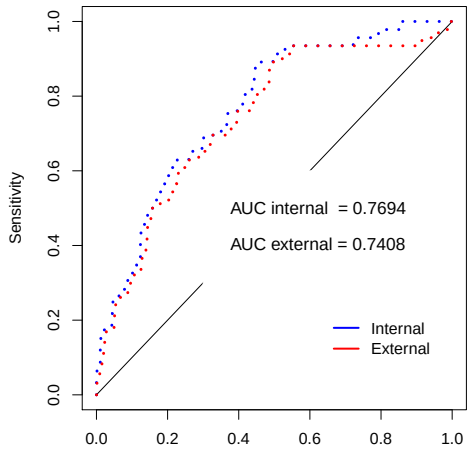


Figure 4.18: ROC-curves with AUC-values for the internal and external validation for the model with CP as response and C_{sd} and Country as covariates.

4.2.2 Variable selection for the cerebral palsy model

In our dataset, we have nine available summary variables from the GMT-software, describing the infants movements. In this section, we investigate if we can find a better model for predicting CP, when considering all nine GMT-variables and using the Lasso method for variable selection. In addition, other variables are available in the dataset, so we investigate if inclusion of some of these variables in the model can improve the fit.

Variable selection for prediction of cerebral palsy with General Movement Toolbox-variables

First, we consider only the GMT-variables in the Lasso analysis. Before we start the analysis, we remove the highly correlated variables. In this way, we can select which of the highly correlated variables that seem most natural to include in the analysis, instead of the Lasso method removing highly correlated variables based on only correlation values. In Section 2.2, we saw that the centroid of motion variables in x- and y-direction were, as expected, highly correlated with the Euclidian distance between them. However, the x- and y-variables were not highly correlated with each other, and there were no high correlation between the mean and standard deviation variables. Due to this, we only remove the Euclidian distance variables, C_{mean} and C_{sd}, from the analysis.

In addition, Q_{mean} was highly correlated with $Q_{sd}, H_{mean}, W_{mean}$ and A_{sd}. In fact, the width and height variables (W and H) are highly correlated with both each other and with the area variables (A). Because of this, we choose to also remove Q_{mean}, H_{mean}, W_{mean}, H_{sd} and W_{sd} from the analysis.

The GMT-variables included in the model were C_{xsd}, C_{ysd}, C_{xmean}, C_{ymean}, Q_{sd}, A_{mean} and A_{sd}. To account for differences between countries, we force the country variables to be a part of the model, i.e. their coefficients can't be shrunken to zero. The

Lasso-analysis is done in R, using the glmnet-package (Friedman et al., 2010).

Figure 4.19 shows the Lasso coefficient path for different values of the tuning parameter, $\log(\lambda)$. The variables that are shrunken to zero last are C_{ysd} and C_{ymean}, while C_{xsd} and C_{xmean} are removed from the model for small values of the tuning parameter.

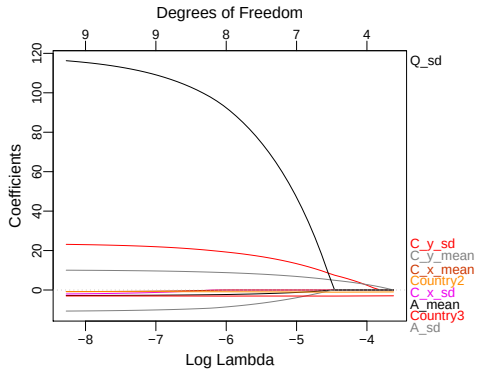


Figure 4.19: The Lasso coefficient path and number of non-zero parameters for different values of the tuning parameter $\log(\lambda)$ for the model with GMT-variables. Note that Country2 =USA and Country3 = India

A k-fold cross validation has been performed to find the value of the tuning parameter that gives the best model. Using 20 folds in the cross validation, we secure that there are some observations of CP in the training sets. Figure 4.20 shows the cross validation curve with upper and lower standard deviation along the λ-sequence for different decision rules for the cross validation. The best model, corresponding to the tuning parameter λ_{min} for deviance and misclassification decision rules and the tuning parameter λ_{max} for the AUC decision rule, includes seven variables. For all decision rules, the model which is within one standard error of the best model, corresponding to the model with the tuning parameter λ_{1se}, include non or only one of the GMT-variables.

Looking at the estimated coefficient values in Table 4.10, we see that the seven variables included in the best model, regardless of the decision rule, is C_{ysd}, C_{ymean}, Q_{sd}, A_{mean}, A_{sd} and the country variables. For the models with tuning parameter λ_{1se}, we see that only the AUC decision rule includes more than the country variables. For this decision rule, also C_{ymean} is included in the model.

Validation of the Lasso model

Next, we will see how well the Lasso model performs at predicting CP. The predicted probabilities for having CP from both the internal and external validation are shown in Figure 4.21. For the internal validation, the estimated Lasso coefficients with deviance and λ_{min} as decision rule from Table 4.10 have been used to estimate the probabilities for

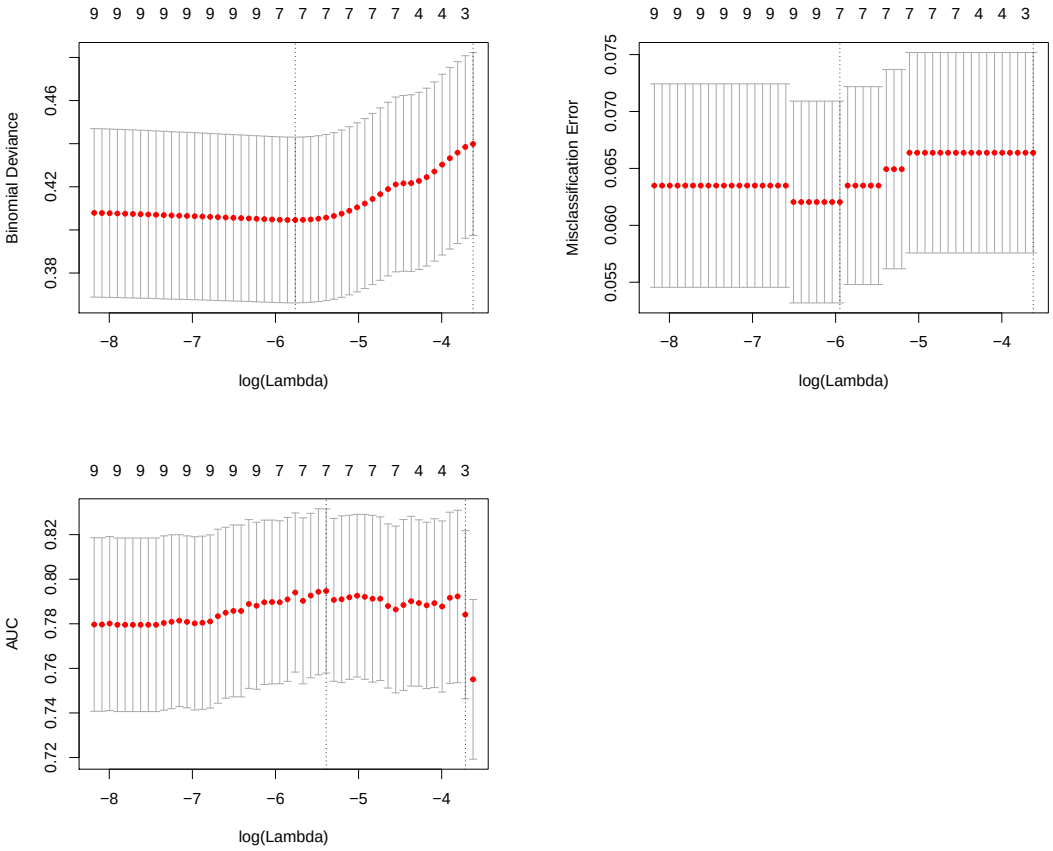


Figure 4.20: Cross validation curve with upper and lower standard deviation with decision rules based on deviance, misclassification error and area under the ROC curve. The two vertical lines displays the values of $\lambda_{min}/\lambda_{max}$ in terms of the decision rule, and the value of λ_{1se}, while the numbers on top displays the number of variables included in the model.

having CP. The model takes the form

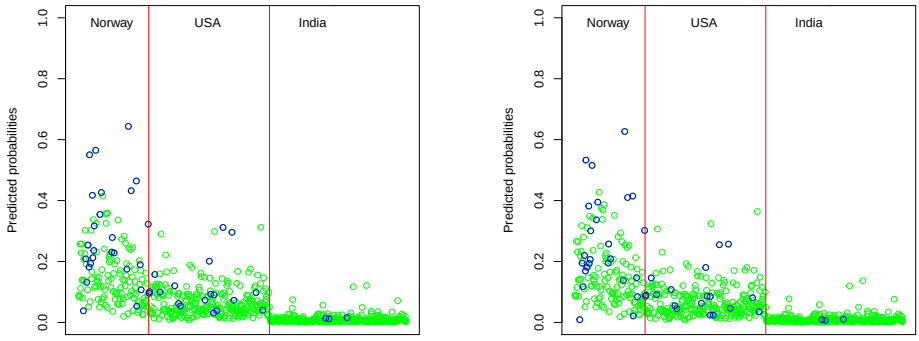
$$logit(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_1 C_{ysd} + \hat{\beta}_2 C_{ymean} + \hat{\beta}_3 Q_{sd} + \hat{\beta}_4 A_{mean} + \hat{\beta}_5 A_{sd} + \hat{\beta}_6 USA + \hat{\beta}_7 India,$$

where $\hat{\pi}_i$ is the predicted probability of infant $i = 1, ..., 693$ having CP. For the external validation, a leave-one-out cross validation has been performed, where a Lasso model was fitted for each of the training sets and used to predict the probability for having CP for the one observation in the test sets.

Also for this model, the predicted probabilities for having CP, $\hat{\pi}_i$, are very similar for the internal and external validation. Comparing these figures to Figure 4.17, we see that including other GMT-variables than the C_{sd}, the model predicts higher probabilities for having CP with more spread. For the Norwegian infants, the highest values of $\hat{\pi}_i$

	λ_{min}			λ_{max}		
	dev	class	auc	dev	class	auc
$\log \lambda$	-5.76	-5.95	-5.39	-3.62	-3.62	-3.72
(Intercept)	-8.41	-8.63	-7.84	-1.60	-1.60	-2.04
C_{xsd}	0	0	0	0	0	0
C_{ysd}	18.3	19.1	16.3	0	0	0
C_{xmean}	0	0	0	0	0	0
C_{ymean}	8.48	8.74	7.82	0	0	0.80
Q_{sd}	85.0	90.9	69.6	0	0	0
A_{mean}	-2.13	-2.26	-1.78	0	0	0
A_{sd}	-7.79	-8.43	-6.18	0	0	0
USA	-0.92	-0.91	-0.96	-0.98	-0.98	-1.00
India	-3.07	-3.07	-3.07	-2.95	-2.95	-2.97

Table 4.10: Estimated coefficients for the models with tuning parameters $\lambda_{min}/\lambda_{max}$ and λ_{1se} from the cross validation of the Lasso-analysis for all three decision rules.



(a) **(b)**

Figure 4.21: Predicted probabilities for having CP for infants diagnosed with CP (blue) and infants diagnosed without CP (green) from the internal (a) and external (b) validation of the Lasso model with GMT-variables.

corresponds to the ones that were diagnosed with CP, but there are also many many that were diagnosed with CP with low values of $\hat{\pi}_i$. For the American and Indian infants, there are no clear separation between the values of $\hat{\pi}_i$ for those diagnose with and without CP.

Figure 4.22 shows the ROC-curve with corresponding AUC-values for both the internal and external validation. With this model, the external AUC-value is 0.794, which corresponds to an acceptable strength of discrimination, according to Lydersen (2012). Using the external validation of this model, the Brier score is 0.054, which is a bit lower than the Brier score for the model with only one GMT-variable. Hence, a model with selected GMT-variables performs better at distinguishing between infants with and without CP, than the model with only one GMT-variable, C_{sd}, as covariate.

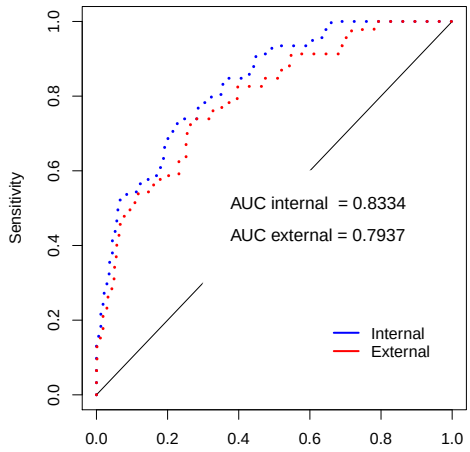


Figure 4.22: ROC-curves and AUC-values for the internal and external validation of the Lasso model with several GMT-variables.

Variable selection for prediction of cerebral palsy with General Movement Toolbox-variables and other available variables

Next, we include some of the other available variables for the 693 infants. These are the birth weight (bw), the gestational age (GA), the length of the video recording (rec_time), the post term age (P_TE_AGE), the trunk area (trunk_area) and the gender of the infants. In addition, there are three dummy-variables for extreme low birth weight and extreme preterm infants. These are listed below.

- Below1000g; 1 if infant has birth weight below 1000g, 0 otherwise.
- Below28GA; 1 if infant was born before 28 weeks' of gestation. 0 otherwise.
- Below1000g28GA; 1 if infant has both birth weight below 1000g and/or born before 28 weeks' gestation, 0 otherwise.

Even though some of these variables are quantities from the GMT-software and some are background variables for the infants, we refer to these variables as the *clinical* variables for simplicity.

Figure 4.23 shows the correlation between the clinical variables. There are high correlation between the birth weight, the gestational age and the three corresponding dummy variables. It is natural that the birth weight and the gestational age are correlated, as early born infants are often of smaller size and hence lower weight than those who are born closer to the term. The dummy variables are categorized based on the birth weight and the gestational age, so these are hence highly correlated. It seems that among these variables, birth weight, below1000g and below28GA are the least correlated variables, with 0.67 at the most. Hence, we choose to drop gestational age and below1000g28GA from the analysis. Figure 4.24 shows the correlation plot of all variables now to be included in

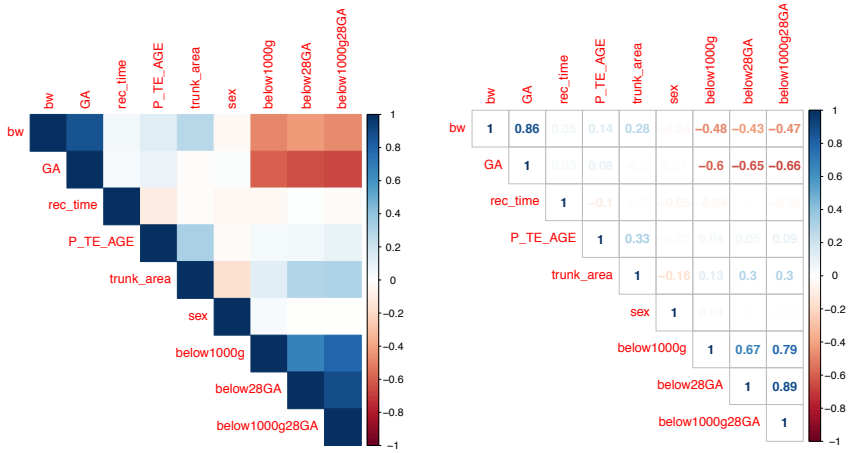


Figure 4.23: Pairwise correlation of the clinical variables.

the Lasso regression model. We see that non of these variables are highly correlated with each other, so we continue our analysis with these variables.

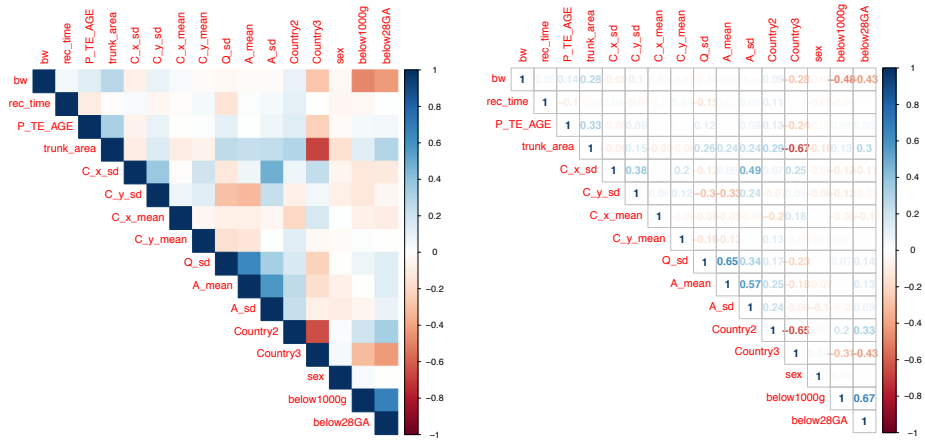


Figure 4.24: Pairwise correlation of the remaining variables that are included in the Lasso analysis.

The Lasso coefficient path when including the clinical variables are shown in Figure 4.25 for a sequence of the log tuning parameter $\log(\lambda)$. Again, C_{ysd} and C_{ymean} are the last variables that are shrunken to zero, and the figure seems quit similar to the one where the clinical variables were not included, Figure 4.19. Among the clinical variables, the gender seems to be most important, as it is the last clinical variable that is shrunken to zero.

Again, we use a 20-fold cross validation to choose which value of the tuning parameter that gives the best model fit. Figure 4.26 shows the cross validated deviance, misclassifi-

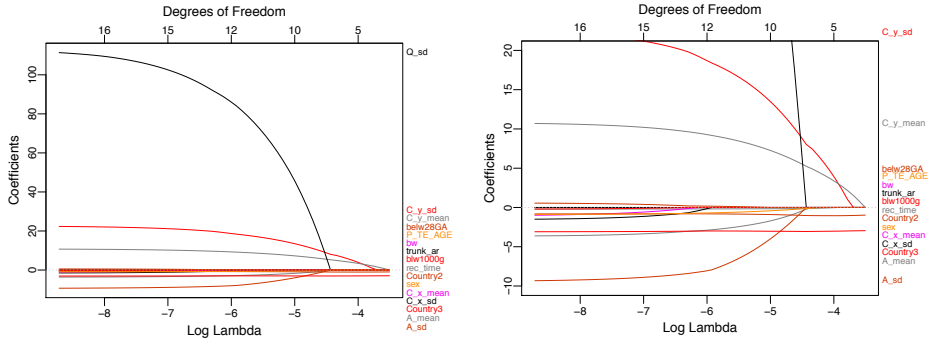


Figure 4.25: The coefficient path in full scale (a) and zoomed in on the y-axis (b) for the Lasso estimates when including GMT-variables and clinical variables in the analysis plotted for a sequence of the log of the tuning parameter.

cation error and AUC-values for the data including both GMT-variables and clinical variables. Again, the three decision rules seem to agree that the model with approximately 10

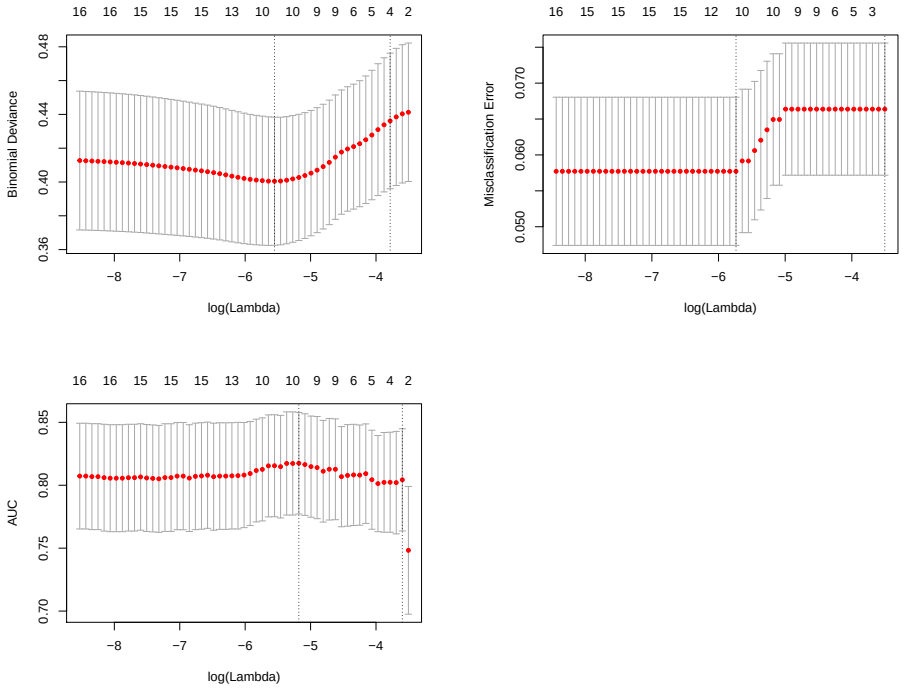


Figure 4.26: Cross validated deviance, misclassification error and AUC-values with standard deviations for the Lasso model including GMT- and clinical variables.

variables gives the best fit. The models that are within one standard error of the best fit include few variables, as for the models with only GMT-variables. Table 4.11 shows the estimated Lasso coefficients with each decision rule for the model with the tuning parameter for the best fit, $\lambda_{min}/\lambda_{max}$ and the model with the tuning parameter corresponding to one standard error within the best fit, λ_{1se}.

	λ_{min}		λ_{max}	λ_{1se}		
	dev	class	auc	dev	class	auc
$\log\lambda$	-5.55	-5.74	-5.18	-3.78	-3.50	-3.60
(Intercept)	-7.32	-7.56	-6.72	-2.93	-1.60	-2.04
bw	0	0	0	0	0	0
rec_time	-0.17	-0.18	-0.16	0	0	0
P_TE_AGE	0	0	0	0	0	0
trunk_area	0	0	0	0	0	0
C_{xsd}	0	0	0	0	0	0
C_{ysd}	16.9	17.7	14.8	1.01	0	0
C_{xmean}	0	0	0	0	0	0
C_{ymean}	8.56	8.89	7.74	2.16	0	0.80
Q_{sd}	72.3	78.7	55.8	0	0	0
A_{mean}	-2.50	-2.69	-2.02	0	0	0
A_{sd}	-6.70	-7.37	-5.00	0	0	0
USA	-0.84	-0.82	-0.88	-1.03	-0.98	-1.00
India	-3.00	-2.99	-3.02	-2.99	-2.95	-2.97
sex	-0.60	-0.65	-0.49	0	0	0
below1000g	0	0	0	0	0	0
below28GA	0.11	0.14	0.04	0	0	0

Table 4.11: The estimated Lasso coefficients for the models with $\lambda_{min}/\lambda_{max}$ and λ_{1se} from the cross validated Lasso-model with both GMT- and clinical variables for different decision rules.

The table shows that among the clinical variables, the length of the recording, gender and the indication variable for extreme preterm infants are included in the best models, for all three decision rules. The centroid of motion variables in x-direction are also shrunken to zero in these models, while all of the other GMT-variables are included in the best models. For the models with tuning parameter λ_{1se}, only C_{ymean} and C_{ysd} are included in the model, in addition to the intercept and the country variables. Comparing these models to the ones without the clinical variables, we see that the same GMT-variables are included in the best models, with quite similar estimated coefficient values. However, comparing Figure 4.26 and Figure 4.20, we see that the AUC-values are generally higher when including the clinical variables. Hence, including clinical variables in the analysis gives higher AUC-value and a better fit to the data.

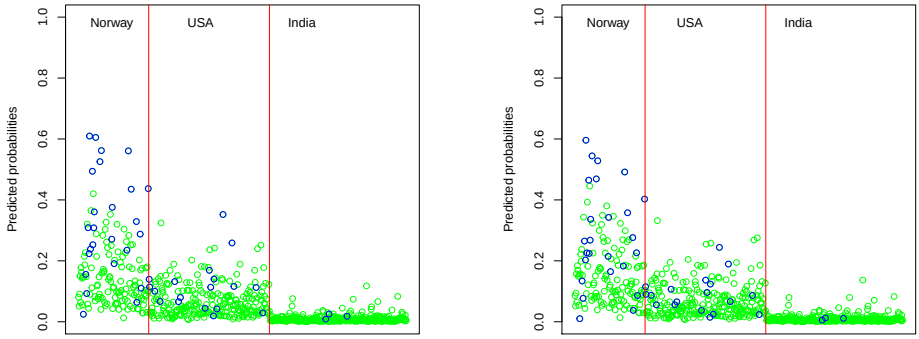
Validation of the Lasso model

Next, we will see how well the Lasso model with clinical variables performs at predicting CP. Figure 4.27 shows the predicted probabilities for having CP, $\hat{\pi}_i$, from the internal and

the external validation. Again, the estimated Lasso coefficients in Table 4.11 has been used to estimate the probabilities for having CP in the internal validation, where the model takes the form

$$logit(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_1 C_{ysd} + \hat{\beta}_2 C_{ymean} + \hat{\beta}_3 Q_{sd} + \hat{\beta}_4 A_{mean} + \hat{\beta}_5 A_{sd} + \hat{\beta}_6 USA + \hat{\beta}_7 India$$
$$+ \hat{\beta}_8 rec\_time + \hat{\beta}_9 sex + \hat{\beta}_{10} below28GA,$$

where $\hat{\pi}_i$ is the predicted probability of infant $i = 1, ..., 963$ having CP. For the external validation we used the leave-one-out cross validation, where a Lasso model was fitted for the training sets, and used to predict the probability for having CP in the test sets. These figures are quite similar to the corresponding figures with only the GMT-variables included in the model, shown in Figure 4.21. Again, the largest values of $\hat{\pi}_i$ corresponds to the ones diagnosed with CP for the Norwegian infants, but there are still many that are diagnosed with CP that have very small values of $\hat{\pi}_i$. For American and Indian infants, there are still no clear separation between the values of $\hat{\pi}_i$ for those that are diagnosed with CP and those that are not.



(a) **(b)**

Figure 4.27: Predicted probabilities for having CP for infants diagnosed with CP (blue) and infants diagnosed without CP (green) from the internal (a) and external (b) validation of the Lasso model with the GMT-variables and the clinical variables.

Comparing the AUC-values from the ROC-curves for the model with clinical variables in Figure 4.28 to the ones from the model without clinical variables shown in Figure 4.22, one can see that inclusion of the clinical variables has increased the external AUC-value a bit, from 0.794 to 0.797. The Brier score for the external validation of the model including clinical variables is 0.053, which is a bit lower than the score for the model without clinical variables. Hence, inclusion of clinical variables improve the model fit a bit.

Inference for the Lasso model

To evaluate the uncertainty of the Lasso estimates, a bootstrap analysis and the method of multi sample-splitting were run. In the bootstrap analysis, there were drawn 1000 boot-

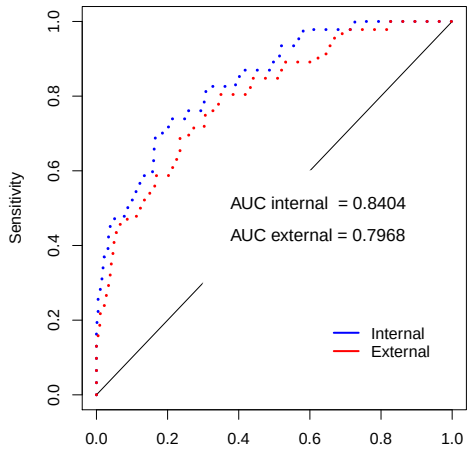


Figure 4.28: ROC-curves and AUC-values for the internal and external validation of the Lasso model with several GMT-variables and clinical variables.

strap samples, and the 20-fold cross validations was performed for each sample. The estimated coefficients of the best model, corresponding to the model with the tuning parameter $\lambda_{min}/\lambda_{max}$, and the estimated coefficients from the model with tuning parameter λ_{1se} were saved for each bootstrap sample. Here, we only focus on the binomial deviance as decision rule, and the models with the tuning parameter λ_{min}, but refer to Appendix B for the same analysis for the other decision rules. Figure 4.29 shows the proportion of the 1000 bootstrap replicates where each of the variables are estimated to a nonzero value. Since the country variables are forced to be included in the model, these are not shown in the figure.

The figure shows that the GMT-variables C_{ymean}, C_{ysd} are nonzero in almost all the 1000 bootstrap replicates. In addition, the Q_{sd} variable is nonzero in about 85%, while the A_{mean} and A_{sd} variables are nonzero in about 70% of the bootstrap replicates. However, the C_{xmean} and C_{xsd} variables are only included in about 50% and 40% of the bootstrap replicates. Among the clinical variables, the gender and the length of the video recording are nonzero in more than 80%, while the remaining clinical variables are included in about half the replicates or less.

Histograms of the estimated Lasso coefficient values from the bootstrap replicates are shown in Figure 4.30. Also here, the coefficients for the country variables are excluded from the figures. From these figures, it is clear that Q_{sd}, C_{ymean} and C_{ysd} are significantly different from zero. The A_{mean} and A_{sd} variables are mostly different from zero. Even though these variables have the interquartile range that borders at zero, the median values are far from zero. For the C_{xmean} and C_{xsd} variables, the median value is very close to zero.

Since the 25th and 75th percentiles for estimated coefficients for Q_{sd}, C_{ymean} and C_{sd} are above zero, the figure indicates that increased values of these variables corresponds to increased logit probability for having CP. The estimated coefficients for the area of

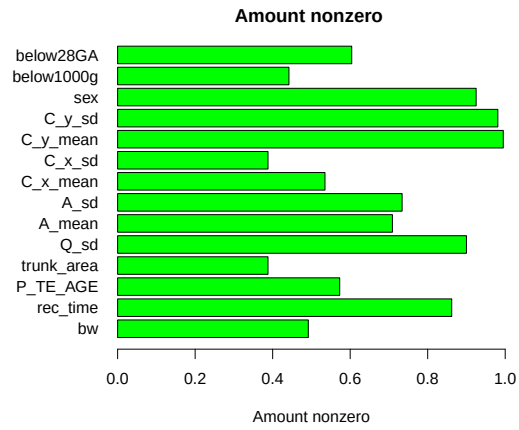


Figure 4.29: Proportion of the 1000 bootstrap replicates where the coefficients from the 20-fold cross validation with binomial decision rule are estimated to be non-zero.

motion variables however, are mostly below zero, so increased values of these variables corresponds to decreased logit probability for having CP.

For the clinical variables, we see that the variables for the length of the video recording, the gender and extreme preterm infants are mostly different from zero. It seems that the estimated coefficients for the length of the video recording and the gender are mostly negative. The estimated coefficients for the indication variable for extreme preterms are mostly positive. This indicates that longer video recordings, being female and being born after 28 weeks' gestation corresponds to a decrease in the logit probability of having CP. The remaining variables have estimated coefficient values both above and below zero, so there is no clear association between any of these variables and having CP.

Table 4.12 shows the results from the multi sample-splitting method for the logistic regression model using a leave-one-out cross validation in the variable selection. Since the method performs a cross validation on only half the dataset, we use a leave-one-out cross validation to include as much data as possible in the training sets.

The adjusted p-values show that only the variable for India is significant at a 5% significance level. The variables with confidence intervals ranging from $-\infty$ to ∞ shows that the variables have been selected less than 50% of the times, so their confidence intervals are not defined. The area of motion variables that have large range between the 25th and the 75th percentiles in the bootstrap analysis does not have defined confidence intervals in the multi sample-splitting. The other GMT-variables, Q_{sd}, C_{ymean}, C_{ysd} and sex and rec_time have defined confidence intervals in the multi sample-splitting. However, most of these confidence intervals range over zero, so the corresponding variables are not significant. The only two variables that doesn't have a confidence interval ranging over zero are the variables for India and C_{ymean}. When adjusting for multiple testing, the adjusted p-values shows that the variable for C_{ymean} is not statistically significant at a 5% significance level, while the variable for India is.

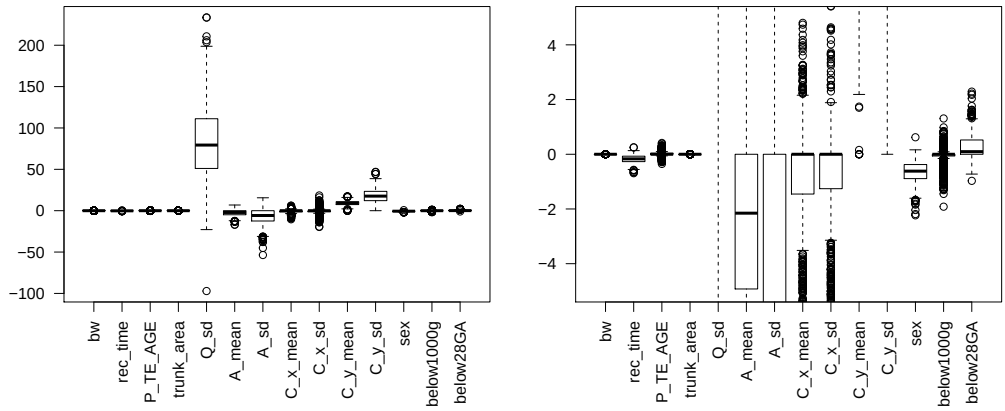


Figure 4.30: The estimated Lasso coefficients in full scale (b) and zoomed in y-axis (a) from the bootstrap replicates from 20-fold cross validation using minimum binomial deviance as decision rule.

	p-value	CI
bw	1.00	$(-\infty, \infty)$
rec_time	1.00	(-0.79, 0.37)
P_TE_AGE	1.00	$(-\infty, \infty)$
trunk_area	1.00	$(-\infty, \infty)$
Q_{sd}	1.00	(-98.2, 263)
A_{mean}	1.00	$(-\infty, \infty)$
A_{sd}	1.00	$(-\infty, \infty)$
C_{xmean}	1.00	$(-\infty, \infty)$
C_{xsd}	1.00	$(-\infty, \infty)$
C_{ymean}	0.32	(1.45, 18.4)
C_{ysd}	1.00	(-4.30, 43.4)
USA	0.73	(-2.27, 0.20)
India	0.03	(-5.71, -1.25)
sex	1.00	(-2.21, 0.71)
below1000g	1.00	$(-\infty, \infty)$
below28GA	1.00	$(-\infty, \infty)$

Table 4.12: P-values adjusted for multiple testing and a 97.5% confidence intervals for the estimated coefficients from the multi sample-splitting with 1000 iterations.

Chapter 5

Discussion

In this thesis, we have used data from the GMT-software for 693 infants with a total of 798 recordings, from three different countries, to find models that predict infants with normal FMs and infants with CP. Fitting a logistic regression model with C_{sd} as covariate and CP as response, we found that there was a significant effect of C_{sd} on the occurrence of CP, where increased values of C_{sd} corresponded to increased logit probability of having CP. For this model, there were no differences in the effect of C_{sd} on the occurrence of CP between countries. Allowing other GMT-variables and other available variables to be included in the model and using the Lasso method for variable selection, the model fit increased and several variables were selected. These selected GMT-variables were C_{ysd}, C_{ymean}, Q_{sd}, A_{mean} and A_{sd}. The other variables were gender, the indication variable for extreme preterm (below 28 weeks' gestation) and the length of the video recordings. When using the FMs as response, we used both a frequentist method and a Bayesian method for fitting a mixed effects logistic regression model with random intercepts, with C_{sd} as covariate. Both approaches found that there was a difference in the effect of C_{sd} on the occurrence of normal FMs between countries, but the effect was only significant for Norwegian infants, where increased values of C_{sd} corresponded to decreased logit probability of having normal FMs.

The results from the logistic regression model with C_{sd} as covariate concur with the results from Adde et al. (2010). However, when allowing other GMT-variables to be included in the model, the model fit increased. The Lasso model showed that it was in fact the y-direction of the centroid of motion variables that had the most effect on having CP, and not the Euclidean distance between the x- and y-variables. Hence, from this model, it seems that more movements in the upper part of the body and uncoordinated movements between upper an lower body parts, leading to large variations of the centroid of motion in the y-direction, corresponds to increased logit probability of having CP. This sounds reasonable, as these movements could be interpreted as non-smooth movements, which are characteristic for abnormal FMs and hence CP (Hopkins and Prechtl, 1984).

The selected Lasso model also showed that increased variation in the quantity of motion corresponds to increased odds of having CP. Large variations of the quantity of motion

can also be interpreted as non-fluent and non-smooth movements. However, larger mean and standard deviation values of the area of motion corresponds to a decrease in the odds of having CP. Hence, infants with fluent and smooth movements on a small area and with small variations of the area of the movements, have lower probability of having CP, which sounds reasonable. In addition, extreme preterm infants are of higher risk of developing CP (Stephens and Vohr, 2009), so it seems logical that this variable is selected in the model. Inclusion of the gender in the Lasso model might be because of the prevalence of CP among the boys and girls in the dataset, as we know little about the risk of having CP between boys and girls. The length of the video recordings are likely more associated with the uncertainty of the GMT-variables, than the probability of having CP.

The result that the logit probability of having normal FMs decreased with increasing C_{sd} values for Norwegian infants concur with the results from Adde et al. (2009). However, there is no reason to expect that the effect of C_{sd} on the occurrence of normal FMs should vary much between countries, so this result might be a consequence of the few cases of abnormal FMs in the dataset.

In a medical setting, this is a large dataset, consisting of 798 recordings of infants at high risk of developing CP. This is a strength of the study, as there are many infants from different continents with both CP and FMs-status. However, there is a limitation with the number of cases of abnormal FMs and CP in the dataset. When considering FMs as response, there are only 103 out of 798 of the recordings of the infants that corresponds to abnormal FMs. For the Indian infants, there are only 16 out of 289 of the recordings corresponding to abnormal FMs. Because of this, there might not be enough data of infants with abnormal FMs to find statistically significant results for the FM-models, especially for the Indian infants. When considering the CP-data, we reduced the dataset such that all 693 infants had only one recording. With CP as outcome, only 46 out of 693 infants are diagnosed with CP, corresponding to less than 7%. Hence, also for the CP-data, finding good models with statistically significant results could be difficult. This is especially difficult for the Indian data, where only 3 out of 298 of the infants are diagnosed with CP, corresponding to less than 1%. In addition, when modeling the FM-data with repeated measurements, the fact that only 595 infants have more than one recording is a challenge.

In order to investigate the uncertainty and stability when modeling a mixed effects logistic regression model with random intercepts for data with few cases of repeated measurements, we performed the simulation study. We found that in general, using more than 20 quadrature points gave quite stable results, but for the dataset with only one or two repeated measurements, there were many extreme values and quite large uncertainties in the estimated values. When adding repeated measurements such that all infants had two repeated measurement, the outliers disappeared and the uncertainty of the estimated coefficient values were considerable reduced. We suspected that the unstable results could be caused by an non-valid approximation of a normal distribution for the posterior, as described in Section 3.4.1. However, in the simulation study, we had at most four repeated measurements per infant. One could question if four repeated measurements could be interpreted as a large cluster size, and is enough to have a good normal approximation. Due to this, one could also question the validity of the simulation study.

When using 50 quadrature points for modeling the frequentist approach for the FM-data, we found that the estimated coefficient values were quite extreme, with large confi-

dence intervals. When applying a Bayesian approach for the same model, the estimated coefficient values as well as the width of the credible intervals were reduced. The Bayesian approach showed that there were little information in the data about the precision of the random intercept and the interaction variables between C_{sd} and country, as their posterior distributions were highly affected by their priors. Hence, as for the frequentist approach, estimation of the coefficients for the random intercept and the interaction variables are difficult with these data. However, the Bayesian approach showed that even though the posterior for the precision of the random intercept was highly affected by the prior distribution, the choice of this prior had little impact on the posteriors for the coefficients of the fixed effects.

Also for the CP-models, we have seen that finding good models with statistically significant results are difficult with these data. Even though we found that the models with several GMT-variables were better for prediction of CP than the model with only the C_{sd} variable, the AUC-values for the Lasso models with and without other available variables were only acceptable according to Lydersen (2012). In addition, the evaluation of the model showed that non of the GMT-variables were statistically significant when adjusting for multiple testing. Hence, in order to find a good model for prediction of CP that can be used in clinical practice, even larger samples are needed. As CP is a rare disease, one should perhaps include new centers in the study, in order to increase the number of infants with CP in the dataset.

Chapter 6

Further work

For further analysis of these data and models, we would like to point out some improvements. First is the prediction of population averaged probabilities for having normal FMs. Using the lme4-package for predictions in the frequentist approach, the population averaged probabilities are not available. In order to get correct population averaged probabilities for validation of the model, one should write own code for performing this.

In the variable selection, we included the variable for the length of the video recording as a covariate. The length of the video recordings could give less uncertainty in the GMT-variables for the infants, as longer video recordings are likely to include more movements, and abnormal movements might be easier to detect. To account for this, the variable for the length of the video recordings could have been included as an interaction with the GMT-variables. In addition, the trunk area have previously been adjusted for, by normalizing all the GMT-variables with respect to the trunk area. In this thesis, we have not considered any of these adjusted variables, we have only included the trunk area as a variable in the Lasso model. For future analysis, one can look further into models for prediction of CP with the adjusted variables, or one could look for other methods to account for the trunk area of the infants.

Also, in this thesis, we have assumed that repeated measurements within each infant is correlated and can not be considered independent of each other. For prediction of CP, we have selected only one of the recording in order to do logistic regression for modeling the data. In this way, we don't use all the information we have available. We have not considered methods for modeling repeated covariates with one response, where the repeated measurements are included. For further work on these data, one could have looked into these type of methods for modeling the data, to include more of the available information.

Bibliography

Adamina, M., Tomlinson, G., Guller, U., 2009. Bayesian statistics in oncology. Cancer 115 (23), 5371–5381.
URL http://dx.doi.org/10.1002/cncr.24628

Adde, L., 2010. Prediction of cerebral palsy in young infants : computer-based assessment of general movements. Ph.D. thesis, Norwegian University of Science and Technology.

Adde, L., Helbostad, J. L., Jensenius, A. R., Taraldsen, G., Grunewaldt, K. H., Støen, R., 2010. Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study. Developmental Medicine & Child Neurology 52 (8), 773–778.
URL http://dx.doi.org/10.1111/j.1469-8749.2010.03629.x

Adde, L., Helbostad, J. L., Jensenius, A. R., Taraldsen, G., Støen, R., 2009. Using computer-based video analysis in the study of fidgety movements. Early Human Development 85 (9), 541–547.
URL http://www.sciencedirect.com/science/article/pii/S0378378209000814

Adde, L., Rygg, M., Lossius, K., Øberg, G. K., Støen, R., 2007. General movement assessment: Predicting cerebral palsy in clinical practise. Early Human Development 83 (1), 13 – 18.
URL http://www.sciencedirect.com/science/article/pii/S0378378206000892

Adde, L., Thomas, N., John, H. B., Oommen, S., Vågen, R. T., Fjørtoft, T., Jensenius, A. R., Støen, R., 2016. Early motor repertoire in very low birth weight infants in india is associated with motor development at one year. European Journal of Paediatric Neurology 20 (6), 918 – 924.
URL http://www.sciencedirect.com/science/article/pii/S1090379816301258

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67 (1), 1–48.
URL https://cran.r-project.org/web/packages/lme4/lme4.pdf

Blangiardo, M., Cameletti, M., 2015. Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley and Sons Inc.

Blauw-Hospers, C. H., Hadders-Algra, M., 2005. A systematic review of the effects of early intervention on motor development. Developmental Medicine and Child Neurology 47 (6), 421–432.
URL http://dx.doi.org/10.1111/j.1469-8749.2005.tb01165.x

Bosanquet, M., Copeland, L., Ware, R., Boyd, R., 2013. A systematic review of tests to predict cerebral palsy in young children. Developmental Medicine and Child Neurology 55 (5), 418–426.
URL http://dx.doi.org/10.1111/dmcn.12140

Box, G. E., Tiao, G. C., 1992. Bayesian inference in statistical analysis. John Wiley and Sons, Inc.

Carlin, B. P., Louis, T. A., 1996. Bayes and empirical Bayes methods for data analysis. Chapman & Hall.

Darsaklis, Snider, L., Majnemer, A., Mazer, B., 2011. Predictive validity of prechtl's method on the qualitative assessment of general movements: a systematic review of the evidence. Developmental Medicine & Child Neurology 53 (10), 896–906.
URL http://dx.doi.org/10.1111/j.1469-8749.2011.04017.x

de Graaf-Peters, V. B., Hadders-Algra, M., 2006. Ontogeny of the human central nervous system: What is happening when? Early Human Development 82 (4), 257–266.

de Kieviet, J. F., Piek, a. P., Aarnoudse-Moens, C. S., Oosterlaan, J., 2009. Motor development in very preterm and very low-birth-weight children from birth to adolescence: A meta-analysis. JAMA 302 (20), 2235–2242.
URL +http://dx.doi.org/10.1001/jama.2009.1708

Dezeure, R., Bühlmann, P., Meier, L., Nicolai, M., 2015. High-dimensional inference: Confidence intervals, p-values and r-software hdi.

Einspieler, C., Prechtl, H. F. R., 2005. Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system. Mental Retardation and Developmental Disabilities Research Reviews 11 (1), 61–67.

Einspieler, C., Prechtl, H. F. R., Bos, A. F., 2004. Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. Mac Keith Press.

Finos, L., Brombin, C., Salmaso, L., 2010. Adjusting stepwise p-values in generalized linear models. Communications in Statistics: Theory and Methods 39 (10), 1832–1846.

Fong, Y., Rue, H., Wakefield, J., 2010. Bayesian inference for generalized linear mixed models. Biostatistics 11 (3), 397412.
URL +http://dx.doi.org/10.1093/biostatistics/kxp053

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.

Grilli, L., Metelli, S., Rampichini, C., 2015. Bayesian estimation with integrated nested laplace approximation for binary logit mixed models. Journal of Statistical Computation and Simulation 85 (13), 2718–2726.
URL http://dx.doi.org/10.1080/00949655.2014.935377

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall.

Hastie, T. J., Tibshirani, R. J., Friedman, J., 2001. The elements of statistical learning : data mining, inference, and prediction. Springer.

Heineman, K. R., Hadders-Algra, M., August 2008. Evaluation of Neuromotor Function in Infancy–A Systematic Review of Available Methods. Journal of developmental and behavioral pediatrics 29 (4), 315–323.

Hopkins, B., Prechtl, H. F. R., 1984. Continuity of neural functions from pretermal to postnatal life. Spastics International Medical Pub., Ch. 12, pp. 179–197.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. Springer New York, New York, NY.

Lindström, K., Bremberg, S., 1997. The contribution of developmental surveillance to early detection of cerebral palsy. Acta Pædiatrica 86 (7), 736–739.
URL http://dx.doi.org/10.1111/j.1651-2227.1997.tb08577.x

Lydersen, S., 2012. Medical statistics: in clinical and epidemiological research. Gyldendal akademisk, Ch. 14.

Marcroft, C., Khan, A., Embleton, N., Trenell, M., Plotz, T., 2015. Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. Frontiers In Neurology 5, 284.

McLellan, T., Goldstein, S., Naglieri, J. A., 2011. Plasticity of the Brain. Springer US, Boston, MA, pp. 1114–1115.
URL http://dx.doi.org/10.1007/978-0-387-79061-9_2179

Oskoui, M., Coutinho, F., Dykeman, J., Jetté, N., Pringsheim, T., 2013. An update on the prevalence of cerebral palsy: a systematic review and meta-analysis. Developmental Medicine and Child Neurology 55 (6), 509–519.
URL http://dx.doi.org/10.1111/dmcn.12080

Pavlou, M., Ambler, G., Seaman, S., Omar, R. Z., 2015. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. BMC Medical Research Methodology 15 (1), 59.
URL http://dx.doi.org/10.1186/s12874-015-0046-6

Platt, M. J., Cans, C., Johnson, A., Surman, G., Topp, M., Torrioli, M. G., Krageloh-Mann, I., 2007. Trends in cerebral palsy among infants of very low birthweight (lt;1500 g) or born prematurely (lt;32 weeks) in 16 European centres: a database study. The Lancet 369 (9555), 43–50.
URL http://www.sciencedirect.com/science/article/pii/S0140673607600300

Prechtl, H. F., Einspieler, C., Cioni, G., Bos, A. F., Ferrari, F., Sontheimer, D., 1997. An early marker for neurological deficits after perinatal brain lesions. The Lancet 349 (9062), 1361–1363.
URL http://www.sciencedirect.com/science/article/pii/S0140673696101823

Prechtl, H. F. R., 1990. New Studies on Movement Assessment in Fetuses and Preterm Infants Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. Early Human Development 23 (3), 151–158.
URL http://www.sciencedirect.com/science/article/pii/0378378290900117

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S., Skrondal, A., 2012. Multilevel and longitudinal modeling using Stata : Vol. 2 : Categorical responses, counts, and survival, 3rd Edition. Stata Press.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics 128 (2), 301 – 323.
URL http://www.sciencedirect.com/science/article/pii/S0304407604001599

Rodríguez, G., 2007. Lecture Notes on Generalized Linear Models. "http://data.princeton.edu/wws509/notes/.

Rosenbaum, P., Paneth, N., Leviton, A., Goldstein, M., Bax, M., 2007. A report: the definition and classification of cerebral palsy - April 2006. Developmental Medicine And Child Neurology 49, 8–14.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2), 319–392.

Rue, H., Ribler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., Lindgren, F. K., 2016. Bayesian computing with inla: A review, provided by the SAO/NASA Astrophysics Data System.

Rufibach, K., 2010. Use of brier score to assess binary predictions. Journal of Clinical Epidemiology 63 (8), 938 – 939.
URL http://www.sciencedirect.com/science/article/pii/S0895435609003631

Self, S. G., Liang, K.-Y., June 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82 (398), 605–610.

Sellier, E., Platt, M. J., Andersen, G. L., Krägeloh-Mann, I., De La Cruz, J., Cans, C., 2016. Decreasing prevalence in cerebral palsy: a multi-site European population-based study, 1980 to 2003. Developmental Medicine and Child Neurology 58 (1), 85–92.
URL http://dx.doi.org/10.1111/dmcn.12865

Spittle, A. J., Boyd, R. N., Inder, T. E., Doyle, L. W., 2009. Predicting Motor Development in Very Preterm Infants at 12 Months' Corrected Age: The Role of Qualitative Magnetic Resonance Imaging and General Movements Assessments. Pediatrics 123 (2), 512–517.
URL http://pediatrics.aappublications.org/content/123/2/512

Stephens, B. E., Vohr, B. R., 2009. Neurodevelopmental Outcome of the Premature Infant. The pediatric clinics of North America 56 (3), 631–646.

Süli, E., Mayers, D., 2003. An Introduction to Numerical Analysis, 8th Edition. Cambridge University Press.

Veelken, N., Just, K., 2013. P7 – 1841 Longterm development of VLBW-infants with cerebral palsy. Results of a population-based study. European Journal of Paediatric Neurology 17, S55.
URL http://www.sciencedirect.com/science/article/pii/S1090379813701867

Yarnell, J., O'Reilly, D., 2013. Epidemiology and disease prevention : a global approach, 2nd Edition. Oxford University Press, Ch. 13, pp. 190–194.

Appendix

A Simulation study

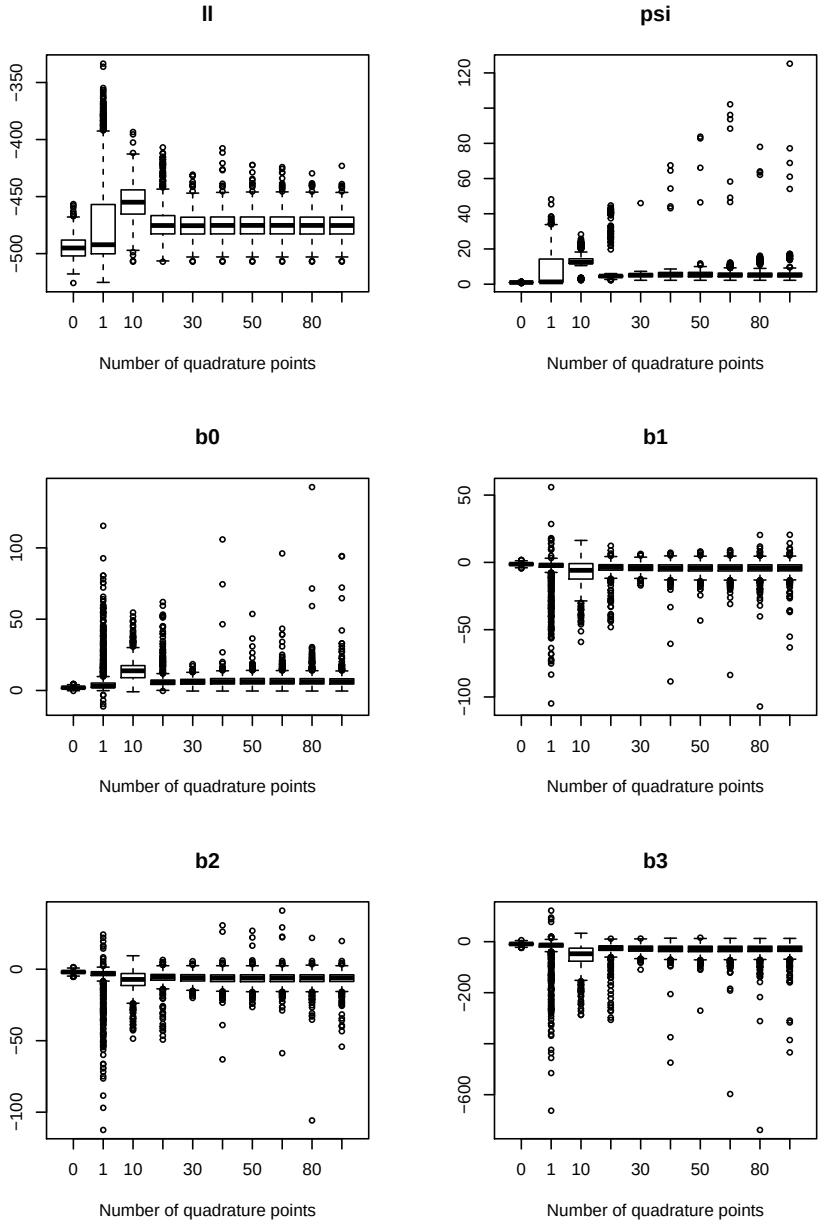


Figure A.1: Estimated values from the simulation study for the log likelihood, ψ and $\beta_0 - \beta_3$ for different number of quadrature points in case 1.

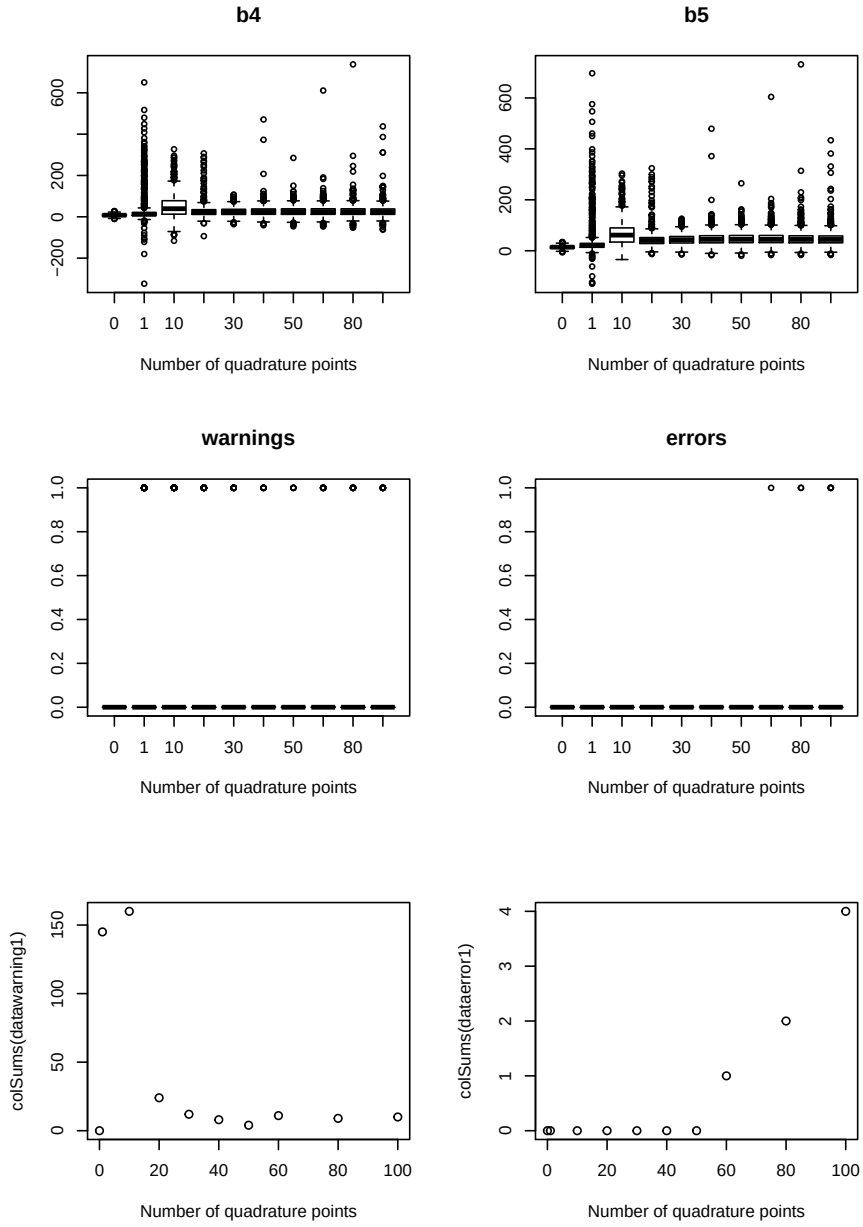


Figure A.2: Estimated values from the simulation study for β_4, β_5, number of warnings and number of errors for different number of quadrature points in case 1.

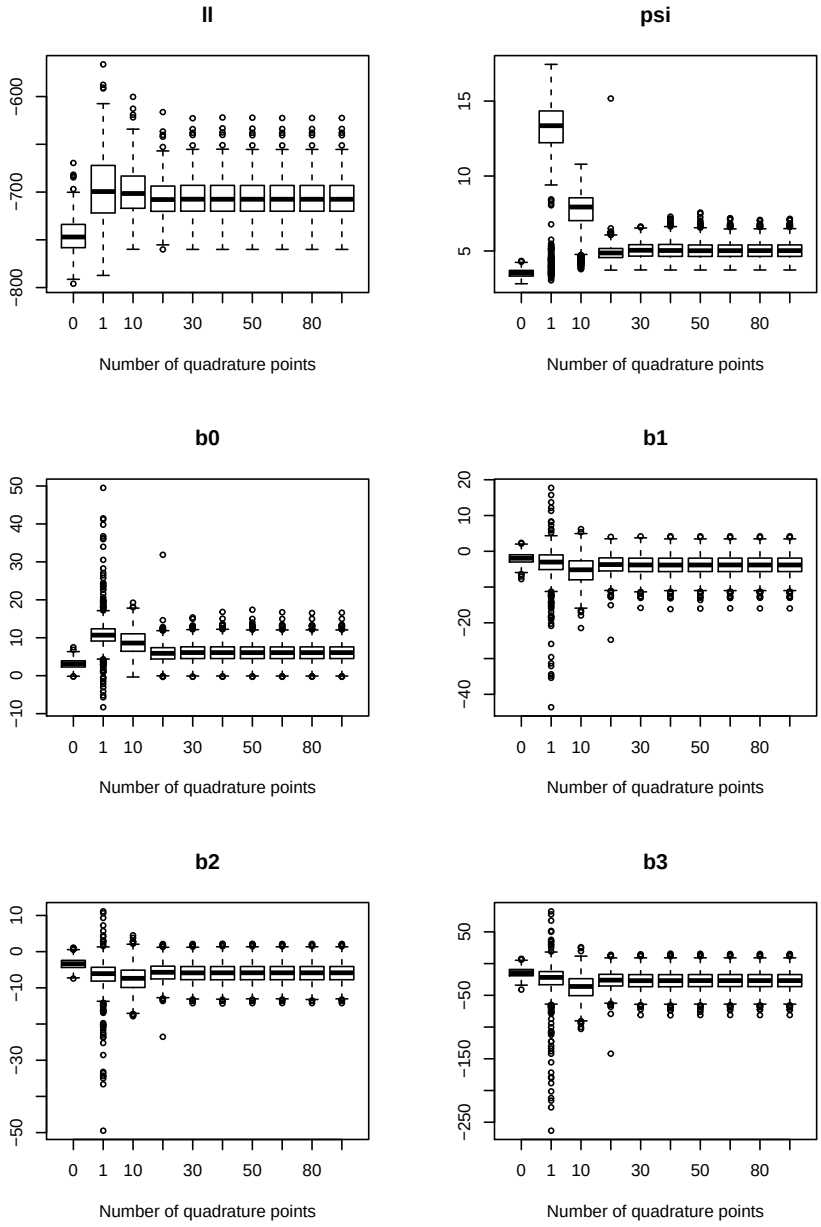


Figure A.3: Estimated values from the simulation study for the log likelihood, ψ and $\beta_0 - \beta_3$ for different number of quadrature points in case 2.

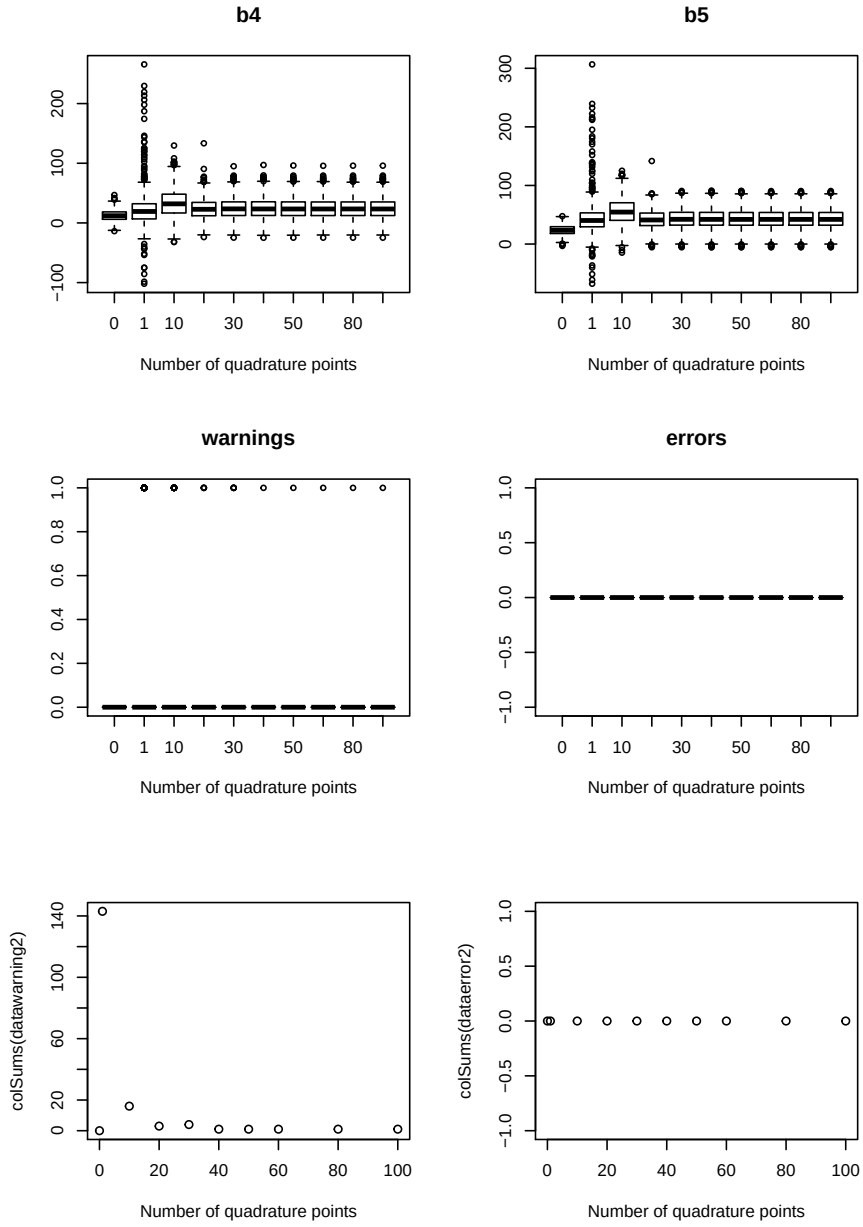


Figure A.4: Estimated values from the simulation study for β_4, β_5, number of warnings and number of errors for different number of quadrature points in case 2.

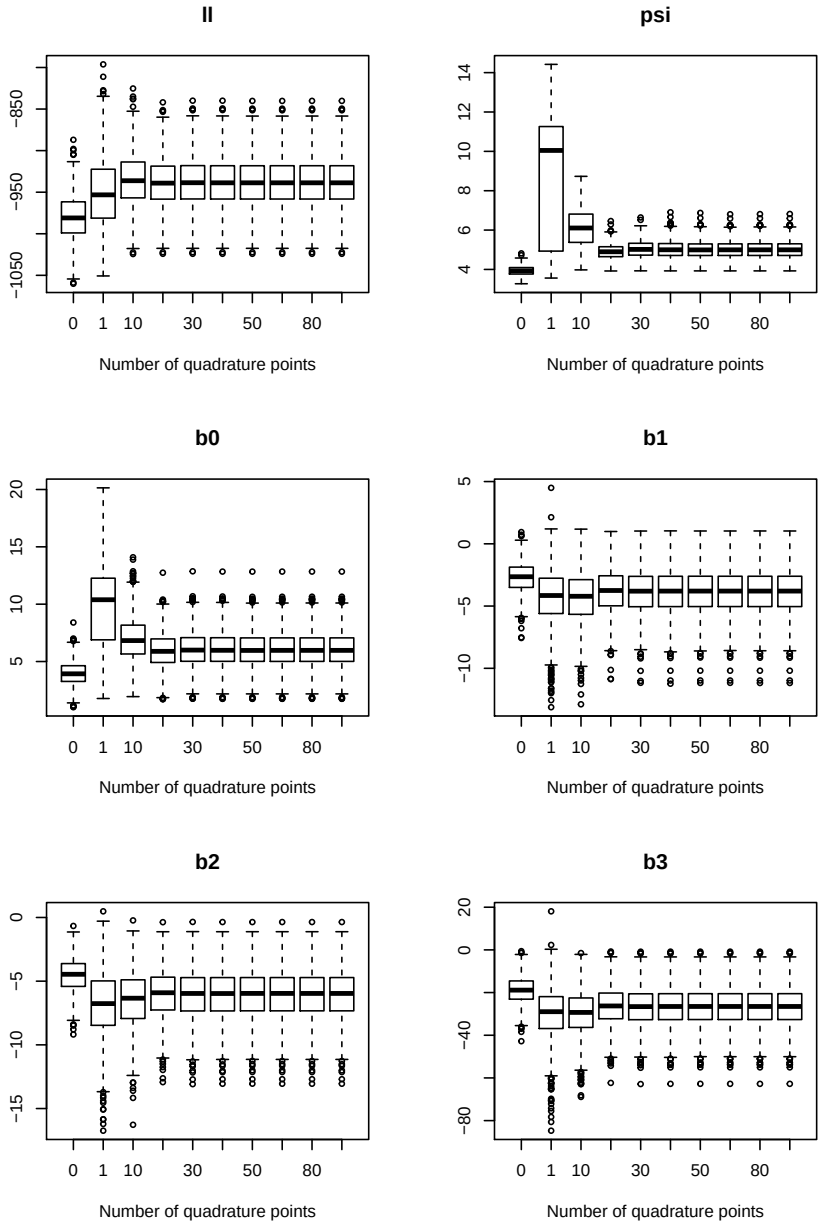


Figure A.5: Estimated values from the simulation study for the log likelihood, ψ and $\beta_0 - \beta_3$ for different number of quadrature points in case 3.

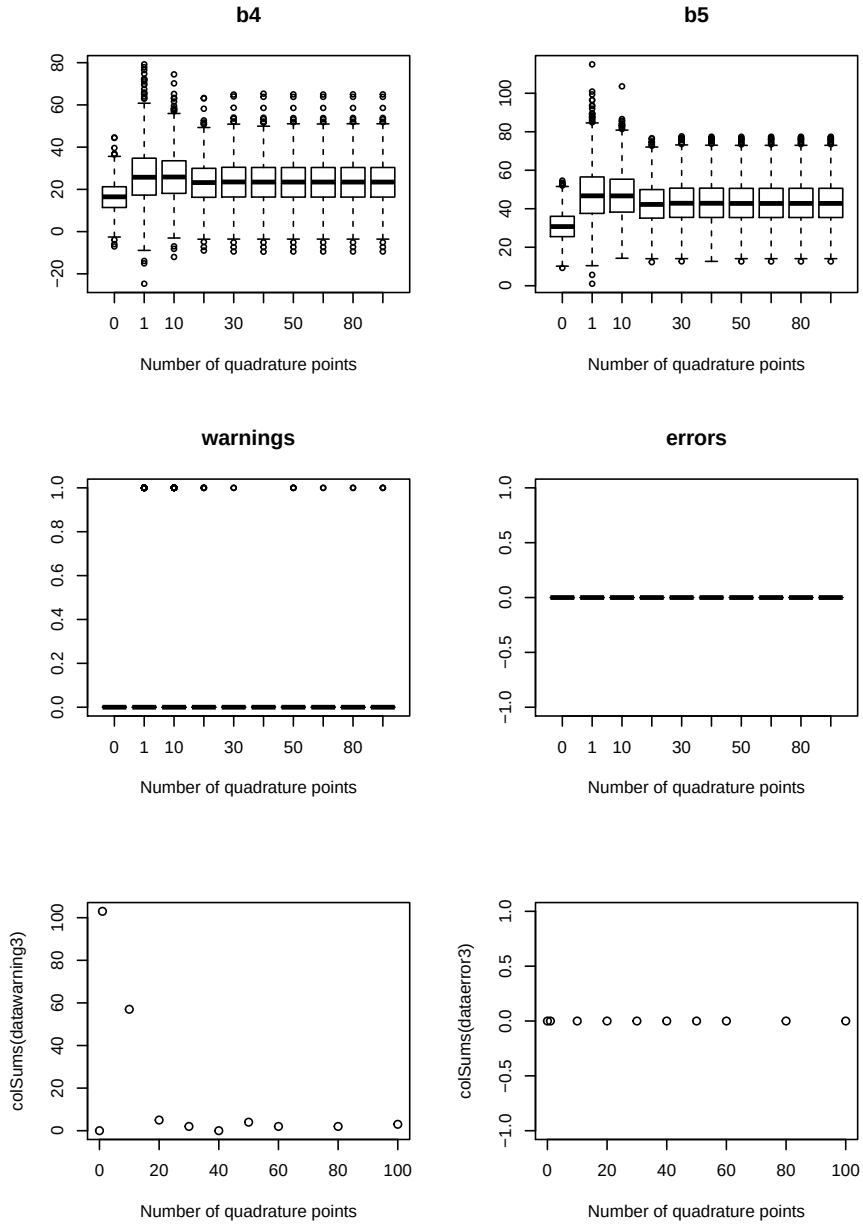


Figure A.6: Estimated values from the simulation study for β_4, β_5, number of warnings and number of errors for different number of quadrature points in case 3.

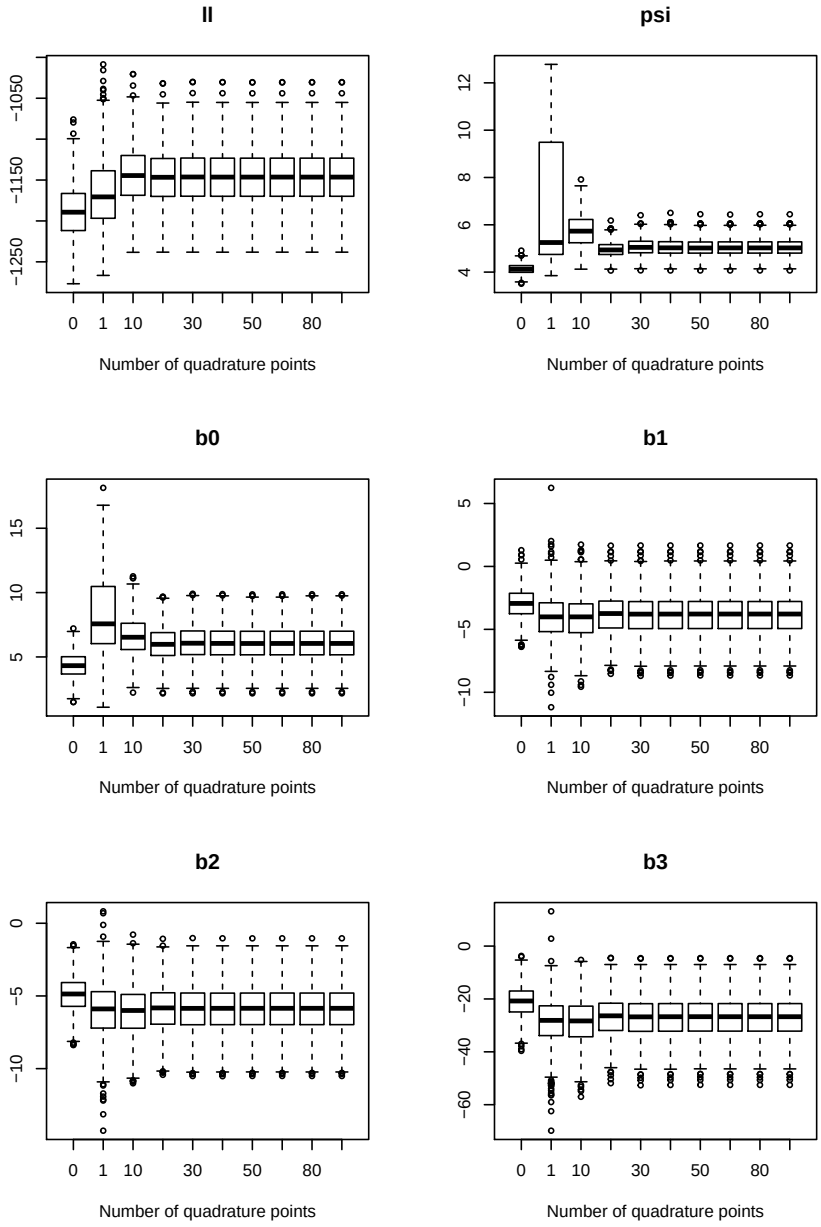


Figure A.7: Estimated values from the simulation study for the log likelihood, ψ and $\beta_0 - \beta_3$ for different number of quadrature points in case 4.

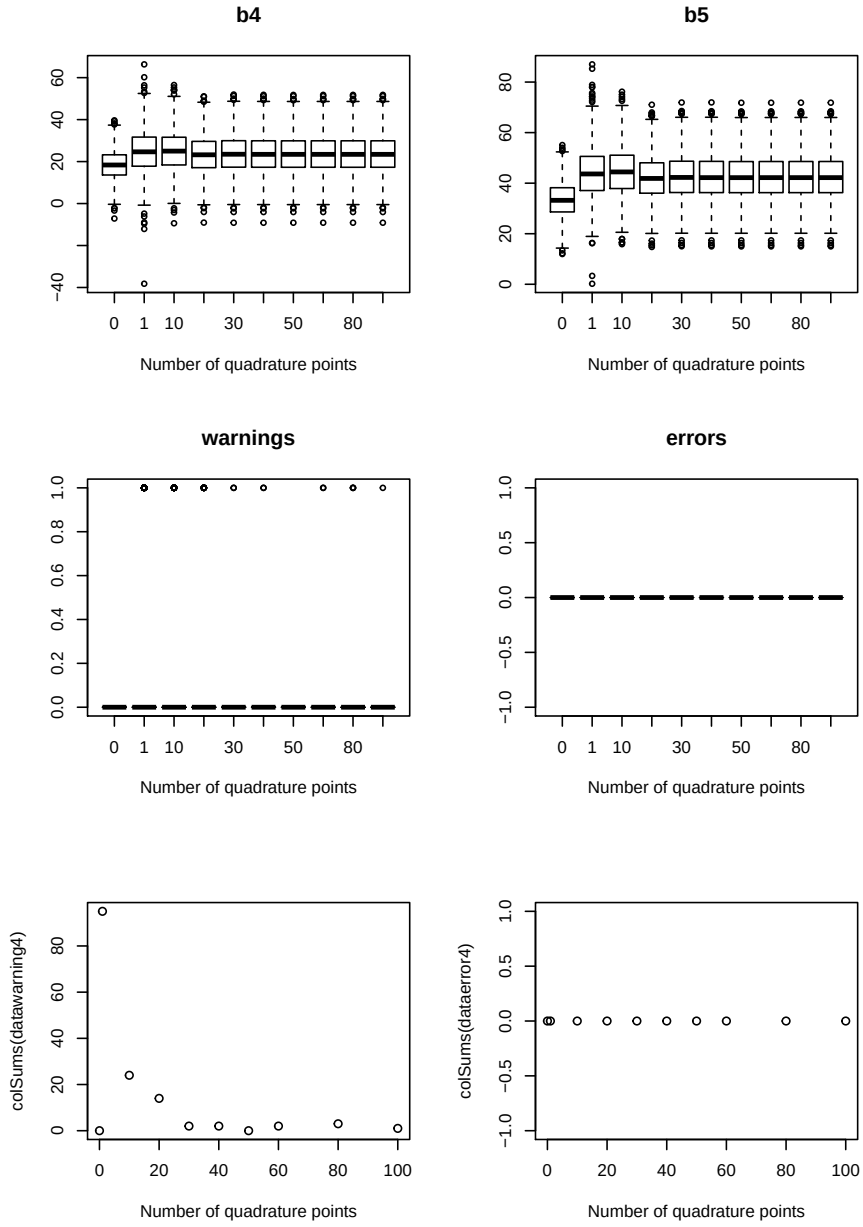


Figure A.8: Estimated values from the simulation study for β_4, β_5, number of warnings and number of errors for different number of quadrature points in case 4.

B Bootstrap

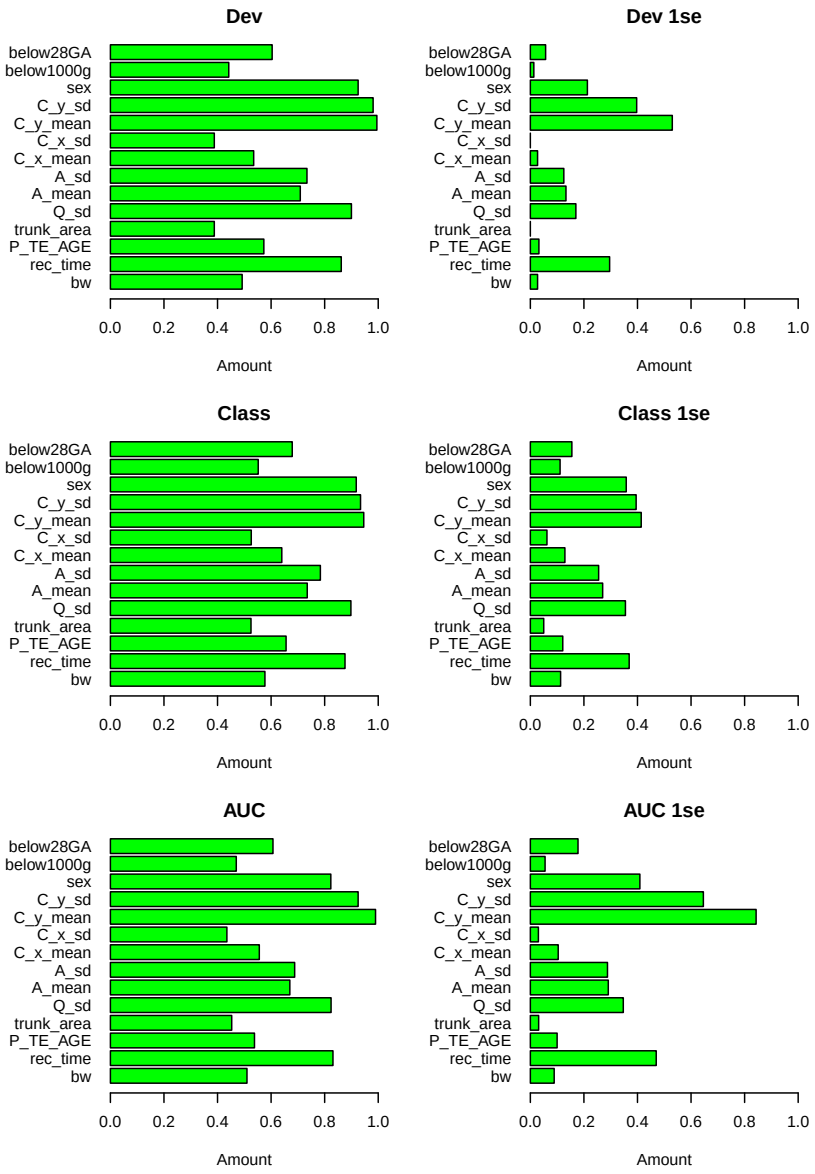


Figure B.1: Amount of nonzero Lasso coefficients from the 1000 bootstrap replicates for minimum/maximum and one standard error within the minimum/maximum for different decision rules; binomial deviance, misclassification error and AUC.

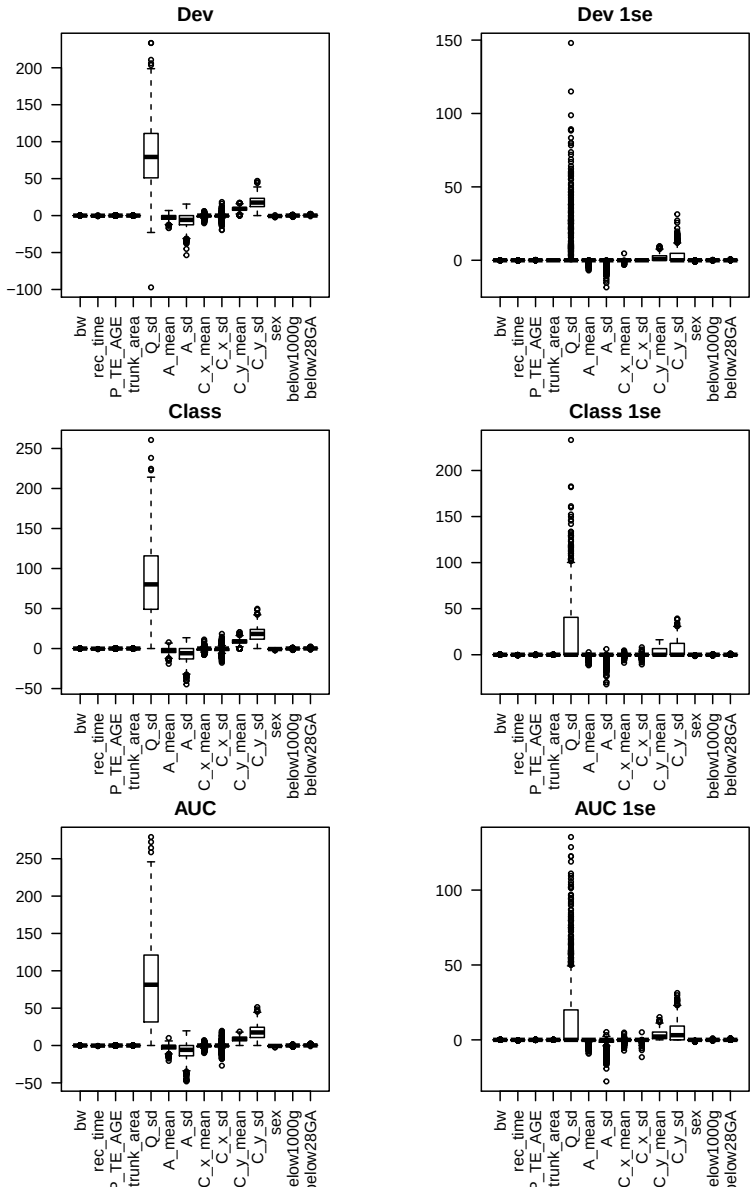


Figure B.2: Estimated Lasso coefficients from the bootstrap replicates using different decision rules with both $\lambda_{min}/\lambda_{max}$ and λ_{1se}.

C R code for simulation study

```
1
2  library("lme4")
3  #load data
4  data <- read.table(file="Master/newdata.txt")
5  #View(data)
6  data$Continent = as.factor(data$Continent)
7  data$FMs_old = as.factor(data$FMs_old)
8  #number of recordings pr infants
9  n = dim(data)[1]
10 data$rep = 1
11 for(i in 2:n){
12   if(data$PID[i]==data$PID[i-1]){
13     data$rep[i] = data$rep[i-1] + 1
14   }
15 }
16
17 #remove recording 3 and 4
18 case1data <- data[1,]
19 teller = 2
20 for(i in 2:dim(data)[1]){
21   if(data$rep[i] !=3 && data$rep[i] != 4){
22     case1data[teller,] = data[i,]
23     teller = teller + 1
24   }
25 }
26
27 case1data$Continent = as.factor(case1data$Continent)
28 case1data$FMs_old = as.factor(case1data$FMs_old)
29
30 #find coefficients
31 FMoldtrue <- glm(formula = FMs_old~com_sd + Continent + com_sd*
       Continent, data = data, family="binomial")
32 beta <- FMoldtrue$coefficients
33 beta
34 psi = 5 # variance of zeta
35
36 #case2data <- read.table(file="Master/case2data2.txt")
37 #case2data <- read.table(file="//Users/martinahall/Documents/
       Master/case2data.txt")
38 #case2data = case2data[order(case2data$PID),]
39 #case2data$FMs_old = as.factor(case2data$FMs_old)
40 #case2data$Continent = as.factor(case2data$Continent)
41
42
43 #case3data <- read.table(file="//Users/martinahall/Documents/
       Master/case3data.txt")
44 #case3data <- read.table(file="Master/case3data2.txt")
```

```
#case3data = case3data[order(case3data$PID),]
#case3data$FMs_old = as.factor(case3data$FMs_old)
#case3data$Continent = as.factor(case3data$Continent)

#case4data <- read.table(file="//Users/martinahall/Documents/
    Master/case4data.txt")
case4data <- read.table(file="Master/case4data2.txt")
case4data = case4data[order(case4data$PID),]
case4data$FMs_old = as.factor(case4data$FMs_old)
case4data$Continent = as.factor(case4data$Continent)

simsim <- function(beta,psi,data, nIter, case){
  simmat <- array(1, dim=c(10,10,nIter)) # quadrature points in
      column, variable in row, iteration in last dim
  rownames(simmat) = c("ll", "psi", "b0", "b1", "b2", "b3", "b4",
      "b5", "warnings", "error")
  FMnVec = rep(NA, nIter)
  FMpVec = rep(NA, nIter)
  likelihood = rep(NA, nIter)
  for(simrep in 1:nIter){
    #for case 1:
    #zeta similar for same PID(person)
    if(case==1){
      zeta = rnorm(n=dim(data)[1], mean = 0, sd = psi)
      for(i in 1:dim(data)[1]){
        if(data$rep[i] ==2){
          zeta[i] = zeta[i-1]
        }
      }
    }
    else{
      #for case 2,3,4
      zeta = rnorm(n=dim(data)[1]/case, mean = 0, sd = psi)
      zeta = rep(zeta,times=rep(case,length(zeta))) #data maa
          vaere sortert forst
    }

    #find pi_ij = exp(xbeta + zeta_j)/(1+exp(xbeta + zeta_j))
    mf = model.frame(formula = FMs_old~com_sd + Continent + com_sd
        *Continent, data = data)
    X  = model.matrix(attr(mf, "terms"), data = mf)
    pi = as.numeric(exp(X%*%beta + zeta)/(1+exp(X%*%beta + zeta)))

    #simulate y
    FMold = rbinom(n=length(pi), size=1, prob=pi)
    FMold = as.factor(FMold)
    #store result
    FMnVec[simrep] = sum(FMold==0)
    FMpVec[simrep] = sum(FMold==1)

```

```
#make dataset
dataset <- data.frame(FMold = FMold, Continent = data$
    Continent, com_sd = data$com_sd, pid = data$PID)

#glm model - for testing random intercept
glmModel <- glm(formula = FMold~Continent + com_sd + Continent
    *com_sd, data = dataset, family = "binomial")
#store likelihood
likelihood[simrep] <- as.numeric(logLik(glmModel))

#catch errors and warnings
myTryCatch <- function(expr) {
  warn <- err <- NULL
  value <- withCallingHandlers(
    tryCatch(expr, error=function(e) {
      err <<- e
      NULL
    }), warning=function(w) {
      warn <<- w
      invokeRestart("muffleWarning")
    })
  list(value=value, warning=warn, error=err)
}

#fit model for different number of quadrature points
vec = c(0,1,10,20,30,40,50,60,80,100)
m = length(vec)
ll = rep(NA, m)
psiVec = rep(NA, m)
b0 = rep(NA, m)
b1= rep(NA, m)
b2 = rep(NA, m)
b3= rep(NA, m)
b4 = rep(NA, m)
b5= rep(NA, m)
warning = rep(NA, m)
error = rep(NA, m)
for(i in 1:m){
  model_random <- myTryCatch(glmer(formula = FMold~Continent +
      com_sd + Continent*com_sd + (1|pid), data = dataset,
      family = "binomial", nAGQ=vec[i]))

  if(is.null(model_random$error)){
    ll[i] = logLik(model_random$value)
    psiVec[i] = as.numeric(getME(model_random$value, "theta"))
    b0[i] = as.numeric(fixef(model_random$value)[1])
    b1[i] = as.numeric(fixef(model_random$value)[2])
    b2[i] = as.numeric(fixef(model_random$value)[3])
    b3[i] = as.numeric(fixef(model_random$value)[4])
    b4[i] = as.numeric(fixef(model_random$value)[5])
```

```
136           b5[i] = as.numeric(fixef(model_random$value)[6])
137           error[i] = 0
138           if(is.null(model_random$warning)){
139             warning[i] = 0
140           }
141           else{warning[i] = 1}
142 
143         }
144         else{
145           error[i] = 1
146           warning[i] = 1
147           ll[i] = NA
148           psiVec[i] = NA
149           b0[i] = NA
150           b1[i] = NA
151           b2[i] = NA
152           b3[i] = NA
153           b4[i] = NA
154           b5[i] = NA
155         }
156 
157       }
158       #save results
159       simmat[1,,simrep] = ll
160       simmat[2,,simrep] = psiVec
161       simmat[3,,simrep] = b0
162       simmat[4,,simrep] = b1
163       simmat[5,,simrep] = b2
164       simmat[6,,simrep] = b3
165       simmat[7,,simrep] = b4
166       simmat[8,,simrep] = b5
167       simmat[9,,simrep] = warning
168       simmat[10,,simrep] = error
169 
170       print(simrep)
171       Sys.time()
172       flush.console()
173     }
174 
175     return(list(simmat = simmat, FMn = FMnVec, FMp = FMpVec,
          Likelihood = likelihood))
176 }#end function
177 
178 #ptm <- proc.time()
179 set.seed(200)
180 ptm <- proc.time()
181 #change value of data and case for different cases
182 result <- simsim(beta,psi,data = case4data, nIter=1000, case=4)
183 proc.time() - ptm
184 saveRDS(result, file = "Master/case4withlogL2.rds")
```

D R-code INLA

```
#install
install.packages("sp") # needed for INLA
install.packages("INLA", repos = "http://www.math.ntnu.no/inla/R/
    testing")
library(INLA)

#load data
mydata = read.table(file="//Users/martinahall/Documents/Master/
    newdata.txt", header = TRUE)
y=mydata$FMs_old
com_sd = mydata$com_sd
Continent = as.factor(mydata$Continent)
PID=mydata$PID

data = list(y=y, com_sd = com_sd, Continent = Continent, PID=PID)

##### Fongs prior #####
formula = y~com_sd + Continent + com_sd*Continent +
  f(PID, model="iid", hyper = list(theta = list(prior = "loggamma"
      , param = c(0.5,0.0164))))
modelFong = inla(formula, data = data, family = "binomial",
    Ntrials = rep(1,length(y)))

summary(modelFong)
#fixed marginal distributions
mF = modelFong$marginals.fixed

##### Default prior #####
formula = y~com_sd + Continent + com_sd*Continent +
  f(PID, model="iid")
modelDef = inla(formula, data = data, family = "binomial", Ntrials
    = rep(1,length(y)))

summary(modelDef)
#fixed marginal distributions
mD = modelDef$marginals.fixed

##### Popular prior #####
formula = y~com_sd + Continent + com_sd*Continent +
  f(PID, model="iid", hyper = list(theta = list(prior = "loggamma"
      , param = c(0.001,0.001))))
modelPop = inla(formula, data = data, family = "binomial", Ntrials
    = rep(1,length(y)))

summary(modelPop)
#fixed marginal distributions
mP = modelPop$marginals.fixed
```

```
41
42 ########################
43 #Plot for all the fixed marginals for different priors for the
       random intercept precision.
44 par(mfrow=c(2,3))
45 plot(mP[[1]], main="Intercept", type="l",ylim=c(0,0.4), xlim=c
       (0,12))
46 lines(mF[[1]], col="red", type="l")
47 lines(mD[[1]], col="blue", type="l")
48 legend("topright",
49         bty = "n",
50         inset=0,
51         cex = 0.6,
52         title="Prior",
53         c("Popular","Fong", "Default"),
54         horiz=FALSE,
55         lty=c(1,1,1),
56         lwd=c(2,2,2),
57         col=c("black","red", "blue"),
58         bg="grey96")
59 plot(mP[[2]], main="C_sd", type="l",ylim=c(0,0.08), xlim=c(-50,0))
60 lines(mF[[2]], col="red")
61 lines(mD[[2]], col="blue")
62 legend("topright",
63         bty = "n",
64         inset=0,
65         cex = 0.6,
66         title="Prior",
67         c("Popular","Fong", "Default"),
68         horiz=FALSE,
69         lty=c(1,1,1),
70         lwd=c(2,2,2),
71         col=c("black","red", "blue"),
72         bg="grey96")
73 plot(mP[[3]], main="USA", type="l",ylim=c(0,0.35), xlim=c(-10,4))
74 lines(mF[[3]], col="red")
75 lines(mD[[3]], col="blue")
76 legend("topright",
77         bty = "n",
78         inset=0,
79         cex = 0.6,
80         title="Prior",
81         c("Popular","Fong", "Default"),
82         horiz=FALSE,
83         lty=c(1,1,1),
84         lwd=c(2,2,2),
85         col=c("black","red", "blue"),
86         bg="grey96")
87 plot(mP[[4]], main="India", type="l",ylim=c(0,0.3), xlim=c(-15,5))
88 lines(mF[[4]], col="red")
```

```
lines(mD[[4]], col="blue")
legend("topright",
       bty = "n",
       inset=0,
       cex = 0.6,
       title="Prior",
       c("Popular","Fong", "Default"),
       horiz=FALSE,
       lty=c(1,1,1),
       lwd=c(2,2,2),
       col=c("black","red", "blue"),
       bg="grey96")
plot(mP[[5]], main="C_sd*USA", type="l",ylim=c(0,0.06), xlim=c
    (-20,60))
lines(mF[[5]], col="red")
lines(mD[[5]], col="blue")
legend("topright",
       bty = "n",
       inset=0,
       cex = 0.6,
       title="Prior",
       c("Popular","Fong", "Default"),
       horiz=FALSE,
       lty=c(1,1,1),
       lwd=c(2,2,2),
       col=c("black","red", "blue"),
       bg="grey96")
plot(mP[[6]], main="C_sd*India", type="l",ylim=c(0,0.05), xlim=c
    (-10,85))
lines(mF[[6]], col="red")
lines(mD[[6]], col="blue")
legend("topright",
       bty = "n",
       inset=0,
       cex = 0.6,
       title="Prior",
       c("Popular","Fong", "Default"),
       horiz=FALSE,
       lty=c(1,1,1),
       lwd=c(2,2,2),
       col=c("black","red", "blue"),
       bg="grey96")


#Plot of precision for the random intercept with different priors
    for the precision. The beta's have default priors.
par(mfrow=c(1,1))
plot(modelPop$marginals.hyper$`Precision for PID`[1:60,], type="l"
    , ylim=c(0,0.95))
lines(x=seq(from=0,to=4, by=0.1), dgamma(x=seq(from=0,to=4, by
```

```
    =0.1), shape=modelPop$all.hyper$random[[1]]$hyper$theta$param
    [1], rate=modelPop$all.hyper$random[[1]]$hyper$theta$param[2])
    , col="black", lty=2)
lines(modelFong$marginals.hyper$`Precision for PID`[1:60,], col="
    red")
lines(x=seq(from=0,to=4, by=0.1), dgamma(x=seq(from=0,to=4, by
    =0.1), shape=modelFong$all.hyper$random[[1]]$hyper$theta$param
    [1], rate=modelFong$all.hyper$random[[1]]$hyper$theta$param
    [2]), col="red", lty=2)
lines(modelDef$marginals.hyper$`Precision for PID`[1:8,], col="
    blue")
lines(x=seq(from=0,to=4, by=0.1), dgamma(x=seq(from=0,to=4, by
    =0.1), shape=modelDef$all.hyper$random[[1]]$hyper$theta$param
    [1], rate=modelDef$all.hyper$random[[1]]$hyper$theta$param[2])
    , col="blue", lty=2)
legend("topright",
       inset=0.01,
        cex = 0.8,
        c("Posterior Popular","Posterior Fong", "Posterior Default"
            , "Prior Popular", "Prior Fong", "Prior Defautl"),
        horiz=FALSE,
        lty=c(1,1,1,2,2,2),
        lwd=c(2,2,2,2,2,2),
        col=c("black","red", "blue"),
        bg="grey96")


# plot of fixed marginals and their priors
par(mfrow=c(2,3))
#use the popular prior for the random intercept
mP = modelPop$marginals.fixed
plot(mP[[1]], main="Intercept", type="l",ylim=c(0,0.45), col="red"
    )
x=seq(from=-5, to=15, by=0.1)
lines(x=x, y=dnorm(x, mean=0, sd=1/sqrt(0)), col="blue")
legend("topright",
       inset=0,
        cex = 0.6,
        c("Posterior","Prior"),
        horiz=FALSE,
        lty=c(1,1,1),
        lwd=c(2,2,2),
        col=c("red", "blue"),
        bg="grey96")
plot(mP[[2]], main="C_sd", type="l",ylim=c(0,0.08), col="red")
x=seq(from=-80, to=25, by=0.1)
lines(x=x, y=dnorm(x, mean=0, sd=1/sqrt(0.01)), col="blue")
legend("topright",
       inset=0,
        cex = 0.6,
```

```
        c("Posterior","Prior"),
        horiz=FALSE,
        lty=c(1,1,1),
        lwd=c(2,2,2),
        col=c("red", "blue"),
        bg="grey96")
plot(mP[[3]], main="USA", type="l",ylim=c(0,0.35), col="red")
x=seq(from=-15, to=10, by=0.1)
lines(x=x, y=dnorm(x, mean=0, sd=1/sqrt(0.01)), col="blue")
legend("topright",
       inset=0,
        cex = 0.6,
        c("Posterior","Prior"),
        horiz=FALSE,
        lty=c(1,1,1),
        lwd=c(2,2,2),
        col=c("red", "blue"),
        bg="grey96")
plot(mP[[4]], main="India", type="l",ylim=c(0,0.3), col="red")
x=seq(from=-20, to=10, by=0.1)
lines(x=x, y=dnorm(x, mean=0, sd=1/sqrt(0.01)), col="blue")
legend("topright",
       inset=0,
        cex = 0.6,
        c("Posterior","Prior"),
        horiz=FALSE,
        lty=c(1,1,1),
        lwd=c(2,2,2),
        col=c("red", "blue"),
        bg="grey96")
plot(mP[[5]], main="C_sd*USA", type="l",ylim=c(0,0.06), col="red")
x=seq(from=-50, to=90, by=0.1)
lines(x=x, y=dnorm(x, mean=0, sd=1/sqrt(0.01)), col="blue")
legend("topright",
       inset=0,
        cex = 0.6,
        c("Posterior","Prior"),
        horiz=FALSE,
        lty=c(1,1,1),
        lwd=c(2,2,2),
        col=c("red", "blue"),
        bg="grey96")
plot(mP[[6]], main="C_sd*India", type="l",ylim=c(0,0.05), col="red
    ")
x=seq(from=-60, to=120, by=0.1)
lines(x=x, y=dnorm(x, mean=0, sd=1/sqrt(0.01)), col="blue")
legend("topright",
       inset=0,
        cex = 0.6,
        c("Posterior","Prior"),
```

```
221         horiz=FALSE,
222         lty=c(1,1,1),
223         lwd=c(2,2,2),
224         col=c("red", "blue"),
225         bg="grey96")
226
227
228 ##### Odds ratios ####
229 #norway
230 data <- within(data, Continent <- relevel(Continent, ref = 1))
231 formula = y~com_sd + Continent + com_sd*Continent +
232   f(PID, model="iid", hyper = list(theta = list(prior = "loggamma"
        , param = c(0.001,0.001))))
233 modelPop = inla(formula, data = data, family = "binomial", Ntrials
      = rep(1,length(y)))
234
235 orNor <- inla.tmarginal(function(x) exp(0.1*x), modelPop$marginals
    .fixed[[2]])
236 inla.zmarginal(orNor)
237
238 #usa
239 data <- within(data, Continent <- relevel(Continent, ref = 2))
240 formula = y~com_sd + Continent + com_sd*Continent +
241   f(PID, model="iid", hyper = list(theta = list(prior = "loggamma"
        , param = c(0.001,0.001))))
242 modelPop = inla(formula, data = data, family = "binomial", Ntrials
      = rep(1,length(y)))
243 summary(modelPop)
244
245 orUSA <- inla.tmarginal(function(x) exp(0.1*x), modelPop$marginals
    .fixed[[2]])
246 inla.zmarginal(orUSA)
247
248 #india
249 data <- within(data, Continent <- relevel(Continent, ref = 3))
250 formula = y~com_sd + Continent + com_sd*Continent +
251   f(PID, model="iid", hyper = list(theta = list(prior = "loggamma"
        , param = c(0.001,0.001))))
252 modelPop = inla(formula, data = data, family = "binomial", Ntrials
      = rep(1,length(y)))
253 summary(modelPop)
254
255 orIndia <- inla.tmarginal(function(x) exp(0.1*x), modelPop$
    marginals.fixed[[2]])
256 inla.zmarginal(orIndia)
257
258
259 ##### Prediction #####
260
261 #internal validation - population averaged
```

```
PID2 =PID+693
data2 = list(y=c(y, rep(NA, length(y))), com_sd = c(com_sd, com_sd
    ), Continent = c(Continent, Continent), PID=c(PID, PID2))
p_mode2 = rep(NA, length(data$PID))

for(i in 1:length(data$PID)){
  link = rep(NA, length(data2$y))
  link[which(is.na(data2$y))] = 1
  modelPop2 = inla(formula, data = data2, family = "binomial",
      Ntrials = rep(1,length(y)), control.predictor = list(compute
       = TRUE, link = link))
  p_mode2[i] = inla.mmarginal(inla.tmarginal(fun = function(x) exp
      (x)/(1+exp(x)),
                        marginal = modelPop2$marginals.linear.
                            predictor[[798+i]]))
  Country[i] = data2$Continent[798+i]
}
write.table(data.frame(p = p_mode2, Country=Country), file="//
    Users/martinahall/Documents/Master/probInternalINLA.txt")
int <- read.table(file="//Users/martinahall/Documents/Master/
    probInternalINLA.txt")

#External validation: leave-one-out cv - population averaged
p_mode2 = rep(NA, length(data$PID))
for(i in 1:length(data$PID)){
  data2 = data
  data2$y[i] = NA
  link = rep(NA, length(data2$y))
  link[which(is.na(data2$y))] = 1
  modelPop2 = inla(formula, data = data2, family = "binomial",
      Ntrials = rep(1,length(y)), control.predictor = list(compute
       = TRUE, link = link))
  p_mode2[i] = inla.mmarginal(inla.tmarginal(fun = function(x) exp
      (x)/(1+exp(x)),
    marginal = modelPop2$marginals.linear.predictor[[i]]))
}
list = list(external = p_mode2, trueY = data$y)
write.table(list, file="//Users/martinahall/Documents/Master/
    INLAExternal.txt")
p_mode2 <- read.table( file="//Users/martinahall/Documents/Master/
    INLAExternal.txt")


##### AUC - pop.avg from validation #####
library("ROCR")
#extern
predE <- prediction(p_mode2$external, data$y)
perfE <- performance(predE,"tpr","fpr")
aucE <- performance(predE,"auc")
# now converting S4 class to vector
```

```
aucE <- unlist(slot(aucE, "y.values"))

#intern
predI <- prediction(int$p, data$y)
perfI <- performance(predI,"tpr","fpr")
aucI <- performance(predI,"auc")
# now converting S4 class to vector
aucI <- unlist(slot(aucI, "y.values"))

#both in one plot
plot(perfI,col="blue",lty=3, lwd=3,ylab="Sensitivity", xlab="1-
    Spesificity")
lines(c(0,1), c(0,1))
plot(perfE,col="red",lty=3, lwd=3,ylab="Sensitivity", xlab="1-
    Spesificity", add=TRUE)


# adding min and max ROC AUC to the center of the plot
Intern<-min(round(aucI, digits = 4))
Extern<-max(round(aucE, digits = 4))
Internt <- paste(c("AUC internal  = "),Intern,sep="")
Externt <- paste(c("AUC external = "),Extern,sep="")
legend(0.3,0.6,c(Internt,Externt,"\n"),border="white",cex=1.1, box
    .col = "white", col=c("blue", "red"))
legend("bottomright",
       bty = "n",
       inset=0.1,
       cex = 1,
       title="",
       c("Internal","External"),
       horiz=FALSE,
       lty=c(1,1,1),
       lwd=c(2,2,2),
       col=c("blue","red"),
       bg="grey96")


#brier score
brier = (1/length(p_mode2$external))*sum((p_mode2$external-p_mode2
    $trueY)^2)
```

E R code Bootstrap

```
mydata2<- read.table(file="Master/newdataCP_new.txt", header =
    TRUE)
library(glmnet)

B=1000
#choose decision rule
type = dev
lambdaMin = TRUE
nfolds = 20

CP = as.factor(mydata2$CP)
Country = as.factor(mydata2$Continent)
C_sd = mydata2$com_sd
C_mean = mydata2$com_mean
Q_mean = mydata2$qom_mean
Q_sd = mydata2$qom_sd
A_mean = mydata2$aom_mean
A_sd = mydata2$aom_sd
H_mean = mydata2$hom_mean
H_sd = mydata2$hom_sd
W_mean = mydata2$wom_mean
W_sd = mydata2$wom_sd
C_x_mean = mydata2$com_x_mean
C_x_sd = mydata2$com_x_sd
C_y_mean = mydata2$com_y_mean
C_y_sd = mydata2$com_y_sd
# clinical variables
sex = as.factor(mydata2$SEX)
below1000g = as.factor(mydata2$BELOW1000GR)
below28GA = as.factor(mydata2$BELOW28GA)
bw = mydata2$BW
GA = mydata2$GA
rec_time = mydata2$rec_time
P_TE_AGE = mydata2$P_TE_AGE
trunk_area = mydata2$Trunk_area

m2 = model.matrix(CP~Country+sex+below1000g+below28GA)[,-1]
x2<- as.matrix(data.frame(bw, rec_time, P_TE_AGE, trunk_area, Q_sd
    , A_mean, A_sd, C_x_mean, C_x_sd, C_y_mean, C_y_sd, m2))

cv <- cv.glmnet(x2, y=CP, alpha=1, family="binomial", penalty.
    factor = c(rep.int(1,11),0,0,1,1,1), standardize = TRUE,
    intercept = TRUE, nfolds = nfolds, type.measure = type)

beta = coef(cv, s="lambda.min")[,1]
lambda = cv$lambda.min
p = length(beta) #antall predictors
```

```
storeMat <- matrix(NA, ncol=(p+1), nrow=B)
colnames(storeMat) <- c(names(beta), "lambda.min")
n=dim(mydata2)[1]

#bootstrap
set.seed(124232)
for(b in 1:B){
  #draw bootstrap samples
  bsample = sample(c(1:n), replace = TRUE)
  data = mydata2[bsample,]
  CP = as.factor(data$CP)
  Country = as.factor(data$Continent)
  C_sd = data$com_sd
  C_mean = data$com_mean
  Q_mean = data$qom_mean
  Q_sd = data$qom_sd
  A_mean = data$aom_mean
  A_sd = data$aom_sd
  H_mean = data$hom_mean
  H_sd = data$hom_sd
  W_mean = data$wom_mean
  W_sd = data$wom_sd
  C_x_mean = data$com_x_mean
  C_x_sd = data$com_x_sd
  C_y_mean = data$com_y_mean
  C_y_sd = data$com_y_sd
  # clinical variables
  sex = as.factor(data$SEX)
  below1000g = as.factor(data$BELOW1000GR)
  below28GA = as.factor(data$BELOW28GA)
  bw = data$BW
  GA = data$GA
  rec_time = data$rec_time
  P_TE_AGE = data$P_TE_AGE
  trunk_area = data$Trunk_area

  m2 = model.matrix(CP~Country+sex+below1000g+below28GA)[,-1]
  x2<- as.matrix(data.frame(bw, rec_time, P_TE_AGE, trunk_area, Q_
      sd, A_mean, A_sd, C_x_mean, C_x_sd, C_y_mean, C_y_sd, m2))

  #fit model
  cv <- cv.glmnet(x2, y=CP, alpha=1, family="binomial", penalty.
      factor = c(rep.int(1,11),0,0,1,1,1), standardize = TRUE,
      intercept = TRUE, nfolds = nfolds, type.measure = type)
  if(lambdaMin == TRUE){
    bbeta = coef(cv, s="lambda.min")[,1]
    blambda = cv$lambda.min
  }
  else{
    bbeta = coef(cv, s="lambda.1se")[,1]
```

```
91      blambda = cv$lambda.1se
92    }
93    #store results
94    storeMat[b,] = c(bbeta, blambda)
95  }
96
97  saveRDS(storeMat, file="Master/bootstrapDev.rds")
```

F R code Multi sample splitting

```
data <- read.table(file="//Users/martinahall/Documents/Master/
    newdataCP_new.txt", header = TRUE)
CP = data$CP
Country = as.factor(data$Continent)
C_sd = data$com_sd
C_mean = data$com_mean
Q_mean = data$qom_mean
Q_sd = data$qom_sd
A_mean = data$aom_mean
A_sd = data$aom_sd
H_mean = data$hom_mean
H_sd = data$hom_sd
W_mean = data$wom_mean
W_sd = data$wom_sd
C_x_mean = data$com_x_mean
C_x_sd = data$com_x_sd
C_y_mean = data$com_y_mean
C_y_sd = data$com_y_sd

#clinical variables
sex = as.factor(data$SEX)
below1000g = as.factor(data$BELOW1000GR)
below28GA = as.factor(data$BELOW28GA)
below1000g28GA = as.factor(data$Below1000g_28GA)
bw = data$BW
GA = data$GA
rec_time = data$rec_time
P_TE_AGE = data$P_TE_AGE
trunk_area = data$Trunk_area

install.packages("hdi")
library("hdi")

mydata = data.frame(Country=Country, C_sd=C_sd, C_mean=C_mean, Q_
    mean=Q_mean, Q_sd=Q_sd, A_mean=A_mean, A_sd=A_sd, H_mean=H_
    mean, H_sd=H_sd, W_mean=W_mean, W_sd=W_sd, C_x_mean=C_x_mean,
    C_x_sd=C_x_sd, C_y_mean=C_y_mean, C_y_sd=C_y_sd, sex=sex,
    below1000g=below1000g, below1000g28GA=below1000g28GA,
    below28GA=below28GA, bw=bw, GA=GA, rec_time=rec_time, P_TE_AGE
    =P_TE_AGE, trunk_area=trunk_area)

# function to use the lasso method in the multi sample-split
lasso.cv.lambda.min=function (x, y, nfolds = 346, grouped = nrow(x
    ) > 3 * nfolds,
                              ...)
{
  fit.cv <- cv.glmnet(x, y, nfolds = 346, grouped = grouped,
```

```
      penalty.factor = c(rep.int(1,11), 0,0,1,1,1), type.measure =
       "dev",
                          ...)
  sel <- predict(fit.cv, type = "nonzero", s="lambda.min")
  sel[[1]]
}

#confidence intervals for logistic model
glm.ci = function (x, y, level = 0.95, ...)
{
  fit.glm <- glm(y ~ x, family = "binomial")
  confint(fit.glm, level = level)[-1, , drop = FALSE]
}

formula = CP~bw + rec_time + P_TE_AGE + trunk_area + Q_sd + A_mean
     + A_sd + C_x_mean + C_x_sd + C_y_mean + C_y_sd + factor(
    Country) + factor(sex) + factor(below1000g) + factor(below28GA
    )
mf = model.frame(formula = formula, data = mydata)
X  = model.matrix(attr(mf, "terms"), data = mf)

#multi sample splitting
hdiRes <- multi.split(x=X[,-1], y=CP, classical.fit = glm.pval,
    classical.ci = glm.ci, model.selector=lasso.cv.lambda.min,
    args.model.selector = list(family="binomial"), B=1000, return.
    nonaggr = TRUE, return.selmodels = TRUE, verbose = TRUE)

#corrected p-values
hdiRes$pval.corr
#confidence intervals
cbind(hdiRes$lci, hdiRes$uc)
```