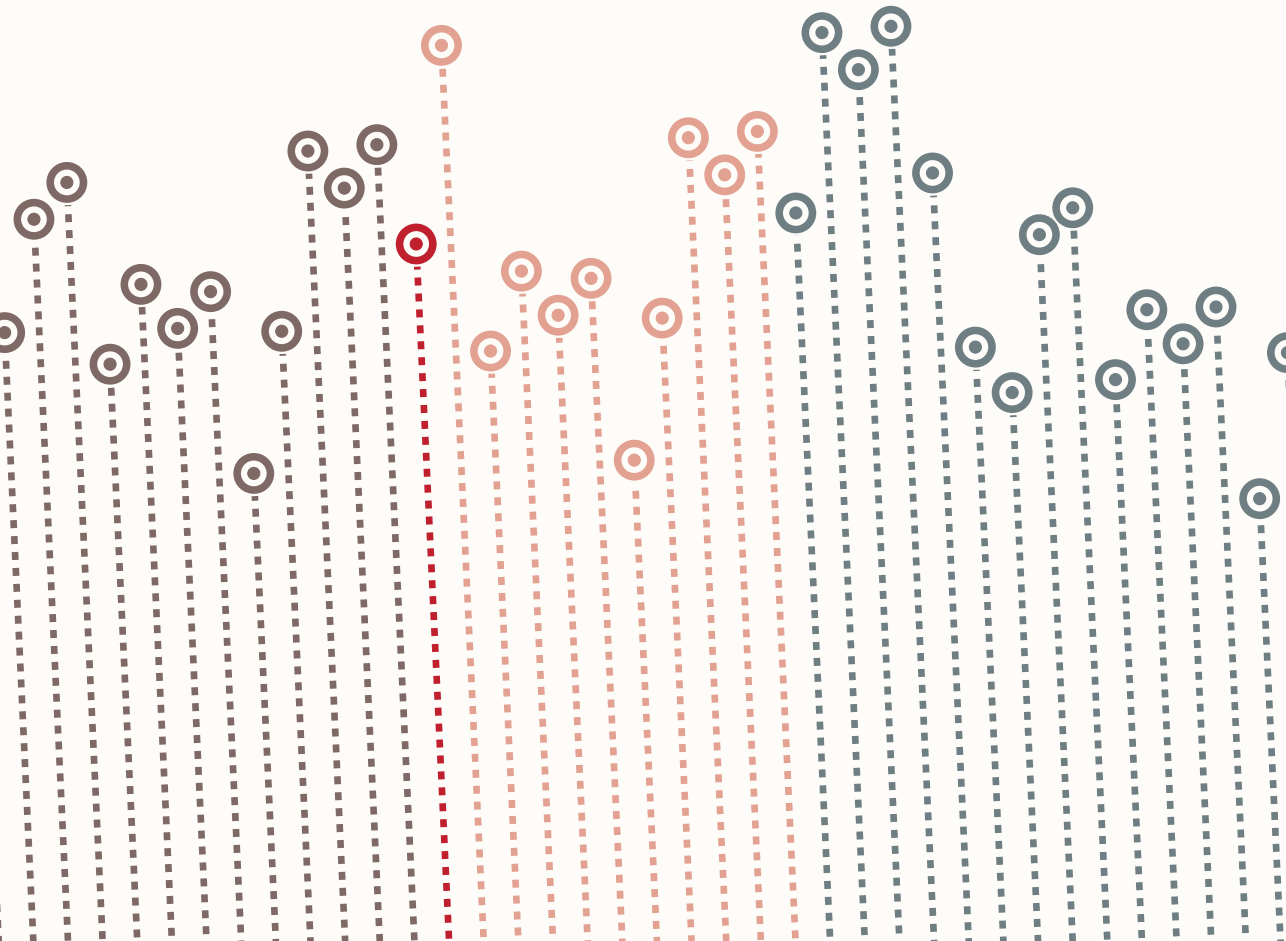
Sten P. **Willemsen**

Multivariate growth models.



Multivariate Growth Models

Sten P. Willemsen

Printed by: Ipskamp Printing (www.ipskampprinting.nl)
Cover design: Tina Siehoff

ISBN: 978-94-028-0856-8

Copyright 2017 © Sten P. Willemsen

All rights reserved. No part of this thesis may be reproduced or transmitted in any form, by any means, electronic or mechanical, without the prior written permission of the author, or when appropriate, of the publisher of the articles.

Multivariate Growth Models

Multivariate groeimodellen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
vrijdag 15 december, om 9.30 uur

door

Sten Paul Willemsen

geboren te Rotterdam

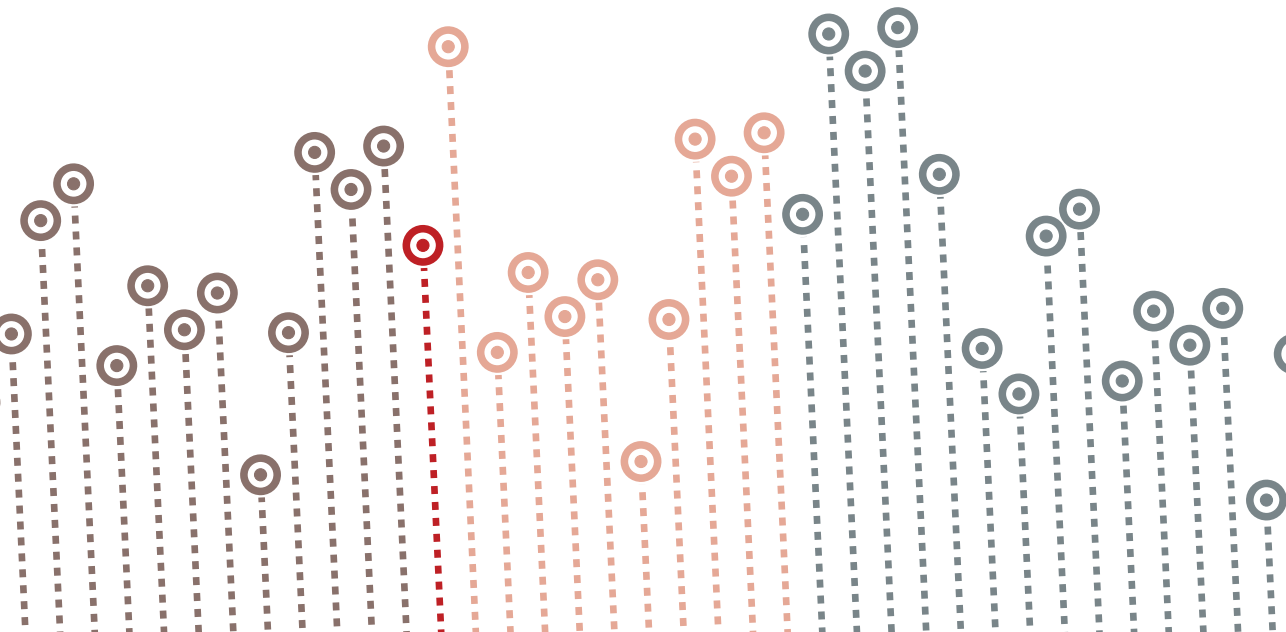
Erasmus University Rotterdam

Promotiecommissie:

Promotor:	Prof. dr. E.M.E.H. Lesaffre
Overige leden:	Prof. dr. ing. P.H.C. Eilers
	Dr. D. Rizopoulos
	Prof. dr. S. van Buuren

Contents

1 Introduction



In this chapter we present the background information for the topics of this thesis. First some general aspects of growth and growth modeling are discussed, to help the reader understand the context of what we are doing. We then explain some of the statistical techniques that we use. Finally an outline of this thesis is given.

1.1 General introduction

This thesis deals with models for growth. This, of course, is a rather generic term and therefore we will explain exactly what is meant by growth models and how the various chapters in this thesis are located within this field.

The term 'growth model' can mean both some kind of model that attempts to describe (human) growth but also more generally a set of statistical models that look at both within-individual and between individual patterns of change [*Curran et al.*, 2010; *Grimm et al.*, 2016]. In this thesis we mostly study human physical growth.

Human growth

First, when we talk about growth, we mean human growth. In human growth studies we study various anthropological measurements and their development over time. Growth is most pronounced in the period before birth where a single cell (called zygote) is transformed into a baby of 3 to 3.5 kilos in a period of approximately forty weeks. In this period the 'gestational age' is used to measure the duration of the pregnancy. To determine the gestational age we start counting from the last day of the last menstrual cycle, as the exact day of conception is usually unknown. As the last day of the last cycle is often not known, the gestational age is usually derived from the first ultrasound scan using reference charts for the crown rump length. This prenatal growth can be subdivided in the embryonic stage, which lasts until the tenth week of gestation, in which the major structures are formed, and the fetal period in which the organs continue to develop. In the postnatal period we distinguish between infancy, childhood and puberty. Immediately after birth, growth is still quite fast but then rapidly decreases until the age of 3. During childhood, the growth rate decreases further, albeit much slower, until there is a growth spurt at puberty triggered by sex

hormones. Thereafter growth decreases again until the adult height is reached. This occurs, on average, at the age of 15 for girls and 17 for boys.

The field of research on human growth is large and diverse [*Johnson*, 2015] and has a long history [*Voss*, 2001]. This interest mostly stems from the fact that a 'normal' growth pattern is seen as a measure of health while abnormal growth can be an indication of a disease. Various subtopics can be distinguished in this field: Sometimes we simply want to describe the natural variation and patterns of human growth. We may wish to make predictions for individual growth outcomes based on earlier observations. Sometimes we want to study what factors influence growth. Finally, we can ask how growth impacts other outcomes later in life. For example, there is evidence that intrauterine growth restriction can cause a multitude of problems at adult ages. This is called the 'fetal origin hypothesis' [*Barker*, 1998].

Growth models in general

While we focused on human physical growth, research on growth occurs in many areas: growths of plants, growth of an economy, etc. Many of the challenges we have seen in the analysis of human growth data are also encountered in other application fields where we also want to compare measurements with a reference or where we also have repeated measurements of some kind. Sometimes the term growth model is even used for any model that deals with any score that changes over time and *the* growth curve model refers to statistical technique (a generalized MANOVA) to model these data [*Potthoff and Roy*, 1964]. In this thesis we also will show an application of a model originally developed for human growth applied on the concentration of a drug in the bloodstream.

1.2 Statistical methods for growth studies

Regression models for growth

Modeling human growth has a long history. Some studies have been based on cross-sectional measurements from individuals with various ages but most research focuses on the growth of a specific individual repeatedly observed in time [*Scammon*, 1927]. Various authors have looked for the most appropriate

mathematical functions that fit the observed measurements best. This resulted, for example, in the Jenss model [*Jenss and Bayley*, 1937], often used for height and weight in children. Most of these models are rather complex and highly nonlinear in their parameters. In the past these growth models were fitted for each individual child separately. The subject-specific parameters obtained were summarized further, e.g.: by calculating the average and standard deviation.

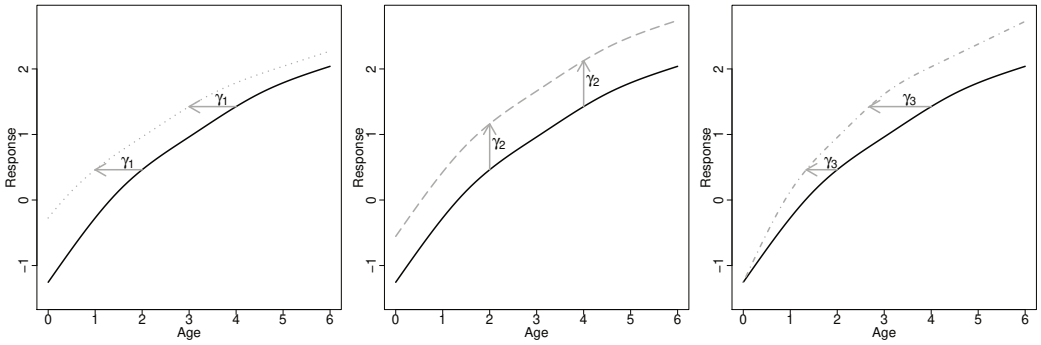
Splines

Some growth models, as discussed above, are highly non-linear and may be difficult to fit to data. It may therefore be useful to look for alternative models that are flexible enough to accommodate the non-linear patterns in the data. One option is to make use of (B-)splines [*De Boor*, 2001]. They are piecewise polynomial functions of low order that are joined together in such a way that they look smooth. The positions where the pieces come together are called knots. Often restrictions are placed on a spline function so that it is linear at the edge of its support. In this case it is called a restricted or natural spline. Because splines are meant to be very flexible there is a danger of over-fitting, especially when the number of knots is large. This can be remedied by placing penalties on the spline coefficients. The result is called a P-spline [*Eilers and Marx*, 1996].

Mixed models

Fitting a model for each individual combined with calculating summary statistics, as described above, is not entirely satisfactory. For example, this approach is not optimal when the study is plagued with missing data. A much better method is formed by mixed effects models [*Laird and Ware*, 1982]. In these types of models we simultaneously model the general pattern of development (formed by the fixed effects) and the way the individuals deviate from this general pattern (which is described by the random effects). In the simplest case the model is linear in all parameters leading to linear mixed effects models. However often nonlinear models are needed then resulting in nonlinear mixed effects models.

A specific type of nonlinear mixed effects model is the SITAR model [*Cole et al.*, 2010b], which will feature prominently in several of the following chapters. This uses a template longitudinal pattern which is described by a spline function.



(a) Horizontal shift (b) Vertical shift (c) Horizontal Stretch

Figure 1.1: Schematic representation of SITAR model with horizontal shift (γ_{i1}), vertical shift (γ_{i2}) and stretch (γ_{i3})

The individual growth profiles are then derived from this template by shifting the curve along the horizontal and vertical axis and by stretching it (see Figure 1.1).

1.3 Multivariate models

Growth is intrinsically multivariate. We mean that a single dimension does not give the total pictue of growth but the proportionality between them has to be considered. The standard approach to growth modeling is however to look at one outcome at a time. This may not always be optimal. For example,an unusual observation might appear normal when examined univariately on its own, but might still be outlying when looked at the total picture [*Gnanadesikan and Kettenring*, 1972]. It is only when we look at the relations between the different outcomes when this becomes apparent. Therefore, we argue that growth should be considered as a multivariate process. Most existing techniques can only handle a single response so here methodological development is needed.

Quantiles, expectiles and standard deviation scores

The methods above focus on the expected value of some random variable. We might also be interested in the extremes of a distribution. We then ask the question

how far away an observation is from the center and the range of values that can still be considered to be within the normal range.

The most familiar example to most people, or at least most parents, are the growth charts for height and weight which are used to check if young children develop normally. Similar growth charts are also used for embryos, for example to a predict the gestational age based on its size. Here often quantiles are used. A quantile says what fraction of observations in the population are expected to be lower. Quantiles expressed as a percentage are called percentiles or centiles. An other generalization that is often used, is formed by standard deviation scores (sds-scores) telling us how many standard deviations a measurement is away from its mean.

One method to compute such quantities is by assuming some parametric distribution. We can then estimate its parameters and derive the quantiles from there. There are however also non-parametric methods that allow for direct computation without the assumption of a specific parametric form.

To reduce the variation and make a comparison with respect to a meaningful peer-group, quantiles are usually conditioned on age and often on sex. In practice, in a data set the children do not have exactly the same age and we need some modeling to estimate the relation between the ages and the quantiles. Often quantile regression is used for this [*Koenker*, 2005]. This method minimizes the weighed absolute residuals, usually with some linear programming technique. There are alternatives to this non-parametric technique that are used to estimate quantiles based on the whole distribution. In the LMS method [*Cole*, 1990] its is assumed that the distribution can be transformed to a normal one after a Box-Cox transformation. Now the parameters of this transformation as well as the mean and the coefficient of variation of the resulting normal distribution are modeled by means of a spline function of age; this results in the so called L (transformation), M (location) and S (scale) curves from which the method derives its name. The generalized additive models for location scale and shape (gamlss) are a generalization of the LMS method that allow us to use many more families of distributions.

Expectiles constitute an interesting alternative to quantiles [*Newey and Powell*, 1987]. They generalize of the mean in the same way that the quantiles generalize the median. The loss is now based on weighted squared residuals instead of weighed absolute residuals. Expectile regression is carried out using the least

asymmetrically weighed squares algorithm (LAWS). They are much easier to compute than quantiles as least squares techniques can be used.

1.4 The Bayesian approach

The Bayesian approach is the basis for several modeling techniques in this thesis. It combines prior believes with the observed data resulting in a posterior distribution. Such a posterior represents our believes after having observed all evidence.

If we want to estimate the parameters in a model we need to be able to find the posterior. A closed-form solution only exists in some special cases for very basic models. Usually a numerical procedure has to be followed to sample from the posterior. Often this is done using Markov Chain Monte Carlo (MCMC) techniques [*Gelfand and Smith*, 1990]. This is a class of models in which the posterior is approached by a sample on which further inference may be based. The sample is drawn iteratively with each value depending on the previous one. As the sample gets larger the approximation to the posterior gets better. Because, unlike the sample of individuals, the sample from the posterior can theoretically be made arbitrarily large when given enough time, one can in principle obtain any precision deemed necessary. In complex models the type of MCMC-method is often the Metropolis-Hastings algorithm or Gibbs sampling [*Geman and Geman*, 1984]. In each iteration of the Metropolis-Hastings algorithm we draw a sample from some proposal distribution; we then either accept or reject this value with some probability that is based on the ratio between the value of the distribution in the proposed point and the value from the previous iteration [*Hastings*, 1970]. When the proposal is rejected we continue to use the previously accepted value. The idea behind Gibbs sampling is that, given same starting values, we can sample from a multivariate distribution by sampling from each of the parameters in turn conditioning on all others (the full conditionals). Eventually the distribution of the sample should converge to the target distribution which is the posterior. There are various software packages that have implemented Gibbs sampling, e.g.: WinBUGS, OpenBUGS and JAGS. The usefulness of Gibbs sampling is based on the fact that generally it is much easier to sample from the full conditionals than it is to sample from the posterior directly. For the full conditionals, often there is even a closed form solution when so called conditional conjugate priors are chosen. When this is not the case other sampling techniques such as Metropolis-Hastings

or a slice sampler can be used. In the slice sampler we first sample a value of the distribution y which is below $f(x')$ where x' is the current value of x. In the next step we uniformly sample a new value of x from all values of for which $f(x)$ is at least equal to y [*Neal*, 2003]. When we use MCMC we will (given some conditions) eventually sample from the posterior but it is not clear how long this will take. Furthermore the sample obtained from an MCMC method is generally correlated. This makes it important to check convergence and asses the effective size of the posterior sample.

1.5 Model evaluation

A simple way to look at model performance is to evaluate the discrepancy between the observed and predicted values. The Root Mean Squared Error (RMSE), which is the standard error of the residuals can be used to this end. However the RMSE is optimistic because models with a lot of parameters always perform better on the estimation sample than on new data. This optimistic behavior can be compensated for by cross-validation: in this technique the data is split into subsets and the residuals of the observations in each subset are based on a model developed on the other subsets. These residuals are then used to calculate the cross-validated RMSE. Because cross-validation involves fitting several models, it is computer intensive.

An other way to evaluate models is using Akaike Information Criterion (AIC). Its is defined as $-2\log(f(y|\hat{\theta})) + 2k$ where y are the observations and $\hat{\theta}$ the maximum likelihood estimator of the parameters vector θ, $f(.)$ the likelihood function of the model considered and k the effective degrees of freedom. In simple models this is equal to the number of parameters. Expressed in words this criterion penalizes the fit of the model (as measured by the log likelihood function) by the model complexity (expressed by the number of parameters). In a Bayesian hierarchical model the number of parameters is not so easy to determine. The Deviance Information Criterion (DIC) was proposed as a substitute for the AIC in a Bayesian setting. The DIC is defined as $\bar{D} + p_D$ with p_D given by $\bar{D} - D(\bar{\theta})$ where the deviance D is defined as $-2\log(f(y|\theta))$, $\bar{D}$ is the average deviance in the posterior samples and $D(\bar{\theta})$ is the deviance calculated at the posterior mode. The effective number of parameters p_D can be seen as a measure of the flexibility of

the model. The DIC may also be expressed as $D(\bar{\theta}) + 2p_D$ making the resemblance with the AIC more clear.

1.6 Outline of this thesis

In this section we introduce the different chapters of this thesis and show how they are situated within the landscape outlined above.

Chapter 2: Modeling height for children born small for gestational age treated with growth hormone

In this chapter two approaches of modeling the height of children are compared. The first is on the original scale, often encountered in the statistical literature. The second approach to modeling is on the scale of the standard deviation scores, which is very common in endocrinology. Using data from a trial that evaluated the effect of growth hormone in children that are born small for gestational age and remain too small later in childhood, we show that the second approach may involve more complex modeling and hence a worse model fit.

Chapter 3: A multivariate Bayesian model for embryonic growth

In this chapter, we generalize the SITAR model in a Bayesian way to multiple dimensions. This multivariate SITAR (MSITAR) model is used to create multivariate reference regions.

The model is illustrated using longitudinal measurements of embryonic growth obtained in the first semester of pregnancy, collected in the Rotterdam Predict study. In addition, we demonstrate how the model can be used to find determinants of embryonic growth.

Chapter 4: Flexible multivariate nonlinear models for bioequivalence problems

In this chapter we modify the MSITAR model that was introduced in chapter 3 to another setting, a bioequivalence study. In this type of study we want to demonstrate that the rate and extend at which the active component of a two drug

formulations becomes available to the body does not differ in an important way. We demonstrate that this model can also be of use to model the concentration of a drug in the bloodstream. Usually this is done using compartment models, which are effectively a type of highly nonlinear mixed-effects models when we the heterogeneity between subjects is taken into account. The MSITAR model as well as a compartment model are estimated on a real data set containing the concentration of a test and reference formulation that was measured in the blood of the antihypertensive drug Losartan. We focus on contrasting the average and individual bioequivalence .

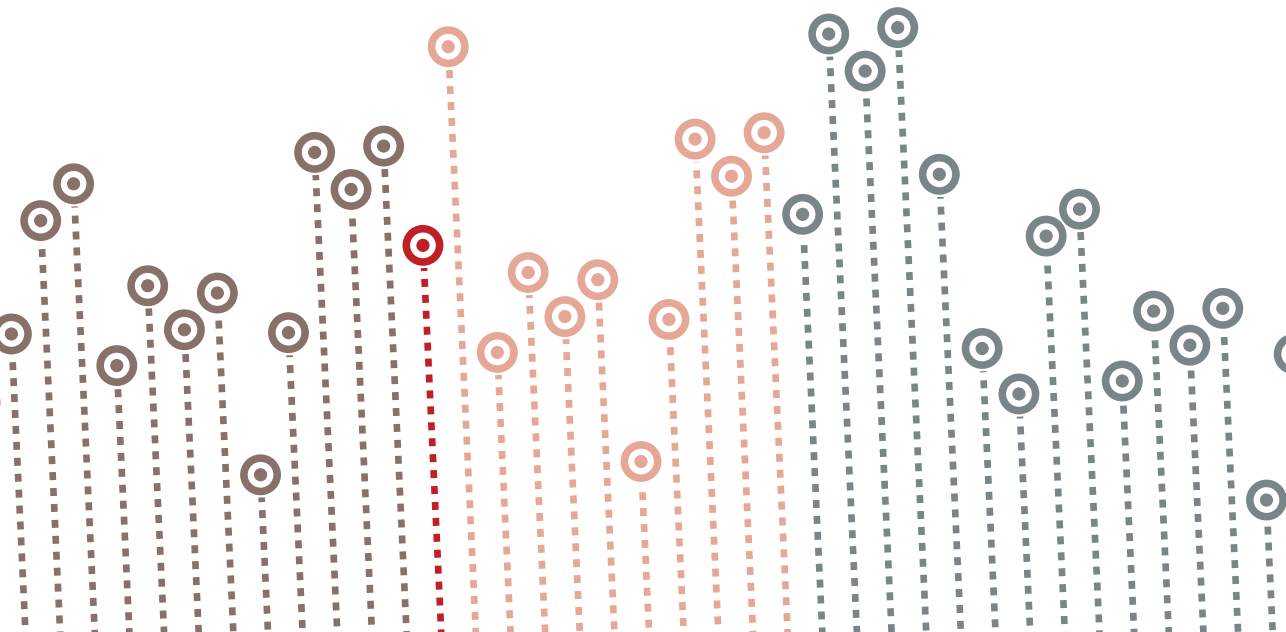
Chapter 5: Tubular expectile contours

There have been various attempts to generalize quantiles to multiple dimensions. One idea (that we will adopt) is to reduce a bivariate problem into a univariate one by projecting the data on a line through the center. When this is done for multiple directions a reference contour is created. Here we will try and integrate these bivariate reference regions with the idea of making them depend on a third variable, thus creating tubular quantile and expectile regions.

Chapter 6: Conclusions and outlook

In this chapter we summarize the main findings from the previous chapters. in addition we have some proposals for further research.

2 Modelling height for children born small for gestational age treated with growth hormone



Abstract:

The analysis of growth curves of children can be done on either the original scale or in standard deviation scores (SDS). The first approach is found in many statistical textbooks, while the second approach is common in endocrinology, for instance in the evaluation of the effect of growth hormone in children that are born small for gestational age (SGA) that remain small later in childhood. We illustrate here that the second approach may involve more complex modeling and hence a worse model fit.

Authors: S.P. Willemsen, M. De Ridder, P.H.C. Eilers, A. Hokken-Koelega, & E. Lesaffre. published in: Statistical methods in medical research (2014)

2.1 Introduction

Children that are born small for gestational age (SGA) and do not show catch-up growth are eligible for treatment with a growth hormone (GH). Although there is some concern that this growth hormone might adversely affect the metabolism, generally the treatment is considered to be safe and effective [*De Zegher et al.*, 2006; *Hokken-Koelega et al.*, 2004; *Kamboj*, 2005]. However, since the cost of treatment with growth hormone is relatively high [*Allen*, 2006], it is of practical interest to know the dose that achieves optimum growth. For children under treatment it is important to predict their adult height in order to avoid unrealistic expectations.

In this paper, we focus on SGA children treated with growth hormone and consider (1) modeling of height as a function of age and (2) the prediction of the adult height. We note that in the endocrinology literature modeling the height is usually done in terms of the standard deviation score (SDS), see e.g. [*Boguszewski et al.*, 1998; *Cooper et al.*, 2001; *Dalton et al.*, 2003; *De Zegher et al.*, 2000; *Hypponen et al.*, 2000; *Kamp et al.*, 2002; *Karaolis-Danckert et al.*, 2006; *Kjaergaard et al.*, 2002; *Nguyen et al.*, 2004; *Ong et al.*, 2002; *Rotteveel et al.*, 2008]. There is only a minority of studies that model the untransformed height [*Davenport et al.*, 2002]. A few studies are based on a combination of both scales [*Bakker et al.*, 2006; *Zemel et al.*, 2002]. In order to model the growth curve of height y in SDS, for each age the mean and standard deviation of y need to be established in the reference population. That is, at age t one determines $\mu(t, s)$ and $\sigma(t, s)$, which are the mean and standard deviation of y respectively, in the reference population at age t and for sex s. A variety of techniques can be employed to determine $\mu(t, s)$ and $\sigma(t, s)$ of which the most prominent seems to be Cole's LMS method, see [*Cole and Green*, 1992; *Wright and Royston*, 1999]. For most of these techniques some kind of smoothing is involved. For a new child with height y_0 determined at age t_0, the height can be expressed in SDS by computing $\frac{y_0 - \mu(t_0, s_0)}{\sigma(t_0, s_0)}$ with s_0 the gender of that child. Note that the reference population most often consists of children with no diagnosed growth retardation problems [*Fredriks et al.*, 2000; *Van Wieringen and Verbrugge*, 1966]. When the distribution of $y_{t,s}$ is not normal, computation of the SDS is slightly more complicated, involving a transformation of the raw scores. Here we will assume that height has a normal distribution conditional on the covariates (as a conditional normal distribution is often used in practice to model height and seems to work quite well.)

Presenting the height in SDS allows to conclude readily how remote the child's height is from its expected height. Further, for a child belonging to the reference population, the expected growth curve in SDS is more or less a horizontal line before it reaches puberty, a phenomenon called 'tracking' [*Foulkes and Davies*, 1981]. Expressing height on the SDS scale aims to produce profiles that are easy to model. However, for children with an abnormal growth pattern, such as SGA children, the SDS profile is typically (highly) non-linear. Thus, a modeling exercise examining factors that influence growth in SGA children will typically be more complex on the SDS scale than on the height scale. This might adversely affect the power of the study. Here we illustrate, using data of a double blinded randomized trial in SGA children conducted in the Netherlands, that it could be preferable to model the height on the original scale.

To model the growth curves, we used a Bayesian approach with WinBUGS (version 1.4.3). The advantage of Markov Chain Monte Carlo (MCMC) techniques over classical non-linear mixed effects maximum likelihood estimation are (1) less dependence of the modeling exercise on the starting values of the parameters; (2) more flexibility with respect to distributional assumptions for the random effects. Further, the Bayesian approach has two additional advantages: (3) taking into account all uncertainty involving the estimation of model parameters when predicting characteristics of individual growth curves and (4) the ability to include prior information into the sampling algorithm. Moreover, (5) Bayesian statistical inference does not need to be based on large sample arguments but is simply derived from the posterior distribution of the parameters.

The paper is organized is follows: Section 2 describes the study and the growth data. In Section 3 we introduce the class of models considered in this exercise. In Section 4 we apply the models to the data and evaluate the models with regard to their prediction characteristics. We conclude the paper with a discussion.

2.2 The Dutch growth hormone trial data

The Dutch growth hormone (DGH) trial [*Sas et al.*, 1999; *Van Pareren et al.*, 2003] was the first randomized double blind trial in the Netherlands that investigated the effect of the dose of growth hormone treatment on the growth of children that were born SGA without spontaneous catch-up growth (so they were still too small at the start if the trial). In this study, SGA is defined as having a birth length

below the third percentile of the normal Dutch population at birth. A number of factors can cause a child to be SGA such as: intrauterine exposure to toxins, infections, chromosomal disorders, congenital defects, metabolic diseases and genetic disorders, the use of tobacco or alcohol by the mother, certain diseases of the mother or certain factors jeopardizing placental perfusion. Often, however, the exact underlying cause remains unclear [*Johnston and Savage*, 2004]. SGA children may show catch-up growth, i.e. while they are initially small, often they grow faster than other children of the same gestational age such that within their first year of age their height lies within the 95% normal range. However, some SGA children experience no catch-up growth and remain too small for their age [*Hokken-Koelinga et al.*, 1995]. These children may experience several medical problems later in life such as: an increased risk of developing hypertension, dyslipidemia, impaired glucose tolerance or diabetes mellitus [*Lee et al.*, 2003].

The DGH trial started in 1991. One group of children was treated with 1 mg GH/(m^2 day) (low dose) and the other with 2 mg /(m^2 day) (high dose) of a synthetic growth hormone (Norditropin®). Inclusion criteria were: boys between three and eleven years of age or girls between three and nine years of age with a birth length below -1.88 SDS, a height below -1.88 SDS at the start of the trial (the SDS scores are computed using the reference values by *Roede and Van Wieringen* [1985]), no spontaneous catch-up growth in the last year, pre-pubertal stage and an uncomplicated neonatal period without severe asphyxia, sepsis or lung problems. Children with endocrine or metabolic disorders, chromosomal disorders or growth failure caused by other disorders and syndromes as well as children that used drugs that could interfere with growth hormones were excluded. Treatment should be stopped as soon as the child reaches adult height. For this purpose adult height was defined as the height of a child at the measurement at which the growth over the previous six months was less than half a centimeter.

In total 79 children were included in the trial, 52 boys and 27 girls. The average age at the start of treatment was 7 years and 3 months. Nine children dropped out of the study before reaching adult height: 4 because of lack of motivation, 3 children stopped after they moved, 1 stopped because of signs of precocious puberty and 1 because of signs of growth hormone insensitivity. For 51 of the 70 children who started growth hormone treatment, an extra height measurement was taken about five years after the treatment was discontinued. This measurement can sometimes be a little higher than the last regular measurement. No

further measurements where taken after this extra measurement, because the investigators were convinced that adult height was reached.

The DGH data have been analyzed before. *Sas et al.* [1999] looked at the height of the children in the five years after the start of this study by comparing the average height SDSs between the treatment groups and with the previous year in all of the first five years of follow up by means of t-tests. The height SDSs were also compared with -1.88 (the 3rd percentile) to see whether the children were within the normal range. The authors concluded that almost all children were within the normal range after five years of GH treatment. Further, there was no significant difference between the average height SDS of the two dose groups at the end of the five year treatment period.

Van Pareren et al. [2003] looked at the 54 children that had reached adult height at the time of their exploration. The average adult height SDS in each treatment arm was compared with the average height SDS before study, and with zero (to see whether the mean adult heights are the same as in the reference population). The mean adult height SDSs were also compared between treatment arms. All comparisons were done using t-tests. In both groups there was an increase in average height SDS but not significantly different between the treatment arms. In the high dose group the average adult height SDS did not differ from zero.

The aims of *Sas et al.* [1999] and *Van Pareren et al.* [2003] differ from the aims in this paper since we want to make use of the whole growth curve to compare treatment doses. In other words we wish to build a growth curve model for the heights.

In Figure 2.1 the height profiles of the children participating in the DGH trial are displayed. The average height in the reference population has been added as a benchmark. When expressed in SDS the profiles look as in Figure 2.2.

2.3 Statistical models for the height profiles

We consider two growth curve approaches to model height as a function of age: (1) the untransformed height (in centimeters) and (2) in SDS. All modeling was done by calling WinBUGS from R using the R2WinBUGS package.

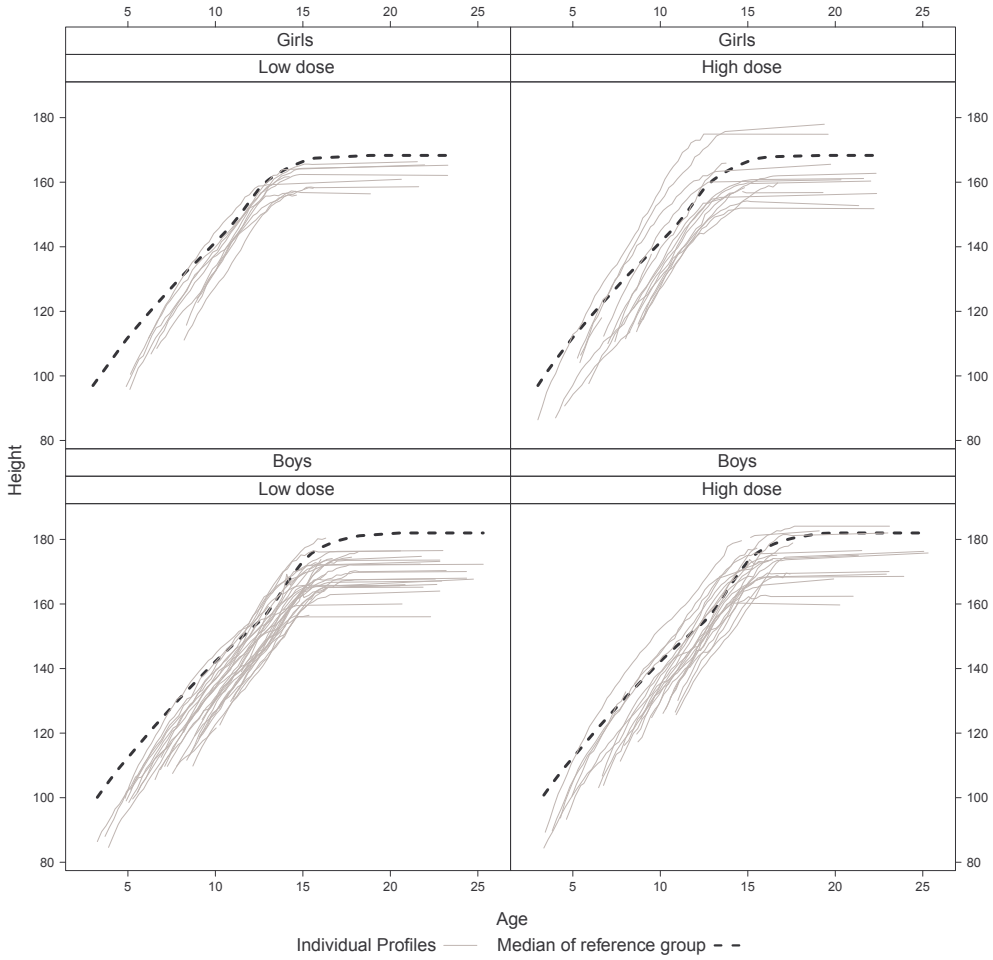


Figure 2.1: DGH trial: height profiles in cm split up according to gender and dose administration

Modeling the height in centimeters

Figure 2.1 suggests a suitable structure for a model of height as a function of age. The height of each child follows an almost linear pattern until the adult height is reached. On closer inspection, the linearity of the initial part of the growth curve can be improved upon by a square root transformation of age (as confirmed by DIC). Hereby the individual growth curves appear approximately as broken

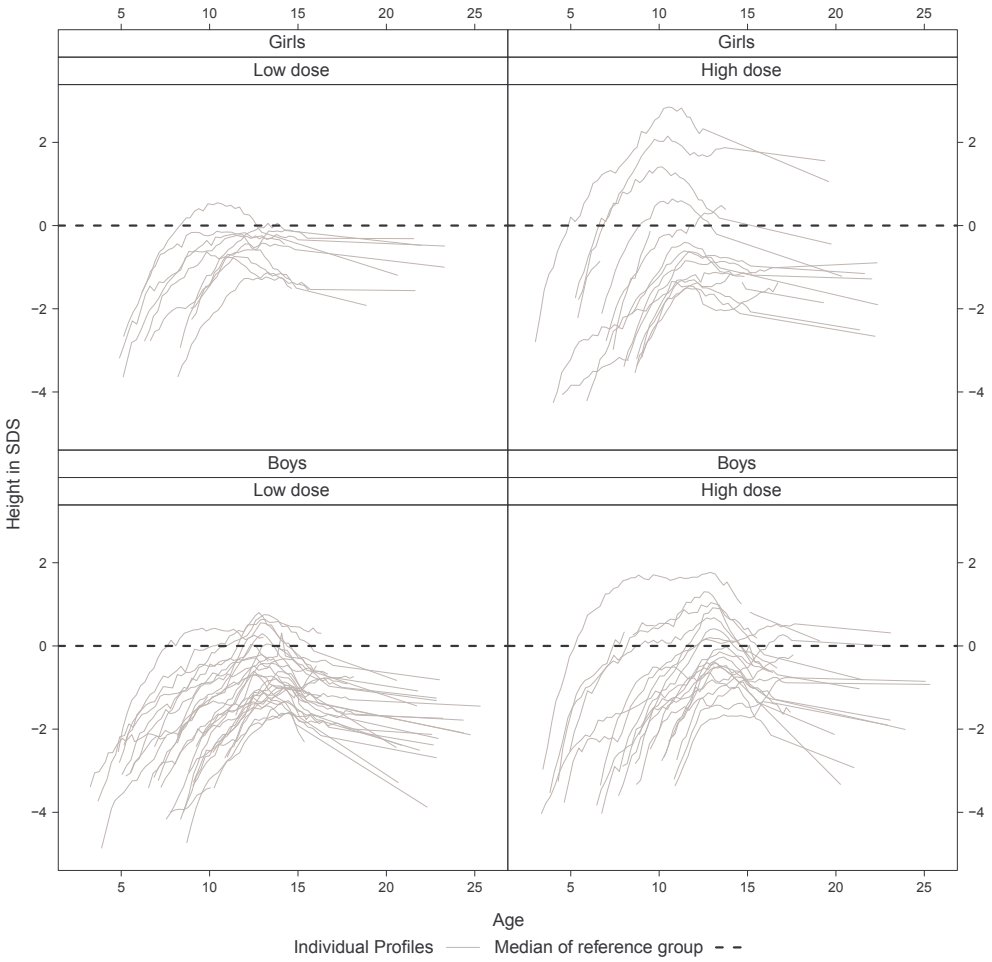


Figure 2.2: DGH trial: height profiles in SDS split up according to gender and dose administration

lines. For this reason, we assumed that the child's height can be described by an initial phase of linear growth (in terms of the square root of age) until maturity is reached at which growth stops and a horizontal line appears. Hence we assumed that an individual's growth curve can be described by an intercept, the slope of the initial phase and the age at which growth stops. Further, we assumed that these three subject specific coefficients have a trivariate normal distribution with

a mean possibly depending on sex, treatment dose, age at start of the treatment and the target height SDS (TH) of the child. The target height is the midparental height, i.e. the average height of the parents, corrected for the sex of the child and corrected for the secular growth trend (meaning that on average each generation is taller than the previous one). To this model measurement error is added. Our model for height now looks as follows:

$$\begin{aligned} y_{ij} &= b_{i1} + b_{i2}\min(t_{ij} - b_{i3}, 0) + \epsilon_{ij}, \\ \boldsymbol{b}_i &= (b_{i1}, b_{i2}, b_{i3})^\intercal \sim \mathrm{N}(\boldsymbol{\mu}_i, \Sigma_b), \\ \mu_{ik} &= \alpha_{0k} + \alpha_{1k}\mathrm{SEX}_i + \alpha_{2k}\mathrm{DOSE}_i + \\ &\qquad \alpha_{3k}\mathrm{SAGE}_i + \alpha_{4k}\mathrm{TH}_i \ (k = 1, 2, 3), \\ \epsilon_{ij} &\sim \mathrm{N}(0, \sigma^2_\epsilon). \end{aligned} \tag{2.1}$$

In (2.1) y_{ij} is the height of the ith child at age_{ij}, $(j = 1, \ldots, n_i; i = 1, \ldots, n)$, the age is transformed $t_{ij} = \sqrt{\mathrm{age}_{ij}} - \sqrt{20}$. Where 20 is the age where all children reach adult height. b_{i1}, b_{i2} and b_{i3} are the subject-specific parameters of the growth curve of the ith child, and indicate respectively the adult height, slope and change point for the ith child. The vector $\boldsymbol{b}_i$ is assumed to have a trivariate normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_b. We assume that μ_{ik}, the kth component of $\boldsymbol{\mu}_i$, is a linear function of covariates with weights α_{hk}, with $h = 0, \ldots, 4$. The measurement error ϵ_{ij} is assumed to follow a normal distribution with zero mean and variance σ^2_ϵ. Dose is coded as -1 for the low dose and 1 for a high dose, sex is coded -1 for boys and 1 for girls. SAGE_i is the age of entry age_{i0}, approximately centered by subtracting 7 from it. By centering the covariates, we obtained better convergence properties for the MCMC chains [*Gilks and Roberts*, 1996].

Vague and conjugate priors were taken for all fixed effect parameters i.e. independent normal priors with zero mean and large (10^6) variance for the αs, an inverse Wishart prior with 3 degrees of freedom and an identity scale matrix for the covariance matrix of the random effect parameters b and an inverse gamma prior with rate 0.001 and mean 0.001 for the residual variance σ^2_ϵ.

Modeling height on the SDS scale

The growth curves in SDS show a more complex behavior than those on the original scale. From the profiles in Figure 2.2 we observe that initially the SGA

children under growth hormone treatment grow much faster than normal children of the same age, so their SDS increase. But afterwards the growth rate (in SDS) slows down. Eventually, the height SDS even decreases, because the children stop growing at an age where some children in the reference population still continue to grow.

Because we do not see an obvious structure in the SDS profiles, the mean of the SDS curves is modeled with a quadratic spline [*Ruppert et al.*, 2003] augmented with subject specific random intercepts, random slopes and a random quadratic part. In this way the number of random terms is the same as for the change point model 2.1. Furthermore, the SDS-growth curves have a quadratic appearance so a quadratic function seems a logical choice when we want to model the growth curves with splines with a low number of knots. We have defined t_{ij} as the square root of age$_{ij}$ after standardization. The number of knots can be increased to make the model more flexible. We included the same covariates as those for modeling the height in centimeters. So we have the following model:

$$\begin{aligned} y_{ij}^S &= c_{i1} + c_{i2}t_{ij} + c_{i3}t_{ij}^2 + \sum_{k=1}^{K} \gamma_k (t_{ij} - \kappa_k)_+^2 + \eta_{ij}, \\ \boldsymbol{c}_i &= (c_{i1}, c_{i2}, c_{i3})^\intercal \sim \mathrm{N}(\boldsymbol{\nu}_i, \Sigma_c), \\ \nu_{ik} &= \beta_{0k} + \beta_{1k}\mathrm{SEX}_i + \beta_{2k}\mathrm{DOSE}_i + \\ &\quad \beta_{3k}\mathrm{SAGE}_i + \beta_{4k}\mathrm{TH}_i \ (k = 1, 2, 3), \\ \eta_{ij} &\sim \mathrm{N}(0, \sigma_\eta^2). \end{aligned} \tag{2.2}$$

y_{ij}^S is the height of the ith child at age$_{ij}$ expressed as SDS with the general Dutch population stratified by sex as a reference [*Roede and Van Wieringen*, 1985]. The coefficients c_{i1}, c_{i2} and c_{i3} denote respectively the intercept, slope and quadratic term for the ith child that are added to the general pattern. The γs are coefficients of the spline function that do not vary across individuals. The notation, $(x)_+^2$ indicates x^2 when $x > 0$ and 0 otherwise. The κ_ks are the knots of the spline. They are not estimated but fixed and placed at equally positioned quantiles. The vector $\boldsymbol{c}_i$ is assumed to have a trivariate normal distribution with mean $\boldsymbol{\nu}_i$ and covariance matrix Σ_c. The mean is allowed to depend linearly on the covariates through the parameters β_{hk} ($h = 0, \ldots, 4$). Measurement error is denoted by η_{ij} and is assumed to follow a normal distribution with mean 0 and variance σ_η^2. Independent normal priors with zero mean variances of 10^6 were used for the γs.

The same vague and conjugate priors were used as for model 2.1 for the other parameters.

Model comparison

The two longitudinal models were compared in two ways. We compared the fit of the models by the Deviance Information Criterion (DIC). In addition we compare the predictive ability using the root mean squared error (RMSE) in prediction. DIC is a Bayesian generalization of AIC [*Spiegelhalter et al.*, 2002a] and is defined as $\hat{D} + 2p_D$, where D is the deviance and p_D is a number that represents the effective number of parameters in the model. The deviance in turn is defined as minus two times the log likelihood. By $\hat{D}$ we indicate that the deviance is evaluated at the posterior mean of the parameters. p_D is defined as $\bar{D} - \hat{D}$, the posterior mean of the deviance minus the deviance evaluated at the posterior mean of the parameters.

A comparison of the DICs of the two growth curve models involves the Jacobian since the response differs in the models (or likelihoods). To make the DICs comparable, we transformed the DIC of model 2.2 to the original height scale with $y_{ij}^S = \frac{y_{ij} - \mu(t_{ij}, s_i)}{\sigma(t_{ij}, s_i)}$, hence:

$$\log[L_S](\cdot) = \log[L_0](\cdot) + \sum_{i}^{n} \sum_{j=1}^{n_i} \log[\sigma(t_{ij}, s_i)], \tag{2.3}$$

with $\log[L_0](\cdot)$ the loglikelihood on the height scale and $\log[L_S](\cdot)$ the log-likelihood on the SDS scale both under model 2.1.

We also compared the models on their predictive ability of the adult height using the observations before the age of twelve. This comparison was done using RMSE, defined as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i,A} - \hat{y}_{i,A})^2}$ where $y_{i,A}$ denotes the observed adult height of the ith child. We defined this as the last measurement we observed. $\hat{y}_{i,A}$ denotes the posterior mean of the adult height of the ith child based on the observations of the child before the age of twelve. In model 2.1 we defined the adult height to be the plateau height that is reached after the change point. In model 2.2 the adult height was taken to be the height at the age of twenty. To generate the predictions for the ith child cross validation was applied using a separate run in WinBUGS for each child. In model 2.1 we used the data of all the children except the ith (the one we want to predict) to sample parameter values $\boldsymbol{\alpha}$,

σ_α^2, Σ_b and $\boldsymbol{b}_i$ (the random effect parameters for all children except the ith). In each iteration of the MCMC algorithm we also sampled the values of the random effects of the ith child, $\boldsymbol{b}_i$, using both the sampled values of the fixed effects parameters and the observed height measurements of the ith child that are taken before the age of twelve. To prevent the observations of the ith child from influencing other parameters than $\boldsymbol{b}_i$ the 'cut' function in WinBUGS was used [*Spiegelhalter et al.*, 2003]. For each sampled value of b_i we also computed the change point. For model 2.2 we computed the RMSE in an analogous fashion.

It would also be interesting to predict not only the adult height itself but also the age on which it is reached. However because model 2.2 does not give such a value we could not perform such a comparison.

2.4 Results

For model 2.1, three initially dispersed chains of 30,000 samples were taken from the posterior after a burn in of the same length. For model 2.2 we took three dispersed chains of 60,000 samples after a burn-in of 60,000. Convergence was checked by visual inspection of the trace plots and the Brooks-Gelman-Rubin diagnostic plots [*Gelman and Rubin*, 1992].

To check the assumptions of the models we performed a number of diagnostics checks. QQ-plots where produced to check the normality of the random effects based on their posterior means. Further, we compared fitted and observed profiles and plotted the residuals as a function of age. The assumptions seemed to hold reasonably well.

Parameter estimates

Parameter estimates (posterior means) and 95% credible intervals for the parameters of model 2.1 are presented in Table 2.1. To aid interpretation the coefficients are printed in bold whenever the credible interval includes 0. The dose has a positive effect on the speed of growth, however the adult height was not significantly different. Quite remarkably, the children that start the therapy at a later age showed a faster growth and seemed to arrive at a comparable adult height as children that start earlier. This might indicate that the timing of the start of the growth hormone therapy is not critical. As we expected, boys attained their adult

height later and also became taller. Children with a higher target height grew faster and tended to become taller.

The parameter estimates and credible intervals for model 2.2 without a spline is presented in Table 2.2 and those with eight interior knots are presented in Table 2.3. Here it is more difficult to attribute a clear meaning to the individual coefficients. Because we were mainly interested in the effect of the growth hormone dose, we computed the effect of dose as a function of age (see the left panel of Figure 2.3). In principle such plots could also have been made for the other covariates but here we omitted this because of lack of space. From a similar plot for the starting age we learned that a later starting age leads to results that are just as good as when therapy is started early as the children grow faster.

The addition of increasingly more knots to the spline allows a more complex relation between age SDS and height to be fit. In the right panel of Figure 2.3 we show how the inclusion of more knots to the model changes the average profile.

Table 2.1: Estimated coefficients and associated credible intervals for the model of the untransformed height scores (Sex is coded -1 for boys and 1 for girls, Dose as -1 for the low dose and 1 for a high dose)

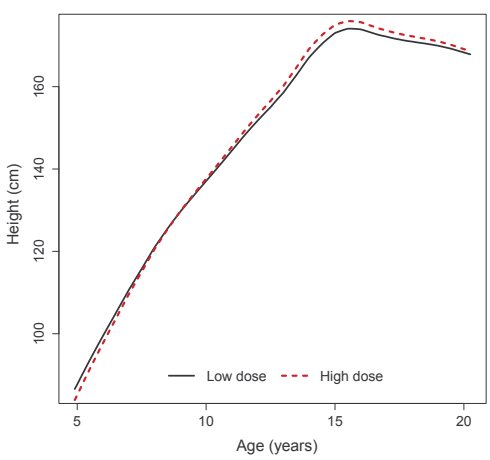
Level	**Variable**	**Estimate**	**95% Credible Interval**
Intercept	Dose	0.67	[-0.68 ; 2.10]
	Sex	-5.57	[-6.79 ; -4.28]
	Age of Entry	-0.19	[-0.81 ; 0.40]
	Target Height	3.17	[1.77 ; 4.67]
Slope	**Dose**	1.31	[0.03 ; 2.55]
	Sex	0.57	[-0.80 ; 1.81]
	Age of Entry	2.66	[2.08 ; 3.26]
	Target Height	2.12	[0.70 ; 3.51]
Change Point	Dose	-0.01	[-0.07 ; 0.06]
	Sex	-0.13	[-0.21 ; -0.07]
	Age of Entry	0.01	[-0.02 ; 0.04]
	Target Height	-0.03	[-0.10 ; 0.04]

Table 2.2: Estimated coefficients and associated credible intervals for the model of the height SDSs without a spline ((Sex is coded -1 for boys and 1 for girls, Dose as -1 for the low dose and 1 for a high dose)

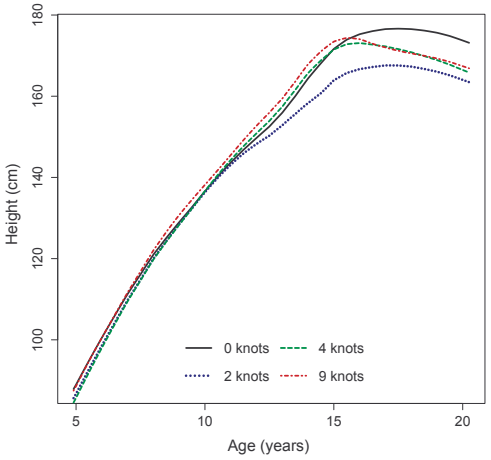
Level	Variable	Estimate	95% Credible Interval
Intercept	Dose	0.08	[-0.10 ; 0.26]
	Sex	0.08	[-0.11 ; 0.27]
	Age of Entry	-0.30	[-0.38 ; -0.22]
	Target Height	0.51	[0.31 ; 0.70]
Linear	Dose	0.08	[-0.02 ; 0.18]
	Sex	-0.09	[-0.19 ; 0.02]
	Age of Entry	0.23	[0.19 ; 0.28]
	Target Height	0.02	[-0.09 ; 0.12]
Quadratic	Dose	-0.04	[-0.12 ; 0.04]
	Sex	0.01	[-0.06 ; 0.08]
	Age of Entry	-0.04	[-0.07 ; -0.00]
	Target Height	-0.02	[-0.10 ; 0.06]

Table 2.3: Estimated coefficients and associated credible intervals for the model of the height SDSs with a spline with 9 knots

Level	Variable	Estimate	95% Credible Interval
Intercept	Dose	0.07	[-0.12 ; 0.25]
	Sex	0.08	[-0.11 ; 0.27]
	Age of Entry	-0.31	[-0.39 ; -0.23]
	Target Height	0.52	[0.32 ; 0.72]
Linear	Dose	0.09	[-0.00 ; 0.20]
	Sex	-0.16	[-0.26 ; -0.05]
	Age of Entry	0.30	[0.26 ; 0.35]
	Target Height	0.01	[-0.10 ; 0.12]
Quadratic	Dose	-0.04	[-0.11 ; 0.03]
	Sex	0.01	[-0.07 ; 0.09]
	Age of Entry	-0.07	[-0.10 ; -0.04]
	Target Height	-0.02	[-0.10 ; 0.06]



(a) Average profile for model 2 with 9 knots for boys with dose and age of entry set at zero



(b) Average profile for model 2 with various numbers of knots with dose and age of entry set at zero

Figure 2.3: Average profiles for model 2

Comparing goodness of fit

The DIC (smaller is better) of model 2.1 converted to the SDS scale was -1 831. For the basic SDS model, i.e. without interior nodes, the DIC was 110. This value can be improved upon by adding nodes to the spline as can be seen from Table 2.4. It is clear from this table that model 2.1 always provides a much better fit to the data than model 2.2 even when a substantial number of nodes are added to the spline.

Based on graphical model diagnostics (not shown) we conclude that the assumptions of our model are reasonable. From the QQ-plots we saw that the assumptions of normality for the subject specific effects and measurement errors hold approximately. When the residuals were plotted against age, we saw that model 2.1 seemed to fit quite well overall. However the variance seems to be larger at early ages. In addition at later ages the model seems to yield predictions that are slightly too low on average. There seemed to be a pattern in the residuals when no knots are in model 2.2. For model 2.2 involving splines there where no patterns in residuals. We could extend the models to incorporate non-normality and heteroskedasticity of the response in our model. However we do not pursue

this route here as it would distract from the main message.

Table 2.4: DIC (lower is better) of the untransformed height model compared with the model of the height SDS model with various numbers of interior knots

Model of untransformed height		Model of height SDS		
Effective df	DIC	Interior knots	Effective df	DIC
220.2	-1 831	0	224.0	110
		1	226.6	5
		2	228.1	-465
		3	229.7	-740
		4	230.6	-751
		5	232.6	-736
		6	233.3	-787
		7	233.2	-806
		8	234.3	-813
		9	235.2	-822

Predictive ability

The RMSE of adult height for model 2.1 was 6.9 cm. For model 2 with 9 knots the RMSE evaluated at 20 years was 8.9 cm. So it appears that model 2.1 performs better than model 2.2. However, because the height does not remain constant after a certain age in model 2.2, for the RMSE depends at the age on which it is evaluated, when an other age than 20 is chosen the RMSE can be larger or smaller. In contrast in model 2.1 the adult height and hence the RMSE is clearly defined. Looking at the predicted profiles (see Figure 2.4) we see that in model 2.2 it is very difficult to predict the change point correctly. In model 2.2 the predicted profiles are often downward sloping for certain age ranges, which is biologically impossible.

2.5 Discussion

Often mixed models are used to fit growth curves. Especially in endocrinology it is customary to do this on the SDS scale. The structure of these models can be

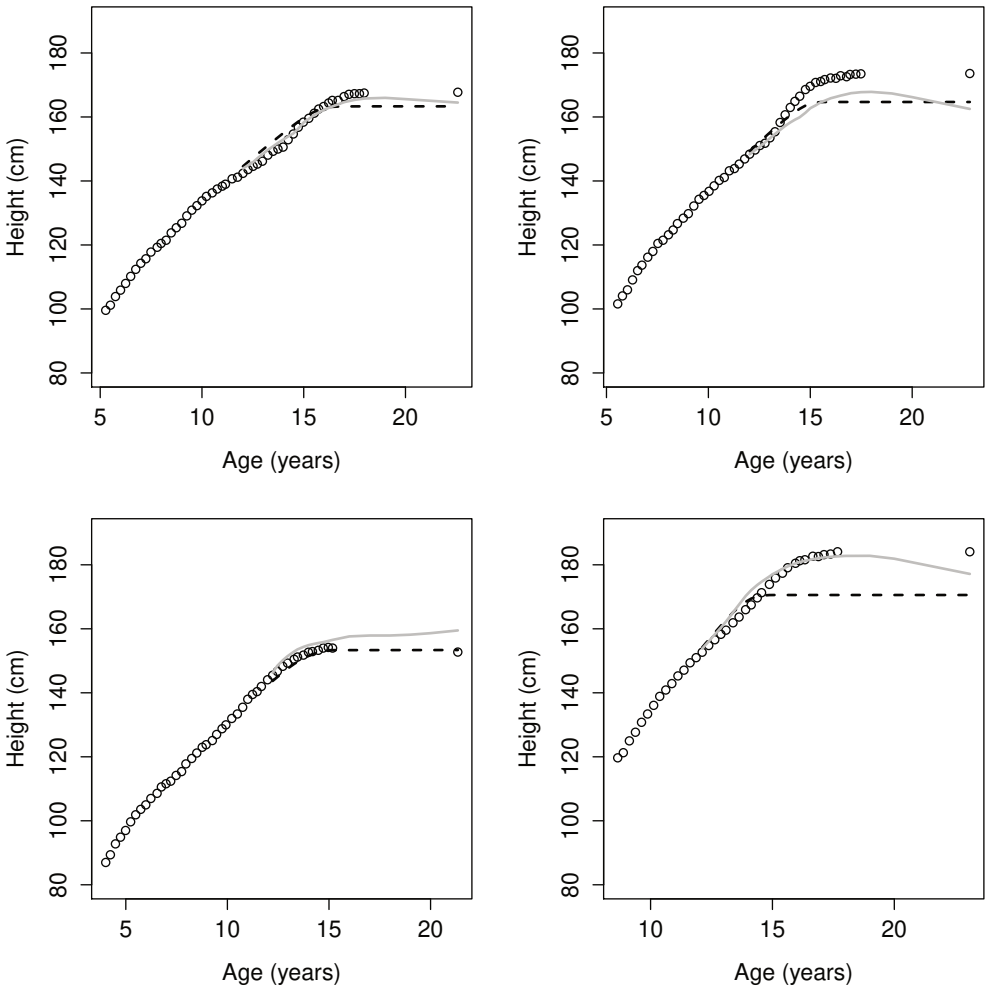


Figure 2.4: Crossvalidated predicted profiles for selected individuals on the cm scale (0=observed, solid = model 1, dashed = model 2 using 9 knots)

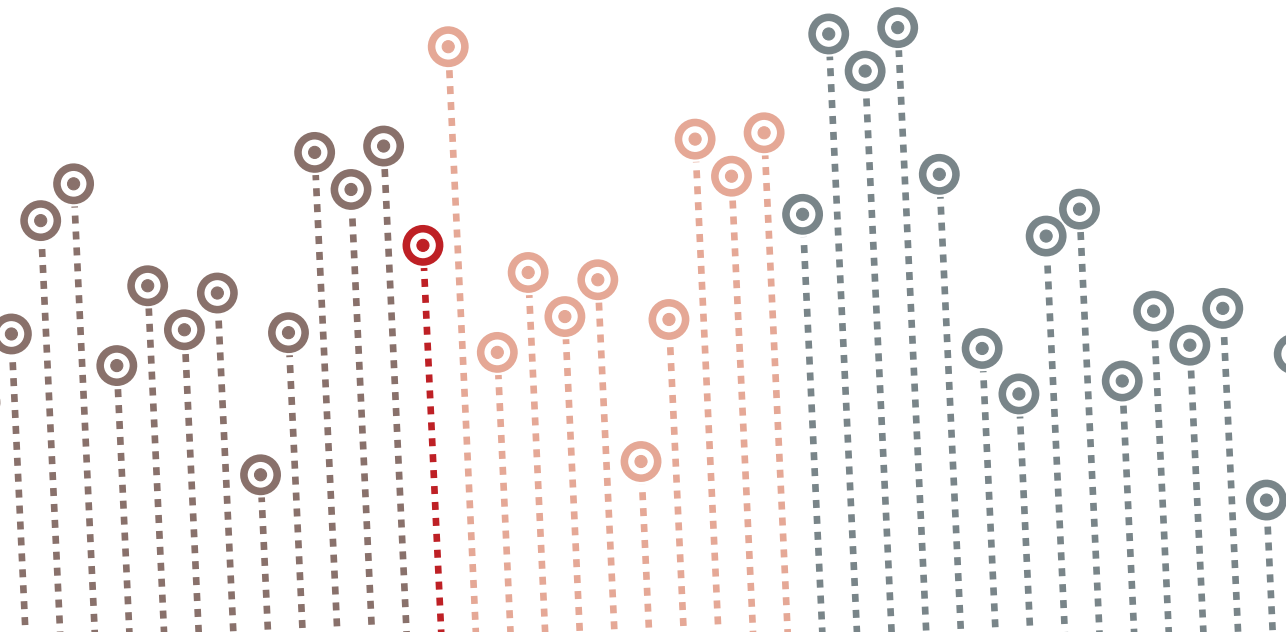
quite complicated. Using the data from the DGH trial we saw that modeling on the original scale was much easier than on the SDS scale. We obtained a good fit using less parameters than when working on the SDS scale and with parameters that are easy to interpret.

It was shown that in our data set that the fit (expressed as the DIC) is better when we modeled height directly. This is because the structure of the model gets

complicated when we transform the outcome to SDS scale for reasons that the studied group of children shows a different growth pattern than the reference population. When predicting adult height the predictions from the model on the SDS scale depend on the age at which we evaluate adult height. Furthermore some of the predicted growth profiles were downward sloping when transformed to centimeters. Note that the models used give relatively low weight to the adult height because once a child reaches its adult height it is not measured anymore. When we would increase the weight of the last measurements, for example by including pseudo observations equal to the last recorded measurement, the model on the original scale performed better in prediction.

While there are some advantages to working with the SDS, in this paper we provide a cautionary note to this standard practice. Depending on the research question and the group being investigated it depends whether modeling should be done on the original scale or on the SDS scale. When we want to make a comparison with a reference group or compare children across different ages, using the SDS scale has clear advantages making direct comparisons possible without additional corrections for age and sex. However in randomized clinical trials where we want to compare between two treatment groups that are comparable in age and sex distribution modeling on the original scale can lead to simpler models. This is especially true if the population being investigated is different than the reference population or for ages around puberty where the SDS is difficult to interpret.

3 A multivariate Bayesian model for embryonic growth



Abstract:

Most longitudinal growth curve models evaluate the evolution of each of the anthropometric measurements separately. When applied to a 'reference population' this exercise leads to univariate reference curves against which new individuals can be evaluated. However, growth should be evaluated in totality, i.e. by evaluating all body characteristics jointly. Recently, Cole et al. suggested the SITAR model which expresses individual growth curves by three subject-specific parameters indicating their deviation from a flexible overall growth curve. This model allows the characterization of normal growth in a flexible though compact manner. In this paper we generalize the SITAR model in a Bayesian way to multiple dimensions. The multivariate SITAR model allows us to create multivariate reference regions which is advantageous for prediction. The usefulness of the model is illustrated on longitudinal measurements of embryonic growth obtained in the first semester of pregnancy, collected in the ongoing Rotterdam Predict study. Further, we demonstrate how the model can be used to find determinants of embryonic growth.

Authors: S.P. Willemsen, P.H.C. Eilers, R.P.M Steegers-Theunissen & E.M. Lesaffre. published in: Statistics in medicine (2016)

3.1 Introduction

The statistical analysis of longitudinal growth curves is an active area of research, with a wide range of applications in medicine, veterinary medicine, biology and other fields. Many models have been suggested in the literature to fit such data. When applied to a normal population, the fitted growth curves can be used to define 'normal growth'. It is important to detect whether subjects exhibit normal growth. For example, a good reference of normal fetal growth during pregnancy can be essential in the management of obstetrical care [*Maršál et al.*, 1996]. In addition, abnormal growth patterns can be an indication for disease later in life [*Nieto et al.*, 1992; *Cameron and Demerath*, 2002]. Many growth models have a parametric, non-linear nature and are applied to specific application areas. Examples are the Gompertz curve, the logistic curve or the Brody curve, see e.g. [*Forni et al.*, 2009].

To relax the parametric nature of the growth curve and hence to broaden the applicability of the growth curve methodology, *Beath* [2007] suggested a flexible approach to model the longitudinal evolution of the growth patterns. This model was later extended by Cole et al. [*Cole et al.*, 2010a] who called it the 'Superimposition by Translation And Rotation' (SITAR) model. This is a shape invariant model with a single fitted 'prototype' curve. Individual curves are obtained from the prototype curve by horizontal and vertical shifts and shrinking or stretching the age scale. However, the model is univariate in nature and therefore cannot completely capture what normal and abnormal growth is when considering multiple dimensions. Multivariate growth models have been suggested in the past, but they either lack flexibility or are difficult to interpret, see e.g. *Goldstein* [1986]; *Reinsel* [1982]; *Macdonald-Wallis et al.* [2012]. In this paper, the SITAR model is generalized to multiple dimensions and is referred to as the Multivariate SITAR (MSITAR) model.

To illustrate the MSITAR model, we use data from an ongoing periconceptional cohort study, called the *Rotterdam Predict study*, conducted in the Department of Obstetrics and Gynaecology and the Department of Clinical Genetics at Erasmus MC University Medical Center, Rotterdam, the Netherlands. In this longitudinal study human embryonic growth in the first trimester of pregnancy is studied. Early growth has been a black box for a long time and it was assumed to occur uniformly, with very small differences between individuals. In fact, this is why

gestational age (GA) is often estimated from embryonic growth measurements. The GA is based on the first day of the last menstrual period. However this day can often not be determined accurately. Lately the idea of uniform growth has been largely abandoned and the thought has taken root that there are important differences in embryonic growth in the first trimester [*Bottomley et al.*, 2009; *Rousian et al.*, 2010; *Van Uitert et al.*, 2013; *Van Uitert et al.*, 2014]. Moreover, these first trimester differences have been related to birth outcomes [*Smith et al.*, 1998; *Mukri et al.*, 2008; *Bukowski et al.*, 2007; *Mook-Kanamori et al.*, 2010; *Carbone et al.*, 2012]. Therefore, the study of first trimester embryonic growth has become increasingly important. Further understanding of normal embryonic growth and the identification of determinants involved will lead to a better and earlier prediction of adverse pregnancy course and future outcomes. Then women with a family history of adverse birth outcomes may be screened at an early stage and timely interventions can be made, if necessary. In this respect, it is important to evaluate growth in its totality and not just look at measurements in separation. Indeed, measurements may appear to evolve in a normal manner when considered separately but when appreciated in a multivariate sense they may stand out as abnormal. In statistical terms, such observations are called multivariate outliers. [*Rousseeuw and Van Zomeren*, 1990, see]. It is known that such outliers cannot be identified using univariate techniques. The MSITAR model has the ability to spot such outlying growth curves, but remains conceptually simple as the univariate SITAR model.

In Section 3.2, we discuss the motivating data set in more detail and highlight the research questions that are of interest. In Section 3.3, we describe the SITAR model exhaustively and show how it can be easily extended to multiple longitudinal series of measurements. In that section we also indicate how multivariate reference regions can be obtained from the MSITAR model. Marginal and conditional reference contours will be discussed. In Section 3.4 we apply the MSITAR approach on data of the Rotterdam Predict study. In that section, we will also contrast the SITAR model applied on each of three parameters to the MSITAR model applied to the three measurements jointly. In Section 3.5 we give some concluding, clinical and statistical remarks. We also outline further extensions of our model.

3.2 Motivating Data Set: The Rotterdam Predict Study

The Rotterdam Predict study is a periconceptional cohort study carried out at the Department of Obstetrics and Gynaecology at Erasmus MC, Rotterdam, the Netherlands. The overall aim of the study is to examine early human growth and to determine factors associated with complications originating in the periconceptional period and the early stages of pregnancy. The Rotterdam Predict study [*Van Uitert et al.*, 2013] is the first of its kind in which the early pregnancy is studied meticulously. In addition to regular, 2D, ultrasound images, three dimensional (3D) ultrasound images are taken.

Every week between the sixth and the thirteenth week of pregnancy, each participating woman receives a 2D and a 3D ultrasound. From these ultrasound measurements, the crown-rump length (CRL), total arc length (TAL) and the embryonic volume (EV) are determined. The CRL is defined as the shortest distance from the crown (top of the head) to the rump (buttocks). The CRL is one of the most important measurements made in early pregnancy as it is often used for pregnancy dating. The TAL is another measure of the distance between crown and rump introduced by Boogers et. al. (unpublished manuscript), but now determined on the outside of the embryo along the dorsal side of the back as depicted in Figure 3.1. The relation between CRL and TAL is a function of the gestational age (GA), which is defined as the number of days since the first day of the last menstrual period (LMP).

The embryonic volume is measured based on the 3D ultrasound images. Using the 3D images, holographic projections of the embryo are made in the Barco I-Space (Barco N.V., Kortrijk, Belgium). Using the V-Scope software an interactive projection is made on the floor and walls using eight projectors in a way that allows depth-perception using stereoscopic imaging [*Koning et al.*, 2009]. From this image, the EV of the embryo was determined based on the grey-level of the voxels (3D pixels) of the embryo. It has been suggested that EV gives a better indication of the GA than CRL *Rousian et al.* [2010]. In addition EV is believed to be an important parameter to detect growth disorders [*Rolo et al.*, 2009]. Unfortunately, not all measurements could be performed using the 3D images because the holographic projector is an expensive piece of equipment that is also used for other research

in Erasmus MC.

For this study women above the age of 18 and before the 8th week of pregnancy were recruited. The cohort is not a random sample of the general population. Rather, women were made aware of the study by posters placed in the Gynaecology and Obstetrics outpatient clinic of the Erasmus MC. A part of the women (23%) did not go to the Erasmus MC outpatient clinic prior to inclusion, but learned about the study through word-of-mouth. As a consequence, the women included in the cohort are likely to have a higher risk of pregnancy complications.

We included 259 eligible singleton pregnancies in the Rotterdam Predict study between 2009 and 2010. 44 women were excluded because of ectopic pregnancy or miscarriage occurring before the 16th week of gestation. These pregnancies were excluded because the course of the pregnancy was different than from regular pregnancies and we are only interested in normal healthy pregnancies. Another 12 pregnancies were excluded because they could not be dated based on the LMP. This leaves 203 pregnancies for the analysis.

Finally, not all of the images could be used for all measurements mostly because of a low image quality or an unfavorable position of the embryo. There are on average approximately 5 available measurements per pregnancy for the CRL and the TAL measurement and approximately 4 measurements for the EV. The EV is often more difficult to measure accurately because sometimes the volume of the fetus cannot be clearly separated from its surroundings on the ultrasound. However, in some cases EV could be determined while CRL or TAL could not be measured.

In Table 3.1 we provide some characteristics of the study sample which reveals that the participants are predominantly highly educated. Because they are recruited mostly from a tertiary hospital they are at a higher risk for pregnancy complications. Therefore, one cannot immediately extrapolate the results of this study to the general population. As a comparison, in the regular hospital population in the Netherlands about 16% of the women giving birth is of non western origin, about 51% is nulliparous and the median age is between 30 and 34 years *De Graaf et al.* [2010]. Notice that we do have some missing values for some variables. The cases with missing data include two cases where the pregnancy was conceived after oocyte donation. For these cases, the relation between maternal characteristics and growth is likely to be different than in regular pregnancies. In addition, there are ten pregnancies for which one or more of the covariates we will

Table 3.1: Descriptive statistics of the selected women from the Predict study (n=203).

Characteristic	Mean / N	(SD/ %)
Age	32.2	(4.8)
Missing	11	(5%)
Ethnicity		
Dutch	150	(74%)
Other western	16	(8%)
Other non-western	30	(15%)
Missing	7	(3%)
Education		
Low	18	(9%)
Middle	56	(28%)
High	113	(56%)
Missing	16	(8%)
BMI	24.6	(4.1)
Missing	8	(4%)
Parity		
Nulliparous	124	(61%)
Multiparous	74	(36%)
Missing	5	(2%)
Periconceptional alcohol use		
Yes	87	(43%)
No	109	(54%)
Missing	7	(3%)
Periconceptional smoking		
Yes	31	(15%)
No	165	(81%)
Missing	7	(3%)

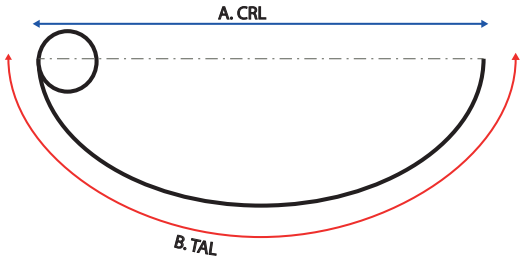


Figure 3.1: Illustration of CRL and TAL: the head of the fetus is on the left and the curved line represents its back. The distance A is the CRL and the distance B is the TAL

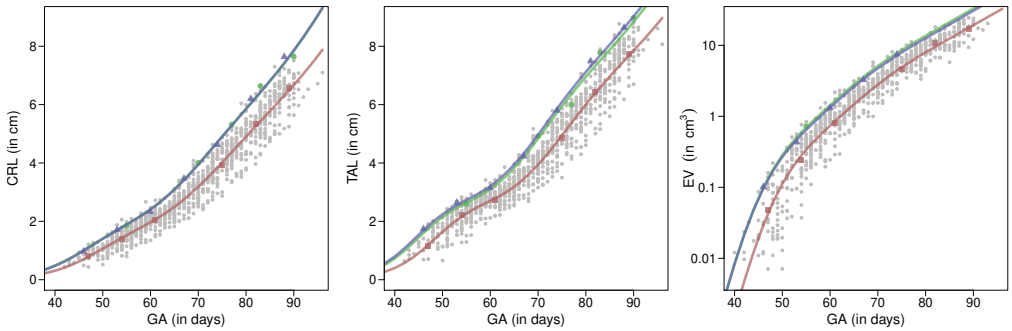


Figure 3.2: Scatterplots of the ultrasound measurements taken in the first three months of pregnancy. Three randomly selected pregnancies are indicated. The estimated profiles for these individuals are obtained from the MSITAR model (See Section 3.4)

consider are missing. So for analyses that use covariates we have 191 pregnancies left. We could also have imputed the missing values within the MCMC scheme. However this increases computational complexity. Because we feel it distracts from the main idea of the paper we did not follow this path.

Here we use the Rotterdam Predict study to illustrate how the multivariate SITAR model can be used to model human embryonic growth. In particular we indicate that our model is better capable to detect abnormal pregnancies.

In Figure 3.2 we show scatter plots of the three features of embryonic growth versus GA, wherein some randomly selected growth profiles are highlighted. Note that we plotted EV on a logarithmic scale. Clearly there is a large amount of correlation between the measurements of each pregnancy, both within each separate outcome but also between the different series.

3.3 The SITAR Model

The univariate SITAR model

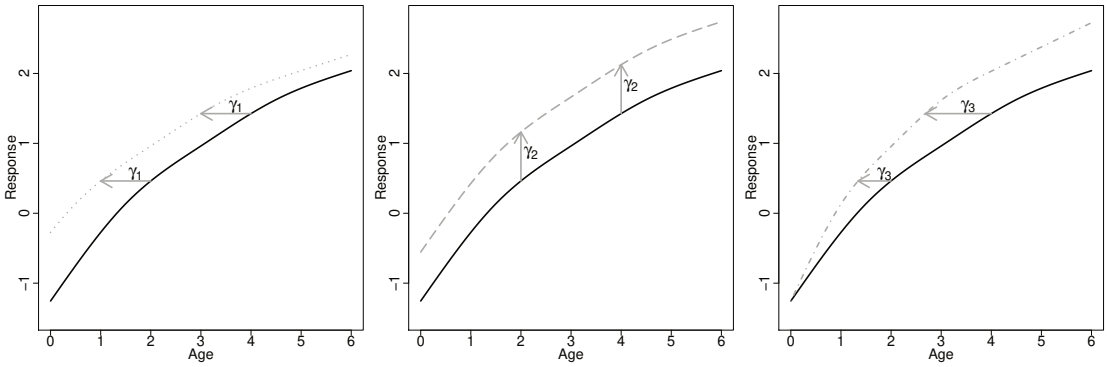
The SITAR model is based on the 'Shape invariant' model of infant growth suggested by *Beath* [2007]. The idea behind the model is that there exists some general growth curve, or prototype function, $f()$ such that all individual growth profiles can be reduced to this general curve by translating them horizontally and vertically and by stretching them horizontally. So we have $y_{ij} = \gamma_{i2} + f(\gamma_{i3}[t_{ij} + \gamma_{i1}]) + \varepsilon_{ij}$. where y_{ij} is the outcome of individual i measured at time t_{ij}. The γs are the subject specific effects and express the subject-specific horizontal shift (γ_{i1}), the vertical shift (γ_{i2}) and the stretch (γ_{i3}) with respect to the general growth curve. ε_{ij} is the measurement error. See Figure 3.3 for a schematic representation of the effect of the random effects. Note that stretching the general curve horizontally is something that is related to accelerated failure time models in which the scale parameter has a similar role as γ_{i3}.

In the SITAR model, the curve of the general pattern is modeled by a natural cubic spline function. A spline of degree d is a smooth function that is piece-wise constructed from polynomials of degree d. For a cubic spline this degree is three. The points where the polynomial functions join are called the inner knots, while the knots at the boundary are the outer knots. A spline is called 'natural' when the second derivative at the outer knots of the spline is zero. The spline can then be linearly extrapolated beyond the outer knots. Proper extrapolation is needed since the horizontal shifts imply that the domain in which the splines will be evaluated are not exactly known beforehand. With κ inner knots the natural cubic spline has $\kappa + 2$ independent coefficients.

Now the univariate SITAR model can be expressed as:

$$\begin{aligned} y_{ij} &= \gamma_{i2} + \boldsymbol{z}_{ij}^T\boldsymbol{\beta} + \varepsilon_{ij}, \\ \boldsymbol{z}_{ij} &= \boldsymbol{B}\left(\exp(\gamma_{i3})(t_{ij} + \gamma_{i1})\right), \ (i = 1, \ldots, N;\ j = 1, \ldots, n_i), \\ \boldsymbol{\gamma}_i &\sim \mathrm{N}(\boldsymbol{0}, \Sigma_\gamma), \\ \varepsilon_{ij} &\sim \mathrm{N}(0, \sigma^2). \end{aligned} \tag{3.1}$$

We made the assumption here that $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3})^T$ is normally distributed and its components are centered at zero to avoid identifiability problems. $\boldsymbol{B}(t)$ is a function that returns the basis of the natural cubic spline, evaluated at t. Thus, $\boldsymbol{z}_{ij}$



(a) Horizontal shift (b) Vertical shift (c) Horizontal Stretch

Figure 3.3: Schematic representation of SITAR model with horizontal shift (γ_{i1}), vertical shift (γ_{i2}) and stretch (γ_{i3})

is a vector of length $\kappa + 2$, as is the regression coefficient vector $\boldsymbol{\beta}$. We also assume that ε_{ij} is independently distributed with variance σ^2 and independent of the γs.

The SITAR model has been successfully applied to different data sets, appears to work well in practice and has parameters that are easily interpretable by clinicians [*Cole et al.*, 2010a].

The multivariate SITAR model

The univariate SITAR model can be easily extended to $K(> 1)$ outcomes. The multivariate SITAR model (MSITAR) is given by:

$$\begin{aligned} y_{ijk} &= \gamma_{i2k} + \boldsymbol{z}_{ijk}^T \boldsymbol{\beta}_k + \varepsilon_{ijk}, \\ \boldsymbol{z}_{ijk} \equiv \boldsymbol{z}_{ijk}(\boldsymbol{\gamma}_{M_i}) &= \boldsymbol{B}\left(\exp(\gamma_{i3k})(t_{ijk} + \gamma_{i1k})\right), \\ &\quad (i = 1, \ldots, N;\ j = 1, \ldots, n_{ik};\ k = 1, \ldots, K), \\ \boldsymbol{\gamma}_{M_i} &\sim \mathrm{N}(\boldsymbol{0}, \Sigma_{M_\gamma}), \\ \varepsilon_{ijk} &\sim \mathrm{N}(0, \sigma_k^2), \end{aligned} \tag{3.2}$$

where y_{ijk} is the kth response of individual i at time t_{ij}, z_{ijk} a basis of a natural cubic spline for the kth response of individual i at time t_{ij}, $\boldsymbol{\beta}_k$ the vector of spline coefficients for the kth response, $\gamma_{ilk}\,(l = 1, 2, 3)$ are the three subject specific effects for individual i and series k $(k = 1, \ldots, K)$, $\boldsymbol{\gamma}_{ik} = (\gamma_{i1k}, \gamma_{i2k}, \gamma_{i3k})^T$ is the

vector of all subject specific effects for outcome k and individual i. When we combine the subject specific effects for all outcomes for an individual we obtain $\boldsymbol{\gamma}_{M_i} = (\boldsymbol{\gamma}_{i1}^T, \ldots, \boldsymbol{\gamma}_{iK}^T)^T$. Σ_{M_γ} is the $3K \times 3K$ joint covariance matrix of all random effects $\boldsymbol{\gamma}_{M_i}$. ε_{ijk} is the measurement error which we assume to be independently distributed and independent of the γs. Finally, σ_k^2 is the measurement error variance for series k. Note that it is assumed that the K outcomes are independent given the random effects.

The multivariate SITAR has the advantage over the univariate SITAR model that the three series of measurements can be modeled jointly. In general, joint modeling of multiple series of growth curves allows for:

- Evaluating relationships between the growth curve series over time.
- Establishing multivariate reference regions and thereby checking whether individuals not only show a normal growth pattern for an individual series but also for, e.g., ratios of series.
- Better prediction of a future growth profile of a series when no or few past observations are available by also incorporating information from the past profiles of other series.

We were also interested in the effect of covariates on the parameters of the model. That is, we assumed that the means of the $\boldsymbol{\gamma}$s could depend on covariates. Let $\boldsymbol{x}_i$ be the vector of covariates for subject i of length n_p. We then assume that $\boldsymbol{\gamma}_{M_i} \sim \mathrm{N}(A\boldsymbol{x}_i, \Sigma_{M_\gamma})$. Here A is a $3K \times n_p$ matrix of parameters with elements $\alpha_{k,p}$. We will also use the notation $\boldsymbol{\alpha}$ to denote $\mathrm{vec}(A) = (\alpha_{1,1}, \alpha_{2,1}, \ldots, \alpha_{3K,n_p})^T$, which is the vectorized form of the matrix A, that is the column matrix that consists of the columns of A that are stacked on top of each other.

MCMC implementation

We used a Bayesian approach to estimate the model parameters, employing vague priors for all parameters. When possible we also opted for priors that are conditionally conjugate. Specifically, we assumed for the αs and βs independent normal priors with zero mean and a standard deviation equal to $\sigma_{\alpha,0} = \sigma_{\beta,0} = 1,000$, which is large relative to the plausible values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For σ^2 an inverse gamma prior was chosen with shape parameter (α_σ) and rate parameter (β_σ) equal to 0.001, i.e. IG(0.001, 0.001). Finally, for Σ_{M_γ} we have chosen an inverted Wishart

prior with degrees of freedom δ equal to three times the number of outcomes and a scale matrix Ψ equal to 0.01 times the identity matrix, i.e. $\text{IW}(3K, 0.01\text{I}_{3K})$. For a motivation of these choices see *Lesaffre and Lawson* [2012, p. 260]. As more (external) information is collected on growth during early pregnancy, we believe that in due time some of these vague priors may be replaced by more oinformative priors.

The posterior distribution is proportional to the product of the likelihood and the priors, and therefore we obtain:

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}_M, \Sigma_{M_\gamma} \mid \boldsymbol{y}) \propto & \prod_{ijk} \text{N}(y_{ijk} \mid \gamma_{i2k} + \boldsymbol{z}_{ijk}(\boldsymbol{\gamma}_{M_i})^T \boldsymbol{\beta}_k, \sigma_k^2) \times \\ & \prod_i \text{N}(\boldsymbol{\gamma}_{M_i} \mid \mathbf{0}, \Sigma_{M_\gamma}) \times \\ & \prod_k \text{N}(\boldsymbol{\alpha}_k \mid \mathbf{0}, \sigma_{\alpha,0}^2 I_{n_p}) \prod_k \text{N}(\boldsymbol{\beta}_k \mid \mathbf{0}, \sigma_{\beta,0}^2 I_{\kappa+2}) \times \\ & \prod_k \text{IG}(\sigma_k^2 \mid \alpha_\sigma, \beta_\sigma) \, \text{IW}(\Sigma_{M_\gamma} \mid \delta, \Psi), \end{aligned} \tag{3.3}$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)^T$, $\boldsymbol{\gamma}_M = (\boldsymbol{\gamma}_{M_1}, \ldots, \boldsymbol{\gamma}_{M_N})^T$ and $\boldsymbol{y} = (y_{111}, y_{121}, \ldots, y_{1n_i1}, y_{211}, \ldots, y_{N,n_{NK},K})^T$.

We implemented a Gibbs sampling procedure to estimate the model parameters. The (block) full conditionals for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ_k^2 and Σ_{M_γ} are given in the supplementary material. Most full conditionals are of standard form and hence standard samplers can be used. However, this is not the case for the full conditional distribution of the subject specific effects, given by:

$$\begin{aligned} \boldsymbol{\gamma}_{M_i} | \boldsymbol{y}, \ldots \sim & N_{3K}(\boldsymbol{\gamma}_{M_i} \mid A\boldsymbol{x}_i, \overline{\Sigma}_\gamma) \times \\ & \prod_{j,k} \text{N}(y_{ijk} \mid \gamma_{i2k} + B(\exp(\gamma_{i3k})(t_{ij} + \gamma_{i1k})\boldsymbol{\beta}_k), \sigma_k^2), \end{aligned} \tag{3.4}$$

where a Metropolis step is needed.

The above model can be estimated in JAGS [*Plummer*, 2003a]. We also implemented the sampler directly into C++ making use of the 'Eigen'[`http://eigen.tuxfamily.org/`] and 'Boost' libraries [`http://www.boost.org/`]. By embedding the program into the R language its output can be post-processed with R functions, such as CODA [*Plummer et al.*, 2006]. For the full conditionals of the

subject specific effects, we used in the C++ program a Random Walk Metropolis algorithm with a multivariate t proposal distribution with 5 degrees of freedom. The acceptance rate of the Metropolis algorithm was tuned so that it was approximately 30%. The source code of our programs can be obtained upon request from the first author.

Model selection and evaluation

Model selection was done using the popular Deviance Information Criterion (DIC) [*Spiegelhalter et al.*, 2002b]. More specifically, we compared the MSITAR model to the ensemble of univariate SITAR models to check the necessity of modeling the three series of growth curve responses jointly.

To check the assumptions of our model, we performed posterior predictive checks (PPC). That is, based on the estimated model we simulated replicated series. We then compared the distribution of the test statistic calculated on the replicates with the test statistic based on the data which we actually observed. Specifically we looked at the chi-squared goodness-of-fit test recommended by *Gelman et al.* [1996]. That is, we drew posterior samples for each of the K outcomes

$$\chi^2_{\mathrm{GM},k} = \sum_{i,j} r^2_{ijk} \qquad \text{with} \qquad r_{ijk} = \frac{y_{ijk} - E(y_{ijk} \mid \widetilde{\boldsymbol{\theta}})}{sd(y_{ijk} \mid \widetilde{\boldsymbol{\theta}})} \tag{3.5}$$

and compared it with $\chi^2_{\mathrm{GM},k,\mathrm{rep}}$ obtained from replacing y_{ijk} by the corresponding sampled value $\widetilde{y}_{ijk}$ from the posterior predictive distribution (PPD) in (3.5), $p(\widetilde{y}_{ijk}|\boldsymbol{y})$. In this expression $\widetilde{\boldsymbol{\theta}}$ stands for the sampled total parameter vector and $E(y_{ijk})$ is defined as $\boldsymbol{B}(\tilde{\gamma}_{i3k}(\tilde{\gamma}_{i1k} + t_{ijk}))\tilde{\boldsymbol{\beta}}_k + \tilde{\gamma}_{i2k}$. We then checked the proportion of samples in which $\chi^2_{\mathrm{GM},k,\mathrm{rep}} \geqslant \chi^2_{\mathrm{GM},k}$. In a similar way we computed the PPC evaluating the skewness and kurtosis of the error terms ε_{ijk}. For instance, the statistic $\chi^2_{\mathrm{SKEW},k} = \sum_{i,j} r^3_{ijk}$ checks the skewness of the errors, which we assumed zero.

Finally, we extended the model so it could accommodate a residual error with excess kurtosis. Specifically, we have replaced the normal residual error by a scaled t-distribution. In our C++ program we simulated from a t-distribution by considering it as a scale mixture of normals. We considered the inverse of the number of degrees of freedom to be uniformly distributed on $[0, \frac{1}{3}]$. With this

prior we give a larger prior probability to smaller numbers of degrees of freedom, that is to distributions that are further away from the normal.

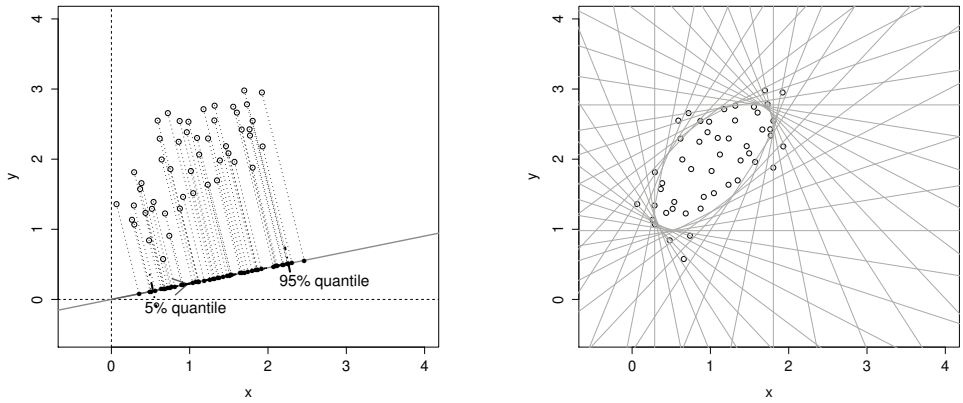
Predictive ability of the model

The predictive ability of the models was evaluated by the root mean squared prediction error (RMSE). We want to know, once we have an estimated model, how well it can predict a future observation of a new pregnancy given the available measurements of this pregnancy. This was evaluated using five-fold cross-validation, whereby the total data set was split into five (approximately) equal parts, each containing 20% of the pregnancies. Each time 80% of the observations served as training set and the remaining 20%-part as validation set. In the validation part, we removed the last value for each combination of pregnancy and outcome, which we will denote by $y_{ik,\text{LAST}}$, which is to be predicted. In the supplemental material we provide the details of the procedure we used to draw samples from the predictive distribution of this last observation. We then took the mean of these samples for each i and k, denoted by $\hat{y}_{ik,\text{LAST}}$ and calculated the RMSE as $\sqrt{\sum_{i=1}^{N_k}(y_{ik,\text{LAST}} - \hat{y}_{ik,\text{LAST}})^2/N_k)}$. Here N_k is the number of pregnancies for which we observed the kth outcome.

Reference contours

Our model allows for the construction of multivariate reference ranges. With a single outcome a reference range is a central predictive interval in which 100(1-q)% of the observations are located. Often equal-tail predictive intervals are taken. This can be generalized to multiple dimensions so we obtain a reference region excluding the most outlying observations and containing the least outlying observations (or for lack of a better word the 'normal' observations). However such an extension is not trivial since univariate quantiles are not easily adapted to the multivariate case [*Serfling*, 2002]. Here we followed the approach of *Kong and Mizera* [2008] and look at contours obtained from a collection of directional quantiles. For a direction indicated by a unit vector $\boldsymbol{s}$, the *qth directional quantile* in the direction of $\boldsymbol{s}$, i.e. $Q(q, \boldsymbol{s})$ of a multivariate random variable $\boldsymbol{y}$ is defined as the quantile of $\boldsymbol{s}^T\boldsymbol{y}$, which corresponds to the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{s}$. To construct the directional quantile contour $Q(q)$, one varies the direction given by

the vector $\boldsymbol{s}$ to cover all angles. Each $Q(q, \boldsymbol{s})$ defines a half-space where $100(1-q)\%$ of the observations lie. By taking the intersection of these half-spaces one obtains the contour $Q(q)$, see Figure 3.4 for a graphical representation of this construction. Note, however, that in this figure only a limited number of directions $\boldsymbol{s}$ has been taken to enhance the legibility of the figure. In general, $Q(q)$ will exclude more than a fraction q of the observations. A contour that excludes a fraction q of the observations is then obtained by an optimization routine to find a value q' for which $Q(q')$ does have this property.



(a) Directional quantiles are the quantiles of the projected points

(b) The directional quantile lines circumscribe a convex region: the reference contour

Figure 3.4: The construction of directional quantiles.

To generate the directional quantile contours based on our model, we generated samples $\boldsymbol{y}$ for a new individual from the posterior predictive distribution of our model and drew a large number of directional quantiles based on this sample. This results in unconditional or marginal contours. It is also possible to derive conditional contours, i.e., when there is interest in the expected range of future 'normal' outcome values for subjects given their history (i.e. past covariates and responses). These values can be generated in much the same way as the unconditional ones. The difference is that we now sample γ_{M_i} from the conditional distribution (3.4) of those individuals given their history.

Conditional quantile contours can also be constructed based on the distribu-

tion of the residuals in a way that avoids sampling. For example, if a multivariate outcome is (conditionally) multivariate normal, the probability contours are elliptical. However this fails to take the uncertainty in the parameter estimates into account properly. We believe this is important in a clinical setting when limited information is available and parameter uncertainty is relatively large.

Effect of time scale

The γ_{i3k} random effects result in stretching the time axis in order to, together with the other random effects, align the curve of the ith individual with the overall 'mean' smooth curve. We argue that this stretch effect is most effective when the first measurement of each individual is taken just after time zero. When the origin of the time axis is remote from the measurement times the stretch effect is rather similar to the horizontal shift effect, so sampling becomes more difficult. However, in human embryonic growth studies, measurements cannot be taken at conception. In fact, in the Rotterdam Predict study, the first measurement was taken between days 40 and 50 since no reliable measurements can be made before this time with the current techniques. For this reason, model (3.2) is based on the GA minus 36 days. In this way the transformed GA at the first occasion that measurements could be taken is small but positive. In general, when we apply a linear transformation on the time scale, we have: $\tilde{t}_{ijk} = \frac{t_{ijk}+a}{b}$ so the second line of model (3.2) becomes $z_{ijk} = \tilde{B}(exp(\tilde{\gamma}_{i3k})(\tilde{t}_{ijk} + \tilde{\gamma}_{i1k})$. Here $\tilde{B}$ is a new spline function that is created by applying the same linear transformation on the nodes of B. When we only change the time scale the effect is similar to increasing the prior variance of the horizontal shifts by a factor $\sqrt{b}$. The effect of a is to change the origin of the time scale. In this case the effect of the transformation is more profound and finding an appropriate prior would be much more complicated than changing the time scale. In theory optimal a and b could be estimated from the data by treating them as extra parameters. However, we have chosen here not to do so to avoid further computational complexity.

3.4 Analysis of the motivating data set

In this section we present the results of our analysis. First we estimated the SITAR model separately for each of the three series (CRL, TAL and EV) and looked at the

correlation between the subject specific effects obtained in these models. We then estimated the joint model for all the outcomes together. In the next subsection we focus on the inclusion of covariates. In all analyses we used $\kappa = 3$ with equidistant nodes, which provided enough flexibility for the prototype growth curve. The univariate models were run for 600,000 iterations (with 300,000 burn-in iterations) and the multivariate model for 1,200,000 iterations (with 600,000 burn-in iterations). For all models three chains were initiated with different starting values. Convergence was checked visually by examining the trace plots of the Markov chains and more formally by the Gelman and Rubin [*Gelman and Rubin*, 1992] test. The effective sample size ranged from approximately 1,200 for the $\boldsymbol{\beta}$ values to 11,000 for the σ^2 values. On an Intel i5-24000 processor running at 3.1 GHz with 8 GB RAM and the 64 bit version of Windows, the time needed to run 1,200,000 iterations for the multivariate model was approximately 18 hours.

Univariate analyses

We used a logarithmic transformation for all responses. This reduced heteroscedasticity in the responses and also ensures that predictions are always positive. Note that the vertical shift on the log scale corresponds to a stretch effect on the original scale.

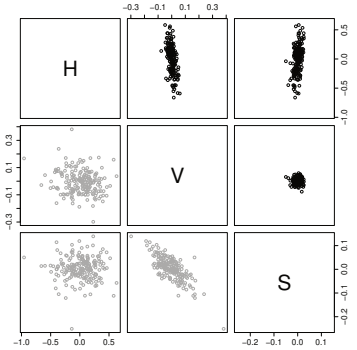
We wanted to investigate the relation between the subject specific effects that belong to the same pregnancy. Therefore we have made scatterplots of the components of the posterior means of the centered effects. By centered effects we mean the deviations of these effects from there expected values (ie. $\gamma_M - XA^T$). In Figure 3.5 we organized the scatterplots of pairs of variables into six scatterplot matrices. In the upper triangular part of each scatterplot matrix we show the pairwise scatterplots of the posterior mean of the centered random effects within the same series (subplots a–c), and the scatterplot between similar effects in different series (subplots d–f). The correlation between the horizontal and vertical effects is always negative. Other than that, there does not seem to be an obvious pattern in the correlations of each outcome. When we look at the correlation between similar effects of different outcomes the very strong correlations of 0.9 between the horizontal shifts stand out (See subplot d). The other correlations are smaller but still positive (0.1–0.4). The relative size of the horizontal shifts, (compared to their sd's, which are not shown here), appear greater than those of

the vertical and stretch effects.

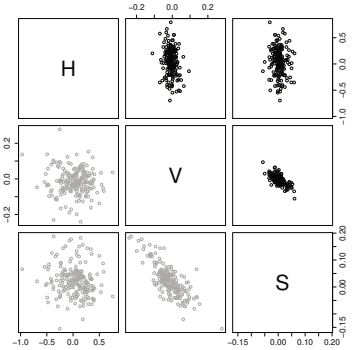
In Table 3.2 we show the results of the posterior predictive checks. Here numbers close to either zero or one indicate that model assumptions might not be met. As is obvious from the leftmost columns of this table there are some deviations from the model assumptions when we assume a normally distributed measurement error. The residuals appear to be platykurtic distributed. This could be remedied by replacing the normal residual error by a t-distribution (this is illustrated in the rightmost part of the table). The posterior mean of the degrees of freedom were 4.9, 3.9 and 3.5 for the CRL, TAL and EV respectively. In the Supplementary Material we provide Q-Q plots of the subject specific effects. There appear to be some, relatively minor, deviations from normality. In particular the stretch-effect of the TAL seems to have heavier tails than expected. An obvious way to remedy this is also to change the distribution of the subject-specific effects. Unfortunately we were unable to fit a model where both the subject specific effects were non-normal. Because changing the residual errors to a scaled t-distribution also improved the Q-Q plots of the subject specific effects, we continued with this model.

Table 3.2: Posterior predictive checks of the **univariate** SITARs

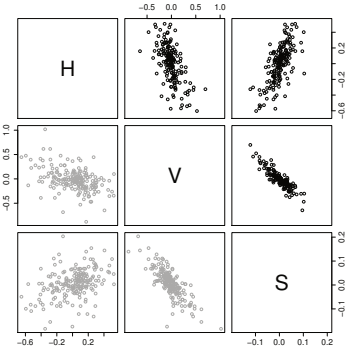
	PPC: Normal errors			PPC: t-distributed errors		
	CRL	TAL	EV	CRL	TAL	EV
χ^2_{GM}	0.48	0.52	0.49	0.48	0.35	0.45
χ^2_{SKEW}	0.27	0.12	0.52	0.41	0.22	0.78
χ^2_{KURT}	0.00	0.00	0.00	0.69	0.61	0.90



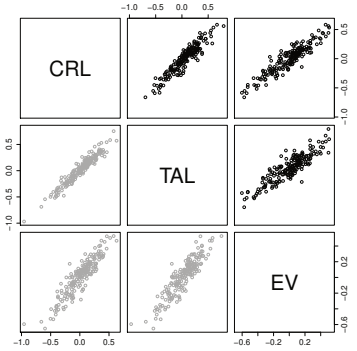
(a) CRL



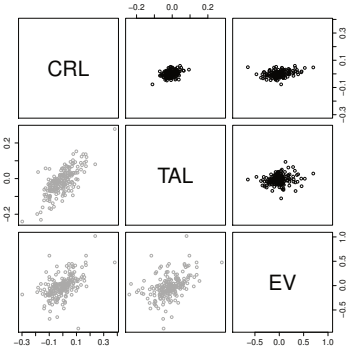
(b) TAL



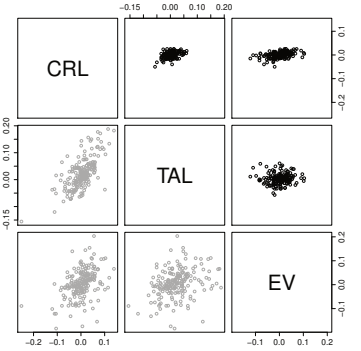
(c) EV



(d) Horizontal Effects



(e) Vertical Effects



(f) Stretch Effects

Figure 3.5: Scatterplots of the posterior means of the subject specific effects $\gamma_M - XA^T$ from the univariate model (upper triangular) and multivariate model (lower triangular). 'H' denotes the horizontal shift, 'V' the vertical shift and 'S' the stretch effect

Multivariate analysis

We then estimated the joint model for all three series together, i.e. the MSITAR model. For this we used the results of our univariate models as starting values of the sampling procedure. The scatterplots of the subject specific effects within each series are shown in the lower triangular subplots of Figure 3.5a–c. The points have slightly fanned out. This is most evident in the vertical and shift effects in the CRL and TAL (see subfigures a and b). The correlation between the horizontal and vertical effects is still negative like in the univariate models. Again, there does not seem to be a pattern in the correlations that involve the stretch effect. Most correlations have the same sign as in the univariate case. Notable exceptions are the relations between the stretch effect and the two other effects in the CRL (Compare the rightmost column and bottom row of subfigure 3.5a). In Figures 3.5d–f we show the scatterplot of similar subject specific effects between similar effects in different series. In subfigure 3.5d we see that, like in the univariate models, the correlation between the horizontal effects is strong (around 0.9). The correlation between the vertical shift and the stretch effects has also become stronger when compared with the univariate models (see subfigures 3.5d and f). As for the univariate models we found some deviations from the model assumptions with the posterior predictive checks (see Table 3.3). Again the observed kurtosis does not match that of the normal distribution. When we replaced the normal distribution with a t-distribution, this problem was solved for CRL and TAL. However, the t-distributions seem to be unable to accurately model the kurtosis of the EV. The posterior mean of the degrees of freedom where 4.5, 4.0 and 3.3 for the CRL, TAL and EV respectively. For three pregnancies we have plotted the estimated profiles based on the MSITAR model with t-distributed errors in Figure 3.2. The estimated profiles from the model with a normal error look very much like these ones and are not shown.

Table 3.3: Posterior predictive checks of the **multivariate** SITAR models

	PPC: Normal errors			PPC: t-distributed errors		
	CRL	TAL	EV	CRL	TAL	EV
χ^2_{GM}	0.50	0.48	0.19	0.50	0.28	0.04
χ^2_{SKEW}	0.37	0.14	0.25	0.22	0.15	0.53
χ^2_{KURT}	0.00	0.00	0.09	0.21	0.70	1.00

Comparison of the univariate and multivariate models

We compared the collection of the three univariate models with the MSITAR model using DIC. We obtained for the multivariate model a DIC of -7,222 with 922 effective parameters (p_D). This value is substantially smaller than for the collection of univariate models for which the DIC is -6,785 with $p_D = 1,030$. Hence the MSITAR model is preferred.

Table 3.4: Root mean squared prediction error of the ensemble of SITAR models and the MSITAR model (both with normal and t-distributed error (MSITAR-T))

	RMSE		
Variable	**SITARs**	**MSITAR**	**MSITAR-T**
log(CRL)	0.051	0.053	0.052
log(TAL)	0.042	0.043	0.044
log(EV)	0.172	0.171	0.172
log(CRL/TAL)	0.042	0.041	0.042

In Table 3.4 we compare the RMSE between the univariate and multivariate models. The RMSEs largely agree. So from these results we must conclude that there does not seem to be any benefit in using the MSITAR model for prediction. However, this general conclusion overlooks some special situations that are highly relevant in practice. Indeed, one advantage of the multivariate model is that it also enables us to make predictions for one of the outcomes when no previous measurements on that kind of outcome are available. For an example we refer to the EV of the pregnancy marked with 'triangles' in Figure 3.2. The reason is that in the MSITAR model the correlation among the series is used to predict the

responses of the missing series while in the univariate models there is nothing that can provide this information. Even when we have an outcome for which we have only a single previous measurement of the same outcome as the outcome which we want to predict, the multivariate model does slightly better. For example when we predict the last measured EV in those pregnancies where we had only one or two previous observations the RMSE of the multivariate models was 0.29 vs 0.33 for the univariate model. For the CRL and the TAL we have more than 2 observations for almost all of the pregnancies.

Covariates

Using the MSITAR model we aimed to look at the effect of periconceptional conditions such as maternal age, nulliparous pregnancies (i.e. first-time pregnancies), smoking and alcohol use of the mother and maternal BMI on CRL, TAL and EV. All continuous covariates were standardized while for the binary variables we used effect coding. In Table 3.5 we show the estimated coefficients. Here we only show the estimated coefficients from the MSITAR model with t-distributed errors, however the results are similar for the univariate models and models with normal distributed errors. We see that the pattern of the coefficients is similar across the different outcomes. It seems that the horizontal effects are most 'significant'.

While the parameters have an obvious meaning, a predictor can influence the response of an outcome in multiple ways. For example smoking at the same time causes children to be 'slow' (ie. behind the curve) and bigger. Therefore, to better appreciate the effect of a particular covariate we plotted the 'average' profiles for a typical pregnancy. In Figure 3.6 we have graphically shown the effect of smoking on the three series.

Reference contours

Unconditional reference contours allow us to determine the range of 'normal' ultrasound images at a particular GA for a new pregnancy. In Figure 3.7 we demonstrate how these contours can be constructed. First, one generates a sample of $\boldsymbol{y}$ for a new individual from the estimated MSITAR model and draws all directional quantiles based on this sample. This is illustrated in Figure 3.7(a) where we have simulated values for CRL, TAL and EV for an embryo of 53 days old

Table 3.5: Regression coefficients (posterior means and standard deviations) of the MSITAR model

	CRL					
	Horizontal effect		**Vertical effect**		**Stretch effect**	
	Coef	**SE**	**Coef**	**SE**	**Coef**	**SE**
Maternal age	0.045	0.021	0.017	0.014	-0.0060	0.0074
Nulliparous pregnancy	-0.056	0.021	0.021	0.013	-0.0028	0.0071
Smoking	-0.085	0.025	0.021	0.017	-0.0085	0.0086
Alcohol use	-0.006	0.024	-0.023	0.013	0.0075	0.0068
BMI	0.025	0.022	-0.013	0.013	0.0018	0.0068
	TAL					
	Coef	**SE**	**Coef**	**SE**	**Coef**	**SE**
Maternal age	0.071	0.022	0.009	0.009	-0.0089	0.0056
Nulliparous pregnancy	-0.053	0.022	0.021	0.009	-0.0069	0.0056
Smoking	-0.091	0.026	0.028	0.011	-0.0169	0.0075
Alcohol use	-0.004	0.022	-0.017	0.008	0.0050	0.0054
BMI	0.008	0.024	0.002	0.009	-0.0048	0.0057
	EV					
	Coef	**SE**	**Coef**	**SE**	**Coef**	**SE**
Maternal age	0.048	0.019	-0.012	0.072	0.0046	0.0170
Nulliparous pregnancy	-0.033	0.019	0.021	0.063	0.0014	0.0155
Smoking	-0.054	0.023	0.205	0.074	-0.0563	0.0185
Alcohol use	-0.008	0.020	0.087	0.056	-0.0283	0.0136
BMI	0.019	0.019	0.052	0.076	-0.0145	0.0185

from the Rotterdam Predict study. Note that this pregnancy appears quite normal univariately with a CRL of 1.45 mm and a TAL of 2.06 mm, but becomes rather unusual when we look at it bivariately. Note also that the fraction of the sample that is outside of the contour of the 5% directional quantiles is actually larger than 10%. As a reference, we have also plotted the quantile contour that does encompass 90% of the observations. Another way to visualize how 'normal' a certain observation is to plot the minimum directional quantile and the direction in which this minimum is reached, i.e. the pair $(\min_s(Q(q,s)), \operatorname{argmin}_s(Q(q,s))$ This is done in Figure 3.7(b). By connecting the values observed at different time

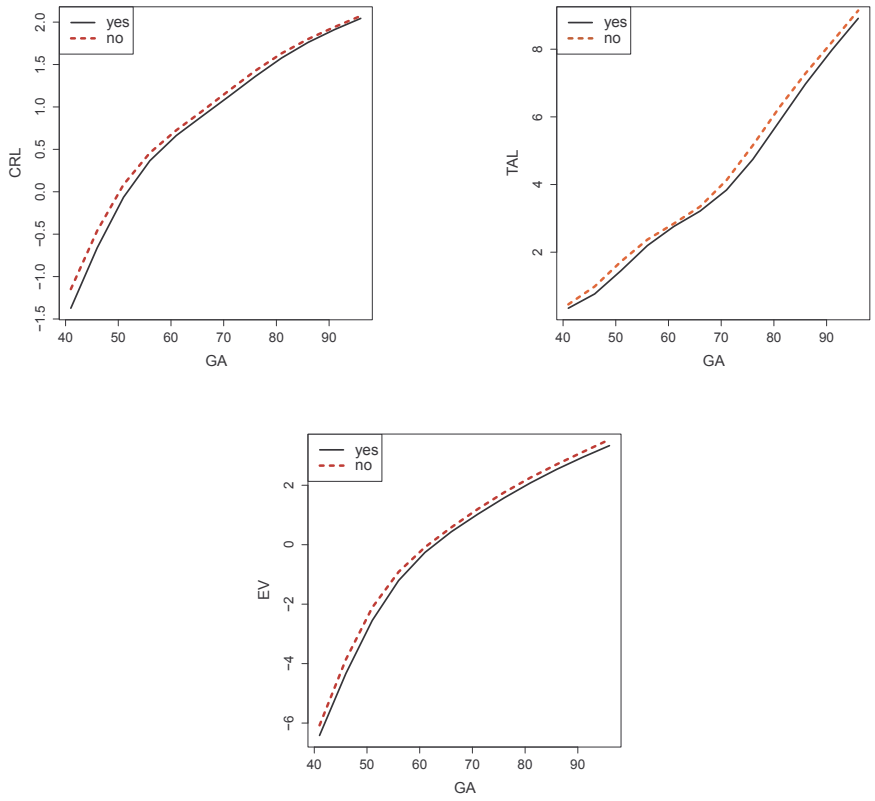
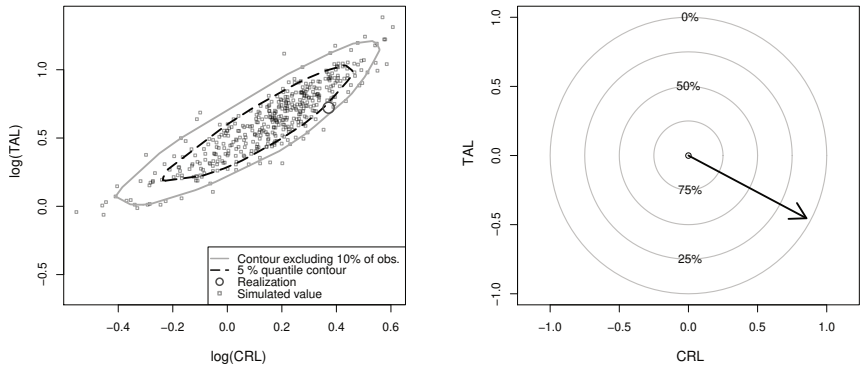


Figure 3.6: Effect of smoking

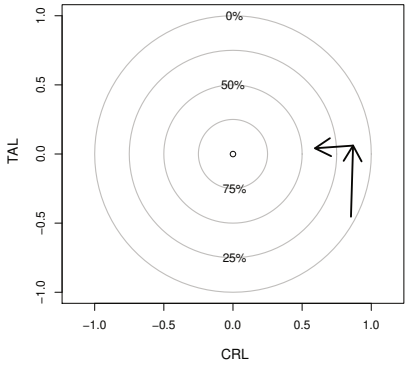
points, as is done in Figure 3.7(c), one can visualize the trajectory of an individual pregnancy relative to the whole group. In this case we see that although the pregnancy is quite outlying at the start, it becomes more ordinary later.

For a conditional reference contour, prediction is done assuming that we already have a number of ultrasound images from a certain pregnancy as well as background information such as smoking status, alcohol use and maternal age. Now we wish to know what measurement values would be considered as normal at a future time point. We again simulate replicate observations from the model but now we use the full conditional (3.4) to simulate the subject specific effects taking into account all the information we have for this pregnancy. Plots of the



(a) Simulated values and reference contours

(b) Direction and size of minimum depth



(c) Following an individual over time

Figure 3.7: Reference contours

conditional reference contours and the evolution of the multivariate quantile over time can be made in much the same way as for the unconditional contours and quantiles and are therefore not shown here.

Simplifying the random effect structure

Triggered by the high correlation between the horizontal effects we investigated if it would be reasonable to assume that these parameters are the same in each of the separate growth curves. Indeed when we simplified the model in this respect the DIC improved by 70 points in the model with normally distributed errors and by 15 points in the model with t-distributed errors.

3.5 Discussion and conclusions

We have extended the SITAR model of Cole et al. to include several outcomes, which we refer to as the MSITAR model. The model is quite flexible in modeling different growth patterns while at the same time remaining easy to interpret. In this paper we have applied this model on embryonic growth data, however it can also be used to model other kinds of growth data and even in completely different fields of study.

We have shown that the MSITAR model provides a better fit to the data of the Rotterdam Predict Study than the collection of univariate SITAR models. Oddly enough this does not result in an improved predictive ability. However a benefit of the MSITAR model is that, in case we have missing values in one of the series, we can recover some of that missing information from the other series. This is very relevant in a clinical setting where it is impossible to make frequent measurements of all outcomes. By means of the MSITAR model we can predict future values for an outcome when we have a single previous measurement for that outcome or even no previous measurements at all as long as we have measurements for other outcomes. Furthermore, the MSITAR model enables one also to look at the relation between the various series which is useful on its own.

Surely other techniques suitable to model flexibly multivariate longitudinal data are available, see e.g. [*Xu et al.*, 2008; *Rosen and Thompson*, 2009; *Macdonald-Wallis et al.*, 2012]. However we argue that the parameters of these models are less interpretable than those of the MSITAR model. A related technique is formed by the class of three-mode models discussed in [*Oort*, 2001]. This approach requires balanced measurements, which is often not the case in growth studies.

We have successfully used the MSITAR model to assess various potential determinants of embryonic growth. However because all three types of effect

influence the outcome simultaneously, we have to calculate the net effect at different time points as we did in Section 3.4, which diminishes the usefulness of the interpretation of the γ parameters.

The Bayesian procedure lends itself to the construction of reference contours in which all parameter uncertainty is taken into account. These can be used to detect multivariate outliers that would otherwise go unnoticed. Furthermore covariates and past measurements of a pregnancy can be incorporated so that the reference values can be individualized.

As is generally known, the Bayesian MCMC estimation procedure allows us to easily change some of model components, which is a great advantage over some other computational approaches. Here we replaced the usual normally distributed error term by t-distributed errors. In addition, we have examined the effect that covariates may have on early embryonic growth. Currently, we are also looking if we can use the subject specific effects in our model to predict birth outcomes. In the Bayesian methodology the individual estimates are automatically shrunken which is advantageous for prediction.

In theory the MSITAR model can be extended to more than three outcomes. However from a practical point we must admit that the model will quickly become too complex to fit. Still, to apply the MSITAR model to more than three dimensions, the current approach might be combined with a dimension reduction technique as in [*Slaughter et al.*, 2009]. Furthermore, here we have analyzed the 'normal' pregnancies. Another approach could be to take basically all pregnancies. This strategy might be preferred because in the early period after conception it is sometimes hard to know whether the pregnancy will be aborted. In fact, one of the reasons for setting up the Rotterdam Predict study is to discover abnormalities. Such a sample will require, however, that we model the random part of the MSITAR model more flexibly. We are currently working on these and other extensions in a Bayesian context.

Appendices

Introduction

In the paper 'A Multivariate Bayesian model for embryonic growth' we presented an extension of the SITAR model so that multiple outcomes could be modeled. Here we show some details that did not fit in the main paper, but which we think are interesting for some readers. First we give the full conditionals of the model. In section 3.B we show how we can sample from the predictive distribution. The last section of the supplementary material we look at the distribution of the subject specific effects and check whether they have a normal distribution.

3.A Full conditionals

We used a Bayesian approach to estimate the model parameters of the model (2) given by the following posterior distribution:

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}_M, \Sigma_{M_\gamma} \mid \boldsymbol{y}) \propto & \prod_{ijk} \mathrm{N}(y_{ijk} \mid \gamma_{i2k} + \boldsymbol{z}_{ijk}(\boldsymbol{\gamma}_{M_i})^\top \boldsymbol{\beta}_k, \sigma_k^2) \times \\ & \prod_i \mathrm{N}(\boldsymbol{\gamma}_{M_i} \mid \mathbf{0}, \Sigma_{M_\gamma}) \times \\ & \prod_k \mathrm{N}(\boldsymbol{\alpha}_k \mid \mathbf{0}, \sigma^2_{\alpha,0} I_{n_p}) \prod_k \mathrm{N}(\boldsymbol{\beta}_k \mid \mathbf{0}, \sigma^2_{\beta,0} I_{\kappa+2}) \times \\ & \prod_k \mathrm{IG}(\sigma_k^2 \mid \alpha_\sigma, \beta_\sigma) \, \mathrm{IW}(\Sigma_{M_\gamma} \mid \delta, \Psi), \end{aligned}$$

The notation used is explained in the main paper. Additionally we use:

$$\begin{aligned} \boldsymbol{\beta} &= (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top, \\ \boldsymbol{\sigma}^2 &= (\sigma_1^2, \ldots, \sigma_K^2)^\top, \\ \boldsymbol{\gamma}_M &= (\boldsymbol{\gamma}_{M_1}, \ldots, \boldsymbol{\gamma}_{M_N})^\top \text{and} \\ \boldsymbol{y} &= (y_{111}, y_{121}, \ldots, y_{1n_i1}, y_{211}, \ldots, y_{N,n_{NK},K})^\top. \end{aligned} \tag{3.1}$$

The (block) full conditionals for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ σ_k^2 and Σ_{M_γ} are:

$$
\begin{aligned}
\boldsymbol{\beta}_k \mid \boldsymbol{y}, \ldots &\sim \mathrm{N}(\overline{\mu}_k, \overline{\Sigma}_{\beta_k}),\\
\text{where} \quad \overline{\mu}_k &= \frac{1}{\sigma^2}\overline{\Sigma}_{\beta_k}\overline{Z}_k\widetilde{\boldsymbol{y}}_k,\\
\overline{\Sigma}_{\beta_k} &= \left(\Sigma_{\beta}^{-1} + \overline{Z}_k^\top \overline{Z}_k/\sigma_k^2\right)^{-1};\\
\boldsymbol{\alpha} \mid \boldsymbol{y}, \ldots &\sim N(\overline{\mu}_\alpha, \overline{\Sigma}_\alpha),\\
\text{where} \quad \overline{\mu}_\alpha &= \overline{\Sigma}_\alpha[(X^\top \otimes \Sigma_{M\gamma}^{-1})\boldsymbol{\gamma}_{M,\text{vec}}],\\
\overline{\Sigma}_\alpha &= [X^\top X \otimes \Sigma_{M\gamma}^{-1} + I_{n_p \times K}\sigma_{\alpha,0}^2]^{-1};\\
\sigma_k^2 \mid \boldsymbol{y}, \ldots &\overset{\text{iid}}{\sim} \mathrm{IG}\left(\alpha_\sigma + \frac{K}{2},\ \beta_\sigma + \frac{\sum_{i,j}\{y_{ijk} - \gamma_{i2k} - B(exp(\gamma_{i3k})(t_{ij} + \gamma_{i1k}))\}^2}{2}\right);\\
\Sigma_{M\gamma} \mid \boldsymbol{y}, \ldots &\sim \mathrm{IW}\left(\delta + n_i,\ (\Psi^{-1} + (\gamma_M - XA^\top)^\top(\gamma_M - XA^\top))^{-1}\right).
\end{aligned}
$$

In the above expressions, $\overline{\mu}_k$ and $\overline{\Sigma}_{\beta_k}$ are the mean and the variance of the full conditional distribution of the spline coefficients $\boldsymbol{\beta}_k$. The $(\sum_1^N n_{ik} \times (\kappa + 2))$ matrix $\overline{Z}_k$ is the spline basis with as rows the $\boldsymbol{z}_{ijk}$ at the current values of the parameters γ_{1k} and γ_{3k}. Further, $\widetilde{\boldsymbol{y}}_k$ is defined as $(y_{11k} - \gamma_{12k}, y_{12k} - \gamma_{12k}, \ldots, y_{1n_1k} - \gamma_{12k}, y_{21k} - \gamma_{22k}, \ldots, y_{Nn_Nk} - \gamma_{N2k})^\top$, $\overline{\mu}_\alpha$ and $\overline{\Sigma}_\alpha$ denote the mean and variance of the full conditionals of $\boldsymbol{\alpha}$. We also used the notation: X for the $N \times n_p$ design matrix $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top$ and $\boldsymbol{\gamma}_{M,\text{vec}} = [\boldsymbol{\gamma}_{M_1}^\top, \ldots, \boldsymbol{\gamma}_{M_K}^\top]^\top$ for the matrix consisting of the stacked rows of $\boldsymbol{\gamma}_M$ and the symbol '$\otimes$' for the Kronecker product.

The above full conditionals are of standard form and hence the usual samplers can be used. However, this is not the case for the full conditional distribution of the subject specific effects, given by:

$$
\boldsymbol{\gamma}_{M_i} | \boldsymbol{y}, \ldots \sim N_{3K}(\boldsymbol{\gamma}_{M_i} \mid A\boldsymbol{x}_i, \overline{\Sigma}_\gamma) \times \prod_{j,k} \mathrm{N}(y_{ijk} \mid \gamma_{i2k} + \gamma_{i3k}B(t_{ij} + \gamma_{i1k})\boldsymbol{\beta}_k, \sigma_k^2), \quad (3.2)
$$

where a Metropolis step is needed.

3.B Sampling from the predictive distribution

We perform cross-validation to assess the predictive ability of the models. We split the data in five approximately equal parts. In turn one of the parts will be the validation set and the other the training sets. For each individual in the

validation set we want to draw samples from the predictive distribution of the last observation in each series given the observations that were made previously and also given the information we have obtained from the individuals in the training set.

We denote the last observation from each combination of pregnancy and outcome by $y_{ik,\text{LAST}}$. This is the observation we want to predict. The earlier observations for this pregnancy on which this prediction will be based (in addition to the information in the training set) are $\boldsymbol{y}_{i,O}$. Here we use the subscript O to denote that these are the Observed measurements of the new pregnancy.

We can draw samples from the predictive distribution of $y_{ik,\text{LAST}}$ as follows:

- Based on the data of the training set, denoted as $\boldsymbol{y}_T$, we have stored the sampled values of all the parameters denoted as $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{N_T}$, where N_T denotes the number of stored samples.

- For a pregnancy in the validation set, we determined the multivariate distribution of $\boldsymbol{\gamma}_{Mi}$ from expression (3.2) in the main paper, based on both the sampled $\tilde{\boldsymbol{\theta}}$ parameters and the observed $\boldsymbol{y}_{i,O}$. This results in N_T sampled parameters $\tilde{\boldsymbol{\gamma}}_{Mi,1}, \ldots, \tilde{\boldsymbol{\gamma}}_{Mi,N_T}$. Note that $\boldsymbol{y}_{i,O}$ is not used to update $\boldsymbol{\theta}$. The idea is that the observations for each pregnancy in the validation set are only used to learn its specific effects but not to update our knowledge of the other parameters.

- Using these subject specific effects, we now predict the values at the last observation of each outcome k of each pregnancy i ($y_{ik,\text{LAST}}$) in the validation sample. That is, we predict the values $y_{ik,\text{NEW}}$ from $N(\tilde{\gamma}_{i2k} + \tilde{\gamma}_{i3k}B(t_{ijk} + \tilde{\gamma}_{i1k}), \tilde{\sigma}_k^2)$. The predicted response is based on the sampled values $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}_{Mi}$ and results in N_T sampled values $\tilde{y}_{ik,\text{LAST},1}, \ldots, \tilde{y}_{ik,\text{LAST},N_T}$. We then took the average over these samples to obtain the predicted values $\hat{y}_{ik,\text{LAST}}$.

The procedure outlined above is followed for all pregnancies using each of the parts as validation sample in turn. We calculated the RMSE for each of the outcomes as $\sqrt{\sum_{i=1}^{N_k}(y_{ik,\text{LAST}} - \hat{y}_{ik,\text{LAST}})^2/N_k}$. Here N_k are the number of pregnancies for which we observed the kth outcome.

3.C Check of the normality assumption of the distribution of the subject specific effects

We have shown that there were advantages of modeling the outcomes together instead of separately and checked the assumptions of our models by using posterior predictive checks. From these checks, we saw that the fit could be improved by changing the distribution of the residual error from a normal distribution to a t-distribution. Because the usual posterior predictive checks focus on the residual error we used another approach to look at the distribution of the subject specific effects. Here, quantile-quantile plots (qq-plots) of the subject specific effects for all models are fitted. In turn we present the qq plot of the univariate model with normally distributed errors (Figure 3.1), the multivariate model with normal errors (Figure 3.2), the univariate models with t-distributed errors (Figure 3.3) and the multivariate model with t-distributed errors (Figure 3.4). The qq-plots compare the sample quantiles of the fitted subject specific effects with the theoretical quantiles of the assumed distribution. Note that we do not actually observe the subject specific effects which reduces our ability to detect deviation from the assumptions.

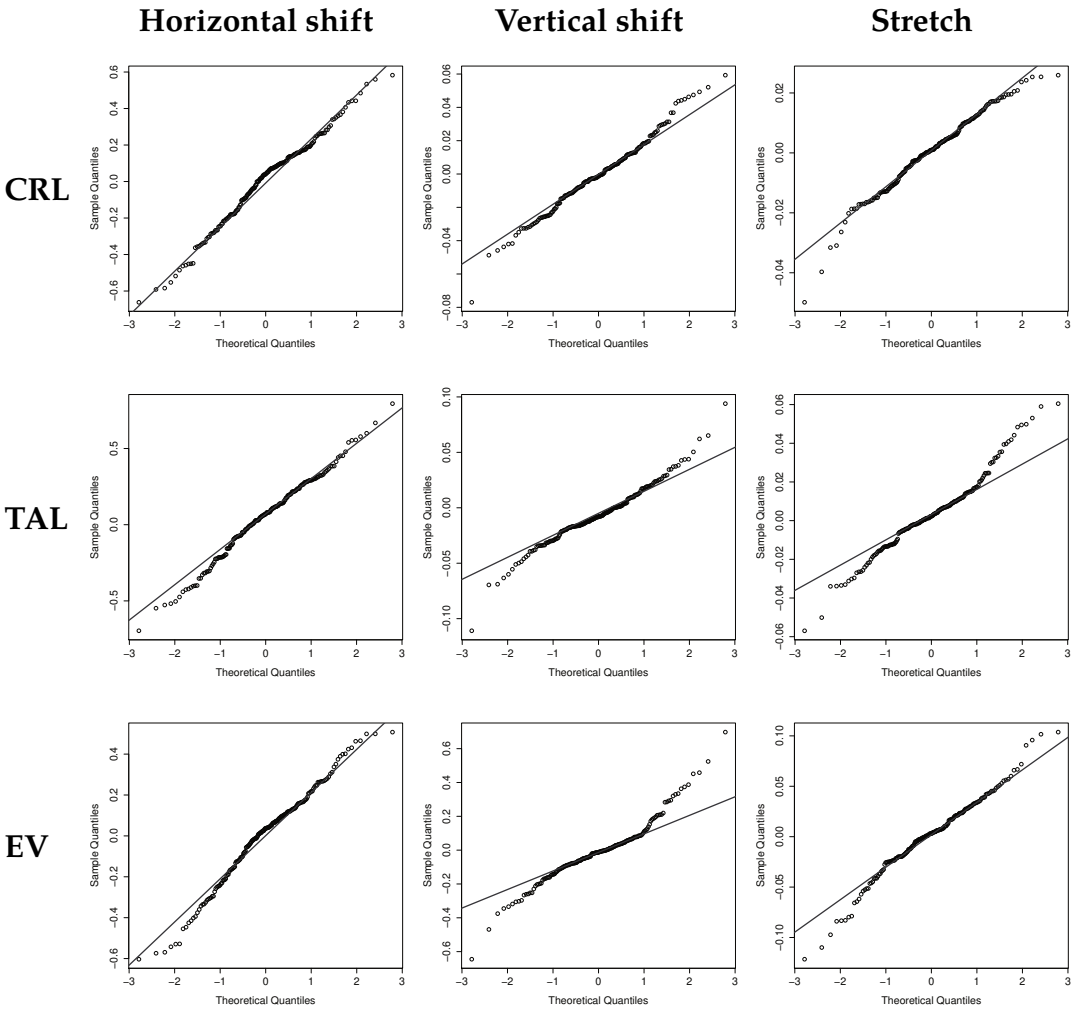


Figure 3.1: QQ-plots for the univariate SITAR models with normal distributed errors

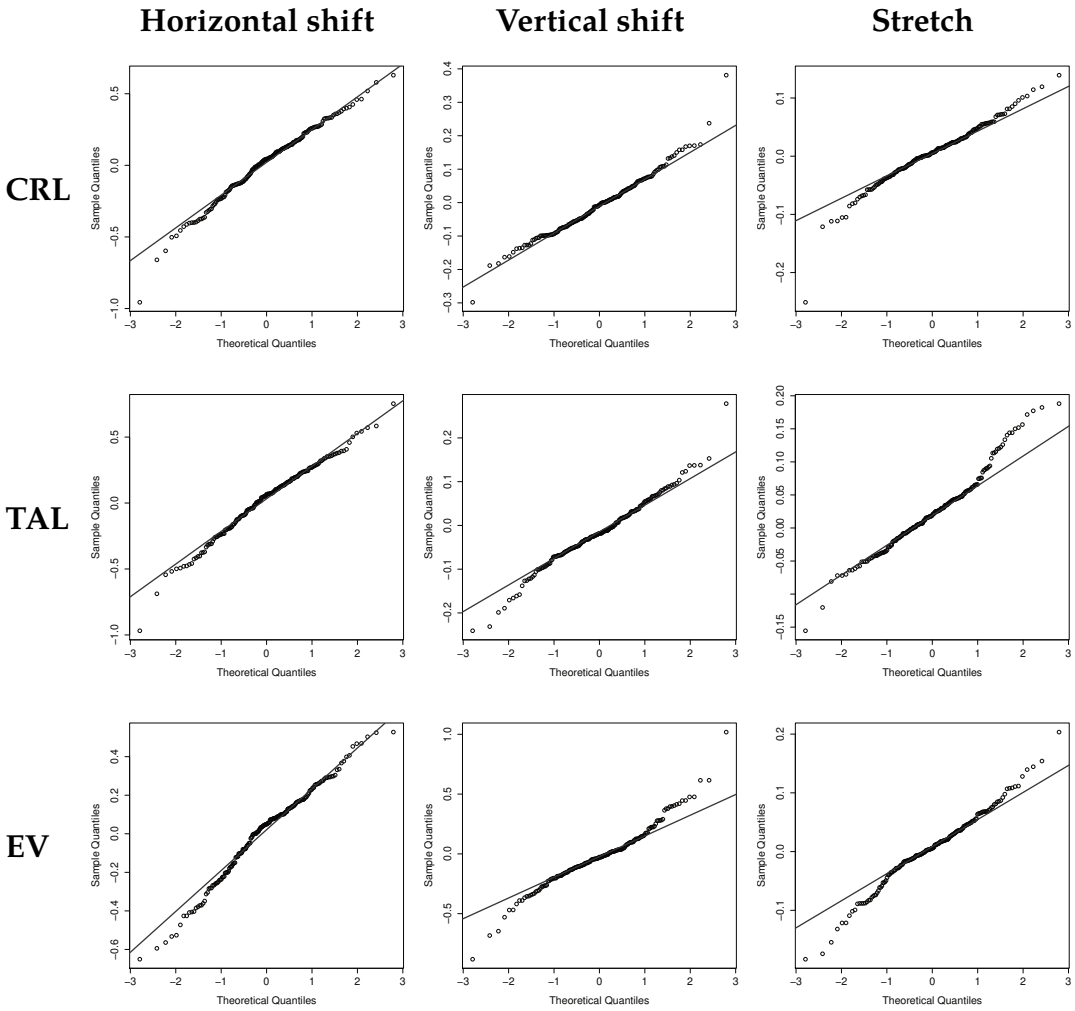


Figure 3.2: QQ-plots for the multivariate SITAR model with normally distributed errors

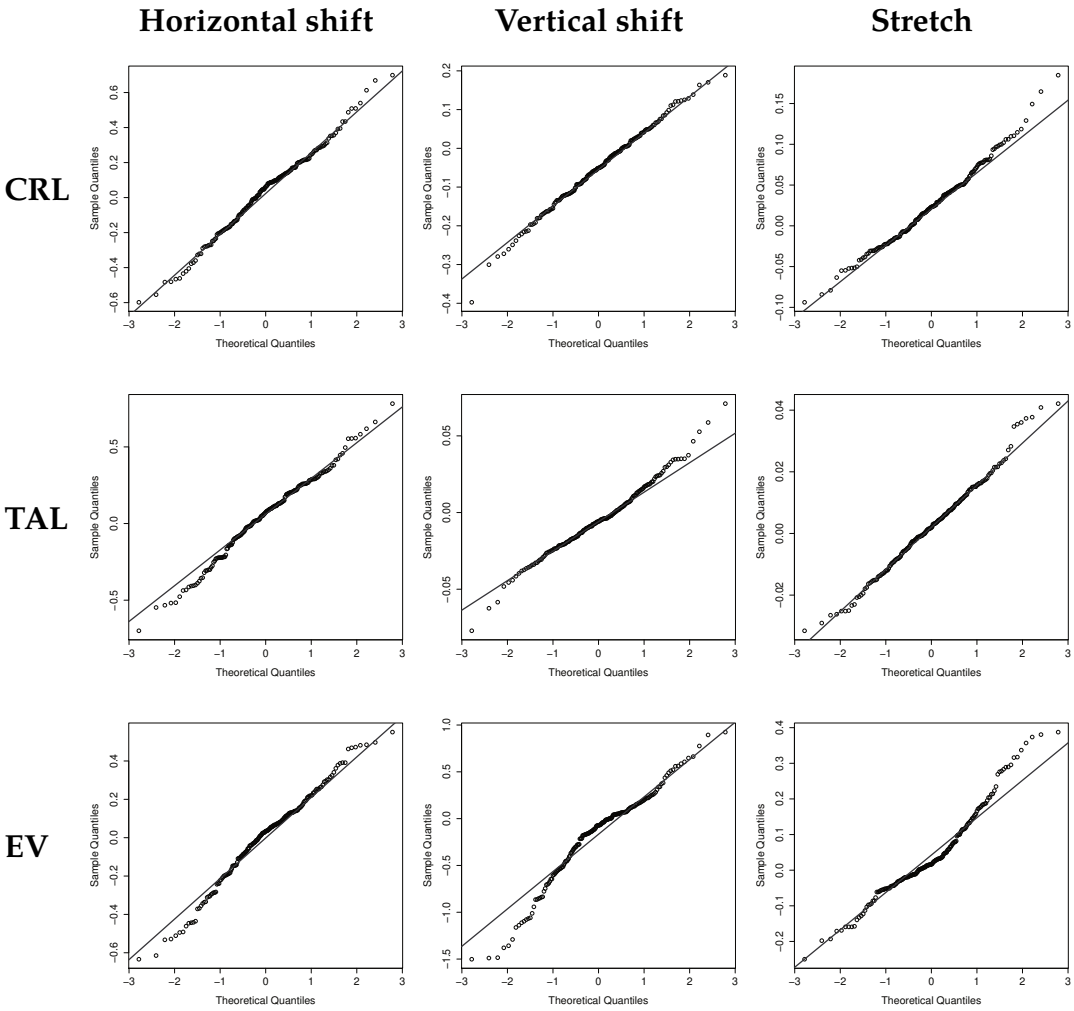


Figure 3.3: QQ-plots for the univariate SITAR models with t-distributed errors

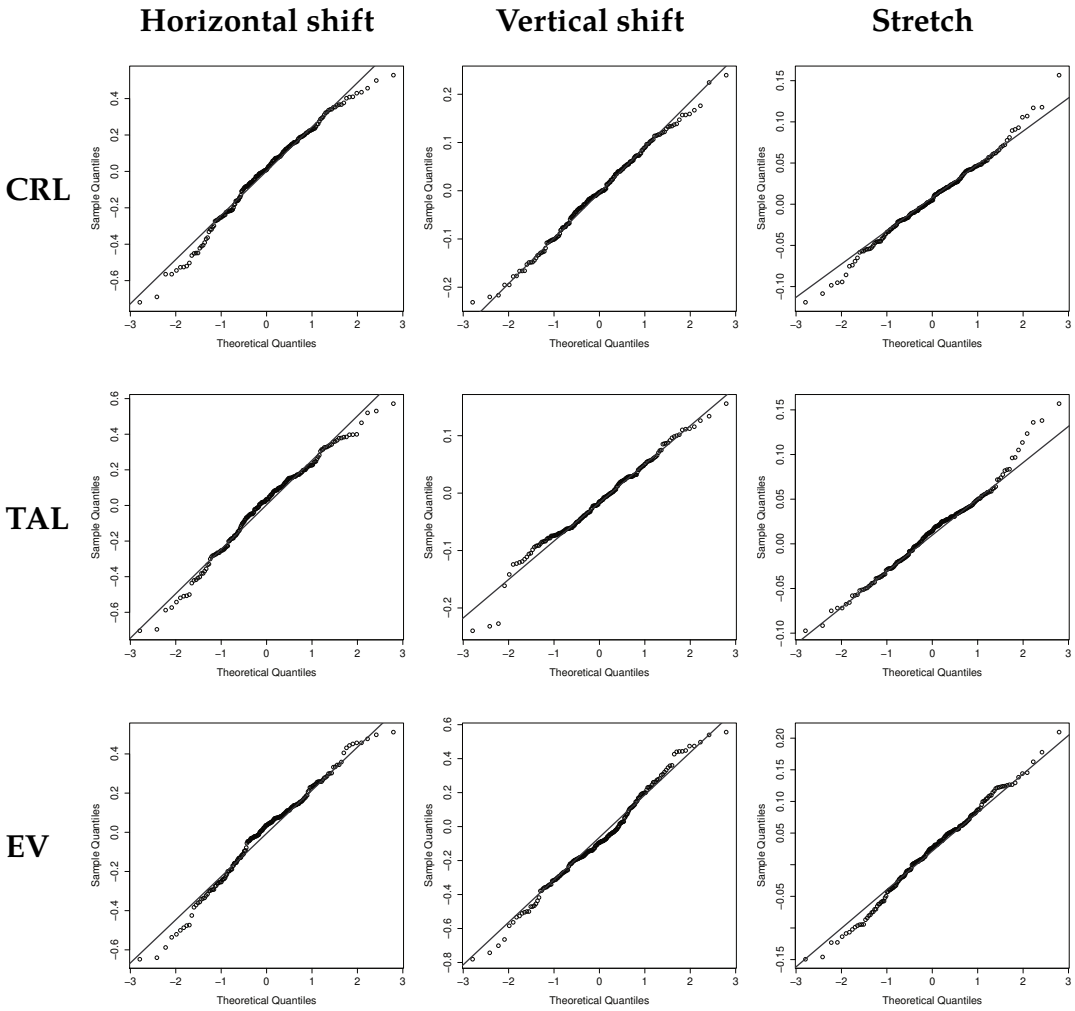
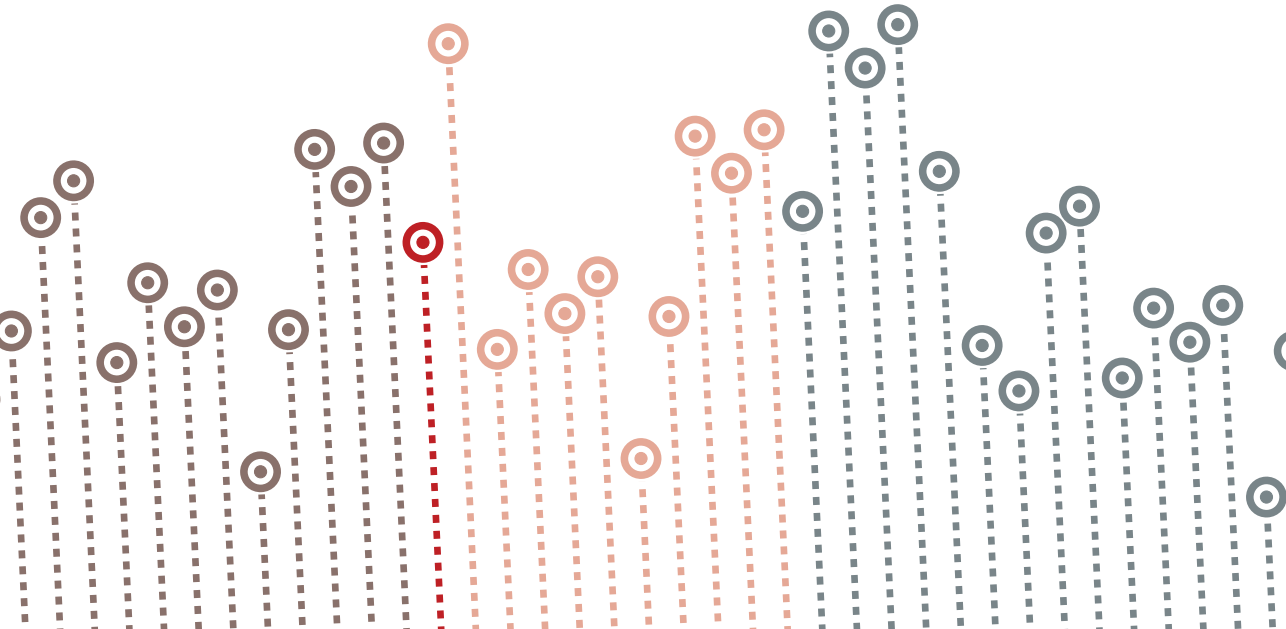


Figure 3.4: QQ-plots for the multivariate SITAR model with t-distributed errors

4 Flexible multivariate nonlinear models for bioequivalence problems



Abstract:

Modeling the concentration of a drug in the bloodstream over time is usually done using compartment models. In pharmacokinetic data they turn into highly nonlinear mixed-effects models (NLMEMs) when we take the heterogeneity between subjects into account. Fitting of NLMEMs can be difficult and may involve complex algorithms, with convergence critically depending on the initial values and maybe requiring data transformations. In this paper, we propose a flexible alternative to the usual parametric compartment models, inspired by the Multivariate SuperImposition by Translation and Rotation (MSITAR) model but adapted to be applicable in this new field. A fully parametric one-compartment nonlinear mixed-effects model is considered for comparison. We make use of a Bayesian approach and illustrate the method on a real data set where the interest lies in contrasting the average and individual bioequivalence of a test and reference formulation of an antihypertensive drug.

Authors: S.P. Willemsen, C.M. Russo, E.M. Lesaffre & D Leão. published in: Statistical Modelling (2017)

4.1 Introduction

The importance of mixed-effects models for analysing longitudinal data and repeated measurements is unquestionable. When the relation between covariates and the response is nonlinear in the parameters the modeling of repeated outcomes leads to nonlinear mixed-effects models (NLMEMs). However fitting these models is notoriously difficult, usually involving iteratively applying numerical integration to maximize the likelihood and requiring well-chosen initial values and even data transformations. Challenging applications are found in the pharmacokinetic area, where highly nonlinear models arise naturally from biological theory about the absorption and elimination of a substance in the body. The nonlinear functions considered, however, may be too restrictive, not allowing much flexibility. In practice, this means that if the nonlinear function is misspecified or if the initial values for the algorithm are chosen badly, a good fit may never be achieved. Important references on nonlinear parametric mixed effects models are [*Kuhn and Lavielle*, 2005; *Wu*, 2004; *Lindstrom and Bates*, 1990]. The particular problem of estimation is intensively discussed by *Pinheiro and Bates* [1995], where different approximations to the log-likelihood are proposed.

Prior to introducing a new drug on the market the manufacturer needs to show, among other things, that it is metabolised in the same way as an equivalent drug that already holds a patent. An example is the proposal of a generic drug as an alternative to a brand-name drug. To achieve this, pharmacokinetic models are used to describe the processes of absorption, distribution and excretion of the drug in the human body. We want to know what portion of the drug reaches the patient's circulation and thus is available to the organs, which is called the bioavailability. When the bioavailability of two drug formulas is equal they are called bioequivalent.

When a series of blood concentration measurements is available from subjects who have taken the original and new drug formulation, pharmacokinetic models could be used to demonstrate that the whole curves of drug concentration as a function of time do not materially differ between the two drugs formulations. The regulatory agencies FDA and EMEA recommend the comparison of two bioavailability parameters: the maximum drug concentration achieved (C_{max}) and the area under the drug concentration-time curve (AUC), see Figure 4.1. To demonstrate population bioequivalence, the ratio of the bioavailability parameters

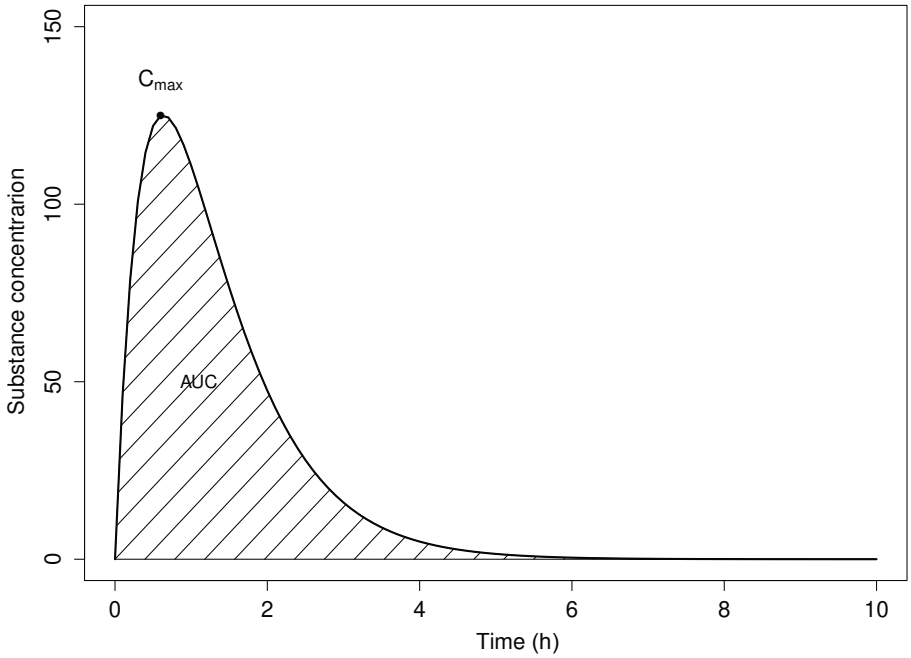


Figure 4.1: Illustration of the Bioavailability parameters C_{max} and AUC.

for the two drugs should be close to one [*U. S. Food and Drug Administration*, 2001; *European Agency for the Evaluation of Medicinal Products*, 2001]. Specifically, FDA and EMEA recommend to check if

$$0.8 \leq \frac{C_{max,2}}{C_{max,1}} \leq 1.25 \text{ and } 0.8 \leq \frac{\text{AUC}_2}{\text{AUC}_1} \leq 1.25. \tag{4.1}$$

with a high degree of certainty. In this equation we use subscript 1 for the reference drug and 2 for the new drug. Usually this is tested by carrying out two one-sided t-tests for each outcome both at the 5% level, which is equivalent to the 90% CI lying entirely within these boundaries.

Notwithstanding this recommendation, there have been attempts to use nonlinear mixed-effects compartmental models to assess bioequivalence in the strict sense [see e.g., *Chen and Huang*, 2013; *Fradette et al.*, 2005; *Hu et al.*, 2004]. A comparison between a non-compartmental analysis (NCA) and a nonlinear mixed-effects model (NLMEM) is presented in [*Dubois et al.*, 2010]. More recently, a stochas-

tic approximation expectation maximization (SAEM) algorithm was proposed by *Dubois et al.* [2011]. The authors also proposed a Wald test to assess bioequivalence in a strict sense.

Note that condition (4.1) only deals with average bioequivalence not with individual bioequivalence. Hence, even when the drug formulations behave similar on average, this does not mean that the formulations behave the same for each individual patient. It is this individual bioequivalence that is relevant for the ease with which patients that currently use the reference drug can switch to a new drug formula.

In this paper we propose a model that is data driven rather than based on a theoretical biological model. Our model relies on the Multivariate SITAR (MSITAR) model, recently proposed by [*Willemsen et al.*, 2015] for embryonic growth data. The MSITAR model is a multivariate extension of the SuperImposition by Translation and Rotation (SITAR) model suggested by *Cole et al.* [2010a]. The price of the flexibility is that the regression parameters no longer have a pharmacokinetic interpretation. However the MSITAR model still allows for performing tests of bioequivalence based on credibility intervals of the bioavailability parameters C_{max} and AUC. Moreover, here the MSITAR model gives a better fit and consequently a greater reliability on the bioequivalence conclusion may be obtained.

This paper unfolds as follows. Section 4.2 introduces the motivating example a pharmacokinetic trial investigating the bioequivalence of two formulas of Losartan. In Section 4.3 we briefly discuss the parametric nonlinear mixed-effects models that are used for pharmacokinetic data and propose a SITAR-type model as an alternative to analyze these nonlinear longitudinal pharmacokinetic data. Section 4.4 details some of our modeling choices, explains how we use the models to evaluate bioequivalence and discusses the way that we have evaluated our models. Section 4.5 is dedicated to the analysis of the application and a discussion is provided in Section 4.6.

4.2 Motivating example

A bioequivalence trial was set up to compare the pharmacokinetic profile of a generic formulation (test drug) of Losartan, an anti-hypertensive drug, and the original Losartan formulation (reference drug). To this end, a crossover randomized study with 24 healthy volunteers was set up. Reference and test

formulations of the drug were administered to each individual on different days in a randomized order, ensuring no carry-over effect. The drug concentration was observed at fixed time points: 0, 20min, 30min, 40min, 60min, 80min, 100min, 2h, 2.5h, 3h, 3.5h, 4h, 4.5h, 5, 5.5h, 6h, 6.5h, 7h, 8h, 9h, 10h, 14h and 24h. In Figure 4.2 we show the pharmacokinetic profiles under both formulations for two randomly selected individuals. The first measurement was taken when the drug was administered and the last measurement was taken 24 hours later. We see that all profiles start at zero from which they increase quickly to reach a peak. From there, they gradually decrease again. Note that there are some individuals whose concentration remains at zero for a short amount of time after the start of the experiment. Also note that there is variation in the height of the peak and the speed at which the concentration diminishes after the peak. In Table 4.1 the means (SD) of the blood-serum measurements at each chosen time point under the two drug formulations are shown.

The measurement technique involved has a detection limit of 2 ng/ml and all values below this threshold are therefore represented as zeros.

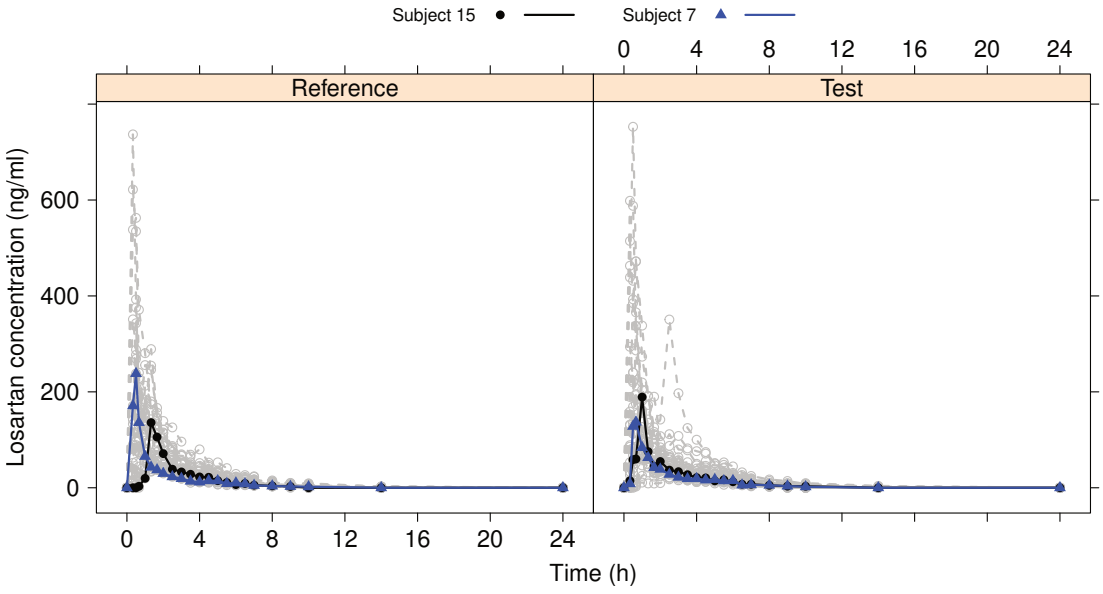


Figure 4.2: Measurements of Losartan concentration in 24 volunteers using two drug formulations, with the measurements of two randomly chosen individuals highlighted.

Table 4.1: Descriptive statistics of the pharmacokinetic Losartan data. The mean, standard deviation and median of the two formulas at all time points are shown.

	Reference drug			Test drug		
Time	mean	SD	median	mean	SD	Median
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.33	134.71	209.15	41.25	138.15	183.67	48.05
0.50	178.28	168.21	106.30	182.72	198.96	120.47
0.66	123.42	89.22	114.94	162.34	143.16	124.80
1.00	119.83	71.02	122.09	125.84	78.69	107.22
1.33	108.83	74.35	72.94	91.95	48.14	85.58
1.66	80.84	42.41	64.83	66.69	40.78	54.98
2.00	67.43	33.33	57.85	58.69	29.46	54.62
2.50	53.07	26.41	44.13	58.71	66.13	40.27
3.00	40.30	19.76	34.90	48.10	38.13	35.92
3.50	32.16	16.36	28.26	37.70	25.35	33.39
4.00	27.32	16.22	22.27	30.77	20.71	24.85
4.50	26.78	11.77	23.92	30.78	17.01	25.36
5.00	22.42	10.93	20.95	24.46	11.90	20.70
5.50	17.12	7.95	15.35	23.54	12.24	19.19
6.00	13.43	6.20	13.16	17.67	8.54	14.55
6.50	12.90	5.70	11.07	13.14	6.62	11.57
7.00	10.71	5.67	9.02	10.45	5.52	8.77
8.00	6.78	3.70	5.53	6.83	3.69	5.93
9.00	4.87	2.56	4.64	4.65	3.25	4.29
10.00	2.92	2.64	2.73	3.06	2.71	2.95
14.00	0.40	0.93	0.00	0.38	1.04	0.00
24.00	0.00	0.00	0.00	0.00	0.00	0.00

4.3 Approaches to assess bioequivalence

In this section we introduce the first order compartment model, which is one of the most used pharmacokinetic models that can assess bioequivalence as specified in (4.1). We then briefly discuss non-compartment analysis, which is a non-parametric alternative to determine bioequivalence. A third alternative to assess bioequivalence is based on a multivariate extension and adaptation of the SITAR model. The correlation among the repeated drug concentrations is taken care of by random effects.

First order compartment model

The absorption and elimination of a substance by the body is usually modelled with a compartment model. One distinguishes between single- and multicompartment models. In the single compartment model, the human body is represented as a single compartment that is absorbing the drug and from which it is naturally eliminated afterwards. When absorption and elimination occur at a constant rate the following equation forms the basis of the first order open compartment model:

$$\frac{dX(t)}{dt} = k_a FD - k_e X(t) \text{ and } Y(t) = X(t)/V. \tag{4.2}$$

Here $X(t)$ is the amount of drug in the blood at time t, F represents the fraction of the drug that is absorbed), D the administered dose, k_a the absorption rate, k_e the elimination rate, V the volume of distribution and $Y(t)$ the drug concentration in the serum. This model (plus measurement error) can be rewritten as:

$$Y(t) = \frac{FDk_a}{Vk_a - Vk_e}[\exp(-k_e t) - \exp(-k_a t)] + \varepsilon.$$

This expression is called the integrated form. Measurement error ε is introduced because the observed concentration levels deviate from the theoretical biological model because of e.g. measurement error. As a function of time t, the drug concentration takes the form of a constant times the difference between two exponential terms.

One clear advantage of this parametric model is the interpretability of its parameters. Although neglected by some authors, the parameters should be positive, which is achieved by working with parameters defined on a logarithm scale: $lKa = log(k_a)$, $lKe = log(k_e)$ and $lCl = log(Cl) = log(k_e V)$. The parameter Cl is called the clearance.

Note that D and F and Cl^{-1} all have the same multiplicative effect on Y and therefore cannot be estimated separately in this experiment. In the remainder they are therefore absorbed in the parameter Cl. We now arrive at the following equation:

$$Y(t) = \exp(lKa + lKe - lCl)\frac{[\exp(-e^{lKe}t) - \exp(-e^{lKa}t)]}{e^{lKa} - e^{lKe}} + \varepsilon. \tag{4.3}$$

The parameters in this case are interpreted as the logarithm of the substance absorption rate (lKa), the logarithm of the substance elimination rate (lKe) and

the logarithm of plasma clearance (lCl) [*Pinheiro and Bates*, 2000]. Notice that in this model the parameters are not fully identified. When lKa and lKe are reversed the profile stays the same. We will assume here that $lKa > lKe$.

The parameters in model (4.3) usually vary between individuals. Here we will extend the model to incorporate the multilevel structure of the data. We assume a lognormal distribution for k_e, k_a and Cl (and hence a normal distribution for lKe, lKa and lCl) as these parameters are usually lognormally distributed in the population. Denote by $\boldsymbol{y}_{ik} = (y_{i1k}, \ldots, y_{im_ik})^\top$ the vector of m_i observed substance concentrations in the ith subject after the administration of the k-th drug, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. In the motivating example, $K = 2$ and $k = 1$ for the reference drug and $k = 2$ for the test drug. Thus the model for the jth response of the kth drug formulation of the ith individual is given by

$$y_{ijk} = \exp(lKa_{ik} + lKe_{ik} - lCl_{ik}) \frac{[\exp(-e^{lKe_{ik}} t_{ij}) - \exp(-e^{lKa_{ik}} t_{ij})]}{e^{lKa_{ik}} - e^{lKe_{ik}}} + \epsilon_{ijk}, \quad (4.4)$$

where $lKa_{ik} = lKa_k + \gamma_{i1k}$, $lKe_{ik} = lKe_k + \gamma_{i2k}$ and $lCl_{ik} = lCl_k + \gamma_{i3k}$; lKa_k, lKe_k and lCl_k are the fixed-effects related to the k-th response; γ_{i1k}, γ_{i2k} and γ_{i3k} are the respective random effects. The usual assumption for the vector of random effects related to the i-th individual is to assume a normal distribution for $\boldsymbol{\gamma}_i$ where $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}^\top, \ldots, \boldsymbol{\gamma}_{iK}^\top)^\top$ and $\boldsymbol{\gamma}_{ik} = (\gamma_{i1k}, \gamma_{i2k}, \gamma_{i3k})^\top$ is the vector of random effects related to the k-th response of the i-th individual for each $k = 1, \ldots, K$ and $i = 1, \ldots, n$. The residual random errors $\epsilon_{ijk} \sim N(0, \sigma_k^2)$ are assumed to be independent of each other and of the random effects. Note that the dependence between the two formulas is introduced by modeling the stacked vector of random effects from both drugs as one (possibly correlated) multivariate normally distributed parameter. Conditional on these random effects the concentrations of the two drug formulas are assumed to be independent.

Figure 4.3 describes the effect of including each of the random effects in a fully parametric nonlinear mixed-effects model. We see that lCl stretches the whole curve upwards. The greater lKe the steeper the decrease at later time points and the lower the initial slope at the start. A higher lKa renders the initial slope steeper.

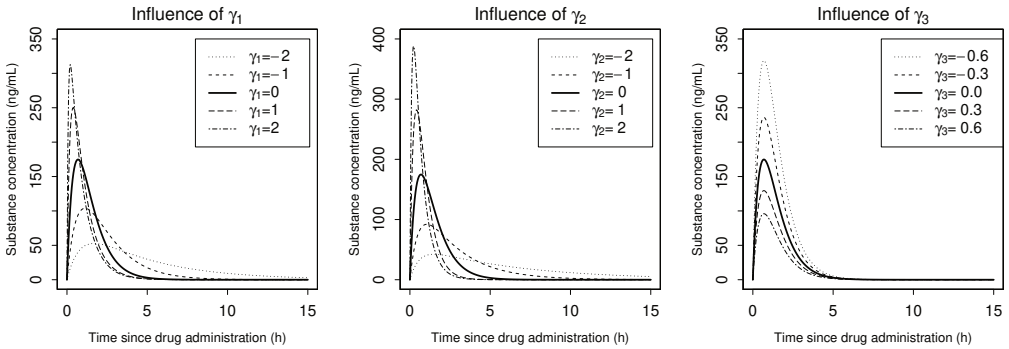


Figure 4.3: Random effects interpretation for the compartment model.

Non-compartmental models

In the previous section we have explained pharmacokinetic compartment models in the context of bioequivalence tests. However in practice, the bioequivalence of drugs is usually assessed by a Non-Compartment Analysis (NCA). In this approach the AUC is obtained by applying a quadrature rule on the observed concentrations and C_{max} is estimated by the maximum of the observed concentrations. This approach does not assume a particular form of parametric model and thus fewer assumptions are required to perform NCA when compared to model-based approaches. On the other hand the number of observations per subject should be sufficiently high [see, for example, *Dubois et al.*, 2010] because in a small sample the estimates may be biased. NLMEMs, which are fully parametric, are often more efficient. In this paper we are also interested in the entire evolution of the drug concentration over time and not only in summary statistics of the profiles. Therefore we will not pursue NCA further in the remainder of this paper.

The SITAR model

The SITAR model proposed by *Cole et al.* [2010b] is a versatile tool to model non-linear functions to longitudinal data sets suggested in the context of growth models. The SITAR approach consists of shifting and/or stretching a longitudinal template curve in different directions, delivering flexible fitted subject-specific profiles guided by the data variability.

Let y_{ij} be the response at the time t_{ij} of individual i. The original SITAR model is then given by the following equation:

$$y_{ij}|\boldsymbol{\gamma}_{Mi}, \boldsymbol{\beta}, \sigma_k^2 \sim N(T_{ij}^\top \boldsymbol{\beta} + \gamma_{i2}, \sigma_k^2),\ i = 1, \ldots, n;\ j = 1, \ldots, m_i,$$
$$T_{ij} = B(\exp(\gamma_{i3})[t_{ij} + \gamma_{i1}]),$$

where T_{ij} is the vector representing the B-spline expansion of a time for individual i at occasion j. The function carrying out the expansion is denoted by B, $\boldsymbol{\beta}$ is the vector of spline coefficients, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3})^\top$ is the vector of random-effects, consisting of γ_{i1} a horizontal shift, γ_{i2} a vertical shift and γ_{i3} a horizontal stretch. Originally a restricted cubic spline was used, which implies that the second derivative of the spline is zero at the outer knots and the spline is linearly extrapolated outside these values. The parameter σ_k^2 is the variance of the residual error of the model. A multivariate version of the SITAR model (MSITAR) was proposed recently by *Willemsen et al.* [2015].

The original SITAR model is however not entirely appropriate to model pharmacokinetic data, therefore we will make a slight modification. In Figure 4.2, it can be seen that some profiles are stretched out. This justifies the use of both horizontal and vertical stretch random effects. As all profiles start at a concentration of zero and asymptotically approach zero, a vertical shift does not make sense. For these reasons we adapted the original SITAR model to our needs here. We also adapted the restrictions of the spline to better match the pharmacokinetic purpose, i.e. we restrict the first derivative (as well as the second) to zero at the right outer knot and define the spline to be zero at and beyond the outer knots. Because all curves start at zero we also restricted the horizontal shift to be negative or zero. A negative shift stands for a delay in the absorption into the bloodstream as is observed in some patients. We now arrive at a modified SITAR model better suited for pharmacokinetic data:

$$\begin{aligned} y_{ij}|\boldsymbol{\gamma}_{Mi}, \boldsymbol{\beta}, \sigma^2 &\sim N(\exp(\gamma_{i2})[T_{ij}^\top \boldsymbol{\beta}], \sigma^2), \\ T_{ij} &= B(\exp(\gamma_{i3})[t_{ij} + \min(\gamma_{i1}, 0)]). \end{aligned} \tag{4.5}$$

The adapted multivariate MSITAR model is now given by:

$$\begin{aligned} y_{ijk}|\boldsymbol{\gamma}_{Mi}, \boldsymbol{\beta}_k, \sigma_k^2 &\sim N(\exp(\gamma_{i2k})[T_{ijk}^\top \boldsymbol{\beta}_k], \sigma_k^2), \\ T_{ijk} &= B(\exp(\gamma_{i3k})[t_{ij} + \min(\gamma_{i1k}, 0)]), \end{aligned} \tag{4.6}$$

where y_{ijk} be the kth response at the time t_{ij} of individual i $i = 1, \ldots, n;\ j = 1, \ldots, m_i$ and $k = 1, \ldots, K$. T_{ijk} is a matrix with bases of cubic splines for the kth

response of the individual i at time j, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kL})^\top$ is the vector of spline coefficients related to the k-th response (here L denotes the degrees of freedom of the spline), $\boldsymbol{\gamma}_{ik} = (\gamma_{1ik}, \gamma_{2ik}, \gamma_{3ik})^\top$ is the vector of random-effects for individual i.

Note that, although the parameters in the MSITAR model have a clear interpretation, their interpretation is no longer based on pharmacokinetic theory.

Figure 4.4 describes the effect of each of the subject-specific effects in the adapted variant of the MSITAR model. Notice that in the one-compartment model the function of the clearance is like the vertical stretch.

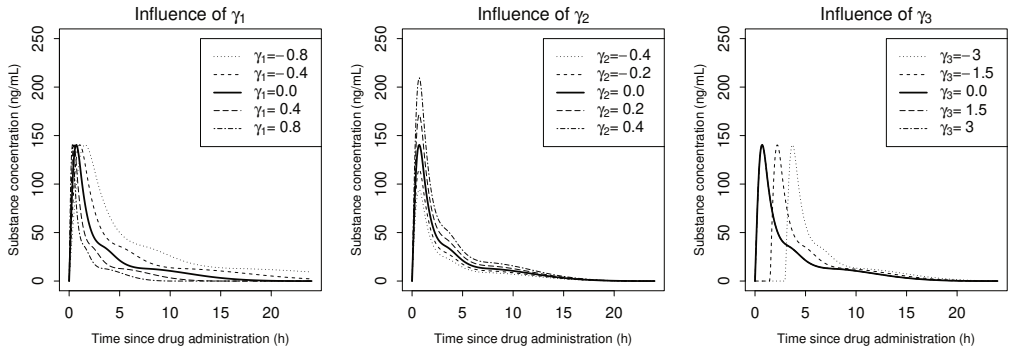


Figure 4.4: Random effects interpretation of the horizontal stretch γ_1, the vertical stretch γ_2 and the horizontal shift γ_3 in the SITAR model.

The first random effect represents a horizontal shift, the second random effect a vertical stretch, and the third random effect a horizontal stretch.

To give a visual representation of both the MSITAR and the compartmental models we have provided the directed acyclic graphs as Appendix 1.

4.4 Modeling aspects

In this section we give further details about how we fitted models (4.4) and (4.6) to the data of the Losartan study. We first discuss the priors used and general aspects of the modeling exercise. Next, we explain how bioequivalence and particularly individual bioequivalence is assessed. Then we show how censoring was allowed in the model. We conclude this section by detailing the diagnostic checks that we performed.

Priors and estimation

A Bayesian approach was used to fit the models and to assess the bioequivalence of the curves. The currently available Bayesian computational tools enabled us to implement models with random effects without much difficulties. For the parametric nonlinear mixed-effects models we have taken independent N(0,1000) priors for the regression coefficients lKa_k, lKe_k and a gamma distribution for the precision of random errors i.e., $\sigma_k^{-2} \sim Gamma(0.001, 0.001)$. For the variance of the random effects we used a scaled inverse Wishart [*Gelman et al.*, 2012]. This means that we use $\gamma_{ipk} = b_{ipk}\xi_{ik}$ where $\boldsymbol{b}_i = (b_{i11}, b_{i21}, \ldots, b_{i3K})^\intercal$ has an inverse Wishart distribution with six degrees of freedom and prior variance Ω and ξ_{ik} has a uniform distribution on [0, 100]. This scaled inverse Wishart leads to an improved convergence relative to the more standard inverse-Wishart.

In the MSITAR model similar independent vague normal priors were taken. For example, for the fixed-effects parameters β_{kl}, $(k = 1, 2; l = 1, \ldots, L)$, we used independent $N(0, 1000)$ priors.

For the MSITAR model we calculated the B-spline basis on a fine grid of values prior to the actual sampling process. In the sampling process linear interpolation was used to evaluate the basis for time points in-between the pre-chosen grid values. This is faster than calculating the basis at each iteration and the error is negligible.

Markov chain Monte Carlo (MCMC) sampling was used to estimate the parameters. The models were fitted in JAGS version 4.2 which was called from R [*Plummer*, 2003b]. For both models we ran approximately 2.4 million iterations of the sampling algorithm, a quarter of which were discarded as burn-in. The data were thinned with a factor 300 to reduce the resulting objects to a manageable size. Appendix 2 shows the JAGS code that was used and in Appendix 3 we provide further details about the sampling procedures in JAGS.

Bioequivalence

To test population bioequivalence of the new formula of Losartan and the reference drug, one could set the random effects to zero and calculate C_{max} and AUC for this 'representative' person. Alternatively one can compare the marginal mean profiles using a double MCMC to integrate out the random effects. That is, within each MCMC iteration we use MCMC integration to average out the random effects. Note

that the marginal mean profile is not equal to the profile calculated at the mean of the random effects (nor is the marginal median equal to the profile calculated at the median of the random effects). In addition to comparing summary measures, we also looked at the entire marginal mean profiles. This enabled us to locate the difference in the curves if any.

In addition to population level bioequivalence, it is also of interest to test individual bioequivalence, see e.g. [*Schall and Luus*, 1993]. Individual bioequivalence tells us whether the behaviour of the test drug is similar to the behavior of the reference drug within patients. This is relevant when for a particular patient who switches from the reference drug to the new formula. In that case we wish to know the particular value of the ratio of the bioequivalence measures between the drug formulations that we will observe. Hence, we propose to use the following measure:

$$\int \frac{O_T(\boldsymbol{y}_{PRED})}{O_R(\boldsymbol{y}_{PRED})} p(\boldsymbol{y}_{PRED}|\boldsymbol{\gamma}_{PRED}, \boldsymbol{\theta}) p(\boldsymbol{\gamma}_{PRED}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{4.7}$$

where O is either the AUC or C_{max} and the indexes T and R specify whether we consider the test or the reference drug. $\boldsymbol{\gamma}_{PRED}$ and $\boldsymbol{y}_{PRED}$ are the random effects and observations for new individuals respectively. So we are sampling from the posterior predictive distribution of our bioequivalence measures. If narrow, this procedure indicates individual bioequivalence. However, we could argue that there is no individual bioequivalence if the 90% credible interval based on this distribution contains either very low or very high values. Unlike for population bioequivalence, there are no generally accepted criteria for individual bioequivalence.

The proposed procedure by the FDA decomposes the within-subject variance into a between-formulation and a within-formulation part [*U. S. Food and Drug Administration*, 2001]. This relates the changes in the bioavailability profile that we observe when an individual switches to a new drug to the changes that we would expect even when the same drug formulation is taken on another occasion. As we do not have any replicates in this study, this method cannot be used.

Detection limit

An additional feature of the motivating data set is that the technique used to measure the drug concentration has a detection limit of 2 ng/ml; in other words,

the drug concentrations are left censored at 2 ng/ml. This left censoring is the mirror image of the right censoring that is common in survival analysis. About 15% of the observations of both treatments is below the detection limit.

It has been shown that dealing with these censored observations explicitly is better than simply substituting zero, the detection limit, or some fraction thereof [*Helsel*, 2006; *LaFleur et al.*, 2011] studied this extra complication in detail. In a MCMC approach the incorporation of censored observations is easily done via data augmentation. We assume that the censored values come from the same distribution as the other observations but every time the values are below the detection limit a zero is observed in the response variable. In the sampling procedure we sample the unobserved values of the response from a truncated distribution. For the stochastic parents of the response sampling can now take place as if it were fully observed.

One should note that censored observations are effectively omitted from the deviance that is calculated by JAGS. Therefore, we had to calculate this ourselves using a likelihood were the censored values appear as terms with $\Phi(2|\mu(.), \tau)$ where Φ is the cumulative density function of a normal tau is the precision and μ is the mean given by the respective models given by equation 4.4 or 4.6.

Diagnostic checks and model comparison

The different models were compared using the deviance information criterion (DIC) [*Spiegelhalter et al.*, 2002c] and the pseudo Bayes factor [*Geisser and Eddy*, 1979]. The pseudo Bayes factor is defined as:

$$PSBF = \frac{\prod_{i,j,k} \text{CPO}_{ijk,\text{MSITAR}}}{\prod_{i,j,k} \text{CPO}_{ijk,\text{COMP}}}$$

Ignoring the subscripts of the model to facilitate the notation, we defined the CPO of observation ijk as:

$$\text{CPO}_{ijk,M} = p(y_{ijk}|\boldsymbol{y}_{(ijk)}) = \int p(y_{ijk}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}_{(ijk)}) d\boldsymbol{\theta}.$$

Here M is the model (MSITAR or COMP) and $\boldsymbol{y}_{(ijk)}$ denotes the data set without observation y_{ijk}. So CPO_{ijk} expresses how surprising y_{ijk} is in the light of the other observations. Large values of $1/\text{CPO}_{ijk}$ indicate the presence of an outlying observation. In, e.g., *Lesaffre and Lawson* [2012, chap. 10], it is shown how the

MCMC output can be used to estimate CPO. From the definition above it is clear that the CPOs can also be used to assess the models for for discrepant observations [see *Geisser*, 1980].

We also performed posterior predictive checks (PPCs). In a PPC the predictive distribution of a discrepancy measure is compared with the observed discrepancy measure. In general, this proceeds as follows. First generate S replicate data sets $\boldsymbol{y}_s^{rep}$ $(s = 1, \ldots, S)$. For discrepancy measure or test statistic $T(\boldsymbol{y}|\boldsymbol{\theta})$ we compare $T(\boldsymbol{y}^{obs}|\boldsymbol{\theta})$ with $T(\boldsymbol{y}^{rep}|\boldsymbol{\theta})$ where $\boldsymbol{y}^{obs}$ represents the observed data set and $\boldsymbol{y}^{rep}$ is a replicated data set. The posterior predictive p-value (PPP-value) is defined as:

$$Pr[T(\boldsymbol{y}_s^{rep}|, \boldsymbol{\theta}) \geqslant T(\boldsymbol{y}^{obs}|\boldsymbol{\theta})|\boldsymbol{y}] \doteq \frac{\#[T(\boldsymbol{y}_s^{rep}|\boldsymbol{\theta}_s) \geqslant T(\boldsymbol{y}^{obs}|\boldsymbol{\theta}_s)]}{S}$$

A PPC below 0.05 or above 0.95 is usually seen as an indication of a problem of the model fit.

Here we chose the omnibus measure suggested in [*Gelman et al.*, 2004] given by:

$$\mathbf{X}^2_{GM} = \sum_{i,j,k} \frac{[y_{ijk} - E(y_{ijk}|\boldsymbol{\theta})]^2}{var(y_{ijk}|\boldsymbol{\theta})}, \tag{4.8}$$

where the expectation and variance are calculated over the samples from the posterior predictive distribution.

The normality of the random effects was assessed visually, by inspecting normal probability plots of the respective posterior means. In addition the Shapiro-Wilk test was performed on the posterior mean of the random effects. In both models the null hypothesis of normality was rejected for one of the random effects.

4.5 Application

Main results and bioequivalence

Population and individual bioequivalence of the two Losartan formulations were analyzed with the parametric compartment and with the adapted MSITAR model. In a global sense the MSITAR model may give a better fit when compared to the compartment model. This is illustrated in Figure 4.5 where the profiles fitted for individual 15 are shown. Marginal mean curves with credible intervals of both models presented in Figures 4.6 and 4.7 enable the comparison of both models

on average. Apparently, the adapted MSITAR model provides more flexible curves than the ones obtained by the compartment model, which permits, for instance, a better fit to the concentration peak and decay. In Table 4.2 population bioequivalence for the Losartan formulations is computed. It is of interest to see that on the whole the CIs of the compartment model are wider than those of the SITAR model (with the exception of AUC for population bioequivalence). For C_{max} the upper bound of the compartment model is above the critical value of 1.25 in contrast to the MSITAR model where we can claim population bioequivalence.

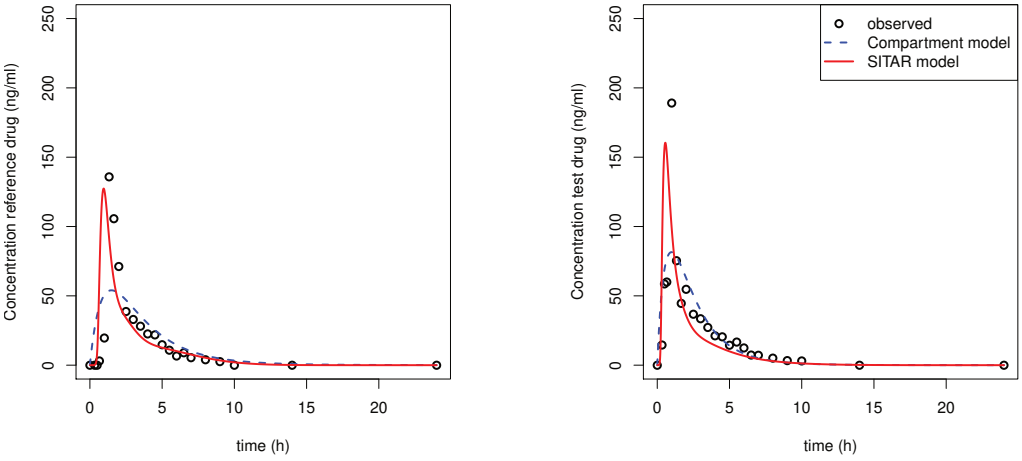


Figure 4.5: MSITAR (solid line) and parametric model (dashed line) fitted to data belonging to individual 15.

Table 4.2: Bioequivalence results for the Losartan trial

Model	Criterion	Mean	SD	5%	50%	95%
parametric	C_{max} ratio	1.08	0.201	0.79	1.07	1.44
MSITAR	C_{max} ratio	0.96	0.096	0.81	0.95	1.12

Model	Criterion	Mean	SD	5%	50%	95%
parametric	AUC ratio	1.00	0.086	0.86	1.00	1.15
MSITAR	AUC ratio	1.01	0.103	0.85	1.00	1.18

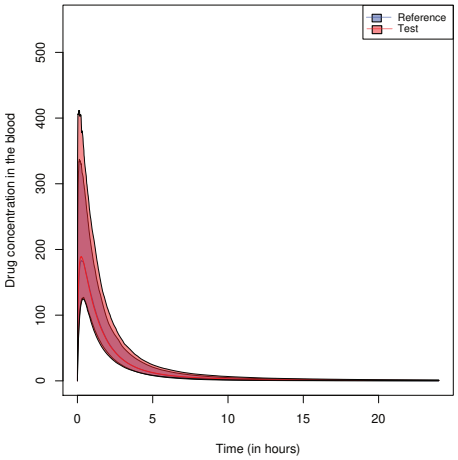


Figure 4.6: Marginal mean profile and 95% CI for the compartment model.

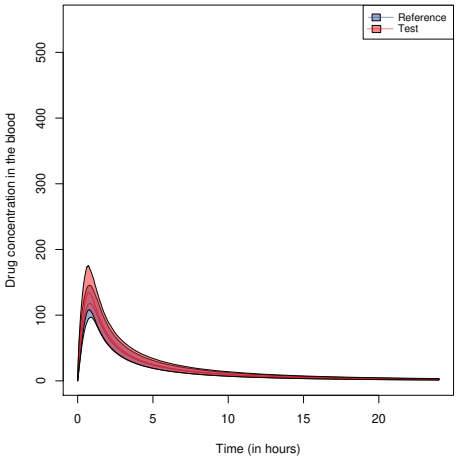


Figure 4.7: Marginal mean profile and 95% CI for the MSITAR model

Table 4.3: Individual bioequivalence results for the Losartan trial

Model	Criterion	Mean	SD	5%	50%	95%
parametric	C_{max} ratio	1.61	7.91	0.25	1.02	4.44
MSITAR	C_{max} ratio	1.16	0.97	0.32	0.95	2.85

Model	Criterion	Mean	SD	5%	50%	95%
parametric	AUC ratio	1.06	0.820	0.58	1.00	1.70
MSITAR	AUC ratio	1.13	0.671	0.41	0.96	2.32

In Table 4.3 we show the results for the individual bioequivalence. The 90% credible intervals are much wider for the individual bioequivalence than those for population bioequivalence. This shows that even when the drugs are population bioequivalent, on an individual basis one cannot be sure that switching will never lead to problems.

The DIC of the compartment model is 9 452 and the DIC of our MSITAR model is 8 359, demonstrating the much better fit of the MSITAR model. The log of the

pseudo Bayes factor was 518 supporting this conclusion.

Diagnostic analysis

To track observations that are difficult to fit, we made use of the CPO. Results for the compartment and MSITAR models are presented in Figures 4.8 and 4.9 respectively. Both models have some problems at time points 2 to 6, i.e. near the peak and generally do well afterwards. They both do not fit well for individual 4 who reaches the maximum concentration for the test drug much later than the other individuals (this is also visible from Figure 4.2). All together, there are more deviating observations in the compartment model than the MSITAR model.

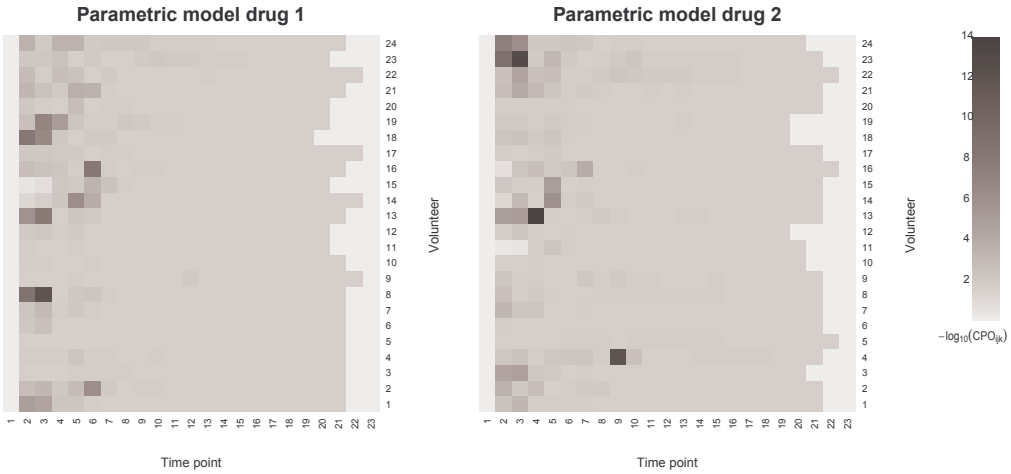


Figure 4.8: Heatmap of the CPO for the compartment model: The CPOs are organized in a matrix with time points on the horizontal axis and the number of the individual on the vertical axis. Darker cells indicate high values of $-log_{10}(CPO_{ijk})$ hence a data point with poor fit.

The value of the omnibus PPP-value was 0.51 for the compartment model and 0.29 for the MSITAR model. So based on this test there seems to be no problem in either model. We could not find any evidence for non-normality of the random effects. However with the small number of volunteers that participated in the trial, normality tests have little power.

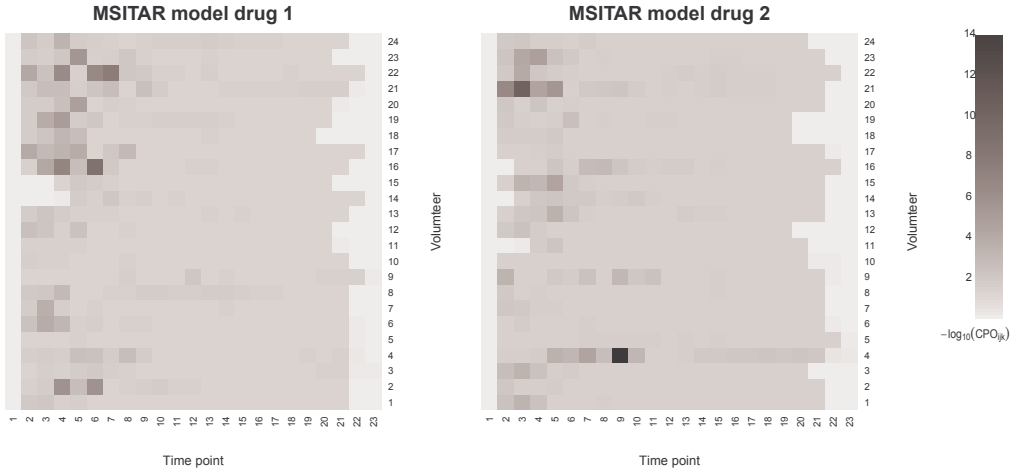


Figure 4.9: Heatmap of the CPO for the MSITAR: The CPOs are organized in a matrix with time points on the horizontal axis and the number of the individual on the vertical axis. Darker cells indicate high values of $-log_{10}(CPO_{ijk})$ hence a data point with poor fit

4.6 Discussion

A bioequivalence analysis is often quite challenging because it involves fitting nonlinear mixed models for correlated data. In this paper, some aspects for population and individual bioequivalence were discussed and two modeling approaches were contrasted: a parametric compartment model which has a biological foundation and a variant of the MSITAR model. This MSITAR model which was originally developed to model human growth had to be adapted substantially to be applicable to model pharmacokinetic data. In this paper we only looked at the one compartment model to keep the comparison simple. Of course we could have attempted to additionally fit more complicated pharmacokinetic models but we believe we would have quickly ran into computationally issues that way.

Using the data of the Losartan bioequivalence trial, it was shown that the adapted MSITAR model can fit the data better than the compartment model. The explanation lies in the flexibility of the MSITAR approach which fits both the absorption peak and the substance elimination, whereas the restricted compartment model seems to prevent the absorption-elimination curve to reach higher concentrations when the decay is relatively fast. However, despite being more flexible and providing a better quality fit, it is important to point that the MSITAR

approach does not provide a biological interpretation of parameters as the parametric compartment model does. It has to be noted though that the compartment model is based on an abstraction of the real world that only holds approximately.

The SITAR model is a kind of semi-parametric model in the sense that the template is modelled by a spline. One could also imagine a fully non-parametric model where each individual curve is described by a separate spline. Such a model would be very flexible but it would also be hard to interpret in which way the individual profiles differ from the 'average' and each other. As will often be the case in practice, we do not have enough data to estimate this fully non-parametric model. In this sense the MSITAR model forms a kind of compromise between the flexibility of non-parametric models on the one hand and interpretability and estimability on the other.

The Bayesian approach permits us to handle special features of the data, such as correlated data and censoring easily. When analysing the population bioequivalence, we concluded that the test formula of Losartan was bioequivalent to the reference formula. Additionally a new way to assess individual bioequivalence was developed based on the posterior predictive distribution of the ratio between the bioavailability measures for new individuals. Here we saw that the credible intervals are a lot wider than for population bioequivalence. A proper investigation of individual bioequivalence requires replicate data which was not available here.

In this paper we concentrate on assessing bioequivalence using the summary statistics AUC and C_{max}. However because we estimate the whole curves for all individuals, an alternative would be the following approach: First, we estimate a model which uses the same spline for both models. We can compare this model (using formal tests or Bayes factors) with a model using two separate splines. If the model with a single spline is just as good as the model with separate splines we can conclude that we have average bioequivalence. When this is the case we can proceed to look at the individuals. When the subject-specific effects do not differ between the models, we also have proven individual bioequivalence.

We noticed heteroskedasticity in the residuals by plotting the residuals of our model (Included as Appendix 4) against the fitted values and against time. This was confirmed by a dedicated PPC yielding PPP-values above 0.99 for both models. However, up to now models that express the variance as a function of time or blood concentration failed to converge. Extensions of above models allowing for

heteroskedasticity will be explored in a subsequent paper.

Acknowledgments

The second author wishes to acknowledge Fundação de Amparo a Pesquisa do Estado de São Paulo, FAPESP, Brazil, for supporting this research.

Appendices

4.A Directed Acyclic graphs

Below we show the directed acyclic graphs (DAGs) of the two models specified above. To prevent the graphs from becoming too cluttered we have omitted a few of the intermediate deterministic nodes. In the graphs i is the (loop) index of the individuals, j of the time points and k of the drugs.

Compartment model

The compartment model for the jth response of the kth drug formulation of the ith individual is given by

$$y_{ijk} = \exp(lKa_{ik} + lKe_{ik} - lCl_{ik}) \times \frac{[\exp(-e^{lKe_{ik}} t_{ijk}) - \exp(-e^{lKa_{ik}} t_{ijk})]}{e^{lKa_{ik}} - e^{lKe_{ik}}} + \epsilon_{ijk}, \tag{4.1}$$

where y_{ijk} be the kth response at the time t_{ijk} of individual i (for each $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ and $k = 1, \ldots, K$); $lKa_{ik} = lKa_k + \gamma_{i1k}$, $lKe_{ik} = lKe_k + \gamma_{i2k}$ and $lCl_{ik} = lCl_k + \gamma_{i3k}$; lKa_k, lKe_k and lCl_k are the fixed-effects related to the k-th response; γ_{i1k}, γ_{i2k} and γ_{i3k} are the respective random effects. Here, $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}^\top, \ldots, \boldsymbol{\gamma}_{iK}^\top)^\top$ and $\boldsymbol{\gamma}_{ik} = (\gamma_{i1k}, \gamma_{i2k}, \gamma_{i3k})^\top$ is the vector of random effects related to the k-th response of the i-th individual for each $k = 1, \ldots, K$ and $i = 1, \ldots, n$. We model $\boldsymbol{\gamma}_i = b_i^\top \xi$ where $b_i \sim N(0, \Sigma_b)$ and the elements of xi are i.i.d. and uniformly distributed on [0, 100]. The residual random errors $\epsilon_{ijk} \sim N(0, \sigma_k^2)$ are assumed to be independent of each other and of the random effects.

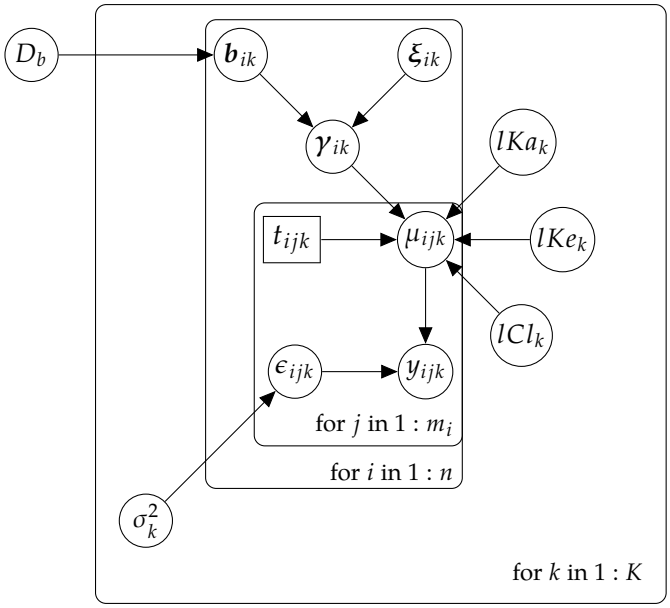


Figure 4.1: Directed acyclic graph for the compartment model. Hyperparameters were omitted.

MSITAR model

The multivariate MSITAR model is given by:

$$\begin{aligned} y_{ijk}|\boldsymbol{\gamma}_{Mi}, \boldsymbol{\beta}_k, \sigma_k^2 &\sim N(\exp(\gamma_{i2k})[T_{ijk}^{\top}\boldsymbol{\beta}_k], \sigma_k^2), \\ T_{ijk} &= B(\exp(\gamma_{i3k})[t_{ijk} + min(\gamma_{i1k}, 0)]), \end{aligned} \tag{4.2}$$

where y_{ijk} be the kth response at the time t_{ijk} of individual i. T_{ijk} is a matrix with bases of cubic splines for the kth response of the individual i at time j, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kL})^{\top}$ is the vector of spline coefficients related to the k-th response (here L denotes the degrees of freedom of the spline), $\boldsymbol{\gamma}_{ik} = (\gamma_{1ik}, \gamma_{2ik}, \gamma_{i3k})^{\top}$ is the vector of random-effects for $i = 1, \ldots, n;\ j = 1, \ldots, m$. We model $\boldsymbol{\gamma} = b_i\xi$ with $b_i \sim N(0, \Sigma_b)$ and the elements of ξ are uniformly distributed between 0 and 100.

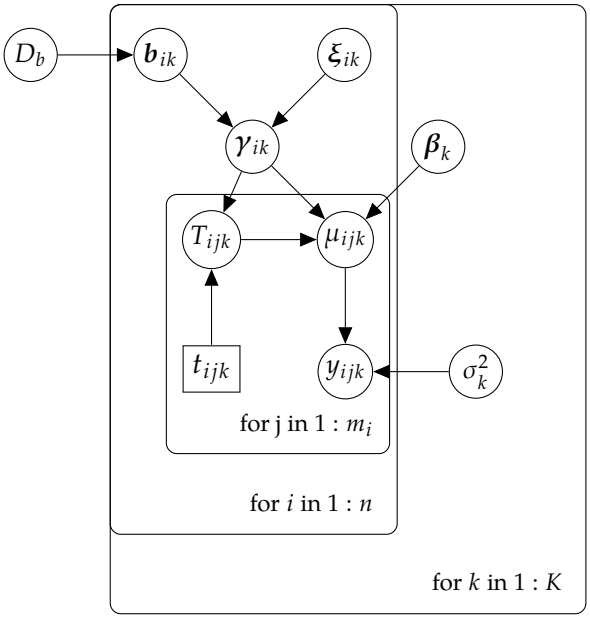


Figure 4.2: Directed acyclic graph for the MSITAR model. Hyperparameters were omitted.

4.B JAGS code

In this appendix we give the JAGS code of the two models that we compared.

Compartment model

The listing below details the compartment model in JAGS. We called JAGS from R using the jagsUI library (which allows us to run several chains in parallel). As explained in section 3.1 of the main text the gammas and phis are not identified. We therefore define k_e as the lowest of the two random effects and k_a as the highest. Note that a few of the deterministic nodes we use in the program to not occur in the main text. They mainly function to make the program clearer and to prevent some lines from becoming too long. Like in the main article we use $i \in 1 \dots n$ as an index of the individuals, $j \in 1 \dots m$ as an index for the measurement time and $k \in 1 \dots K$ as the index of the drug. $lKa_1, lKe_1, \dots, lCl_2$ were renamed to `beta1[1]`, `beta1[2]`, ..., `beta2[3]` because JAGS does not allow the distinction between different symbols by using different numbers of

indices. In this JAGS code m, does not depend on the individual as the data is balanced. Note that the number of drugs K is hard-coded as two. Also the limit of detection is hard-coded.

Listing 4.1: Compartment model

```
model{
  for(i in 1:n)     {
    b[i,1:6] ~ dmnorm(mu.b[1:6], tau.b[1:6, 1:6])
    for (r in 1:6){
      gammas[i, r] <- b[i, r] * xi[r]
    }
    for(k in 1:2){
      lKe[i, k] <- min(phi1[i, k], phi2[i, k])
      lKa[i, k] <- max(phi1[i, k], phi2[i, k])
      ke[i, k] <- exp(lKe[i, k])
      ka[i, k] <- exp(lKa[i, k])
    }
    phi1[i, 1] <- beta1[1] + gammas[i, 1]
    phi2[i, 1] <- beta1[2] + gammas[i, 2]
    phi3[i, 1] <- beta1[3] + gammas[i, 3]
    phi1[i, 2] <- beta2[1] + gammas[i, 4]
    phi2[i, 2] <- beta2[2] + gammas[i, 5]
    phi3[i, 2] <- beta2[3] + gammas[i, 6]
  }  # for(i in 1:n)
  tau.b[1:6, 1:6] ~ dwish(6 * Omega[ , ], 6)
  for(j in 1:m){
    for(i in 1:n){
      for(k in 1:2){
        observed[i, j, k] ~ dinterval(y[i, j, k], 2)
        y[i, j, k] ~ dnorm(mu[i, j, k], tau[k]) #data model: the likelihood
        mu[i, j, k] <- exp(lKe[i, k]+lKa[i, k]-phi3[i, k] ) *
        (exp(-ke[i, k]*time[j])-exp(-ka[i, k]*time[j])) /
        (ka[i, k]-ke[i, k])
        res[i, j, k] <- y[i, j, k] - mu[i, j, k]
      } #  for(k in 1:2)
    } #  for(i in 1:n)
  }  # for(j in 1:m){
  for(k in 1:2){
```

```
    tau[k] ~ dgamma(0.001, 0.001)
  }
  for(u in 1:3){
    beta1[u] ~ dnorm(0, 0.0001)
    beta2[u] ~ dnorm(0, 0.0001)
  }
  for (r in 1:6){
    xi[r] ~ dunif(0, 100)
  }
}
```

MSITAR model model

Now we present the SITAR model. In this model `Bas` represents the restricted cubic spline basis that is evaluated on a fine grid of values from R. This is done by a modified version of the `ns` function from the R splines package. From which the lines that apply the restrictions on the spline basis are replaced by the restrictions that are detailed in the paper. This modified `ns` function is available from the GitHub page of the first author (`https://github.com/stenw/FlexibleBioequivalence`). The `interp.lin` function now interpolates between those values.

Listing 4.2: SITAR model

```
model{
  #subject specific effects
  tau.b[1:6, 1:6] ~ dwish(6 * Omega[,], 6)
  for (r in 1:6){
    xi[r] ~ dunif(0, 100)
  }
  for(i in 1:n) {
    b[i, 1:6] ~ dmnorm(mu.b[1:6], tau.b[1:6, 1:6])
    for (r in 1:6){
      gammas[i, r] <- b[i, r] * xi[r]
    }
    shift1[i] <- min(gammas[i, 5], 0)
    shift2[i] <- min(gammas[i, 6], 0)
```

```
    for(j in 1:m){
      x[i, j, 1] ← exp(gammas[i, 1]) * (time[j] + shift1[i])
      x[i, j, 2] ← exp(gammas[i, 2]) * (time[j] + shift2[i])
      mu[i, j, 1] ← exp(gammas[i, 3] ) *
      inprod(beta1[1:L], B[i,j, 1:L,1])
      mu[i, j, 2] ← exp(gammas[i, 4] ) *
      inprod(beta2[1:L], B[i,j, 1:L,2])
      for(k in 1:2){
        y[i, j, k] ~ dnorm(mu[i, j, k], tau[k] )
        res[ i , j, k] ← y[i, j, k] - mu[i, j, k]
        observed[i, j, k] ~ dinterval(y[i, j, k], 2)
        for(l in 1:L){
          B[i, j, l, k] ← interp.lin(x[i, j, k], grid, Bas[ , l])
        } # for l
      } # for k
    } # for j
  } # for i
  for(k in 1:2){
    tau[k] ~ dgamma(0.001, 0.001)
  }
  for(l in 1:L){
    beta1[l] ~ dnorm(0, 0.0001)
    beta2[l] ~ dnorm(0, 0.0001)
  }
}
```

4.C Sampling procedures in JAGS

To obtain samples from the posterior of our models we use the program 'Just another Gibbs sampler' (JAGS) by Martyn Plummer [*Plummer*, 2003b]. This program can been seen as a reimplementation of WinBUGS in C++, although there are some differences. A big advantage of BUGS-like software, is that it makes sampling from any model relatively easy as one only needs to specify the model and the likelihood, the rest is done by the program. JAGS will then determine the directed acyclic graph, derive the full conditionals and find suitable samplers. The only disadvantage of JAGS or similar software is its black-box nature, i.e. the use is not fully aware how the sampling is done. Below we give details what samplers JAGS has used for the different full conditionals of our models. JAGS converts the program text to a graph of nodes storing information about the distribution of each variable and the resations that exist between them. This graph of nodes is traversed and for each node a suitable sampler is chosen.

The samplers chosen for the MSITAR model are as follows: For the residual variances τ conjugate gamma samplers are used.

$$\tau_k \mid \text{rest} \overset{\text{iid}}{\sim} \text{Gamma}\left(\alpha_\tau + \tfrac{1}{2}\sum_i m_i,\ \beta_\tau + \tfrac{1}{2}\sum_{i,j}\{y_{ijk} - \mu_{ijk}\}^2\right)$$

where n is the number of individuals and m is the number of measurement times so nm is the total number of observations for either drug and

$$\mu_{ijk} = \exp(b_{i2k}\xi_2)T_{ijk}\boldsymbol{\beta}_k$$

is the expected value under the MSITAR model. Here order_i is the covariate that denotes the order in which the drugs are given and α_τ and β_τ are the prior shape and rate parameters for the gamma prior of the τs. For the fixed effects $\boldsymbol{\beta}$ conjugate

normal samplers are used:

$$\begin{aligned}
\boldsymbol{\beta} \mid \text{rest} &\sim \mathrm{N}(\overline{\mu}_{kp}, \overline{\tau}_{\beta_{kp}}^{-1}) \\
\text{where} \quad \overline{\mu}_{kp} &= \overline{\tau}_{\beta_{kp}}^{-1} \tau \sum_{ij} A_{ijkp}^2 B_{ijkp}, \\
\overline{\tau}_{\beta_{kp}} &= \underline{\tau}_{\beta_{kp}} + \sum_{ij} B_{ijkp}^2 \tau_k, \\
A_{ijkp} &= Y_{ijk} - \exp(b_{i2k}) \sum_{l \neq p} \boldsymbol{\beta}_l (T_{ijk})_l, \\
B_{ijkp} &= \exp(b_{i2k}) \boldsymbol{\beta}_l (T_{ijk})_p
\end{aligned}$$

with $(T_{ijk})_l$ the lth element of vector T_{ijk} and $\underline{\tau}_{\beta_{kp}}$ the prior precision of β_{kp}. A_{ijkp} can be thought of as the response when we have removed the effect of the other βs, while B_{ijkp} is the effect that a unit increase of the coefficient has on this outcome. For the precision of the random effects a conjugate Wishart sampler is used, i.e.

$$T_{\boldsymbol{b}} \mid \text{rest} \sim \mathrm{W}\left(\delta + n, \Omega + \sum_i \boldsymbol{b}_i \boldsymbol{b}_i^{\intercal}\right),$$

where δ is the prior number of degrees of freedom, I the number of individuals and Ω the prior variance matrix.

For the random effects $\boldsymbol{b}$, a Metropolis sampler is used. Given the other parameters $\boldsymbol{b}_i$ is sampled from a distribution proportional to:

$$\exp(-\frac{1}{2} \boldsymbol{b}_i' T_{\boldsymbol{b}}^{-1} \boldsymbol{b}_i) \prod_{jk} g(y_{ijk} | \boldsymbol{\beta}_k, \tau_k, \boldsymbol{b}_i, \xi)$$

Here $T_{\boldsymbol{b}}$ is the precision of the bs and we use $g(y_{ijk} | \boldsymbol{\beta}_k, \tau_k \boldsymbol{b}_i, \xi)$ for the density of the MSITAR model conditional on the random effects, that is:

$$\sqrt{\frac{\tau_k}{2\pi}} \exp[-\tfrac{1}{2}(y_{ijk} - \mu_{ijk})^2 \tau_k]$$

For the auxiliary parameters related to the random effect variances ξ and the censored blood concentrations y_{ijk}, slice samplers are used. Given all other prameters ξ is sampled from a distribution proportional to:

$$\prod_{ijk} g(y_{ijk} | \boldsymbol{\beta}_k, \tau_k, \boldsymbol{b}_i, \xi),$$

Finally given the other parameters we can sample the elements of y_{ijk} that are censored from:

$$\frac{\tau_k \phi(\tau_k(y_{ijk} - \mu_{ijk}))}{\Phi(\tau_k(2 - \mu_{ijk}))}$$

where $\phi(x)$ is the pdf and $\Phi(x)$ is the CDF of a standard normal.

In the parametric model most sampler types are identical to the corresponding ones chosen in the MSITAR model, only for the β's slice samplers are chosen instead of a conjugate normal sampler as the response is not linear in the fixed effects in this model.

Sampling now begins with an adaptive phase in which samplers that require it, can change their behavior. For the Metropolis sampler this means that JAGS keeps track of a running mean of the acceptance rate, giving a larger weight to more recent iterations, and the variances of the proposal distribution are adjusted to aim for a target acceptance rate of 0.234. For the slice samplers we similarly keep track of a weighed average of the jumps and the step size (the tuning parameter that controls the initial estimate of the width of the slice) is adjusted accordingly so fewer steps will have to be taken to find the correct slice of the slice. After the adaptive phase we continue the burn-in. When we are convinced we are sampling from the posterior we start storing the sampled values to be used in our further analyses.

4.D Residual plots

In this section we provide the (marginal) residuals plots of the two models. They are defined as:

$$y_{ijk} - \exp(lKa_{ik} + lKe_{ik} - lCl_{ik}) \frac{[\exp(-e^{lKe_{ik}} t_{ijk}) - \exp(-e^{lKa_{ik}} t_{ijk})]}{e^{lKa_{ik}} - e^{lKe_{ik}}},$$

for the compartment model and as

$$y_{ijk} - (\exp(\gamma_{i2k})[T_{ijk}^{\top} \boldsymbol{\beta}_k]$$

for the MSITAR model. The heteroskedasticity is clearly visible.

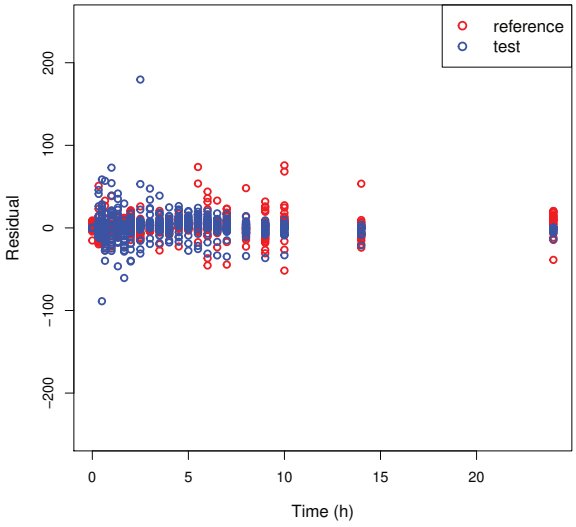


Figure 4.3: Residual plot for the MSITAR model

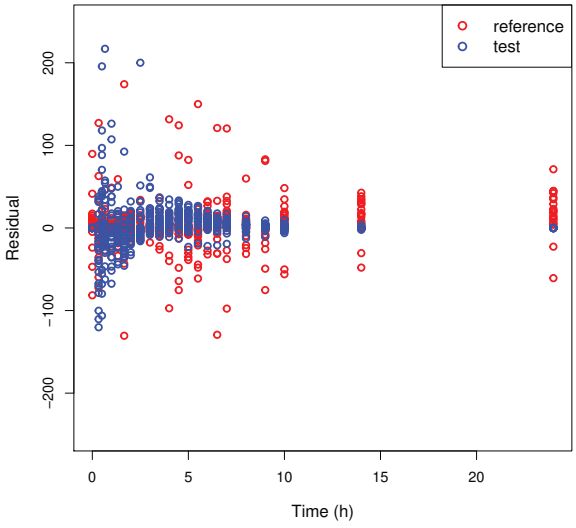
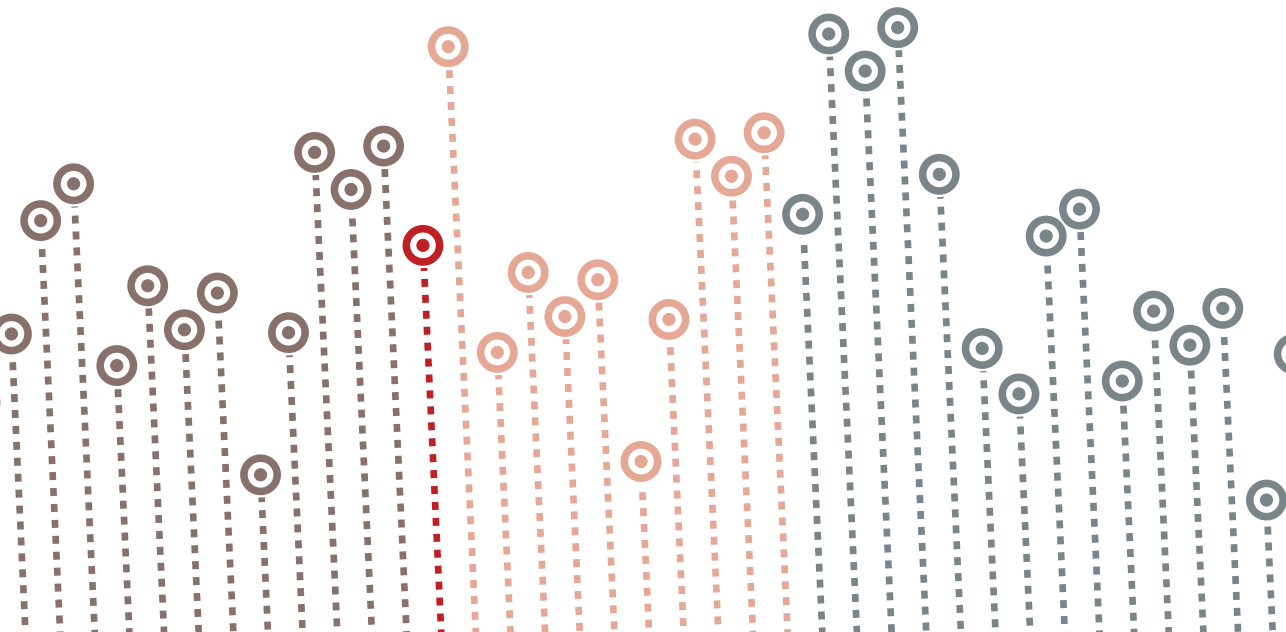


Figure 4.4: Residual plot for the compartment model

5 Two-dimensional longitudinal expectiles



Abstract:

The tails of a distribution (theoretical or observed) can be characterized by extreme quantiles. In growth studies quantile curves, conditional on age, are familiar. Expectile curves are a less well known, but powerful alternative. Expectiles generalize the mean in the same way that quantiles are a generalization of the median. We generalize the concept of expectiles into two dimensions by first projecting the data on a line in certain direction and than calculating the expectiles in the usual way. If we connect the expectiles in several directions a circular expectile contour arises. We can combine these circular contours with the concept of conditional expectiles to create what we call expectile tubes. They are a useful way to visualize how extreme a multivariate observation is, in relation to a distribution.

Authors: S.P. Willemsen & P.H.C. Eilers

5.1 Introduction

In the medical field growth charts are used to describe the distribution of measurements (like height) of a population. A chart of a healthy population without (known) diseases acts as a standard to which individuals are compared. It is known that nutritional problems and diseases are often expressed through abnormal growth patterns. Therefore growth charts have become a standard part of a pediatricians toolkit[*Turck et al.*, 2013]. Usually the references are conditional on a covariate: leading for example trough height-for-age or weight-for-height charts. This is done by making the reference functions of a covariate (like age). Because usually there is no a-priory functional form known it makes sense to use B-spline function. This smoothing can be improved upop when we use a relatively large number of B-splines and put a penalty on the amount that neighboring spline coefficients can differ from each other *Eilers and Marx* [1996].

Different types of measurements are often examined separately (as current reference charts do not permit anything else). However it is much more informative to look at them in relation to each other. A certain weight might be seen as to high when looked at by itself, but it might be well within the norm when seen in relation to the hight of an individual. Univariate outliers do not have to be multivariate outliers and vice versa. In order to detect multivariate outliers we create multivariate reference regions. Here we will concentrate on pairs of variables. In principle we can also study the combination of more variables but visualization becomes problematic.

Growth charts are currently based on quantiles. As we will show, quantiles can be computed as the solution to an optimization problem minimizing an asymmetrically weighed sum of absolute differences [*Koenker*, 2005]. If we the replace the absolute values in the optimization problem by squares we obtain expectiles, an alternative proposed by Newey *Newey and Powell* [1987]. Reference charts based on expectiles can be used in much the same way as charts based on quantiles and have the added advantage of being much easier to compute. Furthermore expectile curves often appear much smoother and lead to less problems like crossing curves, especially in small data sets. [*Schnabel and Eilers*, 2009].

The structure of this paper is as follows: We begin by formally introducing expectiles in Section 5.2. In Section 5.4 we will use the method of *Kong and Mizera* [2012] and adapt it to expectiles We then explain how we can condition

on covariates in Section 5.3. In Section 5.5 we combine circular expectile contours with expectile regression creation what we call 'tubular expectile contours' or 'expectile tubes'. Section 5.6 applies our method on a data set of measurements of the height and weight of Dutch boys. In Section 5.7 we conclude the manuscript with a discussion.

5.2 Expectiles in a nutshell

The τth quantile $q(\tau)$ of a distribution is the smallest value q such that the probability of an observation falling below it is at least τ. Given a sample y of size n, they can be estimated by sorting the observations and then taking the observation in position τn. They can also be obtained as the solution to the minimization problem:

$$\underset{q}{\operatorname{argmin}}\{(1-\tau)\sum_{y_i \le q}(y_i - q) + \tau \sum_{y_i > q}(y_i - q)\}, \tag{5.1}$$

which can be solved by linear programming techniques [*Koenker and D'Orey*, 1987; *Portnoy and Koenker*, 1997]. An advantage of this formulation is that q can be made dependent on explanatory variables as we will show below. The population equivalent expression for (5.1) is:

$$\underset{\mu}{\operatorname{argmin}}\{(1-\tau)\int_{-\infty}^{q}(q-y)dF(y) + \tau\int_{q}^{\infty}(y-q)dF(y)\}$$

the solution is the value q for which

$$(1-\tau)\int_{-\infty}^{q}(y-q)dF(y) = \tau\int_{q}^{\infty}(y-q)dF(y) \tag{5.2}$$

The τth expectile $\mu(\tau)$ of a sample can be defined as the value that minimizes the expression:

$$\underset{\mu}{\operatorname{argmin}}\{(1-\mu)\sum_{y_i \le \mu}(y_i - u))^2 + \tau \sum_{y_i > \mu}(y_i - \mu)^2\},$$

in which the absolute differences in (5.1) are replaced by squared differences. These sample expectiles are estimators of theoretical population expectiles given by:

$$\underset{\mu}{\operatorname{argmin}}\{(1-\tau)\int_{-\infty}^{\mu}(\mu-y)^2dF(y) + \tau\int_{\mu}^{\infty}(y-\mu)^2dF(y)\}$$

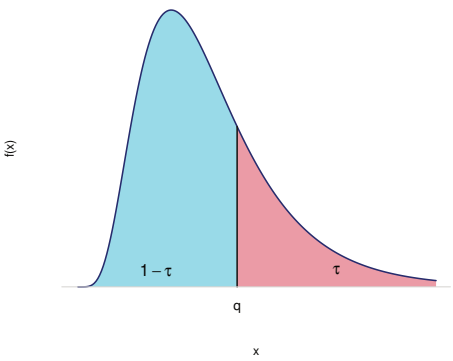


Figure 5.1: Illustration of an quantile. The curve shows the probability density function of an distribution. Hence the colored areas are the fraction of possible outcomes below and above q. For the τth quantile the ratio between them is equal to $\tau/(1-\tau)$

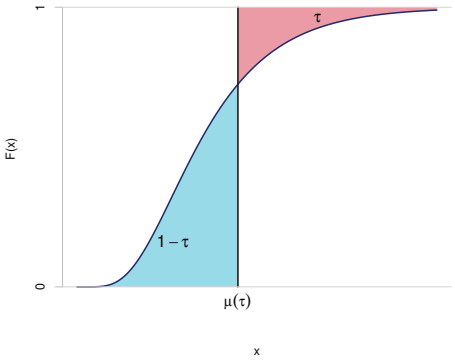


Figure 5.2: Illustration of an expectile. The curve shows the cumulative distribution function. So on horizontal axis we have the possible values the distribution can take while (loosely speaking) the vertical axis can been seen as he fraction of the population with this value of x. Hence the colored areas are the expected shortfall and the expected surplus. The ratio between these is specified by $\tau/(1-\tau)$

the solution is the value μ for which

$$(1-\tau)\int_{-\infty}^{\mu}(y-\mu)dF(y) = \tau\int_{\mu}^{\infty}(y-\mu)dF(y) \tag{5.3}$$

is met. Hence, it is clear that the ratio of τ to $1-\tau$ determines the ratio of the expected distance of a value to the expectile given that the observation falls below the expectile to the expected distance conditional on being above the expectile. An other way to think about expectiles is to see that it is a 'fair' bet to agree to pay an amount τ for each unit that a realization of the distribution is above $\mu(\tau)$ to receive an amount $1-\tau$ for each unit that the outcome is below this boundary.

In figure 5.1 we have visually illustrated an quantiles of the distribution. The curve is the probability distribution function f from which the data originate. The blue part corresponds to the probability mass below the quantile while the pink part lies above the quantile q. These areas correspond the the integrals in (5.2). When balance is obtained in (5.2) the ratio between these areas should be $\tau/(1-\tau)$.

In figure 5.2 we illustrate expectiles in an analogous manner. The curve here

is the cumulative distribution function F. The blue area corresponds to the values between F and the line that denotes the expectile (the think black line) for values of the distribution that fall below the expectile and the pink area are values between the expectile line and F for values that lie above of it. If a point lies further from the expectile, its impact is larger. It can be shown that the integrals in (5.3) corresponds to these two areas. τ controls the ratio between the blue and the pink areas and hence specifies the expectile. Clearly, $\mu(0.5)$ is just the mean. In the remainder of this article we are mainly working with a single expectile at a time and will omit the dependence of the expectiles μ on τ in the notation.

Except in the case that $\tau = 0.5$, there is no closed form solution, however we can use the least asymmetrically weighted squares (LAWS) algorithm. This algorithm was proposed by *Newey and Powell* [1987] who named it Asymmetrical Least Squares or ALS. To avoid the confusion with the desease that uses the same abbreviation it was renamed LAWS.

When we define a weight function:

$$w(\tau, y_i) = \begin{cases} \tau & \text{if } y_i > \mu(\tau) \\ 1 - \tau & \text{if } y_i \leq \mu(\tau) \end{cases} \tag{5.4}$$

(5.2) can now be written as:

$$\underset{\mu}{\operatorname{argmin}} \{ w(\tau, y_i) \sum (y_i - \mu) \}^2 \tag{5.5}$$

The LAWS algorithm finds the solution by iterating between finding the optimal value of $\mu(\tau)$

$$\mu^{(k)} = \frac{\sum_i y_i w_i^{(k-1)}}{\sum_i w_i^{(k-1)}} \tag{5.6}$$

and updating the weights w_i using (5.4) substituting $\mu^{(k)}$ for $\mu(\tau)$. To start the algorithm arbitrary starting weights $w^{(0)}$ are needed (say 0.5 for all observations). Because the objective function is convex it aways converges *Eilers* [1987]. Moreover, convergence is usually quick (Five to ten iterations.)

5.3 Penalized expectile regression

It is easy to see that expectiles can be made conditional on covariates. We simply replace the parameter μ in 5.2 with a linear predictor $x_i\beta$. Hence we arrive at:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i}^{n} w_i(\tau)(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta})^2$$

Where $\boldsymbol{x}_i$ is the vector of covariates belonging to observation i and $\boldsymbol{\beta}$ the vector of coefficients. Again it makes sense to apply smoothing. To keep things simple we describe a situation in which we have a single continuous covariate t (say time) and we conjecture that the quantiles of y smoothly change as t increases. We setup the following optimization problem:

$$\underset{\beta}{\operatorname{argmin}} \sum_{i}^{n} w(t)(y_i - B(t_i)\boldsymbol{\beta})^2 + \lambda_t \left\|D\boldsymbol{\beta}\right\|^2$$

where $B(t_i)$ is a function that performs a B-spline basis expansion for t_i and D is a difference matrix and λ is the penalty parameter that dictates the smoothness of the curves.[*De Boor*, 2001; *Eilers and Marx*, 1996]. Of course it is easy to include additional covariates that do or do not need to be smoothed [*Eilers and Marx*, 2002]. As an illustration of expectile regression we fitted smoothed expectile curves related to 3 values of the asymmetry parameter and 3 values of the penalty on the Dutch boys data. The blue, orange and red lines are respectively based on an asymmetry parameter of 0.929, 0.982 and 0.996 . These values are chosen because they correspond to the values of 1, 1.5 and 2 sd of a normal distribution.

The fully nonparametric method to create reference curves above can be contrasted to the GAMLSS method by *Rigby and Stasinopoulos* [2005] in which a specific parametric form is assumed the parameters of which are modeled by spline functions.

5.4 Directional expectiles and circular expectile contours

As motivated above, it is important to look at variables in relation to each other. To do this we use directional expectiles, adopting a method that was proposed

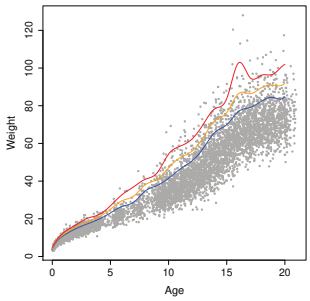


Figure 5.3: Expectile regression with asymmetry corresponding to 1, 1.5 and 2 sd and $\lambda_t = 0.001$

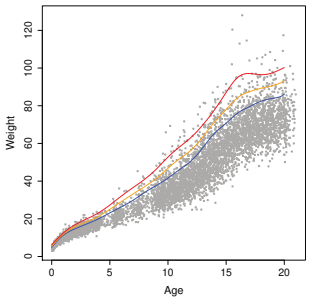


Figure 5.4: Expectile regression with asymmetry corresponding to 1, 1.5 and 2 sd and $\lambda_t = 1$

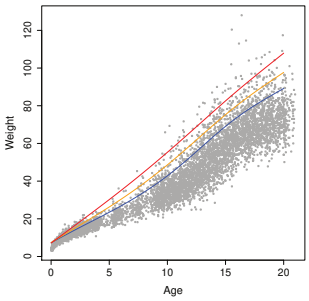


Figure 5.5: Expectile regression with asymmetry corresponding to 1, 1.5 and 2 sd and $\lambda_t = 1000$

for quantiles by *Kong and Mizera* [2012]. Directional quantiles are closely related to the idea of half width depth proposed by Tukey [1975]. For expectiles a similar measure of data depth can be defined [*Giorgi and McNeil*, 2014].

If $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$, $\boldsymbol{y}_i = (y_{i1}, y_{12})$ is a n by 2 dimensional matrix of observations and $\boldsymbol{\theta}$ is a unit vector of length 2 then the τth directional expectile in direction $\boldsymbol{\theta}$ of Y is defined as the the τth expectile of $Y\boldsymbol{\theta}$. In other words we project Y on the line in direction $\boldsymbol{\theta} = (\cos(h), \sin(h))^\top$ for some angle h, and then calculate the expectiles in the usual way. In this manner we reduce the multivariate problem to an univariate one. For the sake of simplicity we restrict ourselves here to a two dimensional response. This is also motivated by the fact that we primarily see these methods as visualization techniques and displaying data with more dimensions becomes difficult. Of course the two variables we will examine can have widely different scales. Here we standardize the variables, that is we subtract the mean and divide by the standard deviation. We now minimize:

$$\sum_i^n w_i(\boldsymbol{y}_i\boldsymbol{\theta} - \mu)^2$$

We can simultaneously look at a large number of directions. If we define a 2 by m matrix of directions $\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m)$ with $\boldsymbol{\theta}_j = (\cos(h_j), \sin(h_j))^T$ where h_j, $j \in 1, 2, \ldots m$ is a number of angles placed on a grid and m is the number of

directions. We now define the projected data as $Z = Y\Theta$. We then arrive at:

$$\sum_{j}^{m}\sum_{i}^{n} w_{ij}(z_{ij} - \mu_j)^2 \tag{5.7}$$

were z_{ij} is the element of Z corresponding to individual (row) i and direction (column) j. We propose that a reference region can be derived from these directional expectiles; To do this we need to calculate the points where the lines defining the expectiles cross.

If we calculate the expectiles in a large enough number of directions the resulting contours look quite good. However, they can be improved upon by applying smoothing. In (5.7) the different angles are essentially unconnected. However we can apply smoothing using splines and penalties.

We will now model the expectile in direction h_j as a $\boldsymbol{c\alpha} = \sum_k^K c_{jk}\alpha_k$, where c_{ik} is the kth component of the circular basis expansion in direction h_j and $\boldsymbol{\alpha}$ is a vector of coefficients . This already reduces the number of parameters. We can further increase the smoothness by putting a penalty on the amount that the coefficients corresponding to neighboring angles can be apart. We penalize by an amount $\lambda_h \|Q\boldsymbol{\alpha}\|^2$, where Q is a circular difference matrix. By circular splines and bases, we mean that they wrap around the origin. A cyclic spline C basis with degrees of freedom n on $[-\pi, \pi]$ consisting of columns c_i with $i \in 1 \ldots n$ can be defined as $c_i(x) = b_i(x) + b_i(x + 2\pi)$. Where b_i is the ith spline function (column) of a regular B spline basis with n degrees of freedom of degree d with $n + 1 + d$ knots equally spaced on $[-\pi, \pi + \frac{2d\pi}{n}]$. This is illustrated in Figure 5.9.

For the penalty matrix we work in a similar way we first set up a regular penalty matrix on a wider range and then ''wrap it around'.

By smoothing in this way, we avoid over-fitting and the resulting contours have fewer sharp corners which we think is more realistic (See Figure 5.11). This is especially the case when little data is available and the expectiles we look at are more extreme. The problem now becomes:

$$\sum_{j}^{m}\sum_{i}^{n} w_{ij}(z_{ij} - \boldsymbol{c}_j\boldsymbol{\alpha})^2 - \lambda_h \|Q\boldsymbol{\alpha}\|^2 \tag{5.8}$$

The constant λ_h specifies the amount of smoothing. When λ_h is large, coefficients that are 'close' to each other cannot be too far apart, so the spline will be very smooth. When it is small, the coefficients are almost independent, so the spline is very flexible.

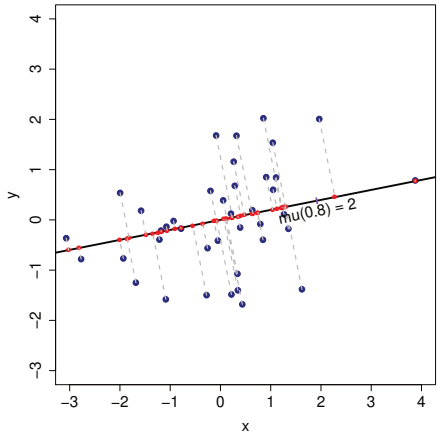


Figure 5.6: Directional expectiles are defined by projecting observations on a line with a specified direction and then calculating the expectiles

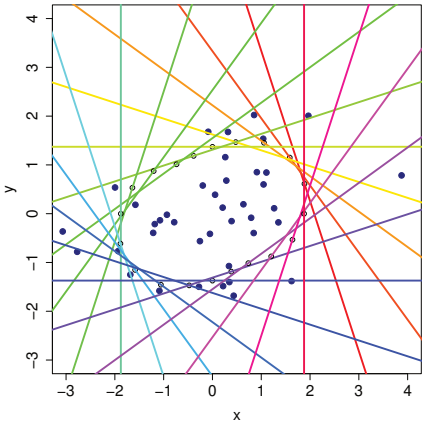


Figure 5.7: Directional expectiles are calculated for a large number of directions

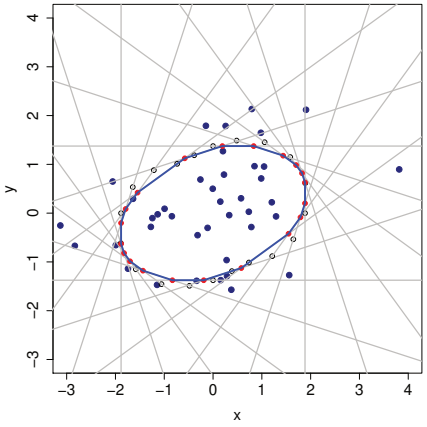


Figure 5.8: The directional expectiles define a convex contour

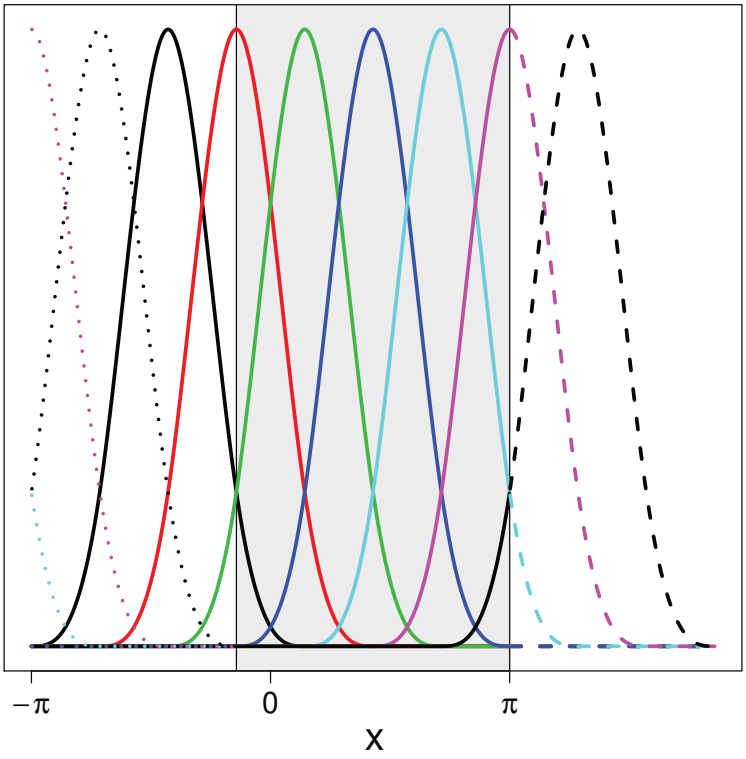


Figure 5.9: Illustration of cyclic splines: To create a cyclic spline of $[-\pi, \pi]$. we first define a regular B-spline. The splines only sum to one in the middle (grey area). If the spines at the right side of the grey area are shifted to the lift by 2π we have created a cyclic spline basis.

$$\left(\begin{array}{ccccccc|cc} 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 \\ \mathbf{1} & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ \mathbf{-2} & \mathbf{1} & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{array}\right)$$

Figure 5.10: Example of a cyclic penalty matrix. The entries in the last two columns wrap around and appear on the first two columns.

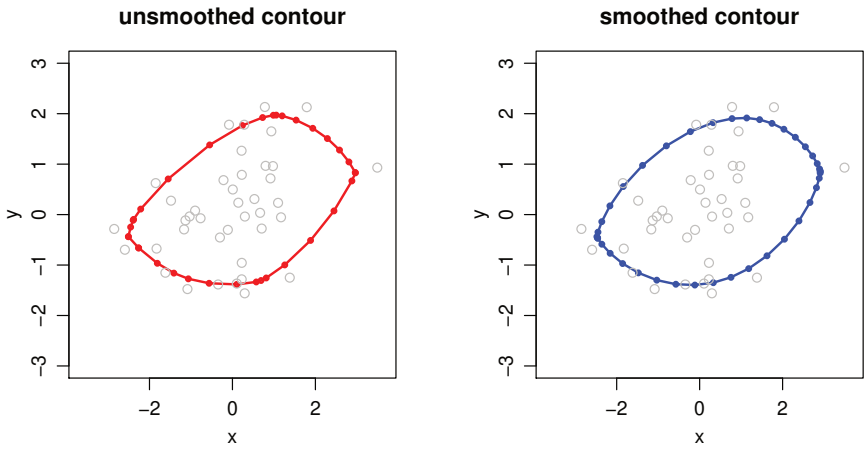


Figure 5.11: Effect of smoothing on expectile contours

In (5.8) we have replaced the single parameter μ with a linear predictor $\boldsymbol{c}_j\alpha$. We now need weighed least squares to update the parameters in the LAWS algorithm. So (5.6) is replaced by the expression

$$\boldsymbol{\alpha}^{(k)} = (C^T W^{(k-1)} C + \lambda_h Q^T Q)^{-1} C^T W^{(k-1)} Y,$$

were C is the circular spline matrix with elements $c_i j$, $W^{(k)}$ is a diagonal matrix with $w_{ij}^{(k)}$ as the diagonal elements.

Without applying smoothing, there can be no more intersections of the lines defining the expectiles then there are data points. So the contours will always be a bit angular. However for quantile contours intersections may be behind an other quantile frontier so (especially for extreme asymmetries) many directions are 'lost'.

Directional contours for quantiles can be made in much the same way as for expectiles. However the expectile contours look much nicer 5.12.

5.5 Tubular expectile regression regression

Smoothing along different angles can be combined with smoothing along an other variable (say age). For this we use multivariate smoothing [*Currie et al.*, 2006]. To do this we work with the tensor product of the P-splines C and B defined in 5.4

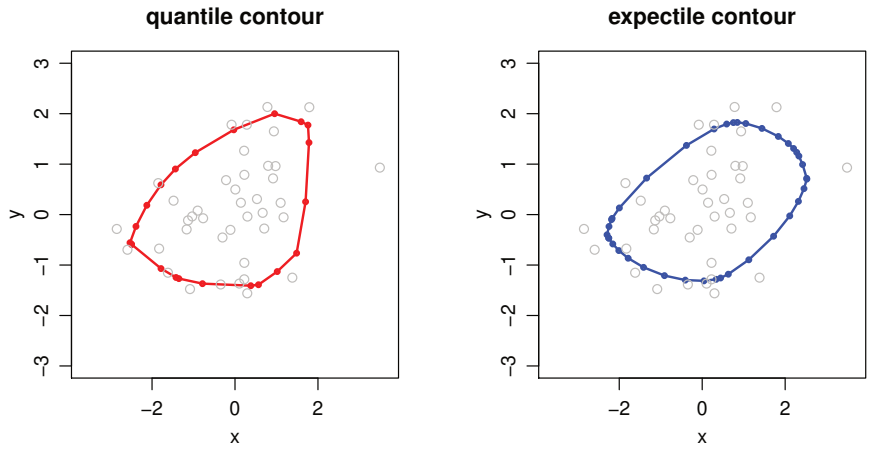


Figure 5.12: A quantile contour and an expectile contour compared. The asymmetry parameter was 0.960 for the directional quantiles and 0.991 for the expectiles (both values corresponding to 1.75 standard deviations on a normal distribution)

and 5.3 respectively. In other words we model the τth expectile in direction θ with a covariate value of t as:

$$\mu_{ij}(x_i, \theta_j) = \sum_k \sum_l \gamma_{ij} C_l(\theta_j) B_k(t_i)$$

We minimize:

$$\sum_i \sum_j w_{ij}(z_{ij} - \mu_{ij}) + \lambda_t \|DG\|_F^2 + \lambda_h \|GQ\|_F^2, \tag{5.9}$$

using the notation $\|.\|_F$ for the Frobenius norm. This equation is similar to (5.8) in that we sum both over the observations and the angles. The difference is that we now take the value of the covariate $\boldsymbol{x}$ of all observations into account and smoothing is done in two directions simultaneously.

Minimizing (5.9) amounts to solving

$$[(C \otimes B)V(B \otimes C) + R]\gamma = (C \otimes B)V \operatorname{vec}(Z) \tag{5.10}$$

for γ. Here V is the diagonal matrix of weights and R is the combined penalty matrix $R = \lambda_x I_m \otimes D^T D + \lambda_\theta C^T C \otimes I_n$. We use array regression [*Currie et al.*, 2006; *Eilers et al.*, 2006] instead of solving 5.10 directly. In this way we prevent calculating and manipulating the large Kronecker products. This makes the calculations both faster and less memory intensive.

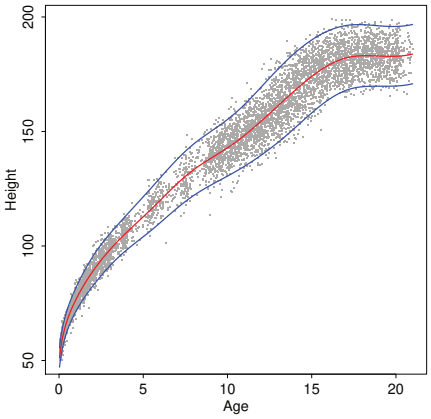


Figure 5.13: Height of Dutch boys with mean (blue) and +/- 2 SD lines (red)

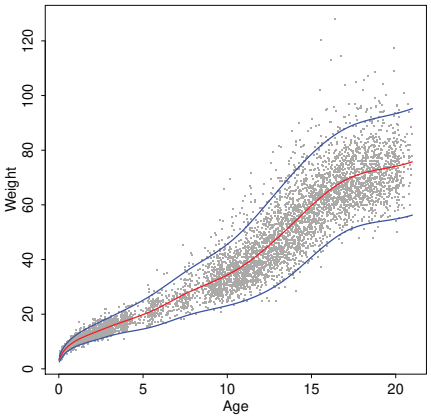


Figure 5.14: Weight of Dutch boys with mean (blue) and +/- 2 SD lines (red)

5.6 Application

We now apply our method on a subset of the data of the Fourth Dutch Growth Study [*Fredriks et al.*, 2000]. We have data on the height and weight of 6.848 Dutch boys. This data was originally collected to construct reference charts using quantiles height and weight separately. Here we will analyze them together. For our analyses we have taken the square root of age as we have more data here and we expect that this is were the curves change fastest.

In figures 5.13 and 5.14 we show weight and height by age. We have modeled the mean of both series by fitting P-spline functions. The absolute values of the residuals from this fit were regression on a P spline expansion of age to obtain the standard deviations as a function of the age for both series [*Altman*, 1993]. The mean and +/- SD lines are added in dark blue and red respectively. With these values we standardize the data points.

These standardized values are now used to estimate the 0.8 and 0.95 expectiles in a large number of directions given the age of he boys. These values are used to calculate the contours which are plotted in figure 5.15 and figure 5.16 for boys of 4, 10 and 16 years respectively. In 5.18 we made a three dimensional visualization of the tubular contours.

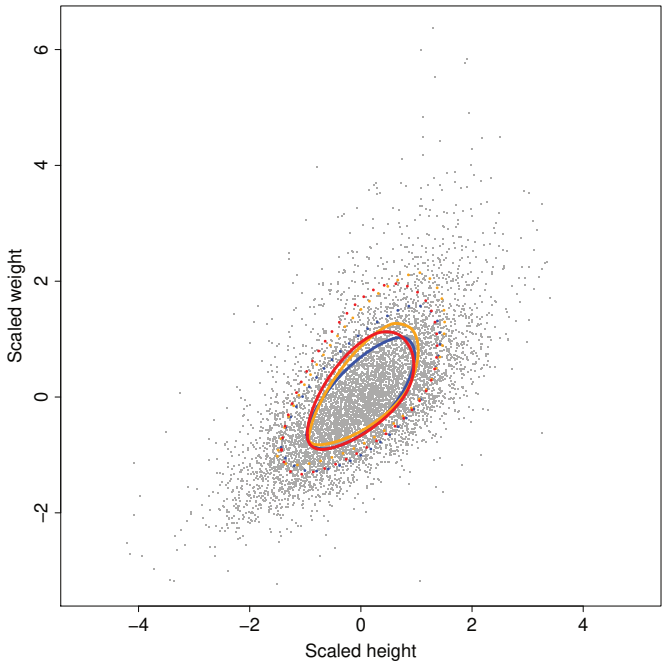


Figure 5.15: Circular expectiles for the Dutch boys data. The expectile contour corresponding to 1 normal standard deviate is solid and that corresponding to 1.5 is dotted. The contours for boys of age 4 are in blue, that of boys of 10 in orange and that of age 16 is in red.

5.7 Discussion

The contribution of this manuscript to the literature is twofold: First, we have developed a new methodology to develop multivariate reference curves based on directional expectiles and expectile regression. Second: this methodology is then extended to take into account a covariate. We smooth in two directions in the direction of a covariate and in the direction of the combination of the responses we are considering.Because we are working in a three dimensional space, data are much sparser then they would be in two dimensions, increasing the need for smoothing.

We imagine that these reference curves could be used by general practitioners, pediatricians and other health care professionals involved in monitoring growth. This would be done in more or less the same way as current reference curves but

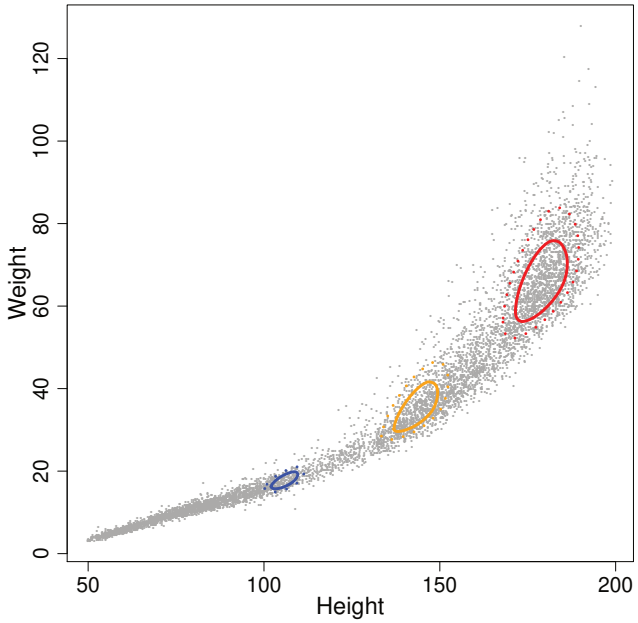


Figure 5.16: Circular expectiles for the Dutch boys data, plotted on the original scale. The expectile contour corresponding to 1 normal standard deviate is solid and that corresponding to 1.5 sd is dotted. The contours for boys of age 4, 10 and 16 years are plotted in blue, orange and red respectively.

providing more insight. Because the contours are three dimensional in nature instead of having them preprinted on paper we imagine a small application either online or on the doctors computer in which the the age and for example weight and height are entered. With this data and then which then displays the 3D contours along with the observations. One can than zoom in and look at the slice of the tube perpendicular to the age axis and see on which contour line an observation is located.

Although we have used expectiles here quantiles can also be used. Basically all one has to do is replace the weight function in the LAWS algorithm to

$$w_i^{(k+1)} = \frac{\tau - a_i}{\sqrt{\epsilon^2 + (y_i - x_i\alpha^{(k)})^2}}$$

Here a_i is 1 when $y_i < x_i\alpha^{(k)}$ and zero otherwise and ϵ is some small constant preventing the denominator to go to zero. Although it has been claimed the LAWS

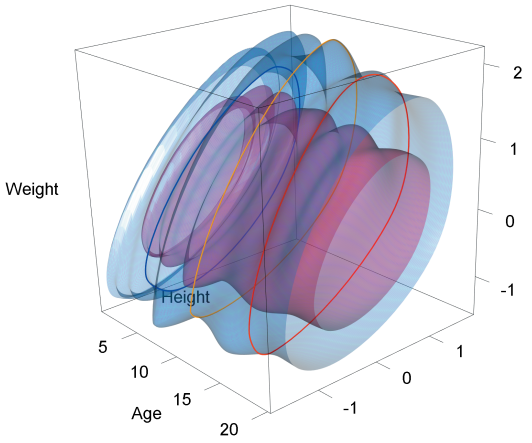


Figure 5.17: Circular expectiles for the Dutch boys data, plotted on the normalized scale.

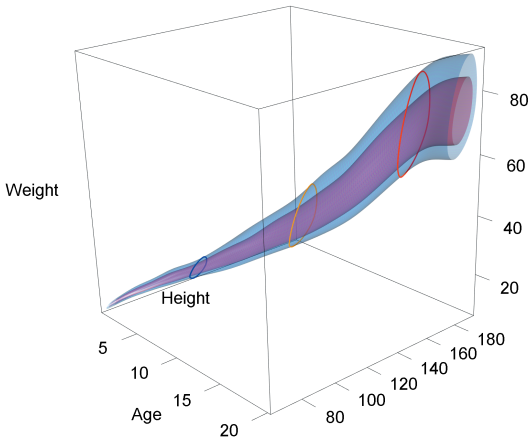
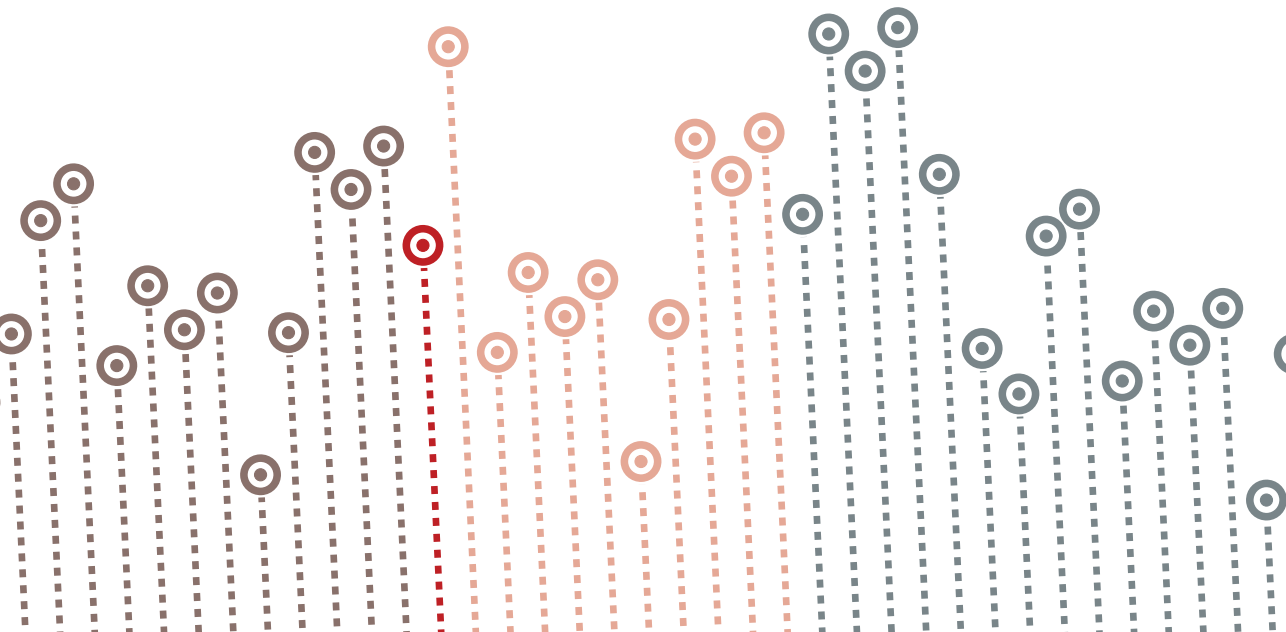


Figure 5.18: A 3D visualization of the tubular contours. The 0.8 expectile contour is displayed in pink, the 0.95 contour in blue.

algorithm is problematic for quantiles this has been refuted by *Schnabel and Eilers* [2013]. It is however true that convergence is much slower, i.e. more iterations of the algorithm are needed.

In this paper we have used tubular expectile contours to make static visualizations. However, when one make interactive plots a whole new range of opportunities opens. Therefore we are currently working on an R package to make this possible. We have visualized the three dimensional tubes using a orthographic projection. However as virtual reality technology is becoming affordable it would also be possible to inspect them in this way. This would require additional software to be written. We have restricted ourselves to two dimensional measurements as otherwise visualization would be impossible. It would be possible to work around this limitation when dimension reduction techniques (like PCA) are used.

6 General conclusions and discussion



In this thesis we have presented some innovations in growth modeling, mostly focusing on multivariate and repeated measures. Below we present the main findings of each chapter, followed by an overall discussion and some directions for further research.

Main findings

In **Chapter 2** we investigated whether it was better to model the height of children on the original scale or on the scale of the standard deviation scores (SDS). Often the SDS scale in growth studies is preferred as one can avoid modeling the full complexity of the overall growth pattern. We investigated this using data on the height of children born small for gestational age which do not show catch up growth and were treated with growth hormone. We demonstrated that the use of standard deviation scores (SDS) can lead to more complicated models. This leads to a fit that might be worse than when modeled on the original scale. Additionally, the model on the SDS scale sometimes showed growth patterns in which the height decreased with age which does not make sense biologically. While in the model on the original scale it is easy to build in the restriction that height is non-decreasing with age, doing so on the SDS scale is much harder.

In **Chapter 3** the SITAR model was generalized to multiple outcomes and estimated using a Bayesian method. The multivariate SITAR (MSITAR) model was used to examine the impact of several risk factors. Additionally, we used the model to create both population as well as subject-specific multivariate reference regions using direction quantiles applied to samples from the posterior predictive distribution. Using cross-validation, we found that the fit of the multivariate model improved upon that of the univariate ones. A benefit of the multivariate model was also that the lost information due to missing values can partly be recovered from the other outcome. Furthermore, a multivariate model allows for the examination of the relation between the outcomes.

In **Chapter 4** we adapted the MSITAR model that was introduced in Chapter 3 to pharmacokinetic data. To be suitable to this new field, some features of the model had to be changed to match the characteristics of the data. Here we concluded that the MSITAR model provided a fit that was somewhat better than that of a traditional compartment model. This demonstrates the usefulness of the SITAR and MSITAR model outside the study of human growth.

In **Chapter 5** we demonstrated how expectiles can be used to study a bivariate outcome (say height and weight) that may depend on a third covariate (e.g. age). Using the expectiles projections along several directions, circular contours can be generated. The dependence on the direction can be smoothed using splines and penalties. A similar smoothing technique can be employed to model the dependence on the covariate. Combining the two we want to smooth along two dimensions; this can be done using multivariate splines and penalties. We conclude that the resulting tubular contours are a good and practically useful way to appraise multivariate growth.

Overall discussion

In most sections we made use of Bayesian methodology. This approach appeals to us based for its own sake. But it also allowed us to incorporate features like censoring or correlated data in an much easier way than what is possible using other methods.

Another theme that returns in most chapters of this thesis is that of multivariate modeling.

Topics for further research

One limitation of our models is that they might be difficult to apply in practice because there is no standard software. This could be easily remedied by creating an R package.

Certain features of the data were not fully incorporated in the models. For example, when modeling the height of children we observed serial correlation in the errors. This can be partly modelled. In the MSITAR model we assumed that the residual errors were uncorrelated. This is probably not realistic. Finally in the application of the MSITAR to pharmacokinetic data there was heteroskedasticity that we have not further attempted to model.

The SITAR model assumes that there is a single template function that applies to everybody. This of course might not be the case. To accommodate different templates in the population a mixture model is needed. An other extension to the MSITAR model is to allow any or all of the outcomes to be categorical. In the

context of human growth one could for example think of modeling developmental stages such as the Carnegie stage of embryonic development or the Tanner stage of the development of sex characteristics alongside other contentious measurements. Using Bayesian methods, inclusion of these additional features in our models is very easy however estimating the models may become very hard as will be discussed below.

In many chapters we used Gibbs sampling to estimate our models. This technique has many advantages but might also be quite slow. In complicated models, with high autocorrelations, very large sample sizes are needed. Especially when combined with other computer intensive methods like cross validation, the running time becomes prohibitive. For this reason other techniques should be investigated that could reduce sampling time. Perhaps the Hamiltonian MCMC technique, implemented in STAN, could be at least a partial solution.

Another issue is the convergence of the models. When we implemented additional features in our models we noticed that they became very sensitive to the starting values or showed bad mixing of the chains. Here also research in new numerical techniques is warranted, perhaps some kind of stochastic annealing could be used.

In this thesis we have stressed that multiple outcomes should be examined in a multivariate way. At the same time we have to admit that the MSITAR model and the tubular expectiles only allow for a limited number of outcomes before they break down. New methods involving dimension reduction are needed when we want to model even higher numbers of outcomes. One direction might be offered by looking at the literature for multi-mode models (or tensor-based data reduction models). In this field one also looks at multiple series of repeated measurements and generalizes traditional dimension reduction techniques to work with these multiple modes). Another possibility would be to try to devise a method that first looks at pairwise combinations of different outcomes and combines this information in a second stage [*Fieuws and Verbeke*, 2006].

When studying growth, we have looked at the development of some trait over time. Often the growth velocity (i.e. the change over time) and the acceleration of growth that are of interest. These could be derived from models based on size but it could also be modeled directly.

In Chapter 2 we predicted the adult height of children using two mixed models. The models aim to fit the whole growth curve well, but are not specifically focused

on predicting the final height. We therefore wonder if we can achieve better prediction using other kinds of models. One option would be a kind of joint model with one part describing childhood growth and the second part linking the subject specific effects of childhood growth to adult height. It would be interesting to see how these and other types of prediction models would compare on predictive accuracy.

In Chapter 5 we estimate expectile and quantiles contours for a single value of the asymmetry parameter at a time. It is easy to see how the method could be extended to look at several values simultaneously. This would then involve smoothing along three dimensions: age, angle and asymmetry. An other route in which the work in this chapter can be extended is by looking at the methods to estimate the expectiles and quantiles. They can be either estimated from a parametric distribution or be fully nonparametric. It would be useful to combine the flexibility of the non-parametric methods with prior knowledge on the approximate parametric shape.

Bibliography

Allen, D., Growth hormone therapy for short stature: is the benefit worth the burden?, *Pediatrics*, *118*(1), 343–348, doi:10.1542/peds.2006-0329, 2006.

Altman, D. G., Construction of age-related reference centiles using absolute residuals, *Statistics in Medicine*, *12*(10), 917–924, doi:10.1002/sim.4780121003, 1993.

Bakker, B., W. Oostdijk, R. Geskus, W. Stokvis-Brantsma, J. Vossen, and J. Wit, Patterns of growth and body proportions after total-body irradiation and hematopoietic stem cell transplantation during childhood., *Pediatr Res*, *59*(2), 259–264, doi:10.1203/01.pdr.0000199550.71887.ba, 2006.

Barker, D. J. P., *Mothers, babies and health in later life*, Elsevier Health Sciences, 1998.

Beath, K. J., Infant growth modelling using a shape invariant model with random effects, *Statistics in Medicine*, *26*(12), 2547–2564, 2007.

Boguszewski, M., et al., Growth hormone treatment of short children born small-for-gestational-age: the nordic multicentre trial., *Acta Paediatr*, *87*(3), 257–263, 1998.

Bottomley, C., A. Daemen, F. Mukri, A. T. Papageorghiou, E. Kirk, A. Pexsters, B. De Moor, D. Timmerman, and T. Bourne, Assessing first trimester growth: the influence of ethnic background and maternal age., *Human Reproduction*, *24*(2), 284–290, 2009.

Bukowski, R., et al., Fetal growth in early pregnancy and risk of delivering low birth weight infant: prospective cohort study., *BMJ*, *334*(7598), 836, 2007.

Cameron, N., and E. W. Demerath, Critical periods in human growth and their relationship to diseases of aging, *American Journal of Physical Anthropology*, *119*(S35), 159–184, doi:10.1002/ajpa.10183, 2002.

Carbone, J. F., M. G. Tuuli, R. Bradshaw, J. Liebsch, and A. O. Odibo, Efficiency of first-trimester growth restriction and low pregnancy-associated plasma protein-a in predicting small for gestational age at delivery, *Prenatal Diagnosis*, *32*(8), 724–729, doi:10.1002/pd.3891, 2012.

Chen, Y. I., and . Huang, C. S., New approach to assess bioequivalence parameters using generalized gamma mixed-effect model (model-based asymptotic bioequivalence test), *Statistics in Medicine*, *33*(5), 786–797, 2013.

Cole, T., and P. Green, Smoothing reference centile curves: the lms method and penalized likelihood., *Stat Med*, *11*(10), 1305–1319, 1992.

Cole, T. J., The lms method for constructing normalized growth standards., *European journal of clinical nutrition*, *44*(1), 45–60, 1990.

Cole, T. J., M. D. Donaldson, and Y. Ben-Shlomo, SITAR–a useful instrument for growth curve analysis, *International Journal of Epidemiology*, *39*(6), 1558–1566, doi:10.1093/ije/dyq115, 2010a.

Cole, T. J., M. D. Donaldson, and Y. Ben-Shlomo, SITAR – a useful instrument for growth curve analysis, *International Journal of Epidemiology*, *39*(6), 1558–1566, 2010b.

Cooper, C., J. Eriksson, T. Forsén, C. Osmond, J. Tuomilehto, and D. Barker, Maternal height, childhood growth and risk of hip fracture in later life: a longitudinal study., *Osteoporos Int*, *12*(8), 623–629, 2001.

Curran, P. J., K. Obeidat, and D. Losardo, Twelve frequently asked questions about growth curve modeling, *Journal of Cognition and Development*, *11*(2), 121–136, 2010.

Currie, I. D., M. Durban, and P. H. Eilers, Generalized linear array models with applications to multidimensional smoothing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 259–280, 2006.

Dalton, V., et al., Height and weight in children treated for acute lymphoblastic leukemia: relationship to CNS treatment., *J Clin Oncol*, *21*(15), 2953–2960, doi:10.1200/JCO.2003.03.068, 2003.

Davenport, M., N. Punyasavatsut, P. Stewart, D. Gunther, L. Sävendahl, and V. Sybert, Growth failure in early life: an important manifestation of turner syndrome., *Horm Res*, *57*(5-6), 157–164, 2002.

Dawes, B., and D. Abrahams, Boost C++ libraries, 2012.

De Boor, C., *A practical guide to splines; rev. ed.*, Applied mathematical sciences, Springer, Berlin, 2001.

De Graaf, J., A. Ravelli, G. Visser, C. Hukkelhoven, W. Tong, G. Bonsel, and E. Steegers, Increased adverse perinatal outcome of hospital delivery at night, *BJOG: An International Journal of Obstetrics & Gynaecology*, *117*(9), 1098–1107, doi:10.1111/j.1471-0528.2010.02611.x, 2010.

De Zegher, F., K. Albertsson-Wikland, H. Wollmann, P. Chatelain, J. Chaussain, A. Löfström, B. Jonsson, and R. Rosenfeld, Growth hormone treatment of short children born small for gestational age: growth responses with continuous and discontinuous regimens over 6 years., *J Clin Endocrinol Metab*, *85*(8), 2816–2821, 2000.

De Zegher, F., K. Ong, L. Ibáñez, and D. Dunger, Growth hormone therapy in short children born small for gestational age., *Horm Res*, *65 Suppl 3*, 145–152, doi:10.1159/000091520, 2006.

Dubois, A., S. Gsteiger, E. Pigeolet, and F. Mentré, Bioequivalence tests based on individual estimates using non-compartmental or model-based analyses: evaluation of estimates of sample means and type I error for different designs, *Pharmaceutical Research*, *27*, 92–104, 2010.

Dubois, A., M. Lavielle, S. Gsteiger, E. Pigeolet, and F. Mentré, Model-based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm, *Statistics in Medicine*, *30*, 2582–2600, 2011.

Eilers, P., Asymmetric least squares: New faces of a scatterplot, *Kwantitatieve Methoden*, *8*, 45–62, 1987.

Eilers, P. H. C., and B. D. Marx, Flexible smoothing with B-splines and penalties, *Statist. Sci.*, *11*(2), 89–121, doi:10.1214/ss/1038425655, 1996.

Eilers, P. H. C., and B. D. Marx, Generalized linear additive smooth structures, *Journal of Computational and Graphical Statistics*, *11*(4), 758–783, 2002.

Eilers, P. H. C., I. D. Currie, and M. Durbán, Fast and compact smoothing on large multidimensional grids, *Comput. Statist. Data Anal.*, *50*(1), 61–76, doi: 10.1016/j.csda.2004.07.008, 2006.

European Agency for the Evaluation of Medicinal Products, Note for guidance on the investigation of bioavailability and bioequivalence., `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003519.pdf`, accessed 12–May–2015, 2001.

Fieuws, S., and G. Verbeke, Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles, *Biometrics*, *62*(2), 424–431, 2006.

Forni, S., M. Piles, A. Blasco, L. Varona, H. N. Oliveira, R. B. Lôbo, and L. G. Albuquerque, Comparison of different nonlinear functions to describe nelore cattle growth, *Journal of Animal Science*, *87*(2), 496–506, doi:10.2527/jas.2008-0845, 2009.

Foulkes, M., and C. Davies, An index of tracking for longitudinal data, *Biometrics*, *Vol 37, No 3*, 439–446, 1981.

Fradette, C., J. Lavigne, D. Waters, and M. Ducharme, The utility of the population approach applied to bioequivalence in patients: Comparison of 2 formulations of cyclosporine, *Therapeutic Drug Monitoring*, *27*(5), 592–600, 2005.

Fredriks, A. M., S. van Buuren, R. J. F. Burgmeijer, J. F. Meulmeester, R. J. Beuker, E. Brugman, M. J. Roede, S. P. Verloove-Vanhorick, and J.-M. Wit, Continuing positive secular growth change in the netherlands 1955-1997, *Pediatr Res*, *47*(3), 2000.

Geisser, S., Discussion on sampling and Bayes' inference in scientific modeling and robustness (by G.E.P. box), *Journal of the Royal Statistical Society A*, *143*, 416–417, 1980.

Geisser, S., and W. F. Eddy, A predictive approach to model selection, *Journal of the American Statistical Association*, *74*(365), 153–160, 1979.

Gelfand, A. E., and A. F. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American statistical association*, *85*(410), 398–409, 1990.

Gelman, A., and D. B. Rubin, Inference from iterative simulation using multiple sequences, *Statistical Science*, *7*(4), pp. 457–472, 1992.

Gelman, A., X. Meng, and H. Stern, Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica*, pp. 733–807, 1996.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, CRC Press, 2004.

Gelman, A., D. A. Van Dyk, Z. Huang, and J. W. Boscardin, Using redundant parameterizations to fit hierarchical models, *Journal of Computational and Graphical Statistics*, *17*, 95–122, 2012.

Geman, S., and D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721–741, 1984.

Gilks, W., and G. Roberts, Markov chain Monte Carlo in Practice, chap. Strategies for improving MCMC, Chapman & Hall/CRC, 1996.

Giorgi, E., and A. J. McNeil, On the computation of multivariate scenario sets for the skew-t and generalized hyperbolic families, *Computational Statistics and Data Analysis*, *100*, 2014.

Gnanadesikan, R., and J. R. Kettenring, Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, pp. 81–124, 1972.

Goldstein, H., Efficient statistical modelling of longitudinal data., *Annals of Human Biology*, *13*(2), 129–141, 1986.

Grimm, K. J., N. Ram, and R. Estabrook, *Growth modeling: Structural equation and multilevel modeling approaches*, Guilford Publications, 2016.

Guennebaud, G., B. Jacob, et al., Eigen v3, http://eigen.tuxfamily.org, 2010.

Hastings, W. K., Monte carlo sampling methods using markov chains and their applications, *Biometrika*, *57*(1), 97–109, 1970.

Helsel, D. R., Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it, *Chemosphere*, *65*(11), 2434–2439, 2006.

Hokken-Koelega, A., W. De Waal, T. Sas, Y. Van Pareren, and N. Arends, Small for gestational age (SGA): endocrine and metabolic consequences and effects of growth hormone treatment., *J Pediatr Endocrinol Metab*, *17 Suppl 3*, 463–469, 2004.

Hokken-Koelinga, A., M. De Ridder, R. Lemmen, H. Den hartog, S. De Muinck Keizer-Schrama, and S. Drop, Children born small for gestational age: Do they catch up?, *Pedeatric Research*, *Vol. 38, No. 2*, 267–271, 1995.

Hu, C., K. H. P. Moore, Y. H. Kim, and M. E. Sale, Statistical issues in a modeling approach to assessing bioequivalence or PK similarity with presence of sparsely sampled subjects, *Journal of Pharmacokinetics and Pharmacodynamics*, *31*(4), 321–339, 2004.

Hypponen, E., S. Virtanen, M. Kenward, M. Knip, and H. Akerblom, Obesity, increased linear growth, and risk of type 1 diabetes in children, *Diabetes Care*, *23:12*, 1755–1760, 2000.

Jenss, R. M., and N. Bayley, A mathematical method for studying the growth of a child, *Human Biology*, *9*(4), 556, 1937.

Johnson, W., Analytical strategies in human growth research: Analytical Strategies in Human Growth Research, *American Journal of Human Biology*, *27*(1), 69–83, doi:10.1002/ajhb.22589, 2015.

Johnston, L., and M. Savage, Should recombinant human growth hormone therapy be used in short small for gestational age children?, *Arch Dis Child*, *89*(8), 740–744, doi:10.1136/adc.2003.034785, 2004.

Kamboj, M., Short stature and growth hormone., *Indian J Pediatr*, *72*(2), 149–157, 2005.

Kamp, G., J. Waelkens, S. de Muinck Keizer-Schrama, H. Delemarre-Van de Waal, L. Verhoeven-Wind, A. Zwinderman, and J. Wit, High dose growth hormone treatment induces acceleration of skeletal maturation and an earlier onset of puberty in children with idiopathic short stature., *Arch Dis Child*, *87*(3), 215–220, 2002.

Karaolis-Danckert, N., A. Buyken, K. Bolzenius, C. D. F. Perim, M. Lentze, and A. Kroke, Rapid growth among term children whose birth weight was appropriate for gestational age has a longer lasting effect on body fat percentage than on body mass index., *Am J Clin Nutr*, *84*(6), 1449–1455, 2006.

Kjaergaard, S., J. Müller, and F. Skovby, Prepubertal growth in congenital disorder of glycosylation type ia (cdg-ia)., *Arch Dis Child*, *87*(4), 324–327, 2002.

Koenker, R., *Quantile Regression*, Econometric Society Monographs, Cambridge University Press, 2005.

Koenker, R. W., and V. D'Orey, Algorithm as 229: Computing regression quantiles, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *36*(3), 383–393, 1987.

Kong, L., and I. Mizera, Quantile tomography: using quantiles with multivariate data, *Statistica Sinica*, *22*, 1589–1610, 2008.

Kong, L., and I. Mizera, Quantile tomography: using quantiles with multivariate data, *Statsitica Sinica*, *22*(4), 1589–1610, 2012.

Koning, A. H. J., M. Rousian, C. M. Verwoerd-Dikkeboom, L. Goedknegt, E. A. P. Steegers, and P. J. van der Spek, V-scope: design and implementation of an immersive and desktop virtual reality volume visualization system., *Stud Health Technol Inform*, *142*, 136–138, 2009.

Kuhn, E., and M. Lavielle, Maximum likelihood estimation in nonlinear mixed effects models, *Computational Statistics & Data Analysis*, *49*(4), 1020–1038, 2005.

LaFleur, B., W. Lee, D. Billhiemer, C. Lockhart, J. Liu, N. Merchant, et al., Statistical methods for assays with limits of detection: Serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer, *Journal of Carcinogenesis*, *10*(1), 12, 2011.

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

Lee, L., S. Chernausek, A. Hokken-Koelenga, P. Czernichow, and "the International SGA Advisory Board", International small for gestational age advisory board consensus development conference statement: Management of short children born small for gestational age, *Pedeatrics*, *Vol. 111 No. 6*, 1253–1261, 2003.

Lesaffre, E., and A. Lawson, *Bayesian Biostatistics*, Statistics in Practice, Wiley, 2012.

Lindstrom, M. J., and D. M. Bates, Nonlinear mixed effects models for repeated measures data, *Biometrics*, *46*(3), 673–687, 1990.

Macdonald-Wallis, C., D. A. Lawlor, T. Palmer, and K. Tilling, Multivariate multilevel spline models for parallel growth processes: application to weight and mean arterial pressure in pregnancy, *Statistics in Medicine*, *31*(26), 3147–3164, doi:10.1002/sim.5385, 2012.

Maršál, K., P.-H. Persson, T. Larsen, H. Lilja, A. Selbing, and B. Sultan, Intrauterine growth curves based on ultrasonically estimated foetal weights, *Acta Paediatrica*, *85*(7), 843–848, doi:10.1111/j.1651-2227.1996.tb14164.x, 1996.

Mook-Kanamori, D., E. Steegers, P. Eilers, H. Raat, A. Hofman, and V. Jaddoe, Risk factors and outcomes associated with first-trimester fetal growth restriction, *The Journal of the American Medical Association*, *303*, 527–534, 2010.

Mukri, F., T. Bourne, C. Bottomley, C. Schoeb, E. Kirk, and A. T. Papageorghiou, Evidence of early first-trimester growth restriction in pregnancies that subsequently end in miscarriage., *BJOG*, *115*(10), 1273–1278, doi:10.1111/j.1471-0528.2008.01833.x, 2008.

Neal, R. M., Slice sampling, *Annals of statistics*, pp. 705–741, 2003.

Newey, W. K., and J. L. Powell, Asymmetric least squares estimation and testing, *Econometrica*, *55*(4), 819–847, 1987.

Nguyen, N., J. Allen, J. Peat, W. Schofield, V. Nossar, M. Eisenbruch, and K. Gaskin, Growth and feeding practices of vietnamese infants in australia., *Eur J Clin Nutr*, *58*(2), 356–362, doi:10.1038/sj.ejcn.1601791, 2004.

Nieto, F. J., M. Szklo, and G. W. Comstock, Childhood weight and growth rate as predictors of adult mortality, *American Journal of Epidemiology*, *136*(2), 201–213, 1992.

Ong, K., M. Preece, P. Emmett, M. Ahmed, D. Dunger, and A. S. Team, Size at birth and early childhood growth in relation to maternal smoking, parity and infant breast-feeding: longitudinal birth cohort study and analysis., *Pediatr Res*, *52*(6), 863–867, 2002.

Oort, F. J., Three-mode models for multivariate longitudinal data, *British Journal of Mathematical and Statistical Psychology*, *54*(1), 49–78, doi:10.1348/000711001159429, 2001.

Pinheiro, J. C., and D. M. Bates, Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics*, *4*(1), 12–35, 1995.

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

Plummer, M., JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, pp. 1–10, 2003a.

Plummer, M., Jags: A program for analysis of Bayesian graphical models using Gibbs sampling, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003b.

Plummer, M., N. Best, K. Cowles, and K. Vines, CODA: Convergence diagnosis and output analysis for MCMC, *R News*, *6*(1), 7–11, 2006.

Portnoy, S., and R. Koenker, The gaussian hare and the laplacian tortoise: computability of squared-error and absolute-error estimation; to appear, *Statistical Science*, *12*(4), 279–300, 1997.

Potthoff, R., and S. Roy, A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, pp. 313–326, 1964.

Reinsel, G., Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure, *Journal of the American Statistical Association*, *77*(377), pp. 190–195, 1982.

Rigby, R. A., and D. M. Stasinopoulos, Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554, 2005.

Roede, M., and J. Van Wieringen, Growth diagrams 1980, netherlands. third nation-wide survey, *T Soc Gezondheidsz*, *63(Suppl)*, 1—-34., 1985.

Rolo, L. C., L. M. M. Nardozza, E. Araujo Júnior, P. M. Nowak, J. Bortoletti Filho, and A. F. Moron, Measurement of embryo volume at 7-10 weeks' gestation by 3d-sonography., *Journal of Obstetrics & Gynaecology*, *29*(3), 188–191, doi: 10.1080/01443610902765204, 2009.

Rosen, O., and W. K. Thompson, A Bayesian regression model for multivariate functional data, *Computational Statistics & Data Analysis*, *53*(11), 3773–3786, doi:10.1016/j.csda.2009.03.026, 2009.

Rotteveel, J., M. Van Weissenbruch, J. Twisk, and H. Delemarre-Van de Waal, Infant and childhood growth patterns, insulin sensitivity, and blood pressure in prematurely born young adults., *Pediatrics*, *122*(2), 313–321, doi:10.1542/peds.2007-2012, 2008.

Rousian, M., A. Koning, R. van Oppenraaij, W. Hop, C. Verwoerd-Dikkeboom, P. van der Spek, N. Exalto, and E. Steegers, An innovative virtual reality technique for automated human embryonic volume measurements, *Human Reproduction*, *25*(9), 2210–2216, doi:10.1093/humrep/deq175, 2010.

Rousseeuw, P. J., and B. C. Van Zomeren, Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, *85*(411), pp. 633–639, 1990.

Ruppert, D., M. Wand, and R. Carroll, *Semiparametric regression*, Cambridge University Press, 2003.

Sas, T., W. De Waal, P. Mulder, M. Houdijk, M. Jansen, M. Reeser, and A. Hokken-Koelega, Growth hormone treatment in children with short stature born small for gestational age: 5-year results of a randomized, double-blind, dose-response trial., *J Clin Endocrinol Metab*, *84*(9), 3064–3070, 1999.

Scammon, R. E., The first seriatim study of human growth, *American Journal of Physical Anthropology*, *10*(3), 329–336, doi:10.1002/ajpa.1330100303, 1927.

Schall, R., and H. G. Luus, On population and individual bioequivalence, *Statistics in Medicine*, *12*(12), 1109–1124, doi:10.1002/sim.4780121202, 1993.

Schnabel, S. K., and P. H. Eilers, Optimal expectile smoothing, *Computational Statistics & Data Analysis*, *53*(12), 4168–4177, 2009.

Schnabel, S. K., and P. H. C. Eilers, Simultaneous estimation of quantile curves using quantile sheets, *AStA Advances in Statistical Analysis*, *97*(1), 77–87, doi: 10.1007/s10182-012-0198-1, 2013.

Serfling, R., Quantile functions for multivariate analysis: approaches and applications, *Statistica Neerlandica*, *56*(2), 214–232, doi:10.1111/1467-9574.00195, 2002.

Slaughter, J. C., A. H. Herring, and J. M. Thorp, A Bayesian latent variable mixture model for longitudinal fetal growth, *Biometrics*, *65*(4), 1233–1242, doi: 10.1111/j.1541-0420.2009.01188.x, 2009.

Smith, G. C., M. F. Smith, M. B. McNay, and J. E. Fleming, First-trimester growth and the risk of low birth weight, *New England Journal of Medicine*, *339*(25), 1817–1822, doi:10.1056/NEJM199812173392504, 1998.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van der Linde, Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*, 583–639, 2002a.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn, *WinBUGS User Manual Version 1.4*, 2003.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde, Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639, doi:10.1111/1467-9868.00353, 2002b.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde, Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639, 2002c.

Tukey, J., Mathematics and the picturing of data, *Proceedings of the International Congress of Mathematicians*, *Canad. Math. Congress*, 523–53, 1975.

Turck, D., et al., World health organization 2006 child growth standards and 2007 growth reference charts: a discussion paper by the committee on nutrition of the european society for pediatric gastroenterology, hepatology, and nutrition, *Journal of pediatric gastroenterology and nutrition*, *57*(2), 258–264, 2013.

U. S. Food and Drug Administration, Guidance for industry: Statistical approaches to establishing bioequivalence, `http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf`, accessed 12–May–2015, 2001.

Van Pareren, Y., P. Mulder, M. Houdijk, M. Jansen, M. Reeser, and A. Hokken-Koelega, Adult height after long-term, continuous growth hormone (gh) treatment in short children born small for gestational age: results of a randomized, double-blind, dose-response gh trial., *J Clin Endocrinol Metab*, *88*(8), 3584–3590, 2003.

Van Uitert, E., et al., An optimal periconception maternal folate status for embryonic size: the rotterdam predict study, *BJOG: An International Journal of Obstetrics & Gynaecology*, pp. n/a–n/a, doi:10.1111/1471-0528.12592, 2014.

Van Uitert, E. M., N. Exalto, G. J. Burton, S. P. Willemsen, A. H. Koning, P. H. Eilers, J. S. Laven, E. A. Steegers, and R. P. Steegers-Theunissen, Human embryonic growth trajectories and associations with fetal growth and birthweight, *Human Reproduction*, *28*(7), 1753–1761, doi:10.1093/humrep/det115, 2013.

Van Wieringen, J., and H. Verbrugge, [growth diagrams of infants. a national survey, 1964-1966], *Maandschr Kindergeneeskd*, *34*(12), 389–414, 1966.

Voss, L. D., The measurement of human growth: A historical review, in *Perspectives in Human Growth, Development and Maturation*, edited by P. Dasgupta and R. Hauspie, pp. 3–15, Springer Netherlands, Dordrecht, doi:10.1007/978-94-015-9801-9_1, 2001.

Willemsen, S. P., P. H. Eilers, R. Steegers-Theunissen, and E. Lesaffre, A multivariate Bayesian model for human growth, *Statistics in Medicine*, *34*, 1351–1365, doi:10.1002/sim.6411, 2015.

Wright, E., and P. Royston, Calculating reference intervals for laboratory measurements., *Stat Methods Med Res*, *8*(2), 93–112, 1999.

Wu, L., Exact and approximate inferences for nonlinear mixed-effects models with missing covariates, *Journal of the American Statistical Association*, *99*(467), 700–709, 2004.

Xu, S., M. Styner, J. Gilmore, and G. Gerig, Multivariate longitudinal statistics for neonatal-pediatric brain tissue development, in *Proceedings*, pp. 69,140C–69,140C–11, 2008.

Zemel, B., D. Kawchak, E. Fung, K. Ohene-Frempong, and V. Stallings, Effect of zinc supplementation on growth and body composition in children with sickle cell disease., *Am J Clin Nutr*, *75*(2), 300–307, 2002.

Summary / Samenvatting

Summary

This thesis deals with models for growth. In human growth studies we study various anthropological measurements and their development over time. Human growth is studied for various purposes: we can describe a population, make individual predictions, examine factors influencing development or study how (abnormal) growth patterns influence outcomes later in life. For this we need good models.

Chapter 1 gives a short introduction to the field of growth modeling. In also introduces some important concepts and models.

In **Chapter 2** we model the height of children born small for gestational age that did not show catch-up growth and were treated with growth hormone. The main question of interest is whether it is better to model the data on the original scale or on the scale of the standard deviation scores (SDSs) as is often done in practice. We find that for this dataset modeling on the original scale produces more accurate outcomes. Moreover some of the growth curves on the SDS scale were downward sloping, which is biologically impossible.

In **Chapter 3** the Multivariate Superimposition by Translation And Rotation model is introduced (MSITAR) as an extension of the SITAR model by Cole. In the SITAR model, there is a general, or template function for the overall growth pattern which is modeled by a spline. From this template the individual growth curves are then derived by shifting it horizontally and vertically and stretching it in the horizontal direction. Here several outcomes are modeled in this way where we allow the subject specific effects of each outcome to be correlated.

In **Chapter 4** we apply the MSITAR model in an other field and demonstrate its use in a bioequivalence study. For this the model had to be slightly adjusted; we work with a vertical stretch instead of a shift and some restrictions are put on the spline to make sure it always begins at zero. Using this model we were able to study both population and individual bioequivalence. The MSITAR model was compared to a compartment model which is traditionally used to assess bioequivalence.

Chapter 5 returns to to topic of human growth. Here we discuss expectiles, an interesting alternative for quantiles. By projecting the data on lines in a large number of directions and calculating expectiles along each directions we can construct a circular reference contour. This contour can be smoothed by using

splines and penalties. This smoothing can also be done when modeling the dependence on a covariate. In this way we create tabular reference regions that can be used to assess growth casually.

Finally **Chapter 6** presents the main findings of each chapter and provides some directions for further research.

Samenvatting

Dit proefschrift gaat over groei. In groeistudies kijken we naar verschillende antropologische metingen en hun ontwikkeling over de tijd. Er zijn verschillende redenen waarom groei bestudeerd wordt: het beschrijven van een populatie, het maken van individuele voorspellingen van groeiuitkomsten, het onderzoeken van factoren die een effect hebben op de groei of bestuderen hoe groei uitkomsten die zich later voordoen beïnvloedt. Hiervoor zijn goede groeimodellen nodig.

In **Hoofstuk 1** geven we een algemene inleiding en we bespreken de context van dit proefschrift en enkele basisbegrippen die nodig zijn om de vervolghoofdstukken te begrijpen.

In **Hoofstuk 2** modelleren we de lengte van kinderen die te klein zijn in verhouding met de zwangerschapsduur (Small for Gestational Age of SGA) die geen inhaalgroei laten zien en die behandeld worden met groeihormoon. We willen onderzoeken of het beter is om een model te schatten op de oorspronkelijke schaal of dt het beter is om de data eerst te converteren naar de standaard deviatie score (SDS) zoals in de praktijk vaak wordt gedaan. We zien als we data van een trial heranalyseren dat het model op de oorspronkelijke schaal betere schattingen geeft. Sommige groeicurven vertonen bovendien een dalend verloop hetgeen biologisch gezien onmogelijk is.

In **Hoofdstuk 3** introduceren we het Multivariate SITAR model als een uitbreiding op het SITAR model van Cole. In dit model, wordt een sjabloon of algemene groeicurve met behulp van een spline gemodelleerd. Van dit sjabloon kunnen we dan de individuele groeicurven afleiden door het horizontaal en verticaal te verschuiven en door het uit te rekken. Hier worden verschillende groeiuitkomsten tegelijkertijd gemodelleerd waarbij we aannemen dat de subject specifieke parameters gecorreleerd kunnen zijn.

In **Hoofdstuk 4** gebruiken we het MSITAR model in een ander toepassingsgebied namelijk in een bioequivalentiestudie. We moesten het hier wel iets voor

veranderen; we werken hier met een verticale uitrekking in plaats ver een verschuiving en we plaatsen restricties op de spline om er zeker van te zijn dat iedereen begint zonder een meetbare concentratie van het medicijn in het bloed. Gebruikmakend van het model onderzoeken we zowel de bioequivalentie op het niveau van de populatie al individuele bioequivalentie. Het MSITAR wordt vergeleken met een compartment model een model dat normaal gebruikt wordt om bioequivalentie te onderzoeken.

Hoofdstuk 5 gaat weer over menselijke groei. We introduceren expectielen, een interessant alternatief voor kwantielen. Door de data te projecteren op lijnen in een groot aantal richtingen en dan in elke richting een expectiel te berekenen creëren we een circulair expectiel contour. We kunnen deze contouren 'gladder' maken door het toepassen van splines. Hetzelfde kunnen we doen wanneer we de contouren ook afhankelijk maken van een andere variabele. Op deze manier worden buisvormige contouren afgeleid die gebruikt kunnen worden om groei te visualiseren en te beoordelen.

In **Hoofdstuk 6** geven we de conclusies van de afzonderlijke hoofdstukken en doen we aanbevelingen voor verder onderzoek.

Dankwoord / Acknowledgements

Bij het maken van dit proefschrift zijn verschillende mensen betrokken geweest. Ik wil deze mensen op deze plek graag daarvoor bedanken.

Theo Stijnen, Hans van der Wouden en Bart Koes door jullie ben ik bij het Erasmus MC begonnen bij de afdelingen Huisartsgeneeskunde en Epidemiologie en daardoor liggen jullie dus aan de basis van dit proefschrift; Ik moest er nog lang over nadenken of ik deze stap echt wou nemen maar ben toch blij dat ik het heb gedaan. Emmanuel Lesaffre, dank dat je mijn promotor hebt willen zijn me en de kans hebt gegeven om dit proefschrift te schrijven. Dimitris Rizopoulos, het nieuwste hoofd van de Biostatistiek, jou wil ik graag bedanken omdat je mij mijn proefschrift af hebt laten maken op de afdeling en voor je advies bij de 'echt moeilijke' statistische problemen.

Wim Hop en Paul Mulder, jullie kon ik in mijn eerste periode bij Biostatistiek alles vragen over alle aspecten van het geven van het geven van een statistisch advies. Heel erg bedankt daarvoor.

Uiteraard wil ik mijn coauteurs bedanken. De artikelen in dit proefschrift zijn er mede dankzij jullie. Cibele, thank you for letting me hijack the pharmacokinetic idea; Finishing it was a lot more work than we both had thought. Paul Eilers bedankt voor je steun bij het schrijven van het laatste artikel en voor de interessante discussies die we hebben gehad.

I also wish to thank all colleagues at the Biostatistics department, for the moments of fun between our hard work and because you would always listen to my complaints if things were not going my way. I really needed your support. Verder bedankt aan de medisch onderzoekers met wie ik heb samengewerkt en

jullie geduld waarmee jullie me de verschillende medische begrippen hebben uitgelegd.

Marek Molas, volgens mij heb ik met jou de langste tijd een kamer gedeeld. Bedankt dat je vaak met me mee wou denken als ik niet direct uit een wiskundig probleem kwam. Ik denk nog vaak aan je.

Ook wil ik natuurlijk mijn paranimfen bedanken. Nicole Erler bedankt dat je (vooral op vrijdag) nog naar mijn verhalen wilde luisteren ook al maakte ik het je soms lastig om alles te volgen. Bas Sala bedankt voor je vriendschap door de jaren heen.

Nano Suwarno, bedankt dat je me hebt geholpen met alle ICT-problemen. Eline van Gent, bedankt voor je ondersteuning. Bij de alle administratieve zaken waar ik in de laatste fase van dit proefschrift mee te maken kreeg, was je hulp onmisbaar. Tina Siehoff, bedankt voor je hulp bij het maken van het omslag.

Ik eindig met het bedanken van mijn familie. Pa, ma, nu denken jullie vast te weten waar ik al die tijd mee bezig ben geweest. Ik ben bang dat dat niet helemaal klopt (daarom duurde het zo lang). Maryna bedankt voor de steun en het geduld met mij als ik veel te lange dagen moest maken. Servaas, je bent nog veel te klein om te weten waar ik mee bezig was. Een kind maakt het schrijven van een proefschrift nog net wat lastiger, maar je bent een verrijking van mijn leven. Ik ben blij dat ik af en toe een pauze kon nemen van het programmeren en het schrijven om met jou te spelen.

About the Author

Sten Paul Willemsen was born on April 25th, 1975 in Rotterdam. In 1993 he obtained his VWO diploma in Brielle. He then started to study econometrics at the Erasmus University in Rotterdam. During this time he did internships at the NEI (Netherlands Economic Institute) and the CPB (Netherlands Bureau for Economic Policy Analysis). In 1999 he obtained his Masters degree in Econometrics.

He then started to work at the IOO (Institute of research on public spending). In 2003 he started to work at Hewitt Associates as an actuarial consultant. Sten started as a statistical consultant and researcher at the Erasmus MC in 2006. Initially he was partially employed by the Department of General Practice but later he switched full time to the Department of Biostatistics. Here he started the work described in this thesis under guidance of prof. Emmanuel Lesaffre. Besides this, he is involved in many consulting projects (both internally as for external clients), is a member of the works council of the health sciences theme and is teaching various courses.

Sten is married to Maryna Burko and together they have a son, Servaas.

Portfolio

Publications

Scientific publications

- Dijk, M. R. V., N. V. Borggreven, S. P. Willemsen, A. H. J. Koning, R. P. M. Steegers-Theunissen, and M. P. H. Koster. Maternal lifestyle impairs embryonic growth: The rotterdam periconception cohort. *Reproductive Sciences*, 0(0), 1933719117728801, 2017. PMID: 28884629.

- Geelen, I. G., N. Thielen, J. J. Janssen, M. Hoogendoorn, T. J. Roosma, S. P. Willemsen, P. J. Valk, O. Visser, J. J. Cornelissen, and P. E. Westerweel. Impact of hospital experience on the quality of tyrosine kinase inhibitor response monitoring and consequence for chronic myeloid leukemia patient survival. *Haematologica*, 2017.

- Geelen, I. G., N. Thielen, J. J. Janssen, M. Hoogendoorn, T. J. Roosma, S. P. Willemsen, O. Visser, J. J. Cornelissen, and P. E. Westerweel. Treatment outcome in a population-based, 'real-world' cohort of patients with chronic myeloid leukemia. *Haematologica*, 102(11), 1842–1849, 2017.

- Herzog, E. M., A. J. Eggink, S. P. Willemsen, R. C. Slieker, K. P. Wijnands, J. F. Felix, J. Chen, A. Stubbs, P. J. van der Spek, J. B. van Meurs, and R. P. Steegers-Theunissen. Early- and late-onset preeclampsia and the tissue-specific epigenome of the placenta and newborn. *Placenta*, 58(Supplement C), 122 – 132, 2017.

- Huijgen, N. A., Y. V. Louwers, S. P. Willemsen, J. H. M. De Vries, R. P. M. Steegers-Theunissen, and J. S. E. Laven. Dietary patterns and the phenotype of polycystic ovary syndrome: the chance of ongoing pregnancy. *Reproductive biomedicine online*, 34, 668–676, 2017.

- Koning, I. V., J. Dudink, I. A. L. Groenenberg, S. P. Willemsen, I. K. M. Reiss, and R. P. M. Steegers-Theunissen. Prenatal cerebellar growth trajectories and the impact of periconceptional maternal and fetal factors. *Human reproduction (Oxford, England)*, 32, 1230–1237, 2017.

- Koning, I. V., J. A. Roelants, I. A. L. Groenenberg, M. J. Vermeulen, S. P. Willemsen, I. K. M. Reiss, P. P. Govaert, R. P. M. Steegers-Theunissen, and J. Dudink. New ultrasound measurements to bridge the gap between prenatal and neonatal brain growth assessment. *AJNR American journal of neuroradiology*, 2017.

- Koning, I. V., A. W. Van Graafeiland, I. A. L. Groenenberg, S. C. Husen, A. T. J. I. Go, J. Dudink, S. P. Willemsen, J. M. J. Cornette, and R. P. M. Steegers-Theunissen. Prenatal influence of congenital heart defects on trajectories of cortical folding of the fetal brain using three-dimensional ultrasound. *Prenatal diagnosis*, 2017.

- Parisi, F., M. Rousian, N. A. Huijgen, A. H. J. Koning, S. P. Willemsen, J. H. M. De Vries, I. Cetin, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. Periconceptional maternal 'high fish and olive oil, low meat' dietary pattern is associated with increased embryonic growth: The rotterdam periconceptional cohort (predict study). *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 2017.

- Parisi, F., M. Rousian, A. H. J. Koning, S. P. Willemsen, I. Cetin, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. Periconceptional maternal biomarkers of one-carbon metabolism and embryonic growth trajectories: the rotterdam periconceptional cohort (predict study). *Fertility and sterility*, 107, 691–698.e1, 2017.

- Parisi, F., M. Rousian, A. H. J. Koning, S. P. Willemsen, I. Cetin, and R. P. M. Steegers-Theunissen. Periconceptional maternal one-carbon biomarkers are associated with embryonic development according to the carnegie stages. *Human reproduction (Oxford, England)*, 32, 523–530, 2017.

- Roelants, J. A., M. J. Vermeulen, I. V. Koning, I. A. L. Groenenberg, S. P. Willemsen, A. C. S. Hokken-Koelega, K. F. M. Joosten, I. K. M. Reiss, and R. P. M. Steegers-Theunissen. Foetal fractional thigh volume: an early 3d ultrasound marker of neonatal adiposity. *Pediatric obesity*, 2017.

- Stocker, M., W. Van Herk, S. El Helou, S. Dutta, M. S. Fontana, F. A. B. A. Schuerman, R. K. Van den Tooren-De Groot, J. W. Wieringa, J. Janota, L. H. Van der Meer-Kappelle, R. Moonen, S. D. Sie, E. De Vries, A. E. Donker, U. Zimmerman, L. J. Schlapbach, A. C. De Mol, A. Hoffman-Haringsma, M. Roy, M. Tomaske, R. F. Kornelisse, J. Van Gijsel, E. G. Visser, S. P. Willemsen, A. M. C. Van Rossum, and N. S. Group. Procalcitonin-guided decision making for duration of antibiotic therapy in neonates with suspected early-onset sepsis: a multicentre, randomised controlled trial (neopins). *Lancet (London, England)*, 2017.
- Van den Berg, C. B., I. Chaves, E. M. Herzog, S. P. Willemsen, G. T. J. Van der Horst, and R. P. M. Steegers-Theunissen. Early- and late-onset preeclampsia and the dna methylation of circadian clock and clock-controlled genes in placental and newborn tissues. *Chronobiology international*, 1–12, 2017.
- Van Dijk, M. R., M. P. H. Koster, S. P. Willemsen, N. A. Huijgen, J. S. E. Laven, and R. P. M. Steegers-Theunissen. Healthy preconception nutrition and lifestyle using personalized mobile health coaching is associated with enhanced pregnancy chance. *Reproductive biomedicine online*, 2017.
- Van Dijk, M. R., E. C. Oostingh, M. P. H. Koster, S. P. Willemsen, J. S. E. Laven, and R. P. M. Steegers-Theunissen. The use of the mhealth program smarter pregnancy in preconception care: rationale, study design and data collection of a randomized controlled trial. *BMC pregnancy and childbirth*, 17, 46, 2017.
- Van Vugt, J. L. A., S. Levolger, A. Gharbharan, M. Koek, W. J. Niessen, J. W. A. Burger, S. P. Willemsen, R. W. F. De Bruin, and J. N. M. IJzermans. A comparative study of software programmes for cross-sectional skeletal muscle and adipose tissue measurements on abdominal computed tomography scans of rectal cancer patients. *Journal of cachexia, sarcopenia and muscle*, 8, 285–297, 2017.
- Willemsen, S. P., C. M. Russo, E. Lesaffre, and D. Leão. Flexible multivariate nonlinear models for bioequivalence problems. *Statistical Modelling*, 1471082X17706018, 2017.
- Doebar, S. C., C. De Monyé, H. Stoop, J. Rothbarth, S. P. Willemsen, and C. H. M. Van Deurzen. Ductal carcinoma in situ diagnosed by breast needle biopsy: Predictors of invasion in the excision specimen. *Breast (Edinburgh, Scotland)*, 27, 15–21, 2016.

- Herzog, E. M., A. J. Eggink, M. Van der Zee, J. Lagendijk, S. P. Willemsen, R. De Jonge, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. The impact of early- and late-onset preeclampsia on umbilical cord blood cell populations. *Journal of reproductive immunology*, 116, 81–85, 2016.

- Koning, I. V., L. Baken, I. A. L. Groenenberg, S. C. Husen, J. Dudink, S. P. Willemsen, M. Gijtenbeek, A. H. J. Koning, I. K. M. Reiss, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. Growth trajectories of the human embryonic head and periconceptional maternal conditions. *Human reproduction (Oxford, England)*, 31, 968–976, 2016.

- Rinkel, W. D., M. H. Aziz, M. J. M. Van Deelen, S. P. Willemsen, M. Castro Cabezas, J. W. Van Neck, and J. H. Coert. Normative data for cutaneous threshold and spatial discrimination in the feet. *Muscle & nerve*, 2016.

- Roelants, J. A., I. V. Koning, M. M. A. Raets, S. P. Willemsen, M. H. Lequin, R. P. M. Steegers-Theunissen, I. K. M. Reiss, M. J. Vermeulen, P. Govaert, and J. Dudink. A new ultrasound marker for bedside monitoring of preterm brain growth. *AJNR American journal of neuroradiology*, 37, 1516–1522, 2016.

- Van Dijk, M. R., N. A. Huijgen, S. P. Willemsen, J. S. Laven, E. A. Steegers, and R. P. Steegers-Theunissen. Impact of an mhealth platform for pregnancy on nutrition and lifestyle of the reproductive population: A survey. *JMIR mHealth and uHealth*, 4, e53, 2016.

- Huijgen, N. A., J. S. E. Laven, C. T. Labee, Y. V. Louwers, S. P. Willemsen, and R. P. M. Steegers-Theunissen. Are dieting and dietary inadequacy a second hit in the association with polycystic ovary syndrome severity? *PloS one*, 10, e0142772, 2015.

- Koning, I. V., I. A. L. Groenenberg, A. W. Gotink, S. P. Willemsen, M. Gijtenbeek, J. Dudink, A. T. J. I. Go, I. K. M. Reiss, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. Periconception maternal folate status and human embryonic cerebellum growth trajectories: The rotterdam predict study. *PloS one*, 10, e0141089, 2015.

- Van Oppenraaij, R. H. F., P. H. C. Eilers, S. P. Willemsen, F. M. Van Dunné, N. Exalto, and E. A. P. Steegers. Determinants of number-specific recall error of last menstrual period: a retrospective cohort study. *BJOG : an international journal of obstetrics and gynaecology*, 122, 835–841, 2015.

- Verkleij, S. P. J., P. A. J. Luijsterburg, S. P. Willemsen, B. W. Koes, A. M. Bohnen, and S. M. A. Bierma-Zeinstra. Effectiveness of diclofenac versus paracetamol in knee osteoarthritis: a randomised controlled trial in primary care. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 65, e530–e537, 2015.
- Willemsen, S. P., P. H. C. Eilers, R. P. M. Steegers-Theunissen, and E. Lesaffre. A multivariate bayesian model for embryonic growth. *Statistics in medicine*, 34, 1351–1365, 2015.
- Zwiers, A. J. M., S. N. De Wildt, Y. B. De Rijke, S. P. Willemsen, N. S. Abdullahi, D. Tibboel, and K. Cransberg. Reference intervals for renal injury biomarkers neutrophil gelatinase-associated lipocalin and kidney injury molecule-1 in young infants. *Clinical chemistry and laboratory medicine*, 53, 1279–1289, 2015.
- Baken, L., B. Benoit, A. H. J. Koning, S. P. Willemsen, P. J. Van der Spek, R. P. M. Steegers-Theunissen, E. A. P. Steegers, and N. Exalto. First-trimester hand measurements in euploid and aneuploid human fetuses using virtual reality. *Prenatal diagnosis*, 34, 961–969, 2014.
- Eindhoven, S. C., E. M. Van Uitert, J. S. E. Laven, S. P. Willemsen, A. H. J. Koning, P. H. C. Eilers, N. Exalto, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. The influence of ivf/icsi treatment on human embryonic growth trajectories. *Human reproduction (Oxford, England)*, 29, 2628–2636, 2014.
- Gijtenbeek, M., H. Bogers, I. A. L. Groenenberg, N. Exalto, S. P. Willemsen, E. A. P. Steegers, P. H. C. Eilers, and R. P. M. Steegers-Theunissen. First trimester size charts of embryonic brain structures. *Human reproduction (Oxford, England)*, 29, 201–207, 2014.
- Van Uitert, E. M., S. Van Ginkel, S. P. Willemsen, J. Lindemans, A. H. J. Koning, P. H. C. Eilers, N. Exalto, J. S. E. Laven, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. An optimal periconception maternal folate status for embryonic size: the rotterdam predict study. *BJOG : an international journal of obstetrics and gynaecology*, 121, 821–829, 2014.
- Verwoerd, A. J. H., W. C. Peul, S. P. Willemsen, B. W. Koes, C. L. A. M. Vleggeert-Lankamp, A. el Barzouhi, P. A. J. Luijsterburg, and A. P. Verhagen. Diagnostic accuracy of history taking to assess lumbosacral nerve root

compression. *The spine journal : official journal of the North American Spine Society*, 14, 2028–2037, 2014.

- Vucic, S., E. De Vries, P. H. C. Eilers, S. P. Willemsen, M. A. R. Kuijpers, B. Prahl-Andersen, V. W. V. Jaddoe, A. Hofman, E. B. Wolvius, and E. M. Ongkosuwito. Secular trend of dental development in dutch children. *American journal of physical anthropology*, 155, 91–98, 2014.

- Willemsen, S. P., M. De Ridder, P. H. C. Eilers, A. Hokken-Koelega, and E. Lesaffre. Modeling height for children born small for gestational age treated with growth hormone. *Statistical methods in medical research*, 23, 333–345, 2014.

- Zwiers, A. J. M., K. Cransberg, Y. B. De Rijke, S. P. Willemsen, A. C. De Mol, D. Tibboel, and S. N. De Wildt. Reference ranges for serum $beta$-trace protein in neonates and children younger than 1 year of age. *Clinical chemistry and laboratory medicine*, 52, 1815–1821, 2014.

- Bondt, A., M. H. Selman, A. M. Deelder, J. M. Hazes, S. P. Willemsen, M. Wuhrer, and R. J. Dolhain. Association between galactosylation of immunoglobulin g and improvement of rheumatoid arthritis during pregnancy is independent of sialylation. *Journal of proteome research*, 12(10), 4522–4531, 2013.

- Kiefte-De Jong, J. C., V. W. V. Jaddoe, A. G. Uitterlinden, E. A. P. Steegers, S. P. Willemsen, A. Hofman, H. Hooijkaas, and H. A. Moll. Levels of antibodies against tissue transglutaminase during pregnancy are associated with reduced fetal weight and birth weight. *Gastroenterology*, 144, 726–735.e2, 2013.

- Reus, A. D., H. El-Harbachi, M. Rousian, S. P. Willemsen, R. P. M. Steegers-Theunissen, E. A. P. Steegers, and N. Exalto. Early first-trimester trophoblast volume in pregnancies that result in live birth or miscarriage. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 42, 577–584, 2013.

- Van Uitert, E. M., N. Exalto, G. J. Burton, S. P. Willemsen, A. H. J. Koning, P. H. C. Eilers, J. S. E. Laven, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. Human embryonic growth trajectories and associations with fetal growth and birthweight. *Human reproduction (Oxford, England)*, 28, 1753–1761, 2013.

- Van Uitert, E. M., N. Van der Elst-Otte, J. J. Wilbers, N. Exalto, S. P. Willemsen, P. H. C. Eilers, A. H. J. Koning, E. A. P. Steegers, and R. P. M. Steegers-Theunissen. Periconception maternal characteristics and embryonic growth trajectories: the rotterdam predict study. *Human reproduction (Oxford, England)*, 28, 3188–3196, 2013.

- Bales, D., N. Van Beek, M. Smits, S. P. Willemsen, J. J. Busschbach, R. Verheul, and H. Andrea. Treatment outcome of 18-month, day hospital mentalization-based treatment (mbt) in patients with severe borderline personality disorder in the netherlands. *Journal of personality disorders*, 26(4), 568–582, 2012.

- Quax, R. A. M., Y. A. De Man, J. W. Koper, E. F. C. Van Rossum, S. P. Willemsen, S. W. J. Lamberts, J. M. W. Hazes, R. J. E. M. Dolhain, and R. A. Feelders. Glucocorticoid receptor gene polymorphisms and disease activity during pregnancy and the postpartum period in rheumatoid arthritis. *Arthritis research & therapy*, 14, R183, 2012.

- Sonnenschein-Van der Voort, A. M., V. W. Jaddoe, R. J. Van der Valk, S. P. Willemsen, A. Hofman, H. A. Moll, J. C. De Jongste, and L. Duijts. Duration and exclusiveness of breastfeeding and childhood asthma-related symptoms. *European Respiratory Journal*, 39(1), 81–89, 2012.

- Sonnenschein-Van der Voort, A. M. M., V. W. V. Jaddoe, R. J. P. Van der Valk, S. P. Willemsen, A. Hofman, H. A. Moll, J. C. De Jongste, and L. Duijts. Duration and exclusiveness of breastfeeding and childhood asthma-related symptoms. *The European respiratory journal*, 39, 81–89, 2012.

- Van den Hooven, E. H., F. H. Pierik, Y. De Kluizenaar, S. P. Willemsen, A. Hofman, S. W. Van Ratingen, P. Y. Zandveld, J. P. Mackenbach, E. A. Steegers, H. M. Miedema, et al. Air pollution exposure during pregnancy, ultrasound measures of fetal growth, and adverse birth outcomes: a prospective cohort study. *Environmental health perspectives*, 120(1), 150, 2012.

- Van der Valk, R. J. P., L. Duijts, M. Kerkhof, S. P. Willemsen, A. Hofman, H. A. Moll, H. A. Smit, B. Brunekreef, D. S. Postma, V. W. V. Jaddoe, G. H. Koppelman, and J. C. De Jongste. Interaction of a 17q12 variant with both fetal and infant smoke exposure in the development of childhood asthma-like symptoms. *Allergy*, 67, 767–774, 2012.

- Brinks, A., R. M. Van Rijn, S. P. Willemsen, A. M. Bohnen, J. A. Verhaar, B. W. Koes, and S. M. Bierma-Zeinstra. Corticosteroid injections for greater trochanteric pain syndrome: a randomized controlled trial in primary care. *The Annals of Family Medicine*, 9(3), 226–234, 2011.

- Durmuş, B., C. J. Kruithof, M. H. Gillman, S. P. Willemsen, A. Hofman, H. Raat, P. H. Eilers, E. A. Steegers, and V. W. Jaddoe. Parental smoking during pregnancy, early growth, and risk of obesity in preschool children: the generation r study. *The American journal of clinical nutrition*, 94(1), 164–171, 2011.

- Gaillard, R., R. Bakker, S. P. Willemsen, A. Hofman, E. A. Steegers, and V. W. Jaddoe. Blood pressure tracking during pregnancy and the risk of gestational hypertensive disorders: the generation r study. *European heart journal*, 32(24), 3088–3097, 2011.

- Heppe, D. H., R. M. Van Dam, S. P. Willemsen, H. Den Breeijen, H. Raat, A. Hofman, E. A. Steegers, and V. W. Jaddoe. Maternal milk consumption, fetal growth, and the risks of neonatal complications: the generation r study. *The American journal of clinical nutrition*, 94(2), 501–509, 2011.

- Molenaar, H. M., R. W. Selles, S. P. Willemsen, S. E. R. Hovius, and H. J. Stam. Growth diagrams for individual finger strength in children measured with the rihm. *Clinical orthopaedics and related research*, 469, 868–876, 2011.

- Rieken, R., J. B. Van Goudoever, H. Schierbeek, S. P. Willemsen, E. A. Calis, D. Tibboel, H. M. Evenhuis, and C. Penning. Measuring body composition and energy expenditure in children with severe neurologic impairment and intellectual disability. *The American journal of clinical nutrition*, 94(3), 759–766, 2011.

- Uijen, J. H. J. M., J. C. Van der Wouden, F. G. Schellevis, S. P. Willemsen, L. W. Van Suijlekom-Smit, and P. J. E. Bindels. Asthma prescription patterns for children: can gps do better? *The European journal of general practice*, 17, 109–115, 2011.

- Van de Geijn, F. E., Y. A. De Man, M. Wuhrer, S. P. Willemsen, A. M. Deelder, J. M. W. Hazes, and R. J. E. M. Dolhain. Mannose-binding lectin does not explain the course and outcome of pregnancy in rheumatoid arthritis. *Arthritis research & therapy*, 13, R10, 2011.

- De Winter, C. F., K. W. Magilsen, J. C. Van Alfen, S. P. Willemsen, and H. M. Evenhuis. Metabolic syndrome in 25% of older people with intellectual disability. *Family Practice*, 28(2), 141–144, 2010.
- Jansen, T. C., J. Van Bommel, F. J. Schoonderbeek, S. J. Sleeswijk Visser, J. M. Van der Klooster, A. P. Lima, S. P. Willemsen, and J. Bakker. Early lactate-guided therapy in intensive care unit patients: a multicenter, open-label, randomized controlled trial. *American journal of respiratory and critical care medicine*, 182(6), 752–761, 2010.
- Selles, R. W., J. M. Zuidam, S. P. Willemsen, H. J. Stam, S. E. Hovius, et al. Growth diagrams for grip strength in children. *Clinical Orthopaedics and Related Research®*, 468(1), 217, 2010.
- Van der Cammen, M. H., H. IJsselstijn, T. Takken, S. P. Willemsen, D. Tibboel, H. J. Stam, R. J. Van den Berg-Emons, et al. Exercise testing of pre-school children using the bruce treadmill protocol: new reference values. *European journal of applied physiology*, 108(2), 393, 2010.
- Veugelers, R., M. A. Benninga, E. A. Calis, S. P. Willemsen, H. Evenhuis, D. Tibboel, and C. Penning. Prevalence and clinical presentation of constipation in children with severe generalized cerebral palsy. *Developmental Medicine & Child Neurology*, 52(9), 2010.
- De Man, Y. A., J. M. Hazes, H. Van der Heide, S. P. Willemsen, C. J. De Groot, E. A. Steegers, and R. J. Dolhain. Association of higher rheumatoid arthritis disease activity during pregnancy with lower birth weight: results of a national prospective study. *Arthritis & Rheumatology*, 60(11), 3196–3206, 2009.
- Logghe, I. H., P. E. Zeeuwe, A. P. Verhagen, R. M. Wijnen-Sponselee, S. P. Willemsen, S. Bierma-Zeinstra, E. Van Rossum, M. J. Faber, and B. W. Koes. Lack of effect of tai chi chuan in preventing falls in elderly people living at home: a randomized clinical trial. *Journal of the American Geriatrics Society*, 57(1), 70–75, 2009.
- Rozendaal, R., E. Uitterlinden, G. Van Osch, E. Garling, S. P. Willemsen, A. Ginai, J. Verhaar, H. Weinans, B. Koes, and S. Bierma-Zeinstra. Effect of glucosamine sulphate on joint space narrowing, pain and function in patients with hip osteoarthritis; subgroup analyses of a randomized controlled trial. *Osteoarthritis and Cartilage*, 17(4), 427–432, 2009.

- Van De Geijn, F. E., M. Wuhrer, M. H. Selman, S. P. Willemsen, Y. A. De Man, A. M. Deelder, J. M. Hazes, and R. J. Dolhain. Immunoglobulin g galactosylation and sialylation are associated with pregnancy-induced improvement of rheumatoid arthritis and the postpartum flare: results from a large prospective cohort study. *Arthritis research & therapy*, 11(6), R193, 2009.

- Van Linschoten, R., M. Van Middelkoop, M. Y. Berger, E. M. Heintjes, J. A. Verhaar, S. P. Willemsen, B. W. Koes, and S. M. Bierma-Zeinstra. Supervised exercise therapy versus usual care for patellofemoral pain syndrome: an open label randomised controlled trial. *Bmj*, 339, b4074, 2009.

- De Man, Y. A., R. J. Dolhain, F. E. Van De Geijn, S. P. Willemsen, and J. M. Hazes. Disease activity of rheumatoid arthritis during pregnancy: results from a nationwide prospective study. *Arthritis Care & Research*, 59(9), 1241–1248, 2008.

- Rozendaal, R. M., B. W. Koes, G. J. Van Osch, E. J. Uitterlinden, E. H. Garling, S. P. Willemsen, A. Z. Ginai, J. A. Verhaar, H. Weinans, and S. M. Bierma-Zeinstra. Effect of glucosamine sulfate on hip osteoarthritis. *Ann Intern Med*, 148, 268–277, 2008.

- Verghese, E., M. Den Bakker, A. Campbell, A. Hussein, A. Nicholson, A. Rice, B. Corrin, D. Rassl, G. Langman, S. Willemsen, et al. Interobserver variation in the classification of thymic tumours–a multicentre study using the who classification system. *Histopathology*, 53(2), 218–223, 2008.

Non-Scientific publications

- Willemsen, S. P. De kunst van variabelen kiezen. *Wetenschap in het ASz*, 8(2), 45, 2016.

- Willemsen, S. P. Bepaling van de steekproefgrootte. *Wetenschap in het ASz*, 6(2), 13, 2014.

Conferences

- IBS Channel network meeting, Hasselt Belgium (2017, oral presentation)

- The 37th Annual Conference of the International Society for Clinical Biostatistics - Birmingham, UK (2016, oral presentation)
- The 29th Workshop on Statistical Modelling, Göttingen, Germany (2014, oral presentation)
- The 28th Workshop on Statistical Modelling, Palermo, Italy (2013, oral presentation)
- The 4th Bayes Conference - Rotterdam, the Netherlands (2013, oral presentation)
- The 33e Annual Conference of the International Society for Clinical Biostatistics - Bergen, Norway (2012, oral presentation)
- The 31st Annual Conference of the International Society for Clinical Biostatistics - Montpellier, France (2010, poster presentation)
- The 28th Annual Conference of the International Society for Clinical Biostatistics - Alexandroupolis, Greece (2007)
- 36th Annual Conference of the International Society for Clinical Biostatistics (ISCB) - Utrecht, Netherlands (2015)
- Variable selection: Spring symposium - Rotterdam, the Netherlands, (2012)
- IBS Channel network meeting, Kerkrade - the Netherlands (2007)

Courses

- GAMLSS in Action, Erasmus MC, Rotterdam, the Netherlands, 2014.
- An Introduction to the Joint Modeling of Longitudinal and Survival Outcomes, Erasmus MC, Rotterdam, the Netherlands, 2013.
- Bayesian inferences for latent Gaussian models using INLA, Utrecht, the Netherlands, 2013.
- Bayesian variable selection, Erasmus MC, Rotterdam, the Netherlands, 2012.
- Bayesian adaptive methods for clinical trials, Erasmus MC, Rotterdam, the Netherlands, 2012.
- Missing data in longitudinal studies: strategies for Bayesian modelling, sensitivity analysis, and causal inference, Rotterdam, the Netherlands, 2011

- Bayesian Methods and Bias Analysis, Rotterdam, the Netherlands, 2010
- Health economic evaluation, Cambridge, UK, 2010
- Short course on Bayesian data analysis, Leuven, Belgie, 2010
- Frailty models: Multivariate survival analysis with applications in medicine, Rotterdam, the Netherlands, 2009
- A Course on Latex, Rotterdam, the Netherlands 2009
- Introduction to R, Rotterdam, the Netherlands, 2008
- Models for Longitudinal and Incomplete Data, Rotterdam, the Netherlands, 2008
- Statistical methods for the analysis of Recurrent Events, Parijs, Frankrijk, 2007
- Multistate model and models for competing risks, Rotterdam, the Netherlands, 2007
- Didactische vaardigheden, Rotterdam, the Netherlands, 2007
- Incomplete Data in Clinical Studies, Rotterdam, the Netherlands, 2007
- Spacial Epidemiology, Rotterdam, the Netherlands, 2006
- Topics in Meta-analysis, Rotterdam, the Netherlands, 2006

Teaching

- Review of Mathematics and Introduction to Statistics, NIHES program, Rotterdam, the Netherlands (2017)
- Linear mixed models, Rotterdam, the Netherlands, (2017)
- Logistic regression (part of Biostatistical Methods II: Classical Regression Models), Rotterdam, the Netherlands (2016)
- Introduction to statistics and SPSS, Dordrecht, the Netherlands (2016, 2017)
- Introduction to R, Rotterdam, the Netherlands, 2015
- Introduction to SAS (part of the NIHES Courses for the quantitative researcher), Rotterdam, the Netherlands, 2009-2014

- Repeated measurements, NIHES program, Rotterdam, the Netherlands (assisting, various years)
- Bayesian statistics, NIHES program, Rotterdam, the Netherlands (assisting, various years)
- Classical methods, NIHES program, Rotterdam, the Netherlands (assisting, various years)
- Modern methods, NIHES program, Rotterdam, the Netherlands (assisting, various years)
- Regression analysis, NIHES summer program, Rotterdam, the Netherlands (assisting, various years)
- Survival analysis, NIHES summer program, Rotterdam, the Netherlands (assisting, various years)
- Methods of clinical and epidemiological research (Thema 4.3) (assisting, various years)
- P-values and confidence intervals, Rotterdam, the Netherlands (assisting, various years)