Prediction of Visual Prognosis to Optimize Frequency of Perimetric Testing in Glaucoma

Susan R. Bryan

2017

Prediction of Visual Prognosis to Optimize Frequency of Perimetric Testing in Glaucoma

ISBN: 978-94-028-0772-1

Cover design: Eoghan Ruadh MacManus
Printing: Ipskamp Printing

Prediction of Visual Prognosis to Optimize Frequency of Perimetric Testing in Glaucoma

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

ingevolge het besluit van het College voor Promoties

in het openbaar te verdedigen op

woensday 04 Oktober 2017 om 9.30 uur

door

Susan Ruth Bryan geboren op 30 March 1987

Promotiecomissie

Promotoren:

Prof.dr. E.M.E.H. Lesaffre
Prof.dr.ing. P.H.C. Eilers

Co-promotoren:

Dr. H.G. Lemij
Dr. K.A. Vermeer

Overige leden:

Prof.dr. D. Garway-Heath
Prof.dr. E. Steyerberg
Prof.dr. H. Vingerling

To my family,
for their never-ending support

Contents

CHAPTER 1

General Introduction

1.1 Introduction

In this chapter, we provide introductory information on glaucoma and the statistical methods used to analyse visual field (VF) data. Both clinical and statistical parts are described in a non-technical way to help the reader understand the basic concepts related to our research. Furthermore, we introduce the motivating data set and explain the aims of this thesis.

1.2 The Human Eye

The human eye is an important organ which is responsible for vision. As seen in Figure 1.1, the anterior part of the eye consists of the iris, the cornea, the pupil, the lens and the anterior chamber. The retina is part of the posterior part of the eye (shown in Figure 1.2). The cornea and the lens focus the light into an image onto the retina. Visual impulses are then transmitted from the retina to the brain via the optic nerve, which is made up of retinal nerve cells and their fibres. The optic nerve head is the point where the ganglion cell axons exit the eye. Since there are no rods or cones in the optic nerve head, this corresponds to a small physiological blind spot in each eye. If any of the structures in the optical pathway are blocked or the eye does not focus the light correctly, visual function will be disturbed. There are several diseases which can affect visual function, such as diabetic retinopathy, cataract or glaucoma.

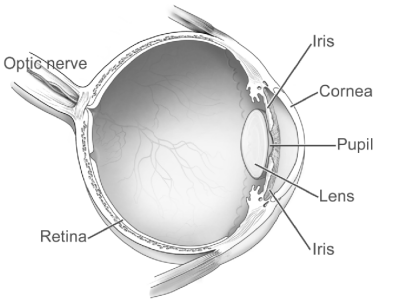

Figure 1.1: Cross-section of the eye. Image courtesy of the National Eye Institute, National Institutes of Health.

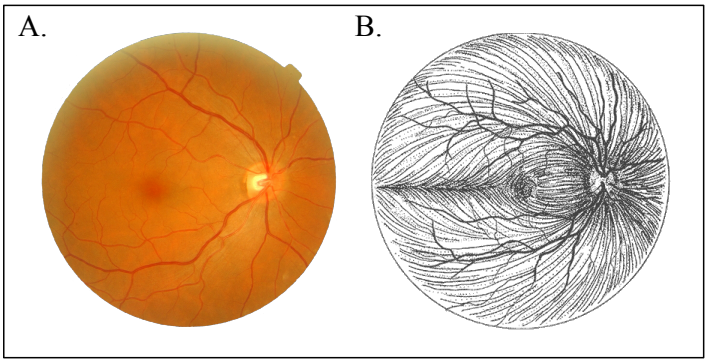

Figure 1.2: (A) Fundus photo of retina and (B) drawing of nerve fibres.

1.3 Glaucoma

The term "glaucoma" covers a wide range of diseases. According to the World Health Organisation (WHO), glaucoma is one of the leading causes of irreversible blindness in the world (Kingman, 2004; Quigley, 2006). In practice, glaucoma applies to all patients with increased intraocular pressure (IOP) and/or glaucomatous damage (Flammer and Meier, 2003). In glaucoma, the nerve cells and nerve fibres progressively die, gradually severing the connection between the eye and the brain. Since the rods and cones are still functioning, the eye is able to detect the light, but the transmission of visual information to the brain is interrupted. One of the most common types of glaucoma is primary open-angle glaucoma (POAG). POAG is a chronic disease, tends to be inherited and may not show up until later in life. There is no cure for it at present, but the disease can be slowed or arrested by treatment. An increased IOP is an important risk factor for glaucoma, which can easily be monitored and treated. However, because most people with glaucoma have no early symptoms or pain from this increased pressure, it is often only detected once long-term visual loss has occurred. Detecting glaucoma at an early stage is important in order to keep the damage as small as possible, with little or no impact on the patient's quality of life. Both structural and functional tests are routinely used in clinical practice for diagnosis and monitoring of glaucoma. Structural tests, provided by several imaging modalities, are objective, fast and reliable. However, measurements obtained from these tests are largely influenced by the variability of optic disc size and the number of nerve fibers among individuals. At present, structural measurements can be used to diagnose early glaucoma, but cannot be used to exclude it. Functional tests, such as perimetry, represent the most important clinical endpoint.

However, perimetry is a subjective test and the results are prone to variability. Knowing the sources of the variability might enable reducing it, and this may improve progression detection in glaucoma. With less variability, smaller changes can be picked up and fewer tests are needed. Although both tests are important, in this thesis we focus on the testing of visual function using perimetry.

1.4 Visual Field Testing

Currently, the Humphrey Field Analyzer (HFA) (Carl Zeiss Meditec, Dublin, CA) is one of the standards of care for glaucoma diagnosis and management. An illustration of the HFA in use is shown in Figure 1.3. The patient sits in front of the device with a fixation target in the center. The patient's chin is rested and the eye that is not being tested is covered. Differential light stimuli are then presented at predefined locations. The patient is required to press a button whenever a light is seen. The light intensity and location are varied. This maps and quantifies visual function. Using one of the default settings, the 24-2 program, this results in a print-out of the VF with sensitivity estimates describing the threshold at each of the 52 locations. The maximum luminance of the Humphrey Field Analyzer perimeter's stimuli is 10,000 abs, which is defined as 0 dB retinal sensitivity. The lowest sensitivity that can be detected by this perimeter is therefore 0 dB, although lower sensitivities could in fact occur if it were not for the limitations of this device. The highest sensitivity that can be detected is 50 dB (0.1 abs). Few humans are however capable of seeing a stimulus less than 40 dB, which is 1/10,000 of the maximum (or 1 abs) (Anderson and Patella, 1999). Thus, for practical purposes, the useful intensity range for white light testing is from 0 to 40 dB.

Figure 1.3: Illustration of the HFA in use.

A common approach to reduce measurement variability is to average multiple measurements. Mean deviation (MD) expresses the overall reduction in sensitivity. This

is averaged across the VF and is relative to a group of healthy age-matched subjects. The visual field index (VFI) describes the visual function as a percentage of a perimetrically normal age-corrected VF. It is an eccentricity-weighted mean of all the scores at all 52 locations (excluding the 2 corresponding to the blind spot), measured from 0 % (blind) to 100 % (normal). Modeling of individual VF points is potentially of greater interest, because it provides additional information such as the spatial nature of the fields, that is otherwise lost in global parameters.

1.5 Glaucoma Study

The Glaucoma Study is a prospective cohort study conducted by the Rotterdam Eye Hospital in the Netherlands. This is an ongoing study which began in 1998. Inclusion criteria included glaucoma diagnosis and an age range of 18 to 85 years. In total, 139 patients, consisting of 80 (57.6%) men and 59 (42.4%) women, were recruited with a mean follow-up of 10.5 years. Follow-up data were collected at approximately 6-monthly intervals. All patients gave their written informed consent for participation. All research procedures followed the tenets set forth in the Declaration of Helsinki. Furthermore, all of the data that was used in this analysis has been made available online at http://rod-rep.com.

1.6 Statistical Approach

Adequate treatment may slow down the disease, possibly even halting its progression. Evaluation of a longitudinal series of visual fields (VF), as measured by standard automated perimetry (SAP), provides a method to detect early evidence of glaucoma and to determine functional deterioration. Hence, treatment strategies can then be optimized to prevent further VF loss. Modelling the sensitivity estimates is beneficial for this evaluation. This helps to extract the systematic part of the data and distinguish it from noise, tries to determine the underlying VF progression from observed measurements, and can be used in the prediction of future VFs.

From a statistical modeling viewpoint the analysis of the VF data presents several challenges. First, VF testing involves measuring the level of differential light sensitivity at different locations. The sensitivity estimates are left-censored due to the limitation of the device used to measure them. Furthermore, the dataset in this study is clustered in nature. The hierarchical structure of the data includes 4 levels, namely, (1) the individual, (2) the eye (3) the hemifield and (4) the location. Correlation of VF measurements within the inferior and superior hemifields, separated by the horizontal meridian, are assumed to be higher than between hemifields due to the structure of the nerve fibres. Furthermore, observations belonging to the same eye are expected to be more alike than observations drawn from an independent eye. The same applies to observations from eyes belonging to the same individual. The easiest but crude way of dealing with the complexity is to either ignore the clustering or

simplify the data structure. For example, by using regression models treating each location as an independent sample. However, this might lead to loss of information and possibility of erroneous conclusions (Liang and Zeger, 1986; Molenberghs and Verbeke, 2005). Multilevel models have been introduced to deal with the phenomenon of data nesting (Leyland and Goldstein, 2001). Multilevel modeling is a generalization of regression methods whereby random coefficients are introduced for each level in the data.
Another difficulty in modeling VF data is the amount and type of measurement error in the sensitivity estimates. This may be due to measurable factors, such as season and time of day, or unknown transient factors, such as fatigue or lack of concentration. Although their magnitudes may vary, these factors are global and affect all locations belonging to the same eye at each visit. In addition, there is an inverse relationship between sensitivity and variability. For example, measurement error in the VFs increases with damage, and hence low sensitivity estimates have high variability. Therefore, it is naive to assume a constant variance over the wide range of sensitivity estimates.

1.7 Aims of this Thesis

The main aim of this thesis is to model this complex data structure in order to obtain better estimates of the true evolution of the sensitivity over time. In order to achieve this we need to deal with two main aspects of the data, namely, (1) time and (2) location. Our aim is to build up a model to incorporate the aforementioned aspects by using a hierarchical model, and to deal with the various sources of measurement error. In this way we can develop a method for modeling the progression of glaucoma. Furthermore, we aim to investigate and evaluate methods to improve the computational difficulties which arise for complex models using standard methods.

1.8 Outline of this Thesis

An introduction to the statistical methods used in this thesis are described in Chapter 2. Regression models are currently a standard method for modeling VF data. In Chapter 3, we show a comparison of different regression models for the fitting and prediction of progression in glaucomatous VFs. We apply classical approaches, as well as robust methods and models which take into account censoring. A problem with regression models is that each location is assumed to be independent. This ignores the correlation between measurements and does not allow borrowing information from the whole dataset. To account for the complex structure of the data, we apply advanced Bayesian hierarchical models in Chapter 4. Furthermore, we apply a recently proposed two-stage approach which allows us to speed up computational time considerably while still benefiting from the full Bayesian approach. In addition, we extend this approach to allow the calculation of model selection and model evaluation criteria. We investigate and compare other methods, such as integrated nested Laplace approximation

(INLA) and hierarchical likelihood (H-likelihood) to the classical one-stage and two-stage approaches in Chapter 5. For the clinical application, we propose to model factors which affect all locations belonging to the same VF as global visit effects (GVEs). Details and an evaluation of the GVE can be found in Chapter 6. Finally, in Chapter 7 we give the conclusions and highlight the major findings and limitations of our research. We also suggest future work related to this topic.

Bibliography

Anderson, D. R. and Patella, V. M. (1999). *Automated Static Perimetry*. Mosby, 2 edition.

Flammer, J. and Meier, E. (2003). *Glaucoma: A Guide for Patients : An Introduction for Care-Providers : A Quick Reference*. Hogrefe & Huber Publishers, 2nd edition.

Kingman, S. (2004). Glaucoma is second leading cause of blindness globally. *Bulletin of the World Health Organization*, 82(11):887–888.

Leyland, A. and Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*. John Wiley and Sons.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized estimating equation models. *Biometrika*, 73:13–22.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics.

Quigley, H.A., B. A. (2006). The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*, 73:262–267.

CHAPTER 2

Introduction to Statistical Methods

2.1 Mixed-Effects Models for Longitudinal Data

Longitudinal data are very common in applied sciences. Longitudinal studies involve repeated measurements of the same variable for each individual in the study over a period of time. Measurements from the same individual are likely to be more correlated than measurements from different individuals. In practice, longitudinal data are often unbalanced, meaning that individuals have an unequal number of measurements and/or that measurements were taken at different points. Due to their unbalanced nature, many longitudinal data sets cannot be analyzed by means of multivariate regression techniques (Verbeke and Molenberghs, 2000). A standard modeling framework for the analysis of longitudinal data, which deals with both the aforementioned issues, is the mixed-effects model. To introduce these models, we let y_{ij} denote the follow-up measurements for the i-th individual ($i = 1,\ldots,n$) at time t_{ij} where j denotes the specific time point $j = 1,\ldots,n_i$. The mixed-effects model can be written as

$$\begin{aligned} y_{ij} &= x_{ij}^{\top}\beta + z_{ij}^{\top}b_i + \varepsilon_{ij}, \\ b_i &\sim N(0,\Sigma_b), \\ \varepsilon_{ij} &\sim N(0,\sigma^2), \end{aligned}$$

where β denotes the vector with the regression coefficients of the design matrix for the fixed effects x_i and z_i denotes row vectors of the design matrix for the random effects b_i. In particular, the fixed and the random effects refer to the population-average and subject-specific effects, respectively. Furthermore, Σ_b is the covariance-variance matrix of the random effects, ε_{ij} are the error terms and σ^2 is the variance of the error. Multilevel models deal with nested data (Leyland and Goldstein, 2001). A mixed-effects model may also be viewed as a multilevel or hierarchical model in which the level-1 observations are nested within the higher level-2 observations. For example, in longitudinal studies repeated observations from an individual are nested within this individual.

2.2 The frequentist approach to statistical inference

In the frequentist approach, the unknown parameter θ is considered fixed and inference is done given observed data. A common way to estimate the model parameters is done using maximum likelihood estimation (MLE). MLE determines the parameter estimates which maximize the likelihood function, $L(\theta \mid y) = p(y \mid \theta)$. For simple models without random effects, the likelihood function can easily be written analytically. For mixed models containing random effects, the marginal likelihood is calculated by integrating over the random effects to obtain the MLEs.

2.3 The Bayesian approach to statistical inference

The main difference between the frequentist and the Bayesian approach is that the latter considers parameters as random variables that are characterized by a prior distribution. In the Bayesian approach, θ is assumed to be a random variable with density function $p(\theta)$. The prior distribution, which we denote as $p(\theta)$, represents the information about an uncertain parameter (Lesaffre and Lawson, 2012). Prior information may be obtained from expert knowledge, historical information, etc. In Bayesian inference, the prior distribution is combined with the probability distribution of the new data to yield the posterior distribution, given by:

$$p(\theta \mid y) = \frac{L(\theta \mid y)p(\theta)}{\int L(\theta \mid y)p(\theta)d\theta}.$$

For more extensive details on the methodology, as well as practical issues, of the Bayesian approach we refer to (Lesaffre and Lawson, 2012).

2.3.1 Bayesian Computational Techniques

For complex models, the posterior distribution often cannot be determined analytically. Calculating the integral by means of numerical integration methods is a practical alternative if only a few parameters are involved, but it becomes difficult for real-life applications where the dimensionality is often high (Lesaffre and Lawson, 2012). These problems can be avoided by using sampling techniques yielding Markov chains. A popular class of sampling algorithms is Markov chain Monte Carlo (MCMC) methods. The two most important MCMC procedures are the Gibbs sampler and the Metropolis(-Hastings) algorithm.

2.3.1.1 Gibbs sampler

The Gibbs sampler, introduced by (Geman and Geman, 1984), uses the property that the joint distribution is uniquely determined by its conditional distributions. In the Bayesian context, this implies for two dimensions that $p(\theta_1, \theta_2 \mid \mathbf{y})$ is uniquely determined by $p(\theta_1 \mid \theta_2, \mathbf{y})$ and $p(\theta_2 \mid \theta_1, \mathbf{y})$. The Gibbs sampler explores the posterior distribution by generating sampled values in a sequential manner. Given θ_1^k and θ_2^k at iteration k, the $(k+1)$-th value for each of the parameters is generated according to an iterative scheme. For the general case with d parameters:

- Set initial values $\theta_1^0, \theta_2^0, \ldots \theta_d^0$
- Sample $\theta_1^{(k+1)}$ from $p(\theta_1 \mid \theta_2^k, \theta_3^k, \ldots, \theta_{(d-1)}^k, \theta_d^k, \mathbf{y})$

- Sample $\theta_2^{(k+1)}$ from $p(\theta_2 \mid \theta_1^{(k+1)}, \theta_3^k, \ldots, \theta_d^k, \mathbf{y})$
- $\ldots$
- Sample $\theta_d^{(k+1)}$ from $p(\theta_d \mid \theta_1^{(k+1)}, \ldots, \theta_{d-1}^{(k+1)}, \mathbf{y})$

Hence, the Gibbs sampler generates a Markov chain which constitutes a dependent sample from the posterior distribution starting from a well chosen iteration (Lesaffre and Lawson, 2012).

2.3.1.2 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is a general MCMC technique to sample from the posterior distribution. In contrast to the Gibbs sampler, the MH algorithm does not require the full conditionals (Lesaffre and Lawson, 2012). The d parameters θ are sampled from a proposal distribution q. A portion of these samples will be accepted to yield a sample from the posterior distribution $p(\theta \mid \mathbf{y})$. This approach samples all d parameters together, according to the following iterative scheme:

- Set initial values θ^0
- Sample $\tilde{\theta}$ from the proposal density $q(\tilde{\theta} \mid \theta)$, with $\theta = \theta^k$
- Calculate the acceptance ratio:

$$\alpha(\theta^k, \tilde{\theta}) = min\left(\frac{p(\tilde{\theta} \mid y) q(\theta^k \mid \tilde{\theta})}{p(\theta^k \mid y) q(\tilde{\theta} \mid \theta^k)}, 1\right)$$

- Update $\theta^{(k+1)} = \tilde{\theta}$ with probability $\alpha(\theta^k, \tilde{\theta})$; otherwise set $\theta^{(k+1)} = \theta^k$

2.3.1.3 Combination of Gibbs sampler and Metropolis-Hastings algorithm

Mathematically, the MH algorithm encompasses Gibbs sampling. The Gibbs sampler is equivalent to the composition of d MH algorithms with acceptance probabilities all equal to 1. Combinations of these two approaches are also possible. For example, by using a Metropolis-within-Gibbs technique.

2.3.2 Software

Both Gibbs sampling and the Metropolis algorithm can be performed with software such as Bayesian inference Using Gibbs Sampling (BUGS) (Lunn et al., 2009). WinBUGS is the Windows version of the program BUGS and is one of the most popular software for Bayesian analysis (Lunn et al., 2000). WinBUGS carries out Bayesian inference by means of MCMC

methods. The whole program can be managed from R, using the R2WinBUGS package. This allows the user to submit the data and model file to WinBUGS, batch processing the MCMC sampling in WinBUGS and returning the samples to R for processing. OpenBUGS is an open source version of WinBUGS that runs on MS Windows. OpenBUGS is very similar to WinBUGS with regards to a lot of aspects. However, there is a greater class of algorithms available in OpenBUGS, hence more flexibility when fitting complex models. Another powerful software for Bayesian analysis using MCMC methods is JAGS (Just Another Gibbs Sampler). JAGS was developed and is maintained by (Plummer, 2003). The main advantage of JAGS over WinBUGS and OpenBUGS is its C++ platform hence can be written via command prompts, while the BUGS family is written in Pascal which is only available for Windows. The R packages R2jags and rjags can be used to manage the whole program from R. Another important statistical software is SAS®. SAS® is a versatile package of programs providing tools for setting up data bases, general data handling, statistical programming, and statistical analyses. Two SAS® procedures allow a Bayesian regression analysis, namely, PROC GENMOD and PROC MCMC. Approaches such as integrated nested Laplace approximations (INLA) and hierarchical likelihood (h-likelihood) provide alternative options. INLA is based on sophisticated Laplace approximations (Rue and Martino, 2007; Rue et al., 2009). Since no sampling is involved, computations appear to be much faster than with MCMC algorithms while keeping a high precision on the posterior estimates of the parameters. H-likelihood was developed by (Lee and Nelder, 1996) and provides a method to model correlated data while avoiding the computation of the marginal likelihood. For a more detailed comparison of software for Bayesian inference, we refer to Lesaffre and Lawson (2012).

Bibliography

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. Wiley & Sons, Chichester, West Sussex.

Leyland, A. H. and Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*. Wiley Series in Probability and Statistics. Wiley, Chichester, New York.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Plummer, M. (2003). *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling*.

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer New York, New York, NY.

CHAPTER 3

Robust and Censored Modeling and Prediction of Progression in Glaucomatous Visual Fields

This Chapter is based on: Bryan, S.R., Vermeer, K.A., Eilers, P.H.C., Lemij, H.G. and Lesaffre, E.M.E.H. (2013). Robust and Censored Modeling and Prediction of Progression in Glaucomatous Visual Fields. *Investigative Ophthalmology & Visual Science*, 54(10):6694 – 6700.

Abstract

Purpose: Classical regression is based on certain assumptions that conflict with visual field data. We investigate and evaluate different regression models and their assumptions in order to determine point-wise visual field (VF) progression in glaucoma and to better predict future field loss for personalised clinical glaucoma management. *Methods:* Standard automated visual fields of 130 patients with primary glaucoma with a minimum of 6 years of follow-up were included. Sensitivity estimates at each VF location were regressed on time with classical linear and exponential regression models, as well as different variants of these models that take into account censoring and allow robust fits. These models were compared for the best fit and for their predictive ability. The prediction was evaluated at 6 measurements (about 3 years) ahead using varying numbers of measurements. *Results:* For fitting the data, the classical uncensored linear regression model had the lowest root mean square error and 95th percentile of the absolute errors. These errors were reduced in all models when increasing the number of measurements used for the prediction of future measurements, with the classical uncensored linear regression model having the lowest values for these errors irrespective of how many measurements were included. *Conclusion:* All models performed similarly. Despite violation of its assumptions, the classical uncensored linear regression model appeared to provide the best fit for our data. In addition, this model appeared to perform the best when predicting future visual fields. However, more advanced regression models exploring any temporal-spatial relationships of glaucomatous progression are needed to reduce prediction errors to clinically meaningful levels.

3.1 Introduction

According to the World Health Organisation (WHO), glaucoma is one of the leading causes of blindness in the world. Adequate treatment may slow down the disease, possibly even halting its progression (Kingman, 2004). Evaluation of a longitudinal series of visual fields (VF), as measured by standard automated perimetry, provides a method to detect early evidence of glaucoma and to determine functional deterioration (Flammer and Meier, 2003). Hence, treatment strategies can then be optimized to prevent further progression of VF loss. Modeling the series of VFs is beneficial for this evaluation. It tries to extract the underlying VF progression from observed measurements and to distinguish it from noise, and can be used in the prediction of future visual fields.

Longitudinal modeling of VF summary parameters, such as the mean deviation (MD) and the visual field index (VFI), has been done before (Bengtsson and Heijl, 2008; Kymes et al., 2012; Artes et al., 2011). Modeling of individual VF points is potentially of greater interest, because it provides additional information such as the spatial nature of the fields, that is otherwise lost in global parameters (Azarbod et al., 2012). Modeling the progression of glaucoma in individual test locations may be of significant interest to best determine

patterns of progression. For example, more centrally located progression would arguably be of greater significance to the patient than more peripherally located progression. In addition, if progression took place in a particular location of the VF that was also affected in the fellow eye, this local progression would be more significant to the patients quality of life than when the local progression took place elsewhere in the field, where no such overlapping defects existed. Determining and predicting the pattern and location of progression would therefore help clinicians to best tailor the management of glaucoma in their patients.

Caprioli et al. (2011) discussed measuring the rate of VF progression in glaucoma for each VF location independently. They explored regression analysis of the sensitivity estimates against time in three models: linear, quadratic, and non-linear exponential. There were some aspects of VF data that were not taken into account in these models. Firstly, the maximum luminance of the Humphrey Field Analyzer (Carl Zeiss Meditec, Dublin, CA) perimeters stimuli is 10000 apostilbs, which is defined as 0 dB retinal sensitivity. The lowest sensitivity that can be detected by this perimeter is therefore 0 dB, although negative values could in fact occur if it were not for the limitations of this device (Anderson and Patella, 1999). It is therefore of interest to use a model that takes censoring at 0 dB into account, as suggested by Russell and Crabb (2011). We think that this is especially true for locations showing more advanced disease progression, with high sensitivity estimates in the early follow-up period, but a relatively large number of 0 dB sensitivity estimates towards the end of the follow-up period. The Tobit model is used to estimate the relationship between variables when there is either left- or right-censoring (or above- and below-censoring) in the dependent variable (Tobin, 1958). Another consideration of regression analysis that deserves attention is that classical least-squares fitting is based on certain assumptions about the measurement error, which in fact may be violated. For example, measurement error in visual fields increases with damage, and hence low sensitivity estimates have high variability. However, values lower than zero cannot be measured. This inherent censoring introduces a positive bias at low sensitivity estimates, which is made worse by the increased variability for low sensitivity estimates (Russell et al., 2012). Another problem in regression analysis is that of outliers or points that do not follow the general trend. As in most practical problems, visual field data may have several outliers. Single outliers may be easily identified in diagnostic tests, but it is generally more difficult to identify multiple outliers (Hawkins, 1980). Therefore, regression analyses that are effective, even in the presence of multiple outliers, need to be considered (Rousseeuw and Leroy, 1987), such as robust methods (Draper and Smith, 1998; Koenker, 2005).

In past research, the main aim has been to compare different statistical models to determine the optimal fit for the VF data. However, the model with the best fit to the data and the best model for prediction may not be the same. More complicated models may fit the data better, but in general, simpler models predict better since they are less likely to over-fit the data. This was illustrated by McNaught et al., who concluded that the best compromise between fitting and predicting sensitivity estimates is the linear model (McNaught et al., 1995). Since glaucoma is a progressive disease and the purpose of modeling the progression

is to be able to evaluate future field loss in a clinical setting, the predictive ability of these models is of importance. Our aim therefore was to introduce these censored and robust regression models for the point-specific evolutions of the visual fields and to evaluate these models, as well as their assumptions, for modeling and predicting point-wise glaucomatous VF progression.

3.2 Methods

The analysis was performed on the VF data of both eyes of 130 individuals from an ongoing study conducted at the Rotterdam Eye Hospital, The Netherlands. The inclusion criteria were glaucoma and a minimum of 15 VFs resulting in a minimum of 6 years of follow-up; the first 15 VFs were used in our analyses. This group consisted of 72 (55%) men and 58 (45%) women. All patients gave their written informed consent for participation. All research procedures followed the tenets set forth in the Declaration of Helsinki. The patients were followed up approximately twice per year. The data set contained 3900 VFs, resulting in time series of 13520 VF points. Descriptive statistics can be found in Table 3.1. All of the data that was used in this study has been made available online at http://orgids.com.

Table 3.1: Descriptive statistics of the study sample

	Mean	Median	Interquartile Range
Baseline age (years)	59.8	61.4	53.3 ; 67.6
Baseline MD (dB)	-8.3	-5.8	-12.8 ; -2.3
Average change in MD (dB/year)	-0.07	-0.04	-0.2; 0.1
Follow-up time (years)	7.7	7.6	7.2 ; 8.2

The VFs were tested by using the Humphrey Field Analyzer with the 24-2, white-on-white test strategy using the Full Threshold algorithm. The response variables of interest were the sensitivity estimates from the 52 VF points (excluding the 2 points that correspond to the blind spot). These locations were analysed as independent samples. The distribution of the observed sensitivity estimates is shown in Figure 3.1. Note that the frequency of points with 0 dB sensitivity was relatively large, which emphasizes the need to specifically address these measurements.

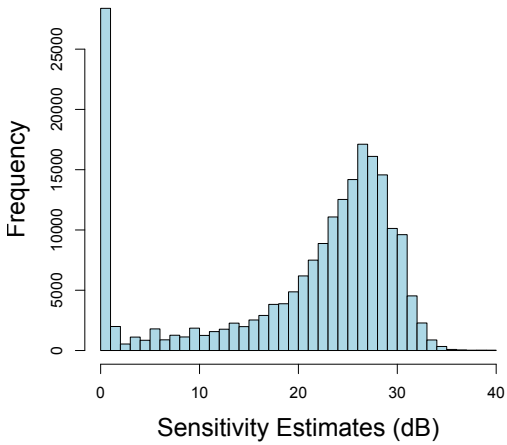

Figure 3.1: Distribution of point-wise sensitivity estimates in the study dataset.

An example of the series of VFs for one eye of a patient can be seen in Figure 3.2, which shows the sensitivity estimates for each location in the VF over a period of 9 years. This example clearly shows little or no progressive damage with little measurement variability in the superior hemifield and markedly progressive changes with relatively large measurement variability in most points of the inferior hemifield. There is also significant measurement variability in the upper, most nasal point of the upper hemifield.

3.2.1 Models

Regression models estimate the relationship between variables. The most straightforward approach is the linear regression model, which minimizes the sum of squared deviations from the regression line and passes through the mean in a straight line. A non-linear alternative is the exponential model. The 2 corresponding statistical models are:

1. First-order linear: $y_i = a + bx_i + \varepsilon_i$
2. First-order exponential: $y_i = e^{a+bx_i} + \varepsilon_i$

where denotes in our current study an index that is unique for every VF location in each eye ($i = 1, \ldots, 13520$), y_i is the measured sensitivity estimate expressed in dB, x_i denotes

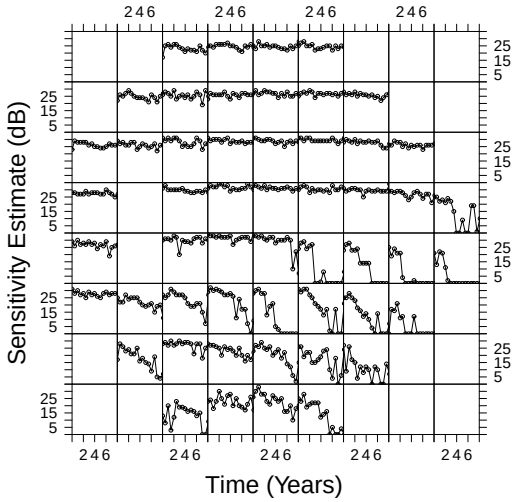

Figure 3.2: Retinal sensitivity estimates over time for each position of the visual field in the left eye of a single glaucoma patient.

the time the VF was acquired, a and b are the model parameters to be estimated and ε is the error term. Here, we assume that the error term is independently and identically normally distributed with a mean of zero and variance σ^2, i.e. $\varepsilon_i \sim N(0, \sigma^2)$.
A censored (Tobit) model in our current analysis takes into account that unseen sensitivity estimates that are indicated on the VF print out as < 0 do not correspond to true zero values: they are smaller than zero but the instrument was unable to determine such sensitivities. The model thus tries to estimate the relationship between time and the latent, true sensitivity value. The relationship between the observed and the latent, true sensitivity value is given by

$$y^* = y \times I(y \geq 0) \; + \; 0 \times I(y < 0).$$

Censoring was here taken into account for the linear model only. The response variable of the exponential model cannot go below 0 dB. Therefore, taking censoring at 0 dB into account in the model will not gain anything.
Linear and exponential models for the median, so-called median regression models were also included. Median regression provides a line which minimizes the sum of absolute values of the deviations from the regression line. Quantile regression, which includes median

regression, provides a method that is less influenced by extreme values. Median regression models can estimate the true rate of change even in cases where a significant fraction of the measurements has large errors. Other robust methods might be considered such as Least Absolute Deviations Regression (Edgeworth, 1887), M-estimators (Huber, 1981), and Robust Regression with Ranked Residuals (Draper and Smith, 1998), however including all these models would be beyond the scope of this paper (Hoaglin et al., 1981).
We evaluated the following 6 models: the classical uncensored linear model, the classical exponential model, the classical censored linear model, the uncensored linear on the median model, the exponential on the median model, and the censored linear on the median model.
Examples of some of the fits are shown in Figure 3.3. The points represent the sensitivity estimates from two locations belonging to the same eye from our data set, and the lines show the fits of the models based on this data. The left frame shows a case in which it is useful to take the censored nature of the data into account. The classical censored linear model takes into account that the measurements are censored at 0 dB, and hence shows the best model fit. This is only beneficial when there is progression, resulting in many 0 dB measurements. The right frame is an example which shows the need for more robust methods such as median regression. Here it is clear that the uncensored linear on the median fit is less influenced by the outlying points, notably during the first 2 years of follow-up, than the other models. These 2 examples show that it is not a priori obvious which model best fits the observed point-wise VF data.

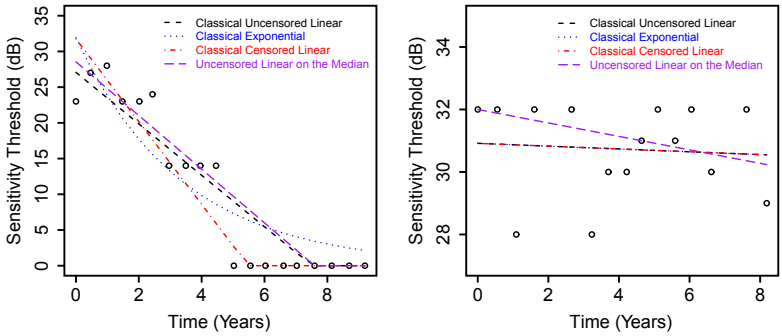

Figure 3.3: Two real examples of sensitivity estimates over time at 2 single location in the visual field. The lines indicate corresponding fits of 4 models.

3.2.2 Truncating the Predicted Response

Conventional models are designed to fit observed data, while censored models are designed to fit the true but unobserved data. To be able to compare predicted values from all models, the predicted values calculated to be less than 0 dB were set to 0 dB for the evaluation of both the fitting and the prediction of the data. Likewise, since the measured sensitivity estimates cannot realistically be above 40 dB, this limit was also set on the predicted values (Anderson and Patella, 1999).

3.2.3 Evaluation of model fit

We employed 2 classical approaches for the overall predictive ability of the models. The first aimed to measure how close the predicted values were to the observed responses in the current data (ie. model fit). This may be done using the Akaike Information Criterion (AIC). AIC is defined as $2k - 2ln(L)$, where k is the number of model parameters and L is the likelihood function evaluated at the estimated parameter values (Akaike, 1981). $-2ln(L)$ measures the model fit to the current data. By adding the model complexity indicated by $2k$, complex models are penalized. Generally, AIC is used to compare non-nested models with the same response, such as the classical uncensored linear and the classical exponential model. However, if the responses differ, one should compare AIC values with caution. In the case when comparing uncensored and censored models, it is not clear whether AIC can be used. An alternative to the AIC is leave-one-out cross-validation. We combined the errors (given by the predicted minus the observed sensitivity estimate) from each sensitivity estimate into summary error measures and compared the root mean square error (RMSE) and the mean absolute error (MAE) for each of the models. We compared both the RMSE and MAE since, by definition, the classical regression models minimize the RMSE while the median regression models minimize the MAE. We evaluated only the classical uncensored linear and classical exponential models by means of AIC, ranking them by percentage of best fits in order to compare our results to previous work (Caprioli et al., 2011), but used leave-one-out cross-validation to calculate the RMSE and MAE for the comparison of all models.
In clinical practice, one is primarily interested in VFs of individual eyes. Hence, interest lies in having a large fraction of patients within an estimated range, rather than looking at the average error over all eyes. Therefore we also computed the 95th percentile of the absolute errors, which is the value below which 95% of the absolute prediction errors may be found, to compare all models.
Since the RMSE and MAE were evaluated as summary measures, these statistics conceal how the models performed at the different levels of the sensitivities. Hence, to evaluate the models in greater detail, we also stratified the MAE over the range of sensitivity estimates to determine how the models perform at different sensitivity levels. In order to further investigate whether the assumptions that are made by the classical regression models hold, we evaluated the whole error distribution over the range of sensitivity estimates.

3.2.4 Evaluation of predictive ability

The second approach aimed to measure how each model performed with respect to predicting future measurements. This involves using the first sequence of data to estimate the model parameters and calculate a future measurement value for that point in the VF. This was done by varying the number of measurements used for the estimation, each time predicting the sensitivity estimates 6 measurements (approximately 3 years) ahead. All the models were evaluated by using the RMSE, MAE and 95th percentile of the absolute errors. As done for the comparison of the models fitting abilities, the MAEs were stratified over the range of sensitivity estimates to compare the models predictive abilities at the different sensitivity levels.

3.3 Results

3.3.1 Evaluation of model fit

When ranking the classical uncensored linear and the classical exponential models based on AIC, the classical uncensored linear model performed best in 57% of the fits, while the classical exponential model had 43% of the best fits. The RMSE, MAE and 95th percentile of the absolute errors for all models, calculated by cross-validation, are listed in Table 3.2. On our dataset, the classical censored linear model had the lowest MAE, although the MAE was almost the same over all the models. The classical uncensored linear model had the lowest RMSE and 95th percentile of the absolute error. All the models performed similarly. Overall, however, the classical uncensored linear model appeared to be the best model for our data.

Table 3.2: Comparison of the fitting ability of the models for the predicted values

Model	RMSE (dB)	MAE (dB)	95th Percentile (dB)
Classical Uncensored Linear	3.65	2.37	7.93
Classical Exponential	3.77	2.39	8.02
Classical Censored Linear	3.72	2.33	8.09
Uncensored Linear on the Median	3.81	2.34	8.26
Exponential on the Median	3.93	2.37	8.41
Censored Linear on the Median	3.83	2.34	8.31

Figure 3.4 shows the fitting ability of the models over different sensitivity levels. This shows a clear trend throughout the models, with low MAEs at 0 dB and in the range from 20 dB to 35 dB. This corresponds to the most frequently observed sensitivity estimates, as shown

in Figure 3.1. The models perform the worst between 0 dB and 25 dB and above 35 dB, where there are very few measurements and hence the models are not being optimized very well at these high sensitivity levels.

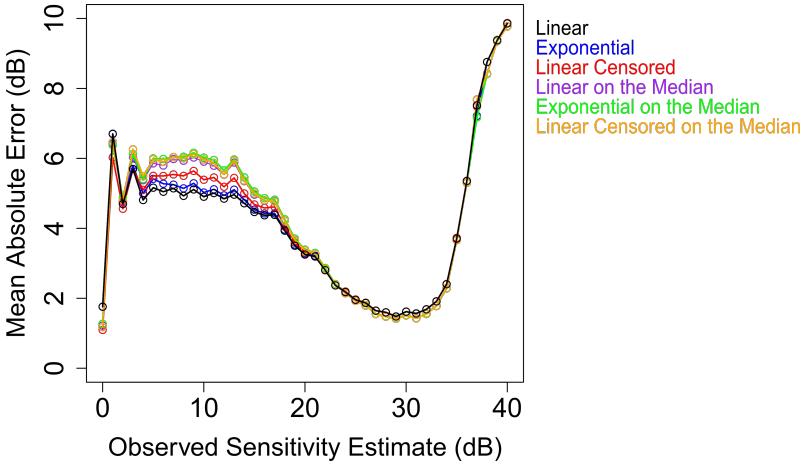

Figure 3.4: Mean absolute error from cross validation to evaluate the fitting ability of each model divided over the range of observed sensitivity estimates.

As a further step, we investigated the error distribution across the entire sensitivity range. Since the models all performed similarly, this was only done for the classical uncensored linear model, which had been found to be the best model. Curves of the 10th, 25th, 50th, 75th and 90th percentiles of the error are plotted for this model in Figure 3.5, superimposed over a smoothed scatter plot showing the observed sensitivity estimates versus the corresponding errors. The median of the errors approximated to zero for sensitivities around 0 dB and 25 dB. From 30 dB down to 10 dB, the median remained close to zero, but the whole distribution became wider. However, from 10 dB down to 0 dB the magnitude of the negative errors decreased due to the truncation at the range, while it continued increasing for the positive errors.

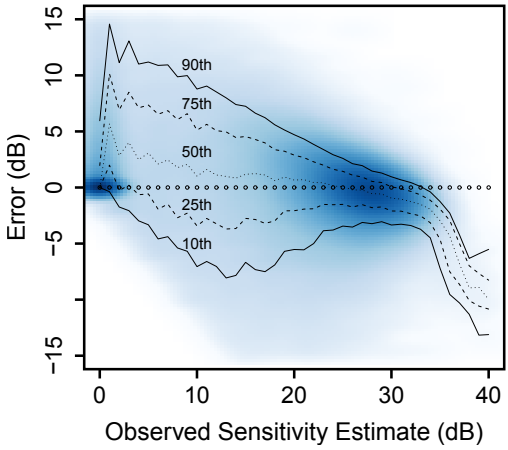

Figure 3.5: Smoothed scatter plot of the observed sensitivity estimates versus the corresponding error from cross validation including percentile curves (10th, 25th, 50th, 75th and 90th) across the range of sensitivity levels for classical uncensored linear regression. The dark blue areas correspond to a high density of observed sensitivity estimates.

3.3.2 Evaluation of predictive ability

For 3-year predictions, RMSE, MAE and 95th percentiles of the absolute errors were estimated for all models. The results are presented in Table 3.3. The classical uncensored linear model had the lowest RMSE, MAE and 95th percentile of the absolute errors. Since the variations from the classical uncensored models all performed worse than classical uncensored models, we have included only the results for the classical uncensored linear and classical exponential regression models in Figure 3.6. Here the MAE and 95th percentiles of the absolute errors for 3-year-predictions have been shown graphically for a varying numbers of time points. When including more measurements, the error was reduced; however the ordering of the models did not change. The classical uncensored linear regression model performed the best when comparing all the models irrespective of how many time points were used.

Table 3.3: Comparison of the predictive ability of the models by using 9 measurements to predict three years (6 measurements) ahead

Model	RMSE (dB)	MAE (dB)	95th Percentile (dB)
Classical Uncensored Linear	6.02	3.95	13.13
Classical Exponential	7.29	4.39	15.00
Classical Censored Linear	6.34	4.05	14.00
Uncensored Linear on the Median	6.34	4.06	14.00
Exponential on the Median	7.58	4.49	16.20
Censored Linear on the Median	6.51	4.12	14.22

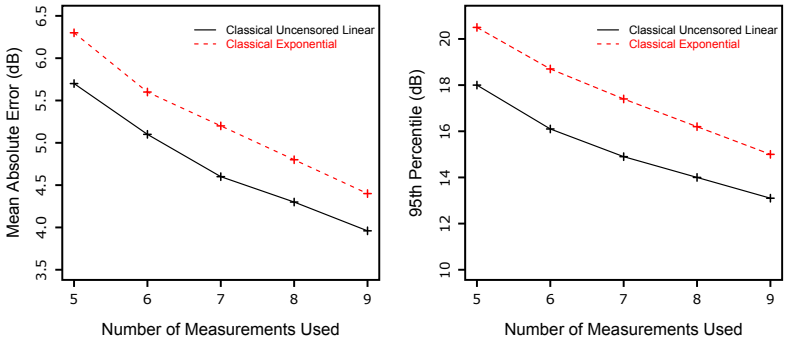

Figure 3.6: Mean absolute and 95th percentiles of 3-year prediction errors for the classical uncensored linear (black) and classical exponential (red) regression models.

Figure 3.7 shows the MAE over the range of sensitivity estimates when using 9 measurements to predict 3 years ahead. A similar trend could be seen throughout all the models as found when comparing the model fits, with the lowest MAEs occurring at 0 dB and between 25 dB and 35 dB. The difference between the models was however more pronounced between 0 dB and 20 dB when predicting than it was when fitting the data. When predicting, the exponential models were more likely to diverge and hence perform the worst within this range.

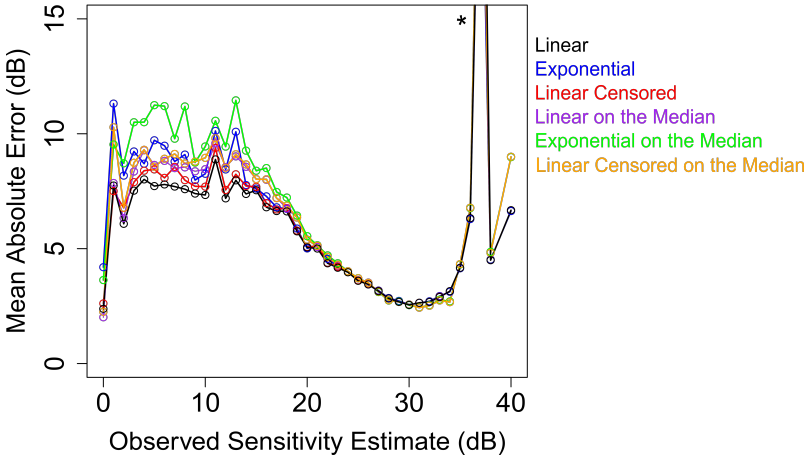

Figure 3.7: Mean absolute error to evaluate the predictive ability of each model divided over the range of observed sensitivity estimates. * MAE ranges from 32.8 dB to 37.0 dB for the different models.

3.4 Discussion

When ranking the classical uncensored linear and classical exponential models, the classical uncensored linear regression model was found to have the best fit most frequently (57% versus 43%). These results differ from those obtained by Caprioli et al. (2011), who concluded that the classical exponential model provided the best fit for their data in a very large majority of cases. This could be due to different implementation of the models. Caprioli et al. specified a mathematical model and not a statistical model. In the latter model it is also specified how the error term enters the model (i.e. as an additive term to the exponential or behind the exponential sign). Without knowing the statistical model which they used, we are unable to determine whether our implementations were the same.

When including all the models, the MAEs were similar, and the classical uncensored linear model again performed the best when comparing the RMSE and 95th percentile of the absolute error. The classical uncensored linear regression model was also found to be the best model for prediction irrespective of how many measurements were used to fit the model. Although the models performed similarly, we conclude that for both fitting purposes, i.e., for determining the rate of progression per test location, and for prediction of future measurements, the classical uncensored linear regression model is recommended. However, even for this model the mean absolute errors were still on the order of 2 dB for fitting, and 4

dB for prediction, which is arguably too high to be of direct clinical use.

A possible reason for the poor discrimination between the goodness of fit of the various fitting and predictive models is the low progression rates found in this study. This might be due to having a younger cohort than other studies, different inclusion and exclusion criteria, or the fairly large proportion of data points with 0 dB retinal sensitivity.

The lowest error for fitting the data was found to be at 0 dB and around 30 dB for all models. This corresponds to the sensitivity estimates which were most frequently seen in this dataset. In the same way, only a few measurements were seen between 0 dB and 25 dB and above 35 dB, where the models all performed the worst. This is also the case when comparing the predictive abilities of the models. It has also been shown that the test-retest variability is highest in this range and hence the values which we assume to be true also have a high error (Russell et al., 2012). However, the difference between the models was more evident between 0 dB and 20 dB when comparing the models for best predictive ability than for best fitting ability. Hence, none of the models were being optimized well in this range, and the result of this problem is even more evident when predicting future measurements. As before, the real value of the sensitivity estimate is unknown. Therefore the measurements being predicted also have error, so an accurate prediction of this measurement is not possible.

The distribution of the residuals was not symmetric and differed over the range of sensitivity estimates, which is a violation of the assumption made by the classical regression models. The bias towards negative errors was however emphasized at the higher end of the range due to there being only a few measurements above 35 dB. The truncation of the negative errors at the lower end of the range was due to the device not being able to measure sensitivity estimates below zero, as well as the truncation of the predicted response which we applied in our analysis. This could be an interesting aspect of the data to investigate further, by using a model which assumes that the distribution of the residuals differ across the range of sensitivity estimates.

In previous research, a quadratic model was often considered in addition to linear and exponential models. Compared to linear and exponential models, these studies did not identify the quadratic model as an improvement (Caprioli et al., 2011; McNaught et al., 1995). In addition, the model is not monotonic and there are no theoretical reasons, e.g. from anatomy or functional models, for using such a quadratic model. We also evaluated the quadratic model but found no indication of its suitability in our application. We therefore argue that this model can safely be excluded in future research on modeling time series of threshold sensitivities.

In this study, censoring was done at 0 dB as suggested by Russell and Crabb (2011). However, recent research found that ganglion saturation can also account for high test-retest variability between 0 dB and 25 dB and hence 0 dB is outside the dynamic range of a ganglion cell (Swanson et al., 2011), and that sensitivities below 10 dB are mainly noise (Junoy Montolio et al., 2012b). For this reason, we also considered censoring and truncating of the predicted response at 10 dB. When censoring and truncation were done at 10 dB rather than 0 dB, the MAE was higher for all the models, however the classical uncensored linear model was still found to be the best model for fitting and for prediction.

For future work, certain aspects of the data need to be further investigated. One of the difficulties in modeling this data is the amount and type of measurement error and variability in the sensitivity estimates. This may be due to blinking, eye movements, flagging attention by the patient, fatigue, inexperience, a delayed reaction time and the patient forgetting to press the button. Factors such as test reliability, technical experience, time of day and season have been shown to influence standard automated perimetry test results (Junoy Montolio et al., 2012). In order to measure the true progression of the disease, the high, sensitivity-dependent, variability would need to be taken into account by incorporating it into the statistical model, or new, more reproducible methods for determining the visual fields need to be developed. The device should flag measurements if they fall outside a clinically acceptable range determined from previous observations. This would indicate that either the measurement is wrong and the test should be redone, or that there is a real change in the rate of progression and the treatment strategy may need to be adjusted.

In addition, one could exploit the spatial nature of the data and capitalize on the specific spatial organization of the nerve fibres in the eye (Denniss et al., 2012; Airaksinen et al., 2008). A problem with using regression models is that each location is treated as an independent sample and we are not able to link locations belonging to the same eye together. This problem can be addressed by using spatial models or mixed-effects models which we propose for future work (Verbeke and Molenberghs, 2000).

Finally, sensitivity estimates at different test locations in the same VF do not all progress at the same speed, and directly modeling summary parameters will not be able to handle this well. Hence it may also be interesting to compare the summary predictions by modeling the MD directly and by predicting each point and consequently calculating the MD. This would elucidate whether modeling the individual points is more informative than modeling global indices (Azarbod et al., 2012).

In conclusion, we have shown that the classical uncensored linear model was found to be the best model for fitting visual field data, as well as for predicting future visual field measurements. However, despite the various theoretical pros and cons of each of the explored models, none of them performed satisfactorily well, neither for fitting, nor for predicting future damage, which is why we need to further explore models that take spatio-temporal relationships more into consideration. Modeling point-wise threshold sensitivity over time is relevant, but more advanced models are needed to further reduce prediction errors and produce clinically meaningful results.

Bibliography

Airaksinen, P. J., Doro, S., and Veijola, J. (2008). Conformal geometry of the retinal nerve fiber layer. *Proceedings of the National Academy of Sciences*, 105(50):19690–19695.

Akaike, H. (1981). Likelihood of a model and information criterion. *Journal of*

Econometrics, 16(1):3–14.

Anderson, D. R. and Patella, V. M. (1999). *Automated Static Perimetry*. Mosby, 2nd edition.

Artes, P. H., O'Leary, N., Hutchison, D. M., Heckler, L., Sharpe, G. P., Nicolela, M. T., and Chauhan, B. C. (2011). Properties of the statpac visual field index. *Investigative Ophthalmology & Visual Science*, 52(7):4030–4038.

Azarbod, P., Mock, D., Bitrian, E., Afifi, A. A., Yu, F., Nouri-Mahdavi, K., Coleman, A. L., and Caprioli, J. (2012). Validation of point-wise exponential regression to measure the decay rates of glaucomatous visual fields. *Investigative Ophthalmology & Visual Science*, 53(9):5403–5409.

Bengtsson, B. and Heijl, A. (2008). A visual field index for calculation of glaucoma rate of progression. *American Journal of Ophthalmology*, 145(2):343–353.

Caprioli, J., Mock, D., Bitrian, E., Afifi, A. A., Yu, F., Nouri-Mahdavi, K., and Coleman, A. L. (2011). A method to measure and predict rates of regional visual field decay in glaucoma. *Investigative Ophthalmology & Visual Science*, 52(7):4765–4773.

Denniss, J., McKendrick, A. M., and Turpin, A. (2012). An anatomically customizable computational model relating the visual field to the optic nerve head in individual eyes. *Investigative Ophthalmology & Visual Science*, 53(11):6981.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley-Interscience, New York, 3rd edition.

Edgeworth, F. Y. (1887). On observations relating to several quantities. *Hermathena*, 6(13):279–285.

Flammer, J. and Meier, E. (2003). *Glaucoma: A Guide for Patients, An Introduction for Care Providers, A Quick Reference*. Hogrefe & Huber Publishers, Cambridge, MA, 2nd edition.

Hawkins, D. M. (1980). *Identification of Outliers*. Monographs on Applied Probability and Statistics. Springer Netherlands, Dordrecht.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1981). *Understanding Robust and Exploratory Data Analysis*. Wiley.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Junoy Montolio, F. G., Wesselink, C., Gordijn, M., and Jansonius, N. M. (2012a). Factors that influence standard automated perimetry test results in glaucoma: Test reliability, technician experience, time of day, and season. *Investigative Ophthalmology & Visual Science*, 53(11):7010–7017.

Junoy Montolio, F. G., Wesselink, C., and Jansonius, N. M. (2012b). Persistence, spatial distribution and implications for progression detection of blind parts of the visual field in glaucoma: A clinical cohort study. *PLoS ONE*, 7(7):e41211.

Kingman, S. (2004). Glaucoma is second leading cause of blindness globally. *Bulletin of the World Health Organization*, 82(11):887–888.

Koenker, R. (2005). *Quantile Regression*. Number 38 in Econometric Society Monographs. Cambridge University Press, Cambridge.

Kymes, S. M., Lambert, D. L., Lee, P. P., Musch, D. C., Siegfried, C. J., Kotak, S. V., Stwalley, D. L., Fain, J., Johnson, C., and Gordon, M. O. (2012). The development of a decision analytic model of changes in mean deviation in people with glaucoma: The COA model. *Ophthalmology*, 119(7):1367–1374.

McNaught, A. I., Crabb, D. P., Fitzke, F. W., and Hitchings, R. A. (1995). Modelling series of visual fields to detect progression in normal-tension glaucoma. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 233(12):750–755.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York, NY, USA.

Russell, R. A. and Crabb, D. P. (2011). On Alternative methods for measuring visual field decay: Tobit linear regression. *Investigative Ophthalmology & Visual Science*, 52(13):9539–9540.

Russell, R. A., Crabb, D. P., Malik, R., and Garway-Heath, D. F. (2012). The relationship between variability and sensitivity in large-scale longitudinal visual field data. *Investigative Ophthalmology & Visual Science*, 53(10):5985–5990.

Swanson, W. H., Sun, H., Lee, B. B., and Cao, D. (2011). Responses of primate retinal ganglion cells to perimetric stimuli. *Investigative Ophthalmology & Visual Science*, 52(2):764–771.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer New York, New York, NY.

CHAPTER 4

Bayesian Hierarchical Modelling of Longitudinal Glaucomatous Visual Fields using a Two-Stage Approach

This Chapter is based on: Bryan, S.R., Eilers, P.H.C., van Rosmalen, J., Rizopoulos, D., Vermeer, K.A., Lemij, H.G. and Lesaffre, E.M.E.H. (2017) Bayesian Hierarchical Modelling of Longitudinal Glaucomatous Visual Fields using a Two-Stage Approach. *Statistics in Medicine*, 36:1735 - 1753.

Abstract

The Bayesian approach has become increasingly popular because it allows to fit quite complex models to data via Markov chain Monte Carlo (MCMC) sampling. However, it is also recognized nowadays that MCMC sampling can become computationally prohibitive when applied to a large data set. We encountered serious computational difficulties when fitting an hierarchical model to longitudinal glaucoma data of patients who participate in an ongoing Dutch study. To overcome this problem, we applied and extended a recently proposed two-stage approach to model these data. Glaucoma is one of the leading causes of blindness in the world. In order to detect deterioration at an early stage, a model for predicting visual fields (VF) in time is needed. Hence, the true underlying VF progression can be determined, and treatment strategies can then be optimized to prevent further VF loss. Since we were unable to fit these data with the classical one-stage approach upon which the current popular Bayesian software is based, we made use of the two-stage Bayesian approach. The considered hierarchical longitudinal model involves estimating a large number of random effects and deals with censoring and high measurement variability. In addition, we extended the approach with tools for model evaluation.

4.1 Introduction

Since the introduction of MCMC sampling by Gelfand and Smith (Gelfand and Smith, 1990) and the development of the BUGS software (Lunn et al., 2009) the Bayesian approach has become tremendously popular in various application areas, but especially to fit models to complex data structures. But with the years it also became clear that MCMC sampling can be computationally quite cumbersome, and even prohibitive, for fitting complex models to relatively large data sets. Several attempts have been made to look for alternative computational procedures and software, with notable examples such as INLA (Rue et al., 2009) and STAN (Hoffman and Gelman, 2014). While this newly developed software can sometimes speed up the computations considerably, the computational gain is not always obvious upfront and for some advanced models the new developments may not be suitable yet. In addition, the majority of the practical Bayesians still use BUGS-related software. In this context, Lunn et al. (2013) proposed to fit a hierarchical model in two stages. The authors claim more model flexibility in this way, but advocate the use of their procedure especially for its computational properties. In this paper we further illustrate the use of the two-stage approach on a far more complex hierarchical data structure of glaucoma patients. In addition, we extend the approach with an additional sampling step to allow the calculation of model selection and model evaluation criteria.

Our modeling approach is motivated by data from the Glaucoma study conducted by the Rotterdam Eye Hospital in the Netherlands. According to the World Health Organization (WHO), glaucoma is one of the leading causes of irreversible blindness in the world

(Kingman, 2004). Adequate treatment may slow down the disease, possibly even halting its progression. Evaluation of a longitudinal series of visual fields (VF), as measured by standard automated perimetry (SAP), provides a way to detect early evidence of glaucoma and to determine functional deterioration. However, due to the subjective nature of this technique, SAP is prone to large variability. In order to measure the true progression of the disease, this variability needs to be taken into account. The Glaucoma study provides a unique database with a long follow up time. Although statistical methods may have existed to model such data, the difficulties in extracting it from the device has made this type of data rare and hence has prevented much research on the topic.

The response variable of interest is the sensitivity estimate describing the level of differential light sensitivity at different locations within each eye. The sensitivity estimates are left-censored due to a limitation of the device. Models which take into account this type of censoring, such as the Tobit model, have been described in the literature (Tobin, 1958). Our interest lies in modeling the latent, true values rather than the observed sensitivity estimates for two reasons. Firstly, clinical interest lies in predicting the disease progression rather than the observed sensitivity estimates. Secondly, using the latent scale allows us to use a simpler model than when directly modeling the observed data. The hierarchical structure of the data consists of 4 levels, namely, (1) the individual, (2) the eye, (3) the hemifield and (4) the location. There is a vast amount of literature that addresses hierarchical mixed-effects models, for both frequentist (Verbeke and Molenberghs, 2000) and Bayesian (Ntzoufras, 2009; Lesaffre and Lawson, 2012) approaches. We model this complex data structure by means of a Bayesian hierarchical mixed-effects model with cross-classified random effects. Hence, we combine both spatial and time effects. One of the difficulties in modeling VF data is the amount and type of measurement error or variability in the sensitivity estimates. This may be due to measurable factors, such as season, time of day and reliability indices, or unknown transient factors, such as fatigue, lack of concentration, or delayed reaction time. Although their magnitudes may vary, these factors affect all locations belonging to the same VF. We propose to model them as Global Visit Effects (GVEs). Furthermore, there is an inverse relationship between sensitivity and variability. For example, measurement error in the VFs increases with damage, and hence low sensitivity estimates have high variability. Therefore, it is naive to assume a constant variance over the wide range of sensitivity estimates. In this paper, we relax this assumption in order to incorporate this relationship. A problem with high dimensional data and complex data structures, is that it is sometimes difficult or even impossible to model them with standard MCMC algorithms. Lunn et al. (2013) proposed a two-stage approach, which allowed us to simplify the problem while still benefiting from the advantages of a full Bayesian model. However, one of the disadvantages of this approach, is that it is not possible to directly obtain the random effects estimates needed for most model comparisons. We address this issue by extending the two-stage approach to be able to determine these estimates.

Our aim is to model this complex data structure in order to obtain better estimates of the true evolution of the sensitivity over time, so that treatment strategies can be optimized to

prevent further progression of VF loss. The structure of the paper is as follows. In Section 5.4.1 we give further details on the motivating data set and introduce the research questions that triggered our modeling approach(es). In Section 5.2 we describe the models used in the analysis. In the subsequent section we briefly review computational aspects of the analysis. Model comparison is dealt with in Section 4.5. In Section 4.6 we apply our models to the Glaucoma study data. Section 4.7 contains a concluding discussion. Further details regarding the modeling approach are provided in an appendix.

4.2 Motivating data set: the Glaucoma study

4.2.1 Description of the project

The Glaucoma study is a prospective cohort study conducted by the Rotterdam Eye Hospital in the Netherlands. This is an ongoing study which began in 1998. Inclusion criteria included glaucoma diagnosis and an age range of 18 to 85 years. In total, 139 patients, consisting of 80 (57.6%) men and 59 (42.4%) women, were recruited with a mean follow-up of 10.5 years. Note that for the statistical analysis of the data, we excluded one patient (see Section 4.6) leading to $N = 138$ patients. Follow-up data were collected at approximately 6-monthly intervals. All patients gave their written informed consent for participation. All research procedures followed the tenets set forth in the Declaration of Helsinki. Furthermore, all of the data that was used in this analysis has been made available online at http://rod-rep.com.

Sensitivity estimates were measured at 52 test locations within each eye, or 26 test locations within each hemifield (excluding two locations corresponding to the blind spot) as shown in Figure 4.1. The VFs were tested using the Humphrey Field Analyzer with the 24-2, white-on-white test strategy using the Full Threshold algorithm. The light source can be attenuated in the range from 1 to 10,000 times. On the decibel (dB) scale an attenuation x is defined as $s = 10 \log_{10}(x)$, or $x = 10^{s/10}$. The lowest sensitivity that can be detected by this perimeter is 0 dB, although negative values could in fact occur if it were not for the limitations of this device. The highest sensitivity that can be detected is 50 dB, however few humans are capable of seeing a stimulus less than 40 dB, which is 1/10,000 of the maximum intensity of the instrument (or 1 asb). Thus, for practical purposes, the useful intensity range for white light testing is from 0 to 40 dB with a background illumination of 31.5 apostilb (asb.) (Anderson and Patella, 1999).

4.2.2 Previous research

Parameters such as the mean deviation (MD) and visual field index (VFI) summarize the 52 sensitivity estimates into single values which can be used by the clinicians when optimizing treatment strategies. Longitudinal modeling of these VF summary parameters has been done

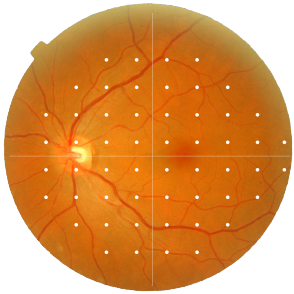

Figure 4.1: Fundus photo of a left human eye with the 54 test locations for the VF test represented by white dots.

before (Bengtsson and Heijl, 2008; Artes et al., 2005; Kymes et al., 2012; Junoy Montolio et al., 2012). Modeling of individual test locations is potentially of greater interest, because it provides additional information such as the spatial nature of the fields which is otherwise lost in global parameters. In previous research, each location was analyzed as an independent sample (McNaught et al., 1995; Caprioli et al., 2011; Bryan et al., 2013). However, separate location-specific regression models are not able to use any information from the data set as a whole. Multilevel mixed-effects models provide a better fit to the data than separate regression models by accounting for group effects and/or within-group correlation (Verbeke and Molenberghs, 2000). This was shown in the context of global VF measurements by Pathak et al. (2013).

In glaucoma, variability is presumably related to fatigue effects and response errors, whereby sensitivity estimates decrease over time (Hudson et al., 1994; Bengtsson and Heijl, 2000). Differences in fatigue effects, between the inferior and superior hemifields within an eye have been demonstrated (Hudson et al., 1994). Furthermore, this effect may differ between the first and second eye at the same visit. The number of false-negative answers has been shown to be higher in eyes with field loss. This may be explained by an increased variability in sensitivity estimates typically found in such eyes (Bengtsson and Heijl, 2000; Russell et al., 2012). A common approach to reduce measurement variability is to average multiple measurements. For example, random uncorrelated measurement errors that are present in the point-wise sensitivity estimates are reduced when calculating summary parameters such as the mean deviation (MD). Other errors, however, are spatially correlated and affect the whole VF. One group of such errors are measurable factors, including season, time of day and reliability indices, which have been evaluated before (Junoy Montolio et al., 2012). Although the effects of these factors are statistically significant, they are rather small

and hence only explain a small part of the observed global variation in VFs.

The inverse relationship between variability and sensitivity has been described in the literature. Henson et al. (2000) found that this relationship is well represented by the function $\ln(SD) = A + B \times sensitivity(dB)$, where A and B are 2.81 dB and -0.066 dB respectively for normal eyes and 3.62 dB and -0.098 dB for glaucomatous eyes. Russell et al. (2012) showed that the distribution of residuals is relatively concentrated at high VF sensitivities (26 to 36 dB) but stretches substantially as the sensitivity estimates decrease to a level of 10 dB. Sensitivity estimates near 10 dB are associated with residuals spanning almost the entire dynamic range of the instrument. This could be caused by a loss of ganglion cells (due to glaucomatous damage), or relocation of the stimulus to the peripheral visual field where there are fewer ganglion cells Swanson et al. (2011). Zhu et al. (2014) describe a method to detect change by means of an inferential statistical model which incorporates the non-stationary variability and uses a mixture of Weibull distributions.

Although there is a wide range of literature which discusses these aspects, the majority of previous work deals with the global indices or treats sensitivity estimates for each location in the VF as an independent sample. Furthermore, these aspects have been addressed separately. Hence, it is clear that an approach which takes into account the complex structure of the data and considers all of the aforementioned problems, is needed. We will address censoring, the hierarchical structure, the global variation as well as the relationship between the variability and sensitivity.

4.3 Statistical Models

Modeling the sensitivity estimates is beneficial for the evaluation of the progression of VF loss. By incorporating biological effects into the model, we aimed to improve the model fit and hence provide a better method for modeling this progression. This was done by building the model up sequentially, resulting in 3 different proposed models.

4.3.1 Censoring

It is important to note that unseen sensitivity estimates are indicated on the VF print out as < 0. They are smaller than zero because the instrument is unable to determine such sensitivities. Thus a model which defines the relationship between time and the latent, true sensitivity estimate is needed. The relationship between the observed sensitivity estimate y and the latent, true sensitivity estimate y^* is given by $y = \max(0, y^*)$.

4.3.2 Hierarchical model

We propose using a Bayesian hierarchical mixed-effects model (Ntzoufras, 2009; Lesaffre and Lawson, 2012) to analyze the glaucoma data. This model is able to take into account both the within subject and between subject variability. Furthermore, we capitalized on the common features within each eye by taking into account the correlation between measurements belonging to the same eye. In addition, correlation of VF measurements within the inferior and superior hemifields, separated by the horizontal raphe, was assumed to be higher than between hemifields. Hence, the hierarchical structure of the data consists of 4 levels, namely, (1) the individual, (2) the eye, (3) the hemifield and (4) the location. The dependent variable in the model is y^*_{iehlt}, which denotes the latent sensitivity estimate for individual $i = 1,\ldots,N$; eye $e = 1,2$; hemifield $h = 1,2$ and location $l = 1,\ldots,26$ at timepoint $t = 1,\ldots,T_i$. The independent variable in the model is time_{it}, which represents the time (in years) between tth measurement and the first measurement for each individual i, ranging from 0 to 10.5 years. Let $\mu_\alpha = (\mu_{\alpha_0}, \mu_{\alpha_1})^T$ correspond to the population-averaged intercept and slope of time. The individual-specific intercept and slope of time for individual i are represented by α_{0i} and α_{1i}, with deviations due to the eye given by γ_{0ie} and γ_{1ie}, due to the hemifield by η_{0ieh} and η_{1ieh}, and due to the location by λ_{0iehl} and λ_{1iehl}. We call this Model 1.

Model 1:

$$\begin{aligned} y^*_{iehlt} &= \alpha_{0i} + \alpha_{1i}\text{time}_{it} + \gamma_{0ie} + \gamma_{1ie}\text{time}_{it} + \eta_{0ieh} + \eta_{1ieh}\text{time}_{it} + \lambda_{0iehl} + \\ &\quad \lambda_{1iehl}\text{time}_{it} + \varepsilon_{iehlt} \\ &= \mu^{(1)}_{iehlt} + \varepsilon_{iehlt}, \end{aligned} \tag{4.3.1}$$

where $\varepsilon_{iehlt} \sim N(0, \sigma^2_\varepsilon)$ and with priors:

$$\begin{aligned} \alpha_i &= \tbinom{\alpha_{0i}}{\alpha_{1i}} \sim N\left(\tbinom{\mu_{\alpha_0}}{\mu_{\alpha_1}}, \Sigma_\alpha\right); \\ \gamma_{ie} &= \tbinom{\gamma_{0ie}}{\gamma_{1ie}} \sim N(\tbinom{0}{0}, \Sigma_\gamma); \\ \eta_{ieh} &= \tbinom{\eta_{0ieh}}{\eta_{1ieh}} \sim N(\tbinom{0}{0}, \Sigma_\eta); \\ \lambda_{iehl} &= \tbinom{\lambda_{0iehl}}{\lambda_{1iehl}} \sim N(\tbinom{0}{0}, \Sigma_\lambda); \\ \mu_{\alpha_b} &\sim N(0, 10^9) \text{ for } b = 0,1; \\ \sigma^2_\varepsilon &\sim \text{IG}(10^{-3}, 10^{-3}); \\ \Sigma_\alpha, \Sigma_\gamma, \Sigma_\eta \text{ and } \Sigma_\lambda &\sim \text{IW}(\text{diag}(1,1), 2). \end{aligned}$$

4.3.3 Global Visit Effect

Junoy Montolio et al. (2012) explicitly modeled the global variations with known factors such as season, time of day and reliability indices. However, we speculated that other transient factors, such as fatigue, lack of concentration, or delayed reaction time may play a more important role. Since all these (as well as possibly other) factors affect all locations belonging to the same VF, we propose to take them together and to call them, as well as model them as the Global Visit Effects (GVEs). In this way, we can account for both the known and the unknown factors. Hence, the GVE accounts for all factors that affect all measurements of the same eye at each visit. The GVE is an eye-specific visit effect. The term "global" relates to the whole VF. To illustrate the importance of these factors, we show in Figure 4.12 the VFs over time of one eye, where all locations have a drastic decrease in sensitivity at around one year. From the longitudinal profiles, it is evident that this decrease is caused by something that affected all VF measurements of that visit, rather than by actual damage. To account for the visit-dependent offset at all locations, or GVE, we included a parameter, ϕ_{iet}, in the following model to capture the offset at the tth time point for eye e within individual i. This gives Model 2.

Model 2:

$$\begin{aligned} y^{*}_{iehlt} &= \mu^{(2)}_{iehlt} + \varepsilon_{iehlt} \\ &= \mu^{(1)}_{iehlt} + \phi_{iet} + \varepsilon_{iehlt} \end{aligned} \tag{4.3.2}$$

From an initial exploratory analysis, we observed a number of spikes in the distribution of the visit effects. To accommodate these spikes, we assumed a t-distribution for ϕ_{iet}. The t-distribution allows greater flexibility in the distribution of random effects compared to the normal distribution, and can handle heavy tails (Lee and Thompson, 2008). Hence, we let

$$\phi_{iet} \sim t(0, \sigma^2_{\phi}, 3),$$

where $t(\mu, \sigma^2, df)$ denotes the generalized t-distribution with mean μ, scale parameter σ^2, and df degrees of freedom. The prior for σ^2_{ϕ} was taken equal to IG($10^{-3}, 10^{-3}$).

4.3.4 Relationship between variability and sensitivity

There is an association between a decline in VF sensitivity and an increase in response variability. However, values lower than 0 dB cannot be measured. This inherent censoring process introduces a positive bias at low sensitivity estimates, which is made worse by the increased variability for low sensitivity estimates. We assumed a linear relationship between the expected values of the sensitivity estimates and the logarithm of the standard deviation. However, since we were interested in modeling the latent sensitivity estimates, we

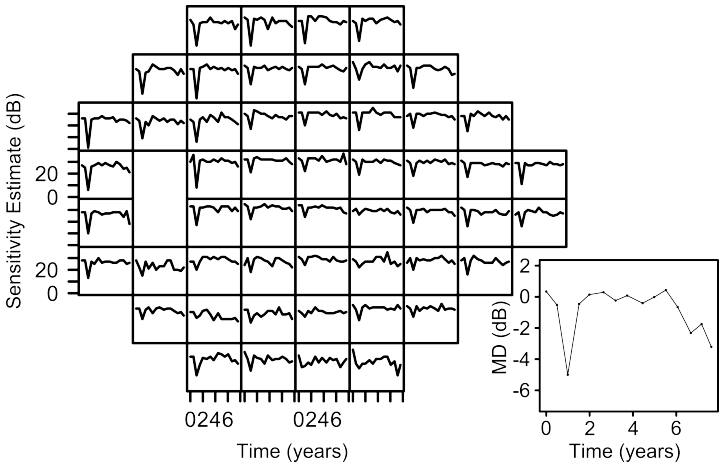

Figure 4.2: Retinal sensitivity estimates over time for each location of the visual field in the left eye of a single glaucoma patient. A decrease in the sensitivity estimates can be seen in all locations at around 1 year. The longitudinal profile of the MD values over time is shown on the right. The visit-dependent decrease is also clear at around 1 year for the MD.

extrapolated this linear relationship for predicted sensitivity estimates below 10 dB. This can be seen in Figure 4.3. In this exploratory analysis, we found that the relationship was well represented by the function $\ln(SD) = A + B \times sensitivity(dB)$, where A and B are 2.60 dB and -0.06 dB respectively. We extended Model 2 to incorporate this relationship in Model 3.

Model 3:

$$y^*_{iehlt} = \mu^{(2)}_{iehlt} + \varepsilon_{iehlt},$$

with

$$\varepsilon_{iehlt} \sim N(0, \sigma^2_{\varepsilon,iehlt}),$$

and

$$\log(\sigma_{\varepsilon,iehlt}) = \mu_{\zeta_0} + \mu_{\zeta_1}\mu^{(2)}_{iehlt}, \tag{4.3.3}$$

where $\sigma_{\varepsilon,iehlt}$ is the standard deviation for individual $i = 1,\ldots,N$; eye $e = 1,2$; hemifield $h = 1,2$; location $l = 1,\ldots,26$ at time $t = 1,\ldots,T_i$. The priors for μ_{ζ_0} and μ_{ζ_1} are both N(0, 10^9). For the other parameters in Model 3 the priors are the same as those for Models 1 and 2. A summary of all parameters and their definitions for all the models is shown in Table 4.1.

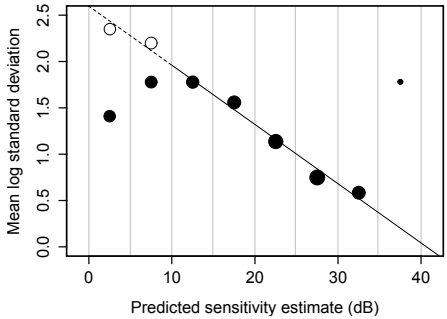

Figure 4.3: Bubbleplot representing the mean logarithm of the standard deviation for different predicted sensitivity estimates determined using linear regression for each location. The predicted values were subdivided into groups with width 5 dB. The empty bubbles correspond to the hypothetical values, corresponding to the censored measurements, for the mean logarithm of the standard deviation for the predicted sensitivity estimates after extrapolation. The bubbles are scaled to the logarithm of the number of observations. Due to the small number of observations within the 35 dB to 40 dB range (0.02% of the total observations), we excluded these measurements in the linear regression.

4.4 Estimation Approach

4.4.1 One-stage approach

The Bayesian approach takes into account the uncertainty in all model parameters and allows prior information to be incorporated. Furthermore, MCMC algorithms allow greater flexibility by relaxing the strong parametric assumptions commonly used in most frequentist hierarchical models (Lunn et al., 2013; Lesaffre and Lawson, 2012). The classical Bayesian approach is one-stage hierarchical modeling, which has the advantage that subject-specific and overall parameters are estimated simultaneously. However, for a (relatively) large data set, this approach can be difficult or even impossible to implement for complex models with standard MCMC software. In our case, we had a total of 45,005 parameters which needed to be estimated. As a consequence, we were unable to achieve convergence in a realistic time frame and we experienced computer memory limitations when using WinBUGS or JAGS. For such situations, a computationally more tractable method is needed.

4.4.2 Two-stage approach

Lunn et al. (2013) proposed two-stage Bayesian hierarchical modeling. The two-stage approach allowed us to simplify the problem by splitting hierarchical models with M levels at level m^*. Independent parameters of interest at level m* are obtained in stage 1 and used as proposal distributions for those parameters in stage 2. Lunn et al. (2013) illustrated this method using models with two and three levels. The glaucoma data exhibit a similar hierarchical structure to that presented by Lunn et al. (2013), but of a more complex nature with four levels. Figure 6.8 illustrates the hierarchical structure of the glaucoma data, as well as the cross-classified random effects, divided into two stages. In our case, we split the levels at the individual level, treating each individual as their own sample. These individuals were then analyzed independently before combining them to obtain population level estimates.

Table 4.1: Summary of parameters included cumulatively in each of the models

Model	Parameter	Definition
1	y^*_{iehlt}	Latent sensitivity estimate
	μ_{α_0}	Population-averaged intercept
	μ_{α_1}	Population-averaged slope
	α_{0i}	Individual-specific intercept
	α_{1i}	Individual-specific slope
	γ_{0ie}	Deviation due to the eye (intercept)
	γ_{1ie}	Deviation due to the eye (slope)
	η_{0ieh}	Deviation due to the hemifield (intercept)
	η_{1ieh}	Deviation due to the hemifield (slope)
	λ_{0iehl}	Deviation due to the location (intercept)
	λ_{1iehl}	Deviation due to the location (slope)
	σ^2_ε	Variance of the measurement error
	Σ_α	Covariance matrix for the individual level
	Σ_γ	Covariance matrix for the eye level
	Σ_η	Covariance matrix for the hemifield level
	Σ_λ	Covariance matrix for the location level
2	ϕ_{iet}	Global visit effect
	σ^2_ϕ	Global Visit Effect variance
3	μ_{ζ_0}	Intercept in logarithm of the standard deviation
	μ_{ζ_1}	Slope in logarithm of the standard deviation
	$\sigma^2_{\varepsilon,iehlt}$	Variance of the measurement error

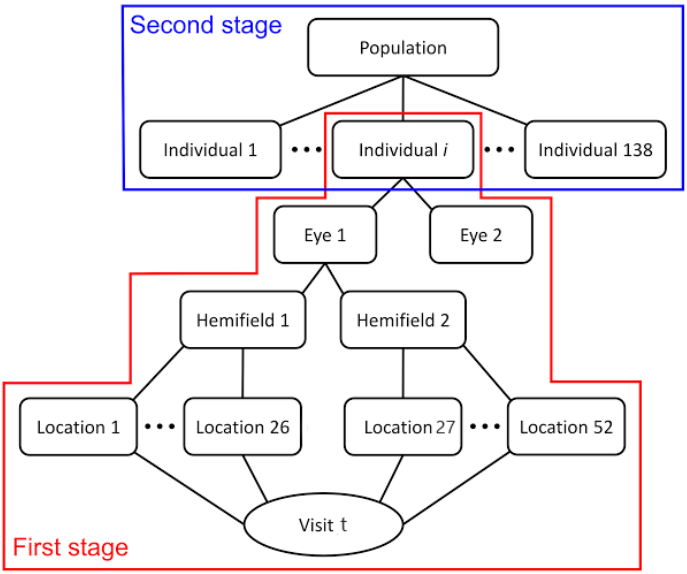

Figure 4.4: Illustration of the hierarchical structure of the data divided into the first and second stages as done in the two-stage approach.

4.4.2.1 First stage

In the first stage, we analyzed each individual separately. This is now shown for Model 3. The first stage model is given by:

$$\begin{aligned} y^*_{iehlt} &= \alpha_{0i} + \alpha_{1i}\text{time}_{it} + \gamma_{0ie} + \gamma_{1ie}\text{time}_{it} + \eta_{0ieh} + \eta_{1ieh}\text{time}_{it} + \\ &\quad \lambda_{0iehl} + \lambda_{1iehl}\text{time}_{it} + \phi_{iet} + \varepsilon_{iehlt}, \end{aligned} \tag{4.4.4}$$

where $\varepsilon_{iehlt} \sim N(0, \sigma^2_{\varepsilon,iehlt})$ and

$$\log(\sigma_{\varepsilon,iehlt}) = \zeta_{0i} + \zeta_{1i}\mu^{(2)}_{iehlt},$$

and

$$\begin{aligned}
\alpha_{bi} &\sim N(0, 10^9) \text{ for } b = 0, 1; \\
\gamma_{ie} &= \binom{\gamma_{0ie}}{\gamma_{1ie}} \sim N(\binom{0}{0}, \Sigma_{\gamma_i}); \\
\eta_{ieh} &= \binom{\eta_{0ieh}}{\eta_{1ieh}} \sim N(\binom{0}{0}, \Sigma_{\eta_i}); \\
\lambda_{iehl} &= \begin{pmatrix} \lambda_{0iehl} \\ \lambda_{1iehl} \end{pmatrix} \sim N(\binom{0}{0}, \Sigma_{\lambda_i}); \\
\phi_{iet} &\sim t(0, \sigma^2_{\phi_i}, 3); \\
\sigma^2_{\phi_i} &\sim \text{IG}(10^{-3}, 10^{-3}); \\
\zeta_{bi} &\sim N(0, 10^9) \text{ for } b = 0, 1.
\end{aligned}$$

Note that vague and independent priors are specified for α_{0i}, α_{1i}, because they are treated as fixed effects in the first stage.

One important detail about the two-stage approach is that it allows the individual variances to differ, i.e. Σ_{γ_i}, Σ_{η_i}, Σ_{λ_i} and $\sigma^2_{\phi_i}$, but also $\sigma^2_{\varepsilon,iehlt}$ since the regression coefficients are now allowed to depend on the subject. In order to prevent the second-stage sampler from getting stuck near local posterior modes, large independent samples are needed from this first stage (Lunn et al., 2013). To achieve this, we ran 200,000 iterations with a burn-in of 150,000 iterations. Using a thinning factor of 10, this resulted in 5,000 stored iterations for each parameter for each individual.

4.4.2.2 Second stage

At the end of the first stage we have the following sampled values for the parameters split up according to whether they will be directly used in the second stage (parameters of interest) or not (nuisance parameters):

- Parameters of interest:
  $\{\alpha_{0i}, \alpha_{1i}, \zeta_{0i}, \zeta_{1i}, \Sigma_{\gamma_i}, \Sigma_{\eta_i}, \Sigma_{\lambda_i}, \sigma^2_{\phi_i}\}$ for $i = 1, \ldots, N$;

- Nuisance parameters:
  $\mathscr{N}_i = \{\gamma_{0i1}, \gamma_{1i1}, \gamma_{0i2}, \gamma_{1i2}, \eta_{0i1,1,}, \ldots, \eta_{1i2,2,}, \lambda_{0i1,1,1,}, \ldots, \lambda_{1i2,2,26}, \phi_{i1,1,}, \ldots, \phi_{i1,T_i}, \phi_{i2,1,}$
  $\ldots, \phi_{i2,T_i}\}$ for $i = 1, \ldots, N$.

The total set of nuisance parameters is then denoted as $\mathscr{N} = \{\mathscr{N}_1, \ldots, \mathscr{N}_N\}$. The split up into parameters of interest and nuisance parameters is motivated as follows. The $\alpha_i = (\alpha_{0i}, \alpha_{1i})^T$ and $\zeta_i = (\zeta_{0i}, \zeta_{1i})^T$ are transferred to the second stage, because we wish to estimate their population-averaged effects assuming a common (normal) distribution with means equal to $\mu_\alpha = (\mu_{\alpha_0}, \mu_{\alpha_1})^T$ and $\mu_\zeta = (\mu_{\zeta_0}, \mu_{\zeta_1})^T$, respectively. Since we decided to split the data at the

individual level for computational reasons, all lower-level random effects are automatically considered to be nuisance. In the case of a meta-analysis, the random effects in the first stage may not be of direct interest. Hence, these terms can then be treated as nuisance parameters as done by Lunn et al. (2013). However, for clinical applications such as ours, these may be important. So, we kept the covariance matrices of those lower-level random effects as input for the second stage to allow us to re-estimate these random effects at a later stage. A Cholesky decomposition of the covariance matrices of the random effects (Σ_{γ_i}, Σ_{η_i} and Σ_{λ_i}) was computed, since each of the entries in the covariance matrices were treated as separate parameters in the second stage. For example, for Σ_{γ_i} the 2×2 full rank lower triangular matrix

$$C_{\gamma_i} = \begin{pmatrix} c_{1\gamma_i} & 0 \\ c_{2\gamma_i} & c_{3\gamma_i} \end{pmatrix},$$

with $c_{1\gamma_i}, c_{3\gamma_i}$ positive diagonal entries such that $C_{\gamma_i} C_{\gamma_i}^T = \Sigma_{\gamma_i}$. We denote the Cholesky decomposition factors for all of the covariance matrices as $\{c_{1\gamma_i}, c_{2\gamma_i}, c_{3\gamma_i}, c_{1\eta_i}, c_{2\eta_i}, c_{3\eta_i}, c_{1\lambda_i}, c_{2\lambda_i}, c_{3\lambda_i}\}$. Hence, the final set of parameters of interest that are directly processed in the second stage are $\{\alpha_i, \zeta_i, \sigma^2_{\phi_i}, c_{1\gamma_i}, c_{2\gamma_i}, c_{3\gamma_i}, c_{1\eta_i}, c_{2\eta_i}, c_{3\eta_i}, c_{1\lambda_i}, c_{2\lambda_i}, c_{3\lambda_i}\}$ for $i = 1, \ldots, N$.

In the second stage we assume that the above parameters of interest share a common distribution. More specifically, we assumed here that $\alpha_i \sim N((\mu_{\alpha_0}, \mu_{\alpha_1})^T, \Sigma_\alpha)$, $\sigma^2_{\phi_i} \sim N(\sigma^2_\phi, \sigma^2_{\sigma^2_\phi})$. However, for ζ_{0i}, ζ_{1i} and all Cholesky decomposition parameters we assumed independent normal distributions, because of computational and memory restrictions. We refer to Section 4.8.2.3 for the priors. The posterior distributions of the aforementioned parameters were then obtained by using their sampled values from the first stage as proposal distributions in a Metropolis-Hastings step. Further details about the two-stage approach are found in (Lunn et al., 2013).

The total set of parameters involved in the second stage is denoted as $\mathscr{P}$ and is given by:

$$\begin{aligned} \mathscr{P} \quad = \quad & \{\alpha_{01}, \alpha_{11}, \ldots, \alpha_{0N}, \alpha_{1N}, \mu_{\alpha_0}, \mu_{\alpha_1}, \Sigma_\alpha, \zeta_{01}, \zeta_{11}, \ldots, \zeta_{0N}, \zeta_{1N}, \mu_{\zeta_0}, \mu_{\zeta_1}, \sigma^2_{\zeta_0}, \sigma^2_{\zeta_1}, \sigma^2_\phi, \\ & \sigma^2_{\sigma^2_\phi}, \mu_{c_{1\gamma}}, \mu_{c_{2\gamma}}, \mu_{c_{3\gamma}}, \mu_{c_{1\eta}}, \mu_{c_{2\eta}}, \mu_{c_{3\eta}}, \mu_{c_{1\lambda}}, \mu_{c_{2\lambda}}, \mu_{c_{3\lambda}}, \sigma^2_{c_{1\gamma}}, \sigma^2_{c_{2\gamma}}, \sigma^2_{c_{3\gamma}}, \sigma^2_{c_{1\eta}}, \sigma^2_{c_{2\eta}}, \sigma^2_{c_{3\eta}}, \\ & \sigma^2_{c_{1\lambda}}, \sigma^2_{c_{2\lambda}}, \sigma^2_{c_{3\lambda}}\}, \end{aligned}$$

where $\{\mu_{c_{1\gamma}}, \mu_{c_{2\gamma}}, \mu_{c_{3\gamma}}, \mu_{c_{1\eta}}, \mu_{c_{2\eta}}, \mu_{c_{3\eta}}, \mu_{c_{1\lambda}}, \mu_{c_{2\lambda}}, \mu_{c_{3\lambda}}\}$ are defined in Section 4.8.2.3.

Three chains were initialized with different starting values for all models determined from the first-stage samples. This was done using the minimum, mean and maximum sampled value for each parameter for every individual. Upon convergence, we computed the posterior mean, median, standard deviation with the equal-tailed 95% credible interval (CI) for all parameters of interest.

4.5 Model Evaluation

Classical approaches to model evaluation are applicable to the results of the first stage, since this stage represents a standard analysis. Hence, we evaluated the models at this stage for each individual separately by using posterior predictive checks (PPC), such as the χ^2 test statistic. We further compared the models after the second stage by using the Deviance Information Criterion (DIC) to determine the overall best model.

4.5.1 Deviance Information Criterion

In a Bayesian framework, a common tool for model evaluation is the DIC proposed by Spiegelhalter et al. (2002) (Spiegelhalter et al., 2002). The DIC is based on a contrast between the deviance at the posterior mean of the parameters and the posterior mean deviance. DIC is defined here as

$$DIC = D(\bar{\mathscr{P}}, \bar{\mathscr{N}}) + 2p_D = \overline{D(\mathscr{P}, \mathscr{N})} + p_D, \tag{4.5.5}$$

with

$$p_D = \overline{D(\mathscr{P}, \mathscr{N})} - D(\bar{\mathscr{P}}, \bar{\mathscr{N}}),$$

where the deviance is denoted as $D(\mathscr{P}, \mathscr{N})$, the posterior expectation of the deviance is denoted as $\overline{D(\mathscr{P}, \mathscr{N})}$ and the posterior means of the parameters $\mathscr{P}$ and $\mathscr{N}$ as $\bar{\mathscr{P}}$ and $\bar{\mathscr{N}}$, respectively.

The definition of the DIC includes the parameters of interest $\mathscr{P}$, which we obtained from the second stage, as well as the random effects $\mathscr{N}$, which are treated as nuisance parameters in the two-stage approach. In practice, DIC is usually computed conditional on the random effects (so-called conditional DIC, see e.g. (Lesaffre and Lawson, 2012)). However, the two-stage approach does not automatically provide estimates for all random effects under the one-stage model.

To address this problem, we propose to complement the two-stage approach with an additional step based on the Method of Composition (Lesaffre and Lawson, 2012) in combination with a Metropolis-within-Gibbs technique. To illustrate our proposed method, we refer again to Model 3. Note that for the calculation of DICs, we need the posterior distribution $p(\mathscr{N} \mid y)$, with y the total set of observed VF data. The required posterior of the random effects can be obtained from the posterior distribution $p(\mathscr{P} \mid y)$ and $p(\mathscr{N} \mid \mathscr{P}, y)$ as

$$p(\mathscr{N} \mid y) = \int p(\mathscr{N} \mid \mathscr{P}, y) p(\mathscr{P} \mid y) d\mathscr{P}. \tag{4.5.6}$$

In practice the integral is replaced by a Monte Carlo estimate, by sampling each element of $\mathscr{N}$ given y and the sampled value of $\mathscr{P}$ and then ignoring the sampled value $\mathscr{P}$. This is in fact an example of the Method of Composition.

Let the vector of all responses for the individual i and eye e be denoted by y_{ie}. Furthermore, let $\mathscr{N}_{ie}$ represent the subset of nuisance parameters in $\mathscr{N}$, for individual i and eye e. For each eye e of each individual i, the computations of our proposed method for sampling the random effects in $\mathscr{N}_{ie}$ consist of the following steps:

Step 1: We use the posterior results of $K = 500$ consecutive iterations of the second stage of the two-stage approach. For iteration k (with $k = 1,\ldots,K$), let $\tilde{\mathscr{P}}^{(k)}$ denote the sampled values of the parameters of interest ($\mathscr{P}$) estimated in the second stage. Conditional on $\tilde{\mathscr{P}}^{(k)}$, we sample the random effects in $\mathscr{N}_{ie}$ (for the eye, hemifield, location and GVE), which were treated as nuisance parameters in the second stage, i.e.

$$\gamma_{0ie}, \gamma_{1ie}, \eta_{0ieh}, \eta_{1ieh}, \lambda_{0iehl}, \lambda_{1iehl} \text{ and } \phi_{iet},$$

and we denote the sampled values as $\tilde{\mathscr{N}}_{ie}^{(k)}$. For each iteration k, the sampling of the random effects is done with the Metropolis-within-Gibbs technique as described in Steps 2A and 2B below.

Step 2A: For $k = 1$ we need to choose initial values for the random effects in $\tilde{\mathscr{N}}_{ie}^{(1)}$. To decrease computational time, the posterior modes of all parameters in $\mathscr{N}_{ie}$ were determined by means of an optimisation routine, and these were used as initial values for $\tilde{\mathscr{N}}_{ie}^{(1)}$ in Step 2B. The optimisation was performed using the `optim()` function in R with a quasi-Newton algorithm. The function which was maximised with respect to $\mathscr{N}_{ie}$ was the logarithm of $p(\mathscr{N}_{ie} \mid \bar{\mathscr{P}}, y_{ie}) \propto p(y_{ie} \mid \bar{\mathscr{P}}, \mathscr{N}_{ie})\, p(\mathscr{N}_{ie} \mid \bar{\mathscr{P}})$. Note that the optimisation procedure is only done for $k = 1$.

Step 2B: A random-walk Metropolis-Hastings algorithm is used to sample the random effects iteratively for each level, to take into account the correlation between the levels, yielding a Metropolis-within-Gibbs sampler. For example, first a sampled value for the eye is obtained given the initial values for the levels lower than the eye (i.e. the hemifield, the location and the GVE). This sampled value is then used in combination with the initial values for the levels lower than the hemifield (i.e. the location and the GVE) to obtain an estimate for the hemifield random effects, and so forth. This Metropolis-within-Gibbs sampler is run for $s = 1,\ldots,S$ iterations, with $S = 200$, and the random effects obtained in the last iteration S are then used as the sampled random effects $\tilde{\mathscr{N}}_{ie}^{(k)}$ in Step 1. The first $S-1$ iterations thus only serve as burn-in samples. For $k = 1$, the estimates obtained in Step 2A are used as initial values (for $s = 1$). For $k = 2,\ldots,K$, the sampled values of the previous iteration of Step 1, i.e. $\tilde{\mathscr{N}}_{ie}^{(k-1)}$ are used as initial values.

In the Metropolis-within-Gibbs sampler, the random effects are sampled for each iteration s (with $s = 1,\ldots,S$) as follows:

1. The random effects for the eye level ($\gamma_{ie}^{(s)}$) are sampled by means of the

Metropolis-Hastings algorithm with target density proportional to

$$\prod_h \prod_l \prod_t p(y_{iehlt} \mid \gamma_{ie}, \eta_{ieh}^{(s-1)}, \lambda_{iehl}^{(s-1)}, \phi_{iet}^{(s-1)}, \tilde{\mathscr{P}}^{(k)}) \times p(\gamma_{ie} \mid \tilde{\mathscr{P}}^{(k)}).$$

2. For hemifield $h = 1,2$, the random effects for the hemifield level ($\eta_{ieh}^{(s)}$) are sampled by means of the Metropolis-Hastings algorithm with target density proportional to

$$\prod_l \prod_t p(y_{iehlt} \mid \gamma_{ie}^{(s)}, \eta_{ieh}, \lambda_{iehl}^{(s-1)}, \phi_{iet}^{(s-1)}, \tilde{\mathscr{P}}^{(k)}) \times p(\eta_{ieh} \mid \tilde{\mathscr{P}}^{(k)}).$$

3. For hemifield $h = 1,2$ and location $l = 1,\ldots,26$, the random effects for the location level ($\lambda_{iehl}^{(s)}$) are sampled by means of the Metropolis-Hastings algorithm with target density proportional to

$$\prod_t p(y_{iehlt} \mid \gamma_{ie}^{(s)}, \eta_{ieh}^{(s)}, \lambda_{iehl}, \phi_{iet}^{(s-1)}, \tilde{\mathscr{P}}^{(k)}) \times p(\lambda_{iehl} \mid \tilde{\mathscr{P}}^{(k)}).$$

4. For visit $t = 1,\ldots,T_i$, the random effects for the GVE ($\phi_{iet}^{(s)}$) are sampled by means of the Metropolis-Hastings algorithm with target density proportional to

$$\prod_h \prod_l p(y_{iehlt} \mid \gamma_{ie}^{(s)}, \eta_{ieh}^{(s)}, \lambda_{iehl}^{(s)}, \phi_{iet}, \tilde{\mathscr{P}}^{(k)}) \times p(\phi_{iet} \mid \tilde{\mathscr{P}}^{(k)}).$$

We let $\tilde{\mathscr{N}}^{(k)}$ denote all of the sampled values of $\mathscr{N}$, i.e. the combined sampled values, $\tilde{\mathscr{N}}_{ie}^{(k)}$, for both eyes from all individuals. Combining the resulting samples of $\mathscr{N}$ with the samples of $\mathscr{P}$ from the second stage yields a set of $K = 500$ posterior samples of the parameters and random effects. These posterior samples are then used to calculate the deviance in iteration k as

$$D(\mathscr{P},\mathscr{N})^{(k)} = -2\sum_{i=1}^{N}\sum_{e=1}^{2}\sum_{h=1}^{2}\sum_{l=1}^{26}\sum_{t=1}^{T_i} \log(p(y_{iehlt} \mid \tilde{\mathscr{N}}^{(k)}, \tilde{\mathscr{P}}^{(k)})),$$

and finally the DIC is calculated from these K deviance values.

4.5.2 Posterior predictive check

Let the vector of all responses for the ith individual be denoted by y_i. Furthermore, we denote all parameters for individual i from the first stage by ψ_i. Then suppose that $\psi_i^1,\ldots,\psi_i^K$ are samples from the converged Markov chain from $p(\psi_i \mid y_i)$. The posterior predictive p-value (PPP) for a discrepancy measure $D_G(y_i \mid \psi_i^k)$ is obtained by contrasting this value to $D_G(\tilde{y}_i^k, \psi_i^k)$ for k in $\{1,\ldots,K\}$, whereby the replicated data $\tilde{y}_i^k$ are sampled from $p(y_i \mid \psi_i^k)$. The following p-value can then be calculated for the i^{th} subject

$$PPP = \frac{1}{K}\sum_{k=1}^{K} I[D_G(\tilde{y}_i^k, \psi_i^k) \leq D_G(y_i, \psi_i^k)]. \tag{4.5.7}$$

Here, the Gelman χ^2-test statistic (Gelman et al., 2013) was used as the discrepancy measure to calculate the PPC for each individual. This is defined as:

$$D_G(y_i, \psi_i^k) = \sum_{e=1}^{2}\sum_{h=1}^{2}\sum_{l=1}^{26}\sum_{t=1}^{T_i} \frac{[y_{iehlt} - E(y_{iehlt} \mid \psi_i^k)]^2}{\text{var}(y_{iehlt} \mid \psi_i^k)}, \tag{4.5.8}$$

where a small value indicates a good model fit. When the PPP is small, say smaller than 0.05, the replicated data fit the assumed model importantly better than the observed data, and hence we conclude in that case that our model does not fit the data well. In this hierarchical structure, such a p-value can be computed for each individual. In case of a good model fit, we expect that the distribution of p-values has a uniform distribution. Now because the data is used twice: once for model fit and once for model evaluation (see e.g. Lesaffre and Lawson (2012)) we will not obtain a uniform distribution for a good model fit, but rather a distribution more clustered around 0.5. For the same reason, the approach does not provide an absolute guarantee of a good model fit. But all improvements of the basic approach, see e.g. (Lesaffre and Lawson, 2012), are computationally too time consuming here.

4.6 Application to the Glaucoma study

For this analysis we included both eyes from the 139 individuals belonging to the Glaucoma study. After excluding VFs with unknown reliability as indicated by the instrument, 138 individuals; 276 eyes; 552 hemifields and 14,352 locations remained. This included 4,760 VFs, resulting in a data set consisting of 247,520 location-specific sensitivity estimates. All analyses were done taking into account censoring.

4.6.1 Results

The two-stage approach is advantageous, as it allows us to do exploratory analyses at the individual level in order to simplify the model before combining the samples in the second stage. An example of the model fits for one location is shown in Figure 4.5. The posterior summary statistics from the second stage are listed in Table 4.2 for each of the models. A lower DIC indicates a preferable model. Using DIC, Model 2 (DIC=1,173,872) performed better than Model 1 (DIC = 1,211,653), with Model 3 (DIC = 1,075,212 performing the best overall. Using the results from Model 3, the population intercept (μ_{α_0}) was 19.82 dB with an average slope (μ_{α_1}) of -0.31 dB per year. The intercept (μ_{ζ_0}) and slope (μ_{ζ_1}) for the logarithm

of the standard deviation was 2.82 dB and -0.08 dB respectively. This corresponds to the 2.60 dB and -0.06 dB which was found in the exploratory analysis shown in Figure 4.3.

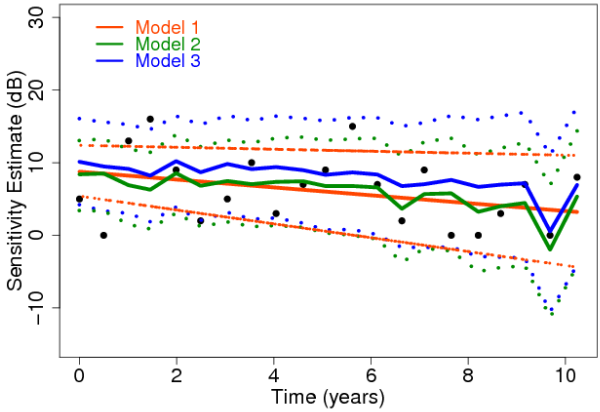

Figure 4.5: Scatter plot representing the retinal sensitivity estimates over time for 1 location of the VF. The lines represent the model fits for each of the 3 models with 95% CIs.

To evaluate the fit of the chosen model(s) we can use the PPC technique. More specifically, we could compute the PPP-value for each individual and then graphically evaluate the distribution of p-values. Figure 4.6 shows the ordered PPP-values for each of the models. Model 1 has a mean $PPP = 0.30$, Model 2 a mean $PPP = 0.30$ and Model 3 a mean $PPP = 0.50$. From this, it appears that Model 3 has the best fit. That the distribution of p-values should deviate from the uniform distribution for a good model was explained above. In fact, we indeed see that for Model 3 the p-values are more clustered around 0.5 than what would be expected if the true P-value had a uniform distribution. For the other two models the distribution of p-values is systematically shifted to the left.

4.6.2 Comparing the models: some reflections

Our model takes into account the hierarchical structure of the data by means of a Bayesian hierarchical model. This allows us to take into account the correlation at each of the 4 levels. The estimated covariance matrices for each level were similar for all models. A large proportion of intercept variability was explained by the individual level random effects

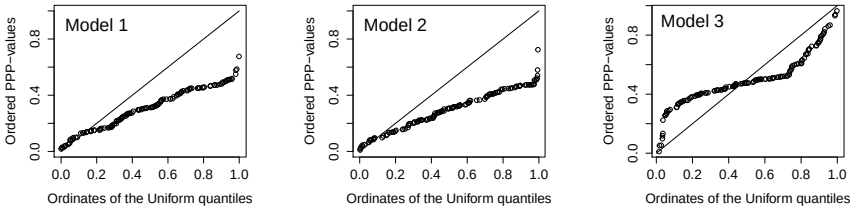

Figure 4.6: Posterior predictive check for each of the models across all individuals

Table 4.2: Posterior summary statistics for the three models by means of the two-stage approach

Model	1			2			3		
Parameter	mean	sd	95% CI	mean	sd	95% CI	mean	sd	95% CI
μ_{α_0}	18.97	0.73	(17.54 ; 20.39)	20.30	0.74	(18.86 ; 21.76)	19.89	0.77	(18.37 ; 21.39)
μ_{α_1}	−0.23	0.05	(−0.33 ; −0.12)	−0.20	0.05	(−0.30 ; −0.11)	−0.30	0.05	(−0.40 ; 0.20)
μ_{ζ_0}							2.82	0.06	(2.69 ; 2.95)
μ_{ζ_1}							-0.08	0.00	(−0.08 ; −0.07)
$\sigma^2_{\varepsilon,iehlt}$	13.46	0.66	(12.20 ; 14.81)	11.53	0.57	(10.45 ; 12.70)			
σ^2_{ϕ}				1.86	0.15	(1.58 ; 2.18)	5.64	0.12	(5.29 ; 5.91)
Σ_α	$\begin{pmatrix} 63.79 & 0.06 \\ 0.06 & 0.20 \end{pmatrix}$			$\begin{pmatrix} 71.11 & -1.73 \\ -1.73 & 0.16 \end{pmatrix}$			$\begin{pmatrix} 76.10 & -0.14 \\ -0.14 & 0.17 \end{pmatrix}$		
Σ_γ	$\begin{pmatrix} 0.44 & -0.02 \\ -0.02 & 0.23 \end{pmatrix}$			$\begin{pmatrix} 0.45 & -0.01 \\ -0.01 & 0.27 \end{pmatrix}$			$\begin{pmatrix} 0.00 & 0.01 \\ 0.01 & 0.96 \end{pmatrix}$		
Σ_η	$\begin{pmatrix} 26.85 & 0.36 \\ 0.36 & 0.23 \end{pmatrix}$			$\begin{pmatrix} 24.73 & 0.19 \\ 0.19 & 0.23 \end{pmatrix}$			$\begin{pmatrix} 25.50 & 0.29 \\ 0.29 & 0.23 \end{pmatrix}$		
Σ_λ	$\begin{pmatrix} 41.41 & 0.40 \\ 0.40 & 0.13 \end{pmatrix}$			$\begin{pmatrix} 40.33 & -0.19 \\ -0.19 & 0.14 \end{pmatrix}$			$\begin{pmatrix} 40.13 & 0.11 \\ 0.11 & 0.09 \end{pmatrix}$		
DIC	1,211,653			1,173,872			1,075,212		

(Model 3: $\Sigma_{\alpha 11} = 76.10$). A small amount of this intercept variability was explained by the eye level random effects (Model 3: $\Sigma_{\gamma 11} = 0.0007$). This is due to the fact that each individual only has 2 eyes and with this approach we assume that only the eyes belonging to the same individual share a common variance distribution. This variability increases for the 4 hemifields per individual (Model 3: $\Sigma_{\eta 11} = 25.50$) and for the 104 locations per individual (Model 3: $\Sigma_{\lambda 11} = 40.13$). The variance of the slope for the hemifield (Model 3: $\Sigma_{\eta 22} = 0.23$) seems to be large compared to the variance of the slope for the location (Model 3: $\Sigma_{\lambda 22} = 0.09$). This shows that locations within a hemifield are correlated with

respect to their change. This is expected, since the nerves within each hemifield belong to the same nerve bundle. Hence, it is likely that there is some correlation between nerves, and hence locations, which are affected by glaucomatous damage. It is important to note that all individuals received therapy throughout the study. The range of slopes may be reduced with therapy, because the intra-ocular pressure (IOP) range is reduced between eyes to somewhat similar levels. Without therapy the individual's IOP would be between 15 and 40 mmHg. With therapy this could be reduced to between 12 and 24 mmHg. Since all individuals in this study are being treated, the reduction in the IOP range could explain the small range of slopes of progression.

We proposed to model measurement errors that affect the point-wise sensitivity estimates within the same VF as GVEs. By correcting for the GVEs, we account for measurable factors, such as season, time of day and the reliability indices, as well as those factors which cannot be measured such as fatigue and delayed reaction time. Including the GVE showed a considerable improvement in the model fit. Hence, by taking into account the GVE we were able to take into account a large part of the variability and obtain better estimates of the true rate of progression. Implementing progression models that incorporate the GVE in clinical care may therefore improve the clinical management of glaucoma. A further evaluation of the GVE by determining the improvement of the model fit due to incorporating the GVE in the model and the effect of including the GVE on estimating the rate of progression has been done in previous work. Furthermore, the improvement of point-wise predictions for future measurements accounting for the GVE has been shown (Bryan et al., 2015).

Including the relationship between variability and sensitivity showed a further improvement in the model fit. The function which describes this relationship was consistent to that found by Henson et al. (2000), however it was not shown previously how to include this relationship in a model, or whether including it would improve the model fit. Other studies have emphasized the importance of this relationship and questioned the validity of assuming a constant variance over the entire range of sensitivity estimates (Russell et al., 2012). Gardiner et al. (2014) showed that clinical VF testing may be unreliable when VF locations have sensitivity estimates below approximately 15-19 dB. By modeling this relationship, we are able to account for the lower sensitivity estimates being less reliable and hence allow the precision throughout the range of sensitivity estimates to differ accordingly.

In clinical practice, it is often difficult to determine which patients progress rapidly towards severe visual impairment and which ones are relatively stable, because of measurement variability. With our current approach, the true rate of progression is much better determined, so that clinicians can specifically target those with rapid progression, thereby significantly reducing the risk of severe visual impairment or blindness.

4.7 Discussion

In this paper, we proposed a method to model point-wise VFs taking into account the complexity of psychophysical testing of visual function in glaucoma. The model is advantageous in dealing with the high measurement variability, and could be extended for the prediction of future VFs. Although it was possible to use the one-stage approach with simplified versions of the model or with smaller datasets, it was not possible to perform these analysis on the full data with a complex model as it was with the two-stage approach. The two-stage approach can be implemented in standard MCMC software. The relevant computations for the first-stage can be carried out in JAGS (Plummer, 2003), WinBUGS or OpenBUGS (Lunn et al., 2009) software and the second-stage carried out in OpenBUGS software (Lunn et al., 2013). However, for the second stage an add-on program is needed. For more details on setting up OpenBUGS for performing the two-stage analyses, we refer to (Lunn et al., 2013). More information regarding the computations done in this paper can be obtained by emailing the first author. These computations can be easily tuned to adapt to other data sets by any practitioner.

The two-stage method is advantageous as it allowed us to do exploratory analysis at an individual level. Hence, we are able to simplify and improve the model before combining it at a population level. Limited simulations showed that the one- and two-stage approaches gave similar results if the variances were the same for all individuals. The two-stage approach assumes a more flexible method. However, there is the additional difficulty in constraining the parameters across individuals. One disadvantage of this approach is that it does not provide the required components to evaluate the fit and predictive ability of the model by means of the Deviance Information Criterion (DIC). In order to calculate the DIC and compare different competing models for our data fitted by using the two-stage approach, we proposed a Monte Carlo scheme based on a Metropolis-within-Gibbs algorithm.

Other issues, which we see as future research directions, is to look at the optimal choice of the level where the data should be split. Extensions include exploiting the spatial nature of the data and capitalizing on the specific spatial organization of the nerve fibres in the eye (Erler et al., 2014).

Acknowledgements

The authors would like to thank Dr. Baoyue Li for his useful discussions which helped with the early stages of this work. The authors would also like to thank Stichting Wetenschappelijk Onderzoek het Oogziekenhuis, Stichting voor Ooglijders and Stichting Glaucoomfonds for the financial support. Furthermore, the authors would like to thank the editor and the referees for their helpful comments and suggestions which drastically improved the quality of the manuscript.

Bibliography

Anderson, D. R. and Patella, V. M. (1999). *Automated Static Perimetry*. Mosby, 2nd edition.

Artes, P. H., Nicolela, M. T., LeBlanc, R. P., and Chauhan, B. C. (2005). Visual field progression in glaucoma: total versus pattern deviation analyses. *Investigative Ophthalmology & Visual Science*, 46(12):4600.

Bengtsson, B. and Heijl, A. (2000). False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? *Investigative Ophthalmology & Visual Science*, 41(8):2201–2204.

Bengtsson, B. and Heijl, A. (2008). A visual field index for calculation of glaucoma rate of progression. *American Journal of Ophthalmology*, 145(2):343–353.

Bryan, S. R., Vermeer, K. A., Eilers, P. H. C., Lemij, H. G., and Lesaffre, E. M. E. H. (2013). Robust and censored modeling and prediction of progression in glaucomatous visual fields. *Investigative Ophthalmology & Visual Science*, 54(10):6694–6700.

Caprioli, J., Mock, D., Bitrian, E., Afifi, A. A., Yu, F., Nouri-Mahdavi, K., and Coleman, A. L. (2011). A method to measure and predict rates of regional visual field decay in glaucoma. *Investigative Ophthalmology & Visual Science*, 52(7):4765–4773.

Erler, N. S., Bryan, S. R., Eilers, P. H. C., Lesaffre, E. M. E. H., Lemij, H. G., and Vermeer, K. A. (2014). Optimizing structure-function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Ophthalmology & Visual Science*, 55(4):2350–2357.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3rd edition.

Henson, D. B., Chaudry, S., Artes, P. H., Faragher, E. B., and Ansons, A. (2000). Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Investigative Ophthalmology & Visual Science*, 41(2):417–421.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hudson, C., Wild, J. M., and O'Neill, E. C. (1994). Fatigue effects during a single session of automated static threshold perimetry. *Investigative Ophthalmology & Visual Science*, 35(1):268–280.

Junoy Montolio, F. G., Wesselink, C., Gordijn, M., and Jansonius, N. M. (2012). Factors that influence standard automated perimetry test results in glaucoma: test reliability, technician experience, time of day, and season. *Investigative Ophthalmology & Visual Science*, 53(11):7010–7017.

Kingman, S. (2004). Glaucoma is second leading cause of blindness globally. *Bulletin of the World Health Organization*, 82(11):887–888.

Kymes, S. M., Lambert, D. L., Lee, P. P., Musch, D. C., Siegfried, C. J., Kotak, S. V., Stwalley, D. L., Fain, J., Johnson, C., and Gordon, M. O. (2012). The development of a decision analytic model of changes in mean deviation in people with glaucoma: the COA model. *Ophthalmology*, 119(7):1367–1374.

Lee, K. J. and Thompson, S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine*, 27(3):418–434.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. Wiley & Sons, Chichester, West Sussex.

Lunn, D., Barrett, J., Sweeting, M., and Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 62(4):551–572.

Lunn, D., Spiegerhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

McNaught, A. I., Crabb, D. P., Fitzke, F. W., and Hitchings, R. A. (1995). Modelling series of visual fields to detect progression in normal-tension glaucoma. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 233(12):750–755.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley & Sons, Hoboken, N.J.

Pathak, M., Demirel, S., and Gardiner, S. K. (2013). Nonlinear, multilevel mixed-effects approach for modeling longitudinal standard automated perimetry data in glaucoma. *Investigative Ophthalmology & Visual Science*, 54(8):5505.

Plummer, M. (2003). *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling*.

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

Russell, R. A., Crabb, D. P., Malik, R., and Garway-Heath, D. F. (2012). The relationship between variability and sensitivity in large-scale longitudinal visual field data. *Investigative Ophthalmology & Visual Science*, 53(10):5985–5990.

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

Swanson, W. H., Sun, H., Lee, B. B., and Cao, D. (2011). Responses of primate retinal ganglion cells to perimetric stimuli. *Investigative Ophthalmology & Visual Science*, 52(2):764–771.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer New York, New York, NY.

Zhu, H., Russell, R. A., Saunders, L. J., Ceccon, S., Garway-Heath, D. F., and Crabb, D. P. (2014). Detecting changes in retinal function: Analysis with non-stationary Weibull error regression and spatial enhancement (ANSWERS). *PLoS ONE*, 9(1):e85654.

4.8 Appendix A: Model Specifications

In this section, the model specifications for Model 3 are given for the one-stage and two-stage approach.

4.8.1 One-Stage Approach

4.8.1.1 Full Model

$$\begin{aligned} y^*_{iehlt} &= \alpha_{0i} + \alpha_{1i}\text{time}_{it} + \gamma_{0ie} + \gamma_{1ie}\text{time}_{it} + \\ &\quad \eta_{0ieh} + \eta_{1ieh}\text{time}_{it} + \lambda_{0iehl} + \lambda_{1iehl}\text{time}_{ij} + \phi_{iet} + \varepsilon_{iehlt} \end{aligned} \tag{4.8.1}$$

where $\varepsilon_{iehlt} \sim N(0, \sigma^2_{\varepsilon,iehlt})$ and $\log(\sigma_{\varepsilon,iehlt}) = \mu_{\zeta_0} + \mu_{\zeta_1}\mu^{(2)}_{iehlt}$.

4.8.1.2 Priors

In the Bayesian procedure prior distributions need to be stipulated for all parameters. When no prior information is available then the prior distribution should reflect this. In this case a vague prior is a natural choice.

$$\begin{aligned} \alpha_i &= \binom{\alpha_{0i}}{\alpha_{1i}} \sim N\left(\binom{\mu_{\alpha_0}}{\mu_{\alpha_1}}, \Sigma_\alpha\right) \\ \gamma_{ie} &= \binom{\gamma_{0ie}}{\gamma_{1ie}} \sim N(\binom{0}{0}, \Sigma_\gamma) \\ \eta_{ieh} &= \binom{\eta_{0ieh}}{\eta_{1ieh}} \sim N(\binom{0}{0}, \Sigma_\eta) \\ \lambda_{iehl} &= \binom{\lambda_{0iehl}}{\lambda_{1iehl}} \sim N(\binom{0}{0}, \Sigma_\lambda) \\ \phi_{iet} &\sim t(0, \sigma^2_\phi, 3) \\ \mu_{\zeta_b} &\sim N(0, 10^9) \text{ for } b = 0, 1 \\ \mu_{\alpha_b} &\sim N(0, 10^9) \text{ for } b = 0, 1 \\ \sigma^2_\phi &\sim IG(10^{-3}, 10^{-3}) \text{ and} \\ \Sigma_\alpha, \Sigma_\gamma, \Sigma_\eta \text{ and } \Sigma_\lambda &\sim \text{IW}(\text{diag}(1,1), 2). \end{aligned}$$

4.8.2 Two-Stage Approach

The full model is the same for both the one-stage and two-stage approach. Priors for μ_α and α_i are also the same for both approaches. However, the covariance matrices for γ_{ie}, η_{ieh} and λ_{iehl}, the scale parameter for ϕ_{iet} as well as ζ_i are subject-specific in the two-stage approach.

4.8.2.1 Full model priors

$$
\begin{aligned}
\alpha_i &= \begin{pmatrix}\alpha_{0i}\\ \alpha_{1i}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_{\alpha_0}\\ \mu_{\alpha_1}\end{pmatrix}, \Sigma_\alpha\right)\\
\gamma_{ie} &= \begin{pmatrix}\gamma_{0ie}\\ \gamma_{1ie}\end{pmatrix} \sim N(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_{\gamma_i})\\
\eta_{ieh} &= \begin{pmatrix}\eta_{0ieh}\\ \eta_{1ieh}\end{pmatrix} \sim N(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_{\eta_i})\\
\lambda_{iehl} &= \begin{pmatrix}\lambda_{0iehl}\\ \lambda_{1iehl}\end{pmatrix} \sim N(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_{\lambda_i})\\
\phi_{iet} &\sim t(0, \sigma^2_{\phi_i}, 3)\\
\zeta_{bi} &\sim N(\mu_{\zeta_b}, \sigma^2_{\zeta_b}) \text{ for } b = 0,1\\
\mu_{\alpha_b} &\sim N(0, 10^9) \text{ for } b = 0,1\\
\sigma^2_{\phi_i} &\sim N(\sigma^2_\phi, \sigma^2_{\sigma^2_\phi})\\
\sigma^2_\phi &\sim \text{IG}(10^{-3}, 10^{-3})\\
\sigma^2_{\sigma^2_\phi} &\sim \text{IG}(10^{-3}, 10^{-3})\\
\mu_{\zeta_b} &\sim N(0, 10^9) \text{ for } b = 0,1\\
\sigma^2_{\zeta_b} &\sim \text{IG}(10^{-3}, 10^{-3}) \text{ for } b = 0,1 \text{ and}\\
\Sigma_{\alpha_i}, \Sigma_{\gamma_i}, \Sigma_{\eta_i} \text{ and } \Sigma_{\lambda_i} &\sim \text{IW}(\text{diag}(1,1), 2).
\end{aligned}
$$

4.8.2.2 First Stage Priors

$$
\begin{aligned}
\alpha_{bi} &\sim N(0, 10^9) \text{ for } b = 0,1\\
\sigma^2_{\phi_i} &\sim \text{IG}(10^{-3}, 10^{-3})\\
\zeta_{bi} &\sim N(0, 10^9) \text{ for } b = 0,1\\
\Sigma_{\alpha_i}, \Sigma_{\gamma_i}, \Sigma_{\eta_i} \text{ and } \Sigma_{\lambda_i} &\sim \text{IW}(\text{diag}(1,1), 2)
\end{aligned}
$$

4.8.2.3 Second Stage Priors

$$
\begin{aligned}
\alpha_i &= \begin{pmatrix}\alpha_{0i}\\ \alpha_{1i}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_{\alpha_0}\\ \mu_{\alpha_1}\end{pmatrix}, \Sigma_\alpha\right)\\
\sigma^2_{\phi_i} &\sim N(\sigma^2_\phi, \sigma^2_{\sigma^2_\phi})\\
\zeta_{bi} &\sim N(\mu_{\zeta_b}, \sigma^2_{\zeta_b}) \text{ for } b = 0,1\\
\mu_{\alpha_b} &\sim N(0, 10^8) \text{ for } b = 0,1\\
\Sigma_\alpha &\sim \text{IW}(\text{diag}(1,1),2)\\
\sigma^2_\phi &\sim \text{IG}(10^{-3}, 10^{-3})\\
\sigma^2_{\sigma^2_\phi} &\sim \text{IG}(10^{-3}, 10^{-3})\\
\mu_{\zeta_b} &\sim N(0, 10^9) \text{ for } b = 0,1\\
\sigma^2_{\zeta_b} &\sim \text{IG}(10^{-3}, 10^{-3}) \text{ for } b = 0,1
\end{aligned}
$$

Using Cholesky decomposition, Σ_{γ_i}, Σ_{η_i} and Σ_{λ_i} become,

$$
\begin{aligned}
c_{r\gamma_i} &\sim N(\mu_{cr\gamma}, \sigma^2_{cr\gamma}) \text{ for } r = 1,2,3\\
c_{r\eta_i} &\sim N(\mu_{cr\eta}, \sigma^2_{cr\eta}) \text{ for } r = 1,2,3\\
c_{r\lambda_i} &\sim N(\mu_{cr\lambda}, \sigma^2_{cr\lambda}) \text{ for } r = 1,2,3\\
\mu_{cr\gamma}, \mu_{cr\eta}, \mu_{cr\lambda} &\sim N(0, 10^9) \text{ for } r = 1,2,3\\
\sigma^2_{cr\gamma}, \sigma^2_{cr\eta}, \sigma^2_{cr\lambda} &\sim \text{IG}(10^{-3}, 10^{-3}) \text{ for } r = 1,2,3.
\end{aligned}
$$

4.9 Appendix B: Methods

In this section, the different stages within the two-stage approach are described in detail. Again, we use Model 3 for illustration purposes. Let

$$\begin{aligned}
\alpha_i &= \{\alpha_{01}, \alpha_{11}, \ldots, \alpha_{0,N}, \alpha_{1,N}\} \\
\zeta_i &= \{\zeta_{01}, \zeta_{11}, \ldots, \zeta_{0,N}, \zeta_{1,N}\} \\
C_{\gamma_i} &= \{c_{1\gamma_1}, c_{2\gamma_1}, c_{3\gamma_1}, \ldots, c_{1\gamma_N}, c_{2\gamma_N}, c_{3\gamma_N}\} \\
C_{\eta_i} &= \{c_{1\eta_1}, c_{2\eta_1}, c_{3\eta_1}, \ldots, c_{1\eta_N}, c_{2\eta_N}, c_{3\eta_N}\} \\
C_{\lambda_i} &= \{c_{1\lambda_1}, c_{2\lambda_1}, c_{3\lambda_1}, \ldots, c_{1\lambda_N}, c_{2\lambda_N}, c_{3\lambda_N}\}.
\end{aligned}$$

To simplify the notation, we let C_i denote all of the Cholesky decomposition parameters in C_{γ_i}, C_{η_i} and C_{λ_i}. In the spirit of Lunn et al. (2013), we are interested in the parameters of interest from the first stage, namely, $\alpha_i, \zeta_i, C_i, \sigma^2_{\phi_i}$. The population averages and covariance matrices for α_i are given by μ_α and Σ_α. The population averages and variances for $\zeta_i, C_i, \sigma^2_{\phi_i}$ are given by $\mu_\zeta, \mu_c, \sigma^2_\phi$ and $\sigma^2_\zeta, \sigma^2_c, \sigma^2_{\sigma^2_\phi}$ respectively, where

$$\begin{aligned}
\mu_\alpha &= \{\mu_{\alpha_{01}}, \mu_{\alpha_{02}}\} \\
\mu_\zeta &= \{\mu_{\zeta_{01}}, \mu_{\zeta_{02}}\} \\
\mu_c &= \{\mu_{c1\gamma}, \mu_{c2\gamma}, \mu_{c3\gamma}, \mu_{c1\eta}, \mu_{c2\eta}, \mu_{c3\eta}, \mu_{c1\lambda}, \mu_{c2\lambda}, \mu_{c3\lambda}\}
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2_\zeta &= \{\sigma^2_{\zeta_{01}}, \sigma^2_{\zeta_{02}}\} \\
\sigma^2_c &= \{\sigma^2_{c1\gamma}, \sigma^2_{c2\gamma}, \sigma^2_{c3\gamma}, \sigma^2_{c1\eta}, \sigma^2_{c2\eta}, \sigma^2_{c3\eta}, \sigma^2_{c1\lambda}, \sigma^2_{c2\lambda}, \sigma^2_{c3\lambda}\}.
\end{aligned}$$

4.9.1 Full Model

The joint posterior distribution for Model 3 by means of the two-stage approach is given by,

$$p(\mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi},\alpha_1,\ldots,\alpha_N,\zeta_1,\ldots,\zeta_N,C_1,\ldots,C_N,\sigma^2_{\phi_1},\ldots,\sigma^2_{\phi_N},\propto$$
$$\mathcal{N}\mid y)p(\mu_\alpha)p(\mu_\zeta)p(\mu_c)p(\sigma^2_\phi)p(\Sigma_\alpha)p(\sigma^2_\zeta)p(\sigma^2_c)p(\sigma^2_{\sigma^2_\phi})\times$$
$$\prod_{i=1}^{N}\Big\{p(y_i\mid\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathcal{N}_i)p(\alpha_i\mid\mu_\alpha,\Sigma_\alpha)\ p(\zeta_i\mid\mu_\zeta,\sigma^2_\zeta)\ p(C_i\mid\mu_c,\sigma^2_c)\times$$
$$p(\sigma^2_{\phi_i}\mid\sigma^2_\phi,\sigma^2_{\sigma^2_\phi})p(\mathcal{N}_i).$$

(4.9.2)

4.9.2 First Stage

We analyse all individuals independently from the joint posterior distribution of each α_i,ζ_i,C_i conditional on y_i alone,

$$p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathcal{N}_i\mid y_i)\propto p(y_i\mid\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathcal{N}_i)p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i})p(\mathcal{N}_i)$$
$$i=1,...,N. \qquad (4.9.3)$$

4.9.3 Second Stage

From distribution (B.1) the posterior distributions for the second stage are given by,

$$p(\mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi\mid\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi},\alpha_1,\ldots,\alpha_N,\zeta_1,\ldots,\zeta_N,C_1,\ldots,C_N,\sigma^2_{\phi_1},\ldots,\sigma^2_{\phi_N},\mathcal{N},y)\ \propto$$
$$p(\mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi)\prod_{i=1}^{N}p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i}\mid\mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi}) \qquad (4.9.4)$$

$$p(\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi}\mid\mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\alpha_1,\ldots,\alpha_N,\zeta_1,\ldots,\zeta_N,C_1,\ldots,C_N,\sigma^2_{\phi_1},\ldots,\sigma^2_{\phi_N},\mathcal{N},y)\ \propto$$
$$p(\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi})\prod_{i=1}^{N}p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i}\mid\mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi}) \qquad (4.9.5)$$

$$p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathscr{N}_i \mid \mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi},y) \quad \propto$$
$$p(y_i \mid \alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathscr{N}_i)p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i} \mid \mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_C,\sigma^2_{\sigma^2_\phi})p(\mathscr{N}_i)$$
$$i = 1,...,N. \qquad (4.9.6)$$

The distributions from (B.3) and (B.4) are available in closed form and hence we can sample from them directly by using standard algorithms. For the distributions (B.5) we use the distributions in (B.2) as the proposal distributions within a Metropolis-Hastings step. For the Metropolis-Hastings algorithm, the target-to-proposal ratio, denoted as R, can be simplified to,

$$\begin{aligned} R(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathscr{N}_i) \quad &\propto \quad \frac{p(y_i \mid \alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathscr{N}_i)}{p(y_i \mid \alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i},\mathscr{N}_i)} \times \\ &\qquad \frac{p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i} \mid \mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi})p(\mathscr{N}_i)}{p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i})p(\mathscr{N}_i)} \\ &= \quad \frac{p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i} \mid \mu_\alpha,\mu_\zeta,\mu_c,\sigma^2_\phi,\Sigma_\alpha,\sigma^2_\zeta,\sigma^2_c,\sigma^2_{\sigma^2_\phi})}{p(\alpha_i,\zeta_i,C_i,\sigma^2_{\phi_i})} \qquad (4.9.7) \\ &\qquad i = 1,...,N. \end{aligned}$$

For more details regarding the two-stage approach, we refer to (Lunn et al., 2013).

4.10 Supplementary Material

In this section we describe the steps used in this chapter. This includes the code used to run each of the stages of the modeling process. Furthermore, we include plots showing the convergence assessment at each of the stages. We focus on the final model (Model 3) which was the most complex model to fit.

4.10.1 A: Data

```
'data.frame': 247520 obs.
 id      : unique eye ranging from 1 to 138
 eye     : unique eye ranging from 1 to 276
 hemi    : unique hemifield ranging from 1 to 552
 pos     : unique location ranging from 1 to 14350
 field   : unique visual field ranging from 1 to 4760
 years   : time in years
 y.cens  : response variable ranging from 0 to 40 with
           censored values as NA
 thres   : response variable ignoring censoring
```

4.10.2 B: Code

4.10.2.1 First Stage

The first stage of the two-stage approach for our specific data set is implemented in jags by means of the **rjags** package.

```
### Model
model.stageOne <- function() {
 for (i in 1:n.obs){
      is.cens[i] ~ dinterval(y.cens[i], -0.000001)
      y.cens[i] ~ dnorm(mu[i],tau[i])
      mu[i] <- lambda[pos[i],1] + years[i] *
           lambda[pos[i], 2] + phi[visit[i]]

  ### Relationship between sentivity and variabilty
      log.Sigma[i] <- zeta[id[eye[hemi[pos[i]]]],1] +
              zeta[id[eye[hemi[pos[i]]]], 2] * mu[i]
      Sigma[i] <- exp(log.Sigma[i])
      tau[i] <- pow(Sigma[i], -2)
    }


      ### Individual Level
```

```
    for(k1 in 1:n.ind){

    for (g in 1:2){alpha[k1, g] ~ dnorm(0, 0.000000001)}
    for (q in 1:2){zeta[k1, q] ~ dnorm(0, 0.000000001)}

    tau.visit[k1] ~ dgamma(0.001,0.001)
    tau.pos[k1,1:2,1:2] ~ dwish(Omega[,], 2)
    tau.hemi[k1,1:2,1:2] ~ dwish(Omega[,], 2)
    tau.eye[k1,1:2,1:2] ~ dwish(Omega[,], 2)

    sigma.field[k1] <- 1/tau.field[k1]
    sigma.pos[k1, 1:2, 1:2] <- inverse(tau.pos[k1,,])
    sigma.hemi[k1, 1:2, 1:2] <- inverse(tau.hemi[k1,,])
    sigma.eye[k1, 1:2, 1:2] <- inverse(tau.eye[k1,,])
  }

  ### Eye Level
  for(k2 in 1:n.eye){
    gamma[k2, 1:2] ~ dmnorm(alpha[id[k2],],
       tau.eye[id[k2],,])
  }

  ### Hemisphere Level
  for(k3 in 1:n.hemi){
    eta[k3, 1:2] ~ dmnorm(gamma[eye[k3], 1:2],
         tau.hemi[id[eye[k3]],,])
  }

  ### Location Level
  for(k4 in 1:n.pos){
    lambda[k4, 1:2] ~ dmnorm(eta[hemi[k4], 1:2],
         tau.pos[id[eye[hemi[k4]]],,])
  }

  ### Visit Level
  for(k5 in 1:n.visit){
    phi[k5] ~ dt(0, tau.visit[v[k5]], 3)
  }

  Omega[1, 1] <- 1
  Omega[1, 2] <- 0
  Omega[2, 1] <- 0
  Omega[2, 2] <- 1

}
```

Parameters of interest need to be followed and saved as an array 'samples'. In our case, the parameters of interest include `alpha`, `zeta`, `sigma.eye`, `sigma.hemi`, `sigma.pos` and `sigma.visit`.

4.10.2.2 Second Stage

In the second stage we use the samples saved in the first stage as proposal distributions for the MH-algorithm. This is implemented by means of the Black Box version of OpenBUGS.

```
### Model
model{
  for (i in 1:N){
    dummy[i] ~ dproposem(theta[i, 1:p],samples[i, 1:p, 1:L])
    theta[i, 1:p] ~ dmnorm(mu[1:p], tau.theta[,])

  }

### Priors
 for(m in 1:13){
  mu[m] ~ dnorm(0, 0.000001)
  }

  mu[14] <- log(sigma.visit)

sigma.field <- 1/tau.visit
tau.visit ~ dgamma(0.001, 0.001)


tau.theta[1,1] <- tau.theta1[1, 1]
tau.theta[1,2] <- tau.theta1[1, 2]
tau.theta[2,1] <- tau.theta1[2, 1]
tau.theta[2,2] <- tau.theta1[2, 2]

tau.theta1[1:2,1:2] ~ dwish(Omega[,],2)
sigma2.theta1 <- inverse(tau.theta1[,])

Omega[1,1]<-1
Omega[1,2]<-0
Omega[2,1]<-0
Omega[2,2]<-1

for(j in 3:14){
tau.theta[j,j] ~ dgamma(0.001, 0.001)
}
```

```
}
```

The samples from this stage need to be saved as coda files to be used in R version 3.0.1.

4.10.2.3 Extension

This stage is done for each eye separately using the parameters estimated in the second stage. This is implemented in R.

Step 2A:

Initial values are determined using the following optimization routine:

```
SIGMA.eye <- rep(list(SIGMA0.eye), length(unique(dat2$eye)))
SIGMA.eye <- do.call(adiag, SIGMA.eye)

SIGMA.hemi <- rep(list(SIGMA0.hemi), length(unique(dat2$hemi)))
SIGMA.hemi <- do.call(adiag, SIGMA.hemi)

SIGMA.pos <- rep(list(SIGMA0.pos), length(unique(dat2$pos)))
SIGMA.pos <- do.call(adiag, SIGMA.pos)

M <- vector()

fr <- function(thetas){

  tht <- relist(thetas, list.thetas)
  eye.int <- tht$eye.int
  eye.slope <- tht$eye.slope
  hemi.int <- tht$hemi.int
  hemi.slope <- tht$hemi.slope
  pos.int <- tht$pos.int
  pos.slope <- tht$pos.slope
  field.offset <- tht$field.offset

  Fixted <- mean(ind.int[ ,unique(id)]) +
            mean(ind.slope[ ,unique(id)])*years


  M <- Fixted + eye.int[eye2] + eye.slope[eye2] * years +
       hemi.int[hemi2] + hemi.slope[hemi2] * years  +
       pos.int[pos2] + pos.slope[pos2] * years  +
       field.offset[field2]
```

```
  proposal <- sum(not.cens * dnorm(thres, mean=M,
                                   sd=exp(mean(b0.sd)+
                                   mean(b1.sd) * M),
                                   log = T)) +
              sum(is.cens * pnorm(thres, mean=M,
                                  sd=exp(mean(b0.sd)+
                                  mean(b1.sd) * M),
                                  log.p = T)) +
              dmvnorm(as.vector(rbind(eye.int, eye.slope)),
                      mean = rep(c(0), length(unique(eye)) * 2),
                      sigma = SIGMA.eye,log=T) +
              dmvnorm(as.vector(rbind(hemi.int, hemi.slope)),
                      mean=rep(c(0), length(unique(hemi)) * 2),
                       sigma=SIGMA.hemi,log=T) +
              dmvnorm(as.vector(rbind(pos.int, pos.slope)),
                      mean=rep(c(0), length(unique(pos)) * 2),
                       sigma = SIGMA.pos,log=T) +
              sum(dgt(field.offset, mu = 0,
                      sigma = mean(field.sd), df=3))

  -(proposal)
}

dgt <- function(x, mu, sigma, df){
       dt((x - mu)/sigma,df = df,log = T)-log(sigma)}

list.thetas <- NULL

list(eye.int = rep(c(0),length(unique(eye))),
     eye.slope = rep(c(0), length(unique(eye))),
     hemi.int = rep(c(0), length(unique(hemi))),
     hemi.slope = rep(c(0), length(unique(hemi))),
     pos.int = rep(c(0), length(unique(pos))),
     pos.slope = rep(c(0), length(unique(pos))),
     field.offset = rep(c(0), length(unique(field))))

thetas <- NULL
thetas <- unlist(as.relistable(list.thetas))


inits <- c(rep(c(0), length(unique(eye))),
   rep(c(0), length(unique(eye))),
           rep(c(0), length(unique(hemi))),
           rep(c(0), length(unique(hemi))),
           rep(c(0), length(unique(pos))),
           rep(c(0), length(unique(pos))) ,
           rep(c(0), length(unique(field))))
```

```
o <-  optim(inits, fr, method = "BFGS",
            control = list(parscale = rep(0.001,
            length(inits))))

gc()

results <- o$par

eye.int1 <- results[1]
eye.slope1 <- results[2]

hemi.int1 <- results[3:4]
hemi.slope1 <- results[5:6]

pos.int1 <- results[7:58]
pos.slope1 <- results[59:110]

field.offset1 <- results[111:length(results)]
```

Step 2B:

A random walk Metropolis-Hastings algorithm is used for each of the levels. This is done iteratively, to take into account the correlation between the levels. This implemented in R version 3.0.1.

```
### Function which can be used for the eye, hemifield and
### location level to estimate the proposal and
### target distributions

like <- function(Fixted, bi.level, bt.level,
 SIGMA.level, level){

   Fixted = Fixted
   bi.level = bi.level
   bt.level = bt.level
   mean.p = Fixted +  bi.level[,1] + bi.level[,2] * years
   mean.t = Fixted +  bt.level[,1] + bt.level[,2] * years
   SIGMA.level = SIGMA.level
   level = level

   proposal = rowsum(not.cens * dnorm(thres,
                    mean = mean.p,
                    sd = exp(b0.sd[i] + b1.sd[i] * mean.t),
```

```
                       log = T), group = level) +
              rowsum(is.cens*pnorm(thres,
                       mean = mean.p,
                       sd = exp(b0.sd[i] + b1.sd[i] * mean.p),
                       log.p = T), group = level) +
              dmvnorm(bi.level[,], mean=c(0,0),
                           sigma=SIGMA.level, log = T)

    target =  rowsum(not.cens * dnorm(thres,
                     mean = mean.t,
                     sd = exp(b0.sd[i] + b1.sd[i] * mean.t),
                     log = T), group = level) +
              rowsum(is.cens * pnorm(thres,
                     mean  =mean.t,
                     sd = exp(b0.sd[i] + b1.sd[i] * mean.t),
                     log.p=T), group = level)+
              dmvnorm(bt.level[,], mean=c(0,0),
                         sigma = SIGMA.level, log = T)

    p = proposal - target
    p
}

### This is run first as an adaption phase to ensure that
### the acceptance rate is between 20-40 percent.
### Once the variance has been determined for the proposal
### distribution at each level,
### this can be run and the results saved.

  for(j in l:u)
  {

    ### RANDOM EFFECTS FOR THE EYE LEVEL

    bi.eye <- rmvnorm(1, mean = c(bt.eye[,1], bt.eye[,2]),
       sigma = SIGMA0.eye * var.change.eye)

    Fixted <- proposal <- target <- vector()

    Fixted <- i.ind.int + i.ind.slope*years +
              hemi.int[j-1,hemi2] +
              hemi.slope[j-1,hemi2] * years +
              pos.int[j-1,pos2] +
              pos.slope[j-1,pos2] * years +
              field.offset[j-1,field2]

    p <- like(Fixted, bi.eye, bt.eye, SIGMA.eye, eye2)
```

```
rho = exp(p)
r.unif = runif(1)

bt.eye[,1] <- eye.int[j-1,] + (bi.eye[,1] -
              eye.int[j-1,])  *  (r.unif < rho)
bt.eye[,2] <- eye.slope[j-1,] + (bi.eye[,2] -
              eye.slope[j-1,]) *  (r.unif < rho)

eye.int[j,] <- as.numeric(bt.eye[,1])
eye.slope[j,] <- as.numeric(bt.eye[,2])

a <- bt.eye[,1]
b <- eye.int[j-1,]
count1.eye[a != b] <- count1.eye[a != b] + 1
count2.eye <- count2.eye + 1


### RANDOM EFFECTS FOR THE HEMIFIELD LEVEL

SIGMA.hemi1 <- rep(list(SIGMA0.hemi), length(unique(hemi)))
SIGMA.hemi1 <- do.call(adiag, SIGMA.hemi1)

Fixted <- proposal <- target <- vector()

Fixted <- i.ind.int + i.ind.slope * years +
          eye.int[j, eye2] +
          eye.slope[j, eye2] * years +
          pos.int[j-1, pos2] +
          pos.slope[j-1, pos2] * years +
          field.offset[j-1, field2]

bi.hemi <- matrix(rmvnorm(1, mean = as.vector(rbind(
                          bt.hemi[,1], bt.hemi[,2])),
                          sigma = SIGMA.hemi1 *
                          rep(var.change.hemi, each = 2)),
                          ncol = 2, byrow = T)

p <- like(Fixted, bi.hemi, bt.hemi, SIGMA.hemi, hemi2)

rho <- exp(p)
r.unif <- runif(2)

bt.hemi[,1] <- hemi.int[j-1,] + (bi.hemi[,1] -
    hemi.int[j-1,]) * (r.unif < rho)
bt.hemi[,2] <- hemi.slope[j-1,] + (bi.hemi[,2] -
    hemi.slope[j-1,]) * (r.unif < rho)
```

```
hemi.int[j,] <- as.numeric(bt.hemi[,1])
hemi.slope[j,] <- as.numeric(bt.hemi[,2])

a <- bt.hemi[,1]
b <- hemi.int[j-1,]

count1.hemi[a != b] <- count1.hemi[a != b] + 1
count2.hemi <- count2.hemi+1

a <- bt.hemi[,1]
b <- hemi.int[j-1,]

count1.hemi[a != b] <- count1.hemi[a != b] + 1
count2.hemi <- count2.hemi+1


### RANDOM EFFECTS FOR THE LOCATION LEVEL

SIGMA.pos1 <- rep(list(SIGMA0.pos), length(unique(pos)))
SIGMA.pos1 <- do.call(adiag, SIGMA.pos1)

Fixted <- proposal <- target <- vector()

Fixted <- i.ind.int + i.ind.slope * years +
          eye.int[j, eye2] +
          eye.slope[j, eye2] * years +
          hemi.int[j, hemi2] +
          hemi.slope[j, hemi2] * years +
          field.offset[j-1, field2]


bi.pos <- matrix(rmvnorm(1,mean = as.vector(rbind(
                         bt.pos[,1], bt.pos[,2])),
                         sigma = SIGMA.pos1 *
                         rep(var.change.pos, each = 2)),
                         ncol = 2, byrow = T)

p <- like(Fixted, bi.eye, bt.eye, SIGMA.eye,
          eye2, n.unif)

rho <- exp(p)
r.unif <- runif(52)

bt.pos[,1] <- pos.int[j-1,] + (bi.pos[,1] -
              pos.int[j-1,]) * (r.unif < rho)
bt.pos[,2] <- pos.slope[j-1,] + (bi.pos[,2] -
```

```
                pos.slope[j-1,]) * (r.unif < rho)

pos.int[j,] <- as.numeric(bt.pos[,1])
pos.slope[j,] <- as.numeric(bt.pos[,2])

a <- bt.pos[,1]
b <- pos.int[j-1,]

count1.pos[a != b] <- count1.pos[a != b] + 1
count2.pos <- count2.pos + 1

### RANDOM EFFECTS FOR THE FIELD LEVEL

Fixted<-proposal<-target<-vector()

Fixted <- i.ind.int + i.ind.slope * years +
          eye.int[j, eye2] +
          eye.slope[j, eye2] * years +
          hemi.int[j, hemi2] +
          hemi.slope[j, hemi2] * years +
          pos.int[j, pos2] +
          pos.slope[j, pos2] * years

bi.field <- rnorm(n = N.field, mean = bt.field,
                  sd = mean(field.sd) * var.change.field)

proposal <- as.numeric(rowsum(not.cens * dnorm(thres,
                              mean = Fixted +
                              bi.field[field2],
                              sd = exp((b0.sd[i]) +
                              (b1.sd[i]) *
                              (Fixted + bi.field[field2])),
                              log = T), group = field2) +
                       rowsum(is.cens * pnorm(thres,
                              mean = Fixted +
                              bi.field[field2],
                              sd = exp((b0.sd[i]) +
                              (b1.sd[i]) *
                              (Fixted + bi.field[field2])),
                              log.p = T), group = field2) +
            as.numeric(dgt(bi.field, mu = 0,
                           sigma = field.sd[i], df = 3)))

target <-   as.numeric(rowsum(not.cens * dnorm(thres,
                              mean = Fixted +
                              bt.field[field2],
                              sd = exp((b0.sd[i]) +
```

```
                                    (b1.sd[i]) *
                                    (Fixted + bt.field[field2])),
                                     log = T), group = field2) +
                              rowsum(is.cens * pnorm(thres,
                                    mean = Fixted +
                                    bt.field[field2],
                                    sd = exp((b0.sd[i]) +
                                    (b1.sd[i]) * (Fixted +
                                    bt.field[field2])),
                                    log.p = T),
                                    group = field2) +
                    as.numeric(dgt(bt.field, mu = 0,
                                sigma = field.sd[i], df = 3)))

  p <- proposal - target
  rho <- exp(p)
  r.unif <- runif(N.field)

  bt.field <- field.offset[j-1,] + (bi.field-
   field.offset[j-1,]) *  (r.unif < rho)
  field.offset[j,] <- as.numeric(bt.field)

  a <- bt.field
  b <- field.offset[j-1,]

  count1.field[a != b] <- count1.field[a != b] + 1
  count2.field <- count2.field + 1
}
```

4.10.3 C: Convergence Assessment

In this section we include plots showing the convergence of certain parameters at each stage of the modeling process.

4.10.3.1 First stage

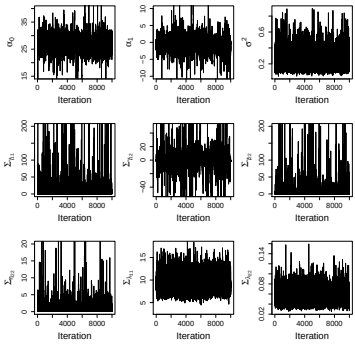

Figure 4.7: Convergence plots for individual 7.

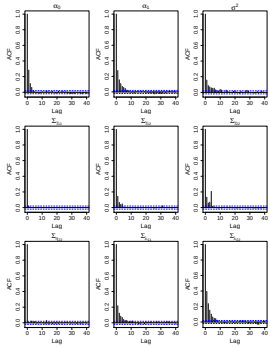

Figure 4.8: Autocorrelation plots for individual 7.

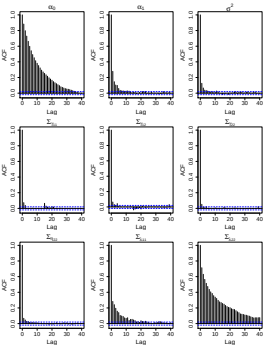

Figure 4.9: Autocorrelation plots for individual 89.

4.10.3.2 Second stage

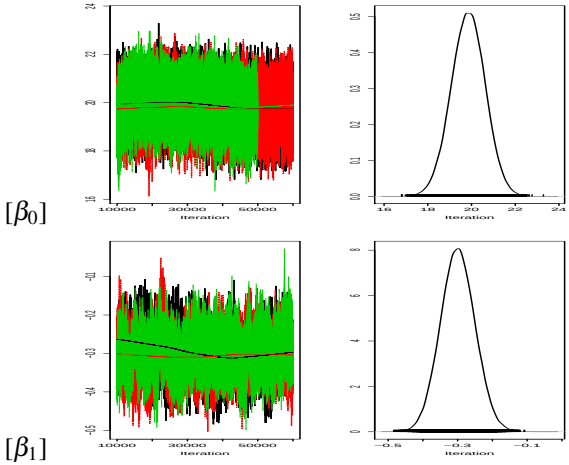

Figure 4.10: Convergence plots.

4.10.3.3 Extension

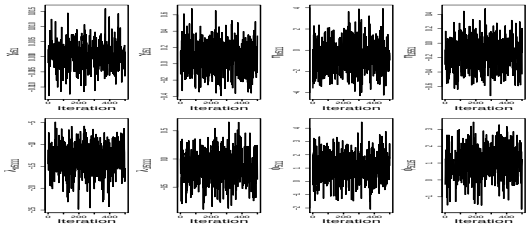

Figure 4.11: Convergence plots for individual 5.

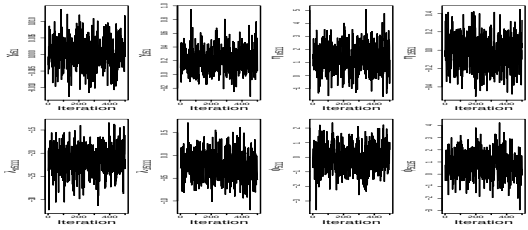

Figure 4.12: Convergence plots individual 9.

CHAPTER 5

Comparison of Estimation Methods for Multilevel Models

This Chapter is based on: Bryan, S.R., Lesaffre, E.M.E.H., van Rosmalen, J.M., Noh, M., Lee, Y. and Rizopoulos, D. Comparison of estimation methods for multilevel models. *Submitted*.

Abstract

The Bayesian approach has become increasingly popular because it allows to fit quite complex models to data via Markov chain Monte Carlo (MCMC) sampling. However, it is also recognized nowadays that MCMC sampling can become computationally prohibitive when applied to a large data set. To overcome these issues we have investigated three possible approaches that address the computational burden in a hierarchical model, namely, the two-stage MCMC approach, Integrated nested Laplace approximation (INLA) and Hierarchical likelihood (H-likelihood) and compare these to the classical one-stage approach. Since these methods approximate the solution of the classical one-stage MCMC approach, something is gained and lost with each approach. The aim of this paper is to compare these three alternative methods from a statistical and practical viewpoint. The motivating data set is an ongoing Dutch glaucoma study. Glaucoma is one of the leading causes of blindness in the world. In order to detect deterioration at an early stage, a model for predicting visual fields (VF) in time is needed. Hence, the true underlying VF progression can be determined, and treatment strategies can then be optimized to prevent further VF loss. A simulation study was done to evaluate and compare the different approaches.

5.1 Introduction

The Bayesian approach has become increasingly popular from a conceptual and practical viewpoint. Firstly, the Bayesian approach allows for prior information to be incorporated and takes into account all uncertainty in the model parameters. Secondly, it allows to fit quite complex models to data via Markov chain Monte Carlo (MCMC) sampling. Furthermore, MCMC algorithms allow greater flexibility by relaxing the strong parametric assumptions commonly used in most frequentist hierarchical models (Lesaffre and Lawson, 2012). The classical Bayesian approach to hierarchical modeling is done in one stage, i.e. subject-specific and overall parameters are estimated simultaneously. However, this approach might be prohibitive in practice with a large number of repeated measurements in combination with a complex hierarchical structure.

The motivating data that we consider in this paper is taken from an ongoing Dutch glaucoma study conducted by the Rotterdam Eye Hospital in the Netherlands. The Glaucoma study provides a unique database of longitudinally measured visual field (VF) data. From a statistical modelling viewpoint the analysis of the VF data presents several challenges. First, VF testing involves measuring the level of differential light sensitivity at different locations within each eye. The sensitivity estimates are left-censored due to the limitation of the device used to measure them. Furthermore, VF data have a hierarchical structure consisting of 4 levels, namely, (1) the individual, (2) the eye, (3) the hemifield and (4) the location. Another difficulty in modelling VF data is the amount and type of measurement error in the sensitivity estimates. This may be due to measurable factors, such as season, time of day and reliability

indices, or unknown transient factors, such as fatigue or lack of concentration. Although their magnitudes may vary, these factors are global and affect all locations belonging to the same eye at each visit. To model these factors, cross-classified random effects for each visit are needed. Furthermore, to accommodate spikes in the distribution of the visit effects, a distribution which allows flexibility and can handle heavy tails should be assumed (i.e. a t-distribution). Including these aspects in the modelling approach has been investigated in previous research (Bryan et al., 2017, 2015). A problem with high dimensional data and complex data structures such as this data set, is that it is sometimes difficult or even impossible to model them with standard MCMC software.
To overcome these issues, we previously analysed the glaucoma data using the two-stage MCMC approach (Lunn et al., 2013; Bryan et al., 2017). In this paper, we compare the two-stage MCMC approach to the classical one-stage MCMC approach. Furthermore, we investigate two alternative approaches, namely, the INLA approach (Rue et al., 2009) and the H-likelihood approach (Lee and Nelder, 1996) and compare them to the classical one-stage approach. Since these methods approximate the one-stage approach, something is gained and lost with each approach. We investigate these approaches from a statistical and practical viewpoint using both real data, as well as data simulated for different scenarios. The rest of the paper is structured as follows. In Section 2 we provide the working model. The estimation approaches are described in Section 3. In the subsequent section we give further details on the motivating data set and apply the computational approaches to the Glaucoma study data. Section 5 describes the simulation study which was done to compare the different approaches. Finally, Section 6 contains a concluding discussion.

5.2 The Bayesian Hierarchical Model

The computational approaches will be introduced via a hierarchical model involving subjects followed-up in time, also called a mixed-effects model. Let y_{it} denote the follow-up measurements for the i-th individual ($i = 1,\ldots,n$) where t denoted the specific time-point $t = 1,\ldots,m_i$. Furthermore, let y_i denote all observations for individual i and y denote the observations for all individuals combined. The Gaussian mixed-effects model can be written as

$$\begin{aligned} y_{it} &= x_{it}^{\top}\beta + z_{it}^{\top}\alpha_i + \varepsilon_{it}, \\ \alpha_i &\sim N(0,\Sigma_\alpha), \\ \varepsilon_{it} &\sim N(0,\sigma^2), \end{aligned}$$

where β denotes the vector with the regression coefficients of the design matrix for the fixed effects x_i and z_i denotes the design matrix for the random effects α_i. In particular, the fixed and the random effects refer to the population-average and subject-specific effects, respectively. Furthermore, Σ_α is the covariance-variance matrix of the random effects, ε_{it} are the error

terms and σ^2 is the variance of the error. Additionally, α_i is assumed to be independent of ε_{it} and the ε_{it} are assumed to be independent of each other for $i = 1,\ldots,n$ and $t = 1,\ldots,m_i$.

5.3 Bayesian Estimation Techniques

5.3.1 Markov chain Monte Carlo techniques

For complex models, the posterior distribution often cannot be determined analytically. Calculating the integral using numerical integration methods is a practical alternative if only a few parameters are involved, but it becomes difficult for real-life applications where the dimensionality is often high Lesaffre and Lawson (2012). These problems can be avoided by using sampling techniques yielding Markov chains. A popular class of sampling algorithms is Markov chain Monte Carlo (MCMC) methods.

5.3.1.1 One-stage approach

The classical MCMC approach to Bayesian hierarchical modeling is done in one stage, which has the advantage that subject-specific and overall parameters are estimated simultaneously. For the model in (1) and assuming a simple structure for α_i which only includes an individual-specific intercept (α_{0i}) and slope for time (α_{1i}), we have for the i-th individual ($i = 1,\ldots,n$) and specific time-point ($t = 1,\ldots,m_i$),

$$y_{it} = \alpha_{0i} + \alpha_{1i}\text{time}_{it} + \varepsilon_{it},$$

where

$$\begin{aligned} \alpha_i &= (\alpha_{0i}, \alpha_{1i})^T \sim N((\beta_0, \beta_1)^T, \Sigma_\alpha), \\ \varepsilon_{it} &\sim N(0, \sigma^2), \end{aligned}$$

with priors $\sigma^2 \sim p(\sigma^2)$, $\beta \sim p(\beta)$ and $\Sigma_\alpha \sim p(\Sigma_\alpha)$. The joint posterior distribution is then given by,

$$p(\alpha, \sigma^2, \beta, \Sigma_\alpha \mid y) \propto (\prod_{i=1}^{n}\prod_{t=1}^{m_i} N(y_{it} \mid \alpha_i, \sigma^2))(\prod_{i=1}^{n} N(\alpha_i \mid \beta, \Sigma_\alpha))p(\sigma^2)p(\beta, \Sigma_\alpha). \quad (5.3.1)$$

However, the MCMC approach can be difficult or even impossible to implement for complex models with standard software. In the motivating data set for this paper, we consider a 4-level model which includes 30858 random effects in total. Computation problems arise when looking at these large data sets, including memory issues, as well as problems

obtaining convergence in a realistic time frame. For these issues, using cluster computing is not a solution since all parameters need to be estimated simultaneously. Hence, a more computationally efficient method is needed for such situations. This is described in more detail by Bryan et al. (2017).

5.3.1.2 Two-stage approach

Having explained in the previous section the complexities of using the full Bayesian approach, in the following subsection we present three alternative procedures that may considerably decrease the computational burden and make it feasible to fit highly complex multilevel mixed models. The first solution we present is the two-stage approach which was previously used to model the glaucoma data by Bryan et al. (2015). The two-stage approach was proposed by Lunn et al. (2013). This approach has been originally suggested in the context of a meta-analysis. To adapt the approach to our setting, we can select a particular level l from the L level multilevel structure, fit the corresponding N_l models at this level. To obtain the final results we then need to appropriately utilize the different MCMC samples.

Lunn et al. (2013) proposed a two-stage approach, which allows us to simplify the problem while still benefiting from the advantages of a full Bayesian model. In the two-stage approach the variances σ_l^2 are forced to differ for each subgroup in level l. For the simplified case described above, we split the model at the individual level, and the model is for the i-th individual ($i = 1,\ldots,n$) and specific time-point ($t = 1,\ldots,m_i$),

$$y_{it} \quad = \quad \alpha_{0i} + \alpha_{1i}\text{time}_{it} + \varepsilon_{it},$$

where

$$\begin{aligned} \alpha_i &= (\alpha_{0i}, \alpha_{1i})^T \sim N((\beta_0, \beta_1)^T, \Sigma_\alpha), \\ \varepsilon_{it} &\sim N(0, \sigma_i^2), \end{aligned}$$

with priors $\sigma_i^2 \sim p(\sigma_i^2)$, $\beta \sim p(\beta)$ and $\Sigma_\alpha \sim p(\Sigma_\alpha)$. The joint posterior distribution is then given by,

$$p(\alpha, \sigma^2, \beta, \Sigma_\alpha \mid y) \propto p(\beta)p(\Sigma_\alpha)(\prod_{i=1}^{n}\prod_{t=1}^{m_i} N(y_{it} \mid \alpha_i, \sigma_i^2))(\prod_{i=1}^{n} N(\alpha_i \mid \beta, \Sigma_\alpha))p(\sigma_i^2) \quad (5.3.2)$$

where α and σ^2 are the collection of all α_is and σ_i^2s respectively.

Note that the classical one-stage approach can be adapted in the same way to allow the variances to differ for each individual i.

5.3.1.3 Stage 1

The two-stage approach allows to simplify the problem by splitting the hierarchical models with K levels at level k, in this case at the individual level. By doing this, we are able to run the analysis for the individuals in parallel, thereby considerably reducing the computational time. Independent parameters of interest at the individual level are estimated in stage 1. In this example, the individual variances σ_i^2 are treated as nuisance parameters. We first analyse each individual independently, to obtain a sample of size B_i for $i = 1,\ldots,N$. The samples are obtained from the joint posterior distribution of each α_i and σ_i^2, conditional on y_i, i.e.

$$p(\alpha_i, \sigma_i^2 \mid y_i) \propto p(y_i \mid \alpha_i, \sigma_i^2) p(\alpha_i) p(\sigma_i^2) \text{ for } i = 1,\ldots,n. \tag{5.3.3}$$

5.3.1.4 Stage 2

The samples obtained from stage 1 are then used as proposal distributions for those parameters in stage 2. In this stage, the joint posterior distribution of β, Σ_α, α and σ^2 is sampled under the full hierarchical model. At each iteration one cycles through the full conditional distributions for β, Σ_α and then each α_i and σ_i^2 jointly. These are given by,

$$p(\beta \mid \Sigma_\alpha, \alpha, \sigma^2, y) \propto p(\beta) \prod_{i=1}^{n} p(\alpha_i \mid \beta, \Sigma_\alpha), \tag{5.3.4}$$

$$p(\Sigma_\alpha \mid \beta, \alpha, \sigma^2, y) \propto p(\Sigma_\alpha) \prod_{i=1}^{n} p(\alpha_i \mid \beta, \Sigma_\alpha), \tag{5.3.5}$$

$$p(\alpha_i, \sigma_i^2 \mid \beta, \Sigma_\alpha, y) \propto p(y_i \mid \alpha_i, \sigma_i^2) p(\alpha_i \mid \beta, \Sigma_\alpha) p(\sigma_i^2), i = 1,\ldots n. \tag{5.3.6}$$

Multivariate normal and inverse Wishart priors can be specified for β and Σ_α respectively. Hence, the distributions from (5.3.4) and (5.3.5) are available in closed form and can hence be sampled directly by using standard algorithms. For the distribution in equation (5.3.6) the distributions in equation (5.3.3) are used as proposal distributions within a Metropolis-Hastings step. For this, the target-to-proposal ratio, denoted as R, is given by, for $i = 1,\ldots,n$:

$$R(\alpha_i, \sigma_i^2) \quad = \quad \frac{p(\alpha_i \mid \beta, \Sigma_\alpha)}{p(\alpha_i)} \tag{5.3.7}$$

Let b_{ij} be a uniformly chosen index from $\{1,\ldots,B_i\}$ at each iteration j. By using flat priors for the α_i in stage 1, these samples are then based solely on the likelihood. Then $\{\alpha_i^{(b_{ij})}, \sigma_i^{2(b_{ij})}\}$ is accepted with probability $\min(1, A')$, where

$$
\begin{aligned}
A' &= \frac{p(\alpha_i^{(b_{ij})} \mid \beta^{(j)}, \Sigma_\alpha^{(j)})}{p(\alpha_i^{(j-1)} \mid \beta^{(j)}, \Sigma_\alpha^{(j)})} \frac{p(\alpha_i^{(j-1)})}{p(\alpha_i^{(c_{ij})})} \\
&= \frac{p(\alpha_i^{(b_{ij})} \mid \beta^{(j)}, \Sigma_\alpha^{(j)})}{p(\alpha_i^{(j-1)} \mid \beta^{(j)}, \Sigma_\alpha^{(j)})}
\end{aligned} \tag{5.3.8}
$$

where $\beta^{(j)}$ and $\Sigma_\alpha^{(j)}$ denote the values of β and Σ_α respectively at iteration j of the Gibbs-Metropolis scheme.

Two-stage models can be performed very quickly, providing scope for rapid exploration of different models. An extension of the two-stage approach to allow for the calculation of model selection and model evaluation criteria by including an additional step based on the Method of Composition in combination with a Metropolis-within-Gibbs technique has been proposed (Bryan et al., 2017).

5.3.2 Methods based on Laplace approximation

5.3.2.1 Integrated nested Laplace approximation

Integrated nested Laplace approximation (INLA) has been proposed as a computationally convenient alternative for MCMC methods (Rue et al., 2009). Laplace approximation is a method for approximating integrals in Bayesian models based on a second-order Taylor approximation of the log posterior, which results in a Gaussian approximation to the posterior. INLA performs a nested version of the classical Laplace approximation combined with modern numerical techniques for sparse matrices, which enables computationally fast approximate Bayesian inference. The method applies only to latent Gaussian models, which are a subset of Bayesian additive models with a structured additive predictor and Gaussian priors for the model parameters in the additive predictor (Rue et al., 2016). The multi-level hierarchical model implemented in this paper can be seen as an instance of a latent Gaussian model, provided that all random effects and model residuals are assumed to be normally distributed.

The model parameters can be divided into the latent Gaussian variables α and β, which have normal prior distributions, and the hyperparameters σ^2 and Σ_α, which do not necessarily have a normal prior. INLA computes an approximation to the posterior in three steps. First the joint posterior distribution of the hyperparameters σ^2 and Σ_α is approximated using Laplace approximation, according to

$$
\tilde{\pi}(\sigma^2, \Sigma_\alpha | y) \propto \left. \frac{\pi(\alpha, \beta, \sigma^2, \Sigma_\alpha, y)}{\tilde{\pi}_G(\alpha, \beta | \sigma^2, \Sigma_\alpha, y)} \right|_{\alpha = \alpha^*(\sigma^2, \Sigma_\alpha), \beta = \beta^*(\sigma^2, \Sigma_\alpha)}, \tag{5.3.9}
$$

where $\tilde{\pi}_G(\alpha,\beta|\sigma^2,\Sigma_\alpha,y)$ is the Gaussian approximation to the full conditional distribution $\pi(\alpha,\beta|\sigma^2,\Sigma_\alpha,y)$, and $\alpha^*(\sigma^2,\Sigma_\alpha)$ and $\beta^*(\sigma^2,\Sigma_\alpha)$ represent the estimated mode of the full conditional distribution as a function of the hyperparameters. The posterior distribution of the hyperparameters $\tilde{\pi}(\sigma^2,\Sigma_\alpha|y)$ is then explored. This step does not require parametric representation, but rather a sufficient exploration to be able to select good evaluation points for the numerical integration of the final step.

In the second step, the posterior distribution of the parameters $\pi(\alpha,\beta|\sigma^2,\Sigma_\alpha,y)$ is approximated using a Gaussian approximation, another Laplace approximation or a simplified version of the Laplace approximation. In the third step, the first two steps are combined by using numerical integration with

$$\tilde{\pi}(\alpha_i|y) = \sum_s \tilde{\pi}(\alpha_i|\sigma_s^2,\Sigma_{\alpha,s},y)\tilde{\pi}(\sigma_s^2,\Sigma_{\alpha,s}|y)\Delta_s$$
$$\tilde{\pi}(\beta|y) = \sum_s \tilde{\pi}(\beta|\sigma_s^2,\Sigma_{\alpha,s},y)\tilde{\pi}(\sigma_s^2,\Sigma_{\alpha,s}|y)\Delta_s,$$

where $\sigma_s^2,\Sigma_{\alpha,s}$ denote the evaluation points obtained in the first step, Δ_s the corresponding area weights, and $\tilde{\pi}(\alpha_i|\sigma_s^2,\Sigma_{\alpha,s},y)$ and $\tilde{\pi}(\beta|\sigma_s^2,\Sigma_{\alpha,s},y)$ are the approximations obtained in the second step. For a more extensive explanation on INLA, we refer to Rue et al. (2009)(Rue et al., 2009).

5.3.2.2 Hierarchical likelihood

The hierarchical likelihood (H-likelihood) approach has been proposed as a conceptually simple and reliable inferential procedure for models with general random effect structures. Consider the h-likelihood of Lee and Nelder (1996) of the form,

$$h = \ell_1(\beta,\theta) + \ell_2(\theta), \tag{5.3.10}$$

where $\ell_1(\beta,\theta) = \log f_{\beta,\theta}(y|\alpha) = \sum_{it}\log f_{\beta,\theta}(y_{it}|\alpha_i)$ and $\ell_2(\theta) = \log f_\theta(\alpha) = \sum_i \log f_\theta(\alpha_i)$. In this section, we use $f_\theta(\cdot)$ to denote the density functions of random variables with parameters θ; the arguments within the brackets can be either conditional or unconditional. Thus, $f_\theta(y,\alpha)$ and $f_\theta(y|\alpha)$ have different functional forms even though we use the same $f_\theta(\cdot)$ to indicate density functions with parameters θ.

This h-likelihood $\exp(h)$ is proportional to the posterior $p(\alpha,\beta,\theta|y)$ when (β,θ) have uniform priors. Maximum likelihood (ML) estimators maximize the log-(marginal) likelihood $m = \log\int\exp(h)d\alpha$, whose asymptotic efficiency is well established. However, the marginal likelihood is often hard to obtain and numerically intractable, especially when the dimensionality of the integral is high. In such cases, Lee and Nelder (2001) considered an adjusted profile likelihood function class $a_\xi(l)$, defined by

$$a_\xi(l) = [l - \frac{1}{2}\log\det\{D(l,\xi)/(2\pi)\}]|_{\xi=\tilde{\xi}}$$

where $D(l,\xi) = -\partial^2 l/\partial\xi^2$ and $\widetilde{\xi}$ solves $\partial l/\partial\xi = 0$. Note that $a_\alpha(h)$ is the Laplace approximation to m (Lee and Nelder, 2001), which often gives accurate approximation (Tierney and Kadane, 1986). In a normal linear mixed model, $a_\beta(m) = a_{\beta,\alpha}(h) = \log f_\theta(y|\hat{\beta})$, where $f_\theta(y|\hat{\beta})$ is the restricted likelihood hood of Patterson and Thompson (1971). This gives the restricted (or residual) ML (REML) estimator for θ. For a detailed discussion see Lee et al. (2006). Here, the function $a_{\beta,\alpha}(h)$ means that both random and fixed effects are simultaneously eliminated from h. Lee and Nelder (2001) showed that $a_{\beta,\alpha}(h) \simeq \log f_\theta(y|\hat{\beta})$ to the first order approximation (Cox and Reid, 1987). The function class $a_\xi()$ is useful in making adjusted profile likelihoods, eliminating nuisance effects ξ, fixed or random or both. Thus, $a_\alpha(h)$ is used to obtain the ML estimators for β and $a_{\beta,\alpha}(h)$ is used to obtain the REML estimators for θ. The h-likelihood approach also allow to estimate the ML estimators for θ by using $a_\alpha(h)$.

Lee and Kim (2017) (Lee and Kim, 2017) showed that asymptotically h-likelihood and Bayesian inferences become equivalent. Thus, with large samples, the optimization procedure based on h-likelihood provides an efficient way of computing relevant quantities for the Bayesian inference without having to resort to MCMC sampling.

5.4 Application

According to the World Health Organization, glaucoma is one of the leading causes of irreversible blindness in the world. Due to the progressive nature of the disease, it is important to detect deterioration at an early stage. Evaluation of a longitudinal series of visual fields (VF), as measured by standard automated perimetry, provides a way to detect early evidence of glaucoma and to determine functional deterioration. Hence, the true underlying VF progression can be determined, and treatment strategies can then be optimized to prevent further VF loss.

5.4.1 Motivating data set: the Glaucoma Study

The Glaucoma Study is a prospective cohort study conducted by the Rotterdam Eye Hospital in the Netherlands. This is an ongoing study which began in 1998. Inclusion criteria included glaucoma diagnosis and an age range of 18 to 85 years. In total, 139 patients, consisting of 80 (57.6%) men and 59 (42.4%) women, were recruited with a mean follow-up of 10.5 years. Follow-up data were collected at approximately 6-monthly intervals. All patients gave their written informed consent for participation. All research procedures followed the tenets set forth in the Declaration of Helsinki. Furthermore, all of the data that was used in this analysis has been made available online at http://rod-rep.com (Bryan et al., 2013; Erler et al., 2014).
Sensitivity estimates were measured at 52 test locations within each eye, or 26 test locations

within each hemifield (excluding two locations corresponding to the blind spot). The VFs were tested using the Humphrey Field Analyzer with the 24-2, white-on-white test strategy using the Full Threshold algorithm. The light source can be attenuated in the range from 1 to 10,000 times. On the decibel (dB) scale an attenuation x is defined as $s = 10\log_{10}(x)$, or $x = 10^{s/10}$. The lowest sensitivity that can be detected by this perimeter is 0 dB, although negative values could in fact occur if it were not for the limitations of this device. The highest sensitivity that can be detected is 50 dB, however few humans are capable of seeing a stimulus less than 40 dB, which is 1/10,000 of the maximum intensity of the instrument (or 1 asb). Thus, for practical purposes, the useful intensity range for white light testing is from 0 to 40 dB with a background illumination of 31.5 asb. (Anderson and Patella, 1999).

5.4.2 Previous research

To account for the aforementioned features of the VF data we recently proposed a multilevel Bayesian mixed effects model with nested and crossed random effects (Bryan et al., 2017). We proposed to model factors (such as season, time of day and fatigue) which affect all locations belonging to the same VF as Global Visit Effects (GVEs) (Bryan et al., 2015). In this way, we can account for both the known and the unknown factors. The sensitivity estimates are left-censored due to a limitation of the device. Let the observed sensitivity estimate be denoted by y^* and the latent, true sensitivity value denoted by y. Let years_{it} represent the time between measurement t and the first measurement for each individual i, ranging from 0 to 10.5 years. We then have, for individual $i = 1,\ldots,N$; eye $e = 1,2$; hemifield $h = 1,2$; location $l = 1,\ldots,26$ and visit $t = 1,\ldots,T_i$,

Full Model:

$$\begin{aligned} y_{iehlt} \quad = \quad & \beta_0 + \beta_1 \text{years}_{it} + \alpha_{0i} + \alpha_{1i}\text{years}_{it} + \gamma_{0ie} + \gamma_{1ie}\text{years}_{it} + \eta_{0ieh} + \\ & \eta_{1ieh}\text{years}_{it} + \lambda_{0iehl} + \lambda_{1iehl}\text{years}_{it} + \phi_{iet} + \varepsilon_{iehlt}. \end{aligned} \tag{5.4.1}$$

Here, $\beta = (\beta_0, \beta_1)^T$ corresponds to the population averaged intercept and slope of time. The individual-specific intercept and slope of time for individual i are represented by α_{0i} and α_{1i}, with deviations due to the eye given by γ_{0ie} and γ_{1ie}, due to the hemifield by η_{0ieh} and η_{1ieh}, and due to the location by λ_{0iehl} and λ_{1iehl}. The GVE is represented by an extra parameter, ϕ_{iet}. This captures the offset at every visit t for each eye e within each individual i. To accommodate the large number of spikes in the distribution of the GVE, we assume a t-distribution for ϕ_{iet} with 3 degrees of freedom.

The combination of the 4-level hierarchical structure with the cross-classified random effects makes fitting the full Bayesian model a challenging task. As a consequence, we were unable to achieve convergence in a realistic time frame and experienced computer memory limitations with the classical one-stage MCMC approach and using standard software such as

WinBUGS or JAGS. To overcome this limitation we used the two-stage approach of Lunn et al. (2013).

5.4.3 Model for comparisons

To compare the four approaches, we use a simplified version of the proposed model which was used by Bryan et al. (2017), as described in Section 5.4.2. Let years_{it} represent the time between measurement t and the first measurement for each individual i, ranging from 0 to 10.5 years. For this simple model, we ignore the censoring since INLA cannot handle the deterministic mapping between y_{iehlt} and $y^{\star}_{iehlt}$. Furthermore, we exclude the GVE since INLA is unable to model the cross-classified random effects with a t-distribution. Hence, we use the observed sensitivity values, y^{*}_{iehlt}. We then have, for individual $i = 1,\ldots,N$; eye $e = 1,2$; hemifield $h = 1,2$; location $l = 1,\ldots,26$ and visit $t = 1,\ldots,T$,

Simplified model:

$$\begin{aligned} y^{*}_{iehlt} &= \alpha_{0i} + \alpha_{1i}\text{years}_{it} + \gamma_{0ie} + \gamma_{1ie}\text{years}_{it} + \\ &\quad \eta_{0ieh} + \eta_{1ieh}\text{years}_{it} + \lambda_{0iehl} + \lambda_{1iehl}\text{years}_{it} + \varepsilon_{iehlt} \end{aligned} \tag{5.4.2}$$

where

$$\begin{aligned} \beta_p &\sim N(0,10^8) \text{ for } p = 0,1;\ \alpha_i = (\alpha_{0i},\alpha_{1i})^T \sim N((\beta_0,\beta_1)^T,\Sigma_\alpha) \\ \gamma_{ie} &= (\gamma_{0ie},\gamma_{1ie})^T \sim N((0,0)^T,\Sigma_\gamma);\ \eta_{ieh} = (\eta_{0ieh},\eta_{1ieh})^T \sim N((0,0)^T,\Sigma_\eta) \\ \lambda_{iehl} &= (\lambda_{0iehl},\lambda_{1iehl})^T \sim N((0,0)^T,\Sigma_\lambda) \text{ and } \varepsilon_{iehlt} \sim N(0,\sigma^2). \end{aligned}$$

We performed two different analyses splitting the two-stage approach splitting the levels in different ways. In the first analysis, we split the levels at the individual level, treating each individual as their own sample. These individuals were then analyzed independently before combining them to obtain population level estimates. The changed priors in the full model are as follows,

$$\begin{aligned} \gamma_{ie} &= (\gamma_{0ie},\gamma_{1ie})^T \sim N((0,0)^T,\Sigma_{\gamma_i});\ \eta_{ieh} = (\eta_{0ieh},\eta_{1ieh})^T \sim N((0,0)^T,\Sigma_{\eta_i}) \\ \lambda_{iehl} &= (\lambda_{0iehl},\lambda_{1iehl})^T \sim N((0,0)^T,\Sigma_{\lambda_i}) \text{ and } \varepsilon_{iehlt} \sim N(0,\sigma_i^2). \end{aligned}$$

In the second analysis, we split the levels at the eye level, treating each eye as their own sample. The changed priors in the full model are as follows,

$$\begin{aligned}\eta_{ieh} &= (\eta_{0ieh}, \eta_{1ieh})^T \sim N((0,0)^T, \Sigma_{\eta_i}) \\ \lambda_{iehl} &= (\lambda_{0iehl}, \lambda_{1iehl})^T \sim N((0,0)^T, \Sigma_{\lambda_i}) \text{ and } \varepsilon_{iehlt} \sim N(0, \sigma_i^2).\end{aligned}$$

Uninformative priors were given for the variances and covariance matrices.

5.4.4 Software

The majority of the programs were run using R (R Foundation for Statistical Computing, 2013). The one-stage approaches, as well as the first stage of the two-stage approach were implemented using the **rjags** package. The second stage of the two-stage approach was run in OpenBUGS (Lunn et al., 2013). The H-likelihood approach was implemented using a self-written program by the 3rd author. INLA was implemented using the **R-INLA** package.

5.4.5 Results

For this analysis we included both eyes from the 139 individuals belonging to the Glaucoma study. This included 4,758 VFs, resulting in a data set consisting of 247,416 location-specific sensitivity estimates. A comparison of the classical one-stage, **R-INLA** and the H-likelihood approaches can be seen in Table 5.1.

Table 5.1: Summary statistics for the simplified model

Parameter	One-stage	R-INLA	H-Likelihood
β_0	20.74 (0.52)	20.77 (0.52)	20.37 (0.55)
β_1	-0.18 (0.03)	-0.18 (0.03)	-0.23 (0.03)
σ^2	12.35	12.39	13.85
Σ_α	$\begin{pmatrix}20.56 & -0.53\\ -0.53 & 0.10\end{pmatrix}$	$\begin{pmatrix}23.53 & -0.35\\ -0.35 & 0.10\end{pmatrix}$	$\begin{pmatrix}30.50 & -0.37\\ -0.37 & 0.13\end{pmatrix}$
Σ_γ	$\begin{pmatrix}15.88 & -0.22\\ -0.22 & 0.06\end{pmatrix}$	$\begin{pmatrix}14.72 & -0.20\\ -0.20 & 0.06\end{pmatrix}$	$\begin{pmatrix}17.49 & -0.04\\ -0.04 & 0.07\end{pmatrix}$
Σ_η	$\begin{pmatrix}35.58 & 0.02\\ 0.02 & 0.08\end{pmatrix}$	$\begin{pmatrix}38.91 & -0.07\\ -0.07 & 0.08\end{pmatrix}$	$\begin{pmatrix}42.23 & 0.23\\ 0.23 & 0.11\end{pmatrix}$
Σ_λ	$\begin{pmatrix}30.20 & -0.36\\ -0.36 & 0.09\end{pmatrix}$	$\begin{pmatrix}31.06 & -0.21\\ -0.21 & 0.10\end{pmatrix}$	$\begin{pmatrix}34.20 & 0.27\\ 0.27 & 0.11\end{pmatrix}$

A comparison of the one-stage and two-stage approaches, splitting the model at the individual level, i.e. allowing σ^2, Σ_γ, Σ_η and Σ_α to differ for every individual i and treating them as nuisance parameters can be seen in Table 5.2.

Table 5.2: Summary statistics when dividing the data at the individual level

Parameter	One-stage	Two-stage
β_0	21.36 (0.60)	20.97 (0.56)
β_1	-0.15 (0.06)	-0.18 (0.04)
Σ_α	$\begin{pmatrix} 32.64 & -0.45 \\ -0.45 & 0.07 \end{pmatrix}$	$\begin{pmatrix} 39.85 & -1.03 \\ -1.03 & 0.14 \end{pmatrix}$

A comparison of the one-stage and two-stage approaches, splitting the model at the eye level, i.e. allowing σ^2, Σ_η and Σ_α to differ for every eye e and treating them as nuisance parameters can be seen in Table 5.3.

Table 5.3: Summary statistics when dividing the data at the eye level

Parameter	One-stage	Two-stage
β_0	21.21 (0.57)	20.88 (0.55)
β_1	-0.15 (0.03)	-0.19 (0.03)
Σ_α	$\begin{pmatrix} 20.13 & -0.31 \\ -0.31 & 0.05 \end{pmatrix}$	$\begin{pmatrix} 22.78 & -0.49 \\ -0.49 & 0.10 \end{pmatrix}$
Σ_γ	$\begin{pmatrix} 28.29 & -0.23 \\ -0.23 & 0.04 \end{pmatrix}$	$\begin{pmatrix} 34.62 & -0.15 \\ -0.15 & 0.04 \end{pmatrix}$

5.4.6 Discussion of the results

The results using the classical one-stage approach, **R-INLA** and H-likelihood gave consistent results. The population averaged-intercepts were found to be 20.74 dB, 20.77 dB and 20.37 dB, respectively for the three approaches. An intercept of approximately 20 dB indicates that the patients within this data set have fairly moderate glaucoma at presentation. The population-averaged slopes were indicative of slow progression and found to be -0.18 dB/year, -0.18 dB/year and -0.23 dB/year while the variances were found to be 12.35, 12.39 and 13.85 for each of the approaches respectively. The two-stage results differ from the above

mentioned approaches because of the assumptions we make when allowing the lower levels of the model to differ per individual (or per eye). These results are consistent with the one-stage approach when making the same assumptions with both approaches. When splitting the model at the individual level, the population-averaged intercepts were found to be 21.39 dB and 20.97 dB and the population-averaged slopes were found to be -0.15 dB/year and -0.18 dB/year for the one-stage approach and the two-stage approach, respectively. When splitting the model at the eye level, the population-averaged intercepts were found to be 21.21 dB and 20.88 dB and the population-averaged slopes were found to be -0.15 dB/year and -0.19 dB/year, respectively. Hence, the results were found to be similar for all models irrespective of which level the model was split at. It is important to note that all individuals received therapy throughout the study and this could explain the small range of slopes of progression.
Our model takes into account the hierarchical structure of the data. This allows us to take into account the correlation at each of the 4 levels. The estimated covariance matrices for each level were similar for the classical one-stage approach, **R-INLA** and H-likelihood. A large proportion of intercept variability was explained by the individual level random effects for the classical one-stage approach, **R-INLA** and H-likelihood ($\Sigma_{\alpha 11} = 20.56$ dB, 23.53 dB and 30.50 dB). A smaller amount of this intercept variability was explained by the eye level random effects ($\Sigma_{\gamma 11} = 15.88$ dB, 14.72 dB and 17.49 dB). This is due to the fact that each individual only has 2 eyes and with this approach we assume that only the eyes belonging to the same individual share a common variance distribution. This variability increases for the 4 hemifields per individual ($\Sigma_{\eta 11} = 35.58$ dB, 38.91 dB and 42.23 dB) and for the 104 locations per individual ($\Sigma_{\lambda 11} =$ 30.20 dB, 31.06 dB and 34.20 dB).
The variances were higher for the level that we split the data at for the one-stage approach when making the same assumptions as the two-stage approach. These results are similar to the variances obtained from the two-stage approach. The individual level intercept variability was found to be 32.64 dB and 39.85 dB for the two approaches when dividing the data at the individual level. The individual level intercept variability was found to be 20.13 dB and 22.78 dB and the eye level intercept variability was found to be 28.29 dB and 34.62 dB for the two approaches when dividing the data at the eye level. The increased variability of these two approaches can be explained by the different model assumptions that are used for these approaches compared to that of the classical one-stage, **R-INLA** and H-likelihood approaches.

5.5 Simulation study

A simulation was set up to evaluate the performance and to compare the different approaches. The design of the simulated data was similar to the models that were fitted on the glaucoma data. A brief description of the simulation study and a discussion of the results follows in the section below. A more extensive description, as well as the results, is given in Appendix A.

5.5.1 Setup

Since the two-stage approach cannot be completely automated, a relatively large simulated data set was used rather than many simulations based on small data sets. A data set consisting of 1000 subjects per scenario was simulated. The follow-up period for each subject was randomly selected between 2 to 8 years equally spaced. The median number of visits was equal to 5. The data was simulated from a linear mixed-effects model (as described in Section 2). The methods were compared as described previously. Namely, the classical one-stage, **R-INLA** and the H-likelihood approaches were compared and the one-stage and two-stage approaches were compared. We varied two factors in the simulation study, (1) the number of levels and (2) the variance of the random effects. This was done for models based on the one-stage and two-stage approaches. This resulted in 8 different simulations. Upon convergence, the posterior mean, median, standard deviation with the equal tail 95% credible interval (CI) for all parameters of interest were computed in order to compare the 4 different approaches.

5.5.2 Discussion of the simulation study results

Although the estimation procedures differ, the classical one-stage, **R-INLA** and H-likelihood approaches produced similar results for all of the simulation studies. However, better estimates for the slope parameters were obtained using the H-likelihood approach for the simulations including 3 levels. The true slope parameters were set to be -1 dB/year. For simulation 3, the slope parameters estimates were -0.95 dB/year, -0.94 dB/year and -0.99 dB/year for the classical one-stage, **R-INLA** and H-likelihood approaches, respectively. For simulation 4, these estimates were -1.34 dB/year, -1.29 dB/year and -0.99 dB/year for the classical one-stage, **R-INLA** and H-likelihood approaches, respectively. The one-stage and two-stage approaches produced consistently similar results for all of the simulation studies.

5.5.3 Practical aspects

When analysing this data set, important aspects such as the left-censoring of the data and Global Visit Effects need to be accounted for (Bryan et al., 2015, 2017). However, for the comparison of these approaches a simplified model was needed. This was due to the fact that it is not possible to use censoring or to use a t-distribution on the cross-classified random effects with R-INLA.

While models using the one-stage, **R-INLA** and the H-likelihood approaches could all be run using one program, the two-stage approach needed to run in R first and then in a 'blackbox' version of OpenBUGS afterwards. Although this approach is computationally quicker since the data can be split into smaller data sets and run in parallel this additional step can be inconvenient, especially for simulation studies where a large number of data sets need

to be analysed.

5.6 Concluding remarks

In this paper we investigated different approaches to model complex high dimensional longitudinal data. Although the parameter estimates for the real data and simulations are similar for the different approaches, each method has specific advantages and disadvantages. The classical one-stage model is a standard method for modelling longitudinal data. However, this method is time consuming with complex models and large data sets and may also imply computational difficulties.

The two-stage approach provides a method to overcome the computational difficulties by allowing the user to split the data into smaller, manageable data sets which can be run in parallel. This reduces the computational burden. However, this adaptation changes the model assumptions by allowing the parameters to vary at lower levels. In some cases this may be in line with the data, however, in other cases when the two-stage is only chosen to reduce computation time, this may be an unintended side effect. Since this is still a new method, there are aspects that need to be taken into account. Firstly, if covariates are included in the model these need to be well thought-out. For example, if factors are the same within one individual, blocks need to be used. Secondly, the two-stage approach works with a 'blackbox' version of OpenBUGS. Currently the only way to perform the second stage is to use a 'blackbox' version of OpenBUGS. Hence, it is not possible to use R to call OpenBUGS. For more information on setting up OpenBUGS for performing the two-stage analyses, we refer to Lunn et al. (2013). Both the one-stage and two-stage approaches can be used in a great variety of hierarchical models, not only for Gaussian hierarchical models. Furthermore, they are able to take into account special features such as a censored response.

R-INLA provides a quick solution, both computationally and with ease of implementation. This approach allows us to overcome the difficulties with computational time. However, **R-INLA** has limitations in what kinds of models can be estimated. For example, it is not possible to use censoring or to use a t-distribution on the cross-classified random effects.

The H-likelihood approach provides a conceptually simple and reliable inferential procedure for models with general random effect structures, for example crossed-classified and multivariate random effects and those with spatial and temporal correlations. Outliers or model misspecifications can cause a severe bias in parameter estimation (Noh and Lee, 2007). The use of a heavy-tailed distribution for responses gives robust estimates against outliers (Noh and Lee, 2007) while that of a heavy-tailed distribution for random effects gives robust analysis against misspecifications of distributional assumptions on random effects (Lee et al., 2006). The h-likelihood approach allows both ML and REML analyses and provides various model checking plots and useful model selection criteria such as conditional AIC for model selection. The R package dhglm allowing heavy-tailed distributions is available at CRAN (Lee and Noh, 2016).

Previous work can be found in the literature comparing combinations of the approaches discussed in this paper. For example, comparisons between the one-stage and two-stage approaches (Lunn et al., 2013); **R-INLA** and OpenBUGS (Carroll et al., 2015) and H-likelihood and Bayesian approaches (Jin Jang et al., 2007), to name a few. To our knowledge, there is no current literature comparing the one-stage, two-stage, **R-INLA** and H-likehihood approaches. Comparisons with other approaches, such as SAS, are described by Lesaffre and Lawson (2012). Furthermore, Browne and Draper (2006) show a comparison of Bayesian and likelihood-based methods for fitting multilevel models. Although these methods were not in the scope of this paper, further comparisons with other approaches could be investigated.

In this paper we have described 4 different estimation techniques for modelling complex hierarchical models. We have investigated the pros and cons for each approach. The decision for which method to use should depend on the required modelling flexibility, computation time and ease of implementation.

Bibliography

Anderson, D. R. and Patella, V. M. (1999). *Automated Static Perimetry*. Mosby, 2nd edition.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Bryan, S. R., Eilers, P. H. C., Lesaffre, E. M. E. H., Lemij, H. G., and Vermeer, K. A. (2015). Global visit effects in point-wise longitudinal modeling of glaucomatous visual fields. *Investigative Opthalmology & Visual Science*, 56(8):4283–4289.

Bryan, S. R., Eilers, P. H. C., van Rosmalen, J., Rizopoulos, D., Vermeer, K. A., Lemij, H. G., and Lesaffre, E. M. E. H. (2017). Bayesian hierarchical modeling of longitudinal glaucomatous visual fields using a two-stage approach. *Statistics in Medicine*, 36:1735–1753.

Bryan, S. R., Vermeer, K. A., Eilers, P. H. C., Lemij, H. G., and Lesaffre, E. M. E. H. (2013). Robust and censored modeling and prediction of progression in glaucomatous visual fields. *Investigative Opthalmology & Visual Science*, 54(10):6694–6700.

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., and Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial and Spatio-temporal Epidemiology*, pages 45–54.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):1–39.

Erler, N. S., Bryan, S. R., Eilers, P. H. C., Lesaffre, E. M. E. H., Lemij, H. G., and Vermeer, K. A. (2014). Optimizing structure-function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Opthalmology & Visual Science*, 55(4):2350–2357.

Jin Jang, M., Lee, Y., Lawson, A. B., and Browne, W. J. (2007). A comparison of hierarchical likelihood and Bayesian approaches to spatial epidemiological modelling. *Environmetrics*, 18:809–821.

Lee, Y. and Kim, G. (2017). Oracle properties of maximum H-likelihood estimators for random effects. *Submitted for publication*.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman and Hall, London.

Lee, Y. and Noh, M. (2016). DHGLM: Double hierarchical generalized linear models. R package version 1.6.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. Wiley & Sons, Chichester, West Sussex, 1st edition.

Lunn, D., Barrett, J., Sweeting, M., and Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 62(4):551–572.

Noh, M. and Lee, Y. (2007). Reml estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, 98:898–915.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

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

Rue, H., Riebler, A., Sorbye, S., Illian, J., Simpson, D., and Lindgren, F. (2016). Bayesian computing with INLA: A review. *arXiv:1604.00860 [stat.ME]*. arXiv: 1604.00860.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

5.7 Appendix

5.7.1 Simulation models

The number of levels in the simulated data set was varied from 2 levels (a and b) to 3 levels (a, b and c). Data sets were simulated based on the model used in the one-stage approach (Models 1 and 2) as well as on the model used in the two-stage approach (Models 3 and 4). For all of the simulations, we let $\beta_0 = 30$ and $\beta_1 = -1$.

Model 1:

Two data sets are simulated from the one-stage approach with 2 levels. Let the simulated response be denoted by y_{abt}. We then simulate y_{abt} for $a = 1,\ldots,1000$, $b = 1,2$ and visit $t = 1,\ldots,T_a$ using,

$$y_{abt} = \alpha_{0a} + \alpha_{1a}\text{years}_{at} + \gamma_{0ab} + \gamma_{1ab}\text{years}_{at} + \varepsilon_{abt} \tag{5.7.1}$$

where,

$$\begin{aligned} \alpha_a &= (\alpha_{0a}, \alpha_{1a})^T \sim N((\beta_0, \beta_1)^T, \Sigma_\alpha) \\ \gamma_{ab} &= (\gamma_{0ab}, \gamma_{1ab})^T \sim N((0,0)^T, \Sigma_\gamma) \text{ and} \\ \varepsilon_{abt} &\sim N(0, \sigma^2). \end{aligned}$$

The variances used to simulate the data were varied to obtain simulated data sets 1 and 2. More specifically:
Simulation 1:

$$\Sigma_\alpha = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \Sigma_\gamma = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix} \textit{ and } \sigma^2 = 1$$

Simulation 2:

$$\Sigma_\alpha = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}, \Sigma_\gamma = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix} \textit{ and } \sigma^2 = 2$$

Model 2:

Two data sets are simulated from the one-stage approach with 3 levels. Let the simulated

response be denoted by y_{abct}. We then simulate y_{abct} for $a = 1,\ldots,1000$, $b = 1,2$, $c = 1,2,3$, and visit $t = 1,\ldots,T_a$ using,

$$\begin{aligned} y_{abct} &= \alpha_{0a} + \alpha_{1a}\text{years}_{at} + \gamma_{0ab} + \gamma_{1ab}\text{years}_{at} + \\ &\quad \eta_{0abc} + \eta_{1abc}\text{years}_{at} + \lambda_{0abc} + \varepsilon_{abct} \end{aligned} \tag{5.7.2}$$

where,

$$\begin{aligned} \alpha_a &= (\alpha_{0a},\alpha_{1a})^T \sim N((\beta_0,\beta_1)^T,\Sigma_\alpha) \\ \gamma_{ab} &= (\gamma_{0ab},\gamma_{1ab})^T \sim N((0,0)^T,\Sigma_\gamma); \\ \eta_{abc} &= (\eta_{0abc},\eta_{1abc})^T \sim N((0,0)^T,\Sigma_\eta) \text{ and} \\ \varepsilon_{abct} &\sim N(0,\sigma^2). \end{aligned}$$

Using this model, and varying the variances, we obtain simulated data sets 3 and 4. More specifically:

Simulation 3:

$$\Sigma_\alpha = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \Sigma_\gamma = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \Sigma_\eta = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \text{ and } \sigma^2 = 1$$

Simulation 4:

$$\Sigma_\alpha = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}, \Sigma_\gamma = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}, \Sigma_\eta = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}, \text{ and } \sigma^2 = 2$$

Model 3:

The two data sets are now simulated from the two-stage approach. The model is the same as in Model 1, however the variance parameters are assumed to differ across individuals as follows,

$$\begin{aligned} \alpha_a &= (\alpha_{0a},\alpha_{1a})^T \sim N((0,0)^T,\Sigma_\alpha) \\ \gamma_{ab} &= (\gamma_{0ab},\gamma_{1ab})^T \sim N((0,0)^T,\Sigma_{\gamma_a}); \end{aligned}$$

Using this model, and varying the variances, we obtain simulated data sets 5 and 6. More specifically:

Simulation 5:

$$\Sigma_\alpha = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \Sigma_{\gamma_a} \sim N \begin{pmatrix} \Sigma_{11\gamma_a} & 0 \\ 0 & \Sigma_{22\gamma_a} \end{pmatrix} \text{ and } \sigma_a^2 \sim N(1,2)$$

where

$$\Sigma_{11\gamma_a} = \sigma_{11\gamma_a} * \sigma_{11\gamma_a} \text{ and } \sigma_{11\gamma_a} \sim N(2,2)$$
$$\Sigma_{22\gamma_a} = \sigma_{22\gamma_a} * \sigma_{22\gamma_a} \text{ and } \sigma_{22\gamma_a} \sim N(2,2)$$

Simulation 6:

$$\Sigma_\alpha = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}, \Sigma_{\gamma_a} \sim N \begin{pmatrix} \Sigma_{11\gamma_a} & 0 \\ 0 & \Sigma_{22\gamma_a} \end{pmatrix} \text{ and } \sigma_a^2 \sim N(2,2)$$

where

$$\Sigma_{11\gamma_a} = \sigma_{11\gamma_a} * \sigma_{11\gamma_a} \text{ and } \sigma_{11\gamma_a} \sim N(5,2)$$
$$\Sigma_{22\gamma_a} = \sigma_{22\gamma_a} * \sigma_{22\gamma_a} \text{ and } \sigma_{22\gamma_a} \sim N(5,2)$$

Model 4:

The two data sets are now simulated from the two-stage approach. The model is the same as in Model 2, however the variance parameters are assumed to differ across individuals as follows,

$$\begin{aligned} \alpha_a &= (\alpha_{0a}, \alpha_{1a})^T \sim N((0,0)^T, \Sigma_\alpha) \\ \gamma_{ab} &= (\gamma_{0ab}, \gamma_{1ab})^T \sim N((0,0)^T, \Sigma_{\gamma_a}); \\ \eta_{abc} &= (\eta_{0abc}, \eta_{1abc})^T \sim N((0,0)^T, \Sigma_{\eta_a}) \end{aligned}$$

The variances used to simulate the data were varied to obtain simulated data sets 7 and 8. More specifically:

Simulation 7:

$$\Sigma_\alpha = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \Sigma_{\gamma_a} \sim N\begin{pmatrix} \Sigma_{11\gamma_a} & 0 \\ 0 & \Sigma_{22\gamma_a} \end{pmatrix}, \Sigma_{\eta_a} \sim N\begin{pmatrix} \Sigma_{11\eta_a} & 0 \\ 0 & \Sigma_{22\eta_a} \end{pmatrix} \text{ and } \sigma_a^2 \sim N(1,2)$$

where

$$\begin{aligned} \Sigma_{11\gamma_a} &= \sigma_{11\gamma_a} * \sigma_{11\gamma_a} \text{ and } \sigma_{11\gamma_a} \sim N(2,2) \\ \Sigma_{22\gamma_a} &= \sigma_{22\gamma_a} * \sigma_{22\gamma_a} \text{ and } \sigma_{22\gamma_a} \sim N(2,2) \\ \Sigma_{11\eta_a} &= \sigma_{11\eta_a} * \sigma_{11\eta_a} \text{ and } \sigma_{11\eta_a} \sim N(2,2) \\ \Sigma_{22\gamma_a} &= \sigma_{22\eta_a} * \sigma_{22\eta_a} \text{ and } \sigma_{22\eta_a} \sim N(2,2) \end{aligned}$$

Simulation 8:

$$\Sigma_\alpha = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}, \Sigma_{\gamma_a} \sim N\begin{pmatrix} \Sigma_{11\gamma_a} & 0 \\ 0 & \Sigma_{22\gamma_a} \end{pmatrix}, \Sigma_{\eta_a} \sim N\begin{pmatrix} \Sigma_{11\eta_a} & 0 \\ 0 & \Sigma_{22\eta_a} \end{pmatrix} \sigma_a^2 \sim N(2,2)$$

where

$$\begin{aligned} \Sigma_{11\gamma_a} &= \sigma_{11\gamma_a} * \sigma_{11\gamma_a} \text{ and } \sigma_{11\gamma_a} \sim N(5,2) \\ \Sigma_{22\gamma_a} &= \sigma_{22\gamma_a} * \sigma_{22\gamma_a} \text{ and } \sigma_{22\gamma_a} \sim N(5,2) \\ \Sigma_{11\eta_a} &= \sigma_{11\eta_a} * \sigma_{11\eta_a} \text{ and } \sigma_{11\eta_a} \sim N(5,2) \\ \Sigma_{22\gamma_a} &= \sigma_{22\eta_a} * \sigma_{22\eta_a} \text{ and } \sigma_{22\eta_a} \sim N(5,2) \end{aligned}$$

5.7.2 Simulation Results

The results for the simulations comparing the classical one-stage, **R-INLA** and H-likelihood approaches are shown in Tables 5.3-5.6. The comparisons between the one-stage and two-stage approaches are shown in Tables 5.7-5.10.

Table 5.4: Results from simulation 1

Parameter	Simulated	One-stage	R-INLA	H-Likelihood
β_0	30.00	30.08 (0.13)	30.01 (0.13)	30.05 (0.13)
β_1	-1.00	-1.03 (0.14)	-1.05 (0.12)	-1.04 (0.13)
σ^2	1.00	1.01	1.01	0.98
Σ_α	$\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix}$	$\begin{pmatrix}11.12 & -0.55\\ -0.55 & 9.46\end{pmatrix}$	$\begin{pmatrix}11.03 & -0.05\\ -0.05 & 9.58\end{pmatrix}$	$\begin{pmatrix}9.58 & 0.03\\ 0.03 & 9.59\end{pmatrix}$
Σ_γ	$\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix}$	$\begin{pmatrix}9.50 & 0.27\\ 0.27 & 9.73\end{pmatrix}$	$\begin{pmatrix}9.43 & 0.02\\ 0.02 & 9.59\end{pmatrix}$	$\begin{pmatrix}9.71 & -0.06\\ -0.06 & 9.78\end{pmatrix}$

Table 5.5: Results from simulation 2

Parameter	Simulated	One-stage	R-INLA	H-Likelihood
β_0	30.00	30.01 (0.18)	29.94 (0.17)	29.96 (0.19)
β_1	-1.00	-0.99 (0.18)	-0.86 (0.17)	-0.99 (0.18)
σ^2	2.00	1.99	2.01	1.99
Σ_α	$\begin{pmatrix}20 & 0\\ 0 & 20\end{pmatrix}$	$\begin{pmatrix}19.21 & -0.15\\ -0.15 & 17.22\end{pmatrix}$	$\begin{pmatrix}19.96 & 0.01\\ 0.01 & 19.01\end{pmatrix}$	$\begin{pmatrix}20.13 & 0.02\\ 0.02 & 19.58\end{pmatrix}$
Σ_γ	$\begin{pmatrix}20 & 0\\ 0 & 20\end{pmatrix}$	$\begin{pmatrix}20.49 & 0.42\\ 0.42 & 19.34\end{pmatrix}$	$\begin{pmatrix}18.15 & -0.05\\ -0.05 & 21.23\end{pmatrix}$	$\begin{pmatrix}19.42 & -0.02\\ -0.02 & 19.51\end{pmatrix}$

Table 5.6: Results from simulation 3

Parameter	Simulated	One-stage	R-INLA	H-Likelihood
β_0	30.00	29.96 (0.12)	29.94 (0.13)	29.95 (0.13)
β_1	-1.00	-0.95 (0.06)	-0.94 (0.13)	-0.99 (0.13)
σ^2	1.00	0.99	0.99	0.99
Σ_α	$\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix}$	$\begin{pmatrix}10.49 & -0.32\\ -0.32 & 10.92\end{pmatrix}$	$\begin{pmatrix}10.70 & -0.02\\ -0.02 & 10.19\end{pmatrix}$	$\begin{pmatrix}9.55 & 0.02\\ 0.02 & 9.88\end{pmatrix}$
Σ_γ	$\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix}$	$\begin{pmatrix}9.89 & -0.11\\ -0.11 & 8.74\end{pmatrix}$	$\begin{pmatrix}11.27 & 0.04\\ 0.04 & 9.48\end{pmatrix}$	$\begin{pmatrix}9.87 & 0.01\\ 0.01 & 9.83\end{pmatrix}$
Σ_η	$\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix}$	$\begin{pmatrix}9.56 & 0.03\\ 0.03 & 10.64\end{pmatrix}$	$\begin{pmatrix}9.43 & 0.001\\ 0.001 & 10.56\end{pmatrix}$	$\begin{pmatrix}10.55 & -0.04\\ -0.04 & 10.46\end{pmatrix}$

Table 5.7: Results from simulation 4

Parameter	Simulated	One-stage	R-INLA	H-Likelihood
β_0	30.00	29.78 (0.17)	29.78 (0.19)	29.88 (0.19)
β_1	-1.00	-1.34 (0.18)	-1.29 (0.19)	-0.99 (0.18)
σ^2	2.00	2.02	2.02	2.01
Σ_α	$\begin{pmatrix}20 & 0\\ 0 & 20\end{pmatrix}$	$\begin{pmatrix}19.60 & 1.77\\ 1.77 & 19.33\end{pmatrix}$	$\begin{pmatrix}19.19 & 0.02\\ 0.02 & 20.75\end{pmatrix}$	$\begin{pmatrix}19.41 & -0.02\\ -0.02 & 19.41\end{pmatrix}$
Σ_γ	$\begin{pmatrix}20 & 0\\ 0 & 20\end{pmatrix}$	$\begin{pmatrix}21.89 & -0.71\\ -0.71 & 20.20\end{pmatrix}$	$\begin{pmatrix}21.83 & 0.06\\ 0.06 & 19.23\end{pmatrix}$	$\begin{pmatrix}20.78 & -0.02\\ -0.02 & 19.17\end{pmatrix}$
Σ_η	$\begin{pmatrix}20 & 0\\ 0 & 20\end{pmatrix}$	$\begin{pmatrix}20.40 & -0.37\\ -0.37 & 20.20\end{pmatrix}$	$\begin{pmatrix}20.37 & -0.02\\ -0.02 & 20.20\end{pmatrix}$	$\begin{pmatrix}19.72 & 0.05\\ 0.05 & 19.41\end{pmatrix}$

Table 5.8: Results from simulation 5

Parameter	Simulated	One-stage	Two-stage
β_0	30.00	30.12 (0.12)	30.17 (0.11)
β_1	1.00	-0.39 (0.10)	-0.40 (0.10)
Σ_α	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} 9.10 & -0.25 \\ -0.25 & 9.56 \end{pmatrix}$	$\begin{pmatrix} 10.73 & -0.81 \\ -0.81 & 10.11 \end{pmatrix}$

Table 5.9: Results from simulation 6

Parameter	Simulated	One-stage	Two-stage
β_0	30.00	30.05 (0.16)	30.10 (0.16)
β_1	1.00	0.46 (0.14)	0.44 (0.14)
Σ_α	$\begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$	$\begin{pmatrix} 19.33 & -0.61 \\ 0.61 & 19.00 \end{pmatrix}$	$\begin{pmatrix} 21.4 & -1.20 \\ -1.20 & 19.75 \end{pmatrix}$

Table 5.10: Results from simulation 7

Parameter	Simulated	One-stage	Two-stage
β_0	30	29.88 (0.11)	29.86 (0.11)
β_1	1.00	0.25 (0.10)	0.26 (0.10)
Σ_α	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	$\begin{pmatrix} 9.80 & -0.27 \\ -0.27 & 9.48 \end{pmatrix}$	$\begin{pmatrix} 7.96 & -0.19 \\ -0.19 & 8.37 \end{pmatrix}$

Table 5.11: Results from simulation 8

Parameter	Simulated	One-stage	Two-stage
β_0	30.00	30.08 (0.15)	30.13 (0.16)
β_1	1.00	0.45 (0.14)	0.41 (0.15)
Σ_α	$\begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$	$\begin{pmatrix} 19.22 & -1.06 \\ -1.06 & 19.65 \end{pmatrix}$	$\begin{pmatrix} 20.24 & -2.61 \\ -2.61 & 19.41 \end{pmatrix}$

CHAPTER 6

Global Visit Effects in Point-wise Longitudinal Modeling of Glaucomatous Visual Fields

This Chapter is based on: Bryan, S.R., Eilers, P.H.C., Lesaffre, E.M.E.H., Lemij, H.G., Vermeer, K.A. (2015). Global Visit Effects in Point-wise Longitudinal Modeling of Glaucomatous Visual Fields. *Investigative Ophthalmology & Visual Science*, 56(8):4283 – 4289.

Abstract

Purpose: One of the difficulties in modeling visual field (VF) data is the sometimes large and correlated measurement errors in the point-wise sensitivity estimates. As these errors affect all locations of the same VF, we propose to model them as Global Visit Effects (GVE). We evaluate this model and show the effect it has on progression estimation and prediction. Methods: VF series (24-2 Full Threshold; 15 biannual VFs per patient) of 125 patients with primary glaucoma were included in the analysis. The contribution of the GVE was evaluated by comparing the fitting and predictive ability of a conventional model, which does not contain GVE, to such a model that incorporates the GVE. Moreover, the GVE's effect on the estimated slopes was evaluated by determining the absolute difference between the slopes of the models. Finally, the magnitude of the GVE was compared to that of other measurement errors. Results: The GVE model showed a significant improvement in both the model fit and predictive ability over the conventional model, especially when the number of VFs in a series is limited. The average absolute difference in slopes between the models was 0.13 dB/year. Lastly, the magnitude of the GVE was more than 3 times larger than the measureable factors combined. Conclusion: By incorporating the GVE in the longitudinal modeling of VF data, better estimates may be obtained of the rate of progression as well as of predicted future sensitivities.

6.1 Introduction

Evaluation of a longitudinal series of visual fields (VF), as measured by standard automated perimetry (SAP), provides a method to detect glaucoma and to determine functional deterioration (Flammer and Meier, 2003). One of the difficulties in modeling VF data is the large measurement variability of VFs, partially due to the inherent subjective nature of such a test (Heijl et al., 1989; Werner et al., 1989; Kutzko et al., 2000). This large variability means that in clinical practice, repeated measurements are performed to confirm real progression.

A learning effect, where the average mean defect significantly decreases between exams, has been shown to be present in VF testing (Heijl et al., 1989; Gloor and Schmied, 1980; Gloor et al., 1980; Kulze et al., 1990). Furthermore, test re-test studies have shown that variability is dependent on defect depth and test location (Heijl et al., 1989).

In glaucoma, variability is presumably related to fatigue effects and response errors. A fatigue effect, whereby sensitivity estimates decrease during an examination, has been demonstrated. This effect has been shown to be different between the inferior and superior hemifields within one eye (Hudson et al., 1994). In addition, this effect may differ between the first and second eye at the same visit. The number of false-negative answers have been shown to be higher in eyes with field loss (Bengtsson and Heijl, 2000). It has also been shown that there is an inverse relationship between variability and sensitivity (Russell et al., 2012).

That is, there is a large amount of variability in eyes with severe damage.

A common approach to reduce measurement variability is to average multiple measurements. For example, summary measures such as the mean deviation (MD) have less variability due to the averaging over the point-wise sensitivity estimates. Other errors, however, are spatially correlated and affect the whole VF. One group of such errors are measurable factors, including season, time of day and reliability indices, which have been evaluated before (Junoy Montolio et al., 2012; Gardiner et al., 2013). Although these factors are statistically significant, they are rather small and hence only explain a small part of the observed global variation in VFs. Junoy Montolio et al. (2012) modelled the visit effect with these known factors. However, we speculate that other transient factors, such as fatigue, lack of concentration, or delayed reaction time may play a more important role. An example of the importance of these factors can be seen in Figure 6.1, where all locations have a drastic decrease in sensitivity in one of a series of visits. From the longitudinal profiles, it is evident that this decrease is caused by something that affected all VF measurements of that visit, rather than by actual damage.

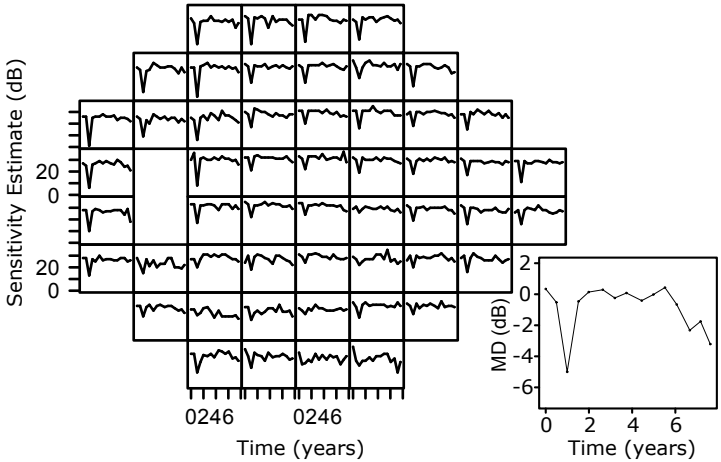

Figure 6.1: Retinal sensitivity estimates over time for each location of the visual field in the left eye of a single glaucoma patient. A decrease in the sensitivity estimates can be seen in all locations at around 1 year. The longitudinal profile of the MD values over time are shown on the right. The visit-dependent decrease is also clear at around 1 year for the MD.

Since all these (as well as possibly other) factors, affect all locations of the same VF, we propose to take them together and to call and model them as Global Visit Effects (GVE). In this way, we can account for both the known and the unknown factors. Hence, the GVE accounts for all factors that affect all measurements of the same eye at each visit. The GVE was inspired by a similar phenomenon, namely, trend estimation of sea levels for large numbers of monitoring stations.

We evaluated the GVE by determining the improvement of the model fit due to incorporating the GVE in the VF model and the effect of including the GVE on estimating the rate of progression. Furthermore, we determined the improvement of point-wise predictions for future measurements accounting for the GVE. Finally, we investigated the magnitude and importance of the GVE by comparing it to influential factors that have been discussed in the literature.

6.2 Methods

6.2.1 Patients and Data

The analysis was performed on a subset of VF data of individuals from an ongoing study conducted at the Rotterdam Eye Hospital, The Netherlands; a full description of the data was provided earlier (Bryan et al., 2013; Erler et al., 2014). All data is available through the Rotterdam Ophthalmic Data Repository at http://rod-rep.com. In brief, the patients were followed up approximately twice per year. The VFs were tested by using the Humphrey Field Analyzer (Carl Zeiss Meditec, Dublin, CA) with the 24-2, white-on-white test strategy by means of the Full Threshold algorithm. The response variables of interest were the sensitivity estimates from the 52 VF points (excluding the 2 points that correspond to the blind spot). All patients gave their written informed consent for participation. The research procedures followed the tenets set forth in the Declaration of Helsinki. We excluded VFs with unknown reliability as indicated by the instrument. Additionally, to simplify the evaluation of the statistical models, we excluded individuals with less than 15 measurements (in either eye). For those individuals with more than 15 measurements, only the first 15 measurements were included in the analysis. The resulting data set consisted of 250 eyes from 125 individuals, resulting in 3750 VFs and 195000 location-specific sensitivity estimates. Descriptive statistics can be found in Table 6.1.

Table 6.1: Descriptive statistics of the study sample

	Mean	Median	Interquartile Range
Baseline age (years)	59.7	61.3	53.2 ; 67.1
Baseline MD (dB)	-7.8	-5.7	-11.7 ; -2.3
Average change in MD (dB/year)	-0.07	-0.04	-0.19 ; 0.13
Follow-up time (years)	7.7	7.6	7.2 ; 8.1

6.2.2 Statistical Modeling

Bayesian models have many advantages over frequentist approaches, such as taking into account the uncertainty in all parameters. Combined with Markov chain Monte Carlo computations, they also allow greater flexibility by relaxing the strong parametric assumptions commonly used in frequentist models. A hierarchical model is able to take into account both the within subject and between subject variability. Moreover, since both eyes were included in the analysis, we are able to capitalize on the common features within each eye by taking into account the correlation between measurements belonging to the same eye. In addition, any correlation of VF measurements within the inferior and superior hemifields, separated by the horizontal meridian, was expected to be higher than between hemifields Asman (1992). Hence, the analysis was done by using a Bayesian hierarchical mixed-effects model (Verbeke and Molenberghs, 2000; Ntzoufras, 2009; Lesaffre and Lawson, 2012). We modeled the hierarchical structure of the data using four levels, namely, (1) the individual, (2) the eye (3) the hemisphere and (4) the location. An example of the mixed-effects model for the four level data structure can be seen in Figure 6.2(A). Furthermore, censoring was taken into account at 0 dB (Tobin, 1958), due to the limitation of the device (Anderson and Patella, 1999). We will refer to this model as the conventional model. To account for the visit-dependent offset at all locations, or GVE, we included a parameter in the model to capture the offset at every visit for each eye within each individual. Hence, this effect accounts for factors that affect all measurements belonging to the same eye at each visit. The impact of this additional parameter is demonstrated in Figure 6.2(B). This model will be referred to as the GVE model.

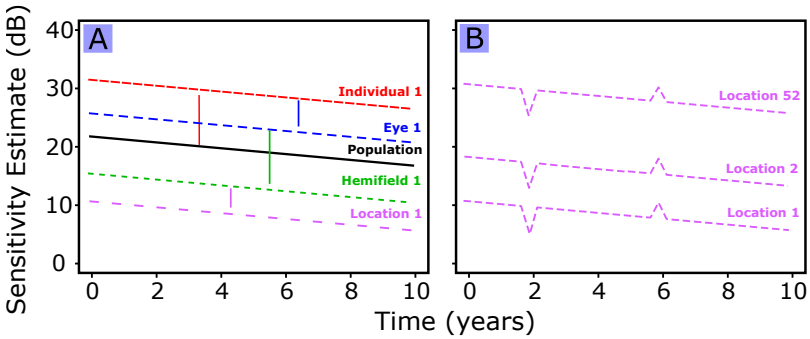

Figure 6.2: (A) The lines represent the estimated evolutions for each of the levels using a hierarchical mixed-effects model. In this example, only the intercept varies, but the model includes slopes in a similar way. (B) The GVE works as an offset for all locations belonging to the same VF. This offset is visible at around 2 years and 6 years in this example.

In classical, one-stage Bayesian hierarchical models, all parameters are estimated simultaneously. For complex models, obtaining results using this approach can be difficult or even impossible due to computational issues. In our case, problems were encountered regarding the running time and computer memory limitations. Thus, the analysis was done by using a recently proposed two-stage approach (Lunn et al., 2013), which allowed us to simplify the computation by splitting the hierarchical model at the individual level. Hence, individuals were analyzed independently before combining them at the population level. Figure 6.3 illustrates the hierarchical structure divided into the two stages. A full description of the models and the computational procedure is given by Bryan et al. (2015) (Bryan et al., 2017).

6.2.3 Model Evaluation

Our aim was to investigate different aspects of the GVE. Namely, how including the GVE affects the model fit, the estimated progression rate, the prediction of future measurements and the magnitude and importance of the GVE. To do this, we compared the conventional model with the GVE model. The models were compared by determining the error for each sensitivity estimate (predicted minus observed) and by combining these errors into summary error measures, namely, the root mean squared error (RMSE) and mean absolute error (MAE) for each model. The 95th percentile of the absolute errors, which is the value below which 95% of the absolute prediction errors may be found, was also computed to compare the models. Non-parametric Wilcoxon (matched paired when applicable) tests were performed to

determine whether the differences between the models were significant.

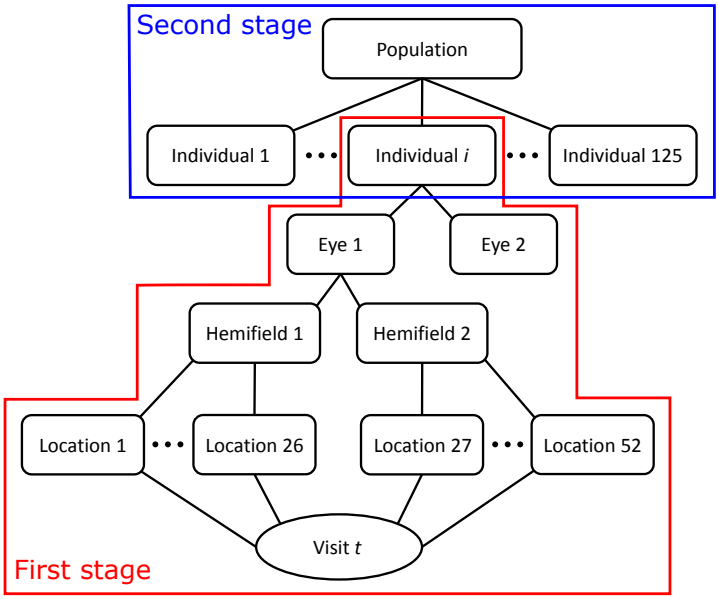

Figure 6.3: Illustration of the hierarchical structure of the data divided into the first and second stages as done in the two-stage approach.

6.2.4 Model Fit

We evaluated the contribution of the GVE by comparing the model fits, for the conventional model and the GVE model, using the RMSE, the MAE and the 95th percentile of the absolute errors. An example of the fits for one eye can be seen in Figure 6.4.

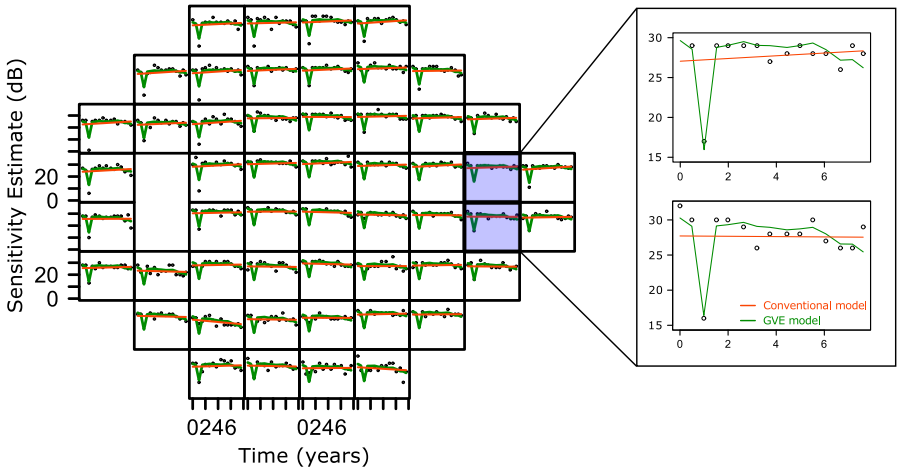

Figure 6.4: Scatter plot representing the retinal sensitivity estimates over time for each location of the VF in an example (left) eye. The lines represent the model fits for the conventional model (orange) and GVE model (green).

6.2.5 Effect on Estimated Progression Rate

To compare the estimated progression rate, or slope, of the conventional model and the GVE model, we first needed to correct for the GVE in the latter model. An example of this is shown in Figure 6.5. The evaluation of the change in slopes was then done by calculating the mean absolute difference of the slopes between the conventional and the GVE model. Additionally, the distribution as well as the 95% confidence interval of the difference between the slopes of the two models was determined.

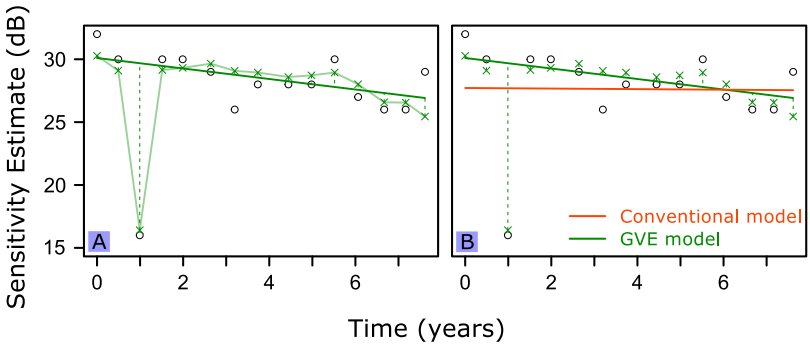

Figure 6.5: A real example showing how the GVE can influence the estimated rate of progression. The open circles represent the retinal sensitivity estimates over time for 1 location in the VF. (A) The crosses represent the fitted data using the GVE model, with the transparent green line showing the model fit over time. The solid green line indicated the slope of the GVE model. (B) The orange line shows the fit for the conventional model. The green line shows the fit for the GVE model after correcting for the GVE estimates.

6.2.6 Prediction of future visual fields

To assess the predictive ability of the model, we performed a 5-fold cross validation. In each fold, the training set included 100 individuals while the testing set contained 25 individuals. Future measurements were then predicted for each location in the VF for each individual in the testing set. The number of measurements used for the estimation of the individual-specific effects was varied (3, 6 and 9 measurements), each time predicting the sensitivity estimate 6 measurements (approximately 3 years) ahead. An example of this can be seen in Figure 6.6. In each of the examples, the GVE model gave a better prediction for the future measurement. The GVE model also appeared to be more robust over the number of measurements included. The models were evaluated by using the RMSE and the 95th percentile of the absolute errors.

6.2.7 Magnitude of the GVE

Junoy Montolio (2012) showed that the time of day, season, reliability indices (number of fixation losses, false negatives and false positive), technical experience and follow up period have a clinically relevant influence on the MD test results (Junoy Montolio et al., 2012). To determine the magnitude and importance of the GVE, we compared it to these factors. We excluded technical experience and follow up period due to the lack of data on these factors.

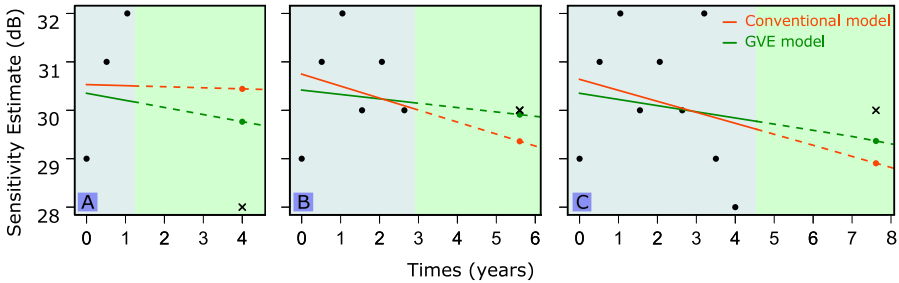

Figure 6.6: Scatterplots represents the retinal sensitivity estimates over time for 1 location of the VF. The predictions were done by using (A) 3, (B) 6 and (C) 9 measurements, which are shown as points in the blue portion of each plot. In each example, the sensitivity estimate 6 measurements ahead, represented by a cross, was predicted. The lines represent the predictions using 3, 6 and 9 measurements to predict 6 measurements ahead for the conventional model (orange) and GVE model (green).

Hence, we focussed on the time of day, season and the reliability indices. For time of day, the tests were stratified into four categories: performed before 10:00 AM, between 10:00 AM and noon, between noon and 2:00 PM, and after 2:00 PM. For season, the tests were also stratified into four categories, of 3 months each (March-May, June-August, September-November, December-February), based on the annual variation of retinal sensitivity (Junoy Montolio et al., 2012). The reliability indices were treated as continuous variables. Reliability indices include the proportion of fixation losses, the proportion of false positives, as well as the proportion of false negatives. We will refer to this model as the fixed factor model. An example of the model fits for one location can be seen in Figure 6.7. We compared the model fits using the RMSE, MAE and 95th percentile of the absolute errors. Furthermore, we determined the magnitude of the GVE compared to the factors by calculating their absolute means. A limitation of the two-stage approach occurs when there is sparse data, such as the season or time of day. Since each individual was analyzed separately, information could not be borrowed from the data set as a whole as done in the classical one-stage approach. Due to this limitation, we used the classical one-stage approach including 50 randomly selected individuals for this analysis.

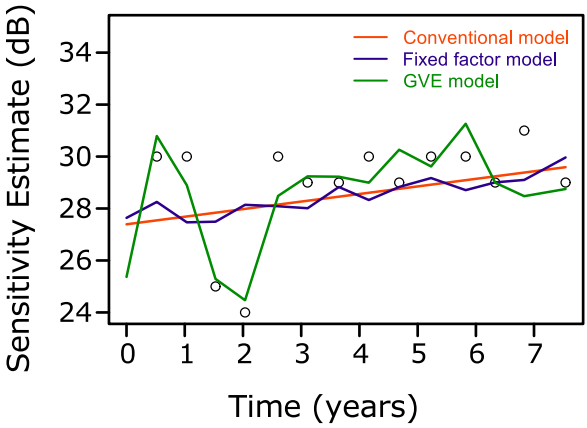

Figure 6.7: Scatterplots representing the retinal sensitivity estimates over time for 1 location of the VF. The lines correspond to the fits of the 3 models, with much larger effects for the GVE (green) than for known, fixed factors (blue).

6.3 Results

6.3.1 Model Fit

Table 6.2 lists the root mean squared errors, mean absolute errors and 95th percentile of the absolute errors for the models, showing that by incorporating the GVE there is an improvement in the model fit. Both the squared errors and absolute errors were significantly smaller for the GVE model than for the conventional model (P<0.001, matched paired).

Table 6.2: Comparison of the fitting ability of the models using the root mean squared error (RMSE), mean absolute error (MAE) and the 95th percentile of the absolute error

Model	RMSE (dB)	MAE (dB)	95th Percentile (dB)
Conventional model	3.23	2.08	7.01
GVE model	2.97	1.91	6.43

6.3.2 Effect on Estimated Progression Rate

When comparing the rate of progression between the conventional model and the GVE model, the mean absolute difference was found to be 0.13 dB/year (P<0.001, matched paired). Figure 6.8 shows the distribution of the differences, including the mean and 95% confidence interval.

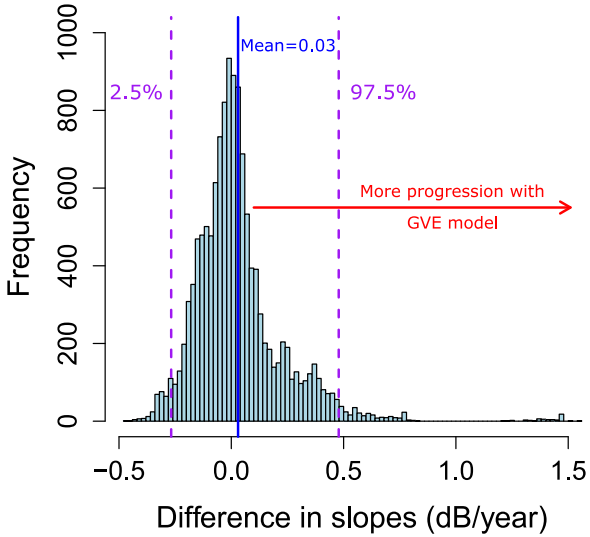

Figure 6.8: Histogram showing the distribution of the differences between the slopes for the conventional model and the slopes for the GVE model. Positive values represent more progression (or smaller slopes) for the GVE model.

6.3.3 Prediction of future visual fields

Table 6.3 lists the RMSE by using 3, 6 and 9 measurements to predict VF sensitivities approximately 3 years ahead (9th, 12th and 15th measurement, respectively). The GVE model showed a significant improvement in the predictions compared to the conventional model, irrespective of how many measurements were used (P<0.001, matched paired). However, the difference between the models predictive abilities decreased as more measurements were included. For the conventional model, there was a significant difference between including 3 and 6 measurements (P<0.001) and between including 6 and 9 measurements (P<0.001). For

the GVE model, these differences were not significant (P=0.08 and P=0.47, respectively).

Table 6.3: Comparison of the predictive ability of the models using the root mean squared error (RMSE), varying the number of measurements included for the estimation. Each time, the measurement 6 visits ahead was predicted

Model	3 measurements (dB)	6 measurements (dB)	9 measurements (dB)
Conventional model	5.63	5.31	4.92
GVE model	4.58	4.65	4.67

6.3.4 Magnitude of the GVE

The fits for the conventional model and the fixed factor model were very similar. This can be seen in Table 6.4, which lists the RMSE, MAE and 95th percentile of the absolute errors for each model. By including the factors, the model fit was slightly improved compared to the conventional model (P<0.001, matched paired). The improvement in the fit was much larger for the GVE model compared to both the conventional model (P<0.001, matched paired) and the fixed factor model (P<0.001, matched paired).

Table 6.4: Comparison of the fitting ability of the models using the root mean squared error (RMSE), mean absolute error (MAE) and the 95th percentile of the absolute error

Model	RMSE (dB)	MAE (dB)	95th Percentile (dB)
Conventional model	3.33	2.14	7.21
Fixed factor model	3.32	2.14	7.18
GVE model	3.07	1.97	6.67

The mean absolute value of each of the factors is shown in Table 6.5, with the most influential factor being season (0.13 dB). The combined value of all factors was found to be 0.23 dB. In contrast, the GVE was found to be 0.85 dB.

Table 6.5: Comparison of the fitting ability of the models using the root mean squared error (RMSE), mean absolute error (MAE) and the 95th percentile of the absolute error

Factor	MAE (dB)
Season	0.13
Time of day	0.09
% Fixation losses	0.00
% False positives	0.10
% False negatives	0.02
Combined	0.23
Global Visit Effect	0.85

6.4 Discussion

In this study, we proposed to model measurement errors that affect the point-wise sensitivity estimates within the same VF as GVEs. By correcting for the GVEs, we accounted for all measureable factors, both known and unknown, that affect all measurements of the same eye at each visit. Furthermore, we evaluated this model and showed the magnitude of the correction for the GVE on progression estimation and prediction of future measurements. The GVE model showed a highly significant improvement in the model fit compared to the conventional model. Hence, by taking the GVE into account, we were able to remove a large systematic component that would otherwise be interpreted as noise and obtain better estimates of the true rate of progression. The mean absolute difference between the rates of progression estimated by the conventional model and the GVE model was shown to be 0.13 dB/year. The distribution was slightly positively skewed. By correcting for the GVE with our model, we found a predominantly higher rate of progression. In other words, ignoring the GVE may obscure the true, most often higher, rate of progression. In this analysis, all measurements were included to determine the difference between the progression rates of the two models.
For this model to be of clinical use, its prediction of future measurements should be better than of the conventional model. By incorporating the GVE in the model, we indeed found significant improvements in the predictions. The GVE model provided more robust predictions than the conventional model, especially when fewer measurements were included. This could be explained by the fact that the GVE model takes factors into account that may either decrease or increase the sensitivity estimates. The GVE has less effect on the estimated slopes when more measurements are included. Thus, the conventional model and the GVE model become more similar as more measurements are included. In clinical practice, the number of visual fields is typically quite small, especially when resources are limited and robust approaches, such as the GVE, are therefore needed to accurately predict glaucomatous

progression.
The importance of the GVE was compared to factors described in the literature that have been shown to be significant. The magnitude of the GVE, however, was shown to be more than 3 times larger than that of the other factors combined (0.85 dB versus 0.23 dB). Our results on those factors were consistent with those found by Junoy Montolio et al. (2012) (Junoy Montolio et al., 2012). Of the known factors, we found season to have the largest effect, with a mean absolute value of 0.13 dB. In agreement with our results, Junoy Montolio et al. (2012) concluded that the number of false positive answers has the largest, or most severe, effect out of the reliability indices. In their study, the MD was determined to be overestimated by 1 dB per 10% of false positive answers. Time of day was found to have a mean absolute value of 0.09 dB. It has, however, been shown that the 24-hour IOP rhythm differs between eyes in glaucoma patients (Liu and Weinreb, 2014). Since the patient's IOP may affect each eye differently at the same time of day, and will vary between individuals, it may be of more interest to determine this factor at an eye-specific level, as done with the GVE. Although the GVE takes both measureable and unmeasurable factors into account, including the known factors allows some explanation of the variation. Hence, a combination of both the GVE and the known factors may be beneficial in the modeling of VF data.
In our model, we determined the GVE for each VF. Thus, the GVE differs per eye for each individual at every visit. In this way, we can take factors into account that affect the two eyes (of the same individual) differently, such as, for example, the fatigue effect. Since the patient has VF tests done on both eyes sequentially, the second eye could be more affected by the lengthy test. Our model can be extended to take the correlation of the GVE between eyes into account. In that way, we could determine the patient-dependent, as well as the eye-dependent GVE. Furthermore, the model may be extended to accommodate that some locations may be affected differently by some of the factors. Although we have shown the effect of the GVE using an advanced Bayesian two-stage model, it is important to note that the GVE can be implemented as a random effect in any point-wise longitudinal model.
In comparison to conventional robust regression which is applied at the location level, the GVE model can be considered a robust approach applied at the VF level. Hence, while the GVE model is able to distinguish between within- and between-field errors, the conventional robust model is not able to do so, because both errors combined in a single error term. By modeling the visit effect explicitly, we exploit the entire VF structure. Hence, the GVE allows us to obtain important additional information, which is needed in order to make any conclusions on the instrument, operator, etc. In this way, we also allow "real" progression to still affect the slope of the model rather than attributing the entire measurement to error as done with the conventional robust model.
An approach which is related to the GVE is the pattern deviation, which uses the cumulative distribution to compute a correction (Anderson and Patella, 1999). The 85th percentile is used to compensate for effects by for instance cataract, which would lead to a general reduction of retinal sensitivity throughout the VF. Hence, the entire VF height is adjusted to the 85th percentile. However, with diffuse loss, the entire VF height tends to be overcorrected

whenever the 85th percentile becomes significantly affected (Artes et al., 2005). The two approaches differ in how the visit effects are expressed. Namely, the percentile correction is treated as a fixed term, while the GVE is treated as a random effect. Because we treat the GVE as a random effect, which has a distribution with mean zero, it is forced to fluctuate around zero. The fixed term of the percentile model does not impose any constraint and hence accounts for both the visit effect and the rate of progression/slope. Hence, in contrast to pattern deviation analysis, the GVE model allows us to estimate the visit effect without disrupting the estimation of the progression.

In clinical practice it is known that one single VF may not be reliable and confirmation is sought in the results from future tests. Clustered VF testing has been shown to identify more rapid progression than evenly spaced follow-up approaches, allowing the clinician to be more confident that the progression is real (Crabb and Garway-Heath, 2012). Our model is in line with these findings and confirms that results which deviate from what is expected may be due to unknown factors that affect the VF measurements on that specific visit rather than representing actual damage.

In conclusion, the GVE has a substantial effect on the point-wise VF sensitivities. In longitudinal VF series, correcting for the GVE provided better estimates of the true rates of progression, which were predominantly higher than with conventional progression analysis. In addition, shorter VF series were required to arrive at relatively accurate VF predictions than with conventional modeling. Implementing progression models that incorporate the GVE in clinical care may therefore improve the clinical management of glaucoma.

Acknowledgments

The authors would like to thank Dimitris Rizopoulos, PhD, for the useful discussions and suggestions which helped improve this manuscript.

Bibliography

Anderson, D. R. and Patella, V. M. (1999). *Automated Static Perimetry*. Mosby, 2nd edition.

Artes, P. H., Nicolela, M. T., LeBlanc, R. P., and Chauhan, B. C. (2005). Visual field progression in glaucoma: total versus pattern deviation analyses. *Investigative Ophthalmology & Visual Science*, 46(12):4600.

Åsman, P. (1992). Glaucoma hemifield test: Automated visual field evaluation. *Archives of Ophthalmology*, 110(6):812–819.

Bengtsson, B. and Heijl, A. (2000). False-negative responses in glaucoma perimetry: Indicators of patient performance or test reliability? *Investigative Ophthalmology & Visual Science*, 41(8):2201–2204.

Bryan, S. R., Eilers, P. H. C., Li, B., Rizopoulos, D., Vermeer, K. A., Lemij, H. G., and Lesaffre, E. M. E. H. (2017). Bayesian hierarchical modeling of longitudinal glaucomatous visual fields using a two-stage approach. *Statistics in Medicine*, 36: 1735–1753.

Bryan, S. R., Vermeer, K. A., Eilers, P. H. C., Lemij, H. G., and Lesaffre, E. M. E. H. (2013). Robust and censored modeling and prediction of progression in glaucomatous visual fields. *Investigative Ophthalmology & Visual Science*, 54(10):6694–6700.

Crabb, D. P. and Garway-Heath, D. F. (2012). Intervals between visual field tests when monitoring the glaucomatous patient: Wait-and-see approach. *Investigative Ophthalmology & Visual Science*, 53(6):2770.

Erler, N. S., Bryan, S. R., Eilers, P. H. C., Lesaffre, E. M. E. H., Lemij, H. G., and Vermeer, K. A. (2014). Optimizing structure and function relationship by maximizing rorrespondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Ophthalmology & Visual Science*, 55(4):2350–2357.

Flammer, J. and Meier, E. (2003). *Glaucoma: A Guide for Patients, An Introduction for Care Providers, A Quick Reference*. Hogrefe & Huber Publishers, Cambridge, MA, 2nd edition.

Gardiner, S. K., Demirel, S., Gordon, M. O., and Kass, M. A. (2013). Seasonal changes in visual field sensitivity and intraocular pressure in the Ocular Hypertension Treatment Study. *Ophthalmology*, 120(4):724–730.

Gloor, B. and Schmied, U. (1980). Findings in glaucomatous visual fields under observation using the "Octopus" automatic perimeter. *Klin. Monatsbl. Augenheilkd.*, (176):545–546.

Gloor, B., Schmied, U., and Fassier, A. (1980). Glaucomatous visual fields analysis of Octopus observation with statistical material. *Klin. Monatsbl. Augenheilkd.*, (117):423–426.

Heijl, A., Lindgren, G., and Olsson, J. (1989). The effect of perimetric experience in normal subjects. *Archives of Ophthalmology*, 107(1):81.

Hudson, C., Wild, J. M., and O'Neill, E. C. (1994). Fatigue effects during a single session of automated static threshold perimetry. *Investigative Ophthalmology & Visual Science*, 35(1):268–280.

Junoy Montolio, F. G., Wesselink, C., Gordijn, M., and Jansonius, N. M. (2012). Factors that influence standard automated perimetry test results in glaucoma: Test reliability, technician experience, time of day, and season. *Investigative Ophthalmology & Visual Science*, 53(11):7010–7017.

Kulze, J. C., Stewart, W. C., and Sutherland, S. E. (1990). Factors associated with a learning effect in glaucoma patients using automated perimetry. *Acta Ophthalmologica*, 68(6):681–686.

Kutzko, K. E., Brito, C. F., and Wall, M. (2000). Effect of instructions on conventional automated perimetry. *Investigative Ophthalmology & Visual Science*, 41(7):2006–2013.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. Wiley & Sons, Chichester, West Sussex.

Liu, J. H. K. and Weinreb, R. N. (2014). Asymmetry of habitual 24-Hour intraocular pressure rhythm in glaucoma patients. *Investigative Ophthalmology & Visual Science*, 55(11):7398–7402.

Lunn, D., Barrett, J., Sweeting, M., and Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 62(4):551–572.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley & Sons, Hoboken, N.J.

Russell, R. A., Crabb, D. P., Malik, R., and Garway-Heath, D. F. (2012). The relationship between variability and sensitivity in large-scale longitudinal visual field data. *Investigative Ophthalmology & Visual Science*, 53(10):5985–5990.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer New York, New York, NY, 1st edition edition.

Werner, E. B., Petrig, B., Krupin, T., and Bishop, K. I. (1989). Variability of automated visual fields in clinically stable glaucoma patients. *Investigative Ophthalmology & Visual Science*, 30(6):1083–1089.

CHAPTER 7

Conclusions

7.1 Concluding Remarks

In this thesis, we explored and developed statistical methods for dealing with high dimensional and complex data. Specifically, we focused on modeling longitudinal VF data from glaucoma patients. In this final chapter we conclude with a summary of principal research contributions in both the statistical methodology and glaucoma research presented in the preceding chapters with a discussion on possible future extensions.

7.2 Summary

Evaluation of a longitudinal series of visual fields (VF), as measured by standard automated perimetry (SAP), provides a way to detect early evidence of glaucoma and to determine functional deterioration. This facilitates the clinician in providing adequate treatment, which may slow down the disease, possibly even halting its progression. However, due to the subjective nature of this technique, SAP is prone to large variability. In order to measure the true progression of the disease, this variability needs to be taken into account.
A simplistic approach to modeling longitudinal VF data is to treat the 52 locations within each eye for every individual as independent samples. In Chapter 3 we compared different regression models, which have been previously investigated in the literature. We evaluated these models with respect to both the fitting and prediction ability for progression in glaucomatous VFs. The models included classical approaches, as well as robust methods and models which take into account censoring.
However, a problem with using separate location-specific regression models is that we are not able to use any information from the data set as a whole. Hierarchical models reduce the variability, caused by randomness, by accounting for group effects and/or within-group correlation. In Chapter 4 we extended this work to account for the complex structure of the data while dealing with censoring and the high measurement variability using advanced Bayesian hierarchical models.
Nowadays it is often of interest to fit hierarchical models to data that are big both in terms of volume and complexity. However, it is also recognized that MCMC sampling can become computationally prohibitive when applying complex models. To overcome these issues, we applied a recently proposed two-stage approach which allowed us to speed up computations considerably while still benefiting from the full Bayesian approach. Furthermore, we extended this approach to allow the calculation of model selection and model evaluation criteria. Other approaches such as integrated nested Laplace approximation (INLA) and hierarchical likelihood (H-likelihood) are also attractive alternative options. In Chapter 5 we investigated and compared these approaches to the classical one-stage and two-stage approaches.
For the clinical application, we proposed to model factors which affect all locations belonging to the same VF as global visit effects (GVEs). Details and extensive evaluations of the GVE were shown in Chapter 6. We evaluated this by determining the improvement of the model fit

due to incorporating the GVE in the model and the effect that this has on estimating the rate of progression. Furthermore, we determined the improvement of point-wise predictions for future measurements. Finally, we investigated the magnitude and importance of the GVE by comparing it to influential factors that have been discussed in the literature.

7.3 Future Research

The work presented in the previous sections provides foundations for interesting extensions. In the following section, we highlight several important aspects which were not discussed in the body of this thesis and which provide potentially interesting directions for future investigation.

7.3.1 Translate the GVE into Clinical Practice

Correcting for the GVE has a number of consequences, including changes in the estimated progression rate and the expected VFs of future visits. The consequences of these changes are still insufficiently mapped. An investigation of these changes, for which patients they occur, and how they would be best used could ultimately improve clinical care. The software that was used for the development of the statistical models described in this thesis is not straightforward and would be difficult to use for a non-statistician. In addition, we have focused on processing off-line VFs, but it would be beneficial to analyze data from a new patient. To allow a better clinical evaluation, it would be useful to have an implementation of the model built in a web environment. This would allow the clinical researcher, or clinician, to directly access and analyse the VFs of a particular patient.

7.3.2 Spatial Nature

An aspect of the data which was only briefly looked at by including the hemifield effect, is the spatial nature of the data. One could exploit this by capitalizing on the specific spatial organization of the nerve fibres in the eye. Knowledge about the physiology of the eye, particularly the topology of the nerve fibres in the retina, suggests that the assumption of a standard neighbourhood structure (e.g. direct horizontal and vertical neighbours) in the visual field data is incorrect. Rather, it seems more plausible to define a neighbourhood that is based on the paths of the nerves (Airaksinen et al., 2008; Denniss et al., 2012). Work has been done on transforming the data from the original x-y-coordinates to a different space, where the location is given in coordinates that are based on the geometry of the retinal nerves (Erler et al., 2014). This could be extended to incorporate the longitudinal nature of the data. Spatial models such as those used for disease mapping could be explored.

7.3.3 Structural and Functional Relationship

Both structural and functional tests are routinely used in clinical practice for diagnosis and monitoring of glaucoma. Functional tests, such as perimetry, represent the most important clinical endpoint, but are highly subjective and variable. This limits the clinical utility and these tests often have to be repeated before a clear result is obtained. This means that a necessary adaptation of the treatment is delayed. Structural tests, provided by several imaging modalities, are objective, fast and reliable, but it has proven to be difficult to show substantial correlation with functional tests. Finding a good correlation would have a significant clinical impact. The research included in this thesis only focused on functional data. It may however be of interest to combine this with the structural data in order to improve detection of disease progression. This can be done by using a joint modeling approach (Rizopoulos, 2012). The formulation of the association structure between the functional and structural measurements are complicated by the fact that we have serial measurements of these two outcomes. In particular, it is not directly evident which features of a patient's longitudinal trajectory of the structural outcome are most predictive for the functional or vice versa. Some potential characteristics that we would like to investigate are the level of structural damage, the rate of progression of the structural damage and suitable summaries of the whole longitudinal trajectory of the structural damage that could be shown to be related to functional deterioration at different time points. This approach might contribute to fighting blindness and severe visual impairment caused by glaucoma.

Bibliography

Airaksinen, P. J., Doro, S., and Veijola, J. (2008). Conformal geometry of the retinal nerve fiber layer. *Proceedings of the National Academy of Sciences*, 105(50):19690–19695.

Denniss, J., McKendrick, A. M., and Turpin, A. (2012). An anatomically customizable computational model relating the visual field to the optic nerve head in individual eyes. *Investigative Ophthalmology & Visual Science*, 53(11):6981.

Erler, N. S., Bryan, S. R., Eilers, P. H. C., Lesaffre, E. M. E. H., Lemij, H. G., and Vermeer, K. A. (2014). Optimizing structure-function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Ophthalmology & Visual Science*, 55(4):2350–2357.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Biostatistics. Chapman and Hall/CRC, Boca Raton.

CHAPTER 8

Nederlandse Samenvatting, Acknowledgements, CV and PhD Portfolio Summary

Nederlandse Samenvatting

Het beoordelen van longitudinale reeksen van gezichtsvelden (visual fields; VFs), gemeten met standaard geautomatiseerde perimetrie (standard automated perimetry; SAP), maakt het mogelijk om de effecten van glaucoom in een vroeg stadium te detecteren en om de functionele veranderingen te bepalen. Dit stelt een clinicus in staat om een adequate behandeling in te zetten, waardoor de ziekte vertraagd wordt en verdere progressie mogelijk zelfs wordt voorkomen. Door de subjectieve manier van meten wordt SAP echter geplaagd door een grote meetvariabiliteit. Om de echte progressie van de ziekte te kunnen meten moet met deze variabiliteit rekening worden gehouden.
Een eenvoudige benadering om longitudinale VF gegevens te modelleren is de 52 locaties binnen elk oog te beschouwen als onafhankelijke gegevens. In hoofdstuk 3 hebben we verschillende regressiemodellen, die eerder in de literatuur beschreven zijn, vergeleken. Deze modellen werden vergeleken op basis van hoe goed zij de data beschrijven en progressie voorspellen in glaucomateuze VFs. De modellen omvatten zowel standaard technieken als robuuste methodes en modellen beide rekening houdende met censurering. Een nadeel van deze aparte, locatie-specifieke regressiemodellen is dat ze niet de samenhang van de ooglocaties in rekening brengen. Dit gebeurt wel met hiërarchische modellen. In hoofdstuk 4 hebben we onze methode uitgebreid om rekening te kunnen houden met de complexe structuur van de data, censurering en de grote meetvariabiliteit, door middel van geavanceerde Bayesiaanse hiërarchische modellen.
Er is een groeiende noodzaak om hiërarchische modellen te fitten op grote en complexe datastructuren. Echter Monte Carlo Markovketen (MCMC) methodes zijn inherent rekenintensief en de huidige computerinfrastructuur van een statistische onderzoeker laat nog niet toe deze modellen steeds in een aanvaardbare tijdspanne op grote complexe datastructuren te schatten. Om dit probleem te ondervangen hebben we een tweetrapsmodel, dat recentelijk is ontwikkeld, toegepast. Hiermee werd de rekentijd sterk beperkt terwijl de voordelen van een Bayesiaanse aanpak nog steeds gelden. Daarnaast hebben we deze nieuwe methode uitgebreid zodat modelselectie-criteria hierop konden worden toegepast. Aantrekkelijke alternatieven, zoals geïntegreerde, geneste, Laplaciaanse benaderingen (integrated nested Laplace approximation; INLA) en H-likelihood werden onderzocht in hoofdstuk 5 en vergeleken met conventionele één- en tweetrapsmodellen.
Voor onze klinische toepassing hebben we allerlei gemeten factoren in rekening gebracht, zowel op subject-als op VF niveau, die een invloed kunnen uitoefenen op het de progressie van glaucoom. We hebben echter ook een latente variabele in het model die alle VFs zou kunnen beïnvloeden, genoemd GVE. Dit GVE (global visit effect) representeert een effect dat zich in de tijd afspeelt en tegelijk op alle locaties in het oog werkt. Een voorbeeld hiervan is de vermoeidheid van de patient op een bepaalde dag. Zo'n effect kan het predictieproces ernstig beïnvloeden en moet daarom in rekening gebracht. De details en een uitgebreide evaluatie van deze GVEs staan beschreven in hoofdstuk 6. We hebben dit geëvalueerd door te bepalen hoe groot het effect is de verbetering van het model is wanneer het GVE hierin wordt opgenomen

en wat het effect hiervan is op de schatting van de progressiesnelheid. Daarnaast is gekeken naar de verbetering van de voorspellingen van punten in het VF. Ten slotte werd de grootte en het belang van het GVE onderzocht door dit te vergelijken met andere belangrijke factoren die bekend zijn uit de literatuur.

Acknowledgements

I would like to thank Prof. Emmanuel Lesaffre for giving me the opportunity to do my PhD in the Netherlands. I would also like to thank Prof. Paul Eilers, Dr Hans Lemij and Dr Dimitris Rizopoulos for their support and contributions to this work. A special thanks to Dr Koenraad Vermeer for teaching me so much over the years. Working with you was a pleasure.

To my former colleagues from the department of Biostatistics and the Rotterdam Ophthalmic Institute, thank you for the useful discussions and support along the way. Specifically, I would like to thank Elrozy for always being there for me and never letting me give up. Thanks also to Eline for her help over the years.

Finally, thanks to my family and friends in South Africa and the UK for their ongoing support. Mom and Dad, thank you for everything you've done for me over the years and for giving me every possible opportunity. Sarah and Stephen, thank you for all of your advice and for always being there for me.

CV

The author was born on March 30, 1987 in Pietermaritzburg (South Africa). She received her Bachelor of Science and Bachelor of Science Honours from the University of Kwa-Zulu Natal (South Africa) in 2008 and 2009 respectively. Afterwhich she received a bursary from the South African Centre for Epidemiological Modelling and Analysis (SACEMA) to persue her Master's degree. In 2012 she received her Master of Science Cum Laude.
In April 2011 she commenced her PhD project under the supervision of Prof. Emmanuel Lesaffre, Prof. Hans Lemij, Prof. Paul Eilers and Dr. Koenraad Vermeer. During this research, she also participated in projects at the department of Biostatistics and the Rotterdam Ophthalmic Institute. In 2015, she started research fellowships at the department of Optometry and Visual Science at City University London and at the Rotterdam Ophthalmic Institute.

PhD Portfolio Summary

Publications

Bryan, S.R., Lesaffre, E.M.E.H., van Rosmalen, J.M., Noh, M., Lee, Y. and Rizopoulos. Comparison of estimation methods for multilevel models. *Submitted*.

Bryan, S.R., Eilers, P.H.C., van Rosmalen, J., Rizopoulos, D., Vermeer, K.A., Lemij, H.G. and Lesaffre, E.M.E.H. (2017). Bayesian Hierarchical Modeling of Longitudinal Glaucomatous Visual Fields using a Two-Stage Approach. *Statistics in Medicine*, 36:1735 - 1753.

Bryan, S.R., Eilers, P.H.C., Lesaffre, E.M.E.H., Lemij, H.G. and Vermeer, K.A. (2015). Global Visit Effects in Point-wise Longitudinal Modeling of Glaucomatous Visual Fields. *Investigative Ophthalmology & Visual Science*, 56(8):4283 – 4289.

Thepass, G., Pel, J.J.M., Vermeer, K.A., O. Creten, O., Bryan, S.R., Lemij, H.G. and van der Steen, J. (2015). The Effect of Cataract on Eye Movement Perimetry. *Journal of Ophthalmology*, vol. 2015, Article ID 425067, 9 pages. doi:10.1155/2015/425067.

Erler, N.S., Bryan, S.R., Eilers, P.H., Lesaffre, E.M., Lemij, H.G., Vermeer, K.A. (2014). Optimizing structure-function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Ophthalmology & Visual Science*, 55(4):2350 - 2357.

Bryan, S.R., Vermeer, K.A., Eilers, P.H.C., Lemij, H.G. and Lesaffre, E.M.E.H. (2013). Robust and Censored Modeling and Prediction of Progression in Glaucomatous Visual Fields. *Investigative Ophthalmology & Visual Science*, 54(10):6694 – 6700.

Conferences

Oral presentations

- 36th Annual Conference of the International Society for Clinical Biostatistics (ISCB) - Utrecht, Netherlands, 2015
- International Biometric Conference (IBC) Channel Meeting - Neijmegen, Netherlands, 2015
- The Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting - Denver, USA, 2015

- IMAGE - Canterbury, UK, 2015
- International Perimetric Society (IPS) - New York, USA, 2014
- 27th International Biometric Conference (IBC) - Florence, Italy, 2014
- Bayesian Biostatistics (BB) - Zurich, Switzerland, 2014
- The Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting - Orlando, USA, 2014
- 4th Bayes - Rotterdam, The Netherlands, 2013
- 33rd Annual Conference of the International Society for Clinical Biostatistics (ISCB) - Bergen, Norway, 2012

Poster presentations

- 34th Annual Conference of the International Society for Clinical Biostatistics (ISCB) - Munich, Germany, 2013
- The Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting - Seattle, USA, 2013

Attended

- The Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting - Fort Lauderdale, USA, 2012

PhD Training

- GAMLSS in action, Rotterdam, The Netherlands, 22-23 January 2014
- Bayesian inferences for latent Gaussian models using INLA, Utrecht, The Netherlands, 24 January 2013
- An introduction to joint modeling of longitudinal and survival data, Rotterdam, The Netherlands, 18 - 19 June 2013
- Survival Analysis, Erasmus Winter Programme, Rotterdam, The Netherlands, 30 January - 3 February 2012
- Variable selection: Spring symposium, Rotterdam, The Netherlands, 9 March 2012
- Bayesian adaptive methods for clinical trials, Rotterdam, The Netherlands, 8 - 9 October 2012
- Bayesian methods in medical research, NIHES, Rotterdam, The Netherlands, 2012
- Bayesian variable selection and model choice for structured additive regression using spike-and-slab priors, Rotterdam, The Netherlands, 28 September 2011

- Missing data in longitudinal studies: strategies for Bayesian modeling, sensitivity analysis, and causal inference, Rotterdam, The Netherlands, 20 - 21 October 2011
- Bayesian Biostatistics, Leiden University, Leiden, the Netherlands, 2011
- Statistical Computing with R, Leiden University, Leiden, the Netherlands, 2011

Assisting Teaching

- Bayesian Statistics (nihes, CE09), 2014
- SAS practical, 2012-2015
- Biostatistical methods II (nihes, EPO3), 2012-2014
- Repeated measurements (nihes), 2013
- Classical Methods in Biostatistics (nihes), 2011
- Introduction to epidemiology, (nihes) summer programme, 2012
- Biostatistical Methods I (nihes, CC02), 2012, 2014

Invited Speaker

- Bayesian Hierarchical Modelling of Longitudinal Glacuomatous Visual Fields using a Two-Stage Approach. Statistics seminar at the University of Kwa-Zulu Natal, South Africa, 2014

Awards

- ARVO Travel Grant (The Association for Research in Vision and Ophthalmology) - Orlando, USA, 2014