Limburgs Universitair Centrum

Faculteit Wetenschappen

Flexible Statistical Modelling: Application to Infectious Diseases and Astronomical Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, Groep Wiskunde
aan het Limburgs Universitair Centrum te verdedigen door

Ziv SHKEDY

Promotors:

Prof. dr. Marc Aerts

Prof. dr. Geert Molenberghs

2003

To the memory of my parents, Ada and Moshe Shkedy

Listening to you I get the music.

Gazing at you I get the heat.

Following you I climb the mountain.

I get the excitement at your feet!

Right behind you I see the millions.

On you I see the glory.

From you I get opinions.

From you I get the story.

(The Who, Tommy, 1969)

Acknowledgment

This work could not be completed without the assistance and support of numerous persons. I would like to take this opportunity to acknowledge their help.

I am deeply grateful to my two supervisors Prof. Dr. Marc Aerts and Prof. Dr. Geert Molenberghs, whose support, advice, guidance and encouragement affected the entire work on this thesis. It was a pleasure to study from you so many things which contributed not only for the "scientific part" of my life but to other parts as well. Thank you so much.

The work on the first part of the thesis is a collaboration with Philippe Beutels and Pierre Van Damme (Centre for the Evaluation of Vaccination, University of Antwerp). I thank them both for many discussions and suggestions and their professional insight about the problem of infectious diseases.

The second part of this work could not be completed without the help and support of Conny Aerts and Leen Decin (Instituut boor Sterrenkunde, Katholieke Universiteit Leuven). Special thanks go to Leen for the pleasant time that I have while working with you on the two papers which lead to the second part of this thesis.

During my stay at LUC I received help from all the academic and the secretarial staff at the Center of Statistics. I am greatly indebted to all of them for their support and patience. I would like, in particular, to thank the members of the "D_3 order" (and alumni): Tomasz Burzykowski, Lien Beunckens, Didier Renard, Veerle Vandersmissen and Suzy Van Sanden whose support, especially in the last few months, contribute to the pleasant atmosphere in which I worked.

Special thanks to my family in Israel whose love made the work on this thesis possible.

Finaly, let me acknowledge the financial support from "Bijzonder Onderzoeksfonds LUC" that allowed me to come to LUC and work on the thesis.

Ziv Shkedy

May 22, 2003

Contents

Chapter 1

Introduction

1.1 Modeling Infection Disease Data

In his historical novel *London* Edward Rutherfurd (1998) devoted a chapter to the plague epidemic which broke out during the Summer of 1665. Rutherfurd (1998) described the spread of the disease from a point of view of a medical doctor, Dr. Meredith, who stayed in London during that awful Summer, tried to treat his patients. Dr. Meredith did not worry at the beginning of the Summer. As Rutherfurd (1998) wrote:
"...Doctor Meredith had not taken much notice of the trouble when a few cases appeared in May. Sporadic visits like this had been a feature of summer in London for centuries...No significant outbreak had occurred, he remained himself, for nearly twenty years and nothing really major since the regime of King James I..."
A few weeks later, while reading the Bill of Mortality, Dr. Meredith realized that the disease broke out:
"...The Bill of Mortality was a document produced every week. In two long columns it noted the numbers who had died, of each of some fifty causes, in the city and surrounding parishes of London. Most of the numbers were small. Apoplexy:1. Dropsy:40. Infants:21. But near the top of the second column, the clerk had pointed to one, terrifying number:1843. And besides it a single, awful word:Plague..."
The disease spread rapidly, as we read on page 896
"..By mid-August the Mortality Bill was four thousand a week; by the end of August, six thousand..."
The numbers in practice were much higher, as Dr. Meredith discovered later. The Mortality Bill suffered from underreporting:
"...Pepys was an official at the Navy Board and , Meredith knew, had access to information of all kind. The real number of deaths is higher than the Mortality Bill shows, Pepys told him. The clerks are falsifying the accounts and some of the poor aren't being counted. The bills show seven and a half last week. And the real figures ? Neared ten, Pepys replied grimly..."
The above quotes from Rutherfurd (1998) are given only for illustration. While reading this chapter from *London*, one can easily imagine the panic in the London population when the Plague killed so many people, while no one really knew how to stop it and not even how it spread out.

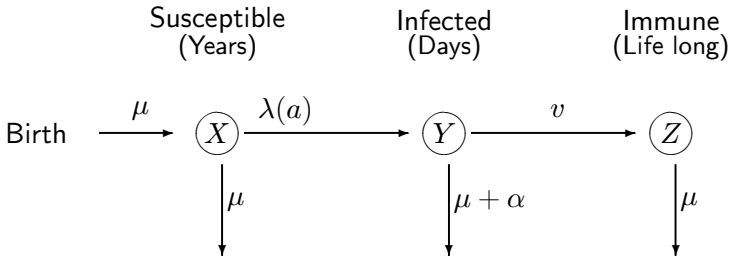


Figure 1.1: *Illustration of the SIR model. The individuals are entered into the susceptible class, then move to the infected class and after recovering they move into the immune class. The parameters μ, λ, α and v will be discussed further in Section 2.1.*

The first part of this thesis is devoted to statistical modeling of infectious diseases data, mainly cross-sectional seroprevalence data. By an infectious disease we mean a disease which is infectious in the sense that an infected host passes through a stage, called the infectious period, during which he/she is able to transmit the disease to susceptible individuals (Becker 1989). Figure 1.1 shows the flow of individuals between the infection classes for a typical childhood infectious disease (Anderson and May 1991). Individuals enter the susceptible class (X) at birth, become infected (Y), recover and gain life long immunity (Z). Figure 1.1 is a graphical representation of the SIR (Susceptible-Infected-Removed) model. Of course, more complicated models, assuming maternal antibody and a latent period can be used to describe the flow of individuals within the disease states. A fundamental parameter of infection disease epidemiology is the force of infection. It is the rate in which susceptibles are transfered from the susceptible to the infection class. It is assumed that the force of infection ($\lambda(a)$ in Figure 1.1) varies across age groups. For example, the force of infection for rubella is much higher at age 0-10 than the force of infection at age 10-20. In the first part of this thesis we focus on statistical models for an age-dependent force of infection. Our starting point will be the three land mark papers of Muench (1394) Griffiths (1974) and Grenfell and Anderson (1985). All of them fit the force of infection taking into account the underlying catalytic model, where the model for the force of infection evolves from the constant model of Muench through the linear model of Griffiths to the flexible polynomial model of Grenfell and Anderson.

In Chapter 2 we present an overview of the mathematical model which describes the transmission dynamics of infectious diseases. In the second part of Chapter 2 we make the link between the mathematical model, which represents the mechanism behind the data, and the statistical models that can be used in order to estimate the force of infection from serological data. Five serological datasets, which will be used in later chapters for illustrations, are presented in the third part of Chapter 2.

Chapter 3 introduced the fractional polynomial approach (Royston and Altman, 1994), in the context of binary regression, as a parametric model for the force of infection.

In Chapter 4 local polynomial models (Fan and Gijbels, 1996) are used in order to estimate the force of infection in a nonparametric fashion.

Since local linear and quadratic polynomial models provide consistent estimates for the probability to become infected before age a and its first derivative, we can derive a consistent estimate for the local force of infection. Furthermore, it will be shown that, using the asymptotic distribution of the estimate of the force of infection, one can choose the

optimal bandwidth, minimizing the asymptotic mean square error of the force of infection. In Chapter 5 we present a simulation study in order to investigate the performance of the local polynomial models compared to the isotonic regression model.
While in Chapters 3 and 4 frequentist methods are used to estimate the force of infection, we shift in Chapter 6 to the framework of hierarchical Bayesian models. Nonlinear and generalized linear models are presented in the first part of Chapter 6. Since several parametric models are fitted to the same data, a model selection procedure, based on the *deviance information criterion* is applied in order to select the model with the best goodness-of-fit. A beta-binomial model, which can be seen as a Bayesian version of the isotonic regression model of Kieding (1991), is presented in the second part of Chapter 6.
In Chapter 7, the last chapter which presents models for the force of infection, we use a Bayesian model with Dirichlet process prior for the prevalence. It will be shown that using a Dirichlet process prior (Ferguson, 1973) for the prevalence, allows us to incorporate parametric models (such as nonlinear or generalized linear models) as a prior model for both the prevalence and the force of infection.
Lastly, Chapter 8 presents a meta-analysis that was conducted to estimate the probability to become hepatitis B carrier. In the first part of the thesis, this is the only chapter which does not deal with the estimation of the force of infection. The probability to become hepatitis B carrier, assumed to be age-dependent, is used in order to describe the flow of individuals from the infected class to the carrier class of the mathematical model. The dataset was previously analyzed by Edmunds *et al.* (1993), which assumed that the probability to become a hepatitis B carrier is constant during the perinatal period (from birth up to 6 months) and drops down exponentially thereafter. We use a hierarchical Bayesian changepoint model (Carlin *et al.* 1992) in order to model the age in which the probability starts to drop down. In this type of model, the end of the perinatal period is not fixed but it is a parameter in the model (the changepoint).

1.2 Modeling Astronomical Data

In her Chapter "Late night thoughts of a classical astronomer" (Babu and Feigelson Eds. 1997) Virginia Trimble (1997) summarized her thoughts before going to a joint conference of statisticians and astronomers, entitled *Statistical Challenges in Modern Astronomy*, held in Penn State university in June 1996, in the following way: "*I arrived at Penn State with a prediction at hand. Nobody is going to learn anything at this meeting*". Her last sentence in the chapter is "*How can we foster collaborations between statisticians and astronomers that will be attractive to both in the sense of advancing basic knowledge in both fields so that each collaborator has something significant to add to his CV at the end ? Participants nodded solemnly and promised to Think About it All*".
In the second part of this thesis we present a research in which we "implemented" the collaborative idea. We focus on the estimation of stellar parameters of a cool star. The estimation procedure consists of a comparison between a theoretical spectrum and the observed spectrum of a star. The calculation of the theoretical spectrum is a time demanding computer procedure in which the model atmosphere of the star is solved and the theoretical spectrum is calculated from the relative transfer equations. This theoretical spectrum consists of dozens of parameters representing the structure of the star under consideration. Once a synthetic spectrum is calculated, it should be compared to the

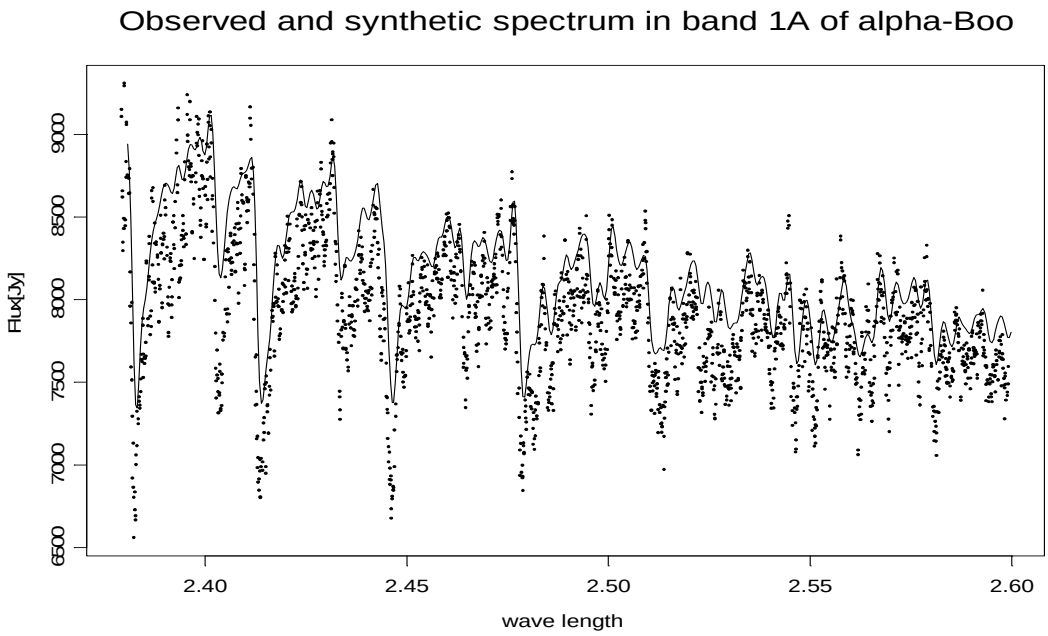


Figure 1.2: *Observed (non rebinned) spectrum (dots) and synthetic spectrum fitted with* $T_{eff} = 4440$*, log g=1.80* $[F_E/H] = 0.00$

observed spectrum. Figure 1.2 shows the observed spectrum of the cool star α-Bootis and a theoretical spectrum (which we call the *synthetic spectrum*). The aim is to choose the synthetic spectrum with the best goodness-of-fit and to estimate the parameters of the star based on this model. Timble (1997) formulated a few questions related to model fitting and goodness-of-fit assessment which we use to introduce the structure of the second part of this thesis.

How do you find your way through a many-dimensional parameter space to a best fit ?

In this thesis we follow the approach introduced in Decin (2000), which we call the *zoom in* approach. The basic idea is to focus on the three most important parameters, effective temperature, gravity and metalicity, and to fix the others in order to calculate a collection of synthetic spectra from which we choose the model(s) with the best goodness-of-fit. Then a new collection of synthetic spectra is calculated based on a more sensitive grid of the affective temperature, gravity and metallicity. The new grid of the parameters is chosen based on the result of the first set of models. Once the new collection of spectra is calculated, it should be compared to the observed spectrum in order to choose the model with the best fit to the data.

What should replace the chi-square as a test for goodness-of-fit ?

Decin (2000) proposed to use the Kolmogorov-Smirnov statistic as model selector. After a short introduction chapter (Chapter 9) which introduces the astronomical and the statistical problems related to the estimation of stellar parameters, we focus, in Chapter 10 and 11, on sensitivity analysis. One should realize that the high resolution observed spectrum (the dots in Figure 1.2) is not used in the analysis since it is not in the same resolution as the synthetic spectrum. For comparison, the high resolution spectrum is reduced (by a rebinning procedure) to the so-called *rebinned data* or the *observed spectrum*. In Chapter 10 we reduce the high resolution spectrum with cubic smoothing splines and base the

model selection procedure on this estimator for the spectrum. We compare the results to those obtained when the rebinned data are used for model selection. The performance of the Kolmogorov-Smirnov statistic as model selector is compared to the performance of the least square criterion as model selector. It will be shown in Chapter 11 that the global Kolmogorov-Smirnov statistic has a local version when lack-of-fit tests are applied to the data.

Suppose the template you are trying to fit itself has uncertainties (e.g. in atomic data for spectral lines); how can this be included in error estimates in Bayesian and frequentist methods ?

Chapters 12 and 13 are devoted to Bayesian analysis. In Chapter 12 we discuss the use of linear mixed models as scatterplot smoothers and apply the method to the observational errors. In Chapter 13 we present an hierarchical Bayesian model for the spectrum. This model accounts for two sources of errors, the first represents the measurement error associated with the observed spectrum and the second represents the accuracy associated with the synthetic spectrum under consideration. The first (STDEV-tag) is included as the variance in the likelihood of the observed spectrum and the second (SPARE-tag) as the variance in the prior distribution of the "true" spectrum. It will be shown that the posterior mean for the spectrum can be expressed as a weighted average between the observed spectrum and the synthetic spectrum. Since there are 125 synthetic spectra under consideration, a model selection procedure, based on the predictive distribution of the spectrum is applied in order to select the model with the best goodness-of-fit.

Part I

Modelling Infectious Diseases Data

Chapter 2

Mathematical and Statistical models for Infectious Diseases Data: An Introduction

2.1 Mathematical Modeling

This section introduces several concepts of infectious diseases, focusing on mathematical models for the disease process. Mathematical models describe the flow of individuals between the different infection states within the population. The goal of these models is to describe the mechanism of the disease, to predict the proportion of individuals at each infection class and to give clear insight about the effect of mass vaccination programs on the population. The discussion in this section is mainly based on Anderson and May (1991), Halloran (1998) and Dietz (1993).

2.1.1 The Basic SIR Model

It is assumed (Anderson 1982, Anderson and May 1991, Bailey 1975) that for simple microparasitic infectious diseases that simulate long-lasting immunity following infection the individuals in the population can be classified, according to their infection status, into three states:

- **Susceptible to infection:** individuals who have not been exposed yet, this is the population at risk. The number of hosts at risk at time t and age a is denoted by $X(a,t)$.
- **Infected and infectious to others:** $Y(a,t)$ is the number of infected hosts at time t and age a.
- **Immune to reinfection:** $Z(a,t)$ is the number of immune hosts at age a and time t.

The total population is given by

$$N(a,t) = X(a,t) + Y(a,t) + Z(a,t).$$

This type of model is called the *SIR* model (Susceptible, Infected, Recovered). It assumes that all newborns are entered directly into the susceptible class and therefore ignores the maternal antibodies period, in which newborns are protected for a few months by maternal antibodies. Another assumption is that the infection, the infectious period and the disease occur simultaneously, i.e. that the *SIR* model ignores the latent period (the period in which the individual is infected but not infectious to others).

The duration of stay within each class varies from class to class. For example, in the case of measles (in the developed countries) the mean age of exposure is approximately 4–5 years, the infectious period is around 7 days and the immunity is life long. These lengths of time hold for most of the childhood infectious diseases, the susceptible period lasts years, the infectious period days and the immunity is assumed to be life long. The proportion of the population within each class depends on the average duration of stay within the class, therefore it is expected that at a given time (or age) the proportion of infectious individuals will be smaller relatively to the proportion within the susceptible and the immune class.

2.1.2 Transmission

A central characteristic of the population dynamics of infection diseases is the transmission of the infection from the infected class (assumed to be infectious in the SIR model) to the susceptible class. The density of new infected individuals at time t depends on the density of infected individuals ($\bar{Y}$), the density of susceptibles ($\bar{X}$) and the rate of effective contact between the two groups (β). The mass-action principle states (Bailey 1975):

$$\text{New cases of infection} = \beta \bar{Y} \bar{X}.$$

The underlying assumption of the mass-action principle is that the infectious and the susceptible individuals are mixing in a random manner (homogeneous mixing), i.e., that β is age and time independent. The *force of infection* (λ), is the rate at which the host moves from the susceptible to the infected class. It is assumed that λ is a linear function of the total number of infected hosts ($\bar{Y}$). As the number of infected hosts increases, the probability to move from X to Y increases (although the probability of transmission β stays constant). The relation between the total number of infected individuals and λ is given by

$$\lambda = \beta \int_0^\infty Y(a,t)da = \beta \bar{Y}.$$

If we assume that the lifetime in the susceptible class is an exponential random variable with parameter λ, then the expected time in the susceptible class is simply $1/\lambda$.

2.1.3 Mathematical Models for Transmission Dynamics

Bailey (1975, 1982), Anderson (1982) and Anderson and May (1991) proposed a set of 3 partial differential equations to describe the flow of the individuals within the population

with respect to time and host age:

$$\begin{aligned} \frac{dX}{da} + \frac{dX}{dt} &= N\mu - (\lambda(a,t) + \mu)X(a,t), \\ \frac{dY}{da} + \frac{dY}{dt} &= \lambda X - (v + \alpha + \mu)Y(a,t), \\ \frac{dZ}{da} + \frac{dZ}{dt} &= vY - \mu Z(a,t). \end{aligned} \tag{2.1}$$

Here, μ is the natural rate of death (the life expectancy is equal to $1/\mu$), $\lambda(a,t)$ is the force of infection for age a at time t, v is the recovery rate and α is the rate of death caused by the disease. Note that this model assumes that the force of infection depends on the host age and the time but all the other parameters in the model are age and time independent. The model can be expressed with age dependent parameters, $v(a)$, $\alpha(a)$ and $\mu(a)$ as well. The change in the number of susceptibles, $X(a,t)$, is equal to the difference between the rate at which the individuals leave the susceptible class $(\lambda(a,t)+\mu)$ times the population at risk $X(a,t)$ and the number of newborns who are entered into the susceptible class $(N\mu)$. The change in the number of infected individuals is the difference between the number of the newly infected $(\lambda(a,t)X(a,t))$ and the number of infected individuals developing immunity or dying $(v+\mu+\alpha)Y(a,t)$. The number of "new" immunes is equal to $vY(a,t)$ while the number of immune hosts that die and leave the population is $\mu Z(a,t)$. Adding all 3 equations gives $dN/da + dN/dt = 0$ which corresponds to the assumption that the total population is constant.

2.1.4 Equilibrium: The Static Model

At equilibrium, the rate of change of each variable in the model does not depend on the time. That is

$$\frac{dX}{dt} = \frac{dY}{dt} = \frac{dZ}{dt} = 0.$$

The system (2.1), now with constant force of infection, can be rewritten as

$$\begin{aligned} \frac{dX}{da} &= N\mu - (\lambda + \mu)X, \\ \frac{dY}{da} &= (\lambda + \mu)X - (v + \alpha + \mu)Y, \\ \frac{dZ}{da} &= vY - \mu Z. \end{aligned} \tag{2.2}$$

If the population is closed, then there are no births or deaths, the parameters $\mu = \alpha = 0$. Therefore, a closed population is analogous to a closed cohort of individuals that enter at birth to the susceptible. Figure 2.1 shows the patterns of the three classes in the *SIR* model for a closed population. The upper panels show the model with $\lambda = 0.1$ and $v = 31.74$. The average duration in the infected class is 11.5 days, much shorter than the average duration in the susceptible class (10 years). This is the reason why the proportion of infected hosts, $y(a)$, is relatively small (compared to $x(a)$). The hosts are infected and recovered within 11.5 days. In this case the proportion of immune hosts is almost a mirror image of the proportion of susceptible since $z(a) \approx 1 - x(a)$, as can be seen from the upper left panel. The lower panels show the model with $\lambda = 0.5$ and $v = 31.74$.

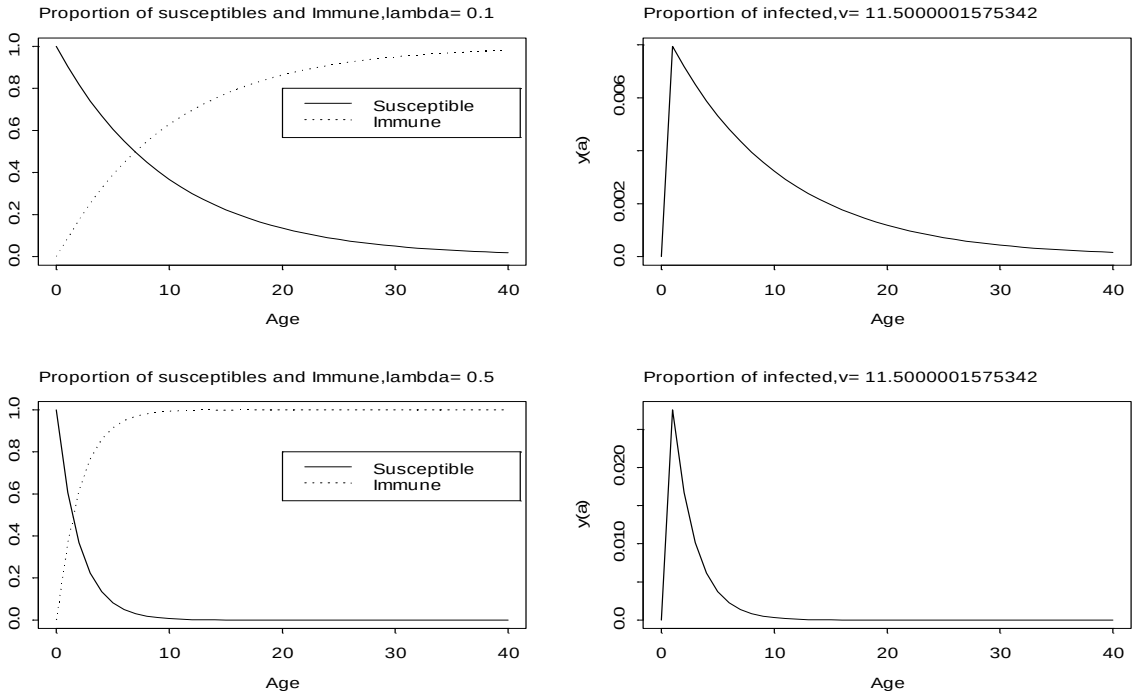


Figure 2.1: *The SIR model. Left panels: $x(a)$ and $z(a)$. Right panels: $y(a)$. Upper panels: the model with average duration of 10 years in the susceptible class and 11.5 days in the infected class. Lower panels: the model with average duration of 2 and 11.5 years for the susceptible and the infected classes respectively.*

2.1.5 Basic (Effective) Reproductive Number

The *basic reproductive rate*, R_0, is the expected number of new infections produced when one infected individual is entering into a completely susceptible population (Halloran 1998).

R_0 =Number of contacts $\times$ Transmission probability per contact $\times$ Duration of infection.

If $R_0 < 1$, it implies that the rate of generation of new cases is smaller than the rate of loss of existing cases, therefore the parasite cannot invade the population. In this case the disease cannot maintain itself and will be eliminated. The *effective reproductive number*, R, is assumed to be linearly dependent on the basic reproductive number:

$$R = R_0 x,$$

where x is the proportion of susceptible hosts,

$$x = \frac{\text{susceptibe hosts}}{\text{total number of hosts}}.$$

R is the expected number of new infections in practice, taking into account that only a proportion x of the individuals are susceptible. At equilibrium, the rate at which susceptible hosts are infected is the same as the rate at which new susceptible hosts appear, so

$R = 1$. Since $R = R_0 x = 1$, we derive

$$R_0 = \frac{1}{x^\star} = \left(\frac{\bar{N}}{\bar{X}}\right).$$

In the *SIR* model a vaccination program implies that one part of the host population, say p, vaccinates at birth and the other part, $(1-p)$, does not. This means that $(1-p)\times 100\%$ from the hosts are entered to the susceptible class while the part that was immunised, $p \times 100\%$, is entered directly into the immune class. Now, if $p \times 100\%$ of the hosts are immunized then at equilibrium $x^\star$ is at most $1 - p$. As mentioned previously, for $R = R_0 x < 1$, the infection cannot spread, so the critical proportion of hosts that has to be vaccinated in order to eliminate the disease is the one with $R_0(1 - p_c) < 1$ or $p_c = 1 - 1/R_0$.

In case that the force of infection is assumed to be age dependent (Dietz 1993) R_0 is given by

$$R_0 = \frac{\int_0^\infty \lambda^2(a) \exp\left[-\int_0^a \mu(s)ds\right] da}{\int_0^\infty \lambda^2(a) \exp\left[-\int_0^a \mu(s) + \lambda(s)ds\right] da},$$

where $\mu(s)$ is the age dependent death rate. Other approximations for R_0 discussed by Keiding (1996) and Farrington *et al.* (2001).

2.1.6 Mixing Pattern

The above assumption about the random mixing of the population usually does not hold. Most populations do not mix in a random fashion but have subgroups that mix more with their own members. For example, different age groups within a school, contacts within households and sexual contacts within the population. Halloran (1998) considered a population with two groups, A and B. The contact pattern between these two groups can be described by the following mixing matrix:

$$C = \begin{pmatrix} \beta_{aa} & \beta_{ab} \\ \\ \beta_{ba} & \beta_{bb} \end{pmatrix}. \tag{2.3}$$

The contact rate for individuals of group A with those of group B is denoted as β_{ab}. Individuals can also contact with individuals from the same group: β_{aa} and β_{bb} denote the within-group contact rate for groups A and B, respectively. Figure 2.2 illustrates the mixing pattern in this simple example. The emphasis in the literature is placed on age-related transmission models. See, for example, Anderson and May (1991), Ferguson, Anderson and Garnett (1996), Halloran *et al.* (1994),Farrington (1990), Farrington (2001), Greenhalgh and Dietz (1994) and Edmunds *et al.* (1996). In an age-related transmission rate model the assumption is that the transmission rate depends on the host age. The population is usually divided into a few age groups and β_{ij} is the transmission rate between hosts from the ith age group and hosts from the jth age group. For example, for a 5 age groups model, C_1 in (2.4) is a matrix which represents a mixing pattern for which individuals are mixing only with individuals from their own age group while C_2 represents a mixing pattern for which individuals are mixing across the age groups in background

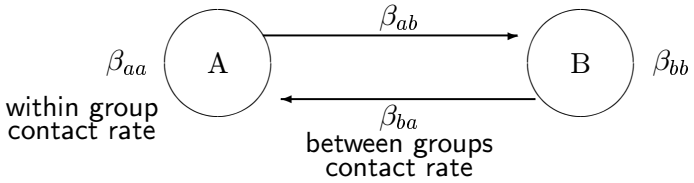


Figure 2.2: *Mixing pattern between two hypothetical sub-populations.* β_{aa} *is the contact rate within group A,* β_{bb} *is the contact rate within group B.* $\beta_{ab} = \beta_{ba}$ *is the contact rate between the two groups.*

rate β_5 and with different rates within the age group. For other structures of the contact matrix see Anderson and May (1985, 1991) and Halloran *et al.* (1994).

$$C_1 = \begin{pmatrix} \beta_1 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 \\ 0 & 0 & \beta_3 & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & 0 \\ 0 & 0 & 0 & 0 & \beta_5 \end{pmatrix}, \quad C_2 = \begin{pmatrix} \beta_1 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \\ \beta_5 & \beta_2 & \beta_5 & \beta_5 & \beta_5 \\ \beta_5 & \beta_5 & \beta_3 & \beta_5 & \beta_5 \\ \beta_5 & \beta_5 & \beta_5 & \beta_4 & \beta_5 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \end{pmatrix} \tag{2.4}$$

The major question is how to estimate the mixing matrix, also called the "Who Acquires Infection From Whom" or WAIFW (e.g., see Anderson and May 1985, 1991).

2.1.7 Estimating the WAIFW Matrix in the Case of Age-Dependent Force of Infection

Consider an age-related transmission model with two age groups as described above. For each age-group there is an age-specific force of infection, λ_i $(i = 1, 2)$. It is assumed that the force of infection at each group is calculated by integrating over the product of age-specific transmission coefficient $\beta(a, a')$ and the number of infected individuals. Using the mixing matrix (2.3) it follows that

$$\lambda(a, t) = \int_0^L \beta(a, a')Y(a't)da. \tag{2.5}$$

Here, L is life expectancy, and $\beta(a, a')$ is the transmission rate between infectious individual at age a' and susceptible at age a The mixing matrix is assumed to be constant and therefore does not change with the progress of the disease. On the other hand, $\lambda(a, t)$ varies with time. If $C(a)$ is known, then $\lambda(a, t)$ can be estimated easily from (2.5). However, in practice $C(a)$ is unknown and has to be estimated. Usually, $C(a)$ can be estimated indirectly in the following way. The age-specific force of infection, λ_i, is estimated from observational data before immunization. From (2.5) we derive

$$\begin{aligned} \lambda_1 &= \beta_{11}\bar{Y}_1 + \beta_{12}\bar{Y}_2, \\ \lambda_2 &= \beta_{21}\bar{Y}_1 + \beta_{22}\bar{Y}_2. \end{aligned}$$

Note that we assume $\beta_{12} = \beta_{21}$. There are three unknowns (the β's) in the two equations above and therefore their values cannot be estimated even if the values of λ_1, λ_2, $\bar{Y}_1$ and $\bar{Y}_2$ are known. The concept of the WAIFW matrix is developed in order to overcome this problem. For two age groups, if, for example, we define

$$C = \begin{pmatrix} \beta & \alpha \\ \\ \alpha & \beta \end{pmatrix},$$

then

$$\begin{aligned} \lambda_1 &= \beta\bar{Y}_1 + \alpha\bar{Y}_2, \\ \lambda_2 &= \alpha\bar{Y}_1 + \beta\bar{Y}_2. \end{aligned}$$

That way, α and β can be estimated from the data before immunization and can then be used to calculate the age-specific force of infection after the vaccination program has been started. For the case with more than 2 age groups, as long as the number of parameters in the WAIFW matrix is equal to the number of age groups, one can use the estimated force of infection in order to estimate the elements of the WAIFW matrix. For more details about this estimating process, see Anderson and May (1985, 1991). Edmunds *et al.* (1997) presented a pioneering study in which they try to estimate the mixing pattern from a sample of 92 individuals. Edmunds *et al.* (1997) used multi-level models to estimate the number of contacts per day adjusted to the age of the individual. However, they mentioned that their study is a pilot study which aims to determine if data of this type can be collected. In practice, Edmunds *et al.* (1997) estimated the number of contacts and not the mixing patterns in their sample.

2.1.8 Example: Hepatitis B

Infection with Hepatitis B (HBV) occurs in many parts of the world. It is estimated that between 1 billion (Edmunds *et al.* 1996) to 2 billion (Van Damme *et al.* (1997,1998) individuals who are alive have been infected with HBV. There are around 300–350 million chronic carriers of the virus all over the world. Van Damme *et al.* (1997,1998) described the three major outcomes of HBV:

- After infection the individuals may present symptomatic acute infection and develop lifelong immunity.

- The infected individuals may become a chronic carrier.
- The infected individuals may die within a few days from infection.

According to the World Health Organization (WHO) the number of carriers increases by 10 million per year. These carriers are susceptible to life long complications. Approximately 25% of the carriers will die.

Edmunds *et al.* (1996) discussed a mathematical model for the dynamics of HBV in developing countries. The model is an extention of the *SIR* model, in which the latent class (H) and the carriers class (C) are added into the model. A set of 5 partial differential equations was used by Edmunds *et al.* (1996) in order to describe the process mentioned above:

$$\begin{aligned}
\frac{dX}{da}+\frac{dX}{dt} &= -(\lambda(a,t)+\mu)X(x,t),\\
\frac{dH}{da}+\frac{dH}{dt} &= \lambda(a,t)X(a,t)-(\sigma+\mu)H(a,t),\\
\frac{dY}{da}+\frac{dY}{dt} &= \sigma H(a,t)-(\mu+\gamma_1)Y(a,t),\\
\frac{dC}{da}+\frac{dC}{dt} &= p\gamma_1 Y(a,t)-(\mu+\gamma_2+w)C(a,t),\\
\frac{dZ}{da}+\frac{dZ}{dt} &= (1-p)\gamma_1 Y(a,t)+\gamma_2 C(a,t)-\mu Z(a,t).
\end{aligned} \tag{2.6}$$

Note that we assume that the vaccination rate is zero. The system (2.6) describes a slightly different process from the one that was discussed by Van Damme *et al.* (1998). The change in the number of susceptibles at age a and time t is equal to the number of susceptibles who are infected or die $(\lambda(a,t)+\mu)X(a,t))$ and transfer into the latent class. As mentioned before, individuals within the latent class are infected but not yet infectious for others. The change in the number of the latent individuals is simply the difference between the "new" latents (arriving from the susceptible class) and the number of individuals who become infectious or die, $(\sigma+\mu)H(a,t)$. The average duration at the latent class is $1/\sigma$. Infectious individuals can become a carrier with probability p or develop life long immunity (with probability $(1-p)$). The process is described in Figure 2.3. The probability to become HBV carrier is assumed to be age dependent. It was first investigate by Edmunds *et al.* (1993) who used maximum likelihood methods to model the probability as function of age. In Chapter 8 we present a reanalysis of the data using hierarchical Bayesian changepoint models to describe the association between the probability and the host age.

2.2 Estimation From Serological Data

So far, the focus in Section 2.1 was placed on the transmission model of the disease. The aim of this Section is to make the connection between the mathematical model and statistical models used to estimate one of the parameters of the model, *the force of infection*. A major effort has been done in the last 68 years to model the force of infection. Although, early models assumed a constant force of infection, the current approach is to model an

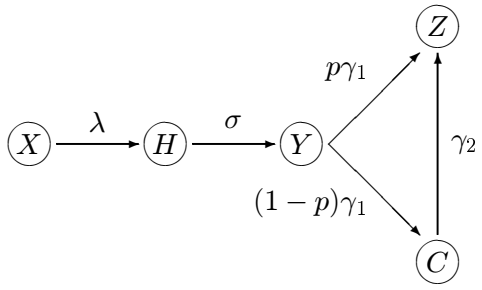


Figure 2.3: *The flow of individuals within the population for the HBV model described in Edmunds et al. (1996). At each state the individual can die at rate μ. At the carrier state the rate of death is $\mu + w$. $1/\sigma$, $1/\gamma_1$ and $1/\gamma_2$ are the average durations in the latent, infectious and carriers states, respectively.*

age-dependent force of infection. In this chapter we will review the literature related to the modeling of the force of infection and we shift the focus of attention from mathematical to statistical modeling. In the epidemiological literature, the proportion of susceptibles at age a and time t is denoted by $x(a,t)$. In order to avoid confusion with statistical notation, which usually use x to denote a random variable, we change notation and use q to denote the susceptible fraction. We assume the disease is irreversible, meaning that the immunity is assumed to be lifelong. We further assume that the mortality caused by the infection is negligible and can be ignored. Let $q(a,t)$ be the fraction of susceptible individuals at age a and time t. Under the assumptions stated above, the partial differential equation which describes the change in the susceptible fraction at age a and time t is given by :

$$\frac{\partial}{\partial a}q(a,t) + \frac{\partial}{\partial t}q(a,t) = -\ell(a,t)q(a,t). \tag{2.7}$$

Here $\ell(a,t)$ is the rate at which susceptible individuals become infected and is called the hazard or the force of infection. Note that (2.7) assumes that the natural death rate is zero up to the life expectancy and thereafter infinity. In a steady state, the time homogeneous form of the model, $\frac{\partial}{\partial t}q(a,t) = 0$ and (2.7) reduces to

$$\frac{d}{da}q(a) = -\ell(a)q(a). \tag{2.8}$$

Differential equation (2.8) describes the change in the susceptible fraction with the host age. Note that the prevalence $\pi(a) = 1 - q(a)$. Throughout this thesis the term "prevalence" is used to mean "prevalence of serological marker of past infection" or "prevalence of immune".

2.2.1 Muench (1934) and Griffiths (1974)

Muench (1934,1959) suggested to model the infection process with a catalytic model, in which the distribution of the time spent in the susceptible class is exponential with rate β. The force of infection in this case, β, is age independent. Precisely, Muench (1934) proposed to model the prevalence by $\pi(a) = k\{1 - \exp(-\beta a)\}$ where $1-k$ is the proportion of the population staying uninfected for lifetime. Under the catalytic model (assuming that $k = 0$) $q(a) = e^{-\int_0^a \beta ds} = e^{-\beta a}$ and $\frac{d}{da}q(a) = -\beta e^{-\beta a}$. Griffiths (1974) proposed a model for measles in which the force of infection increases linearly in the age range $\tau - 10$, $\tau \geq 0$. Specifically, Griffiths (1974) suggested

$$\ell(a) = \begin{cases} \beta_1(a + \beta_0) & a > \tau, \\ 0 & a \leq \tau. \end{cases}$$

In this model the force of infection is zero between $0 - \tau$ years which corresponds to the maternal antibody period. Note that, since Griffiths (1974) specified τ as a parameter in the model, Griffiths' model should be interpreted as a changepoint model. Interestingly, Griffiths (1974) justified his choice of a linear force of infection by using a nonparametric estimate for the force of infection, $\ell(a) = \Delta\pi/(1 - \pi(a))$, which he plotted against age and showed the linear trend of the force of infection. Griffiths (1974) himself mentioned that his model for the prevalence corresponds to a model in which the linear predictor is a quadratic function of age.

2.2.2 Grenfell and Anderson (1985)

Grenfell and Anderson (1985) extended the model further and used polynomial functions to model the force of infection. The advantage of higher order polynomials is their flexible curve shapes. Grenfell and Anderson (1985) did not restrict the force of infection to be constant or linear but gave the data to lead the results. Their model assumes that $\pi(a) = 1 - e^{-\Sigma\beta_i a^i}$ which implies that the force of infection is $\ell(a) = \sum \beta_i i a^{i-1}$. Note that within the framework of generalized linear models (McCullagh and Nelder 1989) for binary response, the model of Grenfell and Anderson (1985) can be fitted using a log link function. In this case, the force of infection is simply the first derivative of the linear predictor. For the general case the solution for (2.8) under the catalytic model is $q(a) = e^{-\gamma(a)}$, where $\gamma(a) = \int_0^a \ell(s)ds$ is the cumulative hazard.

2.2.3 Nonlinear and Generalized Linear Models

One problem that arises when a higher order polynomial model is fitted is that the estimate for the force of infection can get negative. In fact, a force of infection estimate turns negative whenever the estimated probability to be infected before age a is a nonmonotone function.
A possible solution to this problem is to define a nonnegative force of infection, $\ell(a) \geq 0$ for all a, and to estimate $\pi(a)$ under these constraints. Farrington (1990), Farrington *et*

al. (2001) and Edmunds *et al.* (2000) applied this method for measles, mumps and rubella, using a nonlinear model for $\pi(a)$. However, Farrington's method requires prior knowledge about the dependence of the force of infection on age. Other parametric models, fitted within the framework of generalized linear models (GLM) with binomial error, were discussed by Becker (1989), Diamond and McDonald (1992) and Keiding *et al.* (1996) who used models with complementary log-log link function in order to parameterize the prevalence and the force of infection as a Weibull model. Becker (1989) suggested to model a piecewise constant force of infection by fitting a model with log link. For the case that other covariates, in addition to age, are included in the model, Jewell and Van Der Laan (1985) proposed, in the context of current status data, a proportional hazard model with constant force of infection, which can be fitted as a GLM with complementary log-log link. Grummer-Strawn (1993) discussed two parametric models, the first being a Weibull proportional hazard model with complementary log-log link and the second being a log-logistic model with logit link function. For the latter, the proportionality in the model is interpreted as proportional odds.

2.2.4 Keiding (1991)

The first attempt of Griffiths (1974) to estimate the force of infection non-parametrically followed by Farrington (1991), who used a smoothed version of the Griffiths estimator. However, both can lead to a negative estimate for the force of infection. Keiding (1991) proposed a two step approach in which in the first step the prevalence is estimated by isotonic regression (Barlow *et al.* (1972), Robertson *et al.* (1988)) and in the second step a kernel smoother is used in order to estimate the force of infection.
In the first step, Keiding (1991) proposed to estimate the prevalence using an isotonic regression estimate of the observed prevalence. This can be done by applying the pool violator algorithm (PAV) to the data. In this case, the nonparametric maximum likelihood (NPMLE) for the prevalence is a step function. In the second step the force of infection, assumed to be a smooth function of age, is estimated by

$$\hat{\ell}(a) = \frac{1}{h}\int_{a-h}^{a+h} K\left(\frac{x_i - a}{h}\right)\frac{d\hat{\pi}(x)}{1-\hat{\pi}(x^-)},$$

where K is a kernel function, h is the bandwidth and $\hat{\pi}(u)$ is the isotonic regression of the observed prevalence. As discussed in Greenhalgh and Dietz (1994), in case that $\hat{\pi}$ has discontinuities at the points $x_1, x_2, \ldots, x_n$ then

$$\hat{\ell}(a) = \frac{1}{h}\sum_{x_i \in [a-h, a+h]} K\left(\frac{x - a}{h}\right)\frac{\hat{\pi}(x_i) - \hat{\pi}(x_{i-1})}{1-\hat{\pi}(x_{i-1})}.$$

Keiding's method requires the choice of a bandwidth. This issue was not addressed by Keiding who chose the smoothing parameter by visual inspection. Keiding *et al.* (1996) proposed to replace the kernel estimate in the second step with a smoothing spline. However, this method requires to choose a smoothing parameter as well.
In context of current status data, Shiboski (1998) proposed a semiparametric model, based on generalized additive models (Hastie and Tibshirani 1990), in which the dependency of the force of infection and age is modeled nonparametrically and the covariate effect is

the parametric component of the model. Depending on the link function, the model proposed by Shiboski (1998) assumes proportionality; proportional hazard (complementary log-log link) or proportional odds (logit and probit links). Other semiparametric models, assuming a logit link, were proposed by Rossini and Tsiatis (1996).

2.3 Serological Datasets

All the methods mentioned above can be used to estimate the force of infection from serological datasets. In this section we briefly review 5 serological datasets that were used to estimate the force of infection. Figure 2.4 shows these five datasets that will be discussed below.

2.3.1 Rubella and Mumps

The Rubella and the Mumps datasets were used by Farrington (1990) and Farrington *et al.* (2001) to illustrate the use of nonlinear models for estimating the force of infection. The seropervalence samples were collected in UK between November 1986 and December 1987 and reported by Morgen-Capner *et al.* (1988). Seroprevalence samples were collected by 5 public health laboratories in different parts of UK. An average of 250 samples were tested for each year of age from 1-14, two years of age from 15-34, five year of age from 35-44 and 10 years from 45+. For the current analysis data consist of 4230 and 8179 individuals for rubella and mumps, respectively, with age range between 1 to 44 years old.

2.3.2 Hepatitis A

While rubella and mumps are common airborne childhood infectious diseases, the hepatitis A virus (HAV) is mainly (> 95 %) transmitted by the feco-oral route (e.g. through food and water polluted by faeces containing the virus). Transmission is facilitated by poor hygienic living and housing conditions, and is particularly common in developing countries (Hadler, 1991, Beutels *et al.*, 1997). In these countries HAV is mainly a childhood infection, whereas in industrial countries HAV infection occurs during adulthood as well as childhood. In the poorest developing countries, the pattern of high endemicity is characterized by rapid infection at a very young age; over 90% of the children become infected by the age of 5. In 1993 and early 1994, a study of the prevalence of HAV antibodies was conducted in the Flemish Community of Belgium. The purpose of this study was to obtain data on the prevalence of hepatitis A in Flanders and to analyze the epidemiological pattern of HAV. During the study period serum samples were collected from hospitals (non-infectious disease wards) in the Flemish Community. The dataset contains the serological results of 3161 Belgian individuals together with their age in years, ranging from 0.5 to 85 years. The study group was similar in composition to the Flemish population in terms of age.
Keiding (1991) introduced the hepatitis A dataset from Bulgaria to illustrate the use of isotonic regression as an nonparametric approach to estimate the prevalence and the force of infection. The data contains information about 850 individuals with age range from 1

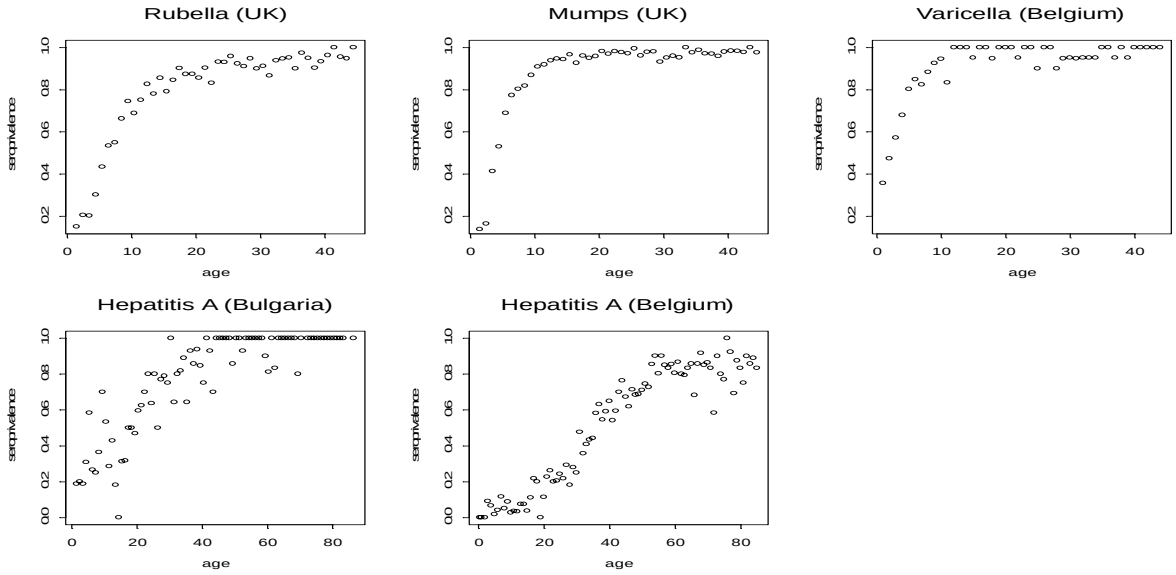


Figure 2.4: *Five cross-sectional seroprevalence datasets.*

to 86 years. This datasets was used by Farrington *et al.* (2001) to illustrate a model with constant force of infection.

2.3.3 Varicella

The Varicella dataset consists of 1673 individuals with age range from 1 to 44 years old that was sampled in Antwerp (Belgium) between October 1999 to April 2000 and reported by Thiry *et al.* (2002). The sera were residuals specimens submitted to laboratories for diagnostic purposes. Sera for age group of 1-12 years were collected from outpatients hospital in Antwerp. Sera for age group 12-16 collected from volunteers in vaccine trails. Sera for age groups older than 16 years were provided by a medical laboratory in Antwerp. The population was stratified by age in order to sample approximately 100 observations per age-group.

Chapter 3

Modeling Forces of Infection Using Fractional Polynomials

3.1 Introduction

The motivation to model the force of infection with higher order polynomials is the flexible curve shapes that these models provide. However, this flexibility in modeling comes at the price of non-monotonicity. Indeed, high order polynomials can predict a model with negative force of infection. Other parametric models, such as the models with constant or linear force of infection of Muench (1959) and Griffiths (1974) or the Weibull models, proposed by Becker (1989), Diamond and McDonald (1992) and Keiding (1996), put the restriction on the shape of the force of infection to be constant or monotone. The force of infection, for these models, cannot be negative. However, this came with the price of flexibility of the model for the force of infection. In this chapter we discuss constrained fractional polynomials (Royston and Altman, 1994) as a parametric models for both the prevalence and the force of infection. This class of models provides highly flexible curve shape for the force of infection. Furthermore, we will show that the parametric models discussed above (all, except the nonlinear model proposed by Farrington 1990,2001) can be fitted within the framework of fractional polynomials as well. In Section 3.2 we describe a general age-dependent model for the force of infection, based on prevalence data. Section 3.3 discusses fractional polynomials as a flexible parametric approach to model the force of infection. The method is applied within the framework of generalized linear models for a binary response. The issue of monotonicity is addressed in Section 3.3. In Section 3.4 we apply the method to the datasets mentioned above. The models in Section 3.4 assume a logistic form for the prevalence $q(a)$ and were fitted with the logit link function. In Section 3.5 we modify this assumption and model the force of infection with fractional polynomials for which $q(a) = \exp(-\gamma(a))$.

3.2 Age-Dependent Force of Infection

Consider an age-specific cross-sectional prevalence sample of size N and let a_i be the age of the ith subject. Instead of observing the age at infection we observe a binary variable Y_i such that

$$Y_i = \begin{cases} 1 & \text{if subject } i \text{ had experienced infection before age } a_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

With $\pi(a_i)$ the probability to be infected before age a_i, $\pi(a_i) = 1-q(a_i)$, the log likelihood is given by

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} Y_i \log\{\pi(a_i)\} + (1 - Y_i) \log\{1 - \pi(a_i)\}. \tag{3.2}$$

Here, $\pi(a) = g^{-1}(\eta(a))$, where $\eta(a)$ is the linear predictor and g is the link function. For binary responses, g is often taken to be a logit link function, $\log(\pi/(1-\pi))$, but other link functions such as the complementary log-log link, $\log(-\log(1-\pi))$, and log link, $-\log(1-\pi)$, can be used as well. The models proposed by Muench (1934,1959), Griffiths (1974) and Grenfell and Anderson (1985) assume g to be the log link function (for $1-\pi$) and $\eta(a) = \sum_{i=0}^{k} \beta_i a^i$, where k is equal to 1 (Muench), 2 (Griffiths) and K (Grenfell and Anderson). Using a model with log link function leads to a simple interpretation of the first derivative of the linear predictor. Indeed, $\eta(a)$ is the cumulative hazard and therefore the force of infection is simply the first derivative of the linear predictor. Under the catalytic model $\pi(a) = 1 - e^{-\eta(a)}$. Using the definition for the hazard rate, we get

$$\ell(a) = \frac{\pi'(a)}{1 - \pi(a)} = \frac{\eta'(a)e^{-\eta(a)}}{e^{-\eta(a)}} = \eta'(a). \tag{3.3}$$

In the general case, when the link function is not restricted to be the log link, the force of infection can still be derived according to (3.3). It is easy to see that for the binomial distribution, the force of infection can be expressed as a product of two functions:

$$\ell(a) = \eta'(a)\delta(\eta(a)). \tag{3.4}$$

Here, $\delta(\cdot)$ is a known function for which the form is determined by the link function g. Table 3.1 shows three possible link functions with their corresponding structure for the force of infection.

3.3 Fractional Polynomial Models for Binomial Data

3.3.1 Motivating Example

Viral hepatitis is a serious problem throughout the world. In Belgium, the most common form of viral hepatitis infection is caused by the hepatitis A virus. We consider a cross-sectional prevalence sample ($N = 3161$), taken in 1993 and at the beginning of 1994

Table 3.1: *General expressions for the force of infection according to different link functions.* Φ *denotes the cumulative distribution function and* ϕ *the density function of the standard normal distribution.*

Link function	$\pi(a)$	$\delta(\eta(a))$	local estimate for $\ell(a)$
log	$1 - e^{-\eta(a)}$	1	$\hat{\eta}'(a)$
Complementary log-log	$1 - e^{-e^{\eta(a)}}$	$e^{\eta(a)}$	$\hat{\eta}'(a)e^{\hat{\eta}(a)}$
logit	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\hat{\eta}'(a)\frac{e^{\hat{\eta}(a)}}{1+e^{\hat{\eta}(a)}}$
probit	$\Phi(\eta(a))$	$\frac{\phi(\eta(a))}{1-\Phi(\eta(a))}$	$\hat{\eta}'(a)\frac{\phi(\hat{\eta}(a))}{1-\Phi(\hat{\eta}(a))}$

from 11 hospitals in Belgium. We consider two generalized linear models with logit and complementary log-log link functions. For the logit model the linear predictor is $\eta(a) = \beta_0 + \beta_1 a + \beta_2 a^3$. This model has a deviance of 82.74 on 83 degrees of freedom. For the complementary log-log model $\eta(a) = log(\beta_0) + \beta_1 a^2 + \beta_2 a^3$. The deviance of this model is 81.41 on 83 degrees of freedom. The force of infection of these models can be derived from Table 3.1. Although both models fit the data well, Figure 3.1 shows that both models predict negative forces of infection at the higher age groups. The motivation to model the force of infection with fractional polynomials is to allow for flexible changes in the force of infection over the age of the host. Indeed, high order conventional polynomials offer a wide range of curve shapes but often fit the data badly at the extremes of the observed age. Moreover, conventional polynomials do not have asymptotes and fit the data poorly whenever asymptotic behavior of the infection process is expected. Royston and Altman (1994) introduced the family of fractional polynomials as a generalization of the conventional polynomial class of functions. In the context of binary responses, a fractional polynomial of degree m for the linear predictor is defined as

$$\eta_m(a,\ \boldsymbol{\beta},\ p_1, p_2 \dots p_m) = \sum_{i=0}^{m} \beta_i H_i(a), \tag{3.5}$$

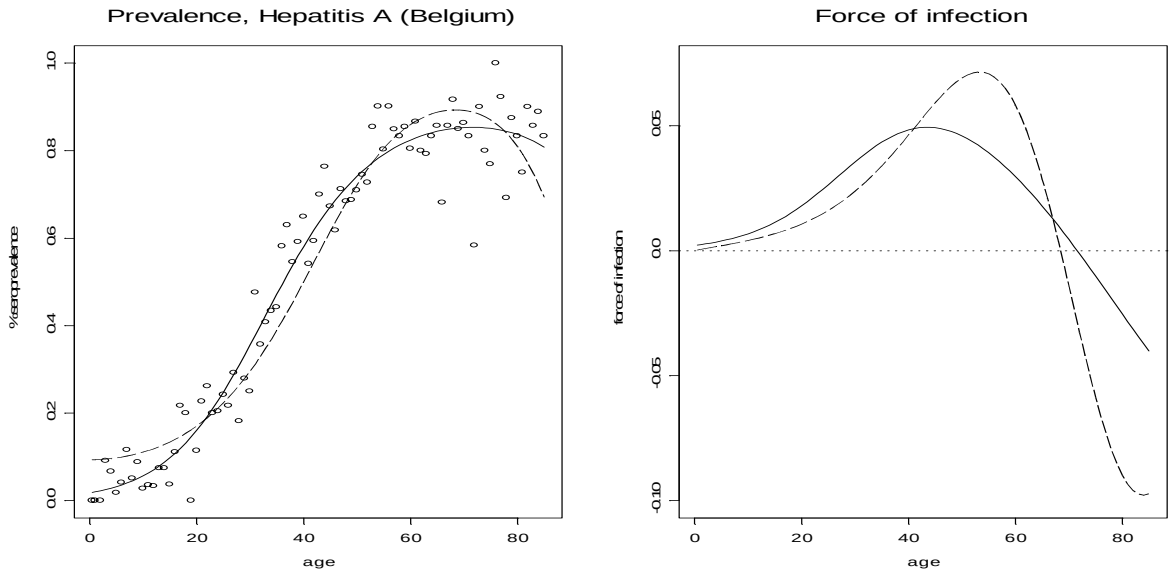


Figure 3.1: *Hepatitis A in Belgium. Left panel: data and estimated models for the prevalence. Right panel: estimated forces of infection. Solid line: model with logit link function. Dashed line: model with complementary log-log link function.*

where m is an integer, $p_1 \leq p_2 \leq \ldots \leq p_m$ is a sequence of powers and $H_i(a)$ is a transformation given by

$$H_i(a) = \begin{cases} a^{p_i} & \text{if } p_i \neq p_{i-1} \\ H_{i-1}(a) \times \log(a) & \text{if } p_i = p_{i-1} \end{cases} \tag{3.6}$$

with $p_0 = 0$ and $H_0 \equiv 1$. Royston and Altman (1994) argued that, in practice, fractional polynomials of order higher than 2 are rarely needed and suggested to choose the value of the powers from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, \max(3, m)\}$. We note that for models with log link function $\eta_1(a, \boldsymbol{\beta}, p = 1)$ is Muench's model, $\eta_2(a, \boldsymbol{\beta}, p_1 = 1, p_2 = 2)$ corresponds to the model proposed by Griffiths (1974) and the models considered by Grenfell and Anderson (1985) have the general form of $\eta_m(a, \boldsymbol{\beta}, p_1, p_2 \ldots, p_m)$ with $p_i = i$ for $i = 1, 2, \ldots, m$.

Table 3.2 shows a selection of parametric models discussed in the literature and their representation as fractional polynomials. For example, the model proposed by Keiding (1996) is a first order fractional ploynomial with $\mathbf{p} = 0$. The model with linear force of infection can be parameterized as a first order fractional polynomial with complementary log-log link for which $\mathbf{p} = 0$ with the constraint that $\beta_1 = 2$. In this case $\ell(a) = 2\beta_0 a$ which implies that the force of infection is zero at birth and increases linearly thereafter. The models presented by Grummer-Strawn (1993) and Jewell and Van Der Laan (1995) included other covariates in addition to age. For these models $\eta(m, p, \beta)$ is the fractional polynomial used to model the dependency of prevalence on age. For the models discussed in Grummer-Strawn, we do not include the adjusted parameter in our analysis since it is assumed that susceptibility is 100% at birth. The models discussed by Grummer-Strawn (1993) and Jewell and Van Der Laan (1995) were used in the context of current status

Table 3.2: *Parametric models for the prevalence presented in the literature and their corresponding force of infection.*

Publication	Fractional polynomial	Link function	Force of infection
Munch (1959),Farrington (2001), Jewell and Van der laan (1995)	$\eta(m=1,p=0,\beta_1=1)$	cloglog	constant
Griffiths(1974)	$\eta(m=1,p=0,\beta_1=2)$	cloglog	linear
Grenfell and Anderson (1985)	$\eta(m=k,p_i=i)$	log	polynomial
Keiding (1996),Becker (1989), Diamond and McDonald (1992), Grummer-Strawn (1993)	$\eta(m=1,p=0,\beta_1\neq 0)$	cloglog	monotone
Grummer-Strawn (1993)	$\eta(m=1,p=0,\beta_1\neq 0)$	logit	flexible

data. When these model are implemented for infection disease data they result constant monotone or flexible force of infection

3.3.2 Model Selection

Within the fractional polynomials framework the deviance of the model with $\eta_1(a,\boldsymbol{\beta},1)$ is taken to be the baseline deviance and improvement by other models is measured by

$$G(m,\mathbf{p}) = D(1,1) - D(m,\mathbf{p}), \tag{3.7}$$

where $D(m,\mathbf{p})$ is the deviance of the model with fractional polynomial of order m and sequence of powers $\mathbf{p} = (p_1, p_2, \ldots p_m)$. Note that a large value of G indicates a better fit. Fitting models within the framework of fractional polynomials requires to start the modeling procedure from first order fractional polynomials. To decide whether a model of first degree is adequate or a second degree model is needed, Royston and Altman (1994) recommend to use the criterion $D(1,\tilde{\mathbf{p}}) - D(2,\tilde{\mathbf{p}}) > \chi^2_{2,0.9}$ where $\tilde{\mathbf{p}}$ is the power sequence for the model that has the best goodness-to-fit (hence, the model with the highest likelihood or, equivalently, the smallest deviance).

3.3.3 Constrained Fractional Polynomials

Although fractional polynomials provide a wide range of curve shapes, there is no guarantee that $\pi(a)$ will be a monotone function of age and therefore fractional polynomials can still result in a negative estimate for the force of infection. It is clear from Table 3.1 that the estimate for the force of infection is negative whenever $\eta'_m(a,\hat{\boldsymbol{\beta}},\mathbf{p}) < 0$ (since

$\delta(\eta_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}))$ is strictly positive). Therefore, one should fit model (3.5) subject to the constraint that $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$, for all ages a in the predefined range. In the framework of fractional polynomials this cannot be done analytically. But in practice, one can fit a large number of fractional polynomials, over a grid of powers, and check for each fitted model whether $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$, for all ages a. In case that a given sequence of powers leads to a negative derivative of the linear predictor, the model is not considered as an appropriate model. This means that we choose the model with the best goodness-to-fit among all fractional polynomials for which $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$.

3.4 Application to the Data

In this section, we apply our method to the datasets mentioned in Chapter 2. For each dataset, first and second order fractional polynomials were fitted and the criterion proposed by Royston and Altman (1994) was used to decide whether the second order model is needed or not. Table 3.3 presents the deviance and gain values for the best first order fractional polynomials. Clearly, for all datasets except the Bulgarian dataset, first order fractional polynomials are not adequate and second order fractional polynomials are required. For the first order models, the gain values in Table 3.3 also indicate that, for all datasets except the Bulgarian dataset, the first order fractional polynomials with $\mathbf{p} = 1$ are not adequate and other powers are needed.

Table 3.3: *Deviance and Gain values for first and second order fractional polynomials with logit link function.*

	First order (**m=1**)				Second order (**m=2**)		
Dataset	df	Deviance	p	$G(1,p)$	df	Deviance	p_1, p_2
Hepatitis A (Be)	83	115.34	0.32	34.21	81	97.61	1.0,1.3
Hepatitis A (Bul)	80	79.51	1	0	78	77.77	1.9,1.9
Varicella	41	50.94	0.07	69.59	39	43.90	-0.7,-0.6
Rubella	41	56.28	0.03	165.13	39	42.34	-0.9,-0.4
Mumps	41	82.31	-0.2	516.88	39	47.94	-1.2,-0.9

3.4.1 Hepatitis A

The upper two panels in Figure 3.2 show the estimated models for the prevalence and the force of infection for hepatitis A in Belgium. The model with the best goodness-of-fit has a gain value of 51.94 and $\mathbf{p} = (1, 1.3)$. For this model the deviance is 97.61 on 81 degrees of freedom. The estimated force of infection reaches a peak at age 40 ($\ell(40) = 0.04159$)

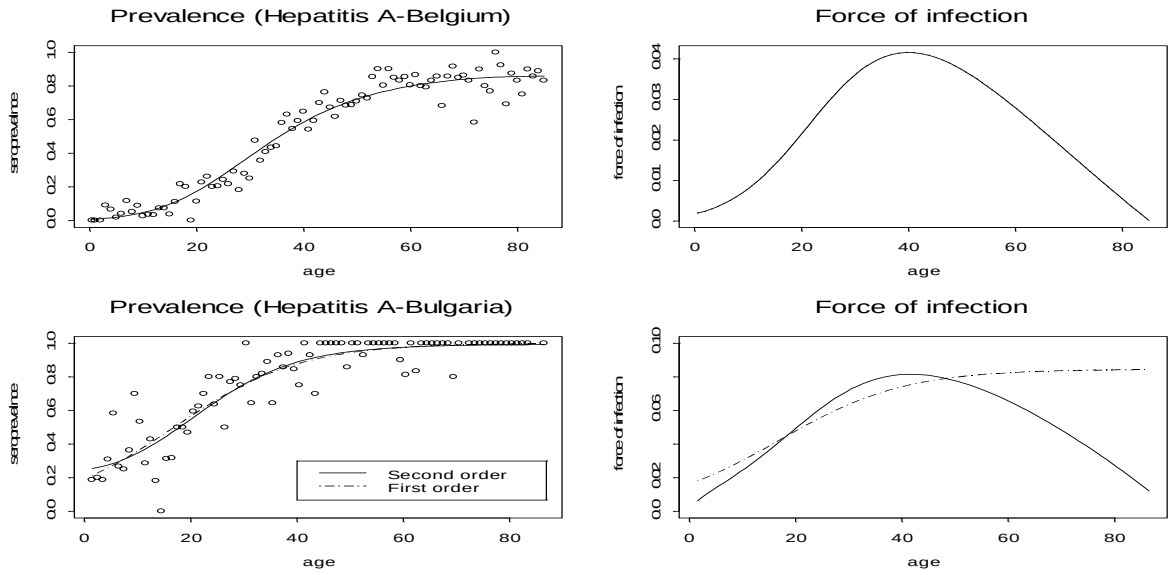


Figure 3.2: *Hepatitis A in Belgium (upper panels) and in Bulgaria (lower panels). Solid line: first order model, dashed-dot line:second order model. Models were fitted with logit link function.*

and drops down thereafter. Figure 3.3 shows the unrestricted profile likelihood surface, $L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, p_1, p_2)$, for this dataset. The point a represents the likelihood's value of the best unrestricted fractional polynomial for which $\mathbf{p} = (1.9, 1.9)$ and the deviance is 79.60 on 81 degrees of freedom. However, this model cannot be retained since it predicts a negative force of infection at older age groups. Point b represents the likelihood's value of the conventional polynomial ($\mathbf{p} = (1, 3)$, deviance equal to 82.74 on 83 degrees of freedom) which was discussed in Section 3.3.1 and will not be considered either. The point c represents the likelihood's value of the best constrained fractional polynomial. Hence, the fractional polynomial presented in Figure 3.2 can be seen as the model that has the best goodness-of-fit among all fractional polynomials satisfying $\eta'(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$.
For the Bulgarian dataset, the second order fractional polynomial with $\mathbf{p} = (1.9, 1.9)$ has a deviance of 77.77 on 78 degrees of freedom. This model suggests that the force of infection is maximal at age 41.5 ($\ell(41.5) = 0.0815$). However, the first order fractional polynomial is to be preferred since $D(1, \mathbf{p}) - D(2, \mathbf{p}) = 1.74$. Interestingly, the first order fractional polynomial with $\mathbf{p} = 1$ and logit link is just a simple linear logistic regression model. For this model $\ell(a) = \beta_1 \pi(a)$ such that it predicts an upward trend for the force of infection.

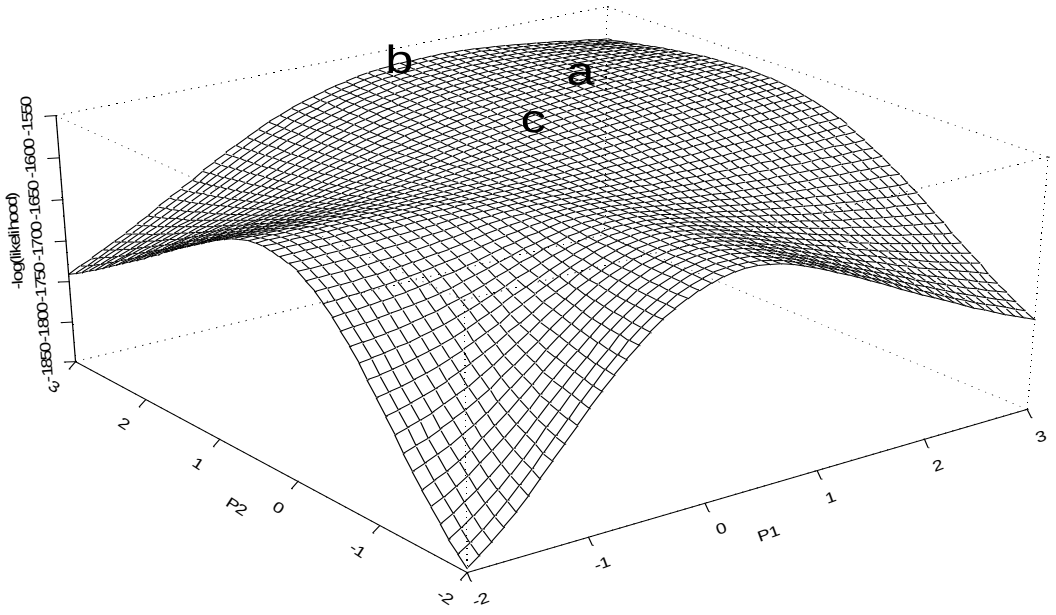


Figure 3.3: *Non-restricted likelihood surface for hepatitis A in Belgium. The points on the surface: (a) the best second order fractional polynomial* $\mathbf{p} = (1.9, 1.9)$*, (b) the conventional polynomial* $\mathbf{p} = (1, 3)$ *and (c) the best restricted fractional polynomial* $\mathbf{p} = (1, 1.3)$*.*

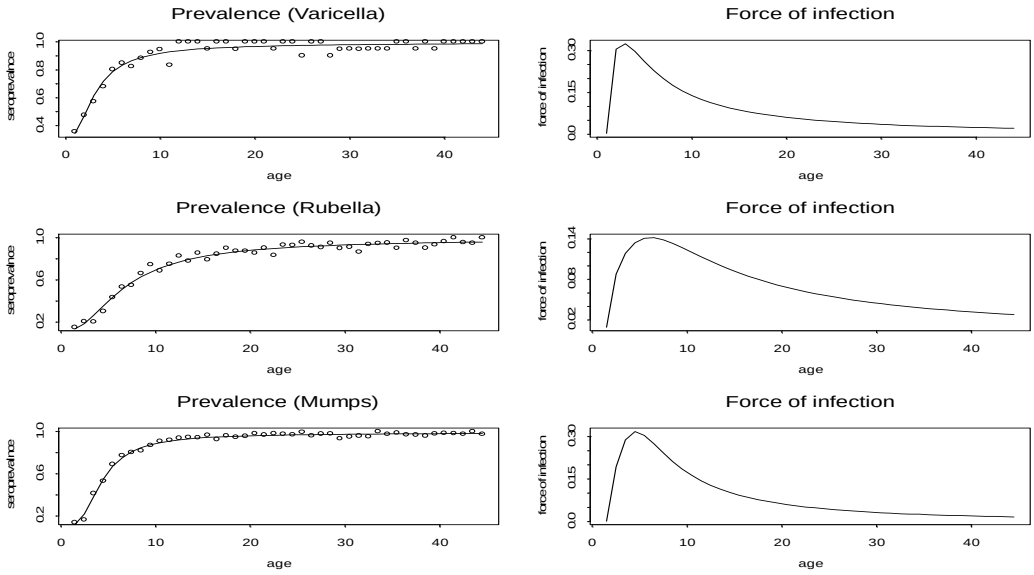


Figure 3.4: *Varicella,rubella and mumps. Second order models with logit link function. Left panels: data and estimated prevalence. Right panels: estimated force of infection.*

3.4.2 Varicella

The upper two panels in Figure 3.4 show the estimated model for both prevalence and force of infection for the Varicella dataset. The deviance of the model is 43.90 on 39 degrees of freedom and $\mathbf{p} = (-0.6, -0.7)$. For varicella, the force of infection reaches a maximum at age 3 with value $\ell(3) = 0.324$ and drops down thereafter. At age 44 the force of infection is estimated to be 0.0214.

3.4.3 Rubella and Mumps

For rubella, the fractional polynomial model with $\mathbf{p} = (-0.9, -0.4)$ has the best goodness-of-fit with a deviance of 42.34 on 39 degrees of freedom. For mumps, the model with the best goodness-to-fit uses power $\mathbf{p}$ =(-1.2,-0.9). For this model, the deviance is 47.94 on 39 degrees of freedom. Figure 3.4 (middle panels) show that for Rubella the force of infection rises to a peak at age 6.5 ($\ell(6.5) = 0.1415$). For Mumps, the force of infection reaches a maximum value at age 4.5, $\ell(4.5) = 0.317$.

3.5 Influence of the Link Function

In previous section, all models were fitted with the logit link function. In this section, we consider models of the general form $\pi(a) = 1 - \exp(-\gamma(a))$. More precisely, for the first

order fractional polynomials we specify

$$\pi(a) = \begin{cases} 1 - \exp\left(-\beta_0 e^{\beta_1 H(a)}\right) & p \neq 0, \\ 1 - \exp\left(-\beta_0 a^{\beta_1}\right) & p = 0. \end{cases} \tag{3.8}$$

For the second order fractional polynomials, we consider the following specification

$$\pi(a) = 1 - \exp\left(-\beta_0 e^{\beta_1 H_1(a) + \beta_2 H_2(a)}\right), \tag{3.9}$$

with corresponding linear predictor

$$\begin{cases} \eta_2(a, \boldsymbol{\beta}, p_1, p_2) = \log(\beta_0) + \beta_1 a^{p_1} + \beta_2 a^{p_2} & \text{if } p_1 \neq p_2, \\ \eta_2(a, \boldsymbol{\beta}, p_1, p_2) = \log(\beta_0) + \beta_1 a^{p_1} + \beta_2 a^{p_1} \log(a) & \text{if } p_1 = p_2. \end{cases} \tag{3.10}$$

We note that the models specified in (3.8) and (3.9) are GLM with a complementary log-log link function. The first order model specified in (3.8) with $p = 0$ implies a Weibull distribution for the time spent in the susceptible class. Such a Weibull model was used by Keiding (1996) to model the force of infection for rubella from an Austrian seroprevalence sample. A model with a constant force of infection is a special case of a first order fractional polynomial with complementary log-log link function with β_1 fixed at value 1; in that case $\eta(a, \boldsymbol{\beta}) = \log(\beta_0) + \log(a)$. Such a model was used recently by Farrington *et al.* (2001) to model the force of infection for hepatitis A in Bulgaria. Furthermore, a model with linear force of infection is a first order fractional polynomial with $\mathbf{p} = 0$ and $\beta = 2$.
Figure 13.7 shows the estimated forces of infection for all datasets, except the bulgarian dataset, when the optimal fractional polynomials were fitted with logit (solid lines) and complementary log-log link functions (dashed lines). We note that although the power sequence changed, the change of the link function has only little influence on the estimated forces of infection. For example, the deviance of the model for varicella is 44.04 on 39 degrees of freedom and $\mathbf{p} = (-1.3, -0.9)$ but the estimated force of infection is the same for the logit and complementary log-log models. Figure 3.6 shows the estimated prevalence and force of infection for hepatitis A in Bulgaria. For the first order models, the best fractional polynomial has a deviance of 82.75 on 80 degrees of freedom and $\mathbf{p} = 0.5$. The force of infection for this model steeply increases with age. Similar to the models with logit link, a second order fractional polynomial is not needed. Since models with different link function are not nested, we use Akaike's information criterion (AIC) for model selection (Akaike, 1974). The smallest value of AIC, 382.83, is obtained for the first order logit model (see Table 4). We note that the upward trend of the force of infection estimated by the first order logit model was already observed by Groeneboom (1991) in his discussion of Keiding's paper.

3.5.1 First Order Fractional Polynomials With $p = 0$

First order models with $p = 0$ are of interest since they lead to known parametric models in the context of survival analysis. The linear predictor for these models is

$$\eta(a) = \beta_0 + \beta_1 \log(a).$$

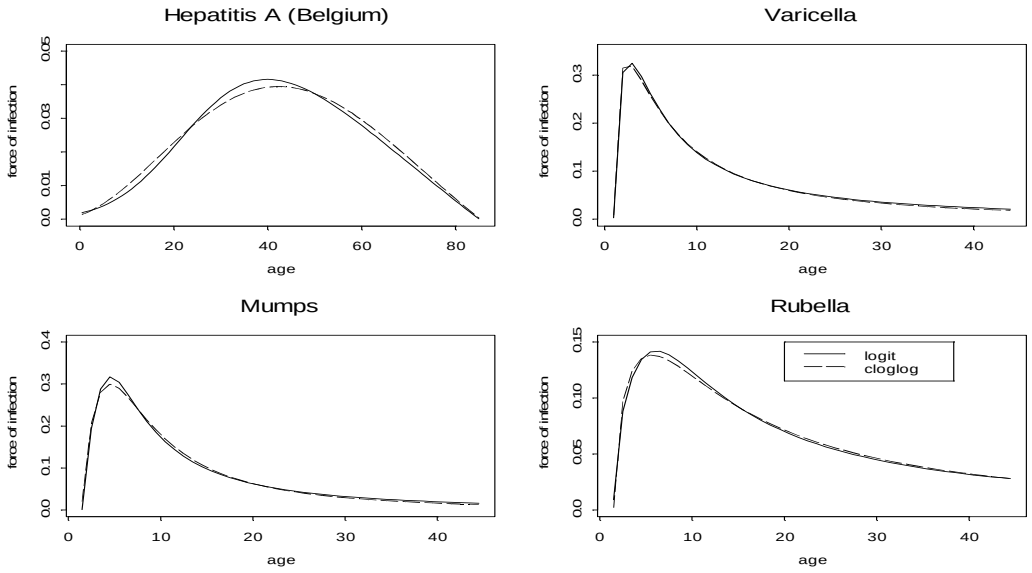


Figure 3.5: *Force of infection for second order fractional polynomial with logit (solid line) and complementary log-log link functions (dashed line). The power sequences are p=(1,1.3), p=(-0.7,-0.6), p=(-0.9,-0.4) and p=(-1.2,-0.9) for hepatitis A (BE), varicella, mumps and rubella respectively.*

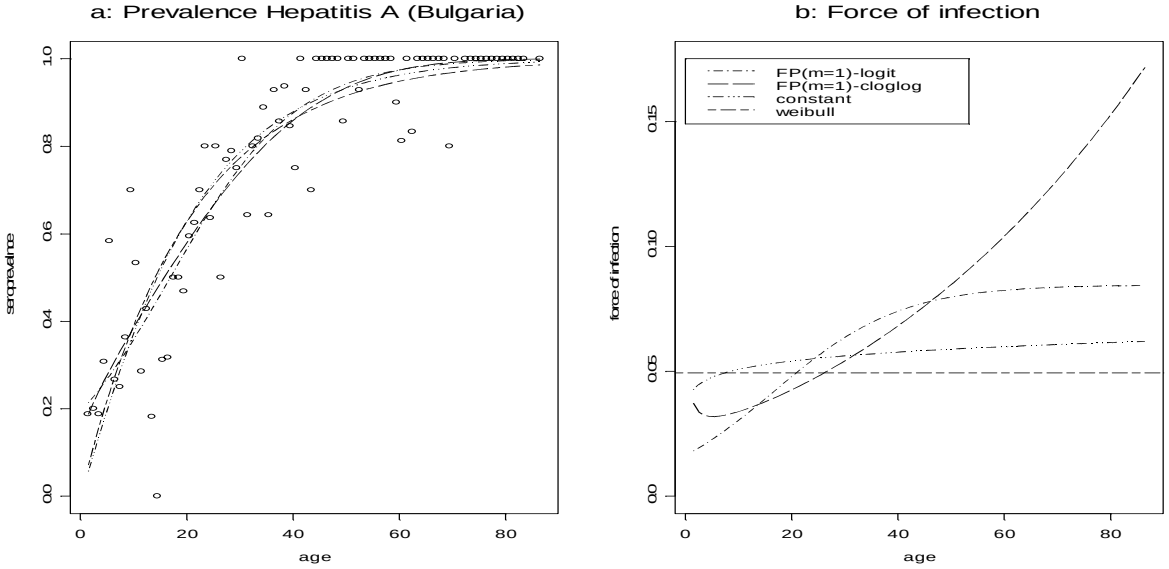


Figure 3.6: *Hepatitis A in Bulgaria, models with complementary log-log link function. First order fractional polynomials with logit and complementary log-log link function (FP(m=1)-logit and FP(m=1)-cloglog respectively). The models with constant and monotone force of infection were both fitted with complementary log-log link function and* $p = 0$.

Table 3.4: *Deviance summaries of the fitted models for the Bulgarian hepatitis A dataset. The Weibull model has 81 degrees of freedom since in this case $p = 0$, the exponential model with constant force of infection has 82 degrees of freedom since we fixed both p and β.*

Model (link)	df	Deviance	**p**	Likelihood	AIC	Force of infection
Second order(logit)	78	77.77	1.9,1.9	375.967	385.96	
First order (logit)	80	79.51	1	376.83	382.83	
Second order(cloglog)	78	79.21	1.3,1.3	376.68	386.68	
First order (cloglog)	80	82.75	0.5	378.45	384.44	
First order (cloglog)	81	94.40	0 ($\beta_1 \neq 1$)	384.27	388.27	Weibull
First order (cloglog)	82	94.67	0 ($\beta_1 = 1$)	384.41	386.41	Constant

For models with complementary log-log link function this implies that

$$\pi(a) = 1 - \exp(-\mu a^{\beta_1}),$$

where $\mu = \exp(\beta_0)$. This is a Weibull model for which the force of infection is given by $\ell(a) = \mu\beta_1 a^{\beta_1 - 1}$. Note that $\beta_1 = 1$ implies a model with constant force of infection while $\beta_1 = 2$ implies a model with linear force of infection. For a model with logit link,

$$\pi(a) = \frac{1}{1 + \beta_0 a^{\beta_1}},$$

which is a log-logistic model with force of infection given by

$$\ell(a) = \frac{\beta_0\beta_1 a^{\beta_1 - 1}}{1 + \beta_0 a^{\beta_1}}.$$

The Weibull model assumes constant or monotone force of infection. The log-logistic models allows for flexible curve shapes for the force of infection.

3.6 Discussion

We have shown that modeling the prevalence and the force of infection with fractional polynomials is a very flexible method, allowing a variety of different types of relationships between the force of infection and age. The method can compete with nonparametric smoothers while keeping the attractive features of parametric models. Furthermore, we have shown that well known parametric models for the distribution of the age at infection, such as exponential, Weibull and log-logistic distributions, can be expressed as a special case of fractional polynomials. For models with complementary log-log link function, the

curve shape of the force of infection depends on the slope of the first order fractional polynomial with $\mathbf{p} = 0$. Therefore, we need to fit the model $\eta_1(a, \boldsymbol{\beta}, \mathbf{p} = 0)$ and to check the parameter estimate for β_1. The force of infection is constant if $\beta_1 = 1$, linear if $\beta_1 = 2$ and monotone if $\beta_1 \neq 1$. Thus, by fitting a large number of fractional polynomials with logit and complementary log-log link function we account for the possibility of constant, linear, monotone or flexible curve shapes for the force of infection. However, we do not require the force of infection to have a specific curve shape in advance, the choice is data-driven.

In case that other covariates, in addition to age, are included, the following semiparametric additive model parameterizes the prevalence as

$$\text{link}(\pi(a)) = \phi(a) + Z\alpha, \tag{3.11}$$

where Z is the additional categorical covariate(s). The nonparametric component of the model, $\phi(a)$, is used to model the dependency of $\pi(a)$ on age while $Z\alpha$, the parametric component of the model, is used to model the covariate effects. In order to ensure a nonnegative estimate for the force of infection, one needs to estimate $\pi(a)$ with a non-decreasing function. This can be done by applying the pool adjacent violators algorithm (Barlow *et al.* 1972 and Robertson *et al.* 1988) to the data. This approach has been followed by Grummer-Strawn (1993) and Shiboski (1998). Within the framework of fractional polynomials, we can replace the nonparametric component of the model with a fractional polynomial

$$\text{link}(\pi(a)) = \eta_m(a, \mathbf{p}, \beta) + Z\alpha,$$

where $\eta_m(a, \mathbf{p}, \beta)$ is the fractional polynomial modeling the dependence on age. Similar to the semiparametric model in (3.11), depending on the link function, this model implies proportionality. For example, suppose that Z is a binary variable, then for models with complementary log-log link we get $\ell(a|Z = 1) = \exp(\alpha)\ell(a|Z = 0)$ and for models with logit link we obtain $\ell(a|Z = 1)/\ell(a|Z = 0) = \alpha q(a|Z = 1)/q(a|Z = 0)$.

All models discussed above are generalized linear models which imply that standard software, such as PROC GENMOD in SAS or the function `glm()` in Splus, can be used. Although our method requires to fit a large number of fractional polynomials and to choose the one with the best goodness-of-fit, the modeling procedure is not time consuming. In fact, the optimal fractional polynomial for each dataset was found in less than 3 minutes.

The models reported in this paper were fitted with a sequence of powers from -2 to 3 with an increment of 0.1. Of course, when a more sensitive grid is used, the final powers of the best model will be slightly different. For example, for hepatitis A (Belgium) the best second order fractional polynomial, fitted using a grid with increment of 0.02, has powers 1.132653 and 1.153061 with deviance 97.44. However, the force of infection is essentially the same as for the model with $\mathbf{p} = (1, 1.3)$ (maximum of absolute differences between the forces of infection is 5.68×10^{-5}). The problem of estimating a negative force of infection was addressed by fitting constrained fractional polynomials, excluding models that lead to negative force of infection as appropriate models. In our opinion, blind use of conventional linear predictors to model the force of infection can yield misleading results. Flexible models should be considered and the family of fractional polynomials offers an interesting choice. They can also be used as an exploratory tool or to perform a sensitivity analysis of a particular parametric model that, for instance, reflects prior information about the force of infection.

Chapter 4

Modeling Forces of Infection Using Monotone Local Polynomials

4.1 Introduction

In Chapter 3, fractional polynomials were fitted within the framework of parametric generalized linear models. The unknown prevalence $\pi(a)$ was modeled linearly using a known link function g, $g\{\pi(a)\} = \eta(a, \boldsymbol{\beta}, \mathbf{p})$, where $\eta(a, \boldsymbol{\beta}, \mathbf{p})$ is a linear combination of age. These models, flexible as they are, are capable to capture only structure of a single peak curve for the force of infection. If the force of infection has a secondary peak, the fractional polynomials will not be able to capture this pattern in the force of infection. In this chapter we introduce monotone local polynomials as a nonparametric alternative to model the prevalence and the force of infection. Within the local polynomials framework the linear predictor is approximated locally, in the neghborhood of some reference age value a, by a polynomial function

$$\eta(x) = \beta_0 + \beta_1(x-a) + \ldots + \beta_p(x-a)^p$$

The parameters $\beta_0, \beta_1, \ldots, \beta_p$ are estimated by maximizing a local version of the log likelihood in equation (13.3). Furthermore, it can be shown (e.g. Fan and Gijbels, 1996, page 195) that a consistent estimator for the k'th derivative of the linear predictor is given by

$$\hat{\eta}^k(a) = k!\hat{\beta}_k(a).$$

This allows us to estimate the prevalence by $\hat{\pi}(a) = g^{-1}\left\{\hat{\beta}_0(a)\right\}$. The force of infection can be estimated by substituting $\hat{\beta}_0$ and $\hat{\beta}_1$ in equation (3.4).
In Section 4.2 we describe the use of local polynomials to estimate $\pi(a)$ and $\ell(a)$, including bandwidth selection. Section 4.3 applies the method on the datasets of rubella, mumps and hepatitis A from Belgium.

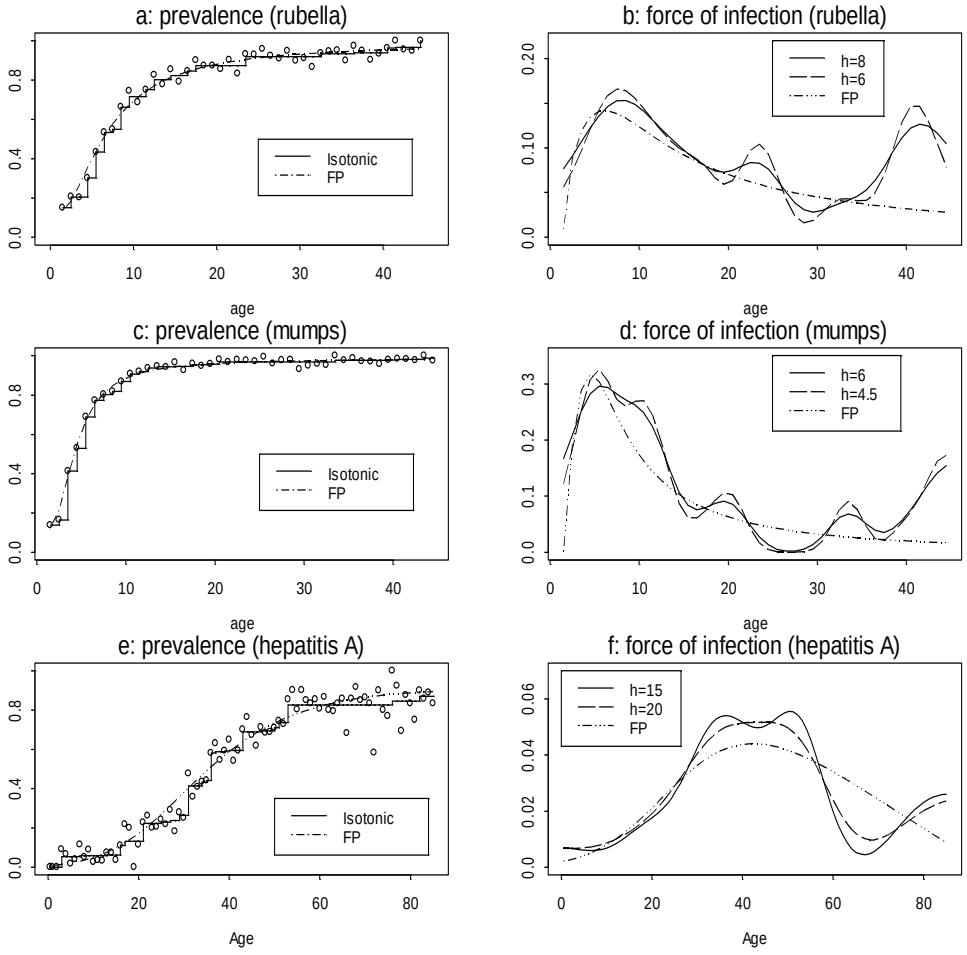


Figure 4.1: Panels a, c and e. Rubella, mumps and hepatitis A, data with estimated probability curve $\hat{\pi}(a)$: isotonic regression (solid line), optimal fractional polynomial (dot-dash line). Panels b, d and f. Estimated force-of-infection curve $\hat{\ell}(a)$: estimate based on optimal fractional polynomial (dotted dashed line). Keiding's (1991) smoothed estimate using the standard normal density function as the kernel function and two bandwidths. The bandwidths for rubella are equal to 8 (solid line) and 6 (longdash line), for mumps $h = 6$ (solid line) and $h = 4.5$ (longdash line), and for hepatitis A $h = 15$ (solid line) and $h = 20$ (longdash line).

4.2 Exploratory Data Analysis

Figure 4.1 shows the approach of Keiding (1991), a step function estimate for $\pi(a)$ and the smooth force of infection estimates using two different bandwidths. The values chosen are $h = 6$ and $h = 8$ for rubella, $h = 4.5$ and $h = 6$ for mumps and $h = 15$ and $h = 20$ for hepatitis A, values were chosen by visual inspection. Although the nonparametric estimate of the probability $\pi(a)$ is consistent under very general conditions, it has the disadvantage of being nonsmooth. The three estimates for the force of infection in the right hand panels have quite different shapes. The parametric models in Figure 4.1 (indicated by FP) lead to models with a unique maximum for the force of infection. These models were estimated using fractional polynomials and will be discussed in detail in Sections 2 and 3. For rubella and mumps, the nonparametric models predict multipeaks models and for hepatitis A, according to different choices for the bandwidth h the smooth nonparametric estimates have one maximum (at age 44.2) or two maxima (at ages 36.3 and 50.2). This illustrates two important issues: i) nonparametric versus parametric models, ii) the critical choice of the bandwidth and the need for an optimal and data-driven bandwidth choice. In this paper, we propose to estimate the force of infection by local polynomials. Compared to the method of Keiding (1991), this approach allows simultaneous estimation of prevalence and force of infection. As a consequence, the estimated probability curve is also smooth. Moreover, local polynomials are known to have several desirable properties like automatic boundary correction (Fan and Gijbels 1996). Whereas Keiding (1991) chose his bandwidth h by visual inspection, we will select an optimal data driven bandwidth, minimizing the mean squared error of the estimated force of infection. According to the principle "smooth then constrain" (Mammen *et al.* 2001), the fitted probability curve is constrained to be monotone leading to a nonnegative estimated force of infection.

4.3 Modeling Age-Dependent Force of Infection with Local Polynomials

Consider an age-specific cross-sectional prevalence sample of size N and let a_i be the age of the ith subject. Instead of observing the age of infection, we observe a binary response indicator Y_i taking the value 1 if subject i had experienced infection before age a_i and 0 otherwise. Let $\pi(a_i)$ denote the probability to be infected before age a_i, so $\pi(a_i) = 1-q(a_i)$. The log-likelihood is given by $L(\boldsymbol{\beta}) = \sum_{i=1}^{N} Q_i \left\{Y_i, g^{-1}\left(\eta(a_i)\right)\right\}$ where Q_i is the contribution of the ith subject to the Bernouilli log-likelihood with success probability $\pi(a_i) = g^{-1}\left\{\eta(a_i)\right\}$, $\eta(a)$ is the linear predictor and g the link function. The functional form describing how the force of infection changes with age is determined by the link function and the parametric structure of the linear predictor. Using a model with a log link as in Muench (1959), Griffiths (1974) and Grenfell and Anderson (1985), $\eta(a)$ is the cumulative hazard and therefore the force of infection is simply the first derivative of the linear predictor . Indeed, under the catalytic model $\pi(a) = 1-e^{-\eta(a)}$, and using the definition for the hazard rate, we get $\ell(a) = \pi'(a)/\{1-\pi(a)\} = \eta'(a)\exp\{-\eta(a)\}/\exp\{-\eta(a)\} = \eta'(a)$. Note however that the log link suffers from the structural defect that the estimated probabilities can exceed unity. In the general case, when the link function is not restricted to be the log link, the force of infection can still be expressed as a product of two functions

as shown in equation (3.4) and Table 3.1 in Chapter 3.

The choice of a link function together with a specification of the functional form of $\eta(a)$ as a function of the age a determines a fully parametric model. When turning to a local polynomial likelihood method in which no specific form for the predictor is assumed, the choice of a particular link function is less important (Fan, Heckman and Wand 1995). The local polynomial likelihood method provides consistent estimates for $\eta(a)$ and $\eta'(a)$, without any parametric restriction on the functional form. They only have to satisfy some smoothness condition (Fan and Gijbels 1996, Chapter 3). Therefore, for a given link function, the local force of infection can be estimated according to (3.4).

Using a kernel K, assigning higher weights to data points in the neighborhood of some fixed age a, and a bandwidth parameter h, the local likelihood estimation is based on maximization of

$$\sum_{i=1}^{n} Q_i \left\{ Y_i, g^{-1}\left(\eta(a_i)\right) \right\} K((a_i - a)/h).$$

The linear predictor is locally approximated by a polynomial of order p, e.g. for $p = 1$ by a linear function $\eta(a_i) \approx \eta(a) + \eta'(a)(a_i - a) = \beta_0(a) + \beta_1(a)(a_i - a)$. The local estimate for $\eta(a)$ is the local intercept, $\hat{\beta}_0(a)$, and the local slope, $\hat{\beta}_1(a)$, is the estimate for the first derivative $\eta'(a)$. Higher order polynomials can be considered as well. The estimation of $\beta_0(a)$ and $\beta_1(a)$ has to be repeated for each value of a. The choice of kernel is less important; typical choices are the symmetrical beta family, $K(u) = (1-u^2)^\gamma / \text{Beta}(0.5, \gamma+1)$ for $|u| \leq 1$, $\gamma = 0, 1, 2 \ldots$, and the Gaussian kernel given by $K(u) = \exp(-u^2/2)/\sqrt{2\pi}$. Throughout the next sections, we will always use the latter Gaussian kernel function. The choice of the smoothing parameter h however is crucial and will be discussed in more detail in what follows. As explained later in this section, it will turn out that in our setting a local quadratic model is of special importance. For a given link function, the local force of infection can be estimated by

$$\hat{\ell}(a) = \hat{\eta}'(a)\delta\left\{\hat{\eta}(a)\right\} = \hat{\beta}_1(a)\delta\left\{\hat{\beta}_0(a)\right\}. \tag{4.1}$$

Table 4.1 presents the local estimates for the force of infection corresponding to different link functions. The local polynomial estimation procedure requires a data driven value for the bandwidth h. Several methods to choose the value of h are discussed in Chapter 4 of Fan and Gijbels (1996). The minimization of the asymptotic mean squared error ($AMSE$) of $\hat{\ell}(a)$ as a function of h leads to an optimal local bandwidth $h(a)$. The asymptotic normality of $\hat{\ell}(a)$ follows from the asymptotic joint normality of $(\hat{\beta}_0(a), \hat{\beta}_1(a))$ and the delta method can be used to derive expressions for the asymptotic bias and variance of $\hat{\ell}(a)$. Indeed, from the main theorem in Fan, Heckman and Wand (1995), it follows that for $p = 2$ (local quadratic) and $h = cn^{-1/7}$ (c some constant)

$$\begin{bmatrix} n^{\frac{3}{7}}(\hat{\beta}_0(a) - \eta(a)) \\ n^{\frac{2}{7}}(\hat{\beta}_1(a) - \eta'(a)) \end{bmatrix} \xrightarrow{D} N(\boldsymbol{b}(c), V(c)), \tag{4.2}$$

where $\boldsymbol{b}(c) = (b_1(c), b_2(c))$ is the asymptotic bias and $V(c)$ the asymptotic covariance matrix. Consider the function φ in (3.4). Using the delta method (some details in the Appendix), we get the following asymptotic normality result for the estimated force of

Table 4.1: *General expressions for the force of infection according to different link functions.* Φ *denotes the cumulative distribution function and* ϕ *the density function of the standard normal distribution.*

Link function	local estimate for $\ell(a)$
log	$\hat{\beta}_1(a)$
Complementary log-log	$\hat{\beta}_1(a)e^{\hat{\beta}_0(a)}$
logit	$\hat{\beta}_1(a)\frac{e^{\hat{\beta}_0(a)}}{1+e^{\hat{\beta}_0(a)}}$
probit	$\hat{\beta}_1(a)\frac{\phi(\hat{\beta}_0(a))}{1-\Phi(\hat{\beta}_0(a))}$

infection:

$$n^{\frac{2}{7}}\left(\varphi(\hat{\beta}_0(a),\hat{\beta}_1(a)) - \varphi(\eta(a),\eta'(a))\right) \xrightarrow{D} N(\gamma,\tau^2), \tag{4.3}$$

where

$$\gamma = \delta\left\{\eta(a)\right\} b_2(c), \tag{4.4}$$

and

$$\tau^2 = \delta^2\left\{\eta(a)\right\} V_{22}(c). \tag{4.5}$$

The results in (4.3)–(4.5) indicate that $\eta(a)$ influences the asymptotic bias and variance of the estimated force of infection only by the term $\delta(\eta(a))$. The asymptotic mean square error of $\hat{\ell}(a)$ is given by

$$AMSE = \delta^2\left\{\eta(a)\right\}\left\{b_2^2(c) + V_{22}(c)\right\}. \tag{4.6}$$

The optimal choice for the constant c of the optimal bandwidth $h = cn^{-1/7}$ is the solution c_{opt} to

$$\frac{\partial b_2^2}{\partial c}(c_{opt}) + \frac{\partial V_{22}}{\partial c}(c_{opt}) = 0. \tag{4.7}$$

It follows from (4.7) that $\eta(a)$ is not directly involved in the determination of the optimal bandwidth which can be obtained by just minimizing the $AMSE$ of $\hat{\beta}_1(a)$ as an estimator for $\eta'(a)$. Fan and Gijbels (1996) explain in their Section 3.3 why the choice $p = 2$ is optimal for estimation of the first derivative $\eta'(a)$. Other odd choices for $p - 1$ are also appropriate but for most applications the choice $p = 2$ suffices. In the context of a binary

response, Fan, Heckman and Wand (1995) showed that for $p = 2$, the optimal constant c_{opt} is given by

$$c_{opt}(a) = \left\{ 27 \frac{\int K^{*2}(z)dz}{\left(\int z^3 K^*(z)dz\right)^2} \frac{\pi(a)(1-\pi(a))g'(\pi(a))^2}{f_A(a)\eta^{(3)}(a)^2} \right\}^{1/7} \tag{4.8}$$

where $f_A(a)$ is the (unknown) density of the age distribution. The factors in (4.8) depending on the so-called equivalent kernel K^* are known and integrate to a constant for a given kernel K (see Section 3.2.2 in Fan and Gijbels 1996). The unknown quantities $\eta^{(3)}(a)$, $\pi(a)$ can be estimated using initial estimators resulting from a global fractional polynomial (or any other flexible parametric) model and the density $f_A(a)$ can be estimated by a kernel estimator.

Note that the choice of the optimal bandwidth as discussed here minimizes the $AMSE$ for the local estimate of the force of infection. Another option is to minimize the $AMSE$ of $\hat{\pi}(a)$ which would optimally lead to a local linear ($p = 1$) instead of a local quadratic ($p = 2$) approach. In that case the optimal bandwidth is $c_{opt}n^{-1/5}$ with for c_{opt} an expression similar to (4.8) with $\eta^{(2)}(a)$ instead of $\eta^{(3)}(a)$.

As mentioned above, we need initial estimators for $\eta^{(3)}(a)$ (or $\eta^{(2)}(a)$ in case of a local linear model) and $\pi(a)$. Using smoothers again would require new bandwidth choices and would make the procedure unnecessarily complicated. At this stage, it is typically sufficient to estimate these unknown quantities based on a flexible parametric model. Here, fractional polynomial models discussed in the previous chapter, will be used.
Keiding's method (1991) to estimate $\pi(a)$ is based on isotonic regression of the observed prevalence on age and results in a step function for $\hat{\pi}(a)$. In practice, the *pool adjacent violator algorithm* (PAV) (Barlow, 1972) can be used to calculate $\hat{\pi}(a)$, which is monotone by construction. Our local polynomial smooth estimate $\hat{\pi}(a) = g^{-1}(\hat{\beta}_0(a))$ can be non-monotone as a function of age a and therefore result in negative estimates for the force of infection. Following Friedman and Tibshirani (1984) and Mammen *et al.* (2001), we suggest to estimate $\pi(a)$ and $\ell(a)$ (using the optimal bandwidth) and then, if necessary, to "isotonize" the estimates by the PAV algorithm. This is in line with the findings of Mammen *et al.* (2001). They showed that constrained smoothing leads to estimates of the form "smooth then constrain". One could also try estimates based on the idea "constrain then smooth" (as in Keiding 1991). For local polynomials this idea does not work: smoothing by polynomials is not monotonicity preserving.

4.4 Application to the Data

Throughout the analysis, the logit link function is used. We start with selecting the best fractional polynomial, by an extensive grid search over powers $\mathbf{p} = (p_1, p_2, \ldots p_m)$. This procedure was discussed in Chapter 3.

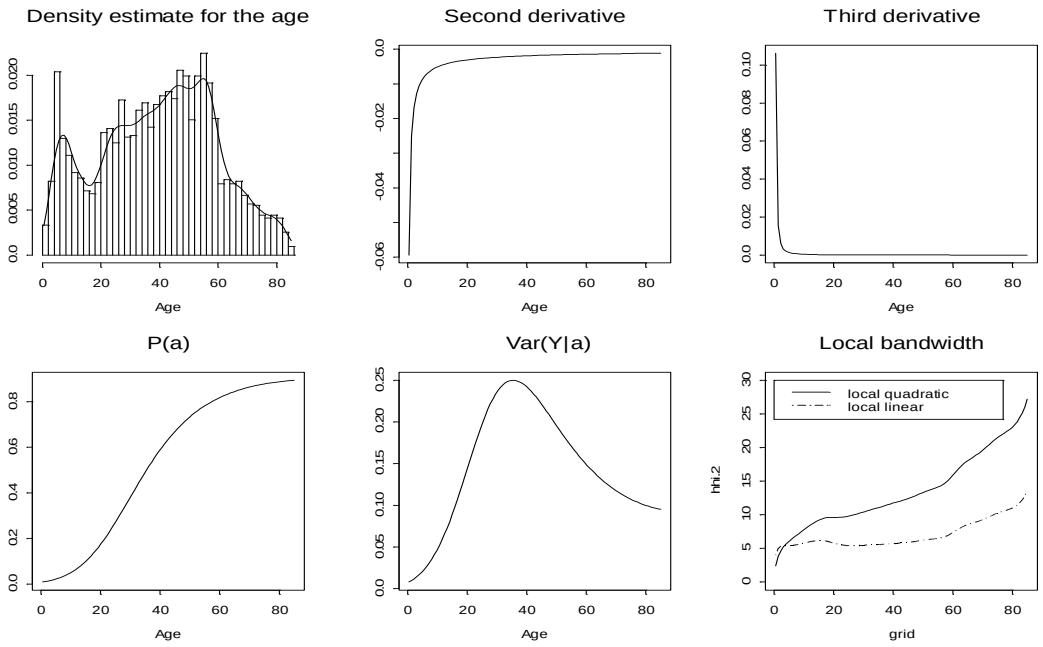


Figure 4.2: *Optimal local bandwidth for hepatitis A. From top left to bottom right: kernel density estimate for $f_A(a)$, estimates for $\hat{\eta}^{(2)}(a)$,$\hat{\eta}^{(3)}(a)$, $\hat{\pi}(a)$, $\widehat{Var}(Y|a) = \hat{\pi}(a)(1-\hat{\pi}(a))$ based on the optimal fractional polynomial of order 2 and the corresponding optimal local linear (dashed line) and local quadratic (solid line) bandwidth estimates.*

As mentioned in the previous section, the computation of the bandwidth of the local polynomial estimator requires estimates for the density $f_A(a)$, initial estimates for $\pi(a)$ and for the second (for $p = 1$) and third derivative (for $p = 2$) of the linear predictor $\eta(a)$. The age density was estimated with a kernel estimator, shown in Figure 4.2. The estimate for the linear predictor of the optimal fractional polynomial model is given by $\hat{\eta}(a) = \hat{\beta}_0 + \hat{\beta}_1 a^{\hat{p}_1} + \hat{\beta}_2 a^{\hat{p}_2}$ such that

$$\hat{\eta}^{(k)}(a) = \hat{\beta}_1 a^{\hat{p}_1 - k} \prod_{i=1}^{k} (\hat{p}_1 - i + 1) + \hat{\beta}_2 a^{\hat{p}_2 - k} \prod_{i=1}^{k} (\hat{p}_2 - i + 1).$$

As an initial estimate for $\hat{\pi}(a)$ we took the probability estimated by the optimal fractional polynomial. The local optimal bandwidth was then estimated according to (4.8). Figure 4.2 shows estimates for the different unknowns (for hepatitis A) in (4.8) and indicates that the local bandwidth which minimizes the $AMSE$ of $\hat{\eta}'(a)$ (with $p = 2$) is higher than the optimal bandwidth that minimizes the $AMSE$ of $\hat{\eta}(a)$ (with $p = 1$) (which is expected and reflects that more data are needed to fit locally a more complicated model). Figures 4.3a and 4.3b show the local linear and local quadratic fit for rubella. Up to age 20 both models predict the same patterns for the force of infection, except at age 1.5 where the local quadratic predicts higher values for the force of infection. From age 20 onwards, the local quadratic model predicts steadily decreasing trend with force of infection equal to zero from age 38 onwards. The local linear model indicates that the force of infection flattens around the value of 0.06 and even slightly increases from age 40 onwards. The estimated models for mumps are shown in Figure 4.3c and 4.3d. Both models predict

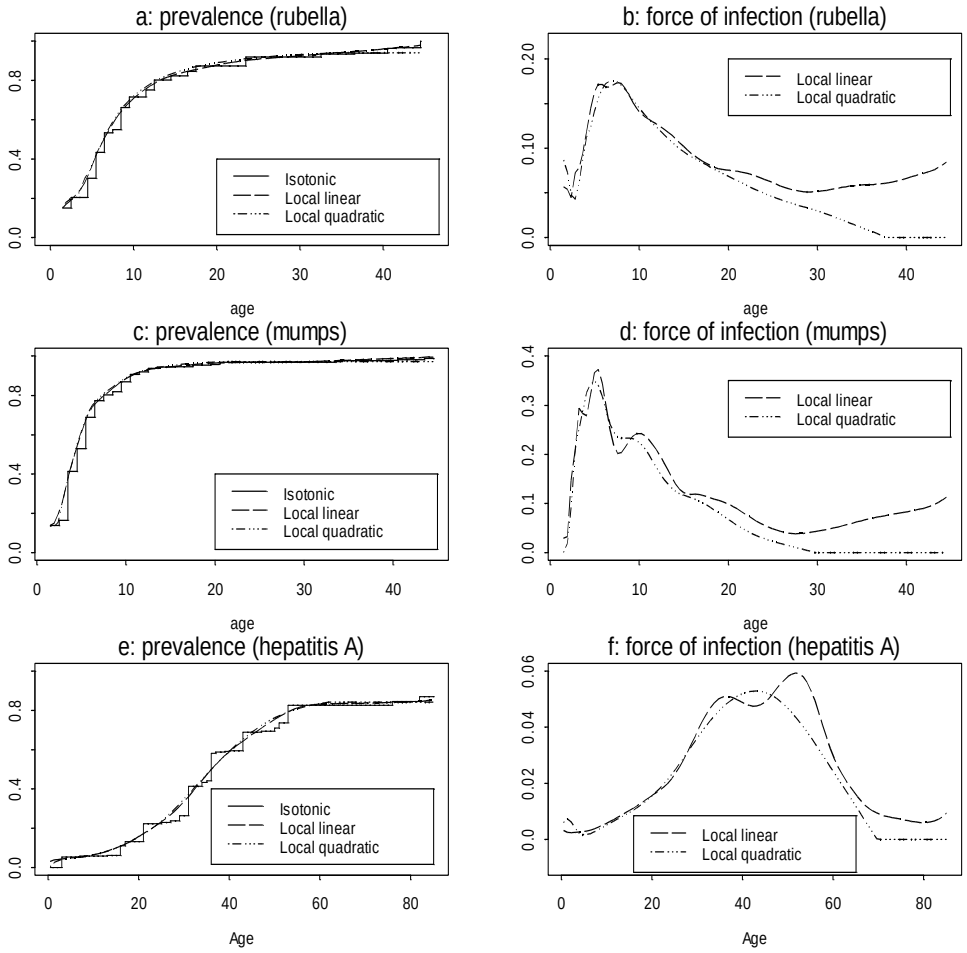


Figure 4.3: *Panels a, c and e. Estimated probability curve $\hat{\pi}(a)$ for rubella, mumps and hepatitis A: isotonic regression (solid line), local linear estimate (dashed line), local quadratic (three dots-dashed). Panels b, d and f. Estimated force-of-infection curve $\hat{\ell}(a)$: constrained local quadratic estimate (three dot-dash) and local linear estimate (dashed).*

a maximum around age 5 (4.97 and 5.4 for the local linear and local quadratic models, respectively), while the local linear model predicts a secondary peak at age 10. Due to the larger bandwidth, this peak is smoothed out by the local quadratic model. Similar to rubella, the local quadratic model predicts a steady decrease of the force of infection becoming zero from age 30 onwards while the local linear model indicates that the force of infection slightly increases as from age 30. Note that for both rubella and mumps the local quadratic estimate for the force of infection is zero at older age groups due to initialy negative estimates for the force of infection. The PAV algorithm was applied to the estimated prevalence at these age groups which leads to a nondecreasing prevalence curve and force of infection equal to zero at these age groups. For hepatitis A, the local linear model has a bimodal form with maxima at 28 and 55 years (Figure 4.1f). Note the close resemblance with Keiding's smoothed estimate with bandwidth $h = 15$, except for the smallest and largest ages where the local linear estimate seems to flatten out more nicely (less boundary effects). The local quadratic polynomial however produces negative estimated forces of infection from age 70.4 onwards. This is due to the larger values (at higher age groups) of the optimal bandwidth that was used to fit the quadratic local polynomial model. Again, the PAV algorithm was applied in order to "monotonize" the probability estimates and as a result the force of infection is estimated to be zero after age 70.4. The force of infection estimated with the local quadratic model shows a unimodal form with a maximum at age 40 and is quite similar to Keiding's smoothed estimate with bandwidth $h = 20$ (except for ages above 75). Since the optimal bandwidth for the local quadratic model is chosen to minimize the local $AMSE$ of the force of infection we recommend the local quadratic method.

To assess the local variability of $\hat{\ell}(a)$, a bootstrap procedure (see e.g. Davison and Hinkley, 1997) was applied to calculate pointwise confidence intervals for $\hat{\ell}(a)$. Specifically, B bootstrap samples were generated by resampling the original data (with replacement, each sample containing N pairs (a_i^*, Y_i^*)) and $(1-2\alpha)\times 100\%$ percentile confidence intervals $(\hat{\ell}^*(a)_{[(B+1)\alpha]}, \hat{\ell}^*(a)_{[(B+1)(1-\alpha)]})$ were calculated, where $\hat{\ell}^*(a)_{[(B+1)\alpha]}$ is the $(B+1)\alpha$th order statistic of the bootstrap replicated local forces of infection $\hat{\ell}_1^*(a), \ldots, \hat{\ell}_B^*(a)$. The same optimal local bandwidth as shown in Figure 4.2 was used (a data driven local bandwidth within each bootstrap run was computationally not feasible). Since the estimation procedure was not constrained for the bootstrap samples, estimates for the force of infection at higher ages might become negative, for both linear and quadratic models. Equivalently to the PAV algorithm, one can define the lower and upper confidence limits to be $\max\{0, \hat{\ell}^*(a)_{[(B+1)\alpha]}\}$ and $\max\{0, \hat{\ell}^*(a)_{[(B+1)(1-\alpha)]}\}$ respectively. Figure 4.4 shows local estimates for $\pi(a)$ and $\ell(a)$ for rubella and mumps together with their bootstrap confidence intervals. The variability of $\hat{\pi}(a)$ increases at older age groups, which can be explained by the smaller sample sizes at these age groups. For rubella, at younger age groups, the lower and upper confidence limits range between 0 to 0.15. Note that for mumps, from age 30 and onwards, the lower and upper confidence limits for the constrained force of infection are both zero.

Figure 4.5 shows the bootstrap estimates for the prevalence and force of infection for hepatitis A. The right hand panels in Figure 4.5 display the corrected pointwise confidence interval for HAV. The confidence intervals obtained from the linear polynomial turned out to be wider than those obtained from the quadratic model. This is due to the larger value of the optimal bandwidth that was used to estimate the quadratic model. Based on its optimal theoretical properties (being optimal for estimating the force of infection) and because of the observed superior accuracy characteristics, we recommend the use of

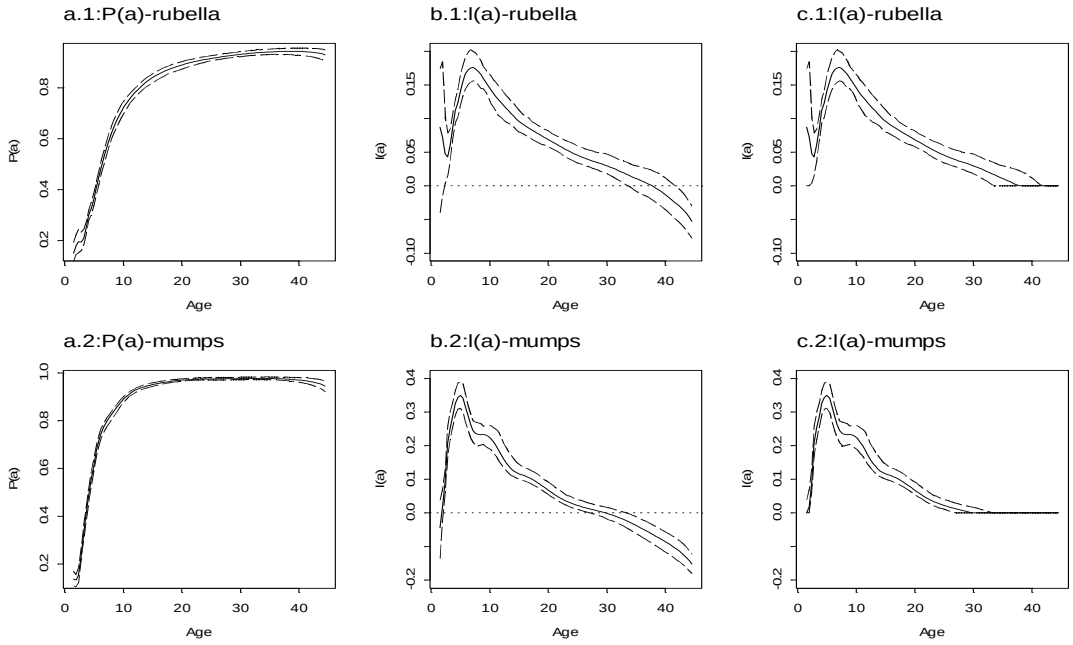


Figure 4.4: *Bootstrap confidence intervals for rubella (upper panels a1, b1, c1) and mumps (lower panels a2, b2, c2). From left to right: local quadratic estimates with confidence intervals for $\pi(a)$, $\ell(a)$ and constrained $\ell(a)$.*

the local quadratic smoother with optimal data-driven bandwidth to estimate the force of infection based on an age-specific prevalence sample.

4.5 Discussion

We have suggested to model the force of infection for rubella, mumps and hepatitis A using the nonparametric technique of local polynomial estimation. Specification of a fully parametric model for the linear predictor will inevitably restrict the shape of the estimated force of infection. This is not always recognized as being possibly too restrictive. It is here where nonparametric methods can contribute to the analysis and, because they are fully unconstrained and highly data-driven, they may reveal aspects of the data which are ignored or hidden by parametric models. Local polynomial estimators are consistent without model assumption (only require sufficient smoothness) and are known to have many desirable properties. This approach also allows simultaneous estimation of prevalence and force of infection. Asymptotic results for the local estimate of the force of infection were derived leading to a data-driven bandwidth selector. According to the principle "smooth then constrain", the fitted probability curve can be constrained to be monotone leading to a nonnegative estimated force of infection. Results from a simulation study, which will be discussed in the next chapter, show that estimates obtained from monotonized local polynomials are less variable than those based on isotonic regression. As an overall conclusion we recommend, based on theoretical considerations and our findings in the data analysis and the simulation study, the use of the local quadratic model to estimate the

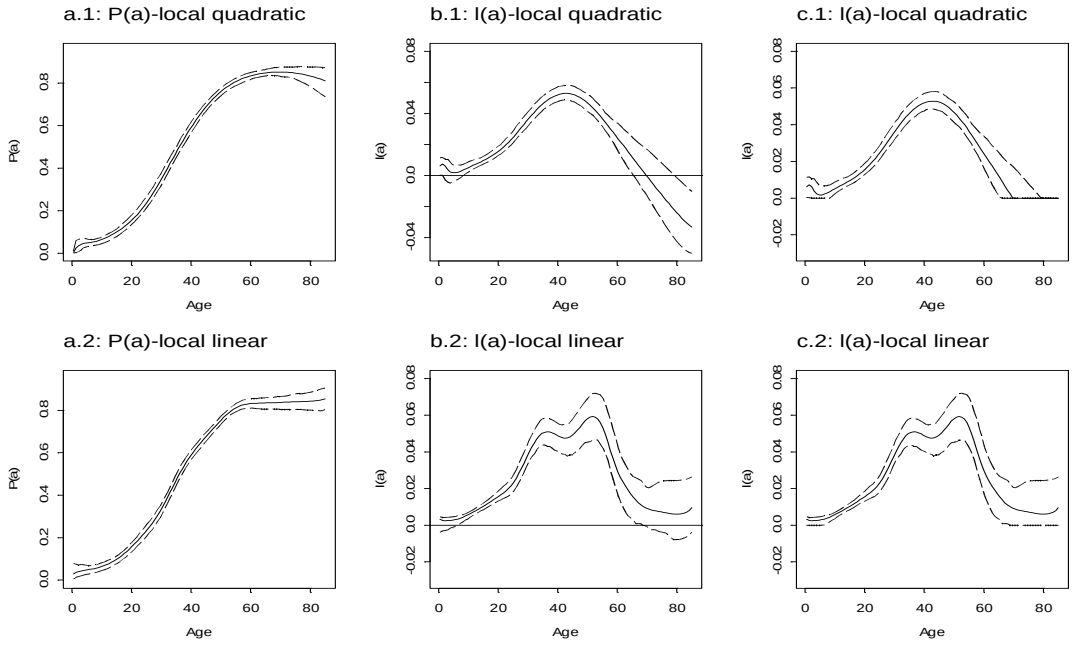


Figure 4.5: *Bootstrap confidence intervals for hepatitis A. From left to right: local estimates with confidence intervals for $\pi(a)$, $\ell(a)$ and constrained $\ell(a)$. Upper panels a1, b1, c1: local quadratic polynomials. Lower panels a2, b2, c2: local linear polynomials.*

force of infection.
In future research it will be examined how such a nonparametric smoothing method can be extended to estimate the force of infection of hepatitis A allowing time heterogeneity. Indeed, in contrast with rubella and mumps, an upward shift in the age at hepatitis A infection has been observed in industrialized countries following overall improvements in hygienic conditions in the second half of the 20th century. Also in Flanders a decrease in prevalence in the youngest age groups (0-14 years) has been observed by comparing the sample from 1993-94 with previous small samples obtained from Belgian first time blood donors in 1979 and 1989 (Beutels *et al.*, 1998). As a consequence of the age shift the assumption of time independence is likely to be violated in relation to hepatitis A. We shall further study this issue when we analyze a new sample to be taken in 2001-2.

Appendix

The following modification of the delta method is used to derive the asymptotic normality result (4.3).

Lemma *Let $\{\boldsymbol{T}_n\} = \{(T_{1n}, T_{2n})\}$ be a sequence of bivariate estimators for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ such that, as $n \to \infty$*

$$\begin{bmatrix} \sqrt{a_n}(T_{1n} - \theta_1) \\ \sqrt{b_n}(T_{2n} - \theta_2) \end{bmatrix} \xrightarrow{D} N(\boldsymbol{b}, V),$$

where a_n and b_n are sequences of constants tending to infinity, $\boldsymbol{b} = (b_1, b_2)$ represents the asymptotic bias and V the asymptotic covariance matrix.
Consider a real-valued function $\varphi(\boldsymbol{t}) = \varphi(t_1, t_2)$ such that $(\partial\varphi/\partial t_1, \partial\varphi/\partial t_2)$ is non-null at $\boldsymbol{t} = \boldsymbol{\theta}$ and continuous in a neighborhood of $\boldsymbol{\theta}$. If $\delta_n = a_n/b_n \to \infty$, then as $n \to \infty$

$$\sqrt{b_n}\{\varphi(\boldsymbol{T}_n) - \varphi(\boldsymbol{\theta})\} \xrightarrow{D} N(\gamma, \tau^2), \tag{4.9}$$

where $\gamma = \dfrac{\partial\varphi}{\partial t_2}(\theta) b_2$ *and* $\tau^2 = \left(\dfrac{\partial\varphi}{\partial t_2}(\theta)\right)^2 V_{22}$.

Proof
We have that

$$\sqrt{b_n}\,\{\varphi(\boldsymbol{T}_n) - \varphi(\boldsymbol{\theta})\}$$
$$= \sqrt{\frac{a_n}{\delta_n}}(T_{1n} - \theta_1)\tilde{\varphi}_1(\boldsymbol{T}_n, \boldsymbol{\theta}) + \sqrt{b_n}(T_{2n} - \theta_2)\tilde{\varphi}_2(\boldsymbol{T}_n, \boldsymbol{\theta})$$

where

$$\tilde{\varphi}_1(\boldsymbol{T}_n, \boldsymbol{\theta}) = \frac{\varphi(T_{1n}, \theta_2) - \varphi(\theta_1, \theta_2)}{(T_{1n} - \theta_1)},$$
$$\tilde{\varphi}_2(\boldsymbol{T}_n, \boldsymbol{\theta}) = \frac{\varphi(T_{1n}, T_{2n}) - \varphi(T_{1n}, \theta_2)}{(T_{2n} - \theta_2)}.$$

Since $\boldsymbol{T}_n \xrightarrow{P} \boldsymbol{\theta}$, it follows that for $i = 1, 2$

$$\tilde{\varphi}_i(\boldsymbol{T}_n, \boldsymbol{\theta}) \xrightarrow{P} \frac{\partial\varphi}{\partial t_i}(\boldsymbol{\theta}).$$

Applying Slutsky's theorem and the fact that $\delta_n \to \infty$ together with $\sqrt{a_n}(T_{n_1} - \theta_1) = O_P(1)$, we get

$$\sqrt{b_n}\,\{\varphi(\boldsymbol{T}_n) - \varphi(\boldsymbol{\theta})\} \xrightarrow{D} N(\gamma, \tau^2).$$

Chapter 5

Estimation From Serological Data: A Simulation Study

5.1 Introduction

Estimation of the prevalence and the force of infection from serological sample is closely related to the problem of estimation from current status data (Keiding 1996). In fact, a seroprevalence sample is a special case of current status data. While in the literature related to estimation from current status data attention is placed on the prevalence, in the context of infectious diseases attention is placed on the estimation of the force of infection (and other parameters of the disease). Consider a seroprevalence sample of size N and let Z_i be the age of immunization of the ith individual, $i = 1, 2, \ldots, N$, A_i be the current age of individual i and $Y_i = I(A_i > Z_i) = 1$ if the individual is immuned, $Y_i = 0$ if not. Note that Y_i is the indicator variable that we defined in equation (3.1) in Chapter 3. Thus, the seroprevalence sample consists of N observations, $(A_1, Y_1), \ldots, (A_N, Y_N)$. Note that A_i is either right or left censored. Let $\pi(a_i) = P(Y_i = 1) = P(A_i > Z_i)$, then the likelihood of $(Y_1, \ldots, Y_N)$ is

$$L = \prod_{i=1}^{N} \pi(a_i)^{Y_i}(1 - \pi(a_i))^{(1-Y_i)}. \tag{5.1}$$

Note that the observation unit in (5.1) is the individual. Alternatively, for data grouped in n unique age groups, $a_1 < a_2 < \ldots < a_n$ the likelihood is

$$L = \prod_{j=1}^{n} \pi(a_j)^{\Sigma_{l=1}^{n_j} Y_{jl}}(1 - \pi(a_j))^{(n_j - \Sigma_{l=1}^{n_j} Y_{jl})}. \tag{5.2}$$

Here, n_j is the sample size in the j'th age group and $N = \sum_{j=1}^{n} n_j$. The nonparametric maximum likelihood estimators (NPMLE) for (5.1) and (5.2) are identical, it is the isotonic regression of the observed prevalence ,$\hat{\pi}(a_j)$, with weights n_j. The NPMLE is a step function with respect to age. Let K be the final number of sets (or the final number of steps), and $\tilde{a}_l, \tilde{a}_2, \ldots, \tilde{a}_K$ be the jump points. Barlow *et al.* (1972) and Robertson *et al.* (1988) discussed the NPMLE for the general case (using the form of the likelihood in (5.2)) and Keiding (1991), Keiding *et al.* (1996) and Greenhalgh and Dietz (1994) discussed

it in the context of infectious diseases (using the form of the likelihood in (5.1)).
In this chapter the performance of the isotonic regression and the local polynomial models are compared by means of a simulation study. In particular the bias, variability and mean square error of the estimator for both the prevalence and the force of infection are evaluated. We will first show, in Section 5.2, that given the number of final levels, the isotonic regression estimators have the same form as the Nadaraya-Watson estimators. Hence, given the final number of levels, the isotonic regression can be expressed as a local constant model. In Sections 5.3 we present the structure of the simulation study, the results are presented in Sections 5.4 and 5.5, where we compare the performance of the models when estimating the prevalence and the force of infection.

5.2 Isotonic Regression and Local Polynomials

Let $\boldsymbol{\pi} = (\hat{\pi}(a_1), \hat{\pi}(a_2), \ldots, \hat{\pi}(a_n))$ be the observed prevalence at age groups $a_1, a_2, \ldots, a_n$ (hence, the unrestricted maximum likelihood estimates) and $n_1, n_2, \ldots, n_n$ be the sample sizes at each age group. It follows that the nonparametric maximum likelihood estimate under order restriction is the isotonic regression of $\hat{\pi}(a_j)$ with weights n_j (Barlow, 1972). We denote the isotonic regression by $\boldsymbol{\pi}^\star = (\hat{\pi}^\star(a_1), \hat{\pi}^\star(a_2), \ldots, \hat{\pi}^\star(a_n))$. Let ϕ_k be the k'th final set, $k = 1, \ldots, K$ and $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{K+1}$ be the jump points with $\tilde{a}_1 = a_1$ and $\tilde{a}_{K+1} = a_n$. For all age groups satisfy $a_{u-1} < \tilde{a}_k < a_u < a_{u+1} <, \ldots, < a_{u+v} < \tilde{a}_{k+1} < a_{u+v+1}$ the NPMLE for the prevalence is the same, $\pi^\star_{a_u} = \pi^\star_{a_{u+1}} = \ldots = \pi^\star_{a_{u+v}}$. It is easy to see that, given the final numbers of sets, the isotonic regression can be expressed as a linear smoother of the observed prevalence,

$$\hat{\boldsymbol{\pi}}^\star = \mathbf{S}\hat{\boldsymbol{\pi}}, \tag{5.3}$$

where $\mathbf{S}_{n\times n}$ is a block diagonal smoothing matrix,

$$\mathbf{S} = \begin{pmatrix} S_1 & 0 & 0 & \ldots & 0 \\ 0 & S_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & S_K \end{pmatrix}. \tag{5.4}$$

for which the jl'th entry is given by

$$[\mathbf{S}]_{jl} = \begin{cases} \frac{n_j}{\sum_{j\in\phi_k} n_j} & \tilde{a}_k < a_j \leq \tilde{a}_{k+1} \quad \text{, age group } j \text{ belong to the } k\text{'th set} \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

Suppose that age groups $j, j+1, \ldots, j+m$ belong to the final set ϕ_k then the rows in the corresponding sub-matrix are identical and given by

$$[\mathbf{S}_k]_{j\cdot} = \left(\frac{n_j}{\sum_{j\in\phi_k} n_j}, \frac{n_{j+1}}{\sum_{j\in\phi_k} n_j}, \ldots, \frac{n_{j+m}}{\sum_{j\in\phi_k} n_j} \right). \tag{5.6}$$

It follows that $\Sigma_j [\mathbf{S}]_{ij} = 1$ and $trace(\mathbf{S}) = K$. Within the framework of nonparametric regression the trace of the smoothing matrix is equivalent to the effective number of parameters (Hastie and Tibshirani, 1991), in our setting it is simply the final number of sets.

Now, one can observe that given the final number of sets and the locations of the jumps, the isotonic regression of $\pi(a_j)$ with weights n_j can be expressed as a local polynomial with $p = 0$ (local constant) for which the local log likelihood (for the individual data) is given by

$$L(\beta_0, a) = \sum_{i=1}^{N} Q_i \left\{ (g^{-1}(\beta_0), Y_i \right\} K(A_i - a), \tag{5.7}$$

where $K(u)$ is the following kernel function

$$K(a) = \begin{cases} 1 & \tilde{a}_k < A_i \leq \tilde{a}_{k+1} \quad \text{if the age of the individual belongs to the } k\text{'th set} \\ 0 & \text{otherwise.} \end{cases} \tag{5.8}$$

Let us assume that the data are sorted by age, $A_1 \leq A_2 \leq, \ldots, \leq A_N$. For $\tilde{a}_k \leq A_i \leq A_{i+1} \leq \tilde{a}_{k+1}$ the local intercepts are the same, $\hat{\beta}_{0,i} = \hat{\beta}_{0,i+1}$, since the weight function is 1. Furthermore, for all s and u such that $\tilde{a}_k \leq A_s \leq \tilde{a}_{k+1} \leq A_u$ it follows that $\hat{\beta}_{0,s} \leq \hat{\beta}_{0,u}$. Fan and Gijbels (1996) showed that for $p = 0$, $\hat{\beta}_0(a)$ is a Nadaraya-Watson estimator, that is

$$\hat{\beta}_0(a) = g \left\{ \frac{\sum K(A_i - a) Y_i}{\sum K(A_i - a)} \right\}. \tag{5.9}$$

Note that for the kernel function specified in (5.8) $\sum K(A_i - a) Y_i / \sum K(A_i - a) = \sum_{i \in \phi_k} Y_i / n_k$, where n_k is the number of individuals for which $\tilde{a}_k < A_i \leq \tilde{a}_{k+1}$ (i.e., the number of individuals in the k'th set). Now, since $\hat{\pi}_{LC}(a) = g^{-1} \left\{ \hat{\beta}_0(a \right\}$ (LC stands for local constant with the kernel function defined in (5.8)), it follows that

$$\hat{\pi}_{LC} = g^{-1} g \left\{ \frac{\sum K(A_i - a) y_i}{\sum K(A_i - a)} \right\} = \hat{\pi}^{\star}(a). \tag{5.10}$$

Thus, the local constant model, with kernel function (5.8) reproduces the isotonic regression. In other words, the isotonic regression should be seen as a local constant model in which the bandwidth is chosen in order to ensure monotonicity.

There are two issues that rise now: (1) the order of the local polynomial and (2) the choice of the bandwidth. Compared to the local linear model the local constant has one local parameter less. However, Fan and Gijbels (1996) showed that both models have the same asymptotic variance but the local linear model has smaller asymptotic bias. The choice of the bandwidth is a crucial point as well. It is here that the concept of smooth then constrain becomes such an attractive modeling approach. If monotonicity is the only consideration in modeling then isotonic regression is a satisfactory estimator. However if in addition to monotonicity one would like to take the mean square error of $\hat{\pi}(a)$ into account then a local linear model with optimal bandwidth is a better choice. Furthermore, if one wishes to minimize the mean square error of the estimator for the force of infection, a local quadratic model is a natural choice, as we have shown in the pervious chapter. Monotonicity can be achieved by applying the PAV algorithm to the unconstrained models, as done in the previous chapter.

5.3 Simulation Structure

We performed a simulation study to investigate the performance of the monotonized local polynomial models compared to the isotonic regression approach. Three test functions were used

$$(1) \quad \pi(a) = \frac{\exp\{-4.98-0.0258a^{1.3958}+0.3081a^{0.9375}\}}{1+\exp\{-4.98-0.0258a^{1.3958}+0.3081a^{0.9375}\}},$$

$$(2) \quad \pi(a) = \frac{0.02a^{1.5}}{1+0.02a^{1.5}},$$

$$(3) \quad \pi(a) = \frac{0.01a^{2.2}}{1+0.01a^{2.2}}.$$

The first test function considered is the estimated fractional polynomial for the hepatitis A example from the previous chapter. The second and the third test functions represent log logistic distributions with force of infection peaked at age 5 and 10 respectively. The test functions and the corresponding forces of infection are shown in Figure 5.1. For each test function three sets of simulation were conducted. In the first the same sample size and age values as in the hepatitis A dataset were used. For the other two sets of simulations, sample sizes at each age group were equal to 75 and 50, respectively. For each set of simulations, $M = 150$ new datasets were generated with the number of infected individuals at age a drawn from the binomial distribution with probability $\pi(a)$. In each simulation run, $\pi(a)$ was estimated by isotonic regression and by monotonized local linear and local quadratic polynomials. For the first test function we used fixed global bandwidths, approximately the global average of the optimal local bandwidths as depicted in Figure 4.2, leading to $h = 7$ for the local linear and $h = 13$ for the local quadratic fits. For the other two test functions we calculate the optimal local bandwidth using the true values of $\pi(a)$ and $\eta^{(k)}(a)$. For Keiding's (1991) smoothed estimate for the force of infection we took $h = 10$. Let $\hat{\pi}_j(a)$ be the estimated probability at age a in the jth simulation, $j = 1, 2, \ldots, M$. The local squared bias is estimated by $\hat{b}^2(a) = \{\bar{\hat{\pi}}(a) - \pi(a)\}^2$, with $\bar{\hat{\pi}}(a) = \Sigma_{j=1}^{M}\hat{\pi}_j(a)/M$ and the local variance is estimated by $\hat{v}(a) = \sum_{j=1}^{M}\{\hat{\pi}_j(a)-\bar{\hat{\pi}}(a)\}^2/M$, leading to the simulation estimate for the local mean squared error MSE, given by $\widehat{MSE}(a) = \hat{b}^2(a) + \hat{v}(a)$.
At each simulation the force of infection was estimated according to (4.1) for the local polynomial models and using a kernel smoother for the isotonic regression. Let $\hat{\ell}_j(a)$ be the estimated force of infection at age a in the jth simulation, $j = 1, 2, \ldots, M$ and $\bar{\hat{\ell}}(a) = \Sigma_{j=1}^{M}\hat{\ell}_j(a)/M$. The local squared bias, variance and mean square error were calculated using $\hat{b}^2(a) = \{\bar{\hat{\ell}}(a) - \ell(a)\}^2$, $\hat{v}(a) = \sum_{j=1}^{M}\{\hat{\ell}_j(a) - \bar{\hat{\ell}}(a)\}^2/M$.

5.4 Results: Prevalence

Figure 5.2 shows the results for the first test function. The three different mean curves $\bar{\hat{\pi}}(a)$ can hardly be distinguished, except in the last 5 age groups (80-85) where the probabilities estimated by the isotonic regression increase. Between age 1 and 80, the local squared bias of the three models is essentially the same. The local variance of the isotonic regression is however much higher than the local variance of monotonized local polynomial models.

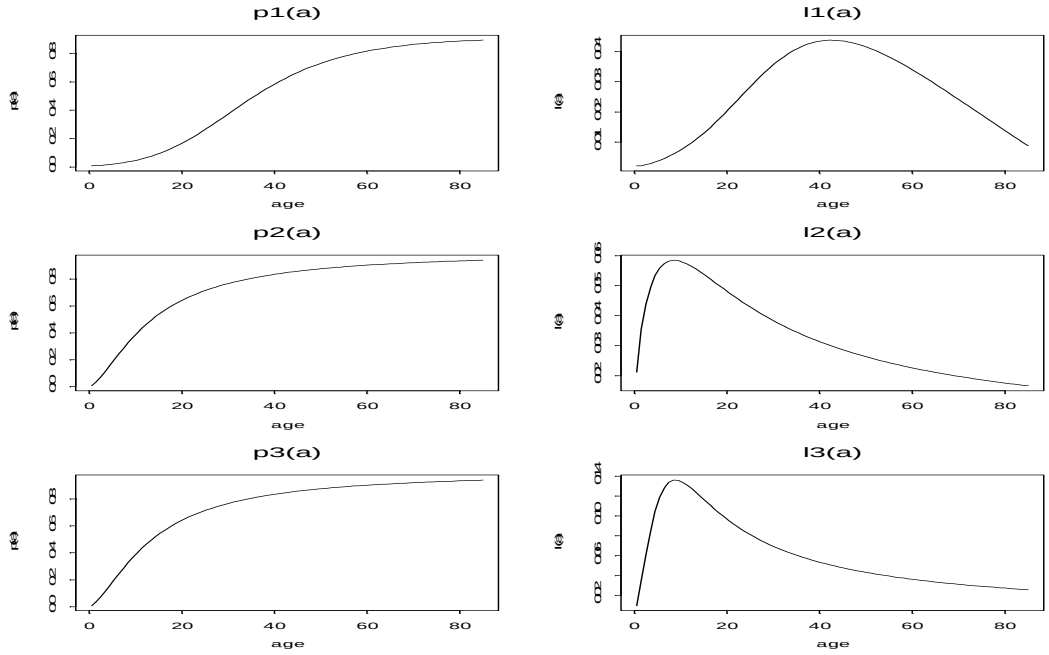


Figure 5.1: *Prevalence and force of infection for the test functions.*

Since the local variance is the dominant term in $\widehat{MSE}(a)$, the isotonic regression model has also higher values for $\widehat{MSE}(a)$. This pattern can also be seen in panel b, which shows the 5% and the 95% quantiles of the isotonic regression and the local polynomial models. The variability in the isotonic regression model is clearly higher than in the local polynomial models. Table 5.1 shows global simulated squared bias, variance and MSE, averaging over all age groups. The global MSE of the isotonic regression is 3.12 times higher than the global MSE of the local linear model and 4.41 times higher than the global MSE of the local quadratic models. The results remain essentially the same for the trimmed (5%) means. The same patterns are observed when sample size at each age group is 75 and 50. For the second test function (see results in Table 5.2, upper row) the local linear model has smaller global MSE for $\pi(a)$. Both local linear and local quadratic models performed much better than the isotonic regression. Similar results obtained for the third test function (see Table 5.3, upper row)

5.5 Results: Force Of Infection

The lower part in Table 5.1 shows that the local quadratic model has the smallest global MSE ,0.78×10^{-4} compared to 1.21×10^{-4} and 8.6×10^{-4} for the local linear model and the isotonic regression respectively. Figure 5.3 displays the simulation results for $\ell(a)$. The variability of $\ell(a)$ increases with age in all models but the local polynomial models have smaller square bias and variance than the isotonic regression model (locally and globally). Note that the pattern of increasing variability in panel b was already observed in the nonparametric bootstrap estimate for the confidence intervals of the force of infection (Figure 4.5). Although these results should be interpreted with some caution (no optimal

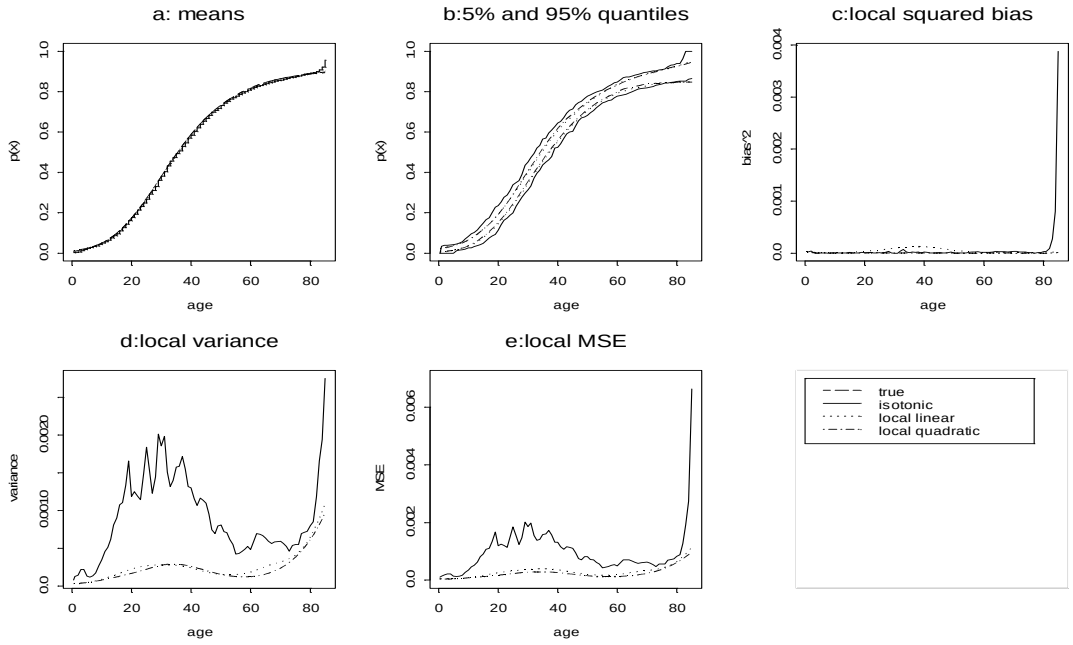


Figure 5.2: *Test function #1. Simulation results for $\pi_A(a)$, including isotonic regression (solid lines), monotonized local linear (dotted lines) and local quadratic (dot-dashed lines) estimates. From top left to bottom right: average probability estimates $\bar{\hat{\pi}}(a)$, 5% and 95% quantiles, simulated squared bias $b^2(a)$, variance $v(a)$ and mean squared error $b^2(a)+v(a)$.*

bandwidths were used), there is a clear preference for the local polynomial models with some advantage for the local quadratic model. When sample size at each age group equal to 75 and 50, the global MSE of all models reduced. But still the local quadratic model has the smallest global MSE.
The lower row in Table 5.2 shows that the local quadratic model performed better than the local linear in terms of the global MSE of the force of infection. Both local polynomial models performed better than the isotonic regression model. Similar patterns were observed for the third test function as shown in the lower row in Table 5.3.

5.6 Discussion

Our simulation study shows that in terms of global MSE for the prevalence and the force of infection, the isotonic regression model cannot compete with the local polynomial models (both linear and quadratic models). Between the local polynomials, the local quadratic model has an advantage in terms of the global MSE of the force of infection.

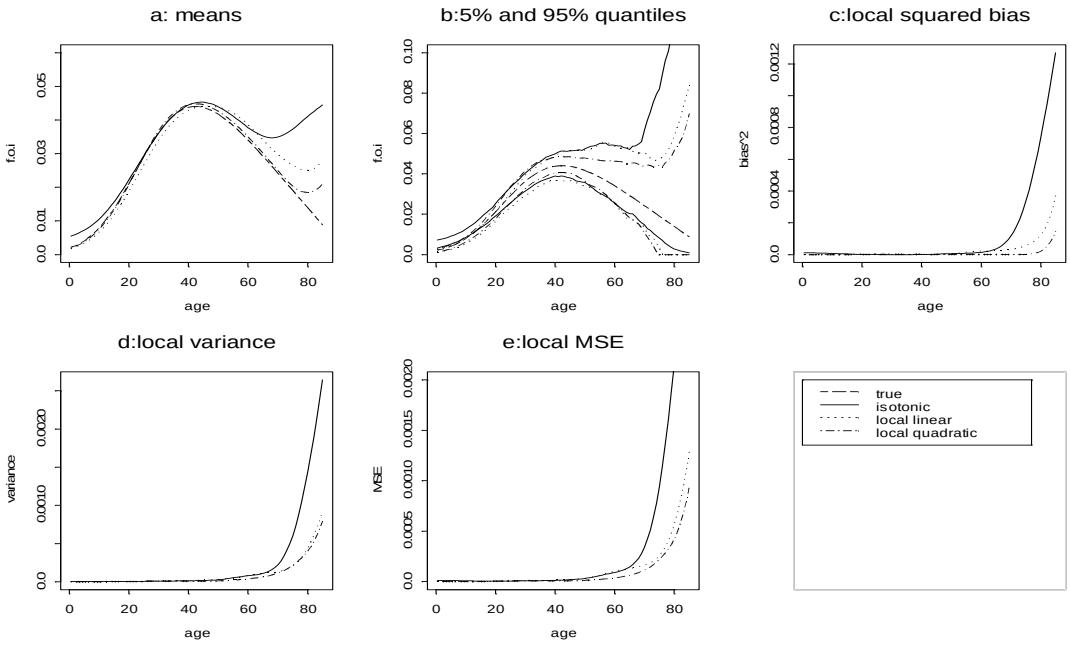


Figure 5.3: *Test function #1. Simulation results for the force of infection $\ell(a)$, including estimates based on isotonic regression (solid lines), monotonized local linear (dotted lines) and local quadratic (dot-dashed lines) models. From top left to bottom right: average probability estimates $\bar{\ell}(a)$, 5% and 95% quantiles, simulated squared bias $b^2(a)$, variance $v(a)$ and mean squared error $b^2(a) + v(a)$. Panels a and b also show the true force of infection $\ell(a)$ (dashed lines).*

Table 5.1: *Simulation results for Test function # 1: global simulated squared bias, variance and mean squared error ($\times 10^4$) for isotonic regression, monotonized local linear and local quadratic fits. The numbers in second line are the trimmed means (trim = 5%). The second column in the table presents the results when sample seizes were the same as in the hepatitis A datasets.*

					$n_j = 75$			$n_j = 50$		
		local linear	local quadratic	isotonic regression	local linear	local quadratic	isotonic regression	local linear	local quadratic	isotonic regression
	$\bar{b}^2$	$0.43^{(a)}$	0.06	0.97	0.44	0.03	0.11	0.42	0.01	0.13
		$0.42^{(b)}$	0.05	0.29	0.43	0.03	0.07	0.4	0.01	0.06
$\pi(a)$	$\bar{v}$	2.3	2.0	8.38	1.08	0.99	5.22	1.45	1.34	6.67
		2.16	1.88	8.26	1.08	0.99	5.14	1.45	1.34	6.6
	$\overline{MSE}$	2.73	1.93	8.53	1.52	1.02	5.33	1.87	1.36	6.8
		2.60	1.56	9.09	1.52	1.02	5.25	1.86	1.36	6.73
	$\bar{b}^2$	0.25	0.06	1.54	0.08	0.02	1.2	0.07	0.01	2.38
		0.18	0.03	1.15	0.05	0.01	0.55	0.05	0.01	1.04
$\ell(a)$	$\bar{v}$	0.89	0.72	7.06	0.3	0.23	3.39	0.4	0.3	8.01
		0.72	0.57	5.56	0.27	0.19	1.92	0.36	0.26	3.77
	$\overline{MSE}$	1.21	0.78	8.6	0.38	0.24	4.58	0.47	0.32	10.39
		0.98	0.60	6.71	0.32	0.21	2.48	0.41	0.27	4.81

(a)-mean, (b)-trimmed mean.

Table 5.2: *Simulation results for Test function # 2: global simulated squared bias, variance and mean squared error ($\times 10^4$) for isotonic regression, monotonized local linear and local quadratic fits. The numbers in the second line are the trimmed means (trim = 5%). The second column in the table presents the results when sample seizes were the same as in the hepatitis A datasets.*

					$n_j = 75$			$n_j = 50$		
		local linear	local quadratic	isotonic regression	local linear	local quadratic	isotonic regression	local linear	local quadratic	isotonic regression
	$\bar{b}^2$	0.52	0.69	0.99	0.36	0.05	0.09	0.4	0.08	0.09
		0.46	0.6	0.67	0.33	0.05	0.06	0.39	0.07	0.04
$\pi(a)$	$\bar{v}$	2.72	3.63	8.46	1.27	1.59	4.79	1.8	2.29	6.28
		2.48	3.32	8.13	1.21	1.5	4.6	1.71	2.17	6.03
	$\overline{MSE}$	3.24	4.32	9.45	1.63	1.64	4.88	2.2	2.37	6.36
		2.98	4.02	9.14	1.55	1.55	4.69	2.1	2.25	6.12
	$\bar{b}^2$	0.51	0.41	1.91	11.21	10.21	11.58	11.06	9.86	12.33
		0.4	0.32	1.42	10.2	9.39	10.64	10.17	9.18	11.49
$\ell(a)$	$\bar{v}$	0.75	1.2	11.02	0.4	0.49	4.87	0.5	0.59	5.86
		0.55	0.6	8.77	0.39	0.36	4	0.46	0.5	4.12
	$\overline{MSE}$	1.27	1.61	12.93	11.61	10.7	16.45	11.56	10.45	18.19
		1.04	1.04	10.19	10.59	9.84	14.64	10.63	9.71	16.08

Table 5.3: *Simulation results for Test function # 3: global simulated squared bias, variance and mean squared error ($\times 10^4$) for isotonic regression, monotonized local linear and local quadratic fits. The numbers in the second line are the trimmed means (trim = 5%). The second column in the table presents the results when sample seizes were the same as in the hepatitis A datasets.*

					$n_j = 75$			$n_j = 50$		
		local linear	local quadratic	isotonic regression	local linear	local quadratic	isotonic regression	local linear	local quadratic	isotonic regression
	$\bar{b}^2$	0.4	0.15	0.11	0.15	0.09	0.03	0.2	0.09	0.02
		0.36	0.12	0.09	0.13	0.08	0.02	0.18	0.09	0.02
$\pi(a)$	$\bar{v}$	2.01	2.37	5.92	0.97	1.12	3.32	1.43	1.74	4.76
		1.79	2.13	5.26	0.86	0.98	2.93	1.26	1.5	4.17
	$\overline{MSE}$	2.41	2.52	6.03	1.12	1.21	3.35	1.64	1.83	4.78
		2.18	2.27	5.37	1	1.07	2.95	1.44	1.59	4.19
	$\bar{b}^2$	8.91	2.43	6.77	2.46	1.9	6.17	4.29	1.96	5.66
		8.27	2.37	6.51	2.31	1.82	5.66	4.08	1.89	5.35
$\ell(a)$	$\bar{v}$	2.18	2.60	27	1.39	0.95	27.55	1.67	1.23	24.52
		2.06	1.91	24.12	1.33	0.83	25.1	1.56	1.13	22.35
	$\overline{MSE}$	11.09	5.03	33.77	3.85	2.85	33.72	5.95	3.19	30.18
		10.5	4.43	30.61	3.72	2.79	30.85	5.77	3.16	27.69

Chapter 6

Hierarchical Nonparametric Bayesian Models for the Force of Infection for Mumps and Rubella

6.1 Introduction

So far the prevalence and the force of infection were estimated within the frequentist framework. In this chapter we estimate the prevalence and the force of infection within the Bayesian framework.
Farrington (1990) and Farrington *et al.* (2001) proposed nonlinear models for the force of infection based on prior knowledge about the relationship between the force of infection and the host age. In Farrington (1991) and Edmunds *et al.* (2000) the force of infection is defined by

$$\ell(a) = (\alpha_1 a - \alpha_3)e^{-\alpha_2 a} + \alpha_3. \tag{6.1}$$

In order to ensure that the force of infection satisfies $\ell(a) \geq 0$, Farrington (1990) constrained the parameter space to be nonnegative ($\alpha_j \geq 0$, $j = 1, 2, 3$).
In a parametric Bayesian framework the prevalence π_i, $i = 1, 2, \ldots, n$, has a parametric form, $\pi(a_i, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a parameter vector. In this case $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ has a deterministic relationship with the predictor a and one may need to constrain the parameters space of the prior distribution $P(\boldsymbol{\alpha})$ in order to achieve monotonicity of the posterior distribution $P(\pi_1, \pi_2, \ldots, \pi_n|\mathbf{y})$, where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and y_i is the number of infected individuals at age a_i. For example, constrain the parameters to be nonnegative as done by Farrington (1990) in a frequentist setting. Hierarchical nonlinear and generalized linear models for prevalence and the force of infection will be discussed in Section 6.3.
In Section 6.4 we turn back to the nonparametric framework. Within the framework of fully nonparametric Bayesian modeling, the problem is to estimate $\pi_1, \pi_2, \ldots, \pi_n$ under the order restriction $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_n$. The prevalence is assumed to be an isotonic nonparametric function satisfying $0 \leq \pi_i \leq 1$. Likewise, in the hierarchical parametric Bayesian models we focus on the posterior distribution of the prevalence $P(\boldsymbol{\pi}|\mathbf{y})$. The n dimensional parameter vector is constrained to lie in a subset S^n of R^n. The constrained set S^n is determined by the order restrictions among the components of $\boldsymbol{\pi}$. In

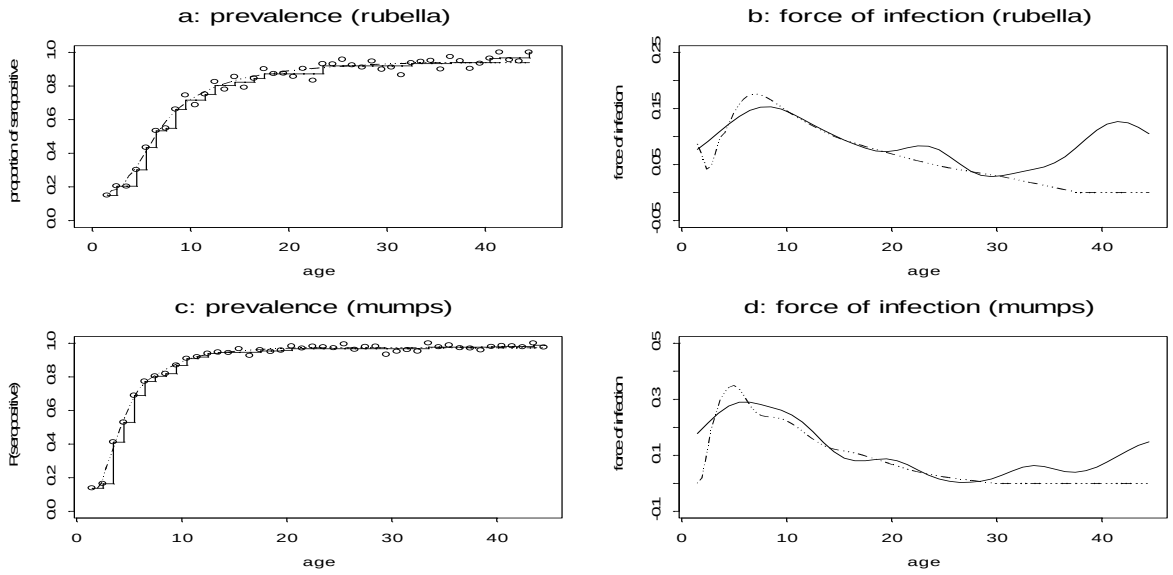


Figure 6.1: *Nonparametric estimate for the prevalence and the force of infection. Solid line: isotonic regression for the prevalence and kernel smoother for the force of infection. Dashed/dotted line: local quadratic model.*

this case it is natural to incorporate the constrains into the specification of the prior distribution, $P(\boldsymbol{\pi})$. In the context of bioassay modeling, Gelfand and Kuo (1991) showed that the constrained posterior distribution has the same form as the unconstrained posterior distribution restricted to the constrained set. This implies that if $P(\boldsymbol{\pi})$ is a product-beta distribution, and the likelihood $P(\mathbf{y}|\boldsymbol{\pi})$ is binomial, then the posterior distribution $P(\pi_i|\mathbf{y}, \boldsymbol{\pi}_{-i})$, $\boldsymbol{\pi}_{-i} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$, is a beta distribution restricted to the interval $[\pi_{i-1}, \pi_{i+1}]$. The hierarchical nonparametric approach will be discussed in Section 6.4. The methods are illustrated on the rubella and mumps datasets which presented briefly in Section 6.2.

6.2 Exploratory Data Analysis

Although isotonic regression and local polynomials models were discussed in Chapter 3, we briefly discuss the results for rubella and mumps in this section as well. Figure 6.1 shows both local polynomials and isotonic regression estimates for $\pi(a)$ and $\ell(a)$. For rubella (see Figure 6.1, panel b), the estimated force of infection, estimated by the local quadratic model, rises steeply to a peak at age 7-8 followed by a steady decrease to zero at older age groups. The two methods result in somewhat different patterns. The force of infection estimated by the kernel estimate predicts a secondary peak at age 24 and a third peak at age 40. Panel d in Figure 6.1 reveals the same patterns for mumps. We note that the second peak at age 10 estimated by the local polynomial is smoothed out by the kernel smoother which also predicts a third peak at age 33.

6.3 Hierarchical Bayesian Models for the Force of Infection

6.3.1 Non-linear Hierarchical Model

The model in (6.1) assumes that the force of infection is zero at birth ($\ell(0) = 0$) and then rises to a peak in a linear fashion followed by an exponential decrease. The peak is reached at an age corresponding to the maximum contact rate of susceptibles with infectious individuals. The parameter α_3 is called the long term residual value of the force of infection. If $\alpha_3 = 0$, then the force of infection decreases to 0 as a tends to infinity. Integrating $\ell(a)$ results in a nonlinear model

$$\pi(a) = 1 - \exp\left\{\frac{\alpha_1}{\alpha_2} a e^{-\alpha_2 a} + \frac{1}{\alpha_2}\left[\frac{\alpha_1}{\alpha_2} - \alpha_3\right]\left[e^{-\alpha_2 a} - 1\right] - \alpha_3 a\right\}. \tag{6.2}$$

In what follows we refer to (6.2) as the exponential model. The average age at infection, the mean of the distribution of the age of infection, is given by $A = \int_0^L (1 - \pi(x))dx$, where L is life expectancy. Following Farrington (1990) we assume that $L = 75$. In case that the data are observed up to a certain age U, $U \leq L$, the average age at infection is given by

$$A = \int_0^U (1 - \pi(x))dx + f(L - U). \tag{6.3}$$

Here, f is the fraction of individuals that remain uninfected which can be estimated from the data by $f = 1 - \pi(U)$. Farrington (1990) estimated unrestricted models for measles, mumps and rubella based on (6.2) and performed sensitivity analysis for f by estimating the model in (6.2) conditional on several values for f. In these analyses the parameter α_1 is no longer a free parameter but can be calculated conditional on the values of α_2, α_3 and f.

In the present study we use hierarchical nonlinear models to estimate the parameters in the exponential model (6.2). Independent binomial distributions are assumed for the number of infected individuals at age a_i

$$y_i \sim \text{Bin}(n_i, \pi_i), \quad \text{for } \; i = 1, 2, \ldots, n, \tag{6.4}$$

where n_i is the sample size at age a_i. The constraints on the parameter space can be incorporated in the hierarchical model by assuming truncated normal distributions for the components of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$, in $\pi_i = \pi(a_i, \boldsymbol{\alpha})$,

$$\alpha_j \sim \text{truncated } N(\mu_j, \tau_j) \quad j = 1, 2, 3.$$

Here, the normal prior distribution is left truncated at 0 to ensure that $\ell(a) \geq 0$. The joint posterior distribution for $\boldsymbol{\alpha}$ can be derived by combining the likelihood and the prior model as

$$P(\boldsymbol{\alpha}|\mathbf{y}) \propto \prod_{i=1}^{n} \text{Bin}(y_i|n_i, \pi(a_i, \boldsymbol{\alpha})) \prod_{i=1}^{3} \frac{1}{\tau_j} \exp\left(\frac{1}{2\tau_j^2}(\alpha_j - \mu_j)^2\right). \tag{6.5}$$

The full conditional distribution of α_i, derived from (6.5), is given by

$$P(\alpha_i|\alpha_j, \alpha_k, k, j \neq i) \propto \frac{1}{\tau_i} \exp\left(\frac{1}{2\tau_i^2}(\alpha_i - \mu_i)^2\right) \prod_{i=1}^{n} \text{Bin}(y_i|n_i, \pi(a_i, \boldsymbol{\alpha})), \tag{6.6}$$

which cannot be simplified further. To complete the specification of the probability model we assume flat hyperprior distributions at the third level of the model, i.e. $\mu_j \sim N(0, 10000)$ and $\tau_j^{-2} \sim \text{gamma}(1000, 1000)$.

6.3.2 Hierarchical Log-logistic Model

The exploratory analysis from Section 6.2 indicates that the force of infection rises to a peak and drops down thereafter. Therefore we can conclude that the time spent in the susceptible class is not an outcome of neither an exponential nor a weibull distribution since these distributions have a constant and a monotone force of infection respectively. In contrast, the log-logistic distribution offers a wide range of curve shapes for the hazard function, which is more capable to capture the common pattern revealed in Figure 6.1 (although, similar to model (6.2), the secondary peaks will be smoothed out). Under the assumption that the time spent in the susceptible class follows a log-logistic distribution, the probability to become infected before age a is given by

$$\pi(a) = \frac{\beta a^{\alpha}}{1 + \beta a^{\alpha}}, \quad \alpha, \beta > 0, \tag{6.7}$$

and the force of infection by

$$\ell(a) = \frac{\alpha\beta a^{\alpha-1}}{1 + \beta a^{\alpha}}. \tag{6.8}$$

The log-logistic model can be fitted as a GLM with $\log(a)$ as a predictor and a logit link function. This leads to a Bayesian logistic regression model (Gilks *et al.* 1996 and Gelman *et al.* 1996) of y with covariate $\log(a)$. We specify the same likelihood as in (6.4) with linear predictor given by

$$\text{logit}(\pi(a)) = \alpha_2 + \alpha_1 \log(a),$$

where $\alpha_2 = log(\beta)$. For the prior model of α_1, we specify $\alpha_1 \sim$ truncated $N(\mu_1, \tau_1)$. We constrain β to be positive by specifying $\alpha_2 \sim N(\mu_2, \tau_2)$. The full conditional distribution of α_1 is

$$P(\alpha_1|\alpha_2) \propto \frac{1}{\tau_1}\exp\left(\frac{1}{2\tau_1^2}(\alpha_1 - \mu_1)^2\right) \prod_{j=1}^{n} \text{Bin}(y_i|n_i, \pi(a_i, \alpha_1, \alpha_2)). \tag{6.9}$$

The full conditional distribution for α_2 can be derived in the same way. The same flat hyperpriors distributions as in the previous section are assumed for the hyperparameters.

6.3.3 Model Selection

Within the Bayesian framework, the unknown parameters are estimated by the posterior mean. However, since the full conditional distributions in (6.6) and (6.9) do not have a closed analytical form, we cannot evaluate it directly. We can approximate it using Markov Chain Monte Carlo (MCMC) methods (Gilks *et al.*, 1996) and generate samples form the full conditional distributions using the Gibbs sampler. The sample averages are taken as the posterior means of the parameters of interest.
A model selection procedure is needed in order to compare between the models mentioned above and to select the best model. Goodness-to-fit and complexity of the models were assessed using the deviance information criterion (DIC) as proposed by Spiegelhalter *et al.* (1998, 2002) and recently used by Erkanli *et al.* (1999), Rahmann *et al.* (1999) and Gelfand *et al.* (2000) for model selection within the Bayesian framework. Spiegelhalter *et al.* (1998, 2002) suggested to measure the effective number of parameters (the complexity)

Table 6.1: *Deviance and goodness-to-fit summaries for the parametric Bayesian models.*

	Rubella				Mumps			
Model	$\bar{D}$	$D(\pi)$	P_D	DIC	$\bar{D}$	$D(\pi)$	P_D	DIC
Exponential ($\alpha_3 = 0$)	64.70	62.70	2.02	66.72	63.10	61.07	2.021	65.13
Exponential ($\alpha_3 > 0$)	61.13	58.13	3.00	64.13	64.62	62.48	2.14	66.76
Log-logistic	58.99	59.70	2.29	61.28	97.26	95.14	2.12	99.37

in the model by the difference between the posterior expectation of the deviance and the deviance evaluated at the posterior expectation of $\boldsymbol{\pi}$, that is

$$P_D = E_{\boldsymbol{\pi}|\mathbf{y}}(D) - D(E_{\boldsymbol{\pi}|\mathbf{y}}(\boldsymbol{\pi})) = \bar{D} - D(\bar{\boldsymbol{\pi}}), \tag{6.10}$$

with deviance given by $D(\boldsymbol{\pi}) = -2\log P(\mathbf{y}|\boldsymbol{\pi}) + 2\log(f(\mathbf{y}))$. The second term in the deviance is a standardizing factor which does not depend on $\boldsymbol{\pi}$; we use -2 log likelihood of the saturated model. Hence, for the models discussed above the binomial deviance is given by

$$D(\boldsymbol{\pi}) = 2\sum_i \left(y_i \log \frac{y_i}{n_i \pi_i} + (y_i - n_i) \log \frac{1 - \frac{y_i}{n_i}}{1 - \pi_i} \right). \tag{6.11}$$

In practice, $D(\boldsymbol{\pi})$ and $\boldsymbol{\pi}$ can be monitored during the MCMC run, $\bar{D}$ is the sample mean of $D(\boldsymbol{\pi})$ while $D(\bar{\boldsymbol{\pi}})$ is the deviance evaluated at the posterior mean. For model selection, Spiegelhalter *et al.* (1998, 2002) suggested to use the *Deviance Information Criterion* (DIC):

$$DIC = \bar{D} + P_D = D(\bar{\boldsymbol{\pi}}) + 2P_D. \tag{6.12}$$

Smaller values of DIC indicate a better fitting model.

6.3.4 Application to the Data

Table 1 presents the deviance summaries of the data and Figure 6.2 shows the fitted models for both the prevalence and the force of infection. Starting with rubella, the first model that was fitted assumes that $\alpha_3 = 0$ in the exponential model in (6.2). For this model the posterior deviance is 64.7 and $P_D = 2.02$, slightly higher than the "true" number of parameters. For the exponential model with $\alpha_3 > 0$, $\bar{D} = 61.13$ and $P_D = 3.00$. The DIC

Table 6.2: *Posterior means for the parameters. Note that the $\boldsymbol{\alpha}$ parameters of the log logistic model and the exponential model are not comparable. The exponential model with $\alpha_3 = 0$ is the model proposed by Farrington (1990) with the assumption that $\ell(a) = \alpha_1 a \exp(-\alpha_2 a)$.*

	Rubella			Mumps		
	Exponential	Exponential	Log	Exponential	Exponential	Log
	$\alpha_3 = 0$	$\alpha_3 > 0$	logistic	$\alpha_3 = 0$	$\alpha_3 > 0$	logistic
α_1	0.067	0.07	1.645	0.139	0.139	2.063
α_2	0.158	0.201	-2.964	0.192	0.198	-2.865
α_3		0.034			0.008	
$f^\star$	0.07	0.044	0.036	0.023	0.021	0.007
A	11.11	10.16	9.86	5.72	5.61	5.05

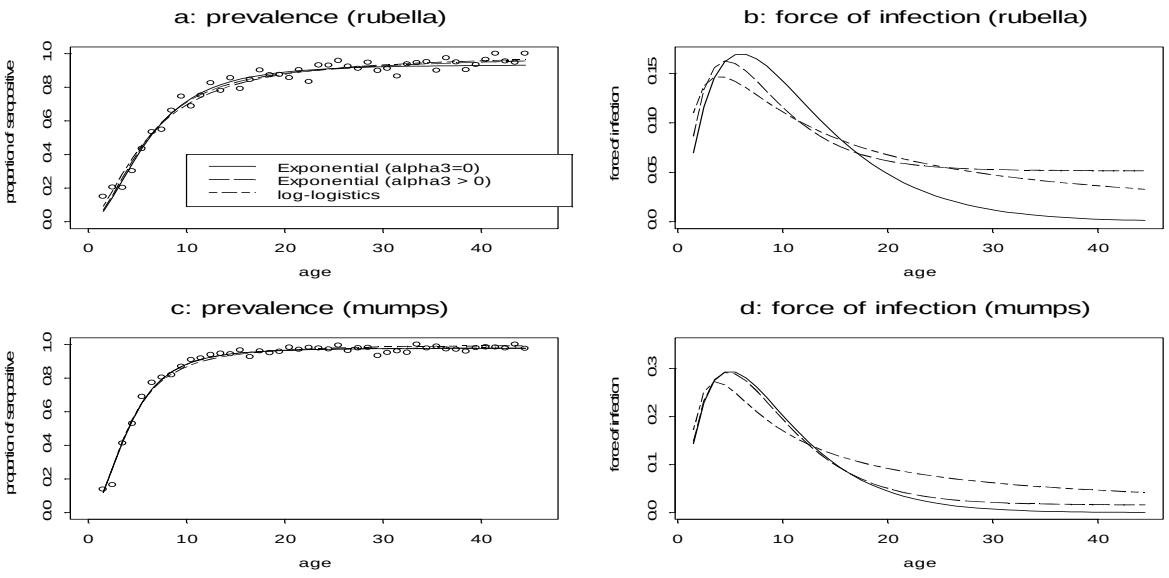


Figure 6.2: *Posterior means for the prevalence (left panels) and the force of infection (right panels). Soild line: exponential model with $\alpha_3 = 0$, long dashed line: exponential model with $\alpha_3 > 0$, dashed line: log-logistic model.*

of this model is 64.13, smaller than the DIC of the first model (66.72) indicating that among the exponential models the second one is to be prefered. However, the log-logistic model with $DIC = 61.28$ has the best goodness-to-fit. For mumps, the model with the lowest value of DIC is the exponential model with $\alpha_3 = 0$ (65.13). Posterior means for the parameters are shown in Table 2. Figure 6.2 shows that there is a substantial difference between the models at the age for which the force of infection reaches its peak and in the level of the force of infection at older age groups. Furthermore, for rubella, the posterior mean of the average age at infection for the exponential model with $\alpha_3 > 0$ is 10.16 and the posterior mean for $f^\star$ is 0.04. When α_3 is not included in the model, the average age of infection increase to 11.11 and $f^\star$ to 0.07. The posterior mean of the average age at infection obtained from the log-logistic model is 9.86. For mumps, the effect of α_3 on $f^\star$ is less substantial, the reason for that is the small value of α_3 that was estimated in the second model (0.008). The smallest value of $f^\star$ is obtained for the log-logistic model (0.0007) with average age at infection equal to 5.053.

6.4 Hierarchical Nonparametric Model

6.4.1 Hierarchical Beta/Binomial Model

In the previous section the prevalence was assumed to have a parametric form $\boldsymbol{\pi}(a, \boldsymbol{\alpha})$, and monotonicity was achieved by constraining the parameter space of $\boldsymbol{\alpha}$. In this section, we assume that $\boldsymbol{\pi}$ is a right-continuous nondecreasing function defined on $[0, \delta]$, $\pi_n \leq \delta \leq 1$, $\delta = 1 - f$. We do not assume any deterministic relationship between π_i and a_i but instead we specify a probabilistic model for π_i at each distinct level of a_i. Since the data are binomial, it is natural to use the product-beta prior (Gelfand and Kuo, 1991) for $\boldsymbol{\pi}$, since it is a conjugate prior for the binomial likelihood and ensures that the posterior distribution of $\boldsymbol{\pi}|\mathbf{y}$ is also beta distribution. A product-beta prior has the form of

$$P_B(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \prod_{i=1}^{n} (\pi_i)^{\alpha_i - 1}(1 - \pi_i)^{\beta_i - 1}, \quad (\alpha_i > 0, \beta_i > 0), \tag{6.13}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$. For the unconstrained case, combining the binomial likelihood and the product-beta prior, leads to the posterior distribution

$$P(\boldsymbol{\pi}|\mathbf{y}, \alpha, \boldsymbol{\beta}) \propto \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{n_i - y_i} \prod_{i=1}^{n} \pi_i^{\alpha_i - 1}(1 - \pi_i)^{\beta_i - 1} = \prod_{i=1}^{n} \pi_i^{y_i + \alpha_i - 1}(1 - \pi_i)^{n_i - y_i + \beta_i - 1}, \tag{6.14}$$

which is Beta$(y_i + \alpha_i, n_i - y_i + \beta_i)$. The problem is to estimate $\boldsymbol{\pi}$ under the order restrictions, $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_n$. Thus, the n dimensional parameter vector is constrained to lie in a subset S^n of R^n. The constrained set S^n is determined by the order among the components of $\boldsymbol{\pi}$. In this case it is natural to incorporate the constraints into the specification of the prior distribution. Gelfand, Smith and Lee (1992) show that the posterior distribution of $\boldsymbol{\pi}$ given the constraints is the unconstrained posterior distribution normalized such that

$$P(\boldsymbol{\pi}|\mathbf{y}) \propto \frac{P(\mathbf{y}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{\int_{S^n} P(\mathbf{y}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\beta})d\boldsymbol{\pi}}, \quad \boldsymbol{\pi} \in S^n. \tag{6.15}$$

Let $S_j^n(\pi_j, j \neq i)$ be a cross section of S^n defined by the constraints for the component π_i at a specified set of π_j, $j \neq i$. In our setting, $S_j^n(\pi_j, j \neq i)$ is the interval $[\pi_{i-1}, \pi_{i+1}]$. It follows from (6.15) that the posterior distribution for π_i is given by

$$\begin{cases} P(\pi_i|\mathbf{y}, \alpha, \beta, \boldsymbol{\pi}_{-i}) \propto P(\mathbf{y}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) & \pi_i \in S_j^n(\pi_j, j \neq i), \\ 0, & \pi_i \notin S_j^n(\pi_j, j \neq i). \end{cases} \tag{6.16}$$

Here, $\boldsymbol{\pi}_{-i} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$. Hence, when the likelihood and the prior distribution are combined, the posterior conditional distribution of $\pi_i|\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}_{-i}$ is the standard posterior distribution restricted to $S_j^n(\pi_j, j \neq i)$, that is $\text{Beta}(y_i+\alpha_i, n_i-y_i+\beta_i)$ restricted to the interval $[\pi_{i-1}, \pi_{i+1}]$ (Gelfand and Kuo, 1991). This means that during the MCMC simulation the sampling from the full conditional distribution can be reduced to interval restricted sampling from the standard posterior distribution (Gelfand, Smith and Lee, 1992).
The hierarchical model we consider is given by

$$\begin{aligned} y_i &\sim \text{Bin}(n_i, \pi_i) && \text{likelihood} \\ \pi_i &\sim \text{Beta}(\alpha_i, \beta_i)I(\pi_{i-1}, \pi_{i+1}) && \text{prior,} \end{aligned} \tag{6.17}$$

where $I(\pi_{i-1}, \pi_{i+1})$ an indicator variable which takes the value of 1 if $\pi_{i-1} \leq \pi_i \leq \pi_{i+1}$ and zero elsewhere. In order to complete the specification of the hierarchical model in (6.17) we need to specify a hyperprior distributions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Note that in the special case that $\alpha_i = \beta_i = 1$ for $i = 1, \ldots, n$ implies that the prior distribution of the prevalence at the i'th age group, condition on π_{i-1} and π_{i+1}, is a uniform distribution over the interval $[\pi_{i-1}, \pi_{i+1}]$, $\pi_i|\pi_{i-1}, \pi_{i+1} \sim \text{Uniform}(\pi_{i-1}, \pi_{i+1})$. However, there is no reason to fix $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be equal to 1, there is no clear way how to choose the hyperprior distribution for the components in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ either. For the analysis presented below we specify non informative distributions for the hyperparameters by specifying a left truncated (at zero) normal distribution with variance equal to 1000 for each one of the components in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ at the third stage of the hierarchical model.

Once the prevalence values are obtained, the problem of estimating the force of infection becomes straightforward. Let $\boldsymbol{\pi}^{(k)}$ be the constrained value of $\boldsymbol{\pi}$, obtained in the k'th iteration of the MCMC simulation. The force of infection $\ell^{(k)}(a)$ can be estimated by $\hat{\ell}^{(k)}(a) = \hat{\pi'}^{(k)}(a)/(1 - \hat{\pi}^{(k)}(a))$. However, since we assume that the force of infection is a smooth function, we smooth $\ell^{(k)}(a)$ with a twice successively third order moving average (Diggle, 1990), i.e. $\ell_S^{(k)}(a) = A\ell^{(k)}(a)$ where $\ell_S(a)$ is the smoothed force of infection and S is the smoothing matrix. The posterior mean of $\ell_S(a)$ is simply $\sum_{k=1}^K \ell_S^{(k)}(a)/K$ where K is the number of MCMC iterations.
The fraction of uninfected individuals can be used in this model to specify the distribution of $\pi(U)$. If our prior assumption is that $f = 0$, then $\pi(U) \sim \text{Beta}(\alpha_n, \beta_n)I(\pi_{n-1}, 1)$, where $I(\pi_{n-1}, 1)$, is an indicator that takes the value of 1 if $\pi_{n-1} \leq \pi_n \leq 1$ and 0 otherwise. In case that we use the prior knowledge that $f > 0$, say $f = f^\star$, then we can truncate the distribution of $\pi(U)$ at the right side with $1 - f^\star$, $\pi(U) \sim \text{Beta}(\alpha_n, \beta_n)I(\pi_{U-1}, 1 - f^\star)$.

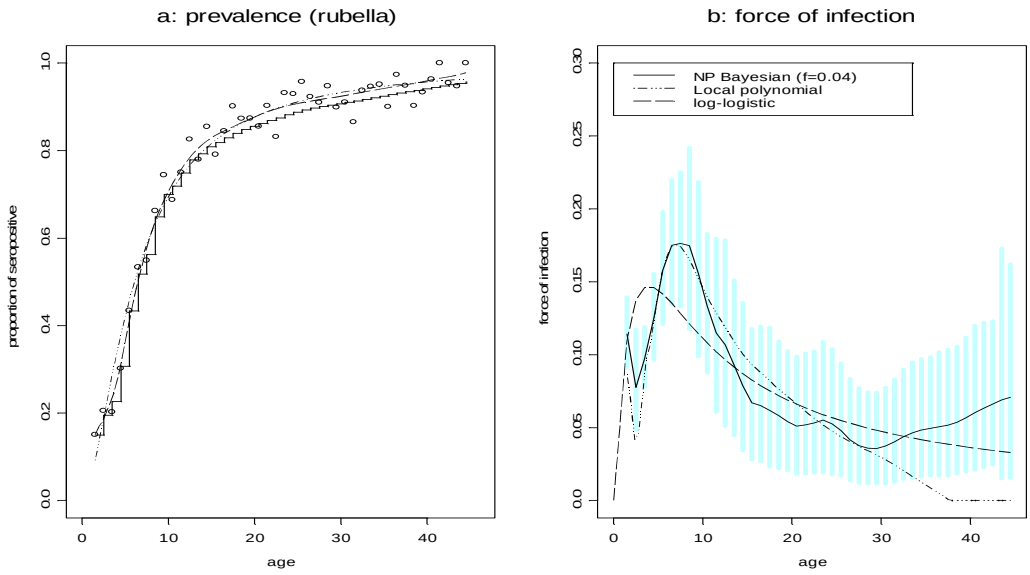


Figure 6.3: *Posterior means for the prevalence and the force of infection (rubella). The gray area in panel b represents the* 95% *credible intervals for the force of infection. NP Bayesian denotes the Beta-binomial model. The parametric model is the log logistic model which, among the parametric models, has the smallest value of DIC.*

6.4.2 Application to the Data

The posterior means for the prevalence and force of infection are shown in Figure 6.3 (rubella) and Figure 6.4 (mumps). For rubella, the nonparametric models indicate essentially the same patterns, although the secondary peak at age 23 is less substantial in the beta-binomial model. In addition, from age 30 onwards, the beta-binomial model predicts a higher force of infection. For mumps, the secondary peak at age 20 was smoothed out by the beta-binomial model. Similar to rubella, the beta-binomial model predicts higher values for the force of infection at the first peak, compared to the parametric model. This can be seen in Figure 6.5 which presents the density estimates for the posterior distribution of the force of infection between age 3.5 and 6.5. Note that the exponential and the beta-binomial models for the force of infection reach a peak at age 4.5 and 5.5 respectively. The beta-binomial model predicts higher values for the force of infection at the ages, 0.36 and 0.29 for the beta-binomial and the exponential models respectively.
The value of f has a substantial influence on the posterior mean of the average age at infection. We fitted the beta-binomial model with several values of f. That is, we truncated the distribution of $\pi(U)$ at the right hand side with $1-f$, $\pi(U) \sim \text{Beta}(\alpha_n, \beta_n) I(\pi_{U-1}, 1-f)$. Table 6.3 presents the results and shows that the posterior mean of A increases with f. This can be seen in Figure 6.7 which shows the 95% credible intervals for the average age at infection. This pattern was observed by Farrington (1990) for the estimated conditional models (see Farrington (1990), Table 3). Note that in the second column in Table 6.3, $\bar{f} = 1 - \bar{\pi}(U)$, is the posterior mean for f. Figure 6.6 shows the estimated forces of infection for several values of f. Note that substantial differences are observed from age 30 and onwards. The force of infection increases with higher values of f.

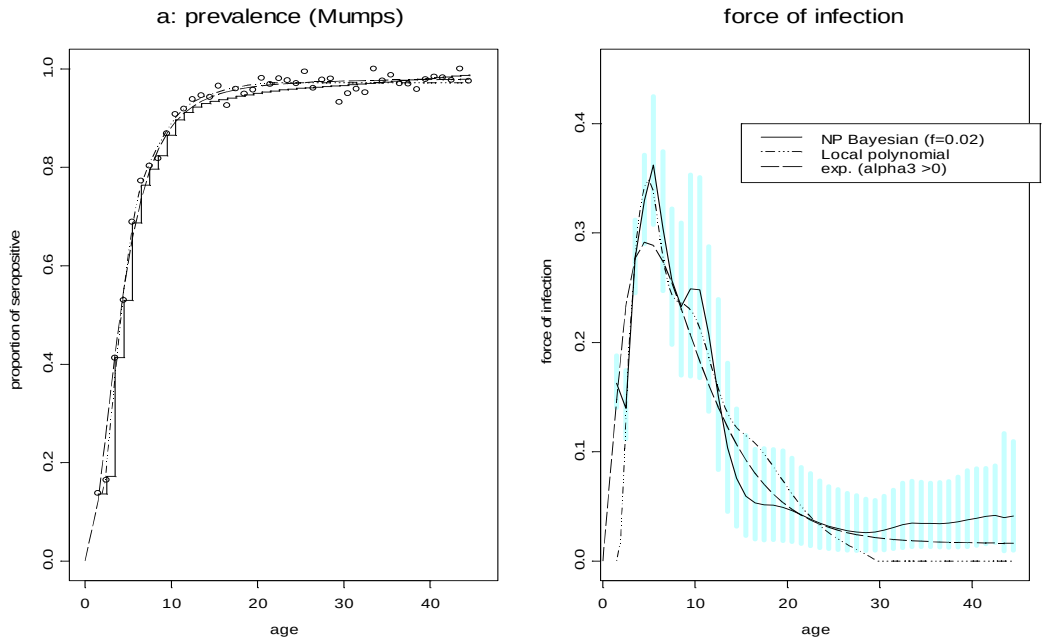


Figure 6.4: *Posterior means for the prevalence and the force of infection(mumps). The gray area in panel b represents the* 95% *credibal intervals for the force of infection. NP Bayesian is the Beta-binomial model. The parametric model is the exponential model with* $\alpha = 0$ *which, among the parametric models, has the smallest value of DIC.*

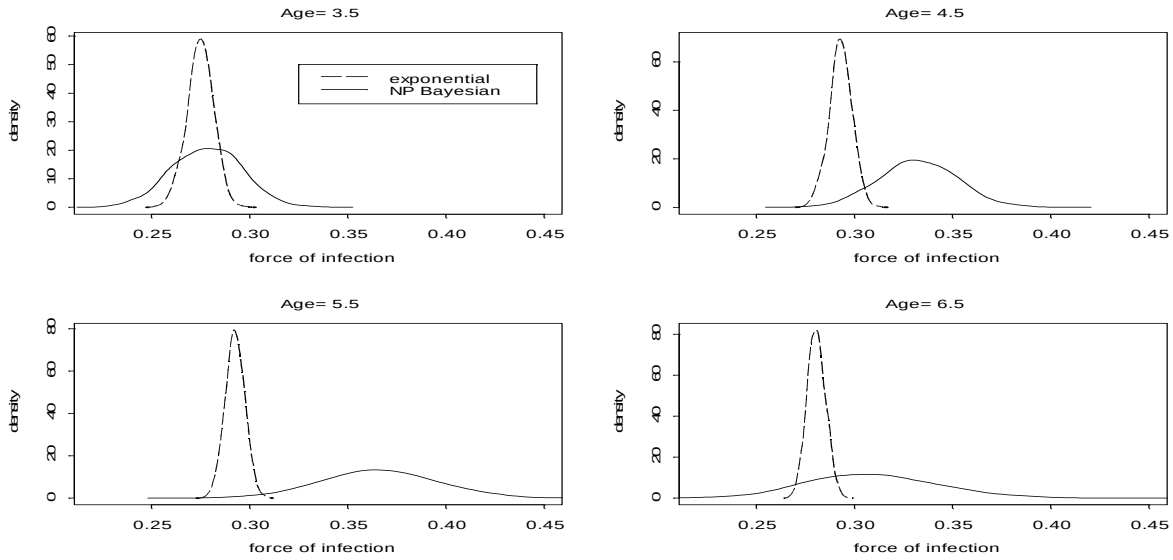


Figure 6.5: *Kernel estimates for the posterior distribution of the force of infection for mumps at ages 3.5 -6.5.*

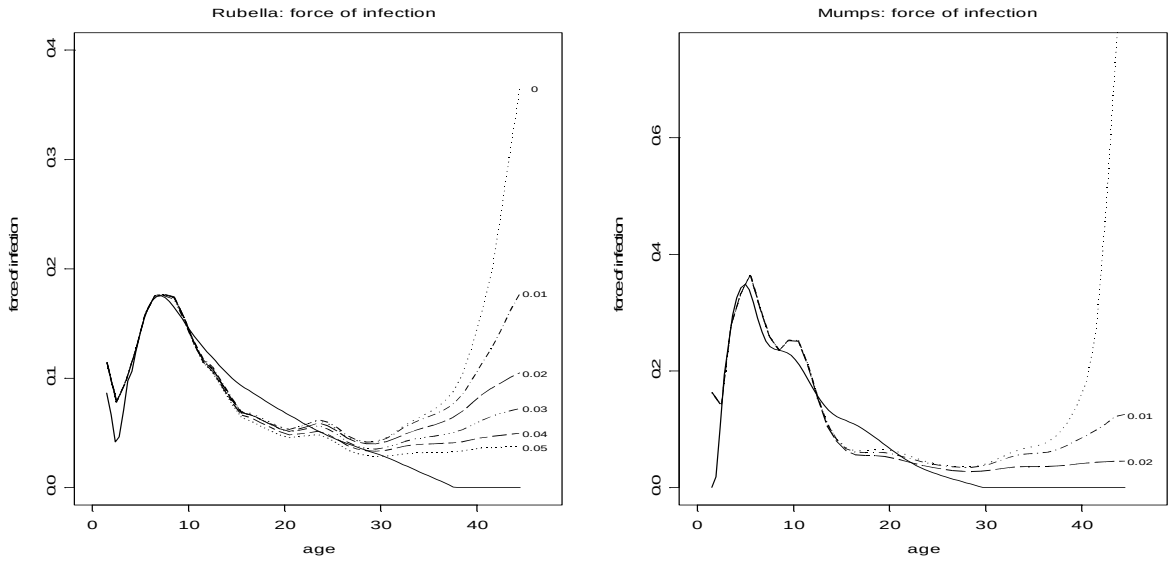


Figure 6.6: *Posterior mean for the force of infection for several values of f. Left Panel: rubella, right panel : mumps. The solid line is the force of infection estimated by local quadratic model. The numbers to the right are the values of f that were used to right truncate the prior distribution of $\pi(U)$.*

6.5 Discussion

The age-dependent force of infection is a basic concept in any epidemiological model for infectious disease. Furthermore, the average age at infection and the basic reproduction number, R_0, depend on the model for the force of infection. In this study we model the prevalence within the framework of hierarchical Bayesian models in order to investigate the posterior distribution of $\ell(x)$ and A. The parametric models are restrictive since they can estimate only a single peak model for the force of infection. However, the beta-binomial model suggests secondary peaks which may be important form an epidemiological point of view. Furthermore, we have shown that compared to the parametric models, the beta-binomial models predict higher values for the force of infection at its maximum.
The problem of estimation under order restrictions was addressed by applying the PAV algorithm to the local polynomial and by choosing a truncated product-Beta for the prior model in the hierarchical beta-binomial model. Both models estimate a nondecreasing prevalence and therefore lead to a nonnegative force of infection, as required. The beta-binomial model is highly sensitive for the values of f. It is necessary to fit the model with several values of f in order to investigate its influence on the posterior mean of A and $\ell(x)$. We specified a product-beta as a prior distribution for the hierarchical nonparametric model. An order Dirichlet distribution as discussed, in the context of binary response, by Ramsey (1972), Gelfand and Kuo (1991) and Qian *et al.* (2000) can be used as well. This issue is discussed in the next chapter.

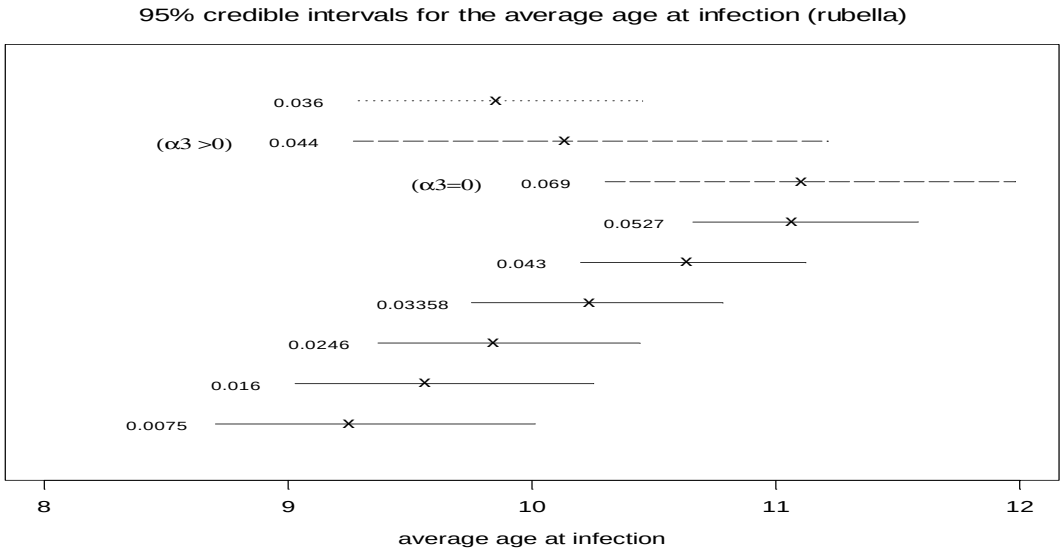


Figure 6.7: *95% credible intervals for the average age at infection. Solid lines: Beta-binomial models, dotted line: log logistic model, long dashed line: exponential model. The numbers to the left of the credible intervals are the posterior means for* f, $1 - \bar{\pi}(U)$.

Table 6.3: *Posterior mean for the average age at infection and* f *obtained from the beta-binomial models.*

Rubella	f	$\bar{f}$	$\bar{A}$
	0.00	0.007	9.27
	0.01	0.016	9.58
	0.02	0.025	9.85
	0.03	0.033	10.24
	0.04	0.043	10.63
	0.05	0.053	11.08
	0.06	0.063	11.55
Mumps	0.00	0.004	5.29
	0.01	0.012	5.58
	0.02	0.021	5.99

Chapter 7

Hierarchical Models with Dirichlet Prior for the Prevalence

7.1 Introduction

In the previous chapter the hierarchical beta-binomial model was used in order to estimate the prevalence and the force of infection. Monotonicity was achieved by incorporating the order constraints into the specification of the prior distribution of $\boldsymbol{\pi}$. We have shown that if $P_B(\boldsymbol{\pi})$ is a product-beta distribution, and the likelihood $P(\mathbf{y}|\boldsymbol{\pi})$ is binomial, then the posterior distribution , $P(\pi_i|\mathbf{y}, \boldsymbol{\pi}_{-i})$, $i = 1, \ldots, n$, is a beta distribution restricted to the interval $[\pi_{i-1}, \pi_{i+1}]$. However, even though, $P_B(\boldsymbol{\pi})$ is conjugate prior to the binomial likelihood, there is no clear way to select the parameters α_i and β_i. In the previous chapter we overcome this problem by specifying a non informative hyperprior distribution for each pair of parameters (α_i, β_i) at the third stage of the hierarchical model. In this chapter we will use a different class of priors for $\boldsymbol{\pi}$, the Dirichlet process prior (Ferguson, 1973). It will be shown that prior information about the prevalence and the force of infection, for example the models proposed by Farrington (1990) and/or the log-logistic model discussed in the previous chapter, can be incorporated naturally into the Dirichlet prior. Discussion and the specification of the Dirichlet prior are given in Section 7.2. The Dirichlet prior is not conjugate for the binomial likelihood and the posterior distribution does not have a closed form. However, we will show in Section 7.3 that the posterior distribution can be estimated using the Gibbs sampler. In Section 7.4 the proposed method will be applied to the rubella and the mumps datasets.

7.2 Dirichlet Process Prior

7.2.1 Definition and Properties of the Dirichlet Distribution

The Dirichlet distribution (Wilks 1962, Ferguson 1973) was proposed by Ramsey (1972) as a prior distribution for $\boldsymbol{\pi}$ in the context of bioassay modeling. Ferguson (1973) defined the Dirichlet process in a more general setting. Disch (1981), Gelfand and Kuo (1991) and

Kuo (1988) discussed the Dirichlet process in the context of bioassay modeling as well. In particular, Gelfand and Kuo (1991) proposed to use a Gibbs sampling procedure (Gilks *et al.* 1996) in order to estimate the posterior distribution. Qian, Lavine and Stow (2000) discussed the use of Dirichlet prior for binary regression when the posterior distribution is estimated using a Gibbs sampler. In this section we define the Dirichlet distribution and discuss several properties of the distribution that will be used in later sections. All proofs for the properties in this section can be found in Wilks (1962).

Definition

Let $(z_1, \ldots, z_k)$ be a k dimensional random variable. The variable $(z_1, \ldots, z_k)$, $z_i \geq 0$, $i = 1, \ldots, k$, is defined to have a k-variate Dirichlet distribution if the joint density of $(z_1, \ldots, z_k)$ is given by

$$f(z_1, \ldots, z_k) = \frac{\Gamma(\Sigma_{i=1}^{k+1}\gamma_i)}{\prod_{i=1}^{k+1}\Gamma(\gamma_i)} z_1^{\gamma_1 - 1} z_2^{\gamma_2 - 1} \ldots z_k^{\gamma_k - 1}(1 - \Sigma_{i=1}^{k} z_i)^{\gamma_{k+1} - 1}. \tag{7.1}$$

We use the notation $(z_1, \ldots, z_k) \sim D(\gamma_1, \ldots, \gamma_k; \gamma_{k+1})$ to denote a k-variate Dirichlet distribution. Note that for $k = 1$ the joint distribution (7.1) reduces to a beta distribution, $z \sim \text{Beta}(\gamma_1, \gamma_2)$. Indeed, the Dirichlet distribution is treated as the k-variate analogue of the beta distribution (or as the multivariate beta distribution). In the following sections we will show that the parameters $(\gamma_1, \ldots, \gamma_k; \gamma_{k+1})$ can be defined using prior knowledge about the prevalence and the force of infection.

Property 1: Relationship between the gamma and the Dirichlet distributions

Suppose that $(z_1, \ldots, z_{k+1})$ are random variables having standard gamma distribution, $z_i \sim \text{Gamma}(\gamma_i, 1), i = 1 \ldots, k+1$. Let

$$t_i = \frac{z_i}{\Sigma_{j=1}^{k+1} z_j}, \quad i = 1, \ldots, k,$$

then the random variable $(t_1, \ldots, t_k)$ has a k-variate Dirichlet distribution, $(t_1, \ldots, t_k) \sim D(\gamma_1, \ldots, \gamma_k; \gamma_{k+1})$. Note that $0 \leq t_1 \leq \ldots \leq t_k \leq 1$.

Property 2: Relationship between the gamma and the beta distributions

Suppose that (z_1, z_2) are independent random variables having standard gamma distribution, $z_i \sim \text{gamma}(\gamma_i, 1), i = 1, 2$, Then the random variable

$$u = \frac{z_1}{z_1 + z_2}$$

has a beta distribution, $u \sim \text{Beta}(\gamma_1, \gamma_2)$.

Property 3: The ordered Dirichlet distribution

Let $(z_1, \ldots, z_{k+1})$ be a random variable having a k-variate Dirichlet distribution, $(z_1, \ldots, z_{k+1}) \sim D(\gamma_1, \ldots, \gamma_k; \gamma_{k+1})$. Let $(t_1, \ldots, t_k)$ be a random variable for which

$$t_i = \Sigma_{j=1}^{i} z_j, \quad i = 1, \ldots, k.$$

The joint density of $(t_1, \ldots, t_k)$ is given by

$$f(t_1, \ldots, t_k) = \frac{\Gamma(\Sigma_{i=1}^{k+1}\gamma_i)}{\prod_{i=1}^{k+1}\Gamma(\gamma_i)} t_1^{\gamma_1 - 1}(t_2 - t_1)^{\gamma_2 - 1} \ldots (t_k - t_{k-1})^{\gamma_k - 1}(1 - t_k)^{\gamma_{k+1} - 1}. \tag{7.2}$$

Note that the t_i satisfy $0 \leq t_1 \leq \ldots \leq t_k$. We refer to the distribution in (7.2) as the ordered k-variate Dirichlet distribution. Note that for $k = 1$ the distribution of t_1 is Beta(γ_1, γ_2).

7.2.2 Specification of the Dirichlet Prior

Similar to the previous chapter we assume that the prevalence at age a_i has the form of

$$\pi_i = \pi(a_i), \qquad i = 1, \ldots, n. \tag{7.3}$$

Here, π_i is considered to be an isotonic nonparametric function, $0 \leq \pi(a) \leq 1$. The aim is to estimate the joint posterior distribution, $P(\pi_1, \ldots, \pi_n|\mathbf{y})$, $\mathbf{y} = (y_1, \ldots, y_n)$, where y_i is the number of infected individuals at the i'th age group.
Let $s_i = \pi_i - \pi_{i-1}$ for $i = 1, \ldots, n+1$ and $\pi_0 = 0$ and $\pi_{n+1} = 1$. Note that since $\boldsymbol{\pi}$ is bounded between 0 and 1, the transformation from $(\pi_0, \pi_1 \ldots, \pi_n, \pi_{n+1})$ to $(s_1, \ldots, s_n)$ is a one to one transformation. Furthermore, since $s_i \in [0, 1]$ and $\sum_{i=1}^{n+1} s_i = 1$ one can specify a Dirichlet for the joint distribution of $\boldsymbol{s}$, $\boldsymbol{s} = (s_1, \ldots, s_{n+1})$.
Ferguson (1973) established his discussion on the Dirichlet distribution by assuming that $z_1, \ldots, z_k$ are independent random variables from a gamma distribution, $z_i \sim$ Gamma$(\gamma_i, 1)$, $i = 1, \ldots, k$. The Dirichlet distribution with parameters $(\gamma_1, \ldots, \gamma_k; \gamma_{k+1})$ was defined as the distribution of $(t_1, \ldots, t_k)$ where $t_i = z_i / \sum_{j=1}^{k} z_j$, $i = 1, \ldots, k$. In our setting, since $\sum_{i=1}^{n+1} s_i = 1$, if we assume that $s_i \sim$ Gamma$(\gamma_i, 1)$ it follows, from property 1, that $\boldsymbol{s} \sim D(\gamma_1, \ldots, \gamma_n; \gamma_{n+1})$. Note that since $\boldsymbol{s}$ has a Dirichlet distribution, it follows from property 3 that $\boldsymbol{\pi}$ has an ordered Dirichlet distribution since $\pi_i = \sum_{j=1}^{i} s_j$. Furthermore, since $s_i \sim$ Gamma$(\alpha_i, 1)$ and $s_{i+1} \sim$ Gamma$(\alpha_{i+1}, 1)$ it follows from property 2 that

$$\frac{\pi_i - \pi_{i-1}}{\pi_{i+1} - \pi_{i-1}} \sim \text{Beta}(\gamma_i, \gamma_{i+1}). \tag{7.4}$$

The beta distribution in (7.4) is defined on the interval $[0, 1]$. Let $\delta_i = (\pi_i - \pi_{i-1})/(\pi_{i+1} - \pi_{i-1})$. Indeed, δ_i has a beta distribution over the interval (0,1). Ramsey (1972) showed that if $z = a + (b - a)y$, where a and b are known constants and y is a random variable which has a beta distribution defined over the interval $[0, 1]$, $y \sim$ Beta(α, β), then the distribution of z is Beta(α, β) over the interval $[a, b]$. Since $\pi_i = \pi_{i-1} + \delta_i(\pi_{i+1} - \pi_{i-1})$ and $\delta_i \sim$ Beta(γ_i, γ_{i+1}) it follows that, given π_{i-1} and π_{i+1}, $\pi_i|\pi_{i-1}, \pi_{i+1} \sim$ Beta(γ_i, γ_{i+1}) over the interval $[\pi_{i-1}, \pi_{i+1}]$. This is consistent with the conditional distribution of π_i discussed in Chapter 6.

7.2.3 The Choice of the Prior Mean for $\boldsymbol{\pi}$

Specification of the Dirichlet prior distribution requires to specify the parameters $(\gamma_1, \ldots, \gamma_n; \gamma_{n+1})$. Since $\boldsymbol{\pi}$ is an ordered Dirichlet random variable, the marginal distribution of π_i is a beta distribution,

$$\pi_i \sim \text{Beta}(A_i, \sum_{j=1}^{n+1} \gamma_j - A_i),$$

where $A_i = \sum_{j=1}^{i} \gamma_j$. Now, suppose that we have a prior guess for $\boldsymbol{\pi}$, $\boldsymbol{F}_0 = (F_{0,1}, \ldots, F_{0,n+1})$, then selecting $F_{0,0} = 0$, $F_{0,n+1} = 1$ and

$$\begin{aligned}
\gamma_1 &= MF_{0,1} \\
\gamma_i &= M(F_{0,i} - F_{0,i-1}) \\
&\vdots \\
\gamma_n &= M(F_{n,1} - F_{n-1,1}) \\
\gamma_{n+1} &= M(1 - F_{n,1})
\end{aligned}$$

results in a marginal distribution for π_i

$$\pi_i \sim \text{Beta}(MF_{0,i}, M - MF_{0,i}),$$

with $E(\pi_i) = F_{0,i}$ and $\text{var}(\pi_i) = F_{0,i}(1 - F_{0,i})/M + 1$. The constant M is a precision parameter which represents the uncertainty about the prior guess of $\boldsymbol{\pi}$. Large value of M reflects "high confidence" in the prior mean for $\boldsymbol{\pi}$. It follows that the distribution of $(s_1, \ldots, s_{n+1})$ is a Dirichlet distribution with the parameters $(\gamma_1, \ldots, \gamma_n; \gamma_{n+1})$,

$$P_D(\boldsymbol{s}) = \frac{\Gamma(\Sigma_{i=1}^{n+1}\gamma_i)}{\prod_{i=1}^{n+1}\Gamma(\gamma_i)} \prod_{i=1}^{n+1} s_i^{\gamma_i - 1}, \tag{7.5}$$

where $s_1 = \pi_1 - 0$ and $s_{n+1} = 1 - \pi_n$. Since $\pi_i = \sum_{j=1}^{i} s_j$ the prior distribution in (7.5) is an ordered Dirichlet distribution. In terms of the prevalence the prior model for $\boldsymbol{\pi}$ is

$$P_D(\boldsymbol{\pi}) = \frac{\Gamma(\Sigma_{i=1}^{n+1}\gamma_i)}{\prod_{i=1}^{n+1}\Gamma(\gamma_i)} \pi_1^{\gamma_1 - 1}(\pi_2 - \pi_1)^{\gamma_2 - 1} \ldots (\pi_n - \pi_{n-1})^{\gamma_n - 1}(1 - \pi_n)^{\gamma_{n+1} - 1}. \tag{7.6}$$

The conditional distribution of π_i, given π_{i-1} and π_{i+1}, is specified in equation (7.4). Note that the main difference between the product-beta to the Dirichlet priors is that the product-beta prior requires to generate a value from $\text{Beta}(\alpha_i, \beta_i)$, where α_i and β_i have a noninformative hyperprior distribution while the Dirichlet prior requires to generate a value from $\text{Beta}(\gamma_i, \gamma_{i+1})$ where γ_i and γ_{i+1} are determined by the prior mean of $\boldsymbol{\pi}$. Similar to Chapter 6, we focus on the conditional distribution, $P(\pi_i|\boldsymbol{\pi}_{-i}, \mathbf{y})$. We have shown in Chapter 6 that if $P(\boldsymbol{\pi})$ is a product-beta, the conditional posterior distribution is truncated beta. With Dirichlet prior, the conditional posterior distribution takes the general from

$$P(\pi_i|\boldsymbol{\pi}_{-i}, \mathbf{y}) \propto P(\mathbf{y}|\pi_i, \boldsymbol{\pi}_{-i}) \times P(\pi_i|\boldsymbol{\pi}_{-i}) = P(y_i|\pi_i) \times P(\pi_i|\pi_{i-1}, \pi_{i+1}). \tag{7.7}$$

The first term in the right hand side in equation (7.7) is the binomial likelihood,

$$P(y_i|\pi_i) = \prod_{j=1}^{n_i} \pi_i^{y_{ij}}(1-\pi_i)^{1-y_{ij}} = \pi_i^{y_i}(1-\pi_i)^{n_i-y_i}. \tag{7.8}$$

Here, n_i is the sample size at age group i, y_{ij} is an indicator variable which takes the value of 1 if individual j is infected and 0 otherwise, $y_i = \sum_{j=1}^{n_i} y_{ij}$. Thus, (7.8) can be seen as the contribution of the i'th age group to the likelihood. The second term is the conditional beta distribution as specified in (7.4).
In contrast with the product-beta prior in Chapter 6, the Dirichlet prior in not conjugate to the likelihood in (7.8). Thus, the posterior distribution in the left hand side in (7.7) is not a beta distribution. According to (7.4) the distribution of $\delta_i|\pi_{i-1}, \pi_{i+1}$ is Beta(γ_i, γ_{i+1}). The posterior distribution can be estimated by using a Gibbs sampler, this procedure will be discussed in Section 7.3

7.2.4 The Choice of $\boldsymbol{F}_0$ and M

The specification of the Dirichlet prior requires to specify a prior curve for the prevalence, $\boldsymbol{F}_0$. This prior distribution of age at infection reflects our knowledge (or expectations) about the prevalence across the age groups. For rubella and mumps we assume that $\boldsymbol{F}_0$ is the log-logistic model and the exponential model respectively. Specifically, we assume

$$\text{rubella:} \quad \boldsymbol{F}_0(a) = \tfrac{1}{1+0.051a^{1.649}},$$

$$\text{mumps:} \quad \boldsymbol{F}_0(a) = 1 - \exp\left\{\tfrac{0.139}{0.192}ae^{-0.192a} + \tfrac{0.139}{0.192^2}\left[e^{-0.192a} - 1\right]\right\}.$$

The prior models (with the prior force of infection) are shown Figure 7.1. These models were chosen as the models with the best goodness-of-fit in the previous chapter. The value of M reflects our confidence in the prior model. For given $\boldsymbol{F}_0$ and M the distribution of $\delta_i|\pi_{i-1}, \pi_{i+1}$ is fully specified. Figure 7.2 shows the prior distributions of $\delta_5, \delta_{10}, \delta_{15}$ and δ_{20} for several values of M. Clearly, as the value of M increase, the prior distribution of δ_i become "more" informative. This issue will be discussed once again in Section 7.5.

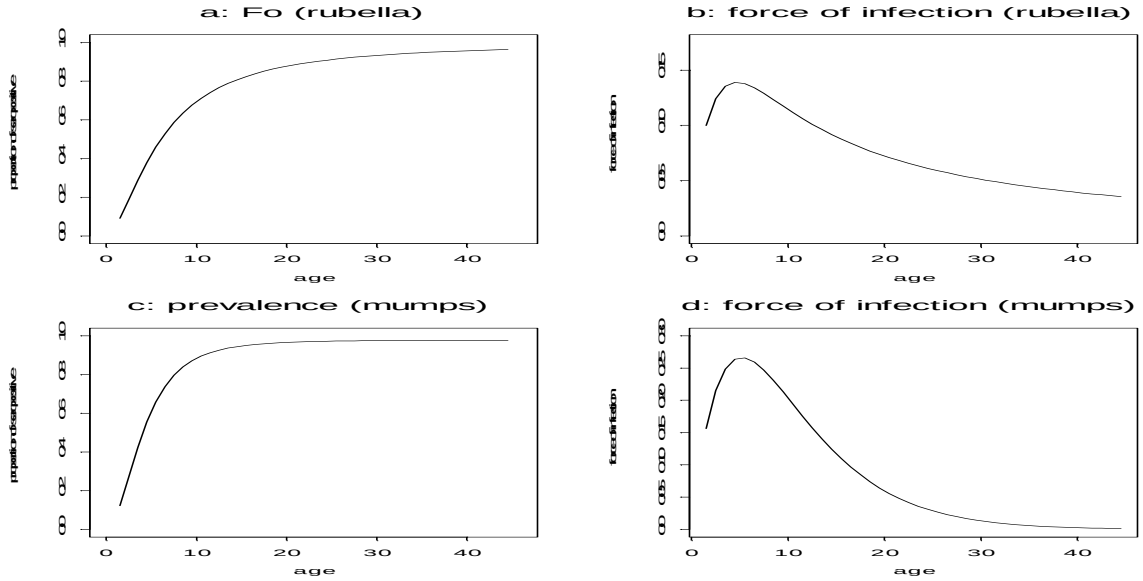


Figure 7.1: *Prior models for the prevalence and the force of infection. Upper row: rubella, lower row: mumps. The model for rubella is a log-logistic model and the model for mumps is Farrington's model with* $\alpha_3 = 0$.

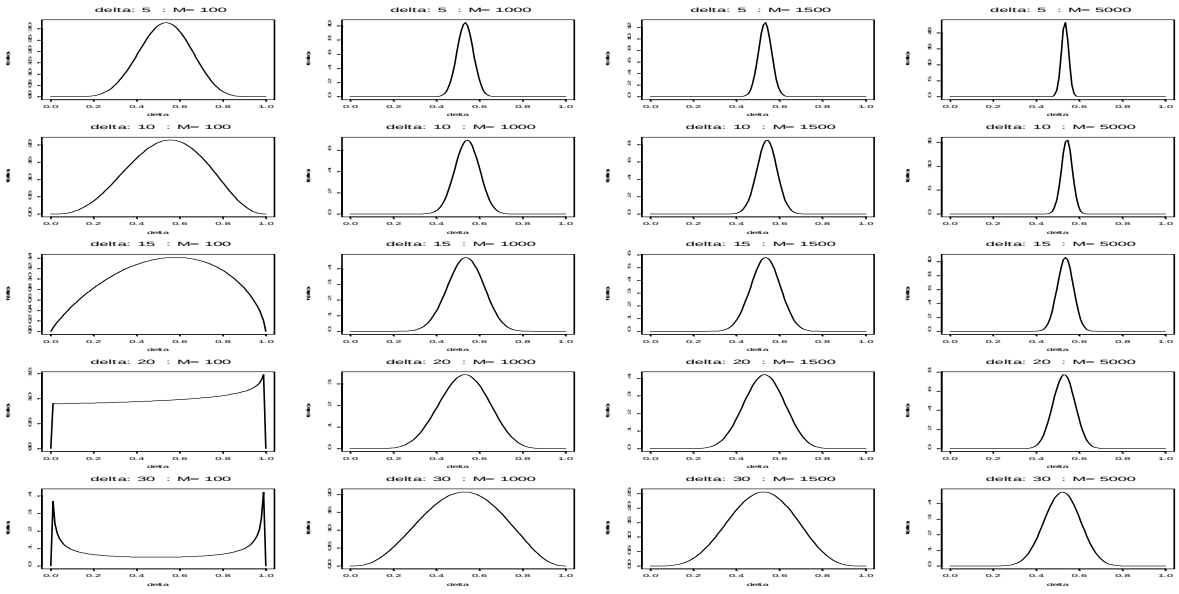


Figure 7.2: *Prior models for* $\delta_5, \delta_{10}, \delta_{15}, \delta_{20}$ *and* δ_{30} *for several values of M.*

7.3 Estimating the Prevalence and the Force of Infection Using the Gibbs Sampler

7.3.1 The Metropolis-Hastings Algorithm

When using ordered Dirichlet prior for $\boldsymbol{\pi}$ the posterior distribution $P(\boldsymbol{\pi}|\mathbf{y})$ cannot be solved analytically. However, we can approximate the posterior distribution using MCMC methods. The Metropolis-Hastings algorithm (Gilks *et al.* 1996) consists of two main loops. An outer loop that runs over the iteration of the chain and an inner loop that runs over the components of $\boldsymbol{\pi}$. At each iteration of the Metropolis-Hastings algorithm we update all components of $\boldsymbol{\pi}$ and proceed for the next iteration of the algorithm. Description of the Metropolis-Hastings algorithm is given below.

- **Step 1**
  Set initial values for $\boldsymbol{\pi}$, $\boldsymbol{\pi}^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \ldots, \pi_n^{(0)}, \pi_{n+1}^{(0)})$, and set the iteration counter of the chain, $\ell = 1$. Note that $\pi_0^{(0)} = 0$ and $\pi_{n+1}^{(0)} = 1$ are constants and do not change from iteration to iteration.

- **step 2**
  Draw a value of δ_i form the beta distribution

$$\delta_i^{(\ell)} \sim \text{Beta}(\gamma_i, \gamma_{i+1}).$$

  - ⋆ **step 2.a**
    Update the value of $\pi_i^{(\ell)}$ by

$$\pi_i^{(\ell)} = \pi_{i-1}^{(\ell)} + \delta_i \times (\pi_{i+1}^{(\ell-1)} - \pi_{i-1}^{(\ell)}).$$

  - ⋆ **step 2.b**
    Calculate the acceptance probability by

$$\alpha(\pi_i^{(\ell)}, \pi_i^{(\ell-1)}) = \min\left\{1, \frac{L(\pi_i^{(\ell)}, y_i)}{L(\pi_i^{(\ell-1)}, y_i)}\right\}.$$

    Here, $L(\pi_i^{(p)}, y_i)$ is the likelihood evaluated at $\pi_i^{(p)}$.

  - ⋆ **step 2.c**
    Draw a random variable u_i from $U(0, 1)$,

$$\text{if} \quad u_i \leq \alpha(\pi_i^{(\ell)}, \pi_i^{(\ell-1)}) \quad \text{accept } \pi_i^{(\ell)},$$

$$\text{if} \quad u_i > \alpha(\pi_i^{(\ell)}, \pi_i^{(\ell-1)}) \quad \text{reject } \pi_i^{(\ell)} \text{ and set } \pi_i^{(\ell)} = \pi_i^{(\ell-1)}.$$

  - ⋆ **step 2.d**
    Repeat for $i = 1, \ldots, n$.

- **Step 3**
  Change the counter from ℓ to $\ell + 1$ and repeat on step 2 until convergence achieved.

Step 2, the inner loop, is the hybrid algorithm (Gamerman, 1997). Note that in each step in the hybrid algorithm we update one component in $\boldsymbol{\pi}$, the others do not change and are treated as fixed values.

7.3.2 The Acceptance Probability

The Metropolis-Hastings algorithm in the previous section is used to generate a sample from the posterior distribution of $\boldsymbol{\pi}$. This is done in two steps. In the first step we generate a value from the prior distribution of δ; in the second step we accept or reject the corresponding value of π with probability $\alpha(\pi^{(\ell)}, \pi^{(\ell-1)})$. This method is known as the rejection method and $\alpha(\pi^{(\ell)}, \pi^{(\ell-1)})$ is the so called acceptance probability. The second step can be considered as a correction step, in this step the value that was drawn from the prior in the first step will be rejected if the data will not support this value. Let $\theta(z)$ be a density of the random variable z. Suppose that sampling a value from $\theta(z)$ is difficult or not possible. The rejection method consists of drawing a value from an auxiliary density, q, from which draws can be made easily. The main idea is to use q to make the drawing from θ. The restriction over q is that there is a constant A such that $\theta(z) \leq Aq(z)$. q is usually called the envelope density.
The simplest version of the rejection method consists of drawing a value of z from q and value of u from $U(0,1)$. The value of z is accepted if $u \leq \theta(z)/Aq(z)$. In our setting $\theta(z)$ represent the posterior distribution of $\boldsymbol{\pi}$ and $q(z)$ is the Dirichlet prior (or more precicly, the prior distribution of $\boldsymbol{\delta}$).
Suppose that t was drawn in iteration $\ell-1$ and z in iteration ℓ. The acceptance probability is given by

$$\alpha(z,t) = \min\left\{1, \frac{\theta(z)q(z,t)}{\theta(t)q(t,z)}\right\}. \tag{7.9}$$

In case that in each iteration the draw is independent of the previous draw, that is $q(t,z) = f(z)$ one can choose $f(z)$ to be the prior distribution of z and in this case $\theta(z)/q(t,z) = \theta(z)/f(z)$. The acceptance probability in this case becomes

$$\alpha(z,t) = \left\{1, \frac{L(z)}{L(t)}\right\},$$

where $L(z)$ is the likelihood evaluated at z. This means that the chain is moved from t to z if $L(z) > L(t)$. If $L(t) > L(z)$ the value of z is accepted with probability $L(z)/L(t)$. Note that even though we draw values from the prior distribution, the rejection method insures that after K iterations we will be able to approximate the posterior distribution. In out setting, $f(\delta_i) = \text{Beta}(\gamma_i, \gamma_{i+1})$, so the value of $\delta_i^{(\ell+1)}$ is drawn from $\text{Beta}(\gamma_i, \gamma_{i+1})$ independent of the value of $\delta_i^{(\ell)}$.

7.4 Application to the Data

We applied the proposed method for the rubella and the mumps datasets. For each dataset we used the prior models as discussed in Section 7.2.4. For the precision parameter we used M=10000, 1000, 500 and 250 (for rubella) and M=5000, 1500, 500 and 250 for mumps.

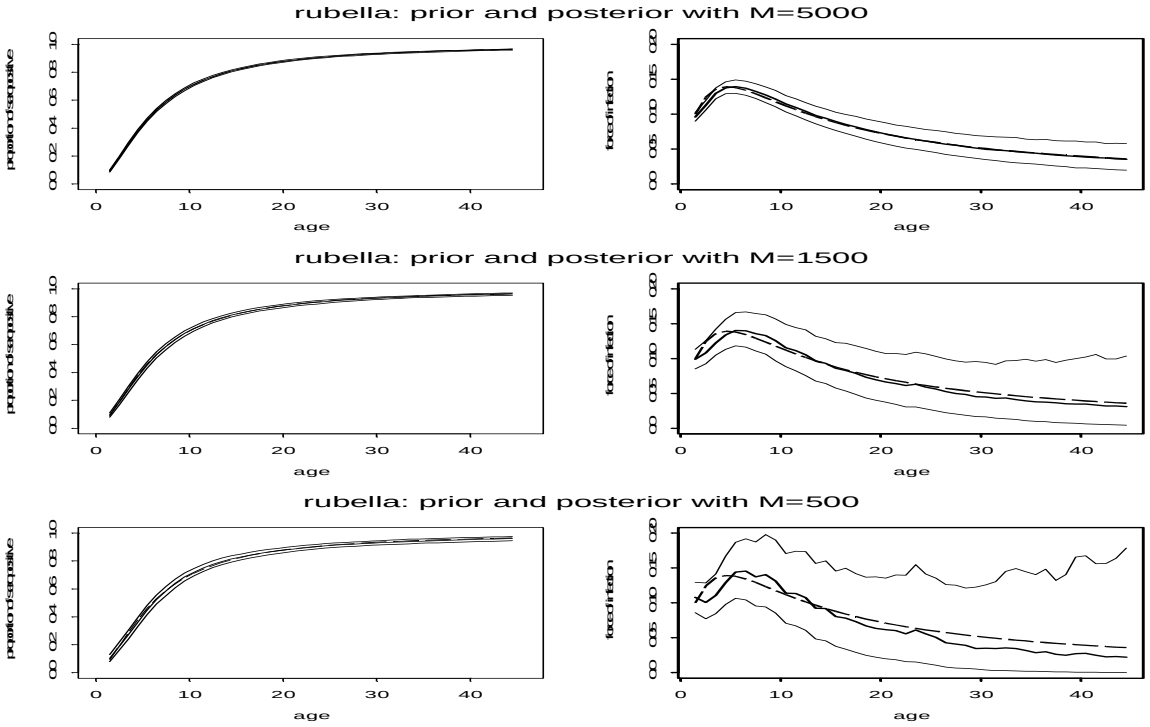


Figure 7.3: *Rubella: prior and posterior prevalence (left panels) and force of infection (right panels) for M=5000, 1500 and 500.*

The results presented here are based on MCMC simulations with 10000 iteration. The first 5000 runs were considered as burn-in period and were discarded form the analysis. Figure 7.3 shows the posterior means for the prevalence and the force of infection. Note that the width of the 95% credible intervals dicrease with the value of M. Figure 7.4 shows the prior curve, the posterior means and the local quadratic estimate for the force of infection. Compared with the prior curve, the posterior force of infection reaches a peak at a higher age with higher values. This pattern supported by the local quadratic estimate for the force of infection. Note that at the older age groups, the levels of the force of infection decrease with the value of M. The results for mumps are shown in Figures 7.5 and 7.6. The posterior force of infection reaches a peak at the same age as the prior curve but the level at the peak of the posterior force of infection is higher than the prior force of infection. The posterior force of infection also indicates on a secondary peak at age 11. From age 30 onward the force of infection is estimated to be zero.
As an informal diagnostic for convergence, we present the trace plot and the autocorrelation function for the force of infection at several ages. Figure 7.7 (from top to bottom) shows the plots for the force of infection at age 5.5, 15.5,20.5 and 30.5 . The low autocorrelation observed in Figure 7.7, which presents the results for the last 5000 iteration of the chain, disappears when we monitor the value of $\ell(a)$ every 5 steps, as shown in Figure 7.8. The trace plots and the autocorrelation function for $\ell(15.5)$ for models with $M = 5000, 1500, 500$ and 250 are shown (from top to bottom) in Figure 7.9. The autocorrelation is decreasing quickly. Although, the smaller the value of M (i.e. the prior becomes "less" informative), the faster the autocorrelation is decreasing.

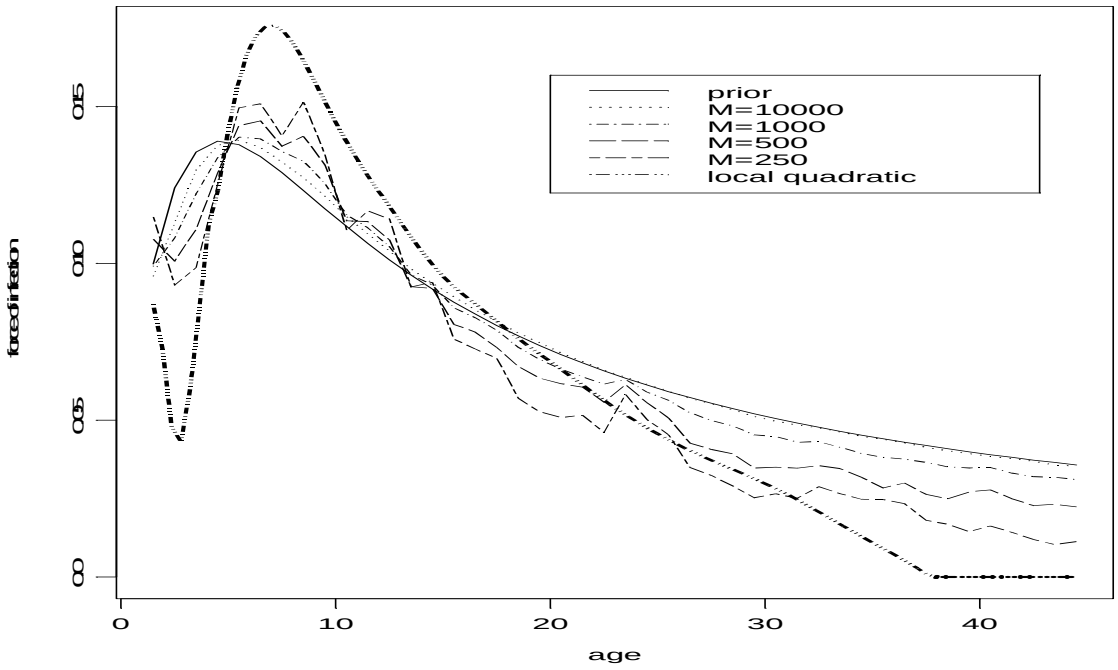


Figure 7.4: *Rubella: posterior means for the force of infection for several values of M.*

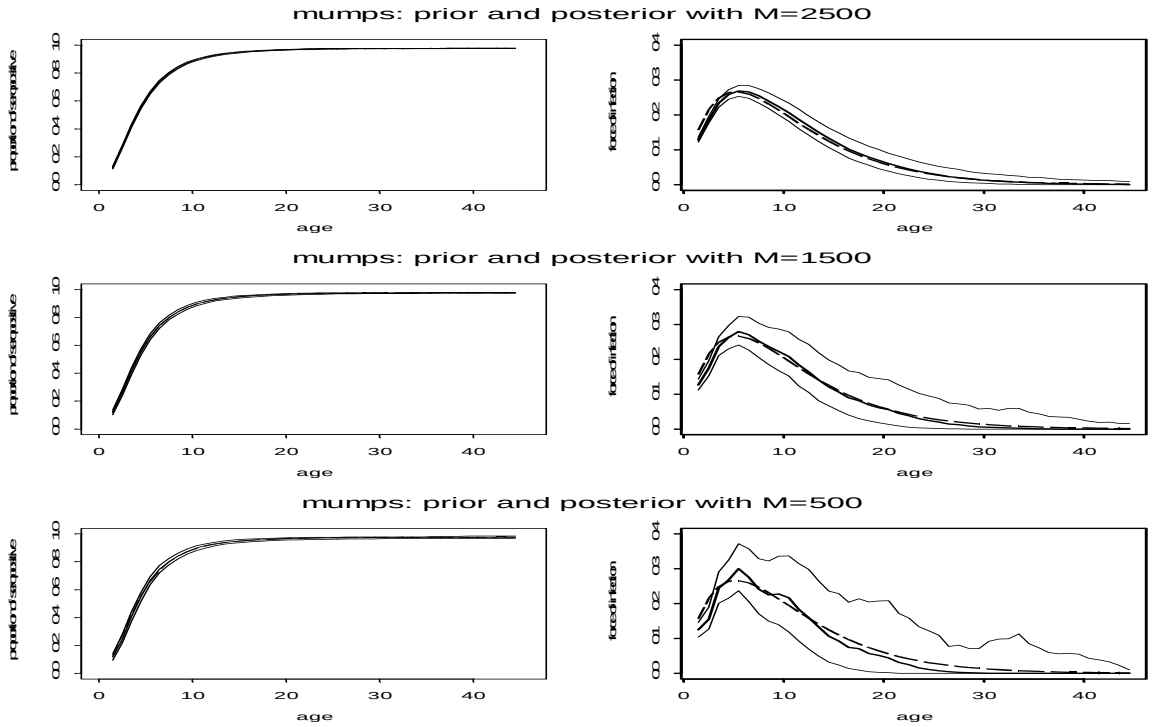


Figure 7.5: *Mumps: prior and posterior prevalence (right panels) and force of infection (left panels) for M=1500, 1500 and 500.*

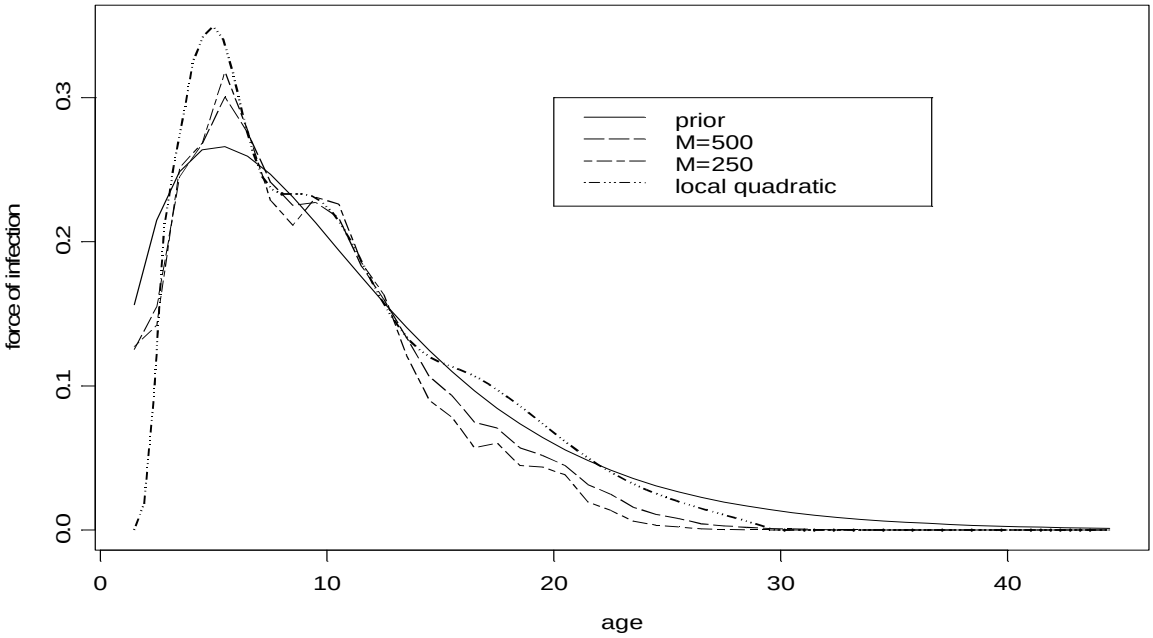


Figure 7.6: *Mumps: posterior means for the force of infection with M=250 and 500 and local quadratic estimate.*

7.5 What Does the Choice of M Actually Do ?

We mentioned in Section 7.2.3 that M is the precision parameter in the prior distribution of δ. Figure 7.10 shows the prior distribution of δ for different age groups when $M = 10000$. At younger age groups the prior distribution symmetric and it centered around the prior mean. As age increase the prior distribution become "less informative" in a sense that the variability around the prior mean increase. The same patterns can be detected in Figure 7.11 for $M = 1500$. Note, that at older age groups the prior distribution is rather flat (compared with the younger age group). When $M = 500$ the prior distribution of δ has higher density near zero and 1 than in the center (see Figure 7.12). Note that for $\delta_i = 0$ $\pi_i = \pi_{i-1}$ while for $\delta_i = 1$ $\pi_i = \pi_{i+1}$. The influence of M on the force of infection is illustrated in Figures 7.13 and 7.14 which show the density estimate for the posterior distribution of the force of infection.

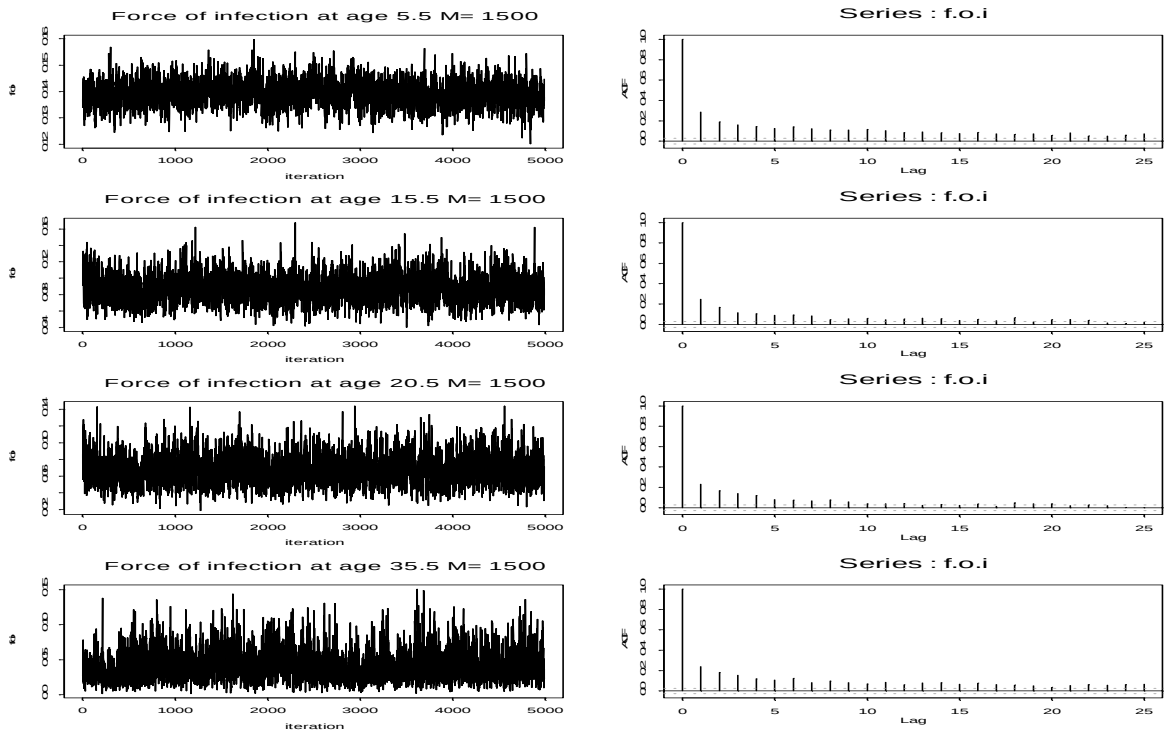


Figure 7.7: *Trace plot (left panels) and autocorrelation functions (right panels) for* $\ell(5.5), \ell(15.5), \ell(20.5)$ *and* $\ell(30.5)$*. The results are based on the last 5000 iterations of the chain for the model with M=1500.*

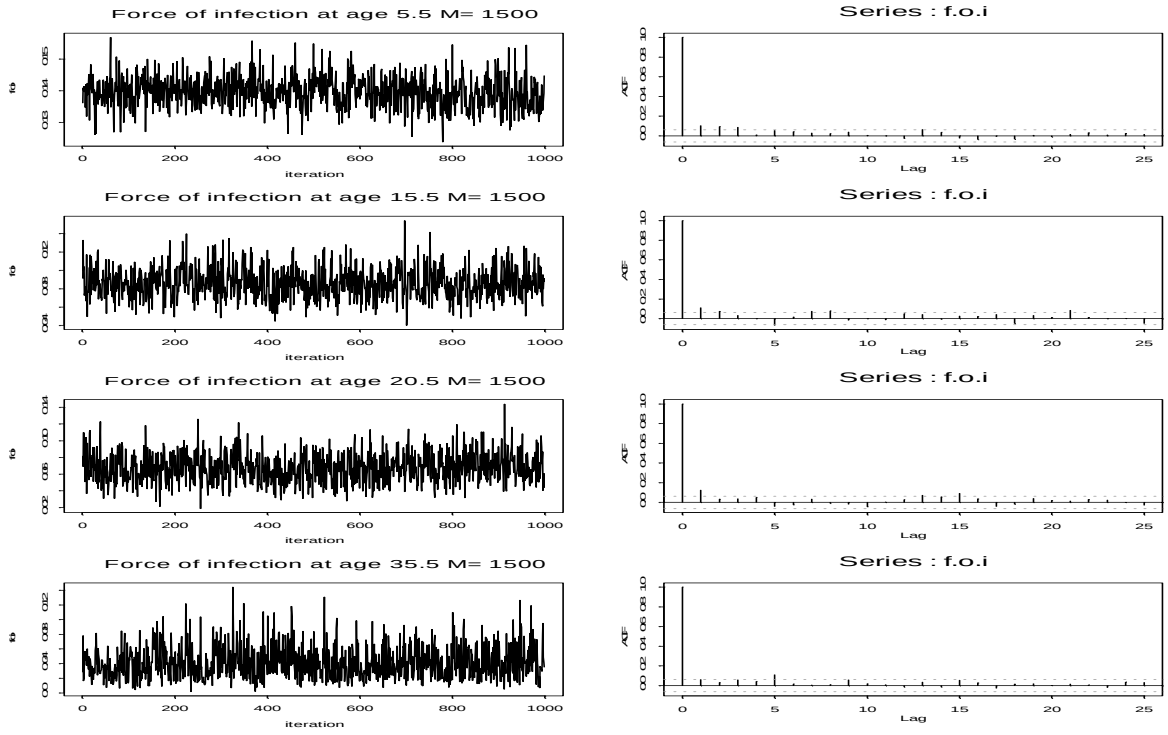


Figure 7.8: *Trace plot (left panels) and autocorrelation functions (right panels) for* $\ell(5.5), \ell(15.5), \ell(20.5)$ *and* $\ell(30.5)$*. The results are based on the last 5000 iterations of the chain* ***when the chain is monitored every 5 iterations*** *for the model with M=1500.*

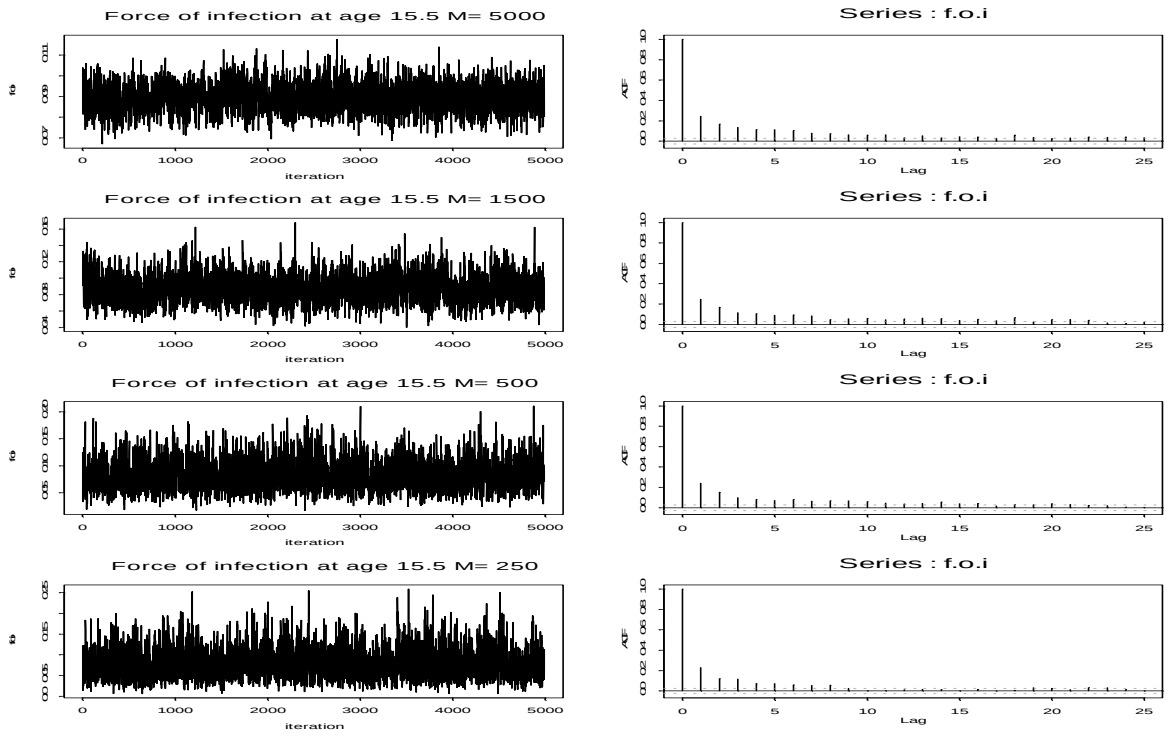


Figure 7.9: *Trace plot (left panels) and autocorrelation functions (right panels) for* $\ell(15.5)$. *The results are based on the last 5000 iterations of the chain for the models with M=5000,1500,500 and 250.*

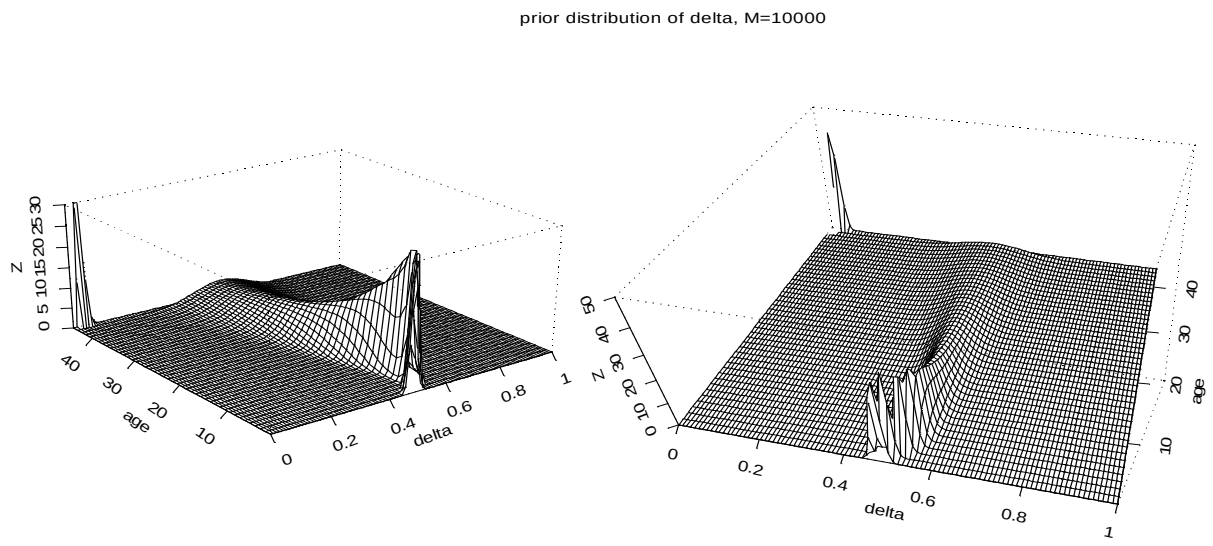


Figure 7.10: *Prior distribution of δ_i with $M = 10000$. Note that both plots show the same distribution from two different points of view.*

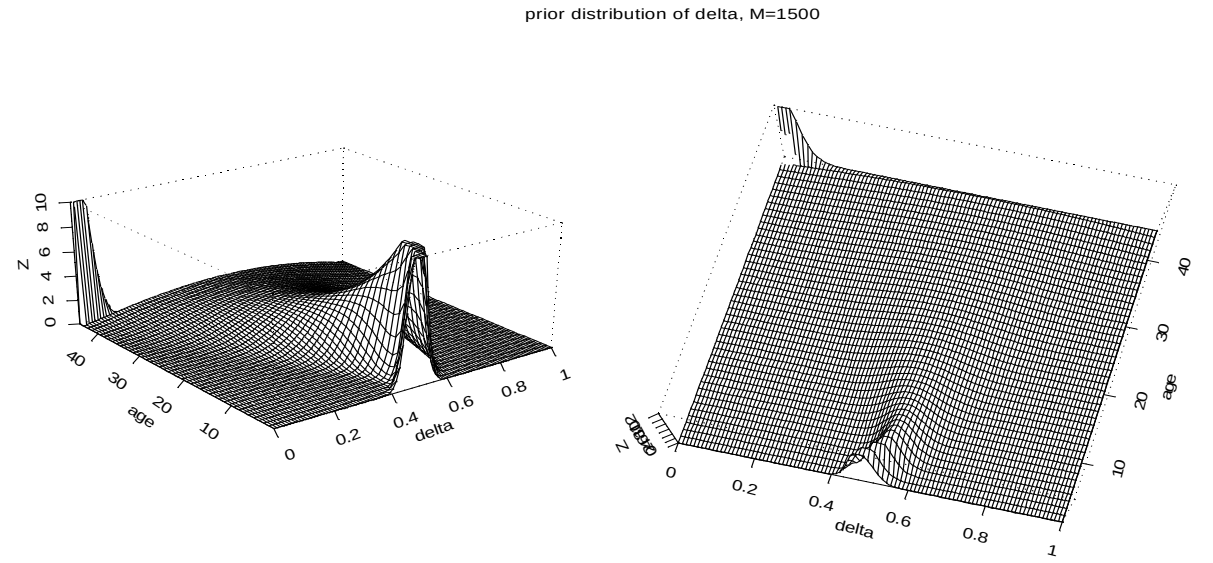


Figure 7.11: *Prior distribution of δ_i with $M = 1500$.*

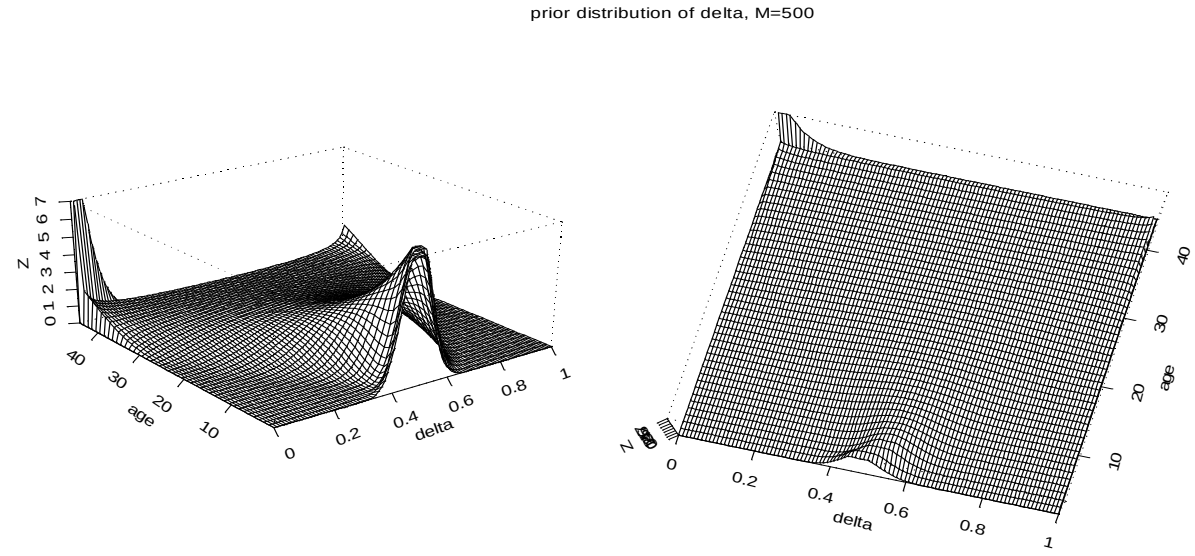


Figure 7.12: *Prior distribution of δ_i with $M = 500$.*

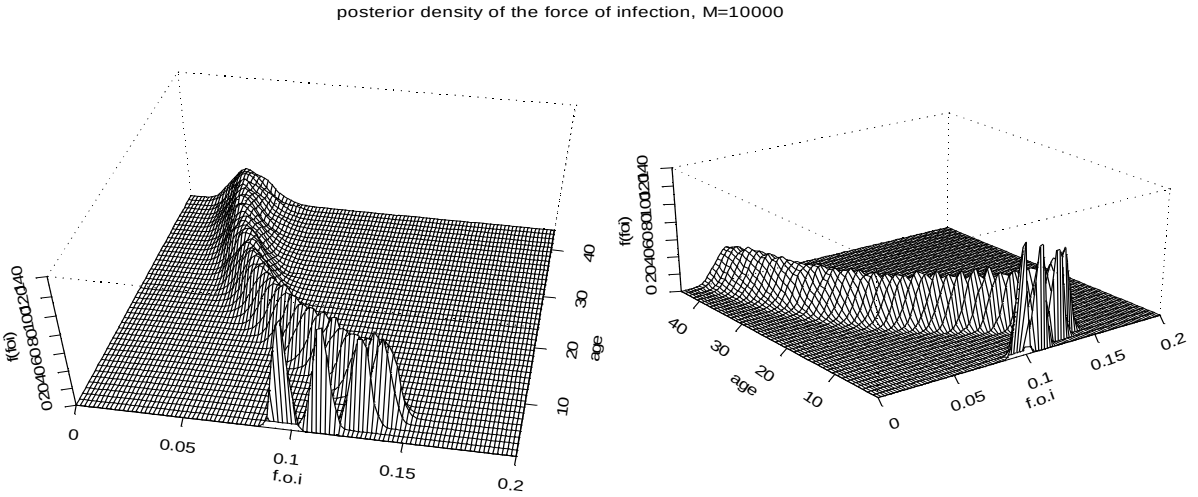


Figure 7.13: *Posterior density of the force of infection with $M = 10000$. Note that both plots show the same distribution from two different points of view.*

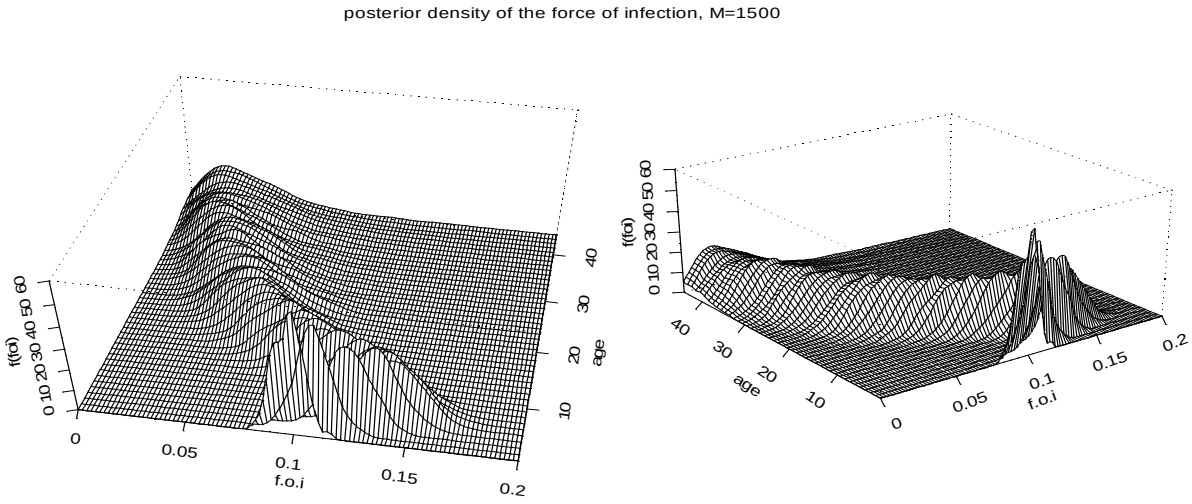


Figure 7.14: *Posterior density of the force of infection with* $M = 1500$.

7.6 Discussion

We have shown that compared with the beta-binomial model the Dirichlet process prior allows us to incorporate prior knowledge about the force of infection in the model. We have shown that parametric models can be used as prior models for the force of infection while the posterior models are nonparametric. This property of the Dirichlet prior distribution makes it attractive for sensitivity analysis. O'Neill (2001), in the discussion for the paper of Farrington *et al.* (2001), arises the question how sensitive the results are for the assumption about the mean structure of the prevalence. Farrington *et al.* (2001) performed a sensitivity analysis for their parametric model by fitting piecewise constant parametric models to the data. Both approaches lead to similar results. Farrington *et al.* (2001) mentioned that the isotonic regression model of Kieding (1991) can be used as well. In the light of the previous chapters, other models such as local polynomials, fractional polynomials, and beta-binomial models can be considered as well. However, there is no doubt, that the model with Dirichlet prior is the most appropriate approach for the sensitivity analysis. It is the only model which takes the parametric model as a prior. The posterior distribution of the force of infection gives us an indication which patterns in the data are missed due to the assumed parametric structure of the force of infection.

Chapter 8

Modeling Age-Dependent Probability to Become Hepatitis B Carrier - A Meta Analysis

8.1 Introduction

In the previous chapter the attention was placed on estimating the rate in which individuals leave the susceptible class for several infection diseases. In this chapter we focus on a different compartment in the transmission model for Hepatitis B. More precisly, the aim of this chapter is to estimate the probability to become hepatitis B carrier. This is done by means of a meta-analysis of a dataset perviously analyzed by Edmunds *et al.* (1993). Edmunds *et al.* (1993) assume that the probability to become a hepatitis B carrier is age dependent with constant probability in the perinatal period (age at infection ranging from 0 to 6 months) and exponential decline thereafter. The data, as well as the scientific question and the model estimated by Edmunds *et al.* (1993) are presented in Section 8.2. In Section 8.3 we present a hierarchical Bayesian changepoint model for the probability to become a carrier. In this model the age at which the probability starts to drop down, is not fixed in advance but treated as a parameter in the model. Section 8.4 presents models that assume different forms for the relationship between age and the probability to become a carrier. In these models, the probability to become a carrier in the perinatal period is not necessarily constant. The deviance information criterion (DIC) is used for model selection. In Section 8.5 we discuss random effects models, taking into account clustering in the data. Section 8.6 is devoted for model criticism and monitoring convergence of the changepoint model.

8.2 The Data

Hepatitis B (HB) represents a major health problem in most of the world. The World Health Organization estimates that about 350 million people are carriers of the HB virus (HBV) and that annually about 0.9 million deaths are caused by hepatitis B. Essentially a relatively virulent pathogen borne by bodily fluids, HBV transmission can occur via a

multitude of routes. Perinatal transmission may occur from an infected mother to her child. Horizontal transmission from person-to-person (mostly from child-to-child) may occur at any time when very small amounts of saliva or blood are transferred via small skin wounds (e.g. impetigo, scabies lesions, abrasions, leg ulcers or infected insect bites). Transmission may also occur during homo- and heterosexual intercourse for which the rate of sexual partner changes and receptive anal intercourse are important risk factors. Finally, parenteral transmission occurs when the virus spreads by penetration of the skin with an infected object, i.e. by needle stick, mucous membrane splash, tatooing, ear piercing, etc. Health care workers and injecting drug users are generally considered key risk groups for this transmission route. Most of the HBV disease burden is due to long-term chronic sequelae of HB, which can culminate in severe inflammation of the liver, leading to cirrhosis and hepatocellular carcinoma. As chronic HB does not become symptomatic until many years (often decades) after the infection, the link with the initial cause, infection with HBV, is often not made.

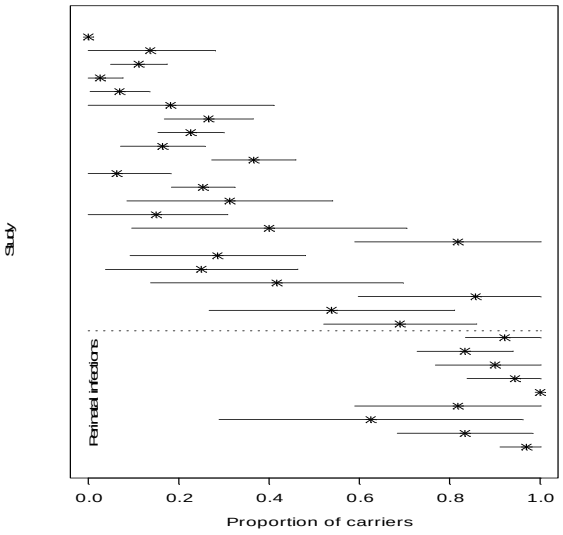
The functional form of the probability to become HBV carrier was investigated by Edmunds *et al.* (1993) who presented a collection of 21 studies from the epidemiology literature. Interestingly, to our knowledge, this is the only dataset that is available about this problem. The data are presented in Table 8.8 in the Appendix. There are 25 unique age groups (with mid-age ranging from 0.25 to 25 years) and the number of different studies within each age group ranges from 1 to 4. For 18 studies there is only one age group. For these studies the available information is (1) the mid-age (the age group) in the ij'th sample, x_{ij}, $i = 1, \ldots, n_j$, $j = 1, 2, \ldots, 25$, where n_j is the number of samples in age group j, (2) the age range in the samples, (3) the number of infected individuals (the sample size, n_{ij}) and (4) the number of carriers, y_{ij}. Three studies, Coursager *et al.* (1987), Marinier *et al.* (1985) and McMahon *et al.* (1985) consist of 6, 3 and 4 age groups respectively. Thus, in total, there is information available from 31 different samples. The data are shown in Figure 8.1.

Edmunds *et al.* (1993) focused on modeling an age-dependent probability to become a carrier. They distinguished between two age periods: (1) the perinatal infection period (0-0.5 years) in which the probability to become a carrier is high and assumed to be constant, (2) age groups older that 0.5 years with a probability that rapidly declines with age and levels in young adults. There are 9 age groups in the perinatal infection period and 22 in the older age groups.

The modeling strategy in Edmunds *et al.* (1993) was as follows. The data (the 31 different samples) was split into two groups according to the two age periods mentioned above and the parameters were estimated separately for each subgroup. By pooling the data in the perinatal infection period, the probability to become a carrier was estimated to be $\hat{\theta}(x) = \sum_{ij} y_{ij} / \sum_{ij} n_{ij} = 0.885$. For the older age groups (0.68-25 years) Edmunds *et al.* (1993) assumed that $\theta(x) = e^{-\beta_1 x^{\beta_2}}$ and estimated the probability by maximizing the log-likelihood $L(\beta_1, \beta_2) = \Sigma_{\ell=10}^{31} y_\ell \log(\theta(x_\ell)) + (n_\ell - y_\ell) \log(1 - \theta(x_\ell))$. Here, x_ℓ is the mid-age in the ℓth age group. The estimated model is given by

$$\theta(x) = \begin{cases} e^{-0.1221} & x \leq 0.5 \\ e^{-0.645 x^{0.455}} & x > 0.5. \end{cases} \tag{8.1}$$

The focus in this chapter is placed on three main issues: (1) assuming that the probability to become a HBV carrier is constant in the perinatal infection period, at what age (denoted



Study	Mid-age	Study No.	N/C	% carrier
Iwarson(1971)	25	21	110 / 0	0
McMahon(1985,4)	25	11	22 / 3	13.6
Neilsen(1971)	23	20	99 / 11	11.1
Beasley(1983)	19.8	19	39 / 1	2.6
McMahon(1985,.)	15	11	58 / 4	6.9
Gambian data(4)	14.96	18	11 / 2	18.2
Gambian data(3)	11.06	17	79 / 21	26.6
Gambian data(2)	7.95	16	128 / 29	22.7
McMahon(1985,2)	7.5	11	61 / 10	16.4
Krugman(1970)	6.5	15	104 / 38	36.5
Coursaget(1987,6)	6	4	16 / 1	6.2
Gambian data(1)	4.07	14	150 / 38	25.3
Colombo(1977)	4	12	16 / 5	31.2
Coursaget(1987,5)	3	4	20 / 3	15
Hayashi(1987)	2.9	12	10 / 4	40
Marinier(1985,2)	2.6	11	11 / 9	81.8
McMahon(1985,1)	2.5	10	21 / 6	28.6
Coursaget(1987,4)	1.5	4	16 / 4	25
Marinier(1985,2)	1.5	10	12 / 5	41.7
Marinier(1985,1)	0.85	10	7 / 6	85.7
Coursaget(1987,3)	0.75	4	13 / 7	53.8
Coursaget(1987,2)	0.68	4	29 / 20	69
Stevens(1975)	0.5	9	38 / 35	92.1
Stevens(1979)	0.5	8	48 / 40	83.3
Schweitzer(1975)	0.5	7	20 / 18	90
Gerety(1977)	0.5	6	18 / 17	94.4
Okada(1975)	0.42	5	8 / 8	100
Coursaget(1987,1)	0.38	4	11 / 9	81.8
Skinhoj(1978)	0.3	3	8 / 5	62.5
Wong(1984)	0.25	2	24 / 20	83.3
Beasley(1981)	0.25	1	33 / 32	97

Figure 8.1: *21 studies presented in Edmunds et al. (1993). ML estimates and* 95% *confidence intervals for the proportions of carriers.*

as $x^{(k)}$) does the probability start to drop down ? (2) Does the probability in the perinatal infection period stay constant or does it decrease immediately from birth ? and (3) Edmunds *et al.* (1993) consider the 31 samples as independent samples. However, the data of three studies consists of more than one age group. Thus, a multilevel model, which assumes that individuals from the same study may be correlated, can be fitted in order to investigate the clustering effect in the data.

8.3 Bayesian Hierarchical Changepoint Model for the Probability to Become a Carrier

The combination of the binomial data with the functional form for the probability in equation (8.1), as assumed by Edmunds *et al.* (1993), implies that a generalized linear model with complementary log-log link can be used to model the probability to become a HBV carrier. More precisely, following Edmunds *et al.* (1993) we assume that

$$\theta(x_{ij}) = \begin{cases} e^{-\alpha_1} & j \leq k \ (x_{ij} \leq x^{(k)}) \\ e^{-\alpha_2 x_{ij}^{\beta_3}} & j > k \ (x_{ij} > x^{(k)}), \end{cases} \tag{8.2}$$

where x_{ij} is the mid-age in the ij'th sample and $x^{(k)}$ is the unknown changepoint. Using the complementary log-log link function, $g(\theta(x)) = \log(-\log(1-\theta(x)))$, we define the linear predictor by

$$\eta(x_{ij}) = \beta_1 I_{ij}^{(k)} + \beta_2(1 - I_{ij}^{(k)}) + \beta_3(1 - I_{ij}^{(k)}) \log(x_{ij}), \tag{8.3}$$

where $I_{ij}^{(k)}$ is an indicator variable which takes the value of 1 if $x_{ij} \leq x^{(k)}$ and 0 elsewhere, $\beta_1 = \log \alpha_1$ and $\beta_2 = \log \alpha_2$.
Bayesian models for changepoint problems were discussed in detail by Carlin, Gelfand and Smith (1992). Models for binomial variables were discussed by Hinkley and Hinkley (1970) and Smith (1975). We assume that $y_{ij} \sim P_1(y|\theta_1)$, $j = 1, 2, 3, \ldots, k$ and $y_{ij} \sim P_2(y|\theta_2)$ $j = k+1, k+2, \ldots, n$. In our setting n is the number of unique age groups. Note that P_1 and P_2 are assumed to be known; in our example P_1 and P_2 are both binomial. The corresponding likelihood is given by

$$P(y|k, \theta_1, \theta_2) = \prod_{j=1}^{k} \prod_{i=1}^{n_j} \theta_1(x)^{y_{ij}} (1 - \theta_1(x))^{n_{ij} - y_{ij}} \prod_{j=k+1}^{n} \prod_{i=1}^{n_j} \theta_2(x)^{y_{ij}} (1 - \theta_2(x))^{n_{ij} - y_{ij}}. \quad (8.4)$$

For a given value of k the linear predictor of the model is given by $\eta(x) = \left[\mathbf{X}_1^{(k)} | \mathbf{X}_2^{(k)}\right]^T \boldsymbol{\beta}$, where

$$\mathbf{X}_1^{(k)} = \begin{pmatrix} 1 & \ldots & 1 \\ 0 & \ldots & 0 \\ 0 & \ldots & 0 \end{pmatrix}^T, \qquad \mathbf{X}_2^{(k)} = \begin{pmatrix} 0 & \ldots & 0 \\ 1 & \ldots & 1 \\ \log(x_{k+1}) & \ldots & \log(x_n) \end{pmatrix}^T,$$

and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. At the second level of the model we assume independent normal priors for β_1, β_2 and β_3, $\beta_\ell \sim \mathrm{N}(\mu_\ell, \sigma_\ell^2)$. The full conditional distribution of β_1 is given by

$$P(\beta_1|\beta_2, \beta_3, k) \propto \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(\beta_1 - \mu_1)^2} \times \prod_{j=1}^{k} \prod_{i=1}^{n_j} \left\{1 - g^{-1}(\eta(x_{ij}))\right\}^{y_{ij}} \left\{g^{-1}(\eta(x_{ij}))\right\}^{n_{ij} - y_{ij}}, \quad (8.5)$$

the full conditional distribution of β_2 is given by

$$P(\beta_2|\beta_1, \beta_3, k) \propto \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(\beta_2 - \mu_2)^2} \times \prod_{j=k+1}^{n} \prod_{i=1}^{n_j} \left\{1 - g^{-1}(\eta(x_{ij}))\right\}^{y_{ij}} \left\{g^{-1}(\eta(x_{ij}))\right\}^{n_{ij} - y_{ij}}. \quad (8.6)$$

At the third level of the model we specify the hyperprior distributions for the variance and the mean of $\boldsymbol{\beta}$. We assume that the hyperparameters $\sigma_1^{-2}, \sigma_2^{-2}$ and σ_3^{-2} are independent gamma, $\sigma_\ell^{-2} \sim \mathrm{Gamma}(0.0001, 0.0001)$, $\ell = 1, 2, 3$. Independent normal flat priors, $N(0, 1000)$, were assumed for the hyperprior means μ_1, μ_2 and μ_3.
In contrast with the model proposed by Edmunds *et al.* (1993), the value of the changepoint is not fixed in advance but it is an additional parameter in the model. Therefore, a prior model has to be specified over the set of all possible changepoints. Following Smith (1975) and Carlin, Gelfand and Smith (1992), we assume a discrete uniform prior model for k, that is

$$P(k) = \frac{1}{n+1} \qquad 0 \leq k \leq n. \quad (8.7)$$

This means that the change in the probability can occur before the first age group or at the design points, i.e. $k \in \{0, 1, 2, \ldots, 25\}$ implying that $x^{(k)} \in \{0, 0.25, 0.3, \ldots, 25\}$. Here, $x^0 = 0$ corresponds to the event "no-change". In this case the probability to become HBV carrier starts to drop down from the first age group. We further assume that k and $\boldsymbol{\beta}$ are

independent so that $P(\boldsymbol{\beta}|k, \boldsymbol{\mu}, \boldsymbol{\sigma}) = P(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\sigma})$. The joint posterior distribution is given by $P(k, \boldsymbol{\beta}|y) \propto P(y|k, \boldsymbol{\beta})P(\boldsymbol{\beta}|\boldsymbol{\sigma}, \boldsymbol{\mu})P(k)$ and the conditional posterior distribution of k is $P(k|\boldsymbol{\beta}, y) \propto P(y|k, \boldsymbol{\beta})$.
Thus, combining the likelihood with the prior and hyperprior models discussed above leads to the hierarchical model

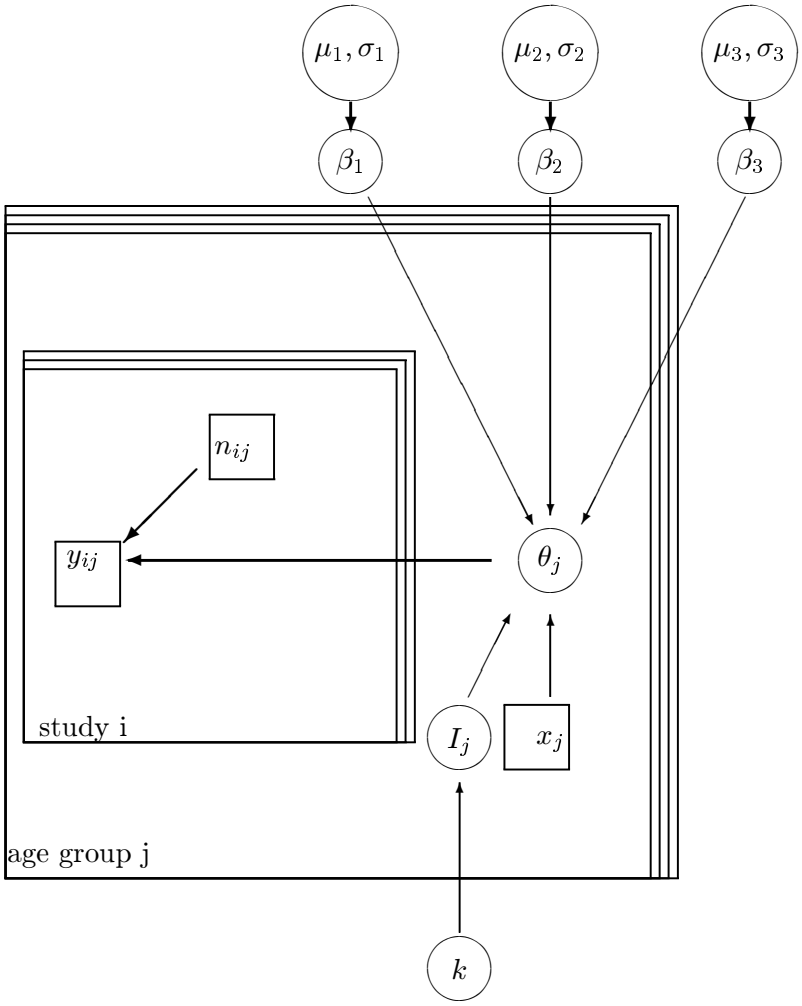


Figure 8.2: *The structure of the changepoint model. Squares and circles represents observed and unobserved components in the model respectively. Thick and thin arrow represents stochastic and deterministic relationships in the model respectively.*

$$
\begin{aligned}
&y_{ij} \sim \text{Bin}(n_{ij}, \theta(x_{ij})) \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, n_j \\
&\log(-log(1-\theta(x_{ij})) = \beta_1 I_{ij}^{(k)} + \beta_2(1 - I_{ij}^{(k)}) + \beta_3(1 - I_{ij}^{(k)}) \log(x_{ij}) \\
&\beta_\ell \sim \text{N}(\mu_\ell, \sigma_\ell^2) \quad \ell = 1, 2, 3 \\
&k \sim \text{discrete uniform}(0, 1, 2, \ldots, 25) \\
&\sigma_\ell^{-2} \sim \text{gamma}(0.0001, 0.0001) \quad \ell = 1, 2, 3 \\
&\mu_\ell \sim \text{N}(0, 10) \quad \ell = 1, 2, 3.
\end{aligned}
\tag{8.8}
$$

For a given value of k, the changepoint model implies that from birth up to a certain age, $x^{(k)}$, the probability $\theta_1(x)$ is age independent. Furthermore, the event "$k = 25$", i.e., "$x^{(k)} = 25$", can be interpreted as "no change" which corresponds to the common mean model which predict constant probability to become a HBV carrier. In this case the second term on the right hand side of (8.4) vanishes. The event "$k = 0$" (i.e., "$x^{(k)} = 0$") implies that the probability to become a HBV carrier drops from the first age group and onwards. In this case the first term in the right hand side is vanished. Finally, the event "$k = 5$" (i.e., "$x^{(k)} = 0.5$") means that the changepoint model corresponds to a constant probability in the perinatal infection period as suggested by Edmunds *et al.* (1993).

8.3.1 Application to the Data

The posterior median for k is 5 (posterior mean is 5.109) with 95% credible interval $(5, 6)$ which corresponds to $\bar{x}^{(k)} = 0.5156$ with 95% credible interval of $(0.5, 0.68)$. Figure 8.3 shows the histograms of the posterior distributions of k and $x^{(k)}$. The prior model for k assumes that $P(k = \ell) = \frac{1}{26}$ for $j = 0, \ldots, 25$ but the histogram shows that the posterior frequencies are $\hat{P}(k = 5) = 0.922$, $\hat{P}(k = 6) = 0.062$, $\hat{P}(k = 7) = 0.001$ and $\hat{P}(k = 8) = 0.015$ (for other values $\hat{P}(k = \ell) = 0$). Figure 8.4 (panel a) shows that in the perinatal period the changepoint model and model proposed by Edmunds *et al.* (1993) gave the same results. In the older age groups, the changepoint model estimates a slightly higher probability than Edmunds' model. However, the fact that the posterior median of k is equal 5 means that the age in which the change occurs, the age in which the probability starts to drop down, is 0.5 (the fifth age group is 0.5). This confirms the assumption made in Edmunds *et al.* (1993) about the length of the perinatal infection period. Table 8.1 shows maximum likelihood estimates, bootstrap estimates (obtain from a nonparametric bootstrap procedure with 2000 bootstrap samples) and the posterior mean of the hierarchical changepoint model. Parameter estimates obtained from the three methods are comparable. Note that the posterior means for the standard error are slightly higher than the maximum likelihood estimates. This reflect the extra variability in the Hierarchical Bayesian model introduced by the prior models.

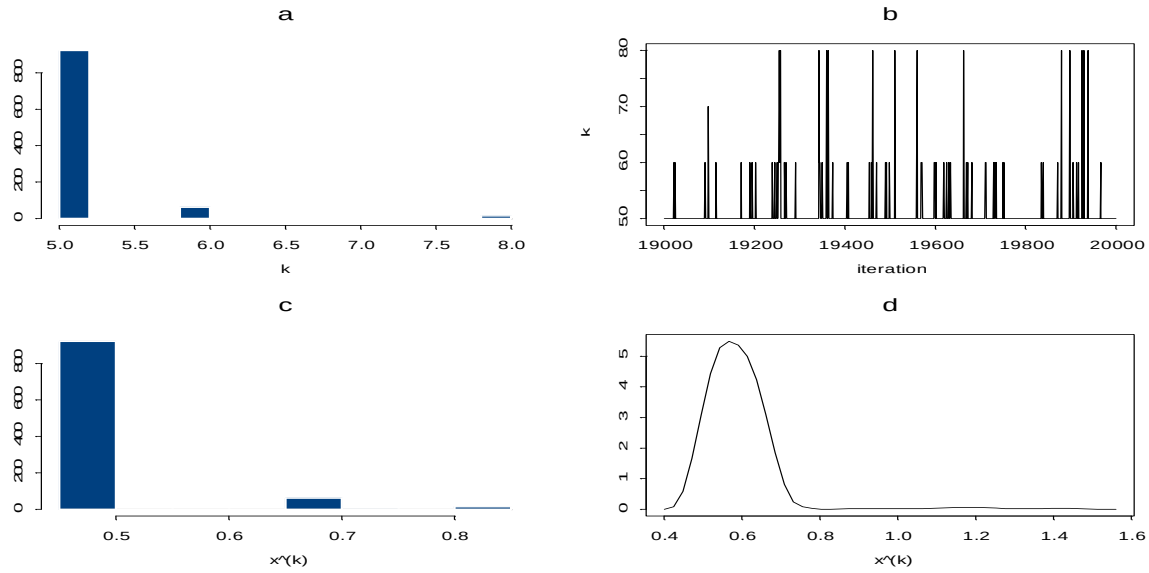


Figure 8.3: *The changepoint model. Panel a : histogram for the posterior distribution of k. Panel b: trace plot (based on the last 1000 iterations) for k, Panel c: histogram for the posterior distribution of* $x^{(k)}$*, Panel d: density estimate for the posterior distribution of* $x^{(k)}$*.*

8.3.2 Sensitivity Analysis for the Distribution of the Changepoint

The models in Section 8.2 assume that the changepoint is a discrete random variable, uniformly distributed over the range $\{0, 0.25, 0.3, \ldots, 25\}$. Thus, the changepoint in this model can occur before the first age group or at the design points. In this section we assume that $x^{(k)} \sim \text{Uniform}(a, b)$, where a and b are the lower and upper bounds for age interval in which the changepoint can be occured. The choice of $(a, b) = (0, 25)$ corresponds to the non informative prior that was used in Section 2. Other choices of (a, b) imposed informative prior for $x^{(k)}$ which restricts the age interval in which the changepoint can be occured. Moreover, we specify continuous uniform priors and therefore allow this changepoint to occur in any point between a and b. Table 8.2 presents the posterior means for several choices of $U(a, b)$. The posterior means for $\boldsymbol{\beta}$ are only slightly different when the prior model for $x^{(k)}$ is changed. The posterior means for $x^{(k)}$ range between 0.51 and 0.59. Note that the posterior mean of the deviance, $\bar{D}$, decreases when the prior $U(a, b)$ is more centered around the value of 0.5, as expected.

Table 8.1: *Parameters estimates for Model 1. For the bootstrap, the parameter estimates are medians of bootstrap distributions based on 2000 nonparametric bootstrap resamples. For the hierarchical Bayesian models, the parameter estimates are the posterior means.*

Parameter	ML	Bootstrap	Full Bayesian
β_1	-2.0988	-2.1062	-2.099
β_2	-0.5206	-0.5212	-0.5189
β_3	0.4735	0.4752	0.4722
$\hat{\sigma}(\beta_1)$	0.2035	0.2108	0.2127
$\hat{\sigma}(\beta_2)$	0.1082	0.1072	0.1119
$\hat{\sigma}(\beta_3)$	0.0482	0.0476	0.0499
k	fixed at 5	fixed at 5	5.083
$x^{(k)}$	fixed at 0.5	fixed at 0.5	0.5132

Table 8.2: *Sensitivity analysis for the prior distribution of* $x^{(k)}$.

Parameter	Discrete uniform	$x^{(k)} \sim U(0, 25)$	$x^{(k)} \sim U(0, 2)$	$x^{(k)} \sim U(0.3, 0.8)$	$x^{(k)} \sim U(0.4, 0.6)$
β_1	-2.099	-2.101	-2.099	-2.105	-2.116
β_2	-0.5189	-0.5175	-0.5209	-0.5238	-0.5244
β_3	0.4722	0.4715	0.4732	0.4744	0.4749
$x^{(k)}$	0.5132	0.5982	0.5974	0.5830	0.5454
$\bar{D}$	82.84	82.8	82.71	82.50	82.41

8.4 Sensitivity Analysis For the Mean Structure

In this section we focus on the question whether $\theta(x)$ is constant during the perinatal period. Five additional hierarchical Bayesian models were fitted, all with complementary log-log link function and with linear predictors given by

(2): $\eta(x_{ij}) = \beta_1 + \beta_2(1 - I_{ij}^{(k)}) + \beta_3 \log(x_{ij})$
(3): $\eta(x_{ij}) = \beta_1 + \beta_2(1 - I_{ij}^{(k)}) + \beta_3 \log(x_{ij}) + \beta_4(1 - I_{ij}^{(k)}) \log(x_{ij})$
(4): $\eta(x_{ij}) = \beta_1 + \beta_3 \log(x_{ij})$
(5): $\eta(x_{ij}) = \beta_1$.
Note that $I^{(k)}$ is the same indicator variable as before. The same prior and hyperprior models as in Section 2 were assumed. Model 2 assumes that the probability starts to drop down from birth but with different intercepts for the perinatal period and the older age group. Model 3 assumes that the probability drops down from birth but allows, in addition to model 2, two different decreasing rates in the two periods. Model 4 assumes that the probability to become a carrier starts to drop down from birth with the same functional form in the two age periods. Model 5, the common mean model, assumes that the probability to become a carrier is constant and therefore age independent. Note that model 4 and 5 are special cases of the changepoint model in which the changepoint is fixed at 0 and 25 years respectively.
We assess the adequacy of the models mentioned above using the deviance information criterion (DIC) proposed by Spiegelhalter *et al.* (1998, 2002) and discussed in Chapter 6.

8.4.1 Application to the Data

The deviance summaries for all models are given in Table 8.3. The first three models are changepoint models. Models 1 and 2 were parameterized with 4 parameters. For these models the effective number of parameters is 3.684 and 4.39 respectively. Model 3 was parameterized with 5 parameters and for this model $P_D = 7.73$. Models 4 and 5 were estimated with 2 and 1 parameters respectively, for these models the effective number of parameters is 2.003 and 1.03 respectively.
Among the five models considered, model 1, which assumes constant probability during the perinatal period, has the smallest DIC-value, 86.52. The posterior mean of the changepoint, $\bar{x}^{(k)}$, in model 1 is 0.51.

Table 8.3: *Deviance summaries for data.*

Model	$\bar{D}$	$D(\bar{\theta})$	P_D	DIC	$\bar{x}^{(k)}$
1	82.84	79.157	3.682	86.522	0.5132
2	83.59	79.190	4.399	87.898	0.5436
3	86.12	78.384	7.735	93.855	0.8802
4	105.30	103.296	2.003	107.303	fixed at 25
5	527.90	526.862	1.037	528.930	fixed at 0

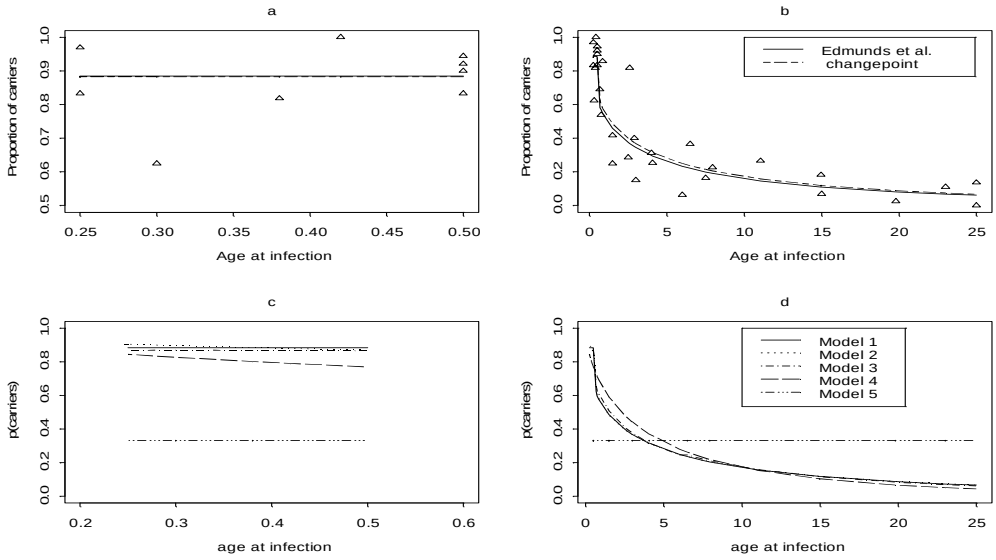


Figure 8.4: *Panel a and b: the changepoint model and the model reported in Edmunds et al. (1993), panel a presents the data and estimated models in the perinatal period. Panel c: Posterior means for $\theta(x)$ in the prenatal period obtained from the different models fitted in Section 8.4. Panel d: Posterior means for $\theta(x)$ for older age groups obtained from the different models fitted in Section 8.4.*

8.5 Random Effects Models

In previous sections the 31 samples were assumed to be independent. However, as mentioned previously, the data were collected from 21 different studies: 18 studies consist of only one sample (or one age group), the other 3 consist of more than one age group. In fact, 14 out of the 31 samples (41.9%) are outcomes of these three latter studies. In this section we investigate this feature of the data by including a study specific random intercept in the model. Thus, the model allows for heterogeneity due to the study.
Such a model can be fitted using generalized linear mixed model (GLMMs) (McCulloch and Searle, 2001) or hierarchical Bayesian model. Within the GLMMs framework it is assumed that

$$\begin{aligned} E(\mathbf{y}_\ell) &= \theta_\ell \quad \ell = 1, 2, \ldots, L, \\ g(\theta_\ell) &= \mathbf{X}_\ell \boldsymbol{\beta} + \mathbf{Z}_\ell \mathbf{u}. \end{aligned} \qquad . \tag{8.9}$$

Here, ℓ is the study's index, $\ell = 1, 2, \ldots, 21$. $\mathbf{y}_\ell$ is the vector of measurements for the ℓ'th study, in our setting it is a scalar for the 18 studies with one age group. $\boldsymbol{\beta}$ is the fixed effects parameter vector and $\mathbf{u}$ is the random component in the model. $\mathbf{X}_\ell$ and $\boldsymbol{Z}_\ell$ are the design matrices for the fixed and random effects respectively. In our setting the distribution of $\mathbf{y}_\ell$ is binomial and in order to complete the specification of the model it is further assumed that $u_\ell \sim N(0, \tau_u^2)$. We consider a complementary log-log normal model. Hence, the linear predictor for the sixth model is given by
(6): $\eta(x_{\ell j}) = \beta_1 I_{\ell j}^{(k)} + \beta_2 (1 - I_\ell^{(k)}) + \beta_3 (1 - I_{\ell j}^{(k)}) \log(x_{\ell j}) + u_\ell$.

Thus, for a given value of k, the conditional mean number of carriers in the ℓth study is

$$\theta(x_{\ell j}|u_\ell) = \begin{cases} e^{-\alpha_1 u_\ell} & j \leq k \ (x_{\ell j} \leq x^{(k)}), \\ e^{-\alpha_2 x_{\ell j}^{\beta_3} u_\ell} & j > k \ (x_{\ell j} > x^{(k)}). \end{cases} \tag{8.10}$$

Within the GLMMs framework, the value of k can be found by a grid search over a set of possible values of k. Note that in this case we assume that the changepoint can occur before the first age-group or only on the design points. The model can be fitted using standard software such as MLWIN or the SAS procedure NLMIXED. Within the hierarchical Bayesian framework the model can be fitted simply by adding an additional prior model for the random effects in (8.8).

8.5.1 Application to the Data

We performed a grid search over the values of $x^{(k)}$ by fitting a sequence of GLMMs, each has different value of $x^{(k)}$. Parameter estimates for the models with $x^{(k)} = 0.5$ and $x^{(k)} = 1$ are reported in Table 8.4. Figure 8.10 (panel a) shows the AIC values for all fitted models. Clearly, the model with $x^{(k)} = 0.5$ has the best fit to the data. The AIC for this model is equal to 158.7 smaller than 175.7, the AIC for the fixed effects model with $x^{(k)} = 0.5$ discussed in the previous section. (marked with plus in panel a).
The full Bayesian model with distinct uniform prior for $x^{(k)}$ could not be implemented due to computational problems. Instead, we use continuous uniform distribution, $x^{(k)} \sim U(0.25, 25)$. For the random effects we assumed $u_\ell \sim N(0, \tau_u^2)$, $\ell = 1, \ldots, 21$. Panel b in Figure 8.10 shows the density estimate for the posterior distribution of $x^{(k)}$. It revels a bimodel with one mode at 0.5 and another mode at 1.25. The posterior mean of $x^{(k)}$ is equal to 0.989. Posterior mean for $\theta(x)$ are shown in Figure 8.6. The three point which marked with arrows in panel a are all outcomes of the same study (the Marinier study). This could be the explanation for the second mode at 1.25. Note that since there are not design points between 0.85 to 1.5, the second mode indicates that the changepoint occurs after 0.85 years and before 1.5 years. This could be simply the influence of the difference between the observed proportions at these design points ($\hat{\theta}_{ML} = 0.86$ at 0.85 years $\hat{\theta}_{ML} = 0.42$ and 0.25 at age 1.5).

8.6 Monitoring Convergence and Model Criticism

Monitoring convergence is needed whenever MCMC simulation is used to approximate the posterior distribution. A major difficulty when MCMC is implemented, is that a random walk can stay in one region for many iterations depending on the starting values. In that cases inference about the parameters is not appropriate. Cowles and Carlin (1996), Gelman and Rubin (1992) and Gelman (1996) discussed different approaches to assess convergence. In this section we applied two methods, the first uses one chain to assess convergence (Geweke's diagnostic) and the second uses parallel chains (Gelman and Rubin). The trace plots for all parameters are shown in Figures 8.7 and 8.8. The first shows the trace plots when the chains are monitored every iteration, the autocorrelation functions indicate on

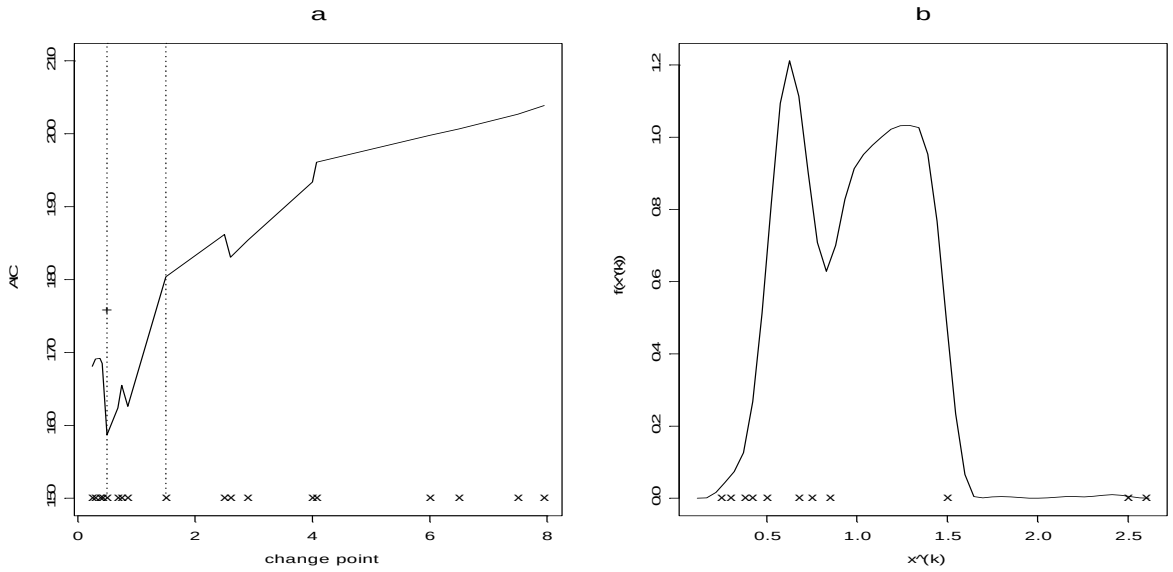


Figure 8.5: *Panel a: AIC values for GLMMs fitted using SAS procedure NLMIXED. The design points are marked with "x' at the bottom of the plot. The plus is marked the AIC value of the fixed effects model with* $x^{(k)} = 0.5$. *Panel b: density estimate for the posterior distribution of* $x^{(k)}$. *Design points are marked with "x" at the bottom of the plot.*

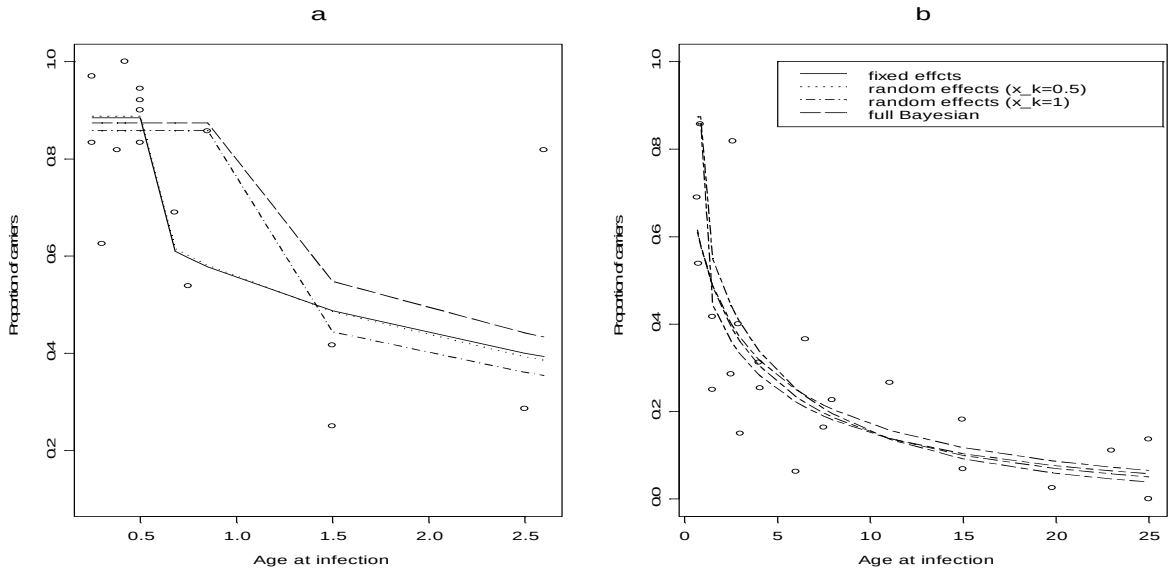


Figure 8.6: *Panel a: Predicted values for* $\theta(x)$ *for age 0.5-2.6 years obtained from the different models. Solid line: fixed effects model, dotted line: random effects model with* $x^{(k)} = 0.5$. *Dotted-dashed line: random effects model with* $x^{(k)} = 1$. *Long-dashed line - full Bayesian model. Panel b: Predicted values for* $\theta(x)$ *when age ranging between 2.6 and 25 years.*

Table 8.4: *Random effects models.*

Parameter	GLMM	GLMM	Full Bayesian
$\hat{\beta}_1$	-2.1315(0.236)	-1.8765 (0.210)	-2.005 (0.247)
$\hat{\beta}_2$	-0.5291(0.187)	-0.3893 (0.257)	-0.751 (0.255)
$\hat{\beta}_3$	0.5040(0.088)	0.4462 (0.119)	0.599 (0.112)
$\hat{\tau}_b^2$	0.092(0.05)	0.1402 (0.072)	0.168 (0.106)
$\hat{x}^{(k)}$	0.5	1	0.989 (0.305)
AIC	158.7	162.6	

slow mixing. The autocorrelation decrease faster when the chain are monitored every 5 iteration (in Figure 8.8), but still high correlation is observed up to lag 10.

8.6.1 Geweke's Diagnostic

Geweke's diagnostic method (Geweke, 1992) is based on a single chain, using time series methodology. Geweke's diagnostic procedure consists of comparison of two pieces of the chain, the first piece contains the first n_A iterations of the chain and the second piece contains the last n_B iterations. Let $\bar{\phi}_n^A$ and $\bar{\phi}_n^B$ be the means of the first and the second pieces respectively. If convergence was achieved then we expect that the difference between $\bar{\phi}_n^A$ and $\bar{\phi}_n^B$ will be relatively small. Geweke (1992) suggested to look at the difference between $\bar{\phi}_n^A$ and $\bar{\phi}_n^B$ divided by the asymptotic standard error of the difference (calculated from the spectrum density of each piece). The result is a Z statistic that converges to $\mathrm{N}(0,1)$ as $n \to \infty$. Geweke (1992) recommended to use $0.1n$ iterations in the first piece and $0.5n$ iterations in the second piece. Figure 8.9 shows Geweke's diagnostic values. Each point in the figures represents comparison between two pieces of the chain, the first contains the first $a\%$ of the iterations and the second contains the last $b\%$. The Z scores, for all parameters, are between ± 1.96, this suggests satisfactory convergence for all parameters.

8.6.2 Diagnostic of Gelman & Rubin

Gelman and Rubin (1992) proposed a diagnostic method based on m parallel MCMC simulations, each one of length n. Their diagnostic statistic is based on two sources of variance: the between-chains variance, B, and the within-chain variance, W. Let ϕ the scalar quantity of interest. The mean of the chain is given by $\bar{\phi}_{i\cdot} = \sum_{j=1}^{n} \phi_{ij}/n$ and the overall mean by $\bar{\phi}_{\cdot\cdot} = \sum_{i=1}^{m} \bar{\phi}_{i\cdot}/m$. The between-chain variance and the within-chain

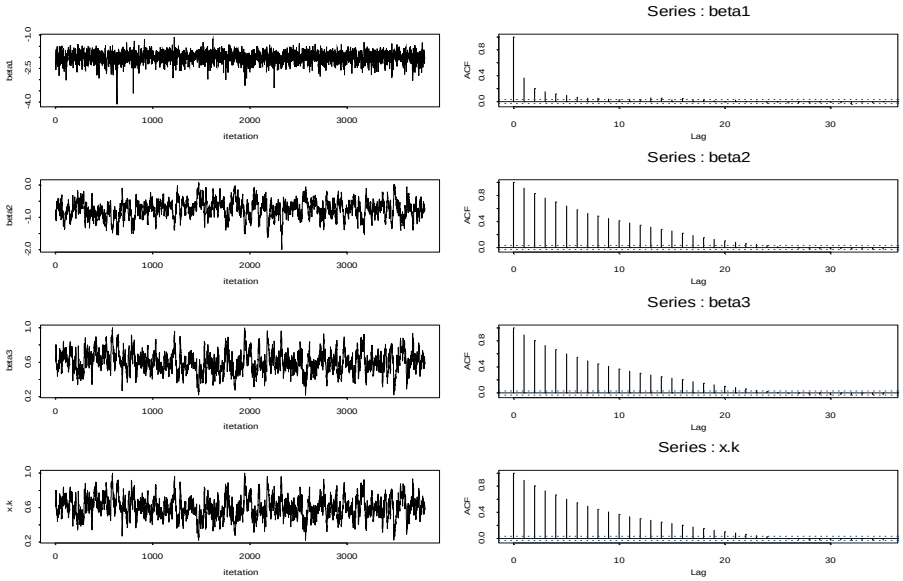


Figure 8.7: *Trace plot and autocorrelations plot for β_1, β_2, β_3 and $x^{(k)}$. Each plot is based on a MCMC sample of 19000 iterations. The chains are monitored every iteration.*

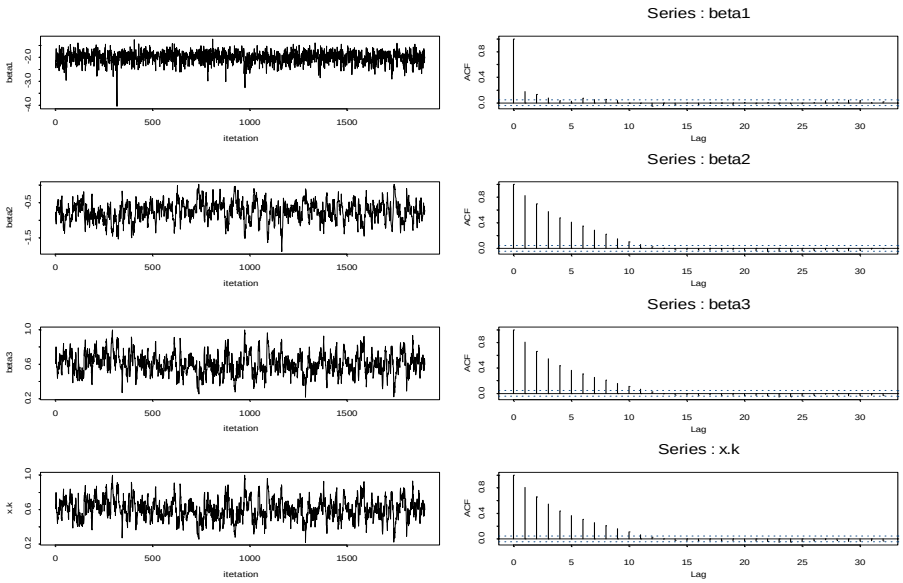


Figure 8.8: *Trace plot and autocorrelatio*

Part II

Modelling Stellar parameters

Chapter 9

Modelling Stellar Atmospheres of Cool Stars: An Introduction

9.1 Introduction

The interior structure of a star can be modelled with a set of differential equations, generally called *the equations of stellar structure* (Collins 1989), which describe the dependency of the state variables of the star, being the pressure, temperature and density on other variables such as the metallicity ([Fe/H]), effective temperature ($T_{\rm eff}$) and gravity ($\log g$). The equations of stellar structure also describe the flow of energy through the star: the energy the star radiates away so profusely from its surface is generally replenished from reservoirs situated in the very hot central region. This requires an effective transfer of energy through the stellar material, which is owing to the existence of a non-vanishing temperature gradient in the star.

Although, many interesting phenomena take place in the stellar interior, all we know about stars rests on the information that we receive from the outer layers of the star — called, by definition, the *atmosphere*. The complete atmosphere can be viewed comprehensively as a transition from the stellar interior to the interstellar medium. In these layers the line and continuum astronomical spectra are formed. So, by studying the stellar spectra, we not only improve our knowledge on the stellar atmosphere itself, but also on structure of the stellar interior. Therefore, astronomers have already been interested for many decades in the construction of reliable *model atmospheres*. The term 'model atmosphere' refers to the construction of mathematical models which provide a description of the dependence of the state variables on parameters as the effective temperature, the gravity or the metallicity. From the model, one often computes the surface flux, called '*the synthetic spectrum*', which gives you the amount of energy radiated away from the surface to the observer in function of the wavelength (or frequency). These synthetic spectra are then compared directly with observational data in order to test the model.

On November 18, 1995 the infrared space observatory (ISO) was launched. With a telescope and four instruments on board, the ISO satellite allowed the astronomical community, for the first time, to observe the cool universe from space in much detail. Before switched off, on May 16, 1998, the ISO satellite made more than 26000 observations.

The infrared radiation of stars was collected by the ISO telescope and was analyzed on

Table 9.1: *Resolution and factors used to shift the sub-bands.*

sub-band	wavelength range [μm]	resolution	factor
1A	2.38 – 2.60	1300	1.007
1B	2.60 – 3.02	1200	1.013
1D	3.02 – 3.52	1500	1.018
1E	3.52 – 4.08	1000	1.011

board by one of four instruments. In this thesis we focus on one of these instruments, the short wavelength spectrometer (SWS). The wavelength covered by the SWS was split into 12 bands, for the analysis presented in the following chapters we use data from 4 bands: 1A, 1B, 1D and 1E. The wavelength and resolution for each band is given in Table 9.1.

The aim of this chapter is to introduce the terminology that will be used throughout the second part of the thesis and to state the scientific questions that will be investigated in the following chapters. We will start with a general description of the main concepts. A more elaborated discussion will be given in later chapters. In Section 9.2 we will describe the problem of estimation of the atmosphere parameters.This will be followed by a description of the observed spectrum (Section 9.3) and the synthetic spectrum (Section 9.4). Model selection criterion, based on the Kolmogorov-Smirnov statistic, will be discussed in Section 9.5.

9.2 Estimation of Stellar Parameters

There are two general approaches to the observational study of stellar atmospheres: analysis and synthesis. Analysis entails measuring detailed features of the spectrum (the ISO-SWS data) being studied and then deducing the parameters of the stellar atmosphere. Synthesis entails specifying atmospheric parameters and calculating the resulting theoretical spectrum: when the synthetic and observed spectra agree, the parameters of the stellar atmosphere are considered to be known.

Throughout this part of the thesis, we focus on the three most important stellar parameters and treat the rest as known. Let $\Omega = (\text{ Teff, log g, [Fe/H]})$ be the parameters of the stellar atmosphere representing the effective temperature, the logarithm of the gravity, and the metallicity respectively. A synthetic spectrum, $\theta^{(m)}$, $m = 1, 2, \ldots, M$, is calculated for specific values of the parameters, $\Omega^{(m)} = (\text{ Teff}^{(m)}, \text{log g}^{(m)}, [\text{Fe/H}]^{(m)})$ and compared with the observed spectrum. The problem is to choose the "best" synthetic spectrum among all spectra under consideration. Figure 9.2 shows a schematic diagram of the procedure. The process of estimating stellar parameters consists of two initial steps: (1) data reduction

of the observed spectrum and (2) the calculation of the collection of synthetic spectra. In the third step, the reduced observed spectrum and the synthetic spectra are compared. The first question which arises when applying this kind of method is how to measure the goodness-of-fit between observed and synthetic spectra. A second important question is then how to assess the uncertainties on the derived stellar parameters.

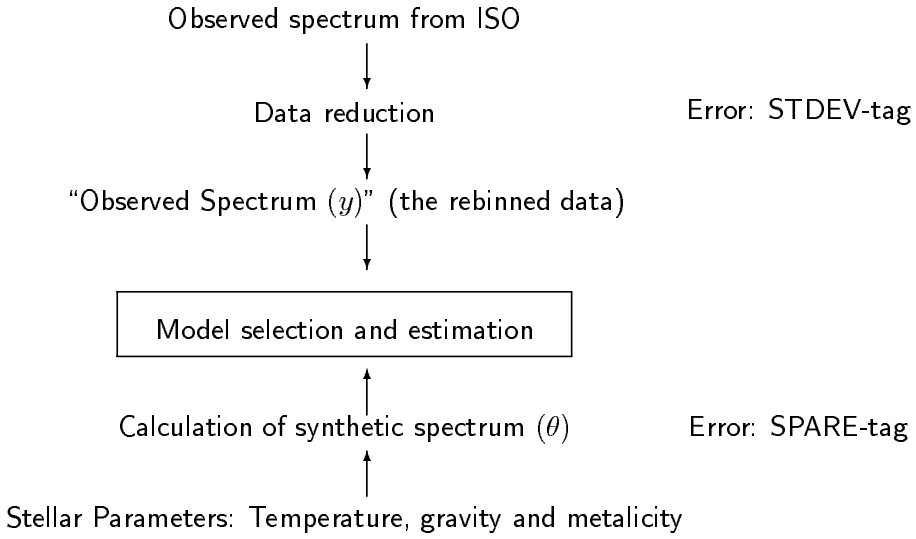


Figure 9.1: *Estimation of stellar parameters. 'STDEV-tag' is the measurement error due to the rebinned procedure. 'SPARE-tag' represents the uncertainty due to the calculation of the synthetic spectrum. By data reduction we mean the process that was used to reduce the raw non rebinned data to the rebinned interpretable data.*

9.3 ISO-SWS Data - The Observed Spectrum

The raw data from the ISO-SWS are reduced using a standard procedure (standard for the astronomical community), called ISO-SWS calibration, which takes into account the resolution in each sub band. The output of the ISO-SWS calibration leads to the reduced data. These data, also called the non rebinned data, are treated as the observed spectrum. Conclusions about the star's atmosphere are based on these data. Figure 9.2 shows the nonrebinned and rebinned data in band 1A of the star alpha Bootis. More details about the data reduction are given in Section 10.1.2 in the next chapter.

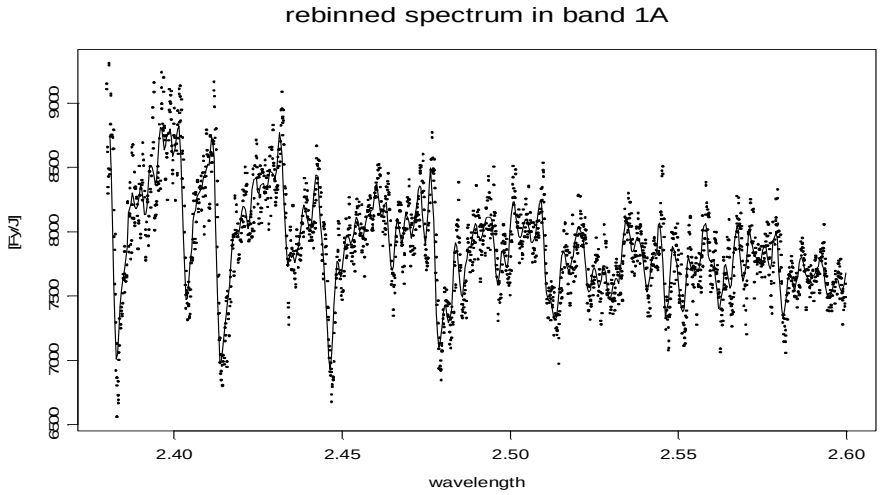


Figure 9.2: *Nonrebinned and rebinned data in band 1A of alpha Bootis.*

9.4 The Collection of Synthetic Spectra

The synthetic spectra used in this study have been generated using model photospheres calculated with the MARCS code, version May 1998. This version is a major update of the MARCS model-photosphere programs first developed by Gustafsson *et al.* (1975). The common assumption of spherical stratification in homogeneous stationary layers, hydrostatic equilibrium and Local Thermodynamic Equilibrium (LTE) were made. Energy conservation was required for radiative and convective flux, where the energy transport due to convection was treated through a local mixing-length theory.
A set of synthetic spectra has been computed over a grid of discrete values in $\Omega = (\mathrm{T_{eff}}, \log \mathrm{g}, [\mathrm{Fe/H}])$. Based on the results reported in (Decin 2000d), the parameter ranges were taken as follows:

$\mathrm{T_{eff}} : 4160K - 4230K - 4300K - 4370K - 4440K,$
$\log \mathrm{g} : 1.20 - 1.35 - 1.50 - 1.65 - 1.80,$
$[\mathrm{Fe/H}] : 0.00 - \ -0.15 - \ -0.30 - \ -0.50 - \ 0.70.$

Each synthetic spectrum was calculated for a different combination of atmospheric parameters. In total, there are thus 125 models. Each of the models got its own model number, specified in Table 9.2. Figure 9.4 and Figure 9.3 show the rebinned data and the synthetic spectra of models 67 and 125.

Table 9.2: *Model numbers associated with the different model parameters of the grid of synthetic spectra.*

log g	[K]					
	4160	4230	4300	4370	4440	
log g = 1.20	1	26	51	76	101	
log g = 1.35	6	31	56	81	106	
log g = 1.50	11	36	61	86	111	[Fe/H] = −0.70
log g = 1.65	16	41	66	91	116	
log g = 1.80	21	46	71	96	121	
log g = 1.20	2	27	52	77	102	
log g = 1.35	7	32	57	82	107	
log g = 1.50	12	37	62	87	112	[Fe/H] = −0.50
log g = 1.65	17	42	67	92	117	
log g = 1.80	22	47	72	97	122	
log g = 1.20	3	28	53	78	103	
log g = 1.35	8	33	58	83	108	
log g = 1.50	13	38	63	88	113	[Fe/H] = −0.30
log g = 1.65	18	43	68	93	118	
log g = 1.80	23	48	73	98	123	
log g = 1.20	4	29	54	79	104	
log g = 1.35	9	34	59	84	109	
log g = 1.50	14	39	64	89	114	[Fe/H] = −0.15
log g = 1.65	19	44	69	94	119	
log g = 1.80	24	49	74	99	124	
log g = 1.20	5	30	55	80	105	
log g = 1.35	10	35	60	85	110	
log g = 1.50	15	40	65	90	115	[Fe/H] = 0.00
log g = 1.65	20	45	70	95	120	
log g = 1.80	25	50	75	100	125	

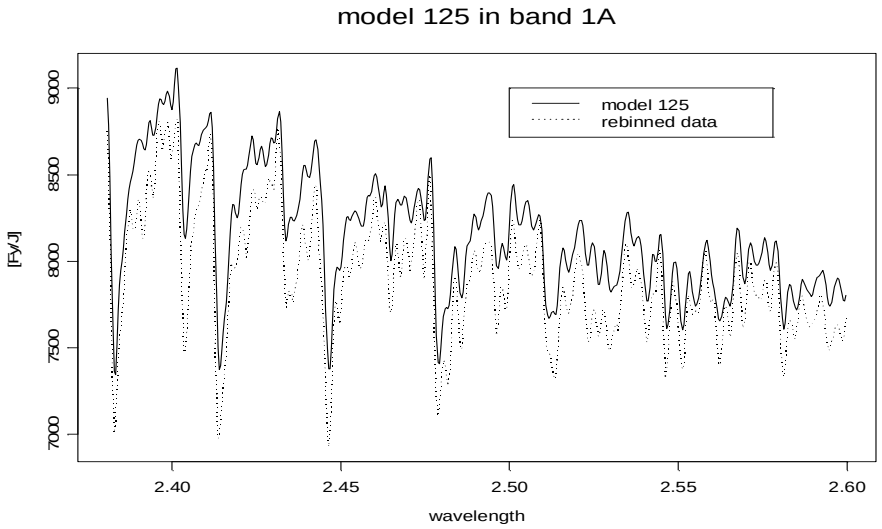


Figure 9.3: *Rebinned data and model 125 in band 1A of alpha Bootis.*

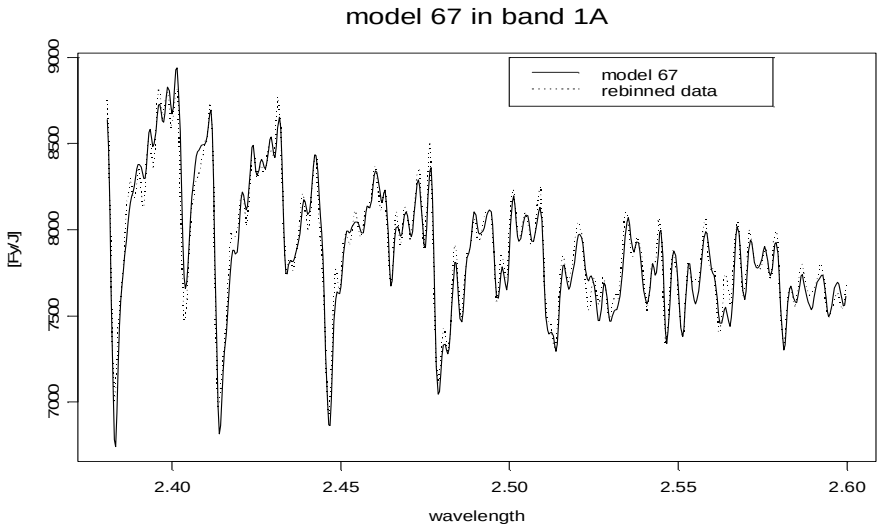


Figure 9.4: *Rebinned data and model 67 in band 1A of alpha Bootis.*

9.5 Model Selection in Decin (2000)

The third step of the estimation procedure requires the selection of the "best" synthetic model among the 125 models under consideration. Decin (2000a, 2000b) proposed to select the synthetic model which minimizes a goodness-to-fit criterion. As a goodness-to-fit measure, Decin (2000a, 2000b) suggested to use the Kolmogorov-Smirnov statistic. Let $y_i = y(t_i)$, $i = 1, \ldots, n$, be the rebinned data in wavelength t_i and let $\theta_i^{(m)}$, $m = 1, \ldots, M$,

be a specific synthetic spectrum. The Kolmogorov-Smirnov statistic is given by

$$\beta^{(m)} = \sqrt{n} \sup_{1 \le k \le n-1} \left| \frac{\sum_{i=1}^{k} \frac{y_i}{\theta_i^{(m)}}}{\sum_{i=1}^{n} \frac{y_i}{\theta_i^{(m)}}} - \frac{k}{n} \right| . \tag{9.1}$$

We define

$$V_i \equiv \frac{y_i}{\theta_i^{(m)}}, \tag{9.2}$$

and

$$Y_k \equiv \frac{\sum_{i=1}^{k} V_i}{\sum_{i=1}^{n} V_i}. \tag{9.3}$$

The Kolmogorov-Smirnov statistic can be rewritten as

$$\beta^{(m)} = \sqrt{n} \sup_{1 \le k \le n-1} \left| Y_k - \frac{k}{n} \right| . \tag{9.4}$$

We further define $\beta_k = |Y_k - k/n|$ such that $\beta^{(m)} = \sqrt{n} \sup_{1 \le k \le n-1} |\beta_k|$. Decin (2000) proposed to select the model which the minimizes the value of $\beta^{(m)}$. The values of the Kolmogorov-Smirnov statistic are 0.0239 and 0.0618 for model 67 and 125 respectively, indicating that model 67 has better goodness-of-fit in band 1A than model 125. This can be seen in Figure 9.5 which plots the values of β_k for the two models. Note that at the peaks, the values of β_k for model 125 are about 2.5 higher than the values of β_k for model 67. This is another indication for the difference in goodness-of-fit of the two models. Figure 9.6 plots the ratio y_i/θ_i (or V_i) for both models. Note how the values of model 67 are distributed around the value of 1 (which indicates that the model fits well) while the values for model 125 are always below 1 with an upward trend (as wavelength increases).

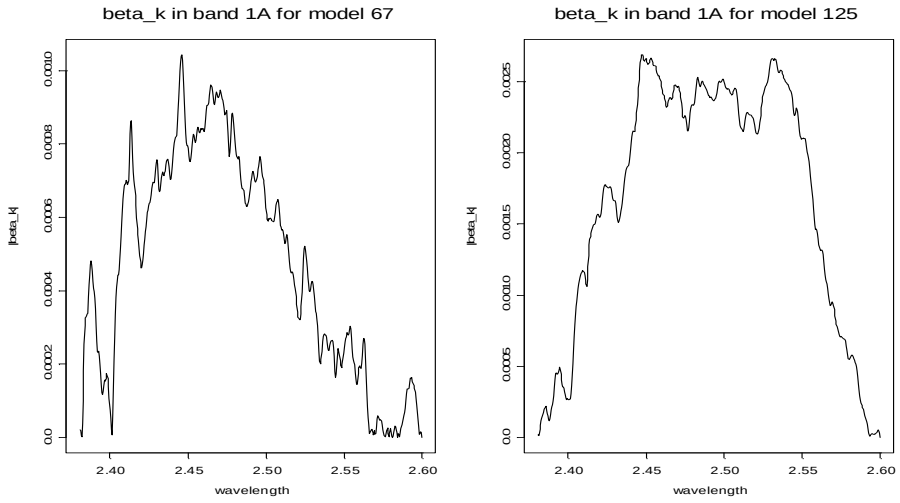


Figure 9.5: *β_k for model 67 (left panel) and model 125 (right panel) in band 1A.*

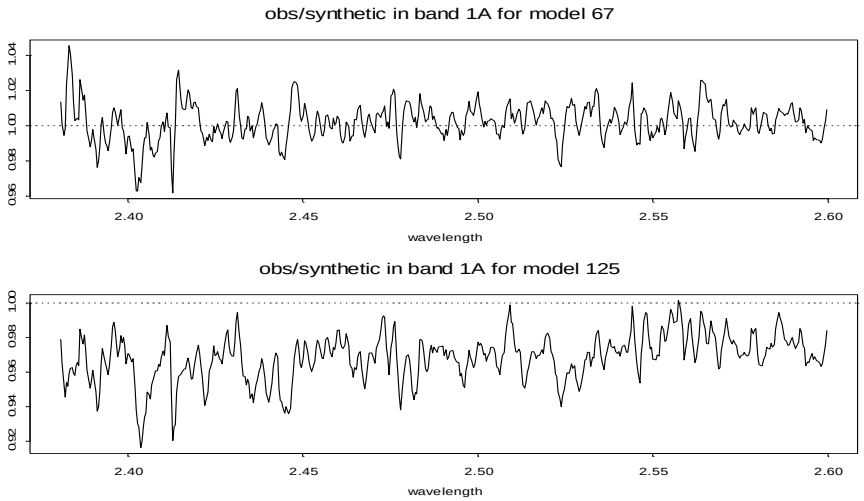


Figure 9.6: *Model 67 (upper panel) and 125 (lower panel).* $\frac{observed\ spectrum}{synthetic\ spectrum}$ *in band 1A.*

9.6 Summary

Several issues will be addressed in the following chapters. First, it will be investigated how sensitive the model selection is for the rebinning procedure. In other words, we will investigate how much the results change when a different method is used to summarize the nonrebinned data (Chapter 10). Second, the Kolmogorov-Smirnov statistic will be compared with the least square criterion (Chapter 10). Third, the Kolmogorov-Smirnov statistic is a measure for global goodness-of-fit, in Chapter 11 we will investigate local properties of the Kolmogorov-Smirnov statistic. Fourth, the rebinned data are a summary of the observed raw spectral data. As such, the rebinned data are subject to measurement error. Another source of error is the accuracy of the SWS itself. These two types of error will be taken into account in Chapters 12 and 13 where we will shift the analysis into the hierarchical Bayesian framework and will use the two sources of variation in the likelihood and the prior model for the spectrum.

Chapter 10

Estimating Stellar Parameters - Nonparametric Estimate for the Spectrum

10.1 Introduction

The use of Kolmogorov-Smirnov statistic to compare the rebinned data with the synthetic spectrum is the state-of-art method in modeling stellar atmospheres. In this chapter, we present a method to determine confidence intervals around the observed spectrum using smoothing splines as an alternative "summary" of the non rebinned data (Section 10.2). This will allow us to perform a sensitivity analysis in order to investigate the sensitivity of the results to the use of both the Kolmogorov-Smirnov statistic and the rebinned data. Using the $2.38 - 4.08\,\mu$m ISO-SWS spectrum of the K2IIIp giant Alpha Bootis (Arcturus, HD 124897), it will be demonstrated, in Section 10.2 that using splines or the rebinned data lead for these infrared data to the same conclusions concerning the derived stellar parameters. In Section 10.3 we will compare the performance of the Kolmogorov-Smirnov statistic to a goodness-of-fit measure based on the least square criterion.
For the remainder of this first section, we give a general description of the model selection criterion used (Section 10.1.1). This is followed by a description of the observational and synthetic data on which the method will be illustrated (Section 10.1.2).

10.1.1 Estimation

Independent of the summary for the observational data, it is natural to estimate Ω with $\Omega^{(\star)}$ for which the synthetic spectrum $\theta^{(\star)}$ is the "closest" to the (summarized) observed spectrum, f. By analogy to linear regression we can estimate Ω by minimizing the residual sum of squares

$$T^{(m)}(f,\theta^{(m)}) = \frac{1}{n}\sum_{t=1}^{n}(f(t)-\theta^{(m)}(t))^2 \quad (10.1)$$

$$= \frac{1}{n}\sum_{t=1}^{n}(\delta^{(m)}(t))^2, \quad (10.2)$$

where $\delta^{(m)}(t)$ is the difference between the observed spectrum f and the m'th synthetic spectrum $\theta^{(m)}$, $\delta^{(m)}(t) = f(t)-\theta^{(m)}(t)$. Hereafter, we refer to this difference as the residual at wavelength t. Hence, the minimizer of $T^{(m)}(f,\theta^{(m)})$ is the least squares estimator for Ω. In practice one can minimize equation (10.2) with a search over a sensitive grid for the parameter vector Ω.

10.1.2 Observational and Synthetic Data

Observational Data y

The observational data for this study consist of near-infrared ($2.38 - 4.08\,\mu$m) spectra of α Boo observed with the SWS (Short Wavelength Spectrometer) on board ISO (Infrared Space Observatory). The spectrometer was used in the SWS observing mode AOT01 (= a single up-down scan for each aperture with four possible scan speeds at degraded resolution) with scanner speed 4, resulting in a resolving power of ≈ 1500. The observation lasted for 6538 sec and was performed during revolution 452[1].

It is important to mention here that from astronomical point of view the rebinned data are treated as the observed spectrum.

The individual sub-band spectra can show jumps in flux level at the band-edges when combining them into a single spectrum. These band-to-band discontinuities can have several causes: uncertainties in flux calibration, the low responsivity at the band edges, pointing errors, and a problematic dark current subtraction in combination with the RSRF (Relative Spectral Response Function) correction, from which the pointing errors are believed to have the largest impact for this high-flux observation. Hence, the individual sub-bands were multiplied with a factor to construct a smooth spectrum (see Table 9.1). These factors were determined by using the SED (Spectral Energy Distribution) of α Boo as constructed in (Decin 2000b) as reference.

Synthetic Data θ

The synthetic spectra used in this thesis have been generated using model photospheres calculated with the MARCS code, version May 1998. For complete discussion of the astronomical aspects of this issue we refer to Decin (2000).

[1]Each observation is determined uniquely by its observation number (8 digits), in which the first three digits represent the revolution number. The observing data can be calculated from the revolution number which is the number of days after 17 November 1995.

The common assumption of spherical stratification in homogeneous stationary layers, hydrostatic equilibrium and Local Thermodynamic Equilibrium (LTE) were made. Energy conservation was required for radiative and convective flux, where the energy transport due to convection was treated through a local mixing-length theory. The mixing-length l was chosen as $1.5\,\mathrm{H_p}$, with $\mathrm{H_p}$ the pressure scale height. Turbulent pressure was neglected. Using the computed model atmospheres, the synthetic spectra were generated by solving the radiative transfer at a high wavelength resolution ($\Delta t \sim 1\,\mathrm{km/s}$, corresponding to $t/\Delta t \sim 330\,000$). With a microturbulent velocity $\xi_t \sim 2\,\mathrm{km/s}$, this means we are sure to sample all lines in the atomic and molecular database in the generation of the synthetic spectrum. This is necessary in order not to overestimate the absorption in regions with a high line density, or vice versa to underestimate it in regions with a low line density. For the line opacity in the ISO-SWS range a database of infrared lines including atoms and molecules has been prepared. For the molecular lines, the same data have been used as in (Decin 2000). The accuracy and completeness of these line lists are discussed in (Decin 2000).

10.2 Estimating the Observed Spectrum Using Smoothing Splines

Since the grid of observational pixel values does not have a fixed resolution, we first want to "summarize" the observational pixel values, and then make a comparison between this summary (denoted as f) and a synthetic spectrum (θ) with the same resolution. The standard way to summarize the input data is by 'rebinning' as explained before. For summarizing the ISO-SWS data to the resolution as specified in Table 9.1, we have applied a flux conserving nonparametric rebinning method — i.e. for each bin the flux value is calculated using the trapezoidal rule — with an oversampling of 4. This means that the used resolution bin is 4 times the grid separation determined by the resolution for a specific wavelength range of the ISO-SWS data. In order to fully recover the intervening flux values it can be shown in the context of 'rectangular filtering' that taking 4 points in an interval of length Δt is enough in order to optimize the signal-to-noise (S/N) ratio (Bracewell 1985).

We consider the model

$$y_i = \mu_i + \varepsilon_i, \tag{10.3}$$

and where $y_i = y(t_i)$ is the observed spectrum at wavelength t_i, $i = 1, \ldots, n$, m_i is the "true" spectrum at wavelength t_i, $\mu_i = \mu(t_i)$ and $\varepsilon_i \sim N(0, \sigma^2)$.
An alternative "summary" of the observed data pixels is given in the context of nonparametric regression. Fitting a model for the observed data pixels within the framework of non-parametric regression allows us to estimate the spectrum without parametric restriction on the mean structure (i.e., we do not need to specify any deterministic relationship between Ω and μ). In particular, we use smoothing splines to estimate $\mu(t)$. Hence, we assume that $\mu(t)$ has at least two continuous derivatives (denoted as $\mu''(t)$) over the wavelength range considered (Whaba 1990). It is true that in some cases smoothed data should not be used for quantitative analysis, but we will show that in our situation the results obtained from smoothing splines or from rebinned data lead to the same

conclusions.

We estimate the spectrum by minimizing the penalized least square criteria (Green and Silverman 1994) given by

$$S(\gamma) = \Sigma_{i=1}^{n}(y_i - \mu(t_i))^2 + \gamma \int (\mu''(t))^2 dt. \tag{10.4}$$

For a given value of γ, the minimizer of equation (10.4), say $f_\gamma(t)$, is a smoothing cubic spline (Whaba 1990). The parameter γ is called the smoothing parameter and it controls the trade-off between the goodness-of-fit of $f_\gamma(t)$ (measured by the residuals sum of squares) and smoothness of $f_\gamma(t)$ (measured by the squared integral). In practice the value of γ is unknown but can be estimated using the cross validation method or the generalized cross validation method (Green and Silverman 1994). Note that the solution for equation (10.4) is defined on the grid of non rebinned data points. Therefore, in order to calculate $T(f, \theta^{(m)})$, we need to calculate the predicted values in the same wavelength grid of the synthetic spectrum. This can be done by using equation (2.22) in (Green and Silverman 1994). An elaborate discussion on smoothing splines is given in Chapter 12.

10.2.1 Confidence Intervals For the Spectrum

One advantage of estimating the spectrum by smoothing splines is related to the calculation of *confidence intervals* around the estimated spectrum. These can then be used to evaluate the variability around the estimated spectrum. Smoothing splines belong to the family of linear smoothers and therefore they can be expressed as

$$\hat{\boldsymbol{f}}_\gamma = \boldsymbol{A}(\gamma)\mathbf{y}. \tag{10.5}$$

Here, $\hat{\boldsymbol{f}}_\gamma = (\hat{f}(t_1), \ldots, \hat{f}(t_n))$, $\mathbf{y} = (y_1, \ldots, y_n)$. $\boldsymbol{A}(\gamma)$ is called the hat (or smoothing) matrix and it depends on the value of γ. Indeed, the hat matrix maps the observational data to their predicted values f_γ, i.e. to the observed spectrum. Details about the structure of the hat matrix for smoothing splines are given in (Green and Silverman 1994). Whaba (1978,1982) showed that a 95% confidence interval for $\mu(t)$ is given by $\hat{f}_\gamma(t) \pm 1.96 \times \hat{\sigma} \times \sqrt{a(\gamma)}$ where $a(\gamma)$ are the diagonal elements of $A(\gamma)$ and $\hat{\sigma}$ is the estimate for σ, $\hat{\sigma} = \sum_{i=1}^{n}(y_i - f_\gamma(t_i)^2/(n - tr(\boldsymbol{A}(\gamma))$.

10.2.2 Application to the Data

In this section, we will apply the method as described in previous section, to the infrared 2.38 – 4.08 μm ISO-SWS data of the star α Boo. In order to test the appropriateness of using smoothing splines instead of the rebinned data, we will compare the obtained "summarized" spectra with a grid of synthetic spectra using $T^{(m)}(f, \theta^{(m)})$ as defined in equation (10.2). A collection of 125 synthetic spectra has been calculated over a grid of discrete values in Ω as described in Chapter 9. Each synthetic spectrum was thus calculated for a different combination of the atmospheric parameters. All models have their own model number as specified in Table 9.2.

Since each sub-band has its own characteristics as explained in Section 10.1.2 the observed spectrum was estimated (by rebinning or by using smoothing splines) separately for each sub-band. We then have calculated $T^{(m)}(f_r, \theta^{(m)})$ and $T^{(m)}(f_s, \theta^{(m)})$ over the grid of synthetic spectra, with f_r and f_s being respectively the rebinned data and the smoothing spline. As an example, we display both the rebinned data and the smoothing spline with the 95 % confidence intervals for band 1A in Figure 10.1. We note that the rebinned data and the smoothing spline do show the same patterns.

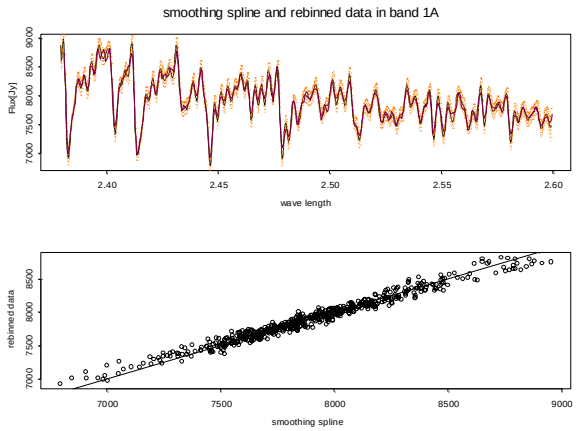


Figure 10.1: *Band 1A. Upper panel: rebinned data and* 95% *confidence intervals for* $\mu(t)$ *based on the smoothing spline. Lower panel: smoothing splines versus the rebinned data in band 1A. The straight line is the bissector.*

Band 1A

Figure 10.2 shows the values of $T^{(m)}(f_s, \theta^{(m)})$ (on a log scale). The vertical lines separate the 125 models by T_{eff}.

Model 67 has the best goodness-of-fit in band 1A. We note that within one temperature level, the models occur in groups of size 5 (according to the value of the gravity). For example, models 1, 2, 3, 4 and 5 have the same effective temperature, being 4160 K, and the same gravity ($\log g = 1.2$) while for models $6 - 10$ $\log g = 1.35$. Trends in the goodness-of-fit are visualized in Figure 10.3. Three patterns are observed: (a) the goodness-of-fit increases with the level of metallicity, (b) a parabolic shape in which the best goodness-of-fit is achieved for models with metallicity between -0.15 to -0.5, and (c) goodness-of-fit decreases with the level of metallicity. For a fixed temperature value, the trend changes more or less from trend (a) via trend (b) to trend (c) when the gravity increases. Sometimes, a trend occurs twice or is absent, but never the order in trends changes. The model having the best goodness-of-fit is always situated at the minimum of a parabolic shape, suggesting that we have reached a local minimum — an equilibrium — in the parameter space.

Table 10.1 shows the 5 models with $T^{(m)}(f, \theta^{(m)})$ having the lowest values, i.e., they have

Table 10.1: *Top 5 models in band 1A. Upper row: ranks based on smoothing spline. Lower row: ranks based on rebinned data.*

Smoother	Model (rank)	T_{eff}	log g	[Fe/H]
smoothing	67 (1)	4300	1.65	−0.50
spline	91 (2)	4370	1.65	−0.70
	43 (3)	4230	1.65	−0.30
	87 (4)	4370	1.50	−0.50
	92 (5)	4370	1.65	−0.50
rebinned	67 (1)	4300	1.65	−0.50
data	43 (2)	4230	1.65	−0.30
	91 (3)	4370	1.65	−0.70
	88 (4)	4370	1.50	−0.30
	92 (5)	4370	1.65	−0.50

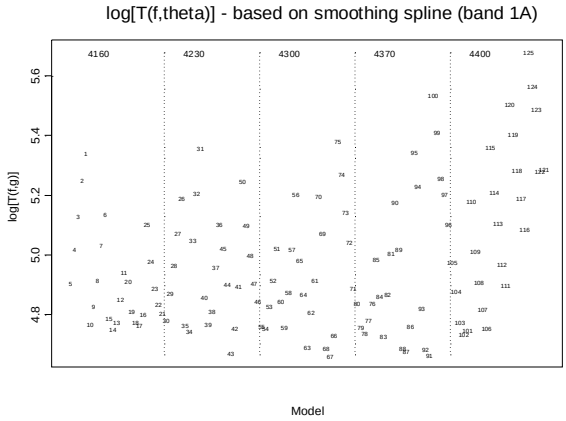


Figure 10.2: $\log T^{(m)}(f_s, \theta^{(m)})$ *(based on smoothing splines) versus the model numbers. The vertical lines separate the models according to the temperature.*

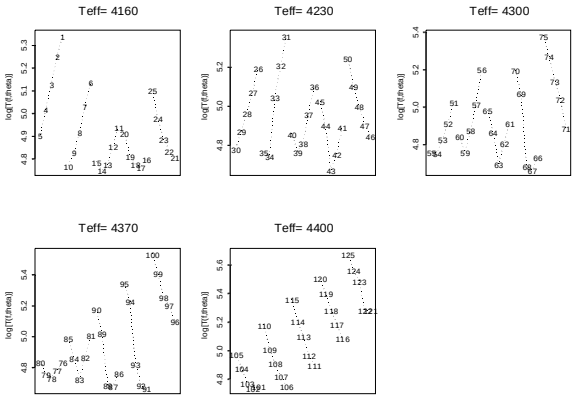


Figure 10.3: *Trends in the goodness-of-fit condition of* $\log T^{(m)}(f_s, \theta^{(m)})$. *The model numbers are as specified in Table 9.2.*

the best goodness-of-fit (upper row based on the smoothing spline, lower row based on the rebinned data). For 4 out of the 5 models log g is 1.65 dex, while the effective temperature ranges between 4230 K to 4370 K. [Fe/H] is between -0.30 dex and -0.70 dex. The second row in Table 10.1 shows the top 5 models in band 1A when $T^{(m)}(f, \theta^{(m)})$ was calculated based on rebinned data. Note that except for the order, the results are the same and the only difference is that model 87 was replaced with model 88 having a somewhat higher metallicity.

As mentioned in Section 10.1, one aim of the analysis is to investigate the behaviour of $T^{(m)}(f, \theta^{(m)})$ when $f(t)$ is either represented by a smoothing spline or by the rebinned data. Figure 10.4 shows the values of $T^{(m)}(f_s, \theta^{(m)})$ versus $T^{(m)}(f_r, \theta^{(m)})$. Clearly, theoretical spectra that fit well when compared to the smoothing spline show also a good goodness-of-fit when being compared to the rebinned data.

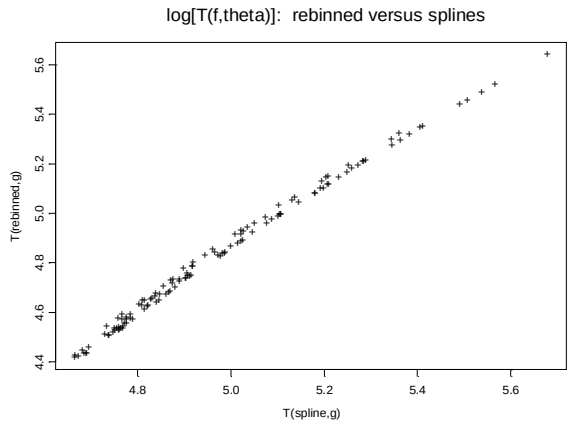


Figure 10.4: *log* $T(f, g^{(m)})$*, splines versus the rebinned spectrum.*

Overall Goodness-Of-Fit

While in previous subsection we have concentrated on band 1A of the ISO-SWS data of α Boo, we now will take the whole 2.38 – 4.08 μm wavelength range into account. Band 1A has a very characteristic footprint, determined by the first overtone CO ($\Delta v = 2$) vibration-rotation bands in this wavelength range (Decin 2000). Molecules absorbing in bands 1B, 1D, and 1E are mainly OH and SiO, while also some atomic features are visible. The absorption pattern of these last molecules is however not as pronounced as for CO ($\Delta v = 2$) in band 1A. Although these CO features can give us already quite a good idea of the temperature and the gravity of the target, it is essential to use the whole 2.38 – 4.08 μm wavelength in order to minimize the uncertainties on the stellar parameters being studied. This is due to the fact that all of these molecular and atomic absorption features have their own characteristic dependence on the atmospheric parameters (e.g., see Decin 2000). Since, however, each sub-band has its own instrumental characteristics, and since the observational data have their largest uncertainties at the band edges (Decin 2002), we will *not* join the whole 2.38 – 4.08 μm wavelength range into 1 spectrum, but we will combine the results obtained from the separate bands.

The values of $T^{(m)}(f, \theta^{(m)})$ were ranked at each band, and for each model we calculate the mean of the ranks (both for the smoothing splines and the rebinned data). This means that the "best" model is the one with the smallest mean rank. For example, model 67 has the lowest value of $T^{(m)}(f_s, \theta^{(m)})$ in band 1A, but this model is ranked 21, 72 and 25 in band 1B, 1D and 1E respectively (hence, the mean rank is 29.75). Overall, the rank of the mean rank of model 67 is 16.5 when smoothing splines are used. The mean ranks of the 125 models (using smoothing splines) are displayed in Figure 10.5. The models with the lowest rank (when smoothing splines are used) are model 63 and 87. Model 63 is ranked 8, 25, 16 and 9 in the 4 bands, respectively, with mean rank being 14.5. When $T^{(m)}(f, \theta^{(m)})$ was calculated based on the rebinned data, model 63 has the smallest mean rank. This model is ranked 8, 16, 6 and 8 in the 4 bands respectively.

Figure 10.6 shows the ranks of the mean ranks of the 125 models based on the rebinned data and smoothing splines (the best is ranked 1 and the worst is ranked 125). Note that

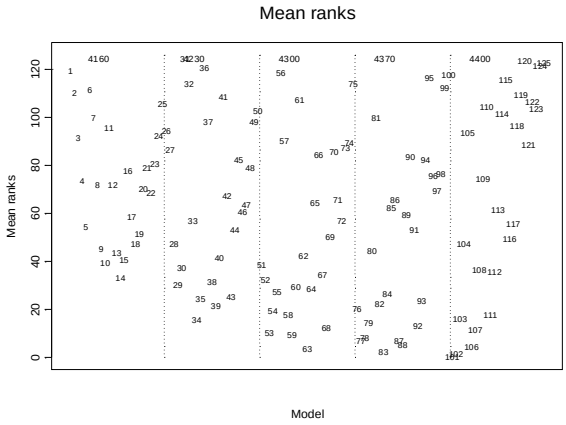


Figure 10.5: *Mean ranks for the 125 synthetic spectrum.* $T^{(m)}(f, \theta^{(m)})$ *is based on smoothing splines.*

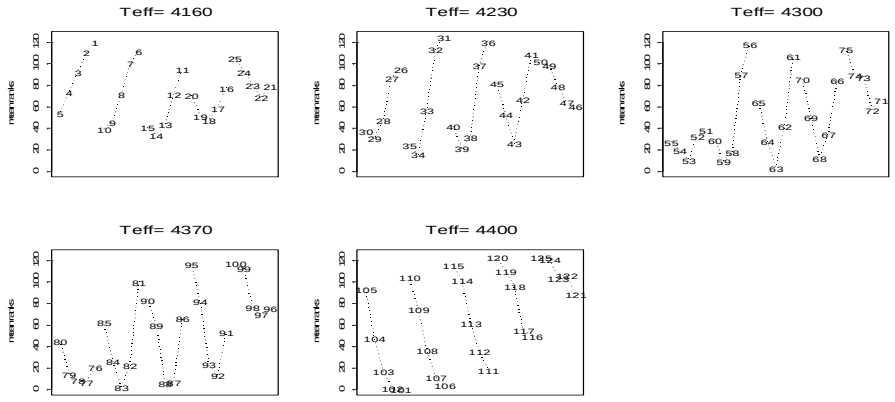


Figure 10.6: *Ranks based on smoothing splines versus ranks based on rebinned data. The bottem left corner is an indication of a better goodness-of-fit. Models in the lower left corner are ranked among the top 10 based on* both *rebinned data and smoothing splines.*

Table 10.2: *The 8 models given in this table are ranked among the top 10 based both on splines and rebinned data. The left rank is the overall rank based on the smoothing spline and the right rank is the overall rank based on the rebinned data.*

Model (ranks)	T_{eff} [K]	log g	[Fe/H]	ranks in band 1a
63 (1.5,1)	4300	1.50	−0.30	(8,8)
87 (1.5,2)	4370	1.50	−0.50	(4,7)
83 (4.5,3.5)	4370	1.35	−0.30	(9,11)
106(4.5,3.5)	4440	1.35	−0.70	(16,17)
92 (3,5)	4370	1.65	−0.30	(5,5)
43 (6.5,6.5)	4230	1.65	−0.30	(3,2)
88 (8,8)	4370	1.50	−0.30	(6,4)
68 (9.5,6.5)	4300	1.65	−0.30	(7,6)

the lower left corner in this figure indicates a better overall goodness-of-fit. We note that the association between the ranks obtained from the two methods is positive (Pearson correlation is 0.96) which means that model selection based on either rebinned data or smoothing splines leads to the *same* conclusions.

10.2.3 Conclusions

The models in the bottom left corner in Figure 10.6 are ranked among the top 10 for both smoothing splines and rebinned data. The stellar parameters for these models are presented in Table 10.2. We note that, except for models 106 and 83, all model are also ranked among the top 10 in band 1A. Based on these models we conclude that the effective temperature ranges between 4230 K and 4440 K, the gravity between 1.35 and 1.65, and the metallicity between −0.7 to −0.3. Note that model 67, ranked first in band 1A (based on smoothing splines and rebinned data), is not in this list. Model 67 has overall ranks of 16 (smoothing splines) and 17 (rebinned data). Our study shows that we can safely replace the rebinned spectrum by a smoothing spline to determine confidence intervals for the parameters.

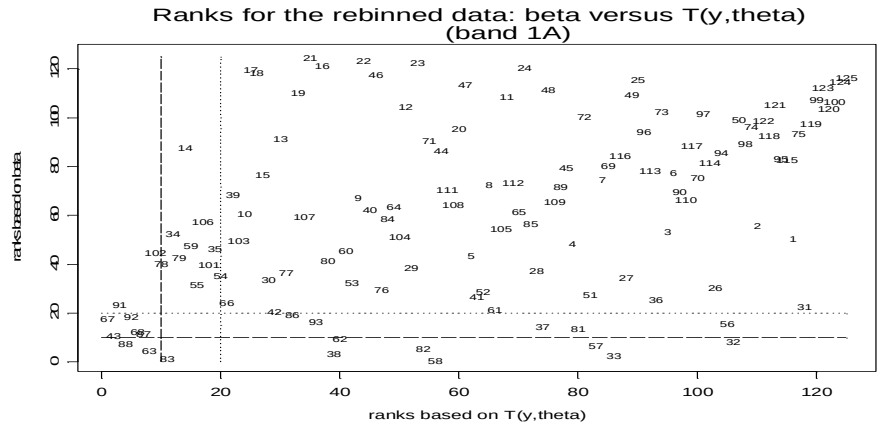


Figure 10.7: *Band 1A. Ranks based on β versus ranks based on $T(y,\theta)$. Ranks are based on rebinned data. The bottom left corner is an indication of a better goodness-of-fit. Models in the lower left corner are ranked among the top 10 (and top 20) based on* both *β and $T(f,\theta)$.*

10.3 Comparison Between Measures for Goodness-Of-Fit

In the previous section, $T(f,\theta)$ was used as a measure for the goodness-of-fit. In this section the analysis discussed above was repeated using the Kolmogorov-Smirnov (β) statistic as a measure for the goodness-of-fit. This statistical test *globally* checks the goodness-of-fit of the observed and synthetic spectra by computing a deviation estimating parameter. Without specifying the distribution function of β, we may summarize that

$$\beta = \sqrt{n} \sup_{1 \le k \le n-1} \left| \frac{\sum_{t=1}^{k} \frac{f(t)}{\theta(t)}}{\sum_{t=1}^{n} \frac{f(t)}{\theta(t)}} - \frac{k}{n} \right| . \tag{10.6}$$

The lower the β-value, the better the accordance between the observed data and the synthetic spectrum. For more details about the use of the Kolmogorov-Smirnov statistic to estimate stellar parameters and their uncertainties we refer to Decin (2000).Hence, the main difference between β and T is that the Kolmogorov-Smirnov statistic β measures a *global* goodness-of-fit, so that local deviations between observations and theoretical data only have a minor influence on the final result, while for $T(f,\theta)$ *local* deviating points are important. Note that a shift in the absolute flux-values (e.g. to simulate a change or uncertainty in the angular diameter) influences T a lot, while β remains almost the same. Since both deviation estimating parameters do stress another point in the goodness-of-fit, a combination of the results based on the two parameters separately can only improve our knowledge on the stellar parameters and their uncertainties.

Figure 10.7 shows the ranks in band 1A. We notice that there is a group of 8 models - 67,92,43,68,87,88,63 and 83 - which have high ranks for both β and $T(y,\theta)$. We calculated the means of the means for the overall goodnes-of-fit. Figure 10.8 shows the ranks of the means based on β versus the ranks based of $T(y,\theta)$. The model with the lowest rank

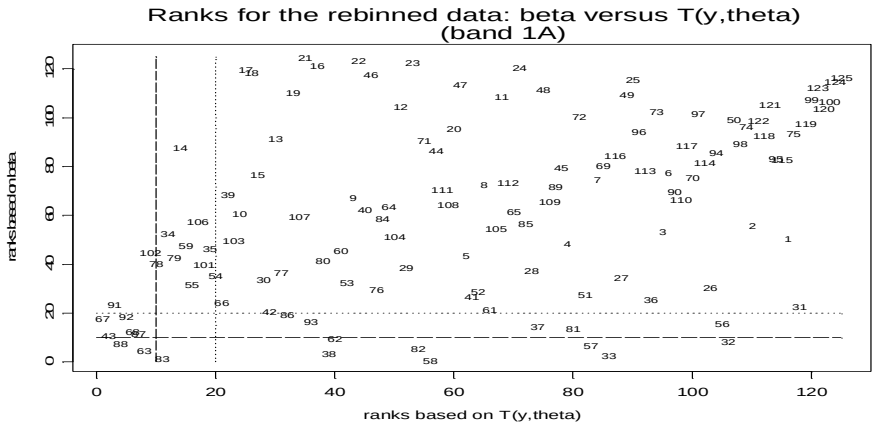


Figure 10.8: *Overall goodness-of-fit. Ranks based on β versus ranks based on $T(f,\theta)$. Ranks are based on rebinned data. The bottom left corner is an indication of a better goodness-of-fit. Models in the lower left corner are ranked among the top 10 (and top 20) based on* both *β and $T(f,\theta)$.*

for β is 93, this model is ranked 39 when the ranks are based on $T(f,\theta)$. Table 10.3 shows the best 5 models, which ranked among the top 20 for both β and $T(f,\theta)$. These 5 models appear in the bottom-left corner in Figure 10.8 which shows the ranks of the mean ranks based on β versus the ranks of the mean ranks based on $T(f,\theta)$. In both cases the rebinned data were used. Note that except model 35, all the other models were identified in the previous section among the best models. Similar results were obtained when β and $T(f,\theta)$ were based on splines.

Note that one indeed can see a correlation between the ranks of the mean ranks based on T and β, but that there are a few outliers mainly in the upper left corner of Figure 10.8. Few models are situated with low T and high β rank of the mean ranks. Inspecting why the deviation estimating parameters do show this trend, learns us that all of these models have a very low rank in T for band 1D and/or band 1E.

10.4 Discussion

Estimation of the stellar parameters requires a reduction of the raw observational data in order to be able to compare the observed spectrum to the synthetic spectrum. Therefore, both the rebinned data and the smoothing spline should be treated as nonparametric smoothers of the observational data, both are the nonparametric estimate for the true spectrum. As such, both are biased to the true spectrum. In Section 10.2 we have shown that the results based on the rebinned data and the smoothing spline are comparable and lead to the same conclusions about the stellar parameters.
In band 1A a clear set of synthetic models were identified by both β and $T(f,\theta)$. Some models which fit well when goodness-of-fit is measured by the Kolmogorov-Smirnov statistic, fit poorly when goodness-of-fit is measured with $T(f,\theta)$. This will be investigated

Table 10.3: *Overall goodness-of-fit. The 5 models given in this table are ranked among the top 20 based both on $T(y,\theta)$ and β.*

Model	rank β	rank $T(y,\theta)$	T_{eff} [K]	log g	[Fe/H]
63	1	14.5	4300	1.50	−0.30
68	6.5	2	4370	1.50	−0.50
43	6.5	4.5	4370	1.35	−0.30
88	8	17	4440	1.35	−0.70
35	18	13	4370	1.65	−0.30

further in the next chapter.

Chapter 11

Estimating Stellar Parameters - Inference With Nonparametric Regression and Model Diagnostics

11.1 Introduction

Within the classical regression framework, estimation is usually followed by inference and model diagnostics. In the previous chapter we focused on model selection and estimation of the stellar parameters. The aim of this chapter is to investigate the local properties of the Kolmogorov-Smirnov statistic and to propose a tool for model diagnostics when this statistic is used for model selection. Choosing the 'best' model (out of a grid) does not necessarily imply that the model is a 'good' representation of the observed data. Hence, when a model is chosen, one can investigate how good the model fits the data. We focus on the variable

$$V_t = \frac{f(t)}{\theta(t)}.$$

Note that V_t was used to construct the Kolmogorov-Smirnov statistic in the previous chapter. Now, if a specific synthetic spectrum is a 'good' model, then we expect that $V_t \approx 1$. That was the main motivation to use the Kolmogorov-Smirnov statistic as a criterion for goodness-of-fit, a good synthetic spectrum implies that $\sum_{t=1}^{k} V_t / \sum_{t=1}^{n} V_t \approx k/n$. Note that if the synthetic spectrum is a good model we expect that a nonparametric smoother (see details in the appendix) of V_t will be flat around 1. This will be a key issue in the following sections. In Section 11.1 we discuss the use of the lack of fit test, proposed by Bowman and Azzalini (1997), in our setting. In Section 11.2 the proposed method is applied to the data.

11.2 Lack-Of-Fit Tests

11.2.1 Test of Hypothesis for the "No Effect" Model

Since the parameter of interest is Ω we wish to test the hypotheses

$$\begin{aligned} H_0 &: \Omega = \Omega_0, \\ H_1 &: \Omega \neq \Omega_0, \end{aligned} \tag{11.1}$$

with Ω_0 representing the "true" stellar parameters of the target being studied. The hypotheses in Eq. (11.1) can be reformulated in terms of the synthetic spectrum,

$$\begin{aligned} H_0 &: \mu(t) = \theta_0(t), \quad \text{for all } t \\ H_1 &: \mu(t) \neq \theta_0(t), \quad \text{for some } t, \end{aligned} \tag{11.2}$$

The rebinned data are used as a summary for the observed spectrum and we assume that $E(f(t)) = \mu(t)$. Thus, the hypotheses in equation (11.2) can be reformulated as

$$\begin{aligned} H_0 &: E(f(t)) = \theta_0(t), \\ H_1 &: E(f(t)) \neq \theta_0(t). \end{aligned} \tag{11.3}$$

Under the null hypothesis in equation (11.3) we expect that $E(V_t) = 1$. Thus, in terms of V_t, we consider two competing models,

$$\begin{aligned} H_0 &: E(V_t) = 1, \\ H_1 &: E(V_t) = \eta(t). \end{aligned} \tag{11.4}$$

Here, $\eta(t)$ is assumed to be a smooth function. The model under H_0 is called the no effect model (Bowman and Azzalini, 1997). In order to test the hypotheses in equation (11.4) one needs to calculate the residuals sum of squares under the two alternatives and to compare between them. Since the mean of V_t under H_0 is 1, the residuals sum of squares under the null hypothesis is

$$RSS_0 = \sum_{t=1}^{n} \{f(t) - 1\}^2 , \tag{11.5}$$

and under H_1

$$RSS_1 = \sum_{t=1}^{n} \{f(t) - \hat{\eta}(t)\}^2 , \tag{11.6}$$

where $\hat{\eta}(t)$ is a linear smoother of V_t. Note that we do not specify any parametric structure for $\eta(t)$ under the alternative in equation (11.4). The underling assumption that we made

is that if a specific synthetic model is not a 'good' model, there is a structure in the rebinned data that this specific model cannot capture. This structure can be captured by the non-parametric smooth function $\hat{\eta}(t)$. In practice, we use the loess method (Cleveland 1979) to model the relationship between V_t and the wavelength. Intuitively, it is clear that for a 'good' synthetic spectrum RSS_0 and RSS_1 have close values. Therefore we will reject the null hypothesis if RSS_0 is sufficiently close to RSS_1. Formally, the test statistics which quantifies the difference between the residuals sums of squares is given by

$$F = \frac{RSS_0 - RSS_1}{RSS_1}. \tag{11.7}$$

Note that if H_0 is correct we expect that F will be small (since $RSS_0 \approx RSS_1$). Hence, we reject the null hypothesis for a large value of F.

In order to proceed further we need to find the distribution of F under the null hypothesis. This can be done using a bootstrap procedure (Davison and Hinkley 1997) which we describe in details in the appendix to this chapter. Briefly, the bootstrap procedure we applied consists of resampling B samples from the original sample. For each bootstrap sample we calculate the value of F. The empirical p-value of the test statistics is simply the proportion of the bootstrap statistics which are larger than the one observed in the original sample. For a given significant level α, one cannot reject the null hypothesis if the $p-\text{value} > \alpha$.

In addition we test the following hypotheses

$$\begin{aligned} H_0 &: E(V_t) = \mu \quad (e.g., \text{constant}), \\ H_1 &: E(V_t) = \eta(t). \end{aligned} \tag{11.8}$$

The null hypothesis in equation (11.8) states that the mean of V_t is constant, but not necessarily 1. Under H_0 in equation (11.8) the residuals sum of squares is

$$RSS_0 = \Sigma_{t=1}^{n} \left\{ f(t) - \bar{f}(t) \right\}^2. \tag{11.9}$$

Here, $\bar{f}(t)$ is the mean of the rebinned data. Note that if we reject the null hypothesis in equation (11.8) the null hypothesis in equation (11.4) will be rejected as well but not vice versa.

It is important to mention here that the hypotheses in equation (11.8) were used only to illustrate the problem. The decision whether a model is significant or not should be based on the hypotheses in equation (11.4).

11.3 Application to the Data

11.3.1 Band 1A

Table 11.1 presents the results for the lack-of-fit tests for the top 20 models in band 1A. For each synthetic spectrum 1000 bootstrap samples (B=1000) were drawn from the original sample as described in the appendix. Whenever the empirical p-value is greater than 0.05

the null hypothesis cannot be rejected. This means that the relationship between V_t and t is assumed to be constant for all models with p-value greater than 0.05. However, under H_0 in equation (11.4) $E(V_t) = 1$. Therefore, we expect that the constant will be close to 1 for a 'good' synthetic spectrum. The bias in the last column in Table 11.1 is defined as

$$bias = (\bar{V}_t - 1) \times 100, \tag{11.10}$$

where $\bar{V}_t$ was estimated under the null hypothesis using the bootstrap procedure (see appendix for more details). Thus, a good synthetic model for the spectrum is one with empirical p-value greater than 0.05 (hence, constant relationship between V_t and t) and small bias (hence, the constant is closed to 1).

The empirical p-values calculated under the null hypothesis in equation (11.4) are all 0. This means that we reject H_0 for all models. When the empirical p-value was calculated under H_0 in equation (11.8) the null hypothesis cannot be rejected for models 58, 83, and 33. For example, model 33 has a p-value of 0.068 (Table 11.1, 4th column) so we do not reject H_0 in equation (11.8). However, this model has a bias of 1.48% (i.e., the constant is estimated to be 1.0148). Figure 11.2 shows the plot of V_t with loess smoothers (with several values of smoothing parameters). Note that, for $\gamma = 0.85$ (the value that was used to calculate the empirical p-value), the loess model is flat (and from this reason the null hypothesis in equation (11.8) cannot be rejected) but it lies above 1 (and from that reason the null hypothesis in equation (11.4) is rejected). This means that, in general, the values of the rebinned data are greater than the values of the synthetic spectrum along the wavelength.

Figures 11.1 and 11.3 show similar patterns for models 58 and 83. As expected, models with a relatively small bias are also ranked among the top 20 when the least square criterion is used. Models 67 and 88 are shown in Figures 11.4 and 11.5. For these models, which ranked among the top 5 when the least square criterion is used, the loess smoother is close to 1 but it is not flat. Therefore, the null hypothesis is rejected although for these two models the bias is relatively small (0.2 and -0.18 for models 67 and 88, respectively). Figure 11.6 shows the results for model 125 (which fit the data poorly according to the least squares criterion). Note how the loess smoother is always below 1 and suggests an increasing trend along the wavelength. When the smoothing parameter for the loess was equal to 0.75 (results are not shown here) H_0 is rejected for all models. The fact that lowering the value of γ results in a rejection of the null hypothesis for all the models indicates that there are systematic patterns left in the data.

11.3.2 Bands 1B, 1D and 1E

The results in bands 1B, 1D, and 1E are similar, the empirical p-values for all models were either zero or very close to zero. Hence, the null hypothesis in equation (11.4) was rejected for all models in the three bands. Figures 11.7 – 11.9 show the plots for models 42, 93, and 85 in band 1B. These three models are ranked first, second and third in band 1B (based on β). For these models there is a clear pattern which can be identified from the plots. This indicates that the synthetic spectra do not follow the same patterns as the rebinned data.

Table 11.1: *Empirical p-values in band 1A. The first column gives the model number, and the second column the rank of the corresponding model determined from the β-value in band 1A for rebinned data. Models which are marked with stars are also ranked among the top 20 based on $T(f_r, \theta)$. Empirical p-values based on a bootstrap with B=1000 and smoothing parameter $\gamma = 0.85$ are given in the second and third column. The fourth column gives the bias as determined from equation (11.10).*

Model	rank (β)	p-value $H_0 : E(V_t) = 1$	p-value $H_0 : E(V_t) = \mu$	Bias under H_0, (%)
58	1	0	0.093	0.986
83	2 ($\star$)	0	0.083	0.460
33	3	0	0.068	1.481
38	4	0	0.015	0.815
63	5($\star$)	0	0.041	0.336
82	6	0	0.017	0.810
57	7	0	0.015	0.987
88	8 ($\star$)	0	0.017	-0.180
32	9	0	0.008	1.928
62	10	0	0.005	0.815
43	11($\star$)	0	0.003	0.215
87	12($\star$)	0	0.002	0.310
68	13($\star$)	0	0.003	-0.262
81	14	0	0.004	1.363
37	15	0	0.000	1.228
56	16	0	0.001	1.896
93	17	0	0.0002	-0.750
67	18($\star$)	0	0.000	0.200
92	19($\star$)	0	0.000	-0.302
86	20	0	0.000	0.651

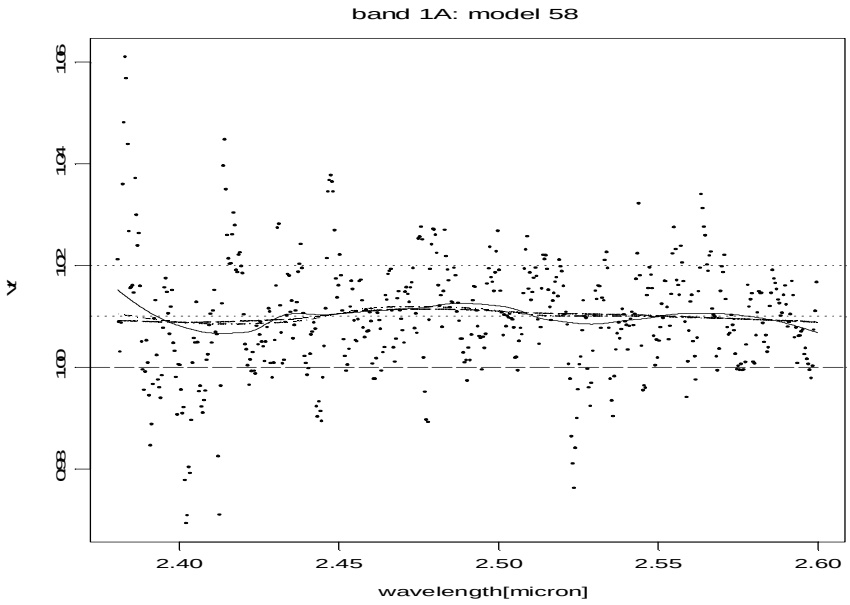


Figure 11.1: *Band 1A: model 58. V_t and the loess smoother with three values of γ. Solid line: $\gamma = 0.5$, dashed line: $\gamma = 0.75$, long-dashed line: $\gamma = 0.85$.*

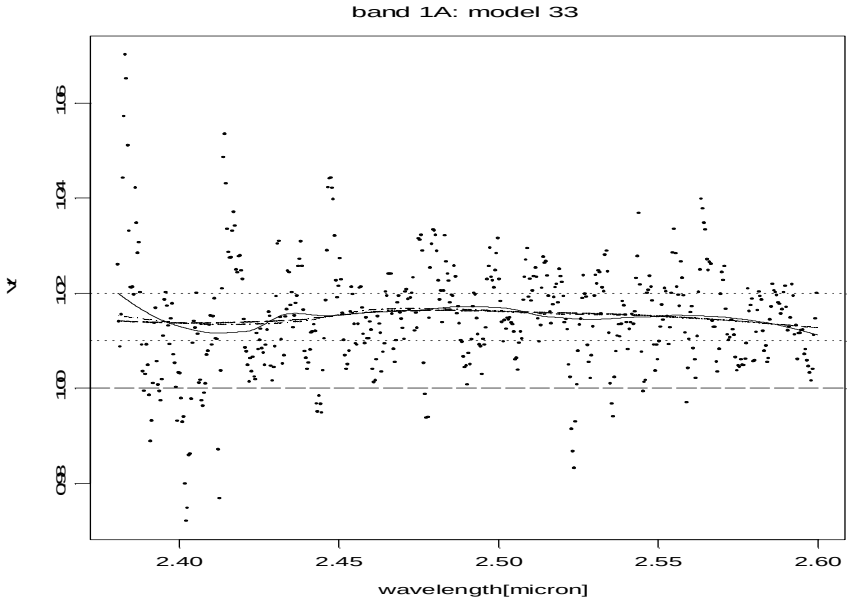


Figure 11.2: *Band 1A: model 33. V_t and the loess smoother with three values of γ.*

11.4 Discussion

What can we learn from the rejection of the null hypothesis by the lack-of-fit test in that many cases? It is clear that this failure can not be solved by relaxing the criteria, e.g.

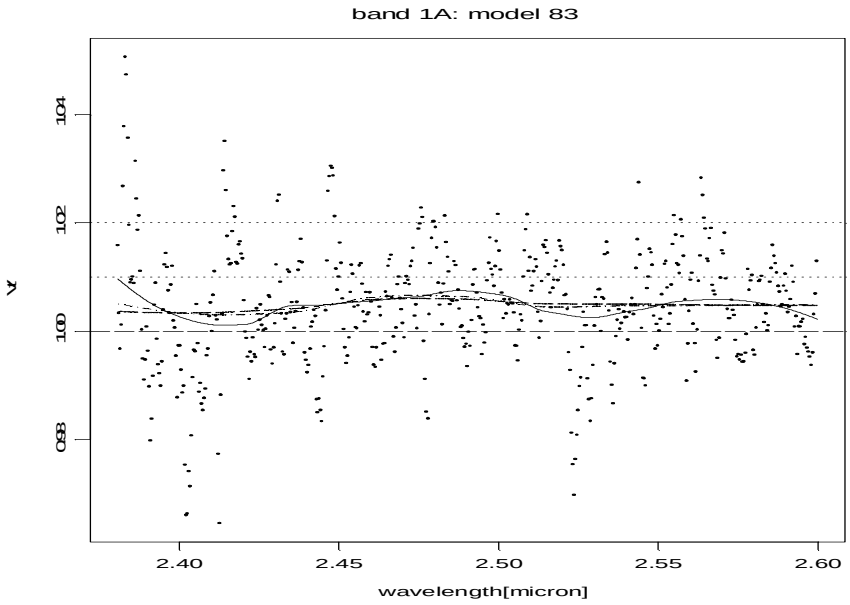


Figure 11.3: *Band 1A: model 83. V_t and the loess smoother with three values of γ.*

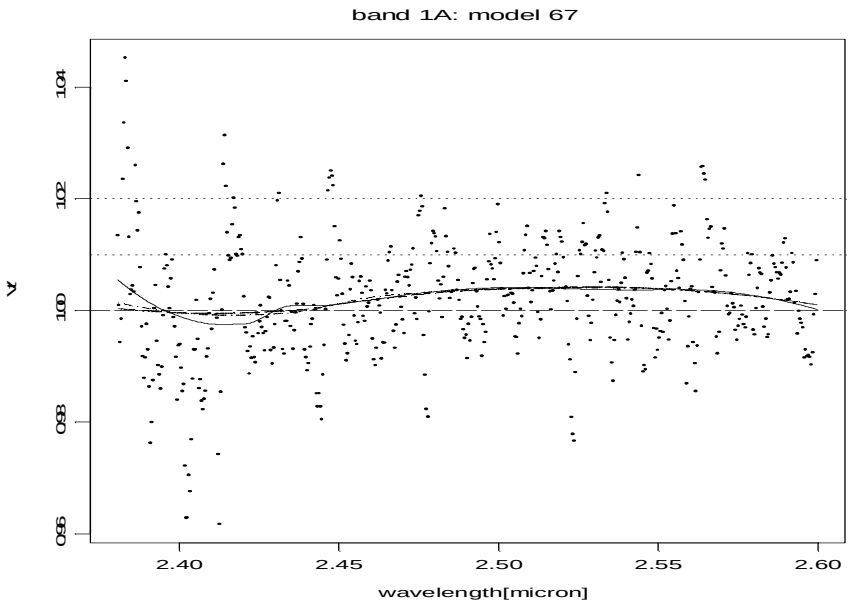


Figure 11.4: *Band 1A: model 67. V_t and the loess smoother with three values of γ.*

by lowering the level of significance α. These lack-of-fit tests are an objective tool to demonstrate that there is still too much of structure left in V_t. This is illustrated, e.g., in Figure 11.10-11.13 where model 68 with a very good goodness-of-fit (for both rebinned data and splines) is depicted in bands 1A, 1B, 1D and 1E. The systematic discrepancy between observations and theory is captured in V_t and its loess smoother, explaining why

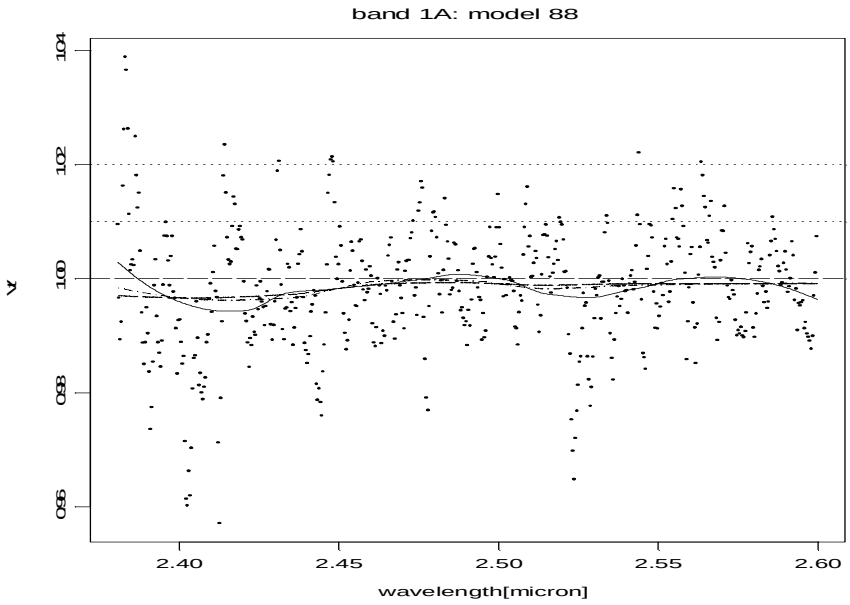


Figure 11.5: *Band 1A: model 88. V_t and the loess smoother with three values of γ.*

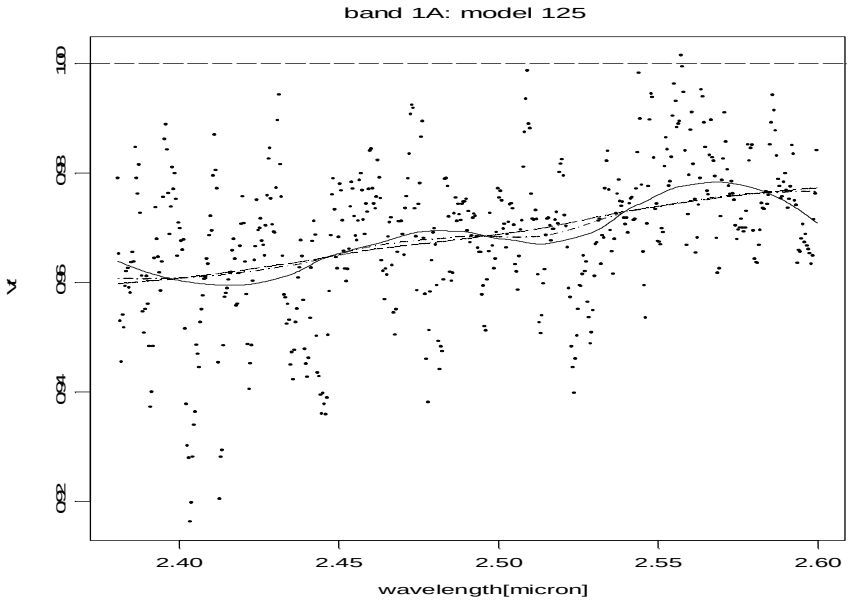


Figure 11.6: *Band 1A: model 125. V_t and the loess smoother with three values of γ.*

the lack-of-fit test rejects the null hypothesis. This systematic problem is not solved by one of the other models in the grid. In general, we may conclude that the systematic rejection of the null hypothesis by the lack-of-fit tests is an indication of a still incomplete knowledge of all the physical mechanisms determining the spectral footprint in the wavelength range considered.

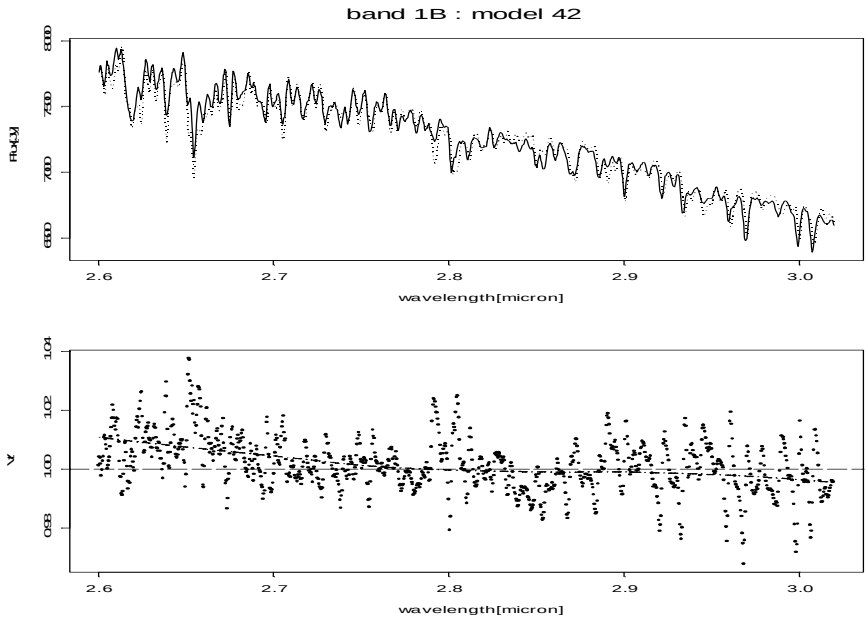


Figure 11.7: *Band 1B, model 44. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

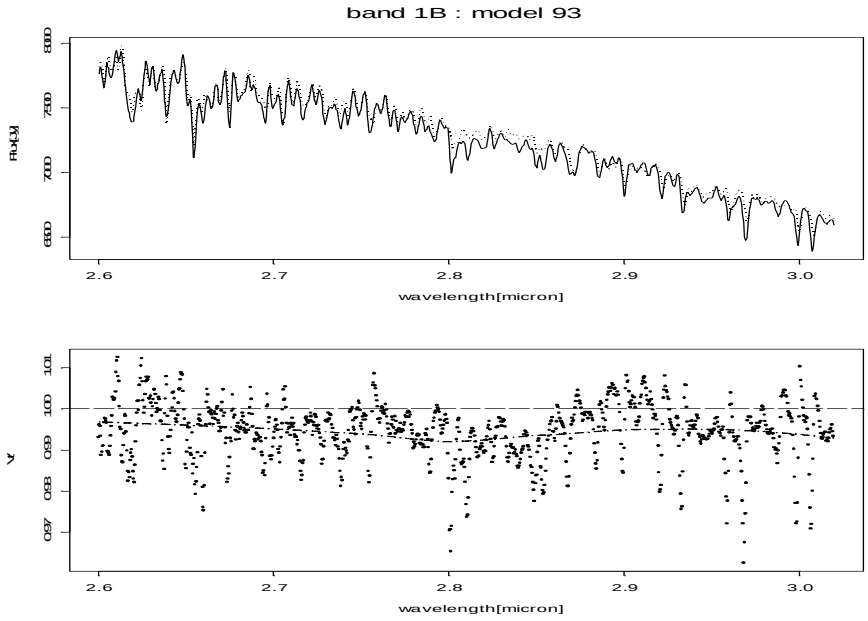


Figure 11.8: *Band 1B, model 93. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

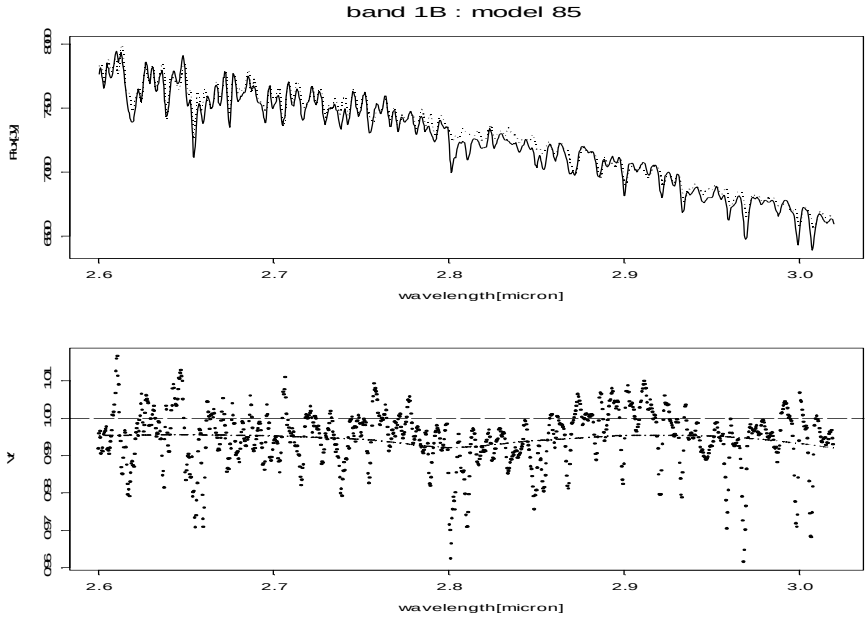


Figure 11.9: *Band 1B, model 85. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

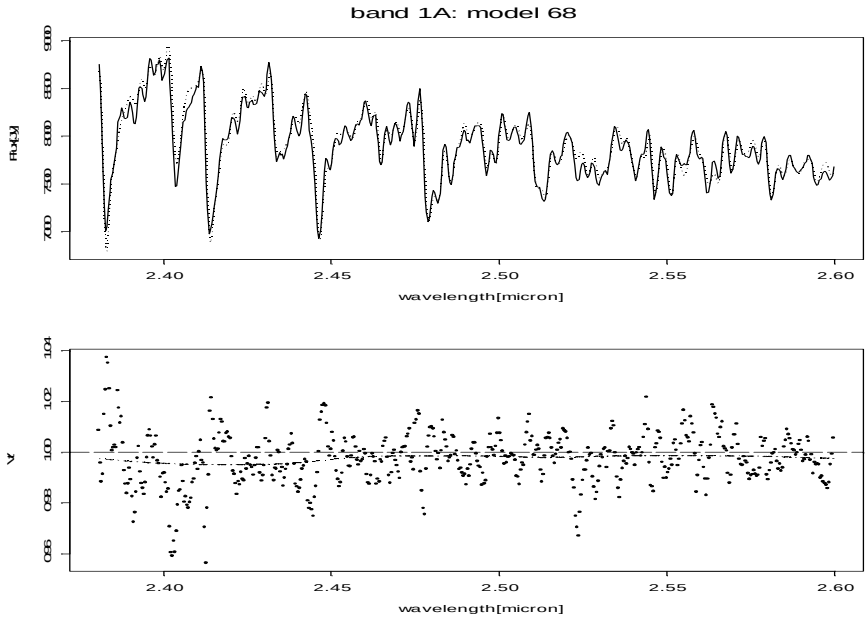


Figure 11.10: *Band 1A, model 68. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

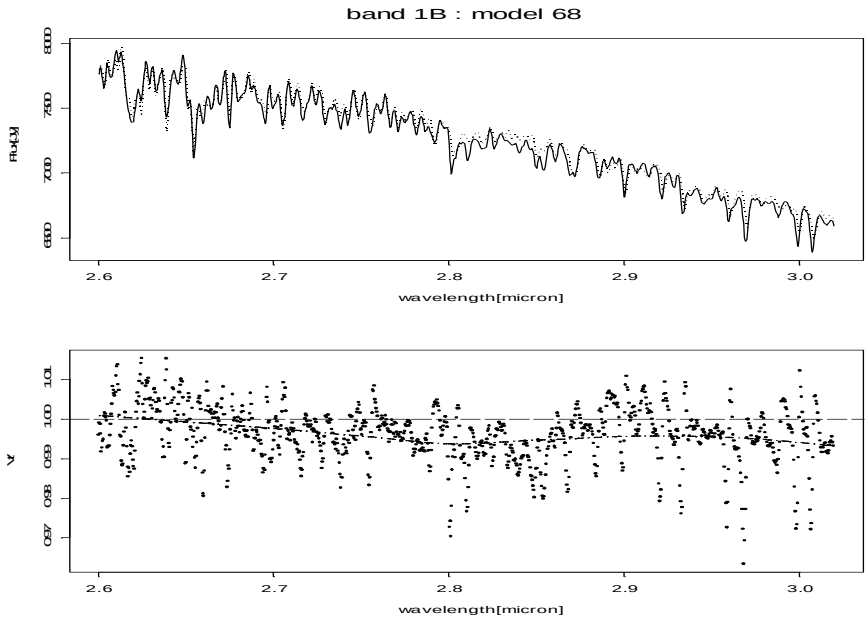


Figure 11.11: *Band 1B, model 68. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

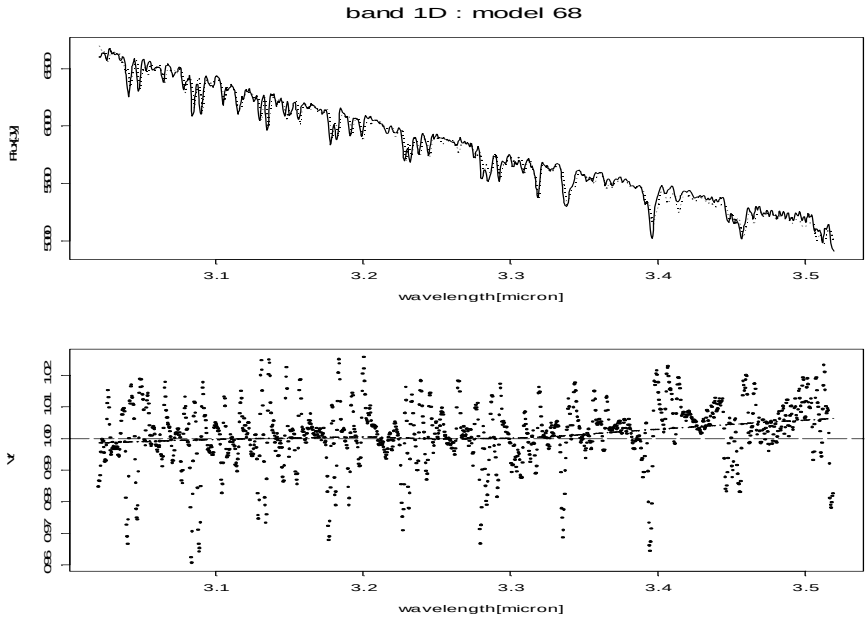


Figure 11.12: *Band 1D, model 68. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

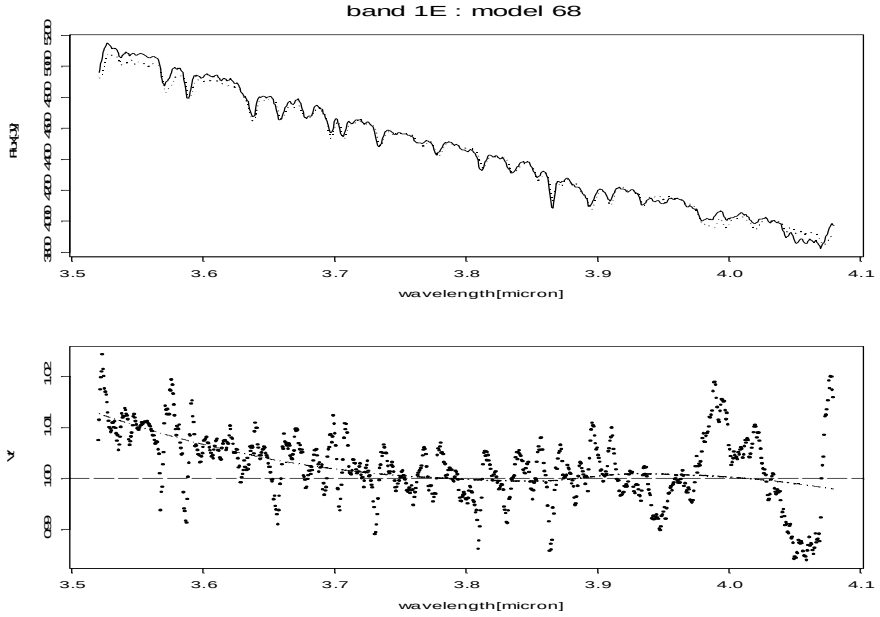


Figure 11.13: *Band 1E, model 68. Upper panel: data and rebinned data. Lower panel:* V_t *and the loess smoother.*

Chapter 12

Smoothing With Hierarchical Linear Mixed Models

12.1 Introduction

Linear mixed models (Verbeke and Molenberhgs 1977, 2000) are commonly used to describe the relationship between a response variable and a predictor(s) when the observations in the datasets are clustered according to a known grouping factor(s). Examples for this type of data include longitudinal data, repeated measured data, multilevel data etc. For all types of data, within the framework of linear mixed models (LMM), it is assumed that the response is normally distributed. In recent years, there is an increasing attention in the statistical literature about the connection between LMM and smoothing splines. The latter are used to estimate nonparametrically an unknown smooth function when the data are assumed to follow a regression model of the form $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$. Due to the Bayesian interpretation of smoothing splines, introduced by Wahba (1978), it can be shown (e.g., Speed 1991) that the solution of the penalized least square problem and the posterior mean of $\boldsymbol{f}$ are identical. This connection allows us to use linear mixed models as scatterplot smoothers.

In Section 12.2 we present two examples of linear mixed models. The first is a model for longitudinal dataset and the second is a model that is used to smooth a scatterplot. The discussion on the connection between LMM and smoothing splines is started in Section 12.3, where we discuss the linear piecewise model of Freedman and Silverman (1989) and illustrate the difference between fixed and mixed effects models in context of smoothing scatterplots. Section 12.3 gives only illustrative examples, the main discussion on the connection between LMM and smoothing splines is given in Section 12.4 and 12.5. It will be shown that the Bayesian formulation of smoothing splines (Wahba 1978, 1983) can be "translated" as a hierarchical Bayesian model with normal likelihood for the response and normal prior for the smooth function. Following Speed (1991) it will be shown that the solution of the penalized least square problem (Green and Silverman 1994) can be expressed as the best linear unbiased predictor (BLUP) obtained from a linear mixed model. The choice of the smoothing parameter is one of the main issues related to nonparametric regression. In the context of smoothing splines, the generalized cross validation method is often used to select the smoothing parameter. In Section 12.6 we discuss the issue of

estimating the smoothing parameter from a LMM point of view. It will be shown that if a LMM is used as a smoother the smoothing parameter is the ratio between the variance components in the model. Furthermore, if a full Bayesian model is fitted, the posterior distribution of the smoothing parameter can be approximated and the smoothing parameter can be estimated by its posterior mean. A simulation study was conducted in order to investigate the influence of the number of knots used in the model and is presented in Section 12.7. The method is applied to the observational errors of α Boo in Section 12.8. We discuss the results and arise a paradox related to the number of parameters in the model in Section 12.9. For the remainder of this section we describe the two sources of observational errors associated with the observed and the synthetic spectra.

12.1.1 Observational Errors

The error propagation of the SWS pipeline reduction package separates pure *statistical* errors from *systematic* errors. Of course, some sources of error (e.g. fringes) will always depend on what sort of and how careful data reduction has been done outside the automatic pipeline. When rebinning the observational data-set, the STDEV of the non-rebinned data-set should be taken into account as weights. This is done for AOT02 (= optimized for sensitive observations of single lines) and AOT06 (= long up-down scan at full instrumental resolution) observations. This is however not done for AOT01 observations since (emission) lines have a smaller STDEV due to the fact that the incoming flux changes during a scan across the line. The observational errors for a rebinned observational data-set are then given by (1) the *statistical* 'STDEV-tag' (σ), which now contains the standard deviation of the points in a certain bin, and (2) the *systematic* 'SPARE-tag' (σ_M), which — once again — contains the offset and gain errors. I.e. the 'SPARE-tag' (σ_M) of the ISO-SWS data corresponds to the statistical term of *accuracy* (= how well we can control systematic errors, how close the result of an experiment comes to the true value), while the 'STDEV-tag' (σ) corresponds to the *precision* (= how well we can overcome random errors). The errors σ and σ_M now have the same order of magnitude (see Figure 12.1). While the 'SPARE-tags' are almost the same for all measurements, the 'STDEV-tag' really makes the distinction between 'good' and 'bad' data-points.

12.2 Linear Mixed Models

Linear mixed models (Verbeke and Molenberghs 1997, 2000) can be formulated as

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{12.1}$$

where $\mathbf{y}$ is a vector of n observed random variables, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known design matrices with dimensions of $n \times p$ and $n \times q$ respectively, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters which are called the fixed effects, $\boldsymbol{b}$ is a $q \times 1$ vector of random effects and $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of unobserved measurement errors. For the random effects and the random error we assume

$$\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D & 0 \\ 0 & \boldsymbol{W} \end{bmatrix} \right).$$

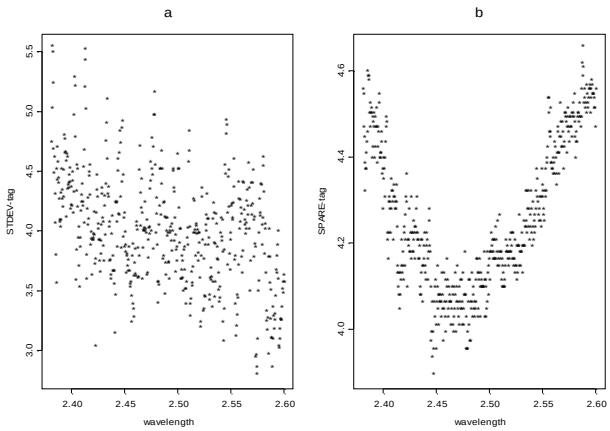


Figure 12.1: *Observational errors in band 1A. Panel a:* statistical *STDEV-tag* σ. *Panel b:* systematic *SPARE-tag* σ_M.

It follows (Searle *et al.* 1992) that, condition on $\boldsymbol{b}$, $\mathbf{y}$ is normal distributed, $\mathbf{y} \sim \mathrm{N}(\boldsymbol{X\beta} + \boldsymbol{Zb}, \boldsymbol{W})$. The marginal distribution of $\mathbf{y}$, $P(\mathbf{y}) = \int P(\mathbf{y}|\boldsymbol{b})P\boldsymbol{b})d\boldsymbol{b}$, given by $\mathbf{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \boldsymbol{V})$ where $\boldsymbol{V} = \boldsymbol{W} + \boldsymbol{ZDZ}'$.

12.2.1 Two Examples of Linear Mixed Models

I: Longitudinal Data Analysis

Pinheiro and Bates (2000) present a dataset of body weights of 16 rats randomized into three treatment groups, each group received a different diet. Each rat was measured at 11 occasions: at the first day of the experiment, and in day 8, 15, 22, 29, 36, 43, 44, 50, 57 and 64. The data are shown in Figure 12.2 which reveals the difference between the response level among the diet group and substantial heterogeneity between the subjects. The parallel subjects profiles suggest a common time effect for the three diet groups.

A possible model for these data is a model which includes a fixed diet effect, a fixed common time trend effect and a random intercept for the subject to account for subject heterogeneity. Such a model can be formulated as special case of the linear mixed model in (12.1) in the following way:

$$\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i. \tag{12.2}$$

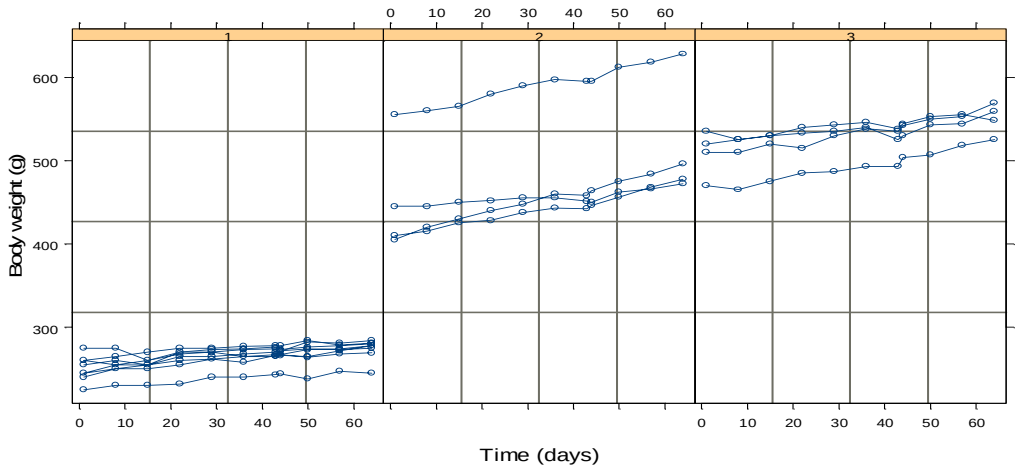


Figure 12.2: *Subject profiles of 16 rats by diet group.*

Here, $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,11})$, $\boldsymbol{\beta} = (\beta_0, \beta_{0,1}, \beta_{0,2}, \beta_1)$ and the design matrices are given by

$$\mathbf{X}_i = \begin{bmatrix} 1 & x_{1i} & x_{2i} & t_{i_1} \\ 1 & x_{1i} & x_{2i} & t_{i_2} \\ . & \ldots & . & . \\ 1 & x_{1i} & x_{2i} & t_{i_{10}} \\ 1 & x_{1i} & x_{2i} & t_{i_{11}} \end{bmatrix} \quad \text{and} \quad \mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}.$$

The variable t_{i_j} denotes the time variable, $t_{i_j} = (1, 5, \ldots, 57, 64)$, x_{1i} and x_{2i} are indicator variables which take the value of 1 if the subject belong the diet group 1 or 2 respectively and zero otherwise. For the random effects we assume $\boldsymbol{b} \sim \mathrm{N}(0, \sigma_b^2 I_{16\times 16})$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 I_{176\times 176})$. Hence, the model in (12.2) can be written as

$$y_{ij} = \begin{cases} \beta_0 + \beta_{0,1} + \beta_1 t_{i_j} + b_i + \varepsilon_{ij}, & \text{diet group 1,} \\ \beta_0 + \beta_{0,2} + \beta_1 t_{i_j} + b_i + \varepsilon_{ij}, & \text{diet group 2,} \\ \beta_0 + \beta_1 t_{i_j} + b_i + \varepsilon_{ij}, & \text{diet group 3.} \end{cases}$$

Predicted and observed means at each time point are shown in Figure 12.4. predicted and observed subjects profiles are shown in Figure 12.3. Both figures indicate that the proposed model fit the data reasonably well.

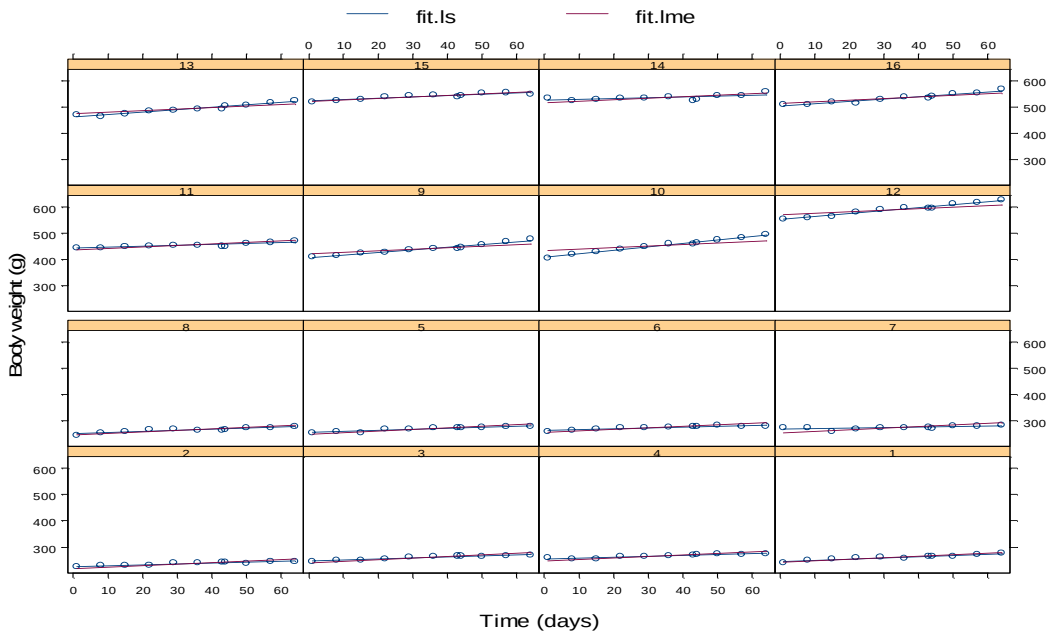


Figure 12.3: *Subject profiles with predicted values. Dotted line: predicted values from the linear mixed model. Solid line: predicted values from individuals simple regression models.*

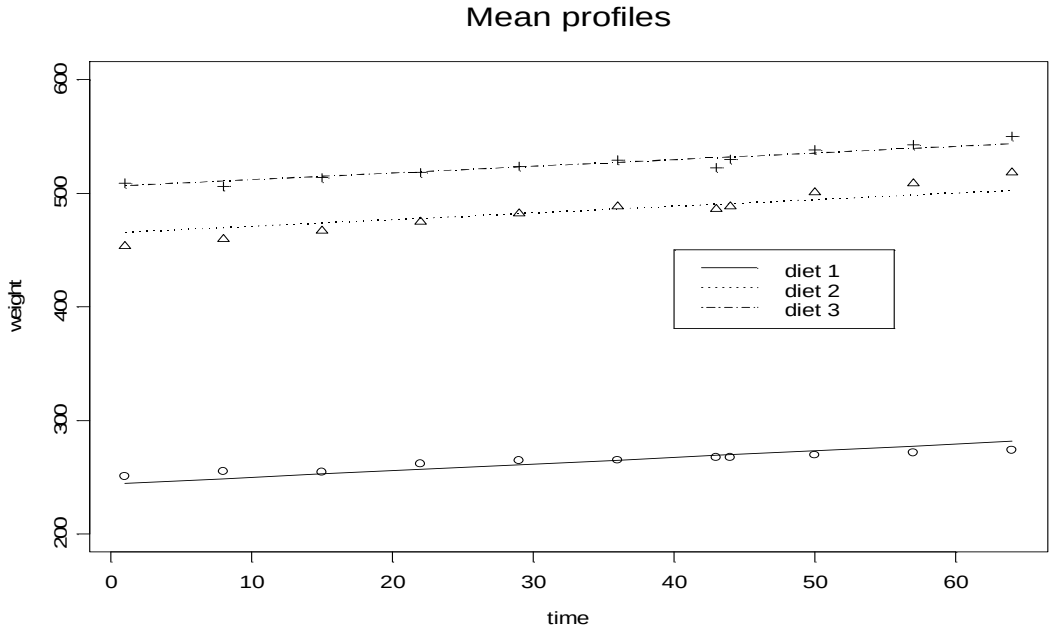


Figure 12.4: *Observed and predicted mean profiles by diet group.*

II: Scatterplots Smoothers

In the previous section, each unit of observation (a rat) has 11 repeated measurements and the random intercept was included in the model in order to capture the subject heterogeneity. Figure 12.5 shows an artifical dataset which was generated from a regression model of the form

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{12.3}$$

where $f(x)$ is a smooth function to be estimated, $x_1 \leq \ldots, \leq x_n$ are the design points and we assume that $\varepsilon_i \sim \mathrm{N}(0, \sigma^2_\varepsilon)$. In vector notation the model is written as $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, where $\mathbf{y} = (y_1, \ldots, y_n)$, $\boldsymbol{f} = (f_1, \ldots, f_n)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. The aim is to estimate the unknown smooth function. For the specific example presented in Figure 12.5, $f(x) = \sin(2\pi(1-x)^2)$. A single dataset with 64 observations was generated by adding to $f(x_i)$, $x_i = i/n$, a random error ε_i sampled from $\mathrm{N}(0, 0.01)$. Compared to the rat dataset in the previous section, the current dataset consists of one observation per unit (x_i, y_i), $i = 1, \ldots, n$. However, the smoother of the data is a linear mixed model with the same form of (12.1). The design matrix for the fixed effects is a 2×64 matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_{63} \\ 1 & x_{64} \end{bmatrix}.$$

$\boldsymbol{Z}$ is a known 64×35 design matrix for the random effects which will be discussed in detail in the next section, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is the vector of the fixed effects and $\boldsymbol{b} = (b_1, \ldots, b_{35})$ is the vector of the random effects which where assumed to follow a normal distribution with mean zero and covariance matrix $\sigma^2_b I_{35\times 35}$. We further assume that $\varepsilon_i \sim \mathrm{N}(0, \sigma^2_\varepsilon)$. The linear mixed model we consider has the form

$$y_i = \beta_0 + \beta_1 x_i + \boldsymbol{Z}_i \boldsymbol{b} + \varepsilon_i. \tag{12.4}$$

The smooth curve (long dashed line) in Figure 12.5 is the predicted curve, $\hat{\mathbf{y}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}}$, obtained from the model in (12.4), where $\tilde{\boldsymbol{b}}$ and $\tilde{\boldsymbol{\beta}}$ are the best linear unbiased predictor (BLUP) estimates for $\boldsymbol{b}$ and the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$. The amount of smoothing (i.e., the smoothing parameter) was chosen automatically using standard linear mixed models methodology which we will discuss in Section 12.6.

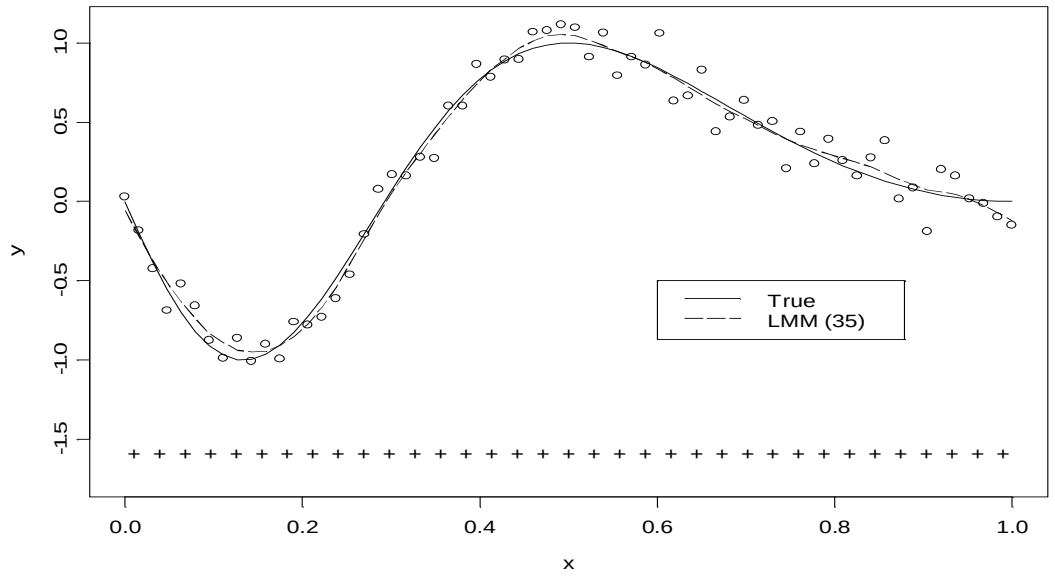


Figure 12.5: *Smoothing scatterplot with linear mixed model. Data (circles), true function (solid line), predicted values for the linear mixed model (long-dashed line). The linear mixed model was fitted with 35 equally spaced knots. The locations of the knots are marked with pluses at the bottom of the figure.*

12.3 Piecewise Linear Smoothing: Freedman and Silverman (1989)

In the previous section, the unknown smooth function f was modeled as $\boldsymbol{f} = \boldsymbol{X\beta} + \boldsymbol{Zb}$ and the smoother for the data was obtained from the linear mixed model $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$. In this section we show that the design matrix for the fixed and the random effects corresponds to the piecewise linear smoother discussed by Freedman and Silverman (1989).
More precisely, the model we consider is

$$f(x_i) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(x_i), \tag{12.5}$$

where x_i are the design points, $i = 1, \ldots, n$, and $\phi_k(x)$, $k = 1, \ldots, K$, are known functions. Note that for $K = 1$ and $\phi_1(x) = x$ the model in (12.5) reduces to a simple linear regression model. The linear piecewise model assumes that the basis function $\phi_k(x_i)$ has the form

$$\phi(x - t_k)_+ = \begin{cases} 0, & x \leq t_k \\ x - t_k & x > t_k \end{cases} \tag{12.6}$$

The linear piecewise model consists of K knots where t_k, $k = 1, 2 \ldots, K$ is the location of the kth knot. The basis function represents a broken line with the knots t_k as a joint point. Figure 12.6 shows an example of the $\phi_k(x)$ with 10 equally spaced knots.

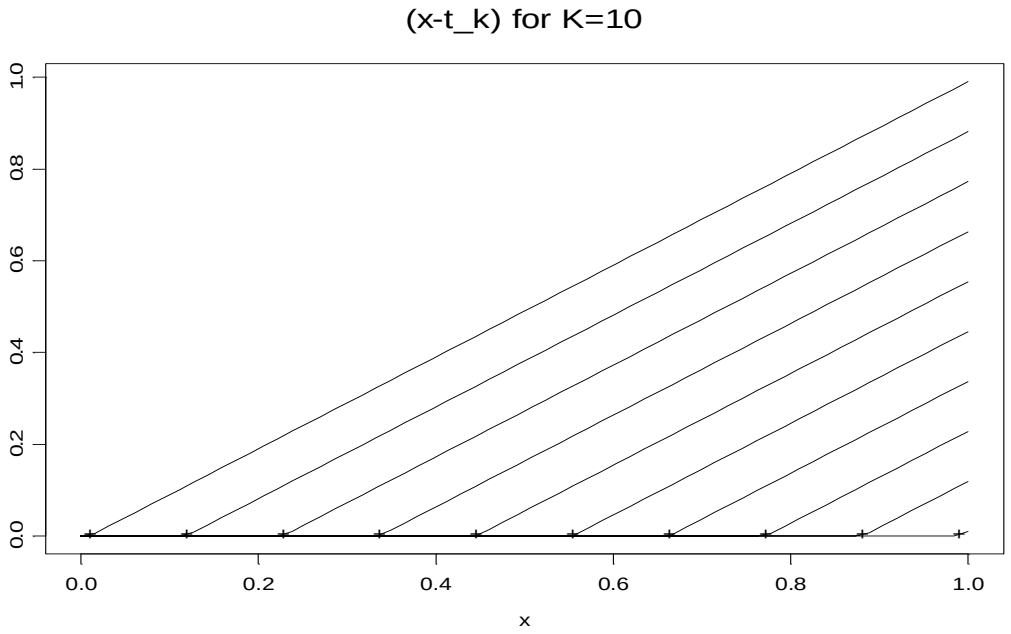


Figure 12.6: $\phi(k) = (x_i - t_k)_+$ *with 10 knots. The locations of the knots are marked with pluses.*

Note that for $t_1 = 0$ the piecewise linear model can be written as

$$f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=2}^{K} \beta_k (x_i - t_k)_+, \tag{12.7}$$

We define two design matrices, an $n \times 2$ design matrix for which the ith row is $\boldsymbol{X}_i = [1, x_i]$ and an $n \times L$ matrix for which the ith row is $Z_i = [(x_i - t_1)_+, \ldots, (x_i - t_L)_+]_{2 \le k \le K}$. The model in (12.7) can be rewritten as

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{Z}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{12.8}$$

with $\boldsymbol{\beta}_1 = (\beta_0, \beta_1)$ and $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_K)$. In the terminology of mixed effects models, the model in (12.8) is a fixed effects model which can be written as $\mathbf{y} = [\boldsymbol{X}|\boldsymbol{Z}]\,[\boldsymbol{\beta}_1|\boldsymbol{\beta}_2]' + \boldsymbol{\varepsilon}$. Figure 12.7 shows the least square fit to the data.

In order to proceed further, let us consider the components of $\boldsymbol{\beta}_2$ as normal distributed random effects (hence, we change notation from $\boldsymbol{\beta}_2$ to $\boldsymbol{b}$), $\boldsymbol{b} \sim \mathrm{N}(0, \sigma_b^2 I_{L \times L})$, $L = K - 1$. Thus, $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$ and $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 I_{n \times n})$. The model has the same form as the linear mixed model defined in (12.1). Note that in this model the unknown smooth function is modeled with two components. The linear part, which is the fixed effects $\boldsymbol{X}\boldsymbol{\beta}$, and the smooth part $\boldsymbol{Z}\boldsymbol{b}$. The conditional mean of $\mathbf{y}$, given the fixed and the random effects, is $E(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{f}$. Figure 12.8 shows the data and three estimated models that were fitted with 10, 20 and 35 knots. The issue of the number of knots will be investigated further by means of a simulation study which we present in Section 12.7. The difference between the fixed effects model and the mixed effects model is also illustrated in Figure 12.9 which shows the fitted values obtained from the two models for 25 datasets which were generated in the same way we described before.

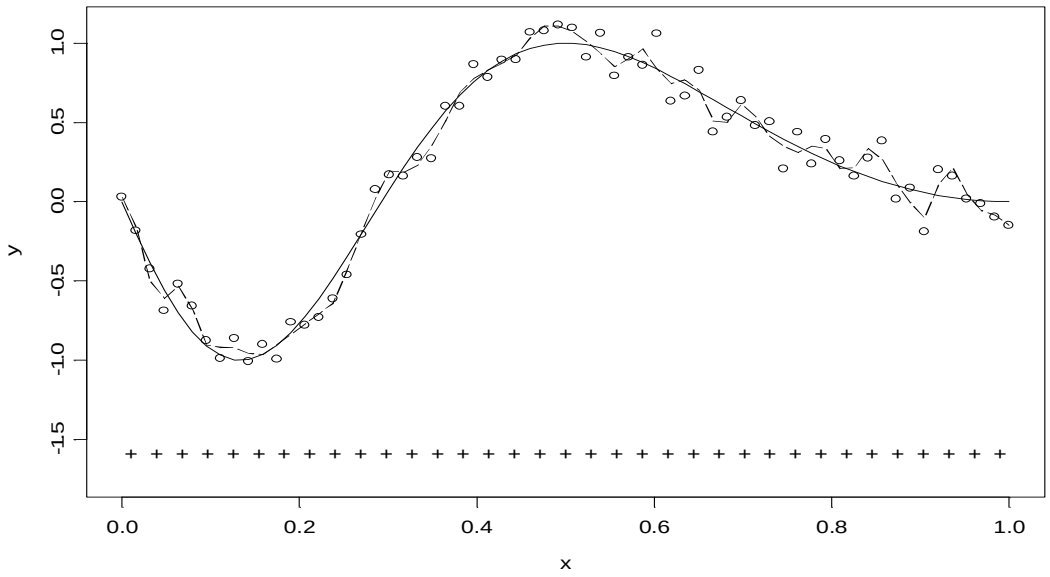


Figure 12.7: *Smoothing the data with the fixed effects model (12.8). Long dashed line: predicted values for the model which was estimated using the least square method. solid line: true function. The model was fitted with 35 equally spaced knots. The locations of the knots are marked with plusess at the bottom of the figure.*

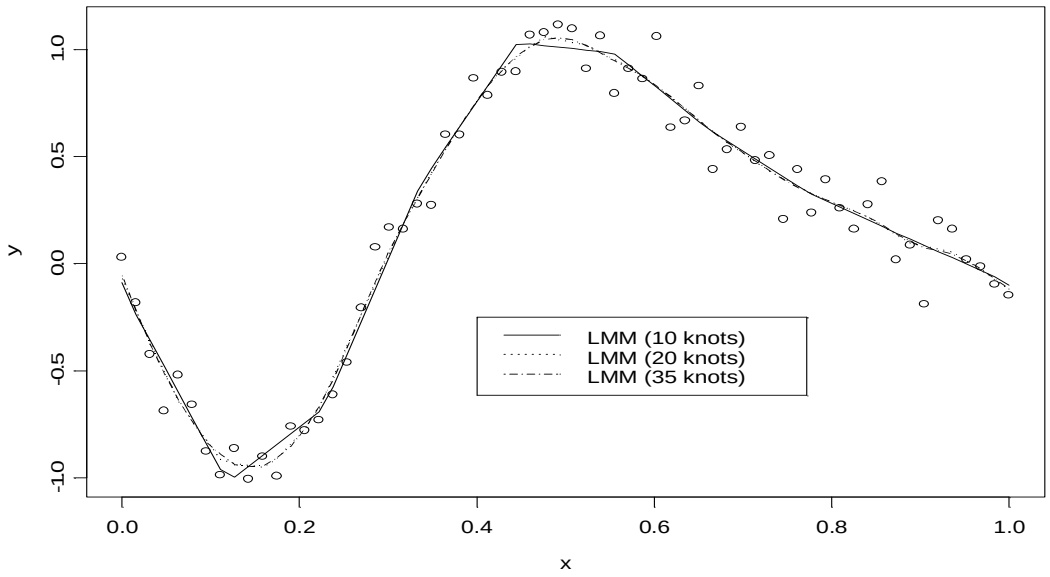


Figure 12.8: *Smoothing the data with the linear mixed effects model (12.8). Long dashed line: model with 35 knots. Dotted line: model with 20 knots. Solid line: Model with 10 knots. In all models the knots are equally spaced.*

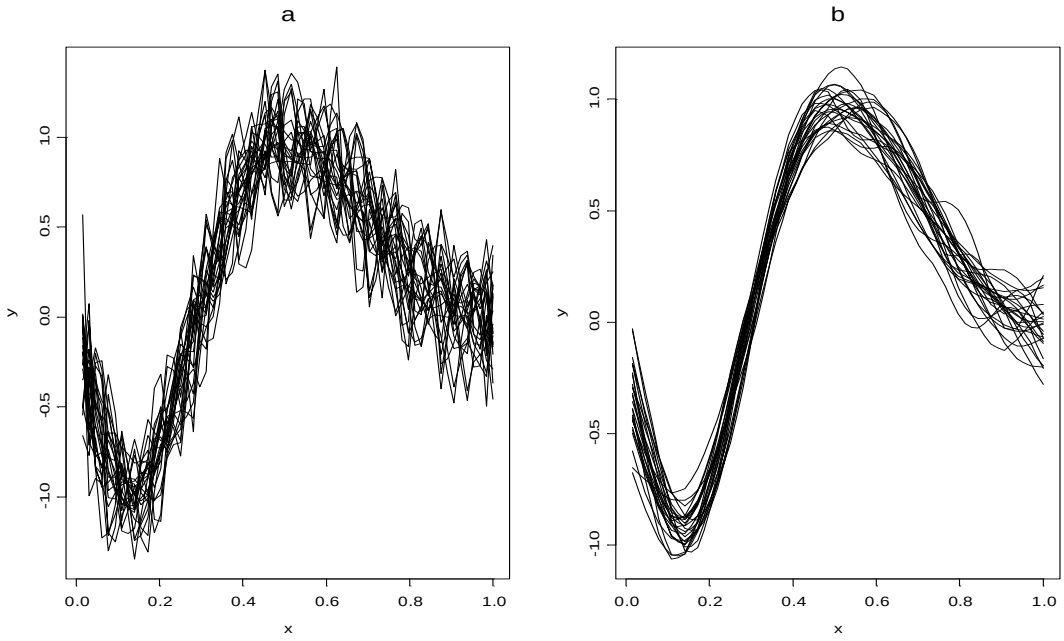


Figure 12.9: *Smoothing the data with the linear fixed effects model (panel a) and linear mixed fixed effects model (panel b) in (12.8) . Both models were fitted with 35 equally spaced knots.*

12.4 Cubic Smoothing Splines As Linear Mixed Models

So far, we showed that linear mixed models can be used as scatterplot smoothers. In this section we will make the connection between linear mixed models and cubic smoothing splines. We will show that the Bayesian interpretation of a smoothing spline, introduced by Wahba (1978,1983) can be formulated as a hierarchical linear mixed model. Moreover, we will show that a cubic smoothing spline, evaluated at the design points, can be expressed as $\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}}$ where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{b}}$ are the BLUP and the BLUE for $\boldsymbol{b}$ and $\boldsymbol{\beta}$, respectively.

12.4.1 Smoothing Splines

We consider the regression model (12.3). The aim is to estimate the unknown smooth function f without imposing parametric structure about the mean of f. Within the parametric regression framework f can be estimated by minimizing the residuals sum of square, $RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$. However, within the nonparametric framework this is not possible, since the choice of $\hat{\boldsymbol{f}} = \mathbf{y}$ brings RSS to zero and $\hat{\boldsymbol{f}}$ will interpolate the data. As Green and Silverman (1994) argued, one needs to consider two different criteria when estimating $\boldsymbol{f}$ nonparametrically: the goodness-of-fit and the smoothness of $\hat{\boldsymbol{f}}$. The basic idea of the penalized least square regression is to pose the estimation problem in a way that it compromises between goodness-of-fit and smoothness. The penalized sum of

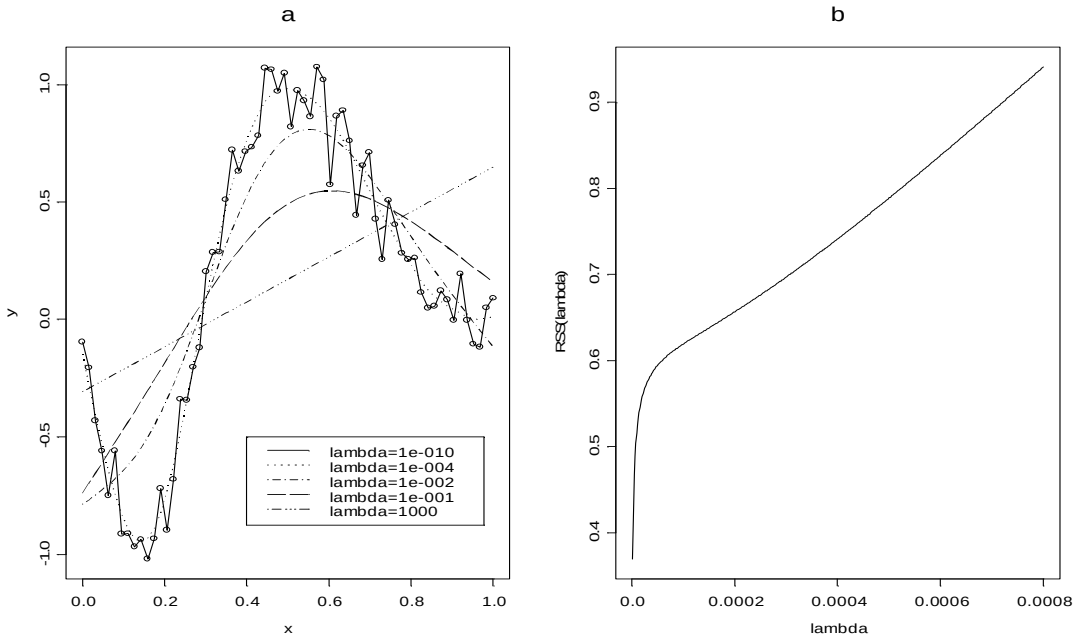


Figure 12.10: *Panel a: Smoothing cubic splines fitted with several values of* λ. *Panel b:* $RSS(\lambda)$ *versus* λ.

squares (Wahba 1978, Green and Silverman 1994) is given by

$$S(\boldsymbol{f}) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx, \tag{12.9}$$

where $\int_0^1 (f^{(m)}(x))^2 dx$ is the roughness penalty which penalizes for the lack of smoothness of $\hat{\boldsymbol{f}}$. Note that if $\hat{\boldsymbol{f}} = \beta_0 + \beta_1 x$ then $\int(\hat{\boldsymbol{f}}'')^2 = 0$. For $m = 2$ and given λ, $\hat{\boldsymbol{f}}$, the minimizer of $S(\boldsymbol{f})$, is a cubic smoothing spline. For other choices of m, the minimizer of $S(\boldsymbol{f})$ is a $2m-1$ order polynomials spline. In this thesis we focus on the case with $m = 2$. The smoothing parameter λ, which needs to be estimated, controls the tradeoff between the smoothness of $\hat{\boldsymbol{f}}$, as measured by $\int_0^1 (f''(x))^2 dx$ and the goodness-of-fit to the data, as measured by $RSS(\lambda) = \sum_i^n (y_i - f(x_i))^2$. When λ tends to be "small" the dominant term of $S(\boldsymbol{f})$ is $RSS(\lambda)$ and when $\lambda \longrightarrow 0$, then $\hat{\boldsymbol{f}}$ converges to an interpolating spline. As λ increases, the squared integral $\int(f''(x))^2 dx$ becomes the dominant term in $S(\boldsymbol{f})$, and $\hat{\boldsymbol{f}}$ becomes more and more "smooth". The $RSS(\lambda)$, on the other hand, increases. As $\lambda \longrightarrow \infty$, $\hat{\boldsymbol{f}}$ converges to a straight line, and in that case $RSS(\lambda)$ is simply the sum of squares of the regression line. Figure 12.10 illustrates this point.

The Hat Matrix

Let R be an $(n-2)\times(n-2)$ tridiagonal matrix with entries $R_{i,j},\ \ i,j = 1,\ldots,n-2$ given by $[R]_{i,i} = 2(h_{i-1}+h_i)/3$, $[R]_{i,i+1} = [R]_{i+1,i} = h_i/3$, and $\boldsymbol{G}$ is an $(n-2)\times(n)$ tridiagonal matrix with entries $[G]_{i,j},\ \ i,= 1,\ldots,n,\ j = 1,\ldots,n-2$ given by $[G]_{i,i} = -1/h_{i-1} - 1/h_i$,

$[G]_{i-1,i} = 1/h_{i-1}$ and $[G]_{i+1,i} = 1/h_i$. Let $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{R}^{-1}\boldsymbol{G}'$, the matrix $\boldsymbol{K}$ satisfy

$$\int_0^1 (f''(x))^2 dx = \boldsymbol{f}'\boldsymbol{K}\boldsymbol{f},$$

and penalized least square problem in (1.9) can be written as $S(\boldsymbol{f}) = (\mathbf{y} - \boldsymbol{f})'(\mathbf{y} - \boldsymbol{f}) + \lambda \boldsymbol{f}'\boldsymbol{K}\boldsymbol{f}$. Equating the first derivative of $S(\boldsymbol{f})$ to zero leads to $\hat{\boldsymbol{f}} = (I - \lambda \boldsymbol{K})^{-1}\mathbf{y} = \boldsymbol{A}(\lambda)\mathbf{y}$. The matrix $\boldsymbol{A}(\lambda)$ is called the hat matrix or the smoothing matrix. For a given value of λ the rows of $A(\lambda)$ act as kernel function. We will discuss this issue further in Section 12.6.3 where it will be shown that $A(\lambda)$ is identical to the hat matrix obtained from a LMM which is used to smooth the data.

The equivalent Degrees of Freedom and the Estimator for σ_ε^2

In the case of parametric regression the number of parameters in the fitted model is equal to $tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$, where $\boldsymbol{X}$ is the design matrix and is also the degrees of freedom of the fitted model. The degrees of freedom of the noise are equal to $tr(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$. By analogy to parametric regression Wahba (1978) suggested that the equivalent degrees of freedom (EDF) are equal to $tr(I - \boldsymbol{A}(\lambda))$. In the parametric case the unbiased estimator for σ_ε^2 is the residual sum of squares divided by the degrees of freedom of the noise. The estimator for σ_ε^2 suggested by Wahba is

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS(\lambda)}{tr(I - \boldsymbol{A}(\lambda))}. \tag{12.10}$$

12.4.2 Estimating the Smoothing Parameter

The main issue which arises when a nonparametric regression model is fitted is the choice of the smoothing parameter. Different choices of λ will lead to different estimated models as we have shown in Chapter 3 where we selected the optimal bandwidth for the local polynomial models. In the context of smoothing splines, an automatic procedure which leads to a data driven smoothing parameters, is the commonly used cross validation method. Note that compared with the optimal bandwidth in Chapter 3 that was chosen in order to minimize the asymptotic mean square error, the smoothing parameter which is chosen by the cross validation method minimizes the prediction square error as we will show in the next section.

The true mean squared error

The true mean squared error, $R(\lambda)$, defined by Craven and Wahba (1979), is given by:

$$R(\lambda) = \frac{1}{n}(\hat{\boldsymbol{f}} - \boldsymbol{f})\prime(\hat{\boldsymbol{f}} - \boldsymbol{f}). \tag{12.11}$$

The estimate of λ which minimizes $R(\lambda)$ is called the optimal smoothing parameter. The minimizer of $R(\lambda)$, can easily be found by grid search. However, in practice finding the minimizer of $R(\lambda)$ is not possible since f is unknown.

Risk approach

Let $ER(\lambda)$ be the expectation of $R(\lambda)$. Then since $\hat{\boldsymbol{f}} = \boldsymbol{A}\mathbf{y}$ (we denote $\boldsymbol{A} = \boldsymbol{A}(\lambda)$)

$$ER(\lambda) = \frac{1}{n}\boldsymbol{f}'(I-\boldsymbol{A})'(I-\boldsymbol{A})\boldsymbol{f} + \frac{\sigma_\varepsilon^2}{n}tr(\boldsymbol{A}'\boldsymbol{A}). \tag{12.12}$$

$ER(\lambda)$ is called the risk function.
An unbiased estimate of $ER(\lambda)$ (Craven and Wahba 1979), $\hat{R}(\lambda)$ is defined by:

$$\hat{R}(\lambda) = \frac{1}{n}\mathbf{y}'(I-\boldsymbol{A})'(I-\boldsymbol{A})\mathbf{y} - \frac{\sigma_\varepsilon^2}{n}tr(I-\boldsymbol{A})'(I-\boldsymbol{A}) + \frac{\sigma_\varepsilon^2}{n}tr(\boldsymbol{A}'\boldsymbol{A}). \tag{12.13}$$

The minimizer of $\hat{R}(\lambda)$ is called the unbiased risk estimate. This method can be used only when σ_ε^2 is known.

The Cross Validation method

The basic idea of the cross validation method is to obtain an estimate to the minimizer of $\hat{R}(\lambda)$ without knowledge of σ^2. Wald and Wahba (1975) suggested using the Cross Validation method. Let $\hat{\boldsymbol{f}}^{-i}$ be the curve that estimates $\boldsymbol{f}$ from the data without the point (x_i, y_i), so $\hat{\boldsymbol{f}}^{-i}$ is the minimizer of:

$$\sum_{j\neq i}(y_j - f(x_j))^2 + \lambda\int(f''(x))^2 d(x) \quad . \tag{12.14}$$

The predicted value for $f(x_i)$ is $\hat{f}^{-i}(x_i)$ and the squared prediction error is $\{y_i - \hat{f}^{-i}(x_i)\}^2$. Summing over i , $i = 1, 2....n$, we get the Cross Validation score, $CV(\lambda)$, which is simply the sum of the squared prediction errors :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}^{-i}(x_i)\}^2 \quad . \tag{12.15}$$

The cross validation estimate for λ is the value that minimizes the cross validation score. In order to calculate $CV(\lambda)$ in (12.15), one needs to fit n models, each one of the models without the point (x_i, y_i). It can be shown (Craven and Wahba 1979, Hastie and Tibshirani 1990, Green and Silverman 1994) that

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - A_{ii}(\lambda)}\right)^2 \quad , \tag{12.16}$$

where $\hat{\boldsymbol{f}}$ is the spline curve calculated from the full data set. When $A_{ii}(\lambda)$ is replaced by its average value:

$$\frac{1}{n}\sum_{i=1}^{n}A_{ii}(\lambda) = \frac{1}{n}tr(A(\lambda))$$

in (12.16), we get the Generalized Cross Validation score, $GCV(\lambda)$, defined by:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}(x_i))^2}{(1 - \frac{1}{n}tr(\boldsymbol{A}(\lambda)))^2} \quad . \tag{12.17}$$

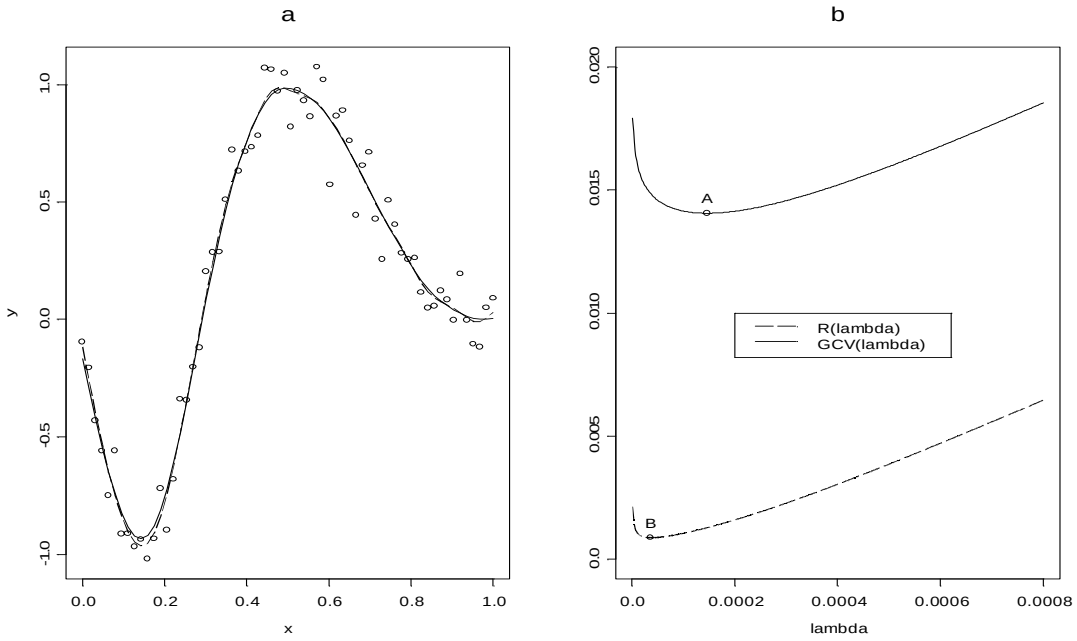


Figure 12.11: *Panel a: Data and fitted models using GVC and R(λ). Panel b: GCV(λ) and R(λ). The point A represents the GCV score for the minimizer of GCV(λ) and the point B the R score for the minimizer of R(λ)*

Note that compared to the variable optimal bandwidth that we calculated for the local polynomial models, the smoothing parameter which minimizing $GCV(\lambda)$ is fixed along the range of the design points. Thus, the amount of smoothing is the same over the range of the predictor. One advantage to use the GCV method is that it does not depend on unknown quantities, such as the density of x or higher order derivatives of f that are needed in order to estimate the optimal bandwidth as we have shown Chapter 3. Figure 12.11 (panel a) shows two fitted model. For the first, the smoothing parameter is the minimizer of $GCV(\lambda)$ and for the second the smoothing parameter is the minimizer of $R(\lambda)$. There is not much difference between the fitted models. Let $\tilde{\lambda}$ be the estimate for λ using the GCV method and let $\lambda^\star$ be the minimizer of $R(\lambda)$. The 'weak GCV theorem' proved by Craven and Wahba (1979) says: as $n \longrightarrow \infty$ then $\lim \frac{ER(\tilde{\lambda})}{ER(\lambda^\star)} \longrightarrow 1$. So the minimizer of $GCV(\lambda)$ is in the neighborhood of the minimizer of $R(\lambda)$. Figure 12.11 (panel b) shows a plot of $R(\lambda)$ and $GCV(\lambda)$ and the chosen values of λ for the two cases.

12.4.3 The Bayesian Interpretation of Smoothing Splines

The Bayesian formulation of the smoothing spline problem was introduced by Wahba (1978), it was discussed further by Wahba (1982,1983), Silverman (1985) and Green and Silverman (1994). The Bayesian formulation of the model is the basis for the construction of the confidence intervals for $\hat{\boldsymbol{f}}$. In this section we will show that the model considered by Wahba (1978) can be formulated as a hierarchical Bayesian linear mixed model.
Wahba (1978) considers the nonparametric regression model as in (12.3). Theorem 2 in Wahba's paper states:

Let $f(x)$ have a prior distribution which is the same as the distribution of the stochastic process $H_\xi(x)$,

$$H_\xi(x) = \sum_{j=1}^{m} \beta_j \delta_j(x) + \sigma_b W(x), \tag{12.18}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m) \sim N(0, \xi T_{m\times m})$,$\sigma_b$ is fixed ≥ 0 and $W(x)$ is a zero mean Gaussian stochastic process with $EW(s)W(t) = Q(s,t)$. Then:

$$\hat{f}(x) = \lim_{\xi\to\infty} E_\xi \{f(x)|y\}$$

with $\lambda = \sigma_\varepsilon^2/n\sigma_b^2$, where E_ξ is expectation of the posterior distribution $f(x)$ with the prior (12.18).

In the theorem above $\hat{f}(x)$ is the minimizer of (12.9), m is the order of the derivative in $\int (f^{(m)})^2 dx$, $\delta_j = x^{j-1}/(j-1)!$ and $\boldsymbol{Q}$ is a $n \times n$ matrix with the ijth entry $Q(x_i, x_j)$,

$$Q(x_i, x_j) = \int_0^1 (x_i - u)_+^{m-1} (x_j - u)_+^{m-1} du/((m-1)!)^2. \tag{12.19}$$

One can conclude for Wahba's theorem that for $\lambda = \sigma_\varepsilon^2/n\sigma_b^2$ the minimizer of (12.9) is equal to the posterior mean of $H_\xi(x)$. For cubic smoothing splines ($m = 2$), Wahba's prior reduces to

$$H_\xi(x) = \beta_0 + \beta_1 x + \sigma_b W(x) = \boldsymbol{X}\boldsymbol{\beta} + \sigma_b W(x), \tag{12.20}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{X}$ is a known matrix for which the ith row is $[1, x_i]$. Let $\boldsymbol{Z}$ be a $n \times n$ known matrix such that $Q(s,t) = \boldsymbol{Z}\boldsymbol{Z}'$. Let $\boldsymbol{b}$ be a $n \times 1$ vector, $\boldsymbol{b} \sim \mathrm{N}(0, I)$ and define $W(x) = \boldsymbol{Z}\boldsymbol{b}$. It is easy to see that $W(x) \sim \mathrm{N}(0, \boldsymbol{Z}\boldsymbol{Z}')$ as required. Since σ_b is fixed, the distribution of $\sigma_b W(x)$ is $\mathrm{N}(0, \boldsymbol{Z}\boldsymbol{Z}'\sigma_b^2)$. Therefore, if we modify the distribution of $\boldsymbol{b}$ to be $N(0, \sigma_b^2 I_{n\times n})$, Wahba's prior can be rewritten as

$$H_\xi(x) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}. \tag{12.21}$$

We remark that $\boldsymbol{Z}\boldsymbol{b}$ has the same distribution as $\sigma_b W(x)$.

Let us focus now on the underlying hierarchical Bayesian model that corresponds to Wahba's prior. In the first stage of the hierarchical model we specify the likelihood as

$$\begin{aligned} y_i &\sim \mathrm{N}(H_\xi(x_i), \sigma_\varepsilon), \quad \text{likelihood,} \\ H_\xi(x) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \quad \text{mean structure.} \end{aligned} \tag{12.22}$$

We replace $f(x)$ with $H_\xi(x_i)$ since, as Wahba's the theorem states, $\hat{\boldsymbol{f}}$ is posterior mean of f with the prior (12.18). Moreover, Green and Silverman (1994), when discussing the results of Wahba(1978, 1983) formulate f as $f(x) = A + Bx + \lambda^{-0.5} \int_0^1 W(s)d(s)$, where $W(s)$ was assumed to be a Brownian motion on $(-\infty, \infty)$ and A and B have noninformative uniform prior. Thus, they assumed that $E(y_i|A, B, W(s)) = A + Bx + \lambda^{-0.5} \int_0^1 W(s)ds$. A similar approach was taken by Wang (1988b), who assumed, in the context of nonparametric mixed effects models, that $\mathbf{y} = \boldsymbol{f} + \boldsymbol{Z}_1\mathbf{u} + \boldsymbol{\varepsilon}$ and reformulated the model as $\mathbf{y} = H_\xi(x) + \boldsymbol{Z}_1\mathbf{u} + \boldsymbol{\varepsilon}$. We keep the prior model as it is in Wahba (1978) but use it as the mean of $\mathbf{y}$ in the likelihood. Thus, we reformulate the model as $\mathbf{y} = H_\xi(x) + \boldsymbol{\varepsilon}$. From this stage

the upper levels of the hierarchical model are straight forward. In the second stage of the hierarchical model we specified the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{b}$

$$\beta_0 \sim \mathrm{N}(0, \xi),$$

$$\beta_1 \sim \mathrm{N}(0, \xi),$$

$$b_i \sim \mathrm{N}(0, \sigma_b^2) \quad i = 1, \ldots, n.$$

In order to complete the specification of the hierarchical model we assume flat hyperprior distributions for ξ, σ_b^2 and σ_ε^2 at the third stage.

$$\xi^{-1} \sim \mathrm{gamma}(0.0001, 0.0001),$$

$$\sigma_b^{-2} \sim \mathrm{gamma}(0.0001, 0.0001),$$

$$\sigma_\varepsilon^{-2} \sim \mathrm{gamma}(0.0001, 0.0001).$$

One should not feel uncomfortable with the non informative hyperprior distribution for the hyperparameters as Wahba herself wrote (page 366 in Wahba 1978): "$\xi = \infty$ corresponds to the diffuse prior for β". Furthermore, in the same context, Green and Silverman (1994) (i.e. page 52) wrote "where A and B have improper uniform distribution on $(-\infty, \infty)$". We interpret Wahba's and Green and Silverman's statements as non informative prior. Figure 12.12 shows the data and the posterior mean and 95% credible intervals for $\boldsymbol{f}$ estimated by the hierarchical LMM.

12.5 That BLUP is a Good Thing

In the previous section, the discussion was started from smoothing splines and we have shown that the Bayesian formulation of the model leads to hierarchical LMM. In this section we discuss the connection between LMMs and cubic smoothing splines from the opposite direction. We start from LMM and show that cubic smoothing splines are just BLUP. Recall that the model we consider is $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}$. We focus on the conditional mean $E(\boldsymbol{b}|\mathbf{y})$. The joint distribution of $\boldsymbol{b}$ and $\mathbf{y}$ (Searle *et al.* 1992) is

$$\begin{bmatrix} \boldsymbol{b} \\ \mathbf{y} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ \boldsymbol{X\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{D} & \boldsymbol{DZ}' \\ \boldsymbol{ZD} & \boldsymbol{V} \end{bmatrix} \right), \tag{12.23}$$

it follows that $\boldsymbol{b}|\mathbf{y} \sim \mathrm{N}(\boldsymbol{DZ}'\boldsymbol{V}^{-1}(\mathbf{y} - \boldsymbol{X\beta}), \boldsymbol{D} - \boldsymbol{DZ}'\boldsymbol{V}^{-1}\boldsymbol{ZD})$. Hence, the conditional mean of $\boldsymbol{b}$ given $\mathbf{y}$ is $\tilde{\boldsymbol{b}} = \boldsymbol{DZ}'\boldsymbol{V}^{-1}(\mathbf{y} - \boldsymbol{X\beta})$. $\tilde{\boldsymbol{b}}$ is called the best linear unbiased predictor of $\boldsymbol{b}$.

The connection between linear mixed models and cubic smoothing splines is due to Speed (1991) in his discussion to the paper by Robinson (1991) "That BLUP is a Good Thing: The Estimation of Random Effects". The best linear unbiased prediction (Searle *et al.*

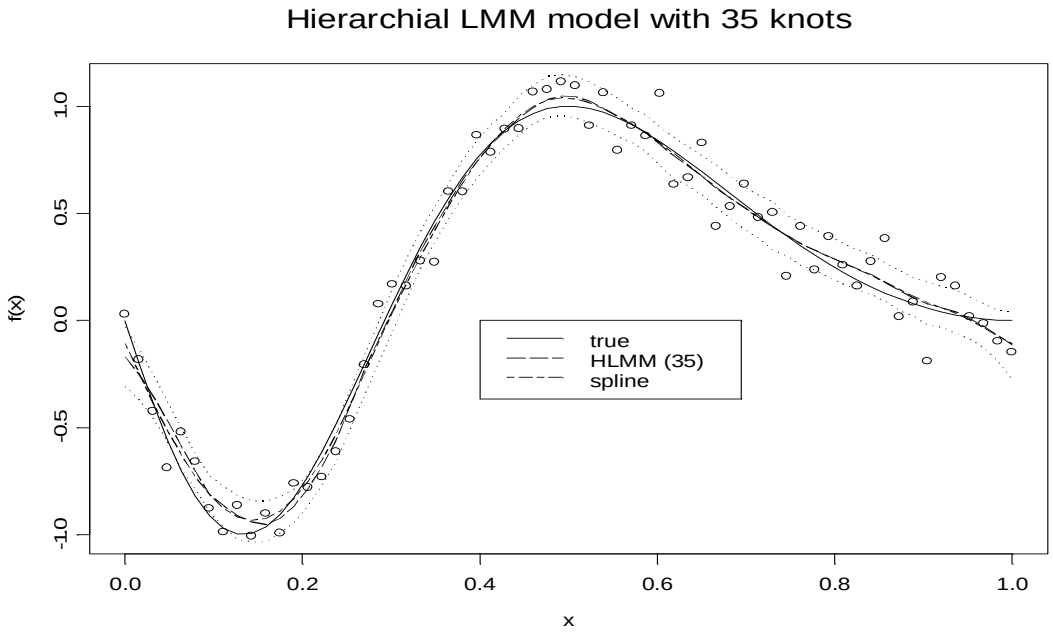


Figure 12.12: *Posterior mean from the hierarchical LMM (fitted with 35 knots) with* 95% *credible intervals.*

1992) can be derived by maximizing the joint density of $\mathbf{y}$ and $\boldsymbol{b}$, $p(\mathbf{y},\boldsymbol{b}) = p(\mathbf{y}|\boldsymbol{b})p(\boldsymbol{b})$, given by

$$(2\pi)^{-\frac{1}{2}n-\frac{1}{2}q}\left(\det\begin{bmatrix}\boldsymbol{D} & 0\\ 0 & \boldsymbol{W}\end{bmatrix}\right)^{-0.5} \times$$

$$\exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{b} & 0\\ 0 & \mathbf{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}\end{pmatrix}'\begin{bmatrix}\boldsymbol{D} & 0\\ 0 & \boldsymbol{W}\end{bmatrix}^{-1}\begin{pmatrix}\boldsymbol{b} & 0\\ 0 & \mathbf{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}\end{pmatrix}\right\}. \qquad (12.24)$$

Maximizing (12.24) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ requires to minimize

$$\begin{pmatrix}\boldsymbol{b} & 0\\ 0 & \mathbf{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}\end{pmatrix}'\begin{bmatrix}\boldsymbol{D} & 0\\ 0 & \boldsymbol{W}\end{bmatrix}^{-1}\begin{pmatrix}\boldsymbol{b} & 0\\ 0 & \mathbf{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}\end{pmatrix}, \qquad (12.25)$$

which leads to the mixed model equations, also called Henderson equations,

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{X}\boldsymbol{W}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{b}} &= \boldsymbol{X}'\boldsymbol{W}^{-1}\mathbf{y},\\ \boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{X}\tilde{\boldsymbol{\beta}} + (\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})\tilde{\boldsymbol{b}} &= \boldsymbol{Z}'\boldsymbol{W}^{-1}\mathbf{y}.\end{aligned} \qquad (12.26)$$

The mixed model equations can be rearranged as

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{Z}+\boldsymbol{D} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\mathbf{y} \\ \boldsymbol{Z}'\boldsymbol{W}^{-1}\mathbf{y} \end{bmatrix}. \tag{12.27}$$

The BLUP estimates for $\boldsymbol{b}$ and $\boldsymbol{\beta}$ are the solution for the simultaneous equations in (12.26): $\tilde{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\mathbf{y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})$ and $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\mathbf{y}$. Note that $\tilde{\boldsymbol{b}}$ is also the empirical Bayes estimate for $\boldsymbol{b}$ (i.e., see page 78 in Verbeke and Molenberghs, 2000).
Now, Speed (1991) showed that the solution of (12.9) has the form of $\hat{f} = \boldsymbol{A}(\lambda)\mathbf{y}$ with the hat matrix

$$\boldsymbol{A}(\lambda) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1} + \boldsymbol{Q}\boldsymbol{V}^{-1}(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}. \tag{12.28}$$

Therefore,

$$\hat{\boldsymbol{f}} = \left[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1} + \boldsymbol{Q}\boldsymbol{V}^{-1}(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right]\mathbf{y},$$

and

$$\hat{\boldsymbol{f}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\mathbf{y} + \boldsymbol{Q}^{\star}\boldsymbol{V}^{-1}(\mathbf{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\mathbf{y}) \tag{12.29}$$

Substituting $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{b}}$ into (12.29) we have

$$\hat{\boldsymbol{f}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Q}\boldsymbol{V}^{-1}(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}). \tag{12.30}$$

Let us consider a linear mixed model with $\boldsymbol{Z} = R = I$ then $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\varepsilon}$. If we further assume that $\boldsymbol{D} = \sigma_{\varepsilon}^2\boldsymbol{Q}/n\lambda$, $\boldsymbol{b} \sim \mathrm{N}(0, \sigma_{\varepsilon}^2\boldsymbol{Q}/n\lambda)$ then the cubic smoothing spline, the solution of the penalized least square problem in (12.9) is a BLUP.
Speed's (1991) formulation for the hat matrix was derived by Wahba (1978), although with the connection to linear mixed models, as she wrote (Wahba 1978, page 367)

"We remark that $\boldsymbol{A}(\lambda)$ is obtained from (2.5) (2.5 in Wahba's paper) and is

$$\boldsymbol{A}(\lambda) = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1} + \boldsymbol{Q}\boldsymbol{M}^{-1}(I - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{M}^{-1}T)^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}".$$

Writing $\boldsymbol{T} = \boldsymbol{X}$, $\boldsymbol{M} = \boldsymbol{V}$ we have $\hat{\boldsymbol{f}} = \boldsymbol{A}(\lambda)\mathbf{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}}$.
It is important to mention that $\boldsymbol{f}$ is not a linear mixed model but $\hat{\boldsymbol{f}}$, the minimizer of (12.9) can be expressed as a BLUP.

12.5.1 The Semiparametric Model of Green and Silverman (1994)

Green and Silverman (1994) discussed a semi parametric model of the form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + g(x_i) + \varepsilon_i \tag{12.31}$$

where g is assumed to be a smooth function, $\boldsymbol{X}_i$ is a design matrix of $p+1$ explanatory variables, x_i is the predictor to which g is related and $\boldsymbol{K}$ is the covariance matrix of the prior distribution of $\boldsymbol{g}$. The random error is assumed to be normally distributed,

$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 \boldsymbol{W})$. In order to avoid confusion in notations we keep the notations of Green and Silverman at this stage and we will change to the LMM notations at the end.
Green and Silverman (1994) decomposed $\boldsymbol{f}$ into two components, $\boldsymbol{f}$ = linear component + smooth component. Note that the semiparametric model allows that x_i are not distinct, since there are other covariates in the model. As we will show later, their setting will be useful if the scatterplot consists of replicate design points, i.e. several values of y are measured in the same design point x. Let $z_1, \ldots, z_s$ be the ordered distinct design points of $x_1, \ldots, x_n$ and $\boldsymbol{N}$ be a $n \times s$ matrix with the entry $N_{ij} = 1$ if $x_i = z_j$ and zero otherwise. The matrix $\boldsymbol{N}$ is called the incidence matrix. It ensures that the smooth part of $\boldsymbol{f}$ will be identical for all value of y sampled at the design point x_i, that is $[\boldsymbol{Ng}]_i$. The penalized sum of square is

$$S(\boldsymbol{\beta}, \boldsymbol{g}) = \sum_{i=1}^{n} w_i (y_i - \boldsymbol{X}_i \boldsymbol{\beta} + [\boldsymbol{Ng}]_i))^2 + \lambda \int_0^1 (g''(x))^2 d(x), \tag{12.32}$$

Green and Silverman (1994) have shown that (12.32) is minimized when $\boldsymbol{\beta}$ and $\boldsymbol{g}$ are the solution for

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{N} \\ \boldsymbol{N}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{N}'\boldsymbol{W}^{-1}\boldsymbol{N} + \lambda \boldsymbol{K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{g} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\mathbf{y} \\ \boldsymbol{N}'\boldsymbol{W}^{-1}\mathbf{y} \end{bmatrix}. \tag{12.33}$$

We now reformulate the semiparametric model as a linear mixed model.

Case I: Distinct Design points

Let us assume that the design points are distinct. In this case $\boldsymbol{N} = I$. We further assume that $\boldsymbol{X}$ is a $n \times 2$ matrix with the ith row $[1, x_i]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Write $\boldsymbol{g} = \boldsymbol{Zb}$ and assume that $\boldsymbol{b} \sim \mathrm{N}(0, \sigma_b^2 I)$. It follows that the model is (12.31) can be expressed as $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$. Note that in our setting there are no covariates except x. The matrix equations in (12.33) become

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{Z} + \lambda \boldsymbol{K} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\mathbf{y} \\ \boldsymbol{Z}'\boldsymbol{W}^{-1}\mathbf{y} \end{bmatrix}, \tag{12.34}$$

where $\boldsymbol{K}$ is the covariance matrix in the prior distribution of $\boldsymbol{b}$. In this case $\hat{g}$, the smooth part of the model, will be estimated by $\boldsymbol{Z}\tilde{\boldsymbol{b}}$, $\boldsymbol{f}$ is estimated by $\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}}$.

Case II: Replicate Design points

In case that scatterplot consists of replicate point, $\boldsymbol{Z}$ is an $s \times s$ matrix. Thus, there are only s random effects in the model. One needs to make sure that all observations, sampled at the same design point, have the same predicted value. This can be done using the incidence matrix $\boldsymbol{N}$. Write $\tilde{\boldsymbol{N}} = \boldsymbol{NZ}$ and $\boldsymbol{g} = \tilde{\boldsymbol{N}}\boldsymbol{b}$ and the matrix equation in (12.33)

can be rewritten as

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}^{-1}\tilde{\boldsymbol{N}} \\ \tilde{\boldsymbol{N}}'\boldsymbol{W}^{-1}\boldsymbol{X} & \tilde{\boldsymbol{N}}'\boldsymbol{W}^{-1}\tilde{\boldsymbol{N}} + \lambda\boldsymbol{K} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{W}^{-1}\mathbf{y} \\ \tilde{\boldsymbol{N}}'\boldsymbol{W}^{-1}\mathbf{y} \end{bmatrix}. \tag{12.35}$$

It is easy to see that in this case the simultaneous equations of the model (i.e. Green and Silverman page 68, equations (4.9) and (4.10)) are identical to the mixed model equations (i.e., Robinson 1991, page 15, equation (1.2)).

12.5.2 Smoothing Correlated Data (Wang Y. 1998)

Wang (1998a) used the relationship between LMM and smoothing spline, in order to smooth correlated data. Similar to Green and Silverman (1994) it is assumed that $\boldsymbol{\varepsilon} \sim (0, \sigma_\varepsilon^2 W)$. Wang (1998a) formulated his model based on results reported in Kimeldrof and Wahba (1970) which showed that the solution of (12.9) has the form of

$$\hat{f}(x) = \sum_{i=1}^{m} d_i \delta_i(x) + \sum_{i=1}^{n} c_i \boldsymbol{Q}(x, t_i), \tag{12.36}$$

with $\boldsymbol{Q}$ and δ_i as defined in Section (12.4.3). Note that this is also the justification of Wahba (1978) for her choice of the prior distribution for $\boldsymbol{f}$. Similar to the previous section we keep the original notation and shift to LMM notations at a later stage. Let $\boldsymbol{T}$ be a $n \times 2$ matrix with the ith row equal to $[1, x_i]$,$i = 1, \ldots, n$, $\boldsymbol{c} = (c_i, \ldots, c_n)$, $\boldsymbol{d} = (d_1, d_2)$, then for cubic smoothing splines ($m = 2$), $\hat{\boldsymbol{f}}$ can be expressed as

$$\hat{f} = \boldsymbol{T}\boldsymbol{d} + \boldsymbol{Q}\boldsymbol{c},$$

where $\boldsymbol{c}$ and $\boldsymbol{d}$ are the solution of

$$\begin{bmatrix} \boldsymbol{T}'\boldsymbol{W}^{-1}\boldsymbol{T} & \boldsymbol{T}'\boldsymbol{W}^{-1} \\ \boldsymbol{Q}'\boldsymbol{W}^{-1}\boldsymbol{T} & \boldsymbol{Q}'\boldsymbol{W}^{-1} + \lambda I \end{bmatrix} \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{Q}\boldsymbol{c} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}'\boldsymbol{W}^{-1}\mathbf{y} \\ \boldsymbol{Q}'\boldsymbol{W}^{-1}\mathbf{y} \end{bmatrix}$$

Now, as Wang (1998) argued, one can consider three different linear mixed models for which $\hat{\boldsymbol{f}}$ can be expressed as a BLUP.

- **Model 1**
  $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{b}_1 + \boldsymbol{\varepsilon}$, where $\boldsymbol{T} = \boldsymbol{X}$, $\boldsymbol{d} = \boldsymbol{\beta}$,$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and $\boldsymbol{b}_1$ is a vector of random effects distributed as $\boldsymbol{b}_1 \sim \mathrm{N}(0, \sigma_\varepsilon^2 \boldsymbol{Q}/n\lambda)$.
- **Model 2**
  $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{b}_2 + \boldsymbol{\varepsilon}$, where $\boldsymbol{T} = \boldsymbol{X}$, $\boldsymbol{d} = \boldsymbol{\beta}$,$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 W)$, $\boldsymbol{b}_2 \sim \mathrm{N}(0, \sigma_\varepsilon^2 \boldsymbol{Q}^+/n\lambda)$ and $\tilde{\boldsymbol{Z}} = \boldsymbol{Q}$. The matrix $\boldsymbol{Q}^+$ is the general inverse of $\boldsymbol{Q}$ so that $\boldsymbol{Q}\boldsymbol{Q}^+\boldsymbol{Q}' = \boldsymbol{Q}$.
- **Model 3**
  $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}_3 + \boldsymbol{\varepsilon}$, where $\boldsymbol{T} = \boldsymbol{X}$, $\boldsymbol{d} = \boldsymbol{\beta}$,$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2 W)$, $\boldsymbol{b}_3 \sim \mathrm{N}(0, \sigma_\varepsilon^2 I/n\lambda)$ and $\boldsymbol{Z}\boldsymbol{Z}' = \boldsymbol{Q}$.

One can observe that model 1 is the same model that was assumed by Speed (1991).

12.5.3 Contracting The Design Matrix for the Random Effects

In previous sections we assumed that there exists a matrix $\boldsymbol{Z}$ such that $\boldsymbol{Z}\boldsymbol{Z}' = \boldsymbol{Q}$ and the LMM was defined using the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Q}$. In this section we present $\boldsymbol{Q}$ and construct the design matrix for the random effects $\boldsymbol{Z}$.
For cubic smoothing splines ($m = 2$) the ijth entry of $\boldsymbol{Q}(s,t)$ is

$$Q(x_i, x_j) = \int_0^1 (x_i - u)_+ (x_j - u)_+ du, \tag{12.37}$$

where $(s-u)_+$ is the truncated power function defined in (12.6). One can approximate the integral with the sum, the ijth entry of $\boldsymbol{Q}$ becomes

$$\tilde{Q}(x_i, x_j) = \sum_{k=1}^{n} (x_i - t_k)_+ (x_j - t_k)_+. \tag{12.38}$$

Note that in this stage we assume that the number of knots is n and the knots are located at the design points. Define an $n \times n$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} (x_1 - t_1)_+ & (x_1 - t_2)_+ & \dots & (x_1 - t_n)_+ \\ (x_2 - t_1)_+ & (x_2 - t_2)_+ & \dots & (x_2 - t_n)_+ \\ . & . & \dots & . \\ . & . & \dots & . \\ (x_n - t_1)_+ & (x_n - t_2)_+ & \dots & (x_n - t_n)_+ \end{bmatrix} \tag{12.39}$$

It follows that $\boldsymbol{Z}\boldsymbol{Z}' = \tilde{\boldsymbol{Q}}$.

Reducing the Number of Knots

Wand (2002) used a model in which the number of basis function, the number of knots K is smaller than n, $K < n$. In this case, $\boldsymbol{Z}$ is an $n \times K$ matrix. The knots, in Wand's model were located at the $(k + 1/K + 2)$,$k = 1 \dots, K$, quantiles of the unique values of x. In our setting the design points are distinct and we choose K knots equally spaced between the 0.01 to 0.99 quantiles of x. Wand (2002) recommends to use $K = \min(n/4, 35)$. The Splus function `smooth.spline()` used the same criterion when $n > 50$. For $n \leq 50$ the number of knots is equal to n and the knots are located at each design point. In Section 12.7 we present a simulation study to investigate the influence of the number of knots on the mean square error of $\hat{\boldsymbol{f}}$ and on the parameter estimate for σ_b.

12.6 From Cubic Smoothing Splines to Linear Mixed Models and Vice Versa

12.6.1 The Relative Precision Factor in a Single Level Linear Mixed Model

The linear mixed model for the bodyweight data, presented in Section 12.2, can be seen as a special case of a multilevel model with a single level. By a single level we mean a single factor which defines the clusters. In the bodyweight example a cluster is a rat. A linear mixed model for scatterplot smoothing is a single level model as well since there is only one variance component associated with the random effects. The relative precision factor, Δ, (Pinherio and Bates 2000) is given by

$$\Delta'\Delta = \frac{\boldsymbol{D}^{-1}}{\frac{1}{\sigma_\varepsilon^2}}.$$

The precision of the random effects is expressed relative to the precision of the measurement error. Pinherio and Bates (2000) show that the matrix Δ, if $\boldsymbol{D}$ is positive definite, exists.

The Bodyweight Example

For the bodyweight example the variance of b_i, $i = 1, \ldots, 16$ is a scalar σ_b^2. The relative precision factor is

$$\Delta = \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_b^2}}.$$

Note that Δ is closely related to the inter-cluster correlation, As Δ approaches zero the inter-cluster correlation approaches 1.

Smoothing Scatterplot

When the LMM is used as a smoother, Δ is simply the smoothing parameter multiplied by $1/n$.

$$\lambda = \frac{1}{n}\Delta^2 = \frac{\sigma_\varepsilon^2}{n\sigma_b^2}.$$

As σ_b^2 increases (for fixed n and σ^2) the smoothing parameter decreases and the BLUP has a tendency to change rapidly. As σ_b^2 decreases the value of the smoothing parameter increase and the estimated model will become "more smooth" Figure 12.13 illustrates the relationship between σ_b^2 and the smoothness of the BLUP. De Boor (1978) defined the penalized least square problem as a weighted average between the residuals sum of squares and the squared integral, $S(\boldsymbol{f}) = (1-p)\sum_i^n (y_i - f(x_i))^2 + p\int_0^1 (f^{(m)}(x))^2 d(x)$, $0 \leq p \leq 1$. Note that for $\lambda = p/(1-p)$ De Boor's penalized least square is identical to the definition in (12.9). Since

$$\lambda + 1 = \frac{n\sigma_b^2 + \sigma_\varepsilon^2}{n\sigma_b^2} = \frac{1}{1-p},$$

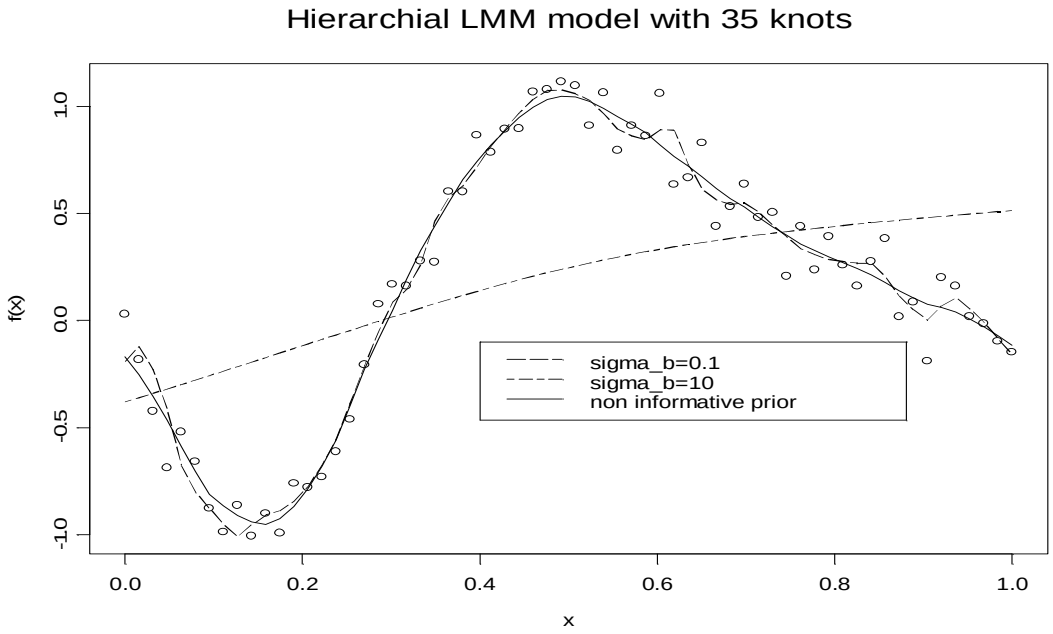


Figure 12.13: *Hierarchical linear mixed model with 35 knots fitted with two fixed values of* σ_b^2. *Long-dashed lines:*$\sigma_b^2 = 0.1$. *Short/long-dashed line:*$\sigma_b^2 = 10$. *Solid line: a non informative hyperprior distribution was specified for* σ_b^2.

we can define the tradeoff proportion TPE as a function of the variance components in the LMM,

$$\text{TPE} = 1 - \frac{n\sigma_b^2}{n\sigma_b^2 + \sigma_\varepsilon^2}.$$

TPE is the proportion explaining how much weight the fitted model puts on the smoothness penalty. As σ_b decreases (for fixed σ_ε) the weight of the squared integral in $s(\boldsymbol{f})$ increases, so the BLUP will be more smooth.

12.6.2 Estimating the smoothing parameter (revisited)

In Section 12.4.2 we discussed the GCV methods as an automatic selection procedure for the smoothing parameter. When a linear mixed model is fitted as a scatterplot smoother one does not need to use any additional procedure in order to select the smoothing parameter. The amount of the smoothing is determined by the ratio of the maximum likelihood estimates for both σ_ε^2 and σ_b^2. Figure 12.14 presents the plot for the $GCV(\lambda)$ and $R(\lambda)$. Two new points appear in this figure: B which represents the ratio between the maximum likelihood estimated (divided by n) from the linear mixed model which was fitted with the function `lme()` in Splus. The point C represents the posterior mean for λ when a full Bayesian hierarchical model was fitted (in WinBugs). Figure 12.15 shows the density estimate for the posterior distribution of λ with the 4 estimated values.

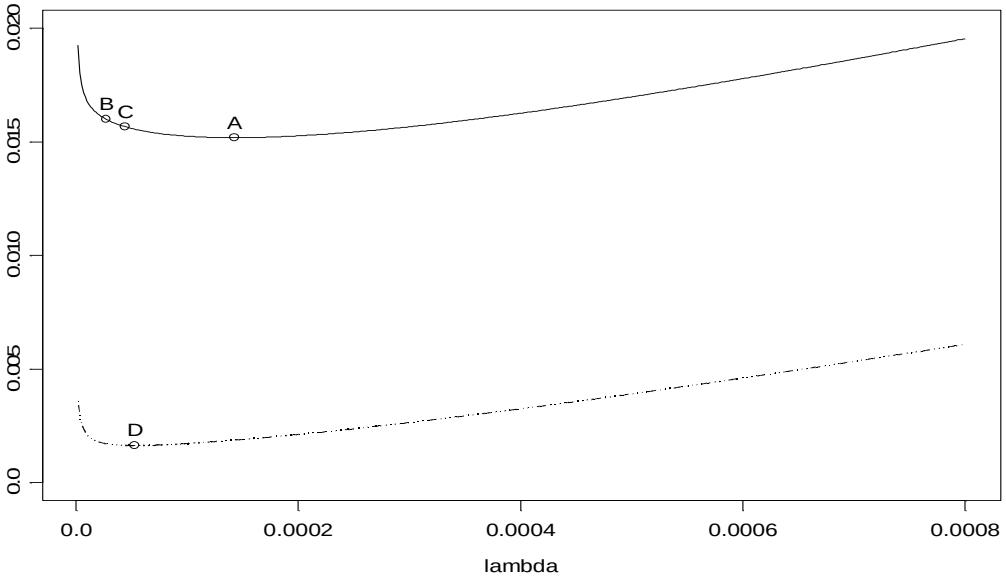


Figure 12.14: *GCV (solid line) and $\hat{R}$ (dotted dashed line). A-the minimizer of GCV. B-the ratio between the variance components in the LMM (divided by n). C-posterior mean of λ. D-the minimizer of $\hat{R}$.*

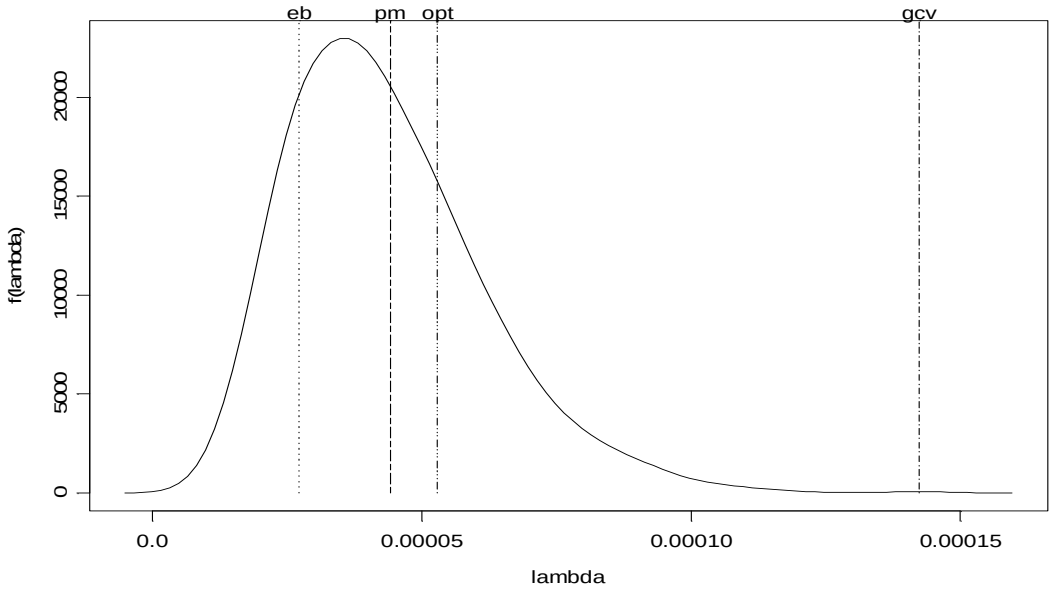


Figure 12.15: *Density estimate for the posterior distribution of λ. eb=the ratio of the variance components in the LMM divided by n, pm=posterior mean of the full Bayesian model, gcv= the estimate for λ which minimizes the GCV score and opt is the minimizer the $R(\lambda)$*

12.6.3 Linear Mixed Models as Kernel Smoothers

Cubic smoothing splines belong the family of linear smoothers. As such, $\hat{\boldsymbol{f}}$ can be expressed as a linear combination of the data, $\hat{\boldsymbol{f}} = \boldsymbol{A}(\lambda)\mathbf{y}$. Hence, the predicted value at x_l can be written as $\hat{f}(x_l) = \sum_{i=1}^{n}[A(\lambda)]_{li}y_i$. The weight $A(\lambda)_{li}$ which associated with the fit at x_0 is called the equivalent kernel at x_0 (Hastie and Tibshirani 1990). In the context of cubic smoothing splines, Silverman (1985) and Green and Silverman (1994) have shown that for large n there exists a weight function $G(x,t)$ such that $\hat{f}(s) = \sum_{i=1}^{n} y_i G(s,x_i)/n$, were $G(s,x_i)$ is a kernel function given by

$$G(s,x) = \frac{1}{g(x)}\frac{1}{h(x)}K\left(\frac{s-x}{h(x)}\right). \tag{12.40}$$

Here, $g(x)$ is the local density of the design points, the local bandwidth $h(x)$ satisfies $h(x) = \lambda^{0.25}n^{-0.25}g(x)^{-0.25}$ and the kernel function is given by

$$K(u) = 0.5\exp(-|u|/\sqrt{2})sin(|u|/\sqrt{2} + \pi/4).$$

Since $\hat{\boldsymbol{f}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{b}$, the hat matrix in (12.28) should be a kernel matrix which asymptotically should be equal to $G(s,x)$. Figure 12.16 shows the asymptotic equivalent kernel for x_{32} (as defined in equation 12.40) corresponding with row 32 of the hat matrix (as defined in equation 12.28). Figure 12.17 shows the rows of the hat matrix plotted against x. It illustrates how the symmetric kernel peakes around the design points and moves along the range of x. Note that since λ is constant and the data are equally spaced, i.e. the density $g(x)$ is a constant, the width of the equivalent kernel function remains constant along the range of x. This can also be seen in Figure 12.18 which presents the equivalent kernel for x_{10}, x_{20}, x_{30} and x_{40} which obtain from the LMM model (solid line) and from the minimizer of $S(\boldsymbol{f})$ in (1.9). Note that we use the same smoothing parameter in both cases.

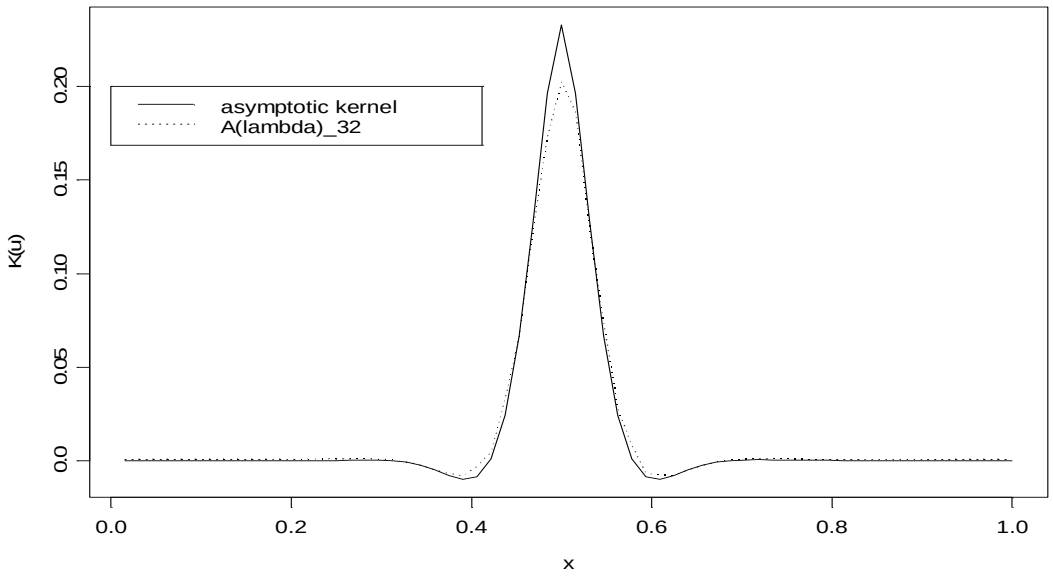


Figure 12.16: *Asymptotic and actual kernel at* x_{32}. $\hat{\lambda} = 2.03 \times 10^{-5}$,$\hat{\sigma}_b^2 = 6.698$ *and* $\hat{\sigma}_\varepsilon^2 = 8.71 \times 10^{-3}$. *Solid line: asymptotic kernel. Dotted line row 32 of the LMM's hat matrix.*

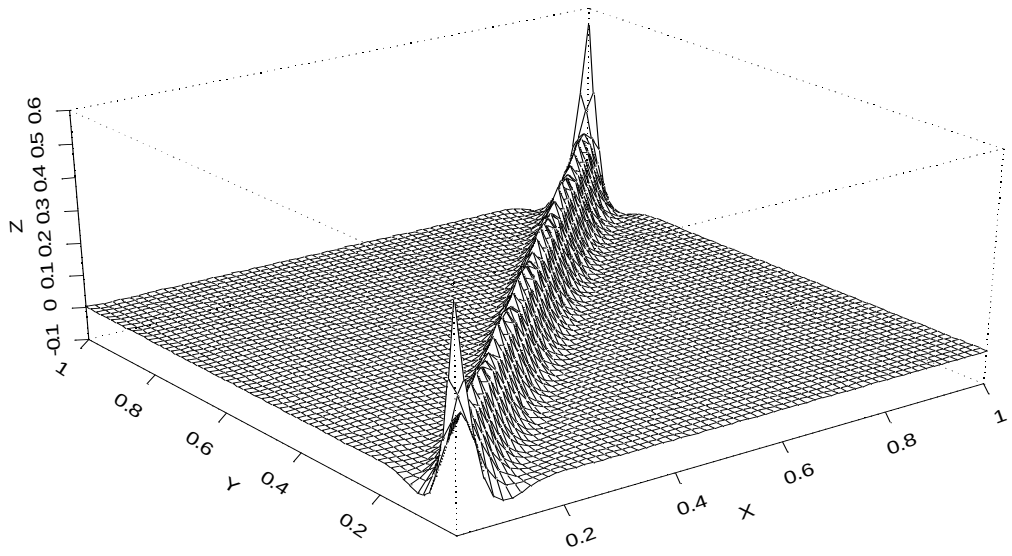


Figure 12.17: *The hat matrix of the LMM with 64 knots.*

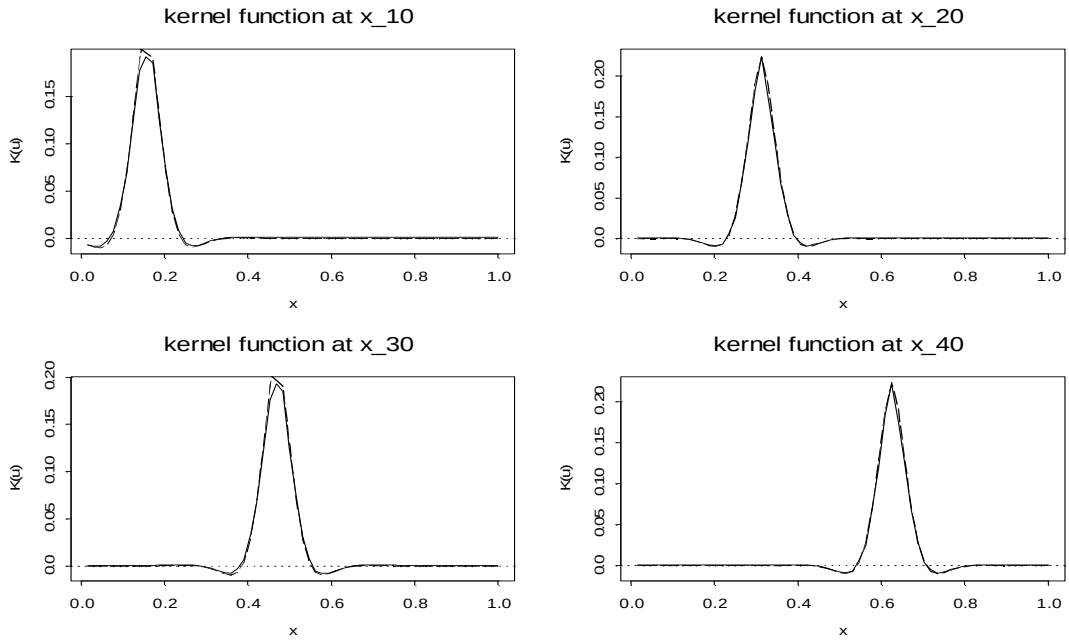


Figure 12.18: *The rows of the hat matrix for x_{10},x_{20},x_{30} and x_{40}. Solid line: rows of the hat matrix of the LMM. Dashed lines: row of $A(\lambda)$ as discussed in Section 12.4.1*

12.7 Simulation Study

A simulation study was conducted in order to investigate the influence of the number of knots on the estimate for $\boldsymbol{f}$. The test functions used in the simulation study were the same as those presented by Craven and Wahba (1979):

1. $f_1(x) = 0.2\beta_{4,15}(x) + 0.7\beta_{5,7}(x) + 0.1\beta_{12,5}(x)$,
2. $f_2(x) = 0.4\beta_{12,7}(x) + 0.6\beta_{4,11}(t)$,
3. $f_3(x) = 0.5\beta_{10,30}(x) + 0.2\beta_{20,20}(x) + 0.3\beta_{30,10}(x)$.

where $\beta_{p,q}$ is the beta density function:

$$\beta_{p,q}(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1} \quad .$$

Figure 12.19 shows the test functions. The designs points x_i were equally spaced, $x_i = \frac{i-1}{n-1}$ and satisfy $x_1 = 0$ and $x_n = 1$.
The simulation study consists of all 3×3 combinations of test function 1,2,3 and $\sigma =$ 0.1, 0.2 and 0.5. Sample size at each experiment was 64 and 250 simulations were conducted for each combination of test function and σ. For each simulated dataset we fitted three linear mixed models with 10, 20 and 35 knots and a smoothing spline with 64 knots (using the option `all.knots=T` in the Splus function `smooth.spline()`).
The performance of the four models (the three LMMs and the smoothing spline) was evaluated by comparing the mean square error, variance and squared bias of $\hat{\boldsymbol{f}}$ and by

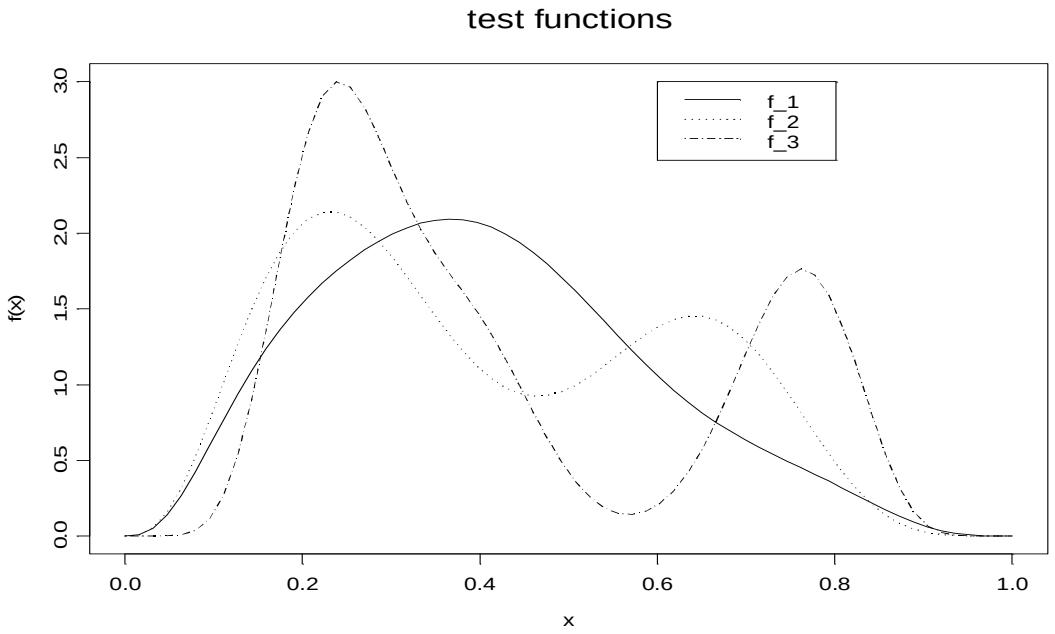


Figure 12.19: *Test functions for the simulation study.*

comparing the parameter estimate for σ.
We estimate the MSE of $\hat{\boldsymbol{f}}$ by $\widehat{MSE}(\hat{\boldsymbol{f}}) = \widehat{bias}^2(\hat{\boldsymbol{f}}) + \widehat{Var}(\hat{\boldsymbol{f}})$. The local squared bias at the point x_i was estimated by $\widehat{bias}_i^2 = (\bar{\hat{f}}(x_i) - f(x_i))^2$, where $\bar{\hat{f}}(x_i) = \sum_{j=1}^{250} \hat{f}_j(x_i)/250$ is the average value of $\hat{f}(x_i)$ over the simulations (j is the simulation number and $\hat{f}_j(x_i)$ is the value of $\hat{f}(x_i)$ in the $j'th$ simulation). The global squared bias was estimated by the average value of the local squared bias $\widehat{bias}^2(\hat{\boldsymbol{f}}) = \sum_{i=1}^{64} \widehat{bias}_i^2/64$. The local variance at the point x_i was estimated by $\widehat{Var}(\hat{f}(x_i)) = \sum_{j=1}^{250}(\hat{f}_j(x_i) - \bar{\hat{f}}(x_i))^2/250$. The global variance was estimated by $\widehat{Var}(\hat{f}) = \sum_{i=1}^{64} \widehat{Var}(\hat{f}(x_i))/64$. For σ_ε we estimate the variance by $\widehat{Var} = \sum_{j=1}^{250}(\hat{\sigma}_{\varepsilon,j} - \bar{\hat{\sigma}}_\varepsilon)/250$ and $\widehat{bias}^2 = (\bar{\hat{\sigma}}_\varepsilon - \sigma_\varepsilon)^2$.

12.7.1 Simulation Results

Global results for the performance of $\hat{\boldsymbol{f}}$ are reported in Table 12.1. Figure 12.22 shows the global MSE in the 9 combinations in the study. It seems that the choice of $K = 10$ is not satisfactory and leads to higher values of MSE compared to $K = 20$ and 35. There is not much difference between the performance of the LMMs when $K = 20$ or 35. It is somewhat surprising when the performance of the LMM models with 20 and 35 knots is compared with the performance of the smoothing spline with 64 knots. For $\sigma > 0.1$ the LMMs perform better. This suggests that compared to the GCV estimate of λ, the estimates for λ obtained by maximizing the likelihood of the LMM leads to a smaller value of $MSE(\hat{\boldsymbol{f}})$. This issue should be investigated further by more extensive simulation study. Figures 12.23-12.27 shows the local MSE, variance and squared bias of the 4 models for some of the experiment combination in the study. The model with $K = 10$ has the poorest local performance. Note that for the second and third test functions, for $\sigma = 0.5$

(Figure 12.25 and 12.26) the local variance of $\hat{\boldsymbol{f}}_{GCV}$ is higher then the local variance of $\hat{\boldsymbol{f}}_{LMM}$ along the range of x. This is the reason why the global MSE of $\hat{\boldsymbol{f}}_{GCV}$ was higher than the global MSE of the estimates obtained from the LMMs.
The results for the estimates for σ_ε under the 4 models are shown in Table 12.2 and Figure 12.28. The LMM with $K = 10$ performed poorly for $\sigma_\varepsilon < 0.5$. In terms of $MSE(\hat{\sigma}_\varepsilon)$ the LMMs with $K = 20, 35$ performed better than the smoothing spline with $K = 64$.

12.8 Application to the Data

We applied the method for the observational errors in band 1A of α-Bootis. The hierarchical Bayesian model was formulated for log(STDEV-tag^2) and log(SPARE-tag^2) separately, in the same way as described in Section 12.4.3 (equation 12.22). Figure 12.20 shows the posterior means and 95% credible intervals for log(STDEV-tag^2) and log(SPARE-tag^2). These posterior means will be used, in Chapter 13, as an input for the hierarchical Bayesian model for the spectrum,

$$\begin{cases} \text{observed spectrum} \sim \text{N}(\text{true spectrum}, \text{STDEV-tag}^2) \\ \text{true spectrum} \sim \text{N}(\text{synthetic spectrum}, \text{SPARE-tag}^2) \end{cases}$$

In this model the ratio SPARE-tag/(STDEV-tag + SPARE-tag) is the shrinkage factor in the model. A bivariate model was formulated for log(STDEV-tag) and log(SPARE-tag),

$$\begin{bmatrix} z_{1_i} \\ z_{2_i} \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta}_1 + \boldsymbol{Z}_i\boldsymbol{b}_1 \\ \boldsymbol{X}_i\boldsymbol{\beta}_2 + \boldsymbol{Z}_i\boldsymbol{b}_2 \end{bmatrix}, \Sigma \right).$$

where z_{1_i} and z_{2_i} are log(STDEV-tag^2) and log(SPARE-tag^2) at wavelength t_i respectively. We assumed that Σ is a diagonal matrix with the entries σ_1^2 and σ_2^2. For the random effects it was assumed that $\boldsymbol{b}_1 \sim \text{N}(0, \sigma_{b_1}^2 I)$ and $\boldsymbol{b}_2 \sim \text{N}(0, \sigma_{b_2}^2 I)$. At each iteration of the MCMC simulation the shrinkage factor was monitored. Figure 12.21 shows the posterior mean for the shrinkage factor with 95% credible intervals. In general, the shrinkage factor increases with wavelength which implies that at the end of band 1A the dominant term in the posterior mean of the spectrum will be the observed spectrum.

12.9 Discussion

In her chapter "Late night thoughts of a classical astronomer" (Babu and Feigelson 1997), Virginia Timble quoted George Gamow who said "with five parameters you can fit an elephant". Paraphrasing Gamow's words we can summarize this Chapter and say: "with four parameters you can smooth a scatterplot".
From a linear mixed model point of view the number of parameters in the models used to smooth the scatterplots in this chapter is four: β_0, β_1, σ_ε and σ_b. If the model is fitted

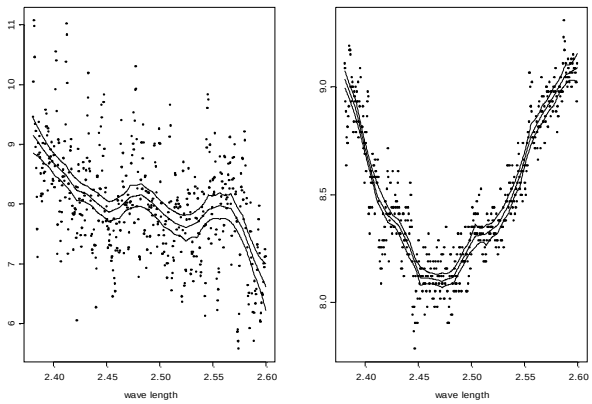


Figure 12.20: Observational errors in band 1A of α-Bootis (on log scale). 'STDEV-tag' with the posterior mean of the mixed model and 95% credible intervals (left) and 'SPARE-tag' with posterior mean and 95% credible intervals (right).

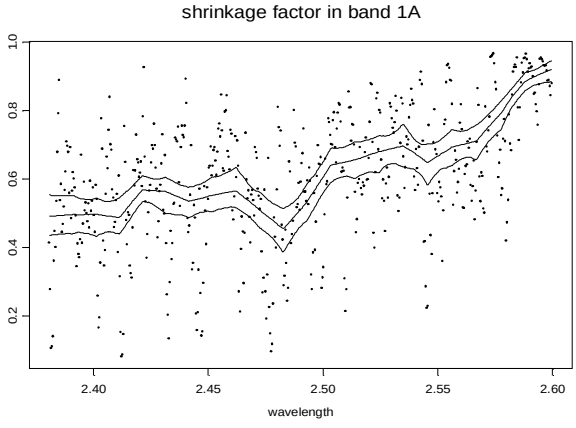


Figure 12.21: *Posterior mean for the shrinkage factor in band 1A and* 95% *credible intervals.*

using a full Bayesian model the number of parameters increases by one, the variance in the hyperprior distribution of $\boldsymbol{\beta}$ (assuming non informative hyperprior distribution for ξ, σ_b and σ_ε). We have shown that the amount of smoothing is determined by the ratio of the variance components divided by the sample size. This means that two models with two different smoothing parameters have the same number of parameters. This view is not acceptable within the framework of nonparametric regression. The smoothing parameter is not just there to control the tradeoff between goodness-of-fit and smoothness but also determines the number of parameters in the fitted model via $tr(\boldsymbol{A}(\lambda))$. As λ decreases the model becomes more complex in the sense that more degrees of freedom are needed to fit the model. From this point of view, the number of parameter in the LMM is the trace of the hat matrix presented in equation (12.29).
Whatever the number of parameters is, four or more, we have shown in this chapter that LMM can be used as scatterplots smoother with automatic choice of λ.
The use of non informative hyperprior distribution in the hierarchical model turns out to be a "good thing". It simply gives the data, observed and unobserved, more chance to speak for themselves in choosing the model to be fitted.

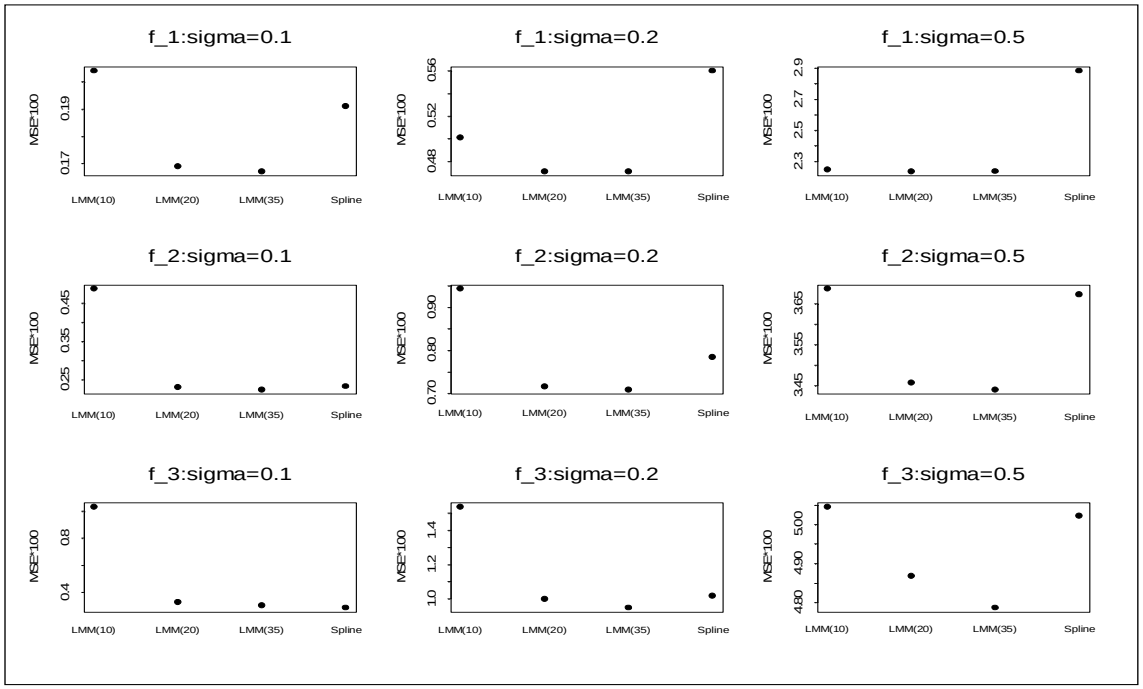


Figure 12.22: *Global MSE for the 9 experiments.*

Table 12.1: *Simulation results for $\hat{\boldsymbol{f}}$: global simulated squared bias, variance and mean squared error ($\times 10^3$).*

Test function	σ	Model	Bias^2	Var	MSE
1	0.1	LMM(10)	0.204	0.128	0.077
		LMM(20)	0.169	0.139	0.030
		LMM(35)	0.167	0.139	0.028
		Spline	0.191	0.159	0.032
	0.2	LMM(10)	0.501	0.407	0.094
		LMM(20)	0.471	0.412	0.059
		LMM(35)	0.471	0.413	0.058
		Spline	0.560	0.497	0.063
	0.5	LMM(10)	2.246	2.046	0.200
		LMM(20)	2.233	2.054	0.180
		LMM(35)	2.235	2.058	0.177
		Spline	2.882	2.692	0.190
2	0.1	LMM(10)	0.487	0.156	0.331
		LMM(20)	0.231	0.203	0.027
		LMM(35)	0.224	0.207	0.017
		Spline	0.233	0.204	0.029
	0.2	LMM(10)	0.943	0.556	0.387
		LMM(20)	0.715	0.632	0.083
		LMM(35)	0.708	0.635	0.073
		Spline	0.784	0.699	0.085
	0.5	LMM(10)	3.686	2.826	0.860
		LMM(20)	3.457	2.858	0.599
		LMM(35)	3.440	2.859	0.581
		Spline	3.672	3.333	0.339
3	0.1	LMM(10)	1.030	0.159	0.871
		LMM(20)	0.327	0.258	0.069
		LMM(35)	0.301	0.289	0.012
		Spline	0.284	0.261	0.023
	0.2	LMM(10)	1.536	0.613	0.923
		LMM(20)	0.999	0.865	0.135
		LMM(35)	0.947	0.874	0.073
		Spline	1.017	0.919	0.099
	0.5	LMM(10)	5.045	3.487	1.559
		LMM(20)	4.868	3.778	1.090
		LMM(35)	4.786	3.779	1.007
		Spline	5.022	4.465	0.558

Table 12.2: *Simulation results for $\hat{\sigma}_\varepsilon$: global simulated squared bias, variance and mean squared error ($\times 10^3$).*

Test function	σ	Model	Mean($\hat{\sigma}$)	Bias^2	Var	MSE
1	0.1	LMM(10)	0.102	0.003	0.110	0.114
		LMM(20)	0.098	0.006	0.106	0.112
		LMM(35)	0.097	0.008	0.107	0.115
		Spline	0.097	0.007	0.133	0.140
	0.2	LMM(10)	0.198	0.006	0.362	0.368
		LMM(20)	0.195	0.024	0.371	0.395
		LMM(35)	0.195	0.026	0.374	0.399
		Spline	0.194	0.038	0.602	0.640
	0.5	LMM(10)	0.487	0.174	2.109	2.283
		LMM(20)	0.486	0.197	2.100	2.296
		LMM(35)	0.486	0.203	2.091	2.295
		Spline	0.482	0.319	2.935	3.254
2	0.1	LMM(10)	0.115	0.231	0.098	0.329
		LMM(20)	0.097	0.008	0.092	0.100
		LMM(35)	0.095	0.025	0.102	0.127
		Spline	0.096	0.013	0.132	0.145
	0.2	LMM(10)	0.206	0.040	0.401	0.442
		LMM(20)	0.194	0.039	0.413	0.451
		LMM(35)	0.192	0.056	0.420	0.476
		Spline	0.193	0.052	0.497	0.549
	0.5	LMM(10)	0.499	0.001	2.349	2.350
		LMM(20)	0.494	0.041	2.364	2.405
		LMM(35)	0.493	0.050	2.355	2.406
		Spline	0.489	0.121	2.739	2.861
3	0.1	LMM(10)	0.141	1.678	0.152	1.830
		LMM(20)	0.101	0.001	0.108	0.109
		LMM(35)	0.092	0.059	0.095	0.154
		Spline	0.094	0.040	0.175	0.215
	0.2	LMM(10)	0.220	0.386	0.429	0.814
		LMM(20)	0.194	0.038	0.379	0.418
		LMM(35)	0.189	0.130	0.399	0.529
		Spline	0.188	0.132	0.672	0.804
	0.5	LMM(10)	0.506	0.038	2.348	2.386
		LMM(20)	0.495	0.030	2.518	2.548
		LMM(35)	0.492	0.060	2.549	2.609
		Spline	0.485	0.235	3.675	3.910

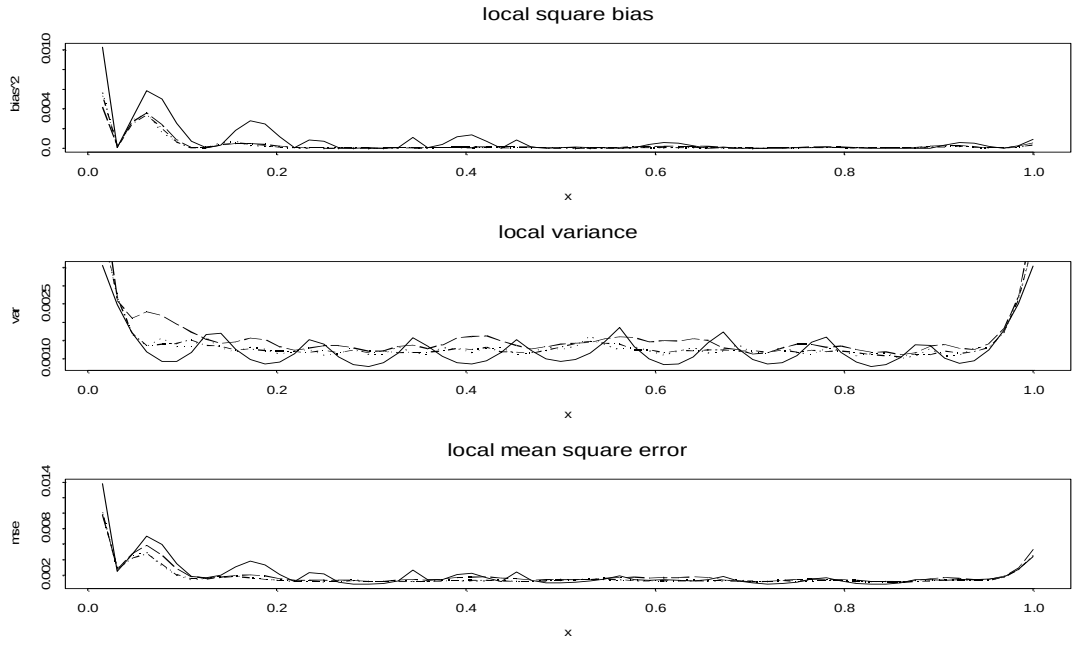


Figure 12.23: *Simulation results for test function 1. Local squared bias, variance and mean square error.* $\sigma = 0.1$. *Solid line: LMM with 10 knots. Dotted line: LMM with 20 knots. Dotted-dashed line: LMM with 35 knots. Long dashed line: smoothing spline with 64 knots. The same types of lines are used Figures 12.24- 12.27.*

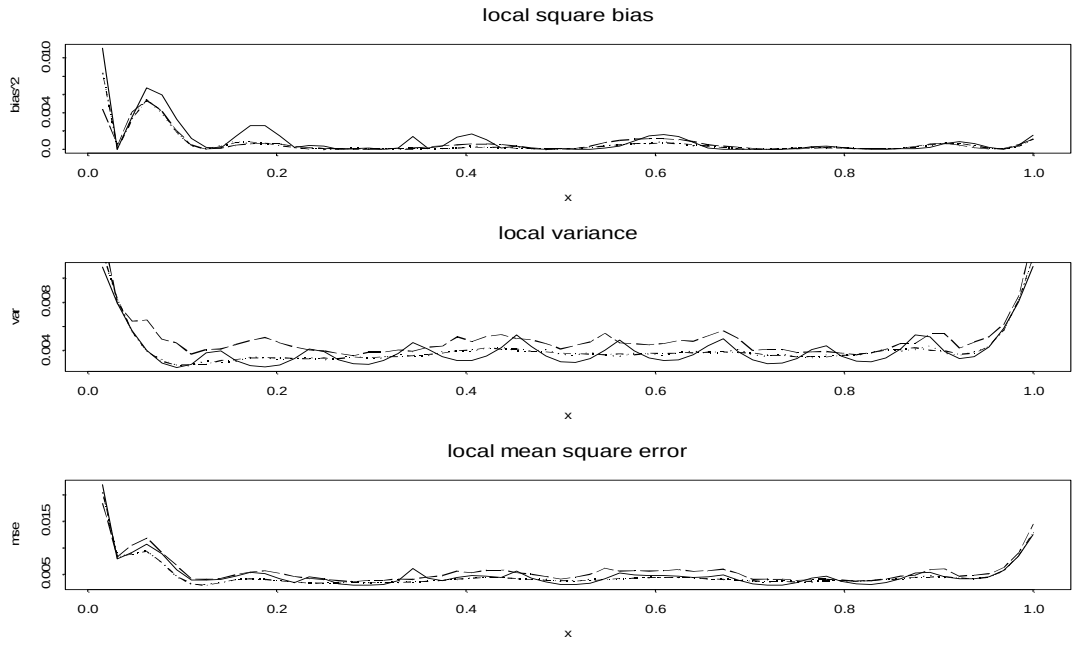


Figure 12.24: *Simulation results for test function 1. Local squared bias, variance and mean square error.* $\sigma = 0.2$.

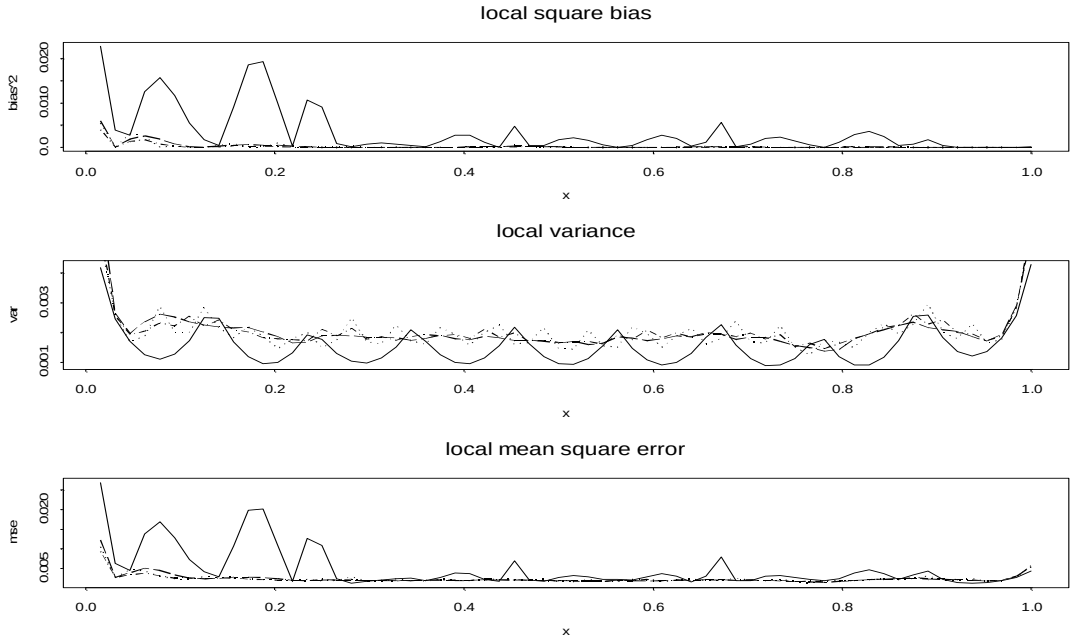


Figure 12.25: *Simulation results for test function 1. Local squared bias, variance and mean square error.* $\sigma = 0.5$.

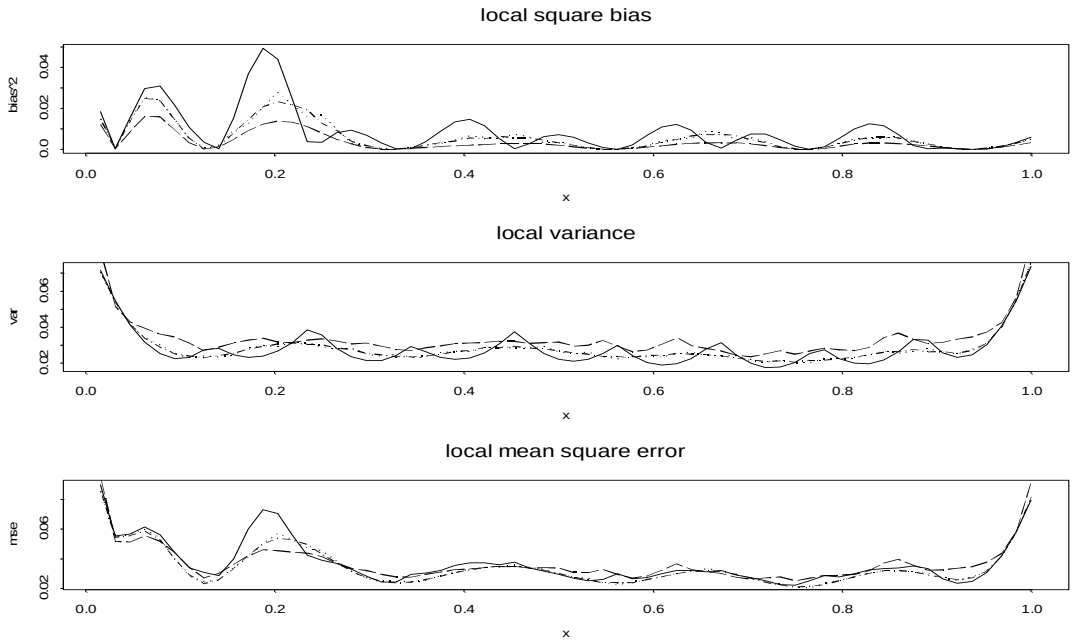


Figure 12.26: *Simulation results for test function 2. Local squared bias, variance and mean square error.* $\sigma = 0.5$.

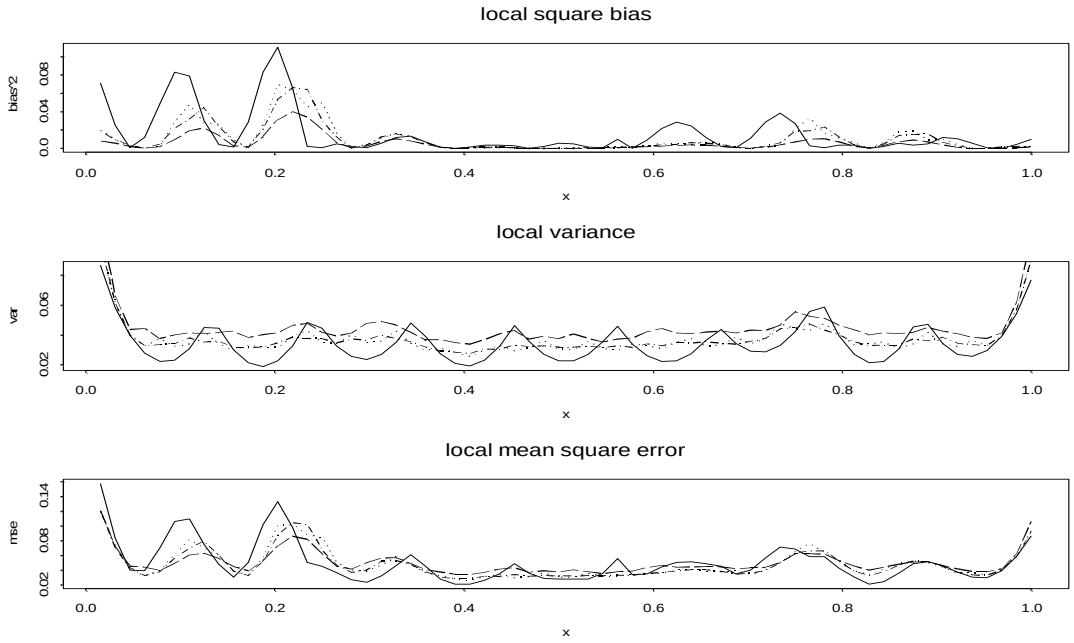


Figure 12.27: *Simulation results for test function 3. Local squared bias, variance and mean square error.* $\sigma = 0.5$.

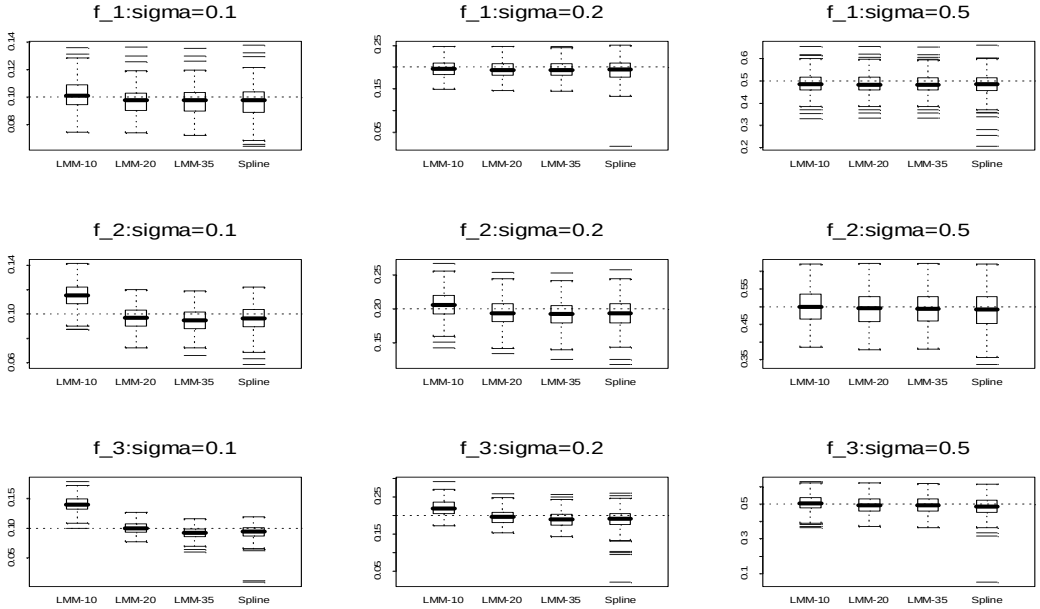


Figure 12.28: *Simulation results for* σ.

Chapter 13

Estimating Stellar Parameters - Hierarchical Bayesian Approach

13.1 Introduction

In this chapter we focus on hierarchical Bayesian models for the spectrum. We wish to construct the posterior distribution of the spectrum by combining the likelihood for the observed spectrum (i.e., the rebinned data), with the prior distribution of the true spectrum.
The attention is placed on hierarchical Bayesian model in which the likelihood for the observed spectrum is specified in the first level and the prior distribution for the spectrum is specified in the second level. In this prior distribution, the synthetic spectrum is used as the prior mean. In the third level of the model we specify the probability models for the observational errors using the models discussed in Chapter 12. Hence, we assume

$$\begin{aligned} &\text{observed spectrum} \sim N(\text{true spectrum}, \sigma^2), && \text{1st level,} \\ &\text{true spectrum} \sim N(\text{synthetic spectrum}^m, \sigma_M^2), && \text{2nd level,} \\ &\sigma_M^2 \sim F_{\sigma_M} \text{ and } \sigma^2 \sim F_\sigma, && \text{3rd level.} \end{aligned}$$

Here, m is the index for the m'th synthetic spectrum, $m = 1, 2, \ldots, 125$. This model accounts for both 'STDEV-tag' and 'SPARE-tags', the first is used as the variance of the likelihood and the second as the variance of the prior model.
In Section 13.2 we will discuss in details the hierarchical Bayesian model for the spectrum. The posterior distribution of the spectrum is discussed in Section13.3. We present the "full" model for the spectrum and discuss the practical problems related to the implementation of this model. A "reduced" model, which uses the synthetic spectrum as the mean of the prior distribution, is used for model selection.
One has to realise that the final synthetic spectrum θ is not the result of the computation of a close analytical expression. Instead, the radiative transfer calculation requires a model atmosphere as input, the synthetic spectrum is the output which is obtained from a convergence process in order to obtain hydrostatic equilibrium and to fulfill the

conservation law of radiative (and convective) energy. When this would not have been the case, i.e. when we could have written $\mu(t) = h(\mathrm{Teff}, \log \mathrm{g}, [\mathrm{Fe/H}], t)$, with h representing an analytical function, then we could have estimated Teff, log g and [Fe/H] directly from the observational spectrum. Instead, we have to calculate each time a synthetic spectrum (and a theoretical model) for a set of $\Omega^{(m)}$ and assess the goodness-of-fit.
Our modeling approach in this chapter is to fit hierarchical Bayesian models for a collection of synthetic specra. Thus, some model selection procedure is needed. We will follow the idea proposed by Laud and Gelfand (1998), who suggested to compare the observed data (y) and hypothetical data (called *replicated data*) which sampled from the posterior predictive distribution. The model which minimizes a predictive discrepancy measure is selected. This will be discussed in Section 13.4.
We apply the method to the $2.38 - 4.08$ ISO-SWS spectrum of the K2IIIp star Alpha Bootis (Arcturus, HD 124897). In the remainder of this section we will discuss the main concepts of Markov Chain Monte Carlo simulations.

13.1.1 Bayesian Inference

Within the Bayesian framework inference is based on the posterior distribution of the unknown parameters in the model given the data. This distribution can be derived analytically (as in the above example) or can be approximated using the Markov Chain Monte Carlo (MCMC) algorithm. A single iteration of the MCMC algorithm (Gilks 1996) consists of sampling the unknown parameters in the models from their full conditional distribution, given the current value of the other parameters in the model and the data. Assume that the distribution of interest is $P(\mu)$, where $\mu = (\mu_1, \ldots, \mu_d)$. We denote the full conditional distribution of μ_i given all other parameters by $P(\mu_i|\mu_{-i})$. One way to implement the MCMC algorithm is the so called Gibbs sampling (Gilks 1996) which can be described as follows:

- Step 1:
  Initialize the iteration counter of the chain ($j = 1$) and the initial values for the parameters $\mu^{(0)} = (\mu_1^{(0)}, \ldots, \mu_d^{(0)})$.
- Step 2:
  Draw a new value $\mu^j = (\mu_1^{(j)}, \ldots, \mu_d^{(j)})$ through successive sampling from the full conditional distributions,

$$\mu_1^{(j)} \sim P(\mu_1|\mu_2^{(j-1)}, \ldots, \mu_d^{(j-1)}),$$

$$\mu_2^{(j)} \sim P(\mu_2|\mu_1^{(j)}, \mu_3^{(j-1)} \ldots, \mu_d^{(j-1)}),$$

$$\vdots$$

$$\mu_i^{(j)} \sim P(\mu_i|\mu_1^{(j)}, \ldots, \mu_{i-1}^{(j)}, \mu_{i+1}^{(j-1)} \ldots, \mu_d^{(j-1)}),$$

$$\vdots$$

$$\mu_d^{(j)} \sim P(\mu_d|\mu_1^{(j)}, \ldots, \mu_{d-1}^{(j)}).$$

- Repeat the second step until convergence.

Assuming that the sampling process is converged after L iterations, the posterior mean of μ can be estimated by MCMC integration:

$$\bar{\mu}_i = \sum_{\ell=1}^{L} \frac{\mu_i^{(\ell)}}{L}.$$

Note that $\bar{\mu}_i$ is simply the sample mean of μ_i which is obtained after L iterations of the Gibbs sampling. In our setting $\bar{\mu}_i$ is the posterior mean of the spectrum at wavelength t_i.

One of the quantities of interest will be a measure of goodness-of-fit, $T^{(m)}(y, \mu^\ell)$ (see Section 13.4). In practice, if we draw L simulations from the posterior distribution of μ we can monitor the value of $T^{(m)}(y, \mu^\ell)$ for each iteration, $\ell = 1, 2, ..., L$ and the posterior mean of $T^{(m)}(y, \mu^\ell)$ is simply $1/L \sum_{\ell=1}^{L} T_\ell^{(m)}(y, \mu^\ell)$.

13.2 Likelihood and Prior Models

As explained in the previous section, we assume a hierarchical Bayesian model in which, at the first stage of the model, the observed spectrum y_i at wavelength t_i, $i = 1, 2, \ldots, n$, is normally distributed with mean μ_i and variance σ_i^2, $y_i = \mu_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma_i^2)$.

Applying Bayes' theorem to our situation, we have as *posterior distribution* for the *spectrum*:

$$\underbrace{P(\mu|y,\theta)}_{\text{posterior}} \propto \underbrace{P(y|\mu,\theta)}_{\text{likelihood}} \times \underbrace{P(\mu|\theta)}_{\text{prior}}, \tag{13.1}$$

or specifying explicitly the observational errors

$$P(\mu|y, \sigma^2, \sigma_M^2, \theta) \propto P(y|\mu, \sigma^2, \sigma_M^2, \theta) \times P(\mu|\theta, \sigma^2, \sigma_M^2). \tag{13.2}$$

The likelihood of the observed spectrum given the parameters in the model is

$$P(y|\mu, \sigma^2) = \prod_{i=1}^{n} N(\mu_i, \sigma_i^2). \tag{13.3}$$

As is argued in Section 13.1, it is not possible to specify the mean function for the observational data as $\mu = h(t)$. We therefore assume that, at the second level of the hierarchical model, the mean of the observational data at wavelength i (μ_i) follows a normal distribution, i.e. we assume that

$$\mu_i = \theta_i + u_i, \tag{13.4}$$

with $u_i \sim N(0, \sigma_{M_i}^2)$ and σ_{M_i} being the *systematic* observational error. This is a crucial step: even if a synthetic spectrum θ is a 'correct' model (in terms of the stellar parameters) following equation (13.4), we assume that, due to the systematic errors, the true spectrum is distributed around θ_i with variance σ_{M_i}.

It follows from equation (13.4) that the *prior distribution* is given by

$$P(\mu|\theta,\sigma_M^2) = \prod_{i=1}^{n} N(\theta_i,\sigma_{M_i}^2). \tag{13.5}$$

Thus, the posterior distribution for the spectrum, given the data and the parameters in the model, is then

$$\begin{aligned} P(\mu|y,\sigma^2,\sigma_M^2,\theta) &\propto P(y|\mu,\sigma^2) \times P(\mu|\theta,\sigma_M^2) \\ &= \prod_{i=1}^{n} N(\mu_i,\sigma_i^2) \times \prod_{i=1}^{n} N(\theta_i,\sigma_{M_i}^2). \end{aligned} \tag{13.6}$$

13.3 Posterior Distribution for the Spectrum

13.3.1 The "Full" Model

Having specified in the previous section the posterior distribution for the spectrum μ, we now will look at the joint posterior distribution of all the parameters in the model:

$$P(\mu,\theta,\sigma^2,\sigma_M^2,\Omega|y) \propto$$
$$\underbrace{P(y|\mu,\sigma^2)}_{\text{likelihood, equation (13.3)}} \times \underbrace{P(\mu|\theta,\sigma_M^2)}_{\text{prior, equation (13.5)}} \times \underbrace{P(\theta|\Omega)}_{\text{distribution of the prior mean } \theta} \times \underbrace{P(\Omega)}_{\text{hyperprior}} . \tag{13.7}$$

In comparison with equation (13.6), we now have to incorporate the distribution of the synthetic spectrum given the atmospheric parameters $P(\theta|\Omega)$, and the prior distribution of Ω, $P(\Omega)$ into the model specified in equation (13.7). Note that equation (13.7) assumes that both σ^2 and σ_M^2 are known, we relax this assumption in Section 13.3.2.

The hierarchical model that we consider at this stage is

$$\begin{array}{ll} y \sim N(\mu,\sigma^2), & \text{likelihood,} \\ \mu \sim N(\theta,\sigma_M^2), & \text{prior,} \\ P(\theta|\Omega), & \text{relationship between} \\ & \theta \text{ and } \Omega \\ P(\mathrm{T}_{eff}), P(\log\,\mathrm{g}), P([\mathrm{Fe/H}]), & \text{hyperpriors.} \end{array} \tag{13.8}$$

Hyperpriors: A literature study for the stellar atmosphere parameters of α Boo is presented in Decin (2000d). In that study, Decin (2000d) found that $\mathrm{T_{eff}}$ ranges between 4060 K and 4628 K, log g goes from 0.90 to 2.60, and [Fe/H] from -0.77 to 0.00. Based on the results reported in Decin (2000d) we can construct the hyperprior distributions for the components in Ω. Since other knowledge about the atmospheric parameters is not available, uniform distributions for the atmospheric parameters over the ranges reported in Decin (2000d) is a reasonable choice for the hyperprior distributions.

Distribution of the Prior Mean: After we have established $P(\Omega)$, we need $P(\theta|\Omega)$ in order to complete the specification of equation (13.8). If the relationship between μ and Ω could be summarised with a deterministic model, say $\mu = h(\Omega, t)$, where $h()$ has some closed form then we could implement the hierarchical model in equation (13.8) as it is and estimate the components in Ω (given the data) with the posterior mean of the hyperprior distributions. In this case the hierarchical model has the following form

$$\begin{array}{ll}
y \sim N(\mu, \sigma^2), & \text{likelihood,} \\
\mu \sim N(\theta, \sigma_M^2), & \text{prior for } \mu, \\
\theta = h(\Omega, t), & \text{relationship between } \theta \text{ and } \Omega \text{ and } t, \\
\mathrm{T_{eff}} \sim \mathrm{Uniform}(T_l, T_u) & \text{hyperprior for } \mathrm{T_{eff}}, \\
\log \mathrm{g} \sim \mathrm{Uniform}(g_l, g_u) & \text{hyperprior for log g,} \\
[\mathrm{Fe/H}] \sim \mathrm{Uniform}(m_l, m_u) & \text{hyperprior for [Fe/H].}
\end{array} \tag{13.9}$$

Since there is no deterministic relationship between θ and Ω and t we cannot specify the mean of the prior distribution using standard methods (i.e. to model θ with linear, generalised linear or non-linear models). This last point is critical since it implies that we need to adopt a two-stage approach in which in the first stage we calculate a collection of models for the synthetic spectrum over a grid of discrete values in Ω, $\Omega^{(1)}, \ldots, \Omega^{(M)}$, and in the second stage we use the models $\theta^{(m)}$, $m = 1, \ldots, M$, as the prior mean of μ in equation (13.5). In this approach the value of Ω (given the data) is not estimated with the posterior means of the hyperprior distributions, but instead we select a model (or models) from the collection of models calculated in the first stage. Thus, our two-stage approach implies that a model selection procedure should be used in order to select the 'best' synthetic spectrum. This issue is discussed further in Section 13.4.

13.3.2 The Reduced Model

For the m'th combination of Ω we calculate $\theta^{(m)}$ and consider a 'reduced' posterior distribution

$$\begin{aligned}
&P(\mu, \theta^{(m)}, \sigma^2, \sigma_M^2 | y) \\
&\quad \propto \quad P(y|\mu, \sigma^2, \theta^{(m)}) \times P(\mu|\theta^{(m)}, \sigma_M^2), \\
&\quad \propto \quad P(\mu|y, \sigma^2, \sigma_M^2, \theta^{(m)}),
\end{aligned} \tag{13.10}$$

where $P(\mu|y, \sigma^2, \sigma_M^2, \theta^{(m)})$ is the posterior distribution of the spectrum μ given $\Omega^{(m)}$, and $\theta^{(m)}$ is the prior mean of μ (equation (13.5)). We note that since for the m'th combination $P(\theta^{(m)}|\Omega^{(m)}, m) = P(\Omega^{(m)}|m) = 1$, the move from equation (13.7) to equation (13.11) is straightforward.

Specification of the Reduced Model

We focus at the posterior distribution of the spectrum μ at wavelength t_i given y_i, $\theta_i^{(m)}$, σ_i, and σ_{M_i}. For the remainder of this section we drop the superscript m and subscript i. The posterior distribution in equation (13.11) can be derived by combining the likelihood and the prior in equation (13.3) and equation (13.5). Since the prior in equation (13.5) is conjugate to the normal likelihood in equation (13.3) the posterior distribution of the spectrum is also normal. Formally, the likelihood and the prior can be expressed respectively by

$$P(y|\mu,\sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right), \tag{13.11}$$

and

$$P(\mu|\theta,\sigma_M^2) \propto \exp\left(-\frac{1}{2\sigma_M^2}(\mu-\theta)^2\right). \tag{13.12}$$

It follows from equation (13.11) and equation (13.12) that the posterior distribution of μ is

$$P(\mu|y,\theta,\sigma^2,\sigma_M^2) \propto \exp\left(-\frac{1}{2\delta^2}(\mu-\theta_1)^2\right), \tag{13.13}$$

which is a normal distribution with mean θ_1 and variance δ^2 given by

$$\theta_1 = \frac{\frac{1}{\sigma_M^2}\theta + \frac{1}{\sigma^2}y}{\frac{1}{\sigma_M^2}+\frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\delta^2} = \frac{1}{\sigma^2}+\frac{1}{\sigma_M^2}. \tag{13.14}$$

This normal/normal likelihood/prior model is discussed in detail by Gelman *et al.* (1995). The result in equation (13.14) means that the posterior mean of the spectrum θ_1 (equation (13.13)) is a weighted average between the synthetic spectrum and the observed spectrum. It can be shown that

$$\theta_1 = \theta + (y-\theta)\underbrace{\frac{\sigma_M^2}{\sigma^2+\sigma_M^2}}_{\text{shrinkage ratio}}. \tag{13.15}$$

Hence, if the SPARE-tag (σ_{M_i}, containing the *systematic* measurement error) is relatively large compared to the STDEV-tag (σ_i, containing the *statistical* measurement error) the posterior mean of the spectrum at wavelength i shrinks towards the observed spectrum at wavelength i and vice versa. If σ_i is relatively large compared to σ_{M_i} the posterior mean of the spectrum shrinks towards the synthetic spectrum.

Contracting the Variance Function

In the previous sections we assume that both σ^2 and $\sigma_{M_i}^2$ are known. In practice, we can estimate σ^2 and $\sigma_{M_i}^2$ and therefore, from a statistical point of view, they should be treated as random variables and not as constant. We use two different approaches to model the variance components in the model.

The first can be considered as an empirical Bayes approach (Carlin and Louis 1996). Using the estimates for the measurement errors, we first specify a model for both σ^2 and $\sigma_{M_i}^2$. We then estimate this model and plug in the predicted values in the hierarchical model of equation (13.8). Specifically, we smooth the data using a hierarchical linear mixed model

(Verbeke and Molenberghs 1997,2000). This allows us to estimate a smooth function for the variance components in a nonparametric fashion as was described in the previous chapter. Figure 13.1 (which we present here, for convenience, once again) displays the measurement errors in band 1A (in log scale). The shrinkage ratio, $\sigma_M^2/(\sigma^2+\sigma_M^2)$, is shown in panel *c*. Note that for wavelengths smaller than or equal to $2.4\,\mu$m the mean of the shrinkage ratio is 0.5 while for wavelengths greater than or equal to $2.58\,\mu$m the mean of the shrinkage ratio increase to 0.87. This means that at the beginning of band 1A the posterior mean is an average between the observed and synthetic spectrum, while the weight of the observed spectrum increase with the wavelength.

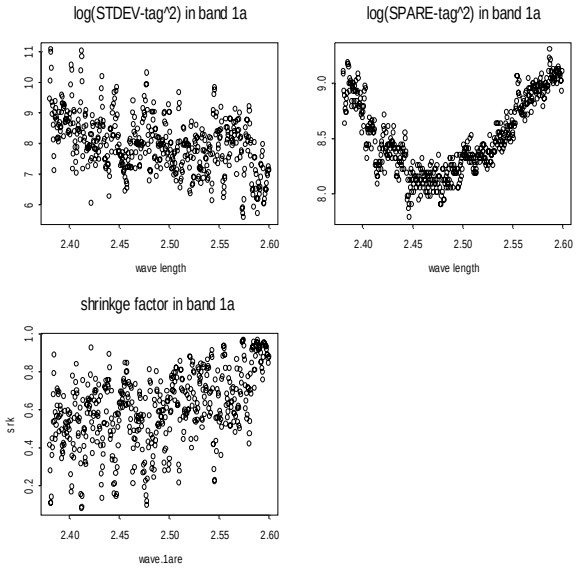


Figure 13.1: *Measurement and systematic errors in band 1A (in log scale). Upper left: STDEV-tag (σ). Upper right: SPARE-tag (σ_M). Bottom left: the shrinkage ratio* $\sigma_M^2/(\sigma^2+\sigma_M^2)$.

The second approach specifies a hyperprior model for the variance components. For example, for 'STDEV-tag' (σ) we assume

$$\sigma_i^2 \sim N(\phi_i, \delta_\sigma^2), \tag{13.16}$$

where ϕ_i is a polynomial function of the wavelength. These models were specified as hyperprior distribution of σ_i^2 and $\sigma_{M_i}^2$ in the hierarchical model. Note that compared to the first approach — in which the variance function is estimated in the first step and used in the hierarchical model in the second step — we now estimate the variance functions and the posterior mean of the spectrum simultaneously.

13.4 Model Selection

13.4.1 Measures for Goodness-Of-Fit

Using equation (13.11), we can predict μ conditional on y and $\theta^{(m)}$. The selection of a "bad" set of stellar parameters $\Omega^{(m)}$ (and thus $\theta^{(m)}$), will result in μ being rather different from y, while a "good" choice of $\Omega^{(m)}$ will bear a nice resemblance. Following Gelman *et al.* (1995) and Carlin and Louis (1996), one can use as measure for the goodness-of-fit a weighted χ^2 statistic given by

$$T^{(m)}(y,\mu) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\mu_i,\theta^{(m)})^2}{\mathrm{var}(y_i|\mu_i,\theta^{(m)})}. \tag{13.17}$$

We note that $T^{(m)}(y,\mu)$ measures the discrepancy between the observed data y and the expected mean accounting to the variability in the model. Both the systematic observational error and the measurement observational error influence $T^{(m)}(y,\mu)$: σ^2 is $\mathrm{var}(y|\mu)$, and since $\mu \sim N(\theta^{(m)}, \sigma_M^2)$ the denominator is influenced from both σ^2 and σ_M^2.

13.4.2 Posterior Predictive Distribution

Criteria for *model selection* are discussed in Laud and Ibrahim (1995) and Gelfand and Ghosh (1998). Both proposed to use the posterior predictive distribution to measure the discrepancy between *replication of the data* and the *observed data* y. This means that based on the specification of the model for the synthetic spectrum θ one can investigate how close the observed data y are to hypothetical data that would have been observed if a new sample was generated under the specific synthetic spectrum model.

Let y_i be the observed data at wavelength t_i and μ_i^ℓ the current value of μ_i at the ℓ'th MCMC iteration. We can simulate n hypothetical replications from the data given the current value of μ_i^ℓ. We denote these values by y_i^{rep}, $i = 1,2\ldots,n$. From these n replications $P(y_{rep}|\mu,\theta,y)$ can be constructed. Note that y_i^{rep} is the spectrum that could have been observed in the next observation of the star. Formally the posterior predictive distribution is given by

$$P(y^{rep}|y) = \int P(y^{rep},\mu,\theta)\,d\mu\,d\theta = \int P(y^{rep}|\mu,\theta,y)P(\mu,\theta|y)\,d\mu\,d\theta. \tag{13.18}$$

Once a sample of replicated data is obtained from the predictive distribution, we can compare y^{rep} with the observed data. If the m'th synthetic spectrum is reasonably accurate the hypothetical replication and the observed data should look the same.

13.4.3 Predictive Model Selection Under Squared Error Loss

A good model for the synthetic spectrum, among the models under consideration, should make a prediction close to what has been observed. Thus, a synthetic spectrum model that leads to a small discrepancy between the replication and the observed data is considered

to be a "good" model. A measure for the discrepancy, based on "squared error loss" is proposed by Laud and Ibrahim (1995):

$$\begin{aligned} L_m^2 &= E[(y^{rep} - y)^T (y^{rep} - y)] \\ &= E\sum_{i=1}^{n} (y_i^{rep} - y_i)^2. \end{aligned} \tag{13.19}$$

Moreover, Laud and Ibrahim (1995) and Gelfand and Louis (1998) showed that L_m^2 can be expressed as a sum of two terms,

$$L_m^2 = \sum_{i=1}^{n} [E(y_i^{rep} - y_i)^2 + \mathrm{var}(y_i^{rep})] \tag{13.20}$$

$$= G(m) + P(m). \tag{13.21}$$

This expected squared error loss of the replicated data L_m^2 can then be used as a criterion for model selection: both Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) suggested to select a model from a collection of M candidates by minimising the expected squared error loss of the replicated data. Hence, the procedure proposed by Gelfand and Ghosh (1998) requires to calculate L_m^2 over the model collection. We notice that

$$L_m^2 = \sum_{i=1}^{n} (\eta_i^{(m)} - y_i)^2 + \sum_{i=1}^{n} \sigma_i^{2(m)}, \tag{13.22}$$

where $\sigma_i^{2(m)} = \mathrm{var}(y_i^{rep}|y, m)$ and $\eta_i^{(m)} = E(y_i^{rep}|y, m)$. We note that in our setting $\eta_i^{(m)} = E(y_i^{rep}|y, \theta^{(m)})$. In equation (13.21), $G(m)$ measures the goodness-of-fit and $P(m)$ is a penalty term that measures the complexity of the m'th model. In our setting the complexity for all candidate models is the same since all synthetic spectra are calculated with the same number of parameters. If we assume that both σ_i and σ_{M_i} are known then the model that minimises $G(m)$ is selected, otherwise the model that minimises L_m^2 is selected.

13.5 Application to the Data

For each model, an MCMC simulation with 11000 iterations (the first 5000 were used as burn-in period) was used to calculate the posterior mean of μ and $T^{(m)}(y, \theta)$.

13.5.1 Measure for the Goodness-Of-Fit

Results for the best 10 models (and also for the models which ranked 15, 20, 25, 50, 75, 100 and 125) are given in Table 13.1. The model with the lowest value of $T^{(m)}(y, \mu)$ is 67 with $T^{(67)}(y, \mu) = 497.6$. Model 125 has the highest value with $T^{(125)}(y, \mu) = 2508$. Posterior means as calculated with equation (13.13) and 95 % credible intervals are presented in Figure 13.2. We notice that the first 8 models produce more or less the same results, but there is a substantial increment in $T^{(m)}(y, \mu)$ from model 104 (which is ranked 9) and onwards. Figure 13.3 shows the density estimate for the posterior distribution of $T^{(m)}(y, \mu)$. The density of $T^{(78)}(y, \mu)$ (for model 78) is located to the right relatively to the densities of the other models 67, 43 and 91 (which are ranked among the top 3). This illustrates once again that these models have a better goodness-of-fit than model 78.

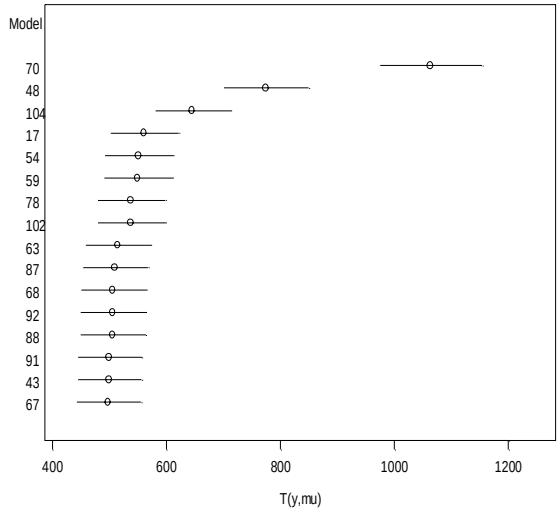


Figure 13.2: Posterior means and 95% credible intervals for $T(y, \mu)$ for 12 models in band 1A.

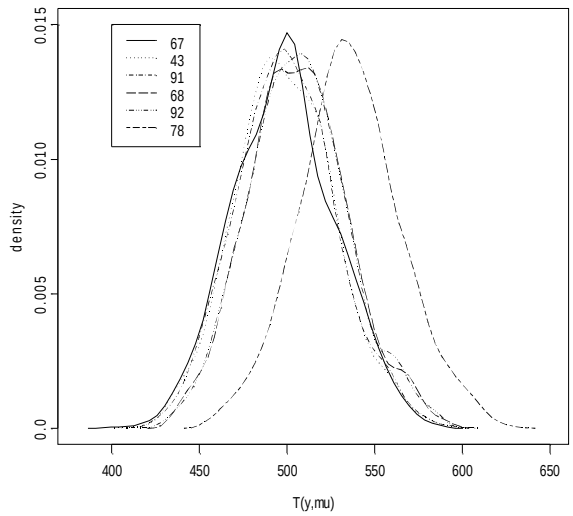


Figure 13.3: Kernel density estimate for the posterior distribution of $T^{(m)}(y, \mu)$.

13.5.2 Expected Squared Error Loss

Model 67 has the smallest value of L_m^2 (3.435×10^6) while model 125 is the one that has the highest value, $L_{125}^2 = 10.57 \times 10^6$. Figure 13.4 and Figure 13.5 show the observed spectrum,

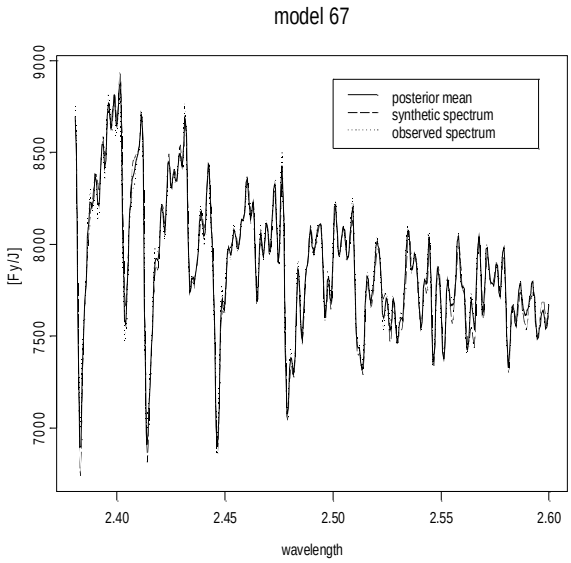


Figure 13.4: Observed spectrum of α Boo (dotted line), synthetic spectrum of model 67 (dashed line) and posterior mean for the spectrum (full line) in band 1A.

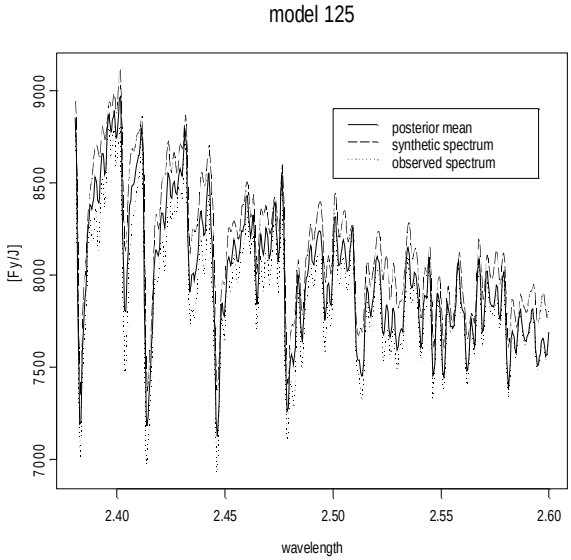


Figure 13.5: Observed spectrum of α Boo (dotted line) , synthetic spectrum of model 125 (dashed line) and posterior mean for the spectrum (full line) in band 1A.

the synthetic spectrum and the posterior mean (calculated with the posterior distribution as given in equation (13.13)) for the spectrum for models 67 and 125 respectively. For model 67 the posterior mean and the observed spectrum lay close to each other along the

whole wavelength range. The discrepancies between the observed and the posterior mean of the spectrum are much more enhanced for model 125. Note how the posterior mean for the spectrum is always between the observed spectrum and the synthetic spectrum. It is also clear that for both models the observed spectrum is more dominant at the end of band 1A. Especially for model 125 in Figure 13.5 the posterior mean and the observed spectrum become closer when approaching the end of the band. Based on this model selection criterion, model 67 with stellar parameters $T_{\rm eff} = 4300\,{\rm K}$, log $g = 1.65\,{\rm dex}$ and $[{\rm Fe/H}] = -0.50\,{\rm dex}$ is selected as giving the 'best' representation of the band 1A ISO-SWS data of α Boo.

13.5.3 Variance Function

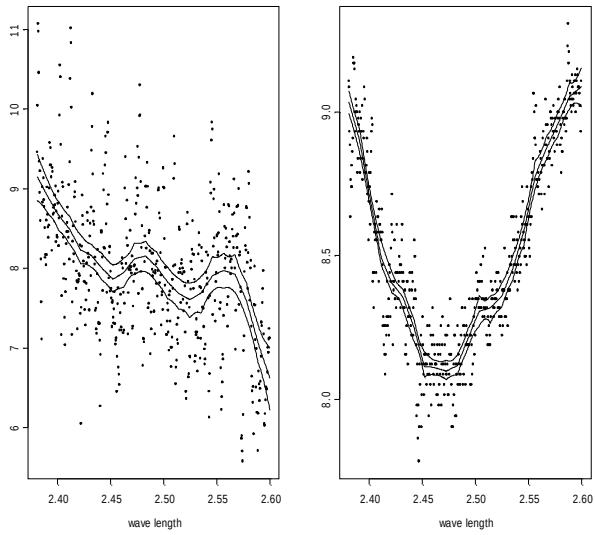


Figure 13.6: *Variance functions. The models were fitted by applying a* linear mixed *model for the data. Left panel: STDEV-tag in band 1A with the estimated model and* 95% *credible intervals. Right panel: SPARE-tag in band 1A with the estimated model and* 95% *credible intervals.*

Figure 13.6 shows the posterior means and the 95% credible intervals for $\log(\sigma^2)$ and $\log(\sigma_M^2)$. Figure 13.7 shows the estimated models obtained for the *parametric polynomial* fitting. Both the non-parametric smoothers and the polynomial models suggest the same pattern. Although, for $\log(\sigma^2)$ the polynomial model smooths out the wave that was estimated by the non-parametric smoother. Both models indicate on the same increasing pattern of the shrinkage factor. Table 13.2 presents the goodness-of-fit measures when polynomial models were used for the variance functions. The patterns are the same as detected from Table 13.1. However, since we observe problems of convergency when the polynomial models are used to model the measurement errors, the first method based on non-parametric linear mixed models is preferred to model the measurement errors. Hence, one estimates the variance function for both σ and σ_M and plug it in the hierarchical

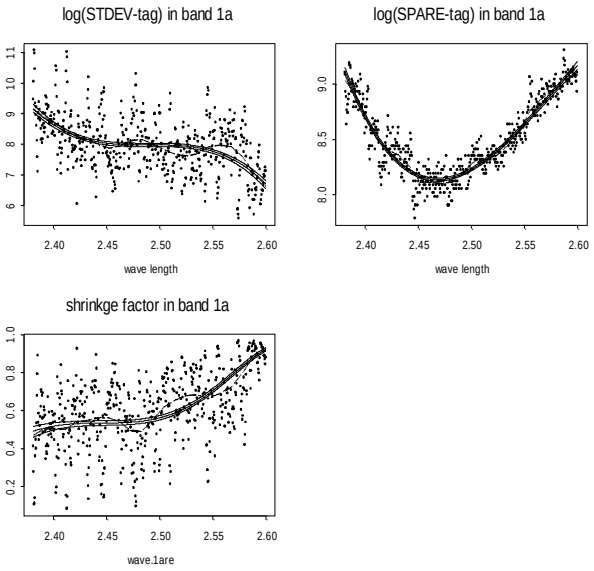


Figure 13.7: *Variance functions. The models were fitted by applying a* linear parametric polynomial *model for the data. Upper left : STDEV-tag in band 1A with the estimated model and* 95% *credible intervals. Upper right: SPARE-tag in band 1A with the estimated model and* 95% *credible intervals. Bottom left: shrinkage factor in band 1A with* 95% *credible intervals.*

model. This reduces computation time substantially and solves the problems related to the convergence of the model.

The advantage to use the polynomial models for the variance function as hyperprior models is that this approach allows us to model the spectrum and the variance function simultaneously. However, this came with a price of convergency problems and computation time that were detected when this approach was used. 11000 iterations took 481 seconds when this approach was used compared with 75 seconds for 11000 iterations when the first approach was used. Another advantage of using the first approach to estimate the variance functions with a linear mixed model is that we estimate the variance function without imposing a parametric structure in advance but give the data to lead the results.

13.6 Discussion

Estimating the stellar atmospheric parameters introduced a problem of model selection in which we had to select a synthetic spectrum from a collection of 125 models. Frequentist methods based on a Kolmogorov-Smirnov test and χ^2 statistics for goodness-of-fit are not able to incorporate the *statistical* and *systematic* measurement errors within the model selection process. We have shown that a hierarchical Bayesian model with a normal model for the likelihood and conjugate normal prior is capable to take into account both the *statistical* measurement error (σ) and the *systematic* measurement error (σ_M). Using

the Bayesian weighted χ^2 statistics to assess the goodness-of-fit, the results based on the band 1A ISO-SWS data of α Boo are as follows: T_{eff} ranges between 4230 and 4370 K, log g ranges between 1.50 and 1.65 dex and [Fe/H] ranges between -0.30 and -0.70 dex. For model selection we have used the predictive squared error loss function. The parameters of the model with the best representation of the ISO-SWS data are $T_{eff} = 4300$ K, log g $= 1.65$ dex and [Fe/H] $= -0.50$ dex.

Including σ and σ_M gives the same results for α Boo in band 1A as in Chapter 10 when the least squares criterion was used for model selection. A sensitivity analysis was performed and the scale of σ_M was reduced by a factor of 1/2. The value of L_m^2 was changed but the results remain the same. Moreover, the same results were obtained when both σ and σ_M were reduced by a factor of 1/2. Further investigation about the performance of the Kolmogorov-Smirnov statistics within the Bayesian frame work is needed. This will allow us the investigate how the measurement errors influence the results when the model selection criterion is based on the Kolmogorov-Smirnov statistic. Furthermore, fitting the model within the framework of Hierarchical Bayesian models will also allow us to investigate the joint distribution of $T^{(m)}(y, \mu)$ and $\beta^{(m)}$ for a specific model of a synthetic spectrum. This could give us deeper insight in the different results between the least squares and the Kolmogorov-Smirnov method that were found in Chapter 10.

Table 13.1: *Measures for the goodness-of-fit T_N for some selected models. The model was estimated using the predicted value of the* linear mixed *model for the variance functions. The expected loss values $G(m)$ are given in units of* 10^6.

Rank	Model	T_{eff}	log g	[Fe/H]	T_N	Expected loss $G(m)$
1	67	4300	1.65	−0.50	497.6	3.423
2	43	4230	1.65	−0.30	499.3	3.435
3	91	4370	1.65	−0.70	499.2	3.429
4	88	4370	1.50	−0.30	505.4	3.447
5	92	4370	1.65	−0.50	505.7	3.466
6	68	4300	1.65	−0.30	505.6	3.462
7	63	4300	1.50	−0.30	509.5	3.457
8	87	4370	1.50	−0.50	514.3	3.479
9	102	4440	1.20	−0.50	537.5	3.512
10	78	4370	1.20	−0.30	537.5	3.492
15	59	4300	1.20	0.00	549.1	3.581
20	54	4300	1.20	−0.15	551.0	3.514
25	17	4160	1.65	−0.50	560.3	3.582
50	104	4440	1.20	−0.15	644.9	3.955
75	48	4230	1.80	−0.30	774.4	4.492
100	70	4300	1.65	−0.00	1063.0	5.496
125	125	4400	1.80	0.00	2508.0	10.57

Table 13.2: *Measures for the goodness-of-fit for some selected models. The model was estimated using the* parametric polynomial *model for the variance function. The expected loss values (L_m^2, $G(m)$, and $P(m)$) are given in units of* 10^6.

Rank	Model	T_{eff}	log g	[Fe/H]	T_N	Expected loss L_m^2,$G(m)$ and $P(m)$
1	67	4300	1.65	−0.50	498.9	5.006, 3.323, 1.683
2	43	4230	1.65	−0.30	500.2	5.017, 3.333, 1.684
3	91	4370	1.65	−0.70	500.9	5.014, 3.331, 1.683
4	88	4370	1.50	−0.30	507.1	5.060, 3.370, 1.690
5	92	4370	1.65	−0.50	507.3	5.072, 3.384, 1.691
6	68	4300	1.65	−0.30	507.1	5.068, 3.378, 1.690
7	63	4300	1.50	−0.30	510.1	5.071, 3.377, 1.694
8	87	4370	1.50	−0.50	514.4	5.110, 3.410, 1.700
9	102	4440	1.20	−0.50	538.0	5.219, 3.498, 1.721
10	78	4370	1.20	−0.30	537.4	5.201, 3.479, 1.722
15	59	4300	1.20	0.00	549.4	5.325, 3.590, 1.735
20	54	4300	1.20	−0.15	549.6	5.255, 3.521, 1.734
25	17	4160	1.65	−0.50	559.5	5.361, 3.617, 1.744
50	104	4440	1.20	−0.15	641.6	6.030, 4.192, 1.837
75	48	4230	1.80	−0.30	763.0	7.245, 5.216, 2.028
100	70	4300	1.65	0.00	1017.0	10.17, 8.019, 2.600
125	125	4440	1.80	0.00	581.0	5.493, 3.723, 1.777

Chapter 14

Conclusions and Further Research

14.1 Modeling Infection Disease Data

The aim of this research was to develop new methodology to estimate the force of infection form seroprevalence data. Muench (1943) wrote: "The thing to do, then, is to find out what curve describes the growth of the summation data and to find its derivative, which will be the rate at which the curve is rising at different ages". In the first part of this thesis we follow this approach. Although, we did not model the summation data but model the binary outcome obtained from serological surveys. Figure 14.1 shows the modeling strategy that we used in the first part of the thesis. All models discussed in Chapter 3-7 are related to existing models. The local polynomial model combines the idea of the polynomial models of Grenfell and Anderson (1985) with the nonparametric approach of Keiding (1991). The fractional polynomial models combine the idea of the GLMs for the force of infection and monotonicity of Farrington's (1991) model. The beta-binomial model is the Bayesian version of Keiding's (1991) model while the Dirichlet process model combines the parametric models (in the prior) with nonparametric estimation of the posterior distribution.

All models discussed in Chapters 3–7 should be seen as part of the toolbox for estimating the force of infection. These models were not introduced as a replacement for existing models. On the contrary, they can be used together with others to give us a more accurate idea about the relationship between age and the force of infection. Materials from Chapters 3–7 are discussed in Shkedy *et al.* (2003a, 2003b).

The local polynomial models, discussed in Chapter 4, were introduced as a nonparametric alternative to the estimation of the force of infection. The selection of the optimal band width, which minimizes the mean square error of the force of infection, overcame the main difficulty of the isotonic regression model proposed by Keiding (1991). Keiding's method requires the selection of a bandwidth for the kernel smoother. However, the bandwidth cannot be chosen by an automatic procedure. Furthermore, the local polynomial model provides a smooth estimator for both the prevalence as well as the force of infection. Thus, the second step in Keiding's method can be surpassed. Although the topic of sample design for serological data is out of the scope of this thesis, it is important to mention that fitting

a local polynomial to the data requires large sample sizes. This is the main disadvantage of the method. For many serological datasets the sample size at older age groups is relatively small. This could lead, especially when local quadratic model is fitted, to a nonmonotone prevalence. Therefore, after applying the *pooling adjacent violators* (PAV) algorithm, it could lead to a zero estimate for the force of infection at older age groups.

Initialy, the fractional polynomial models were developed as a parametric model to estimate the second and the third derivatives that are needed in order to calculate the optimal bandwidth for the local polynomial models. However, these models can stand on their own as a parametric approach to model the force of infection. They provide flexible curve shapes for the force of infection without the need to assume anything about the relationship between the force of infection and age prior to the analysis. Furthermore, we have shown that models with constant, linear, Weibull and log-logistic form for the force of infection can be expressed as fractional polynomials. The models in Chapter 3 are presented without confidence intervals for the force of infection. This is definitely a topic worthy of further exploration.

As we argued in Chapter 6, the beta-binomial model should be seen as the Bayesian version of Keiding's model. As such, it still has the problem of crossing the bandwidth, in the second step of the model, when a smooth force of infection is estimated. We have shown that the beta-binomial model is sensitive to the proportion of noninfected individuals in the population. However, even with these difficulties of the model, one can use the beta-binomial model to estimate the posterior distribution of the average age at infection and to obtain credible intervals for the force of infection.

The Dirichlet process prior models, discussed in Chapter 7, combine parametric models in the prior with nonparametric estimation of the posterior distribution. Therefore, they provide a useful tool for sensitivity analysis. One can fit a parametric model to the data and then use it as a prior model for the force of infection and for the prevalence. By fitting a Dirichlet model with several values of the precision parameter M we can assess to which extent the data support the prior model. As a tool for sensitivity analysis the Dirichlet model is attractive since it does not require to fit several parametric models to the data but rather to fit a single model and then to assess the validity of the proposed model by estimating the posterior distribution of the force of infection.

In Chapter 8, hierarchical Bayesian changepoint model was used in order to estimate the probability to a become hepatitis B carrier. The sensitivity analysis of the model was based on the initial model proposed by Edmunds *et al.* (1993). Therefore, the probability to become a carrier decreases to zero at older ages. Alternatively, one can use a model in which the probability to become a carrier stays constant, at a low level, from a certain age onwards. This will be subject for further investigation. Taking into account that the dataset presented in Edmunds *et al.* (1993) is the only one available on this problem, an effort should be made to collect model data by literature research and by adding more data points to the dataset, especially at older age groups.

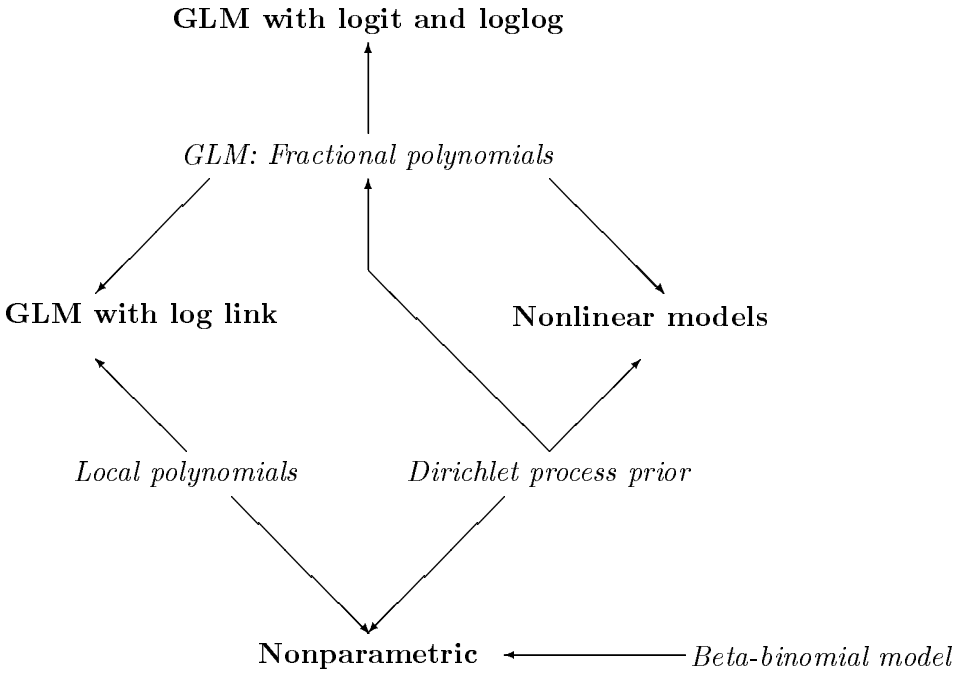


Figure 14.1: *Modeling Strategy.*

14.1.1 Further Research

Incorporating Other Covariates in the Model

The models discussed in the first part of the thesis consist of only one covariate in the model, *age*. As we have pointed out in Chapter 3, when fractional polynomials are used to model the force of infection, one can include covariates in the model using generalized additive model of the form $\text{link}(\pi(a)) = \phi(a) + \text{covariates}$. In this case, $\phi(a)$ is modeled parametrically as a fractional polynomial. The case in which $\phi(a)$ is modeled in a non-parametric way and other covariates include in the model implies fitting a semiparametric model. An alterative for the local polynomial could be a GLMM in which $\phi(a)$ is modeled as a mixed-effects model, i.e.,

$$\text{link}(\pi(a)) = \boldsymbol{X\beta} + \boldsymbol{Zb} + \text{covariates},$$

with the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ as discussed in Chapter 12. The model in this case can be seen as a generalized linear additive model. As Wand (2002) showed, such a model can be implemented using the SAS macro `GLIMMIX`. Of course, one needs to ensure that the estimated prevalence will be a monotone function. This can be done by including an intermediate step during the estimating procedure in which the PAV algorithm is applied in order to ensure monotonicity.

Age-Time Dependent Force Of Infection

In Chapters 3–7, all models were fitted under the assumption that the disease is in a steady state. For the hepatitis A examples we know that this assumption is not likely to be true. Age-time dependent models for the force of infection can be fitted when a series of serprevalence data is available. A possible model could be a proportional hazard model in which the time effect is included an additional covariate. For example, the models discussed by Ades and Nokes (1993) and by Nagelkerke *et al.* (1999). Within the fractional polynomials framework, $\phi(a)$ can be modeled as a fractional polynomial with time as a covariate. In other words, the fractional polynomial could replace the isotonic function in the model proposed by Nagelkerke *et al.* (1999). Local polynomials can be used as well, in this case one needs to iterate between local GLM (in which $\phi(a)$ is estimated) and a global GLM (in which the parametric part is estimated).

Random Effects Models

When the seroprevalence data are clustered, for example when seroprevalence data from different European countries are available, one can account for the intra-cluster correlation by including a random effect for the cluster. A fractional polynomial model for this case has the form

$$\text{link}(\pi(a)) = \phi(a) + b_j, \quad j = 1 \ldots, K,$$

where K is the number of clusters. The cluster-specific force of infection can be calculated by $\ell(a)_K = \phi(a)'\delta(\phi(a))$. Note that for a random intercept model, $\phi(a)'$ does not involves b_j. The population force of infection (or the marginal force of infection) can be derived by $\ell(a)_K = \phi(a)' \int \delta(\phi(a))db$. Multilevel models (Goldstein 1995) which account for both smoothness and cluster heterogeneity are given by

$$\text{link}(\pi(a)) = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{b} + \boldsymbol{Z}_2\mathbf{u}.$$

Here, $\boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{b}$ is the smoother for the force of infection and $\boldsymbol{Z}_2\mathbf{u}$ accounts for cluster heterogeneity. Wang (1998b) fitted such a model to normally distributed data.

14.2 Modeling Astronomical Data

The aim of this research was to investigate the properties of the rebinned data and the Kolmogorov-Smirnov statistic by means of a sensitivity analysis and to develop a tool for model diagnostics when the Kolmogorov-Smirnov statistic is used for model selection. A second goal was to incorporate observational error in the model selection procedure.

Chapter 10 was devoted to sensitivity analysis for both the smoothing procedure (the rebinned data) as well as for the model selection procedure. Smoothing splines were used as an alternative nonparametric estimator for the spectrum and we have shown that using these model selection procedures, the results based on the rebinned data and smoothing splines are the same. Other smoothers can be used as well. In particular, wavelets (Hart, 1997 and Efromovich, 1999) has the advantage to adapt to local feature of curves. As

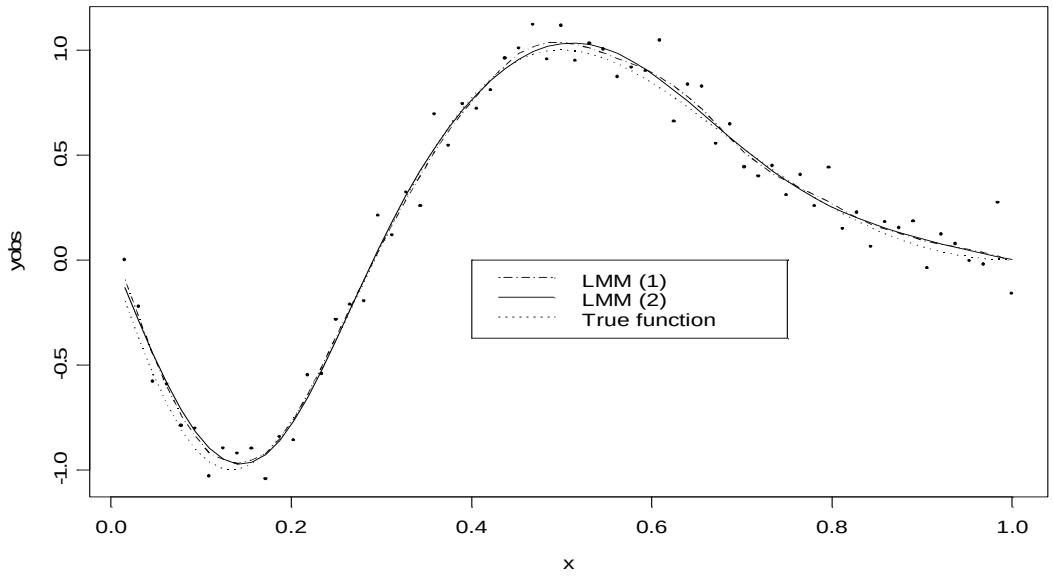


Figure 14.2: *Data, true function (dotted line) and two linear mixed smoothers. LMM(1): smoothing spline (dotted-dashed line). LMM(2): linear mixed model with kernel matrix for the random effects (solid line). Both models were fitted with 35 knots.*

argued by Hart (1997), wavelets have no problem adapting to jump discontinuities. Thus, by using wavelets we will be able to capture the peaks of the observed spectrum more faithfully.

The lack-of-fit test presented in Chapter 11 is based on the methodology proposed by Bowman and Azzalini (1997). The main advantage of Bowman and Azzalini's test is simplicity. Hart (1997) discuss lack-of-fit test based on the order selection procedure. This method can be applied to the ratio between the observed and the synthetic spectra (V_t) and can be used as a "local" version of the Kolmogorov-Smirnov test.

The least squares criterion and the Kolmogorov-Smirnov statistic select the same group of models in band 1A. However, the difference between the two model selection criteria in the other bands should be investigates further.

Modeling the spectrum within the framework of hierarchical Bayesian models allows us to incorporate the observational error into the model. The model selection procedure, based on the squared error loss, leads to the same results as the least squares criterion. We discuss other issues related to the Bayesian model in the following section. Materials in Chapters 10–13 arer discussed in Shkedy *et al.* (2003C) and Decin *et al.* (2003).

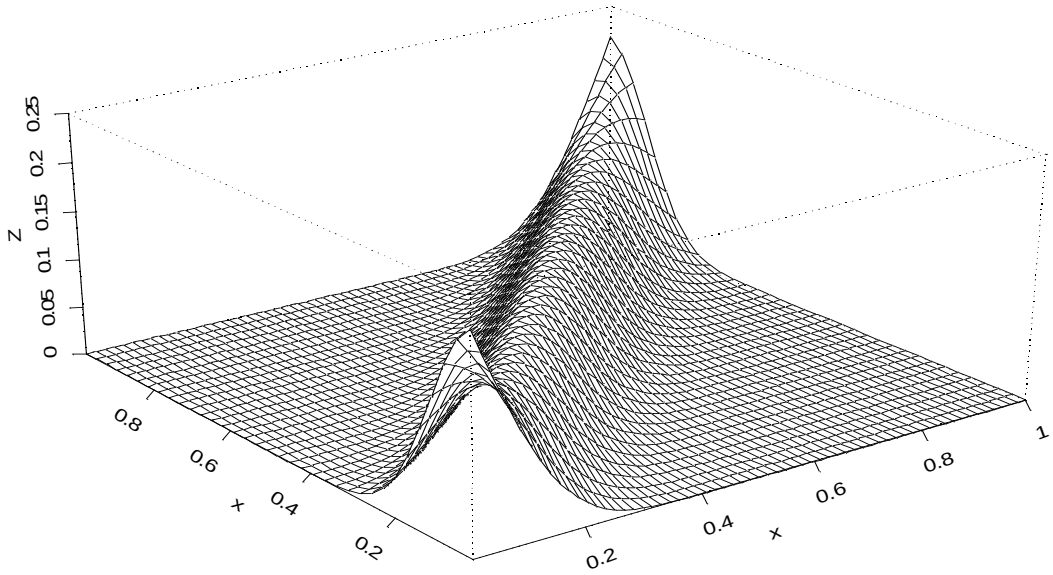


Figure 14.3: *Design matrix for the random effects in LMM(2).*

14.2.1 Further Research

Data Reduction and Bias

From a statistical point of view, the rebinned data should be considered as a nonparametric estimate of the spectrum. The rebinning procedure is the smoothing procedure. We have shown that using (statistical) standard smoothing procedures, such as smoothing splines, leads to the same results in terms of model selection. A linear smoother for the model $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the high resolution spectrum, $\boldsymbol{\mu}$ is the true spectrum and $\boldsymbol{\varepsilon}$ is a random error with $E(\boldsymbol{\varepsilon}) = 0$, can be expressed as $\hat{\boldsymbol{f}} = \boldsymbol{A}\mathbf{y}$. It follows that $E(\hat{\boldsymbol{f}}) = \boldsymbol{A}\boldsymbol{\mu}$, so the bias associated with $\hat{\boldsymbol{f}}$ is $(I - \boldsymbol{A})\boldsymbol{\mu}$. How can we take this bias into account when the rebinned data are used as the input for model selection ? Furthermore, in order to ensure that the synthetic and the observed spectra have the same resolution, the rebinning procedure is applied to the synthetic spectrum as well. How can we adjust the model selection procedure for this source of bias ?

Using a Dirichlet Distribution in the Prior Model of the Spectrum

The final goal of the analysis is to find a point estimate for Ω and to be able to calculate confidence intervals around $\hat{\Omega}$. Let us assume that we will be able to reduce the number of candidate synthetic spectra from M to P. For example, the P best models based on the analysis presented in Chapters 10–13. Thus, the parameter space is reduced to P triples of $(\mathrm{T_{eff}}, \log g, [F_e/H])$. Consider a Bayesian model in which the likelihood is given

by $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2)$ and the prior distribution is

$$\mu_i \sim \mathrm{N}\left(\sum_{p=1}^{P} \delta_p \theta_i^{(p)}, \sigma_{Mi}^2\right),$$

where $0 \leq \delta_p \leq 1$, $\sum_{p=1}^{P} \delta_p = 1$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_P)$ is a P-variate Dirichlet random variable. In such a model, the prior mean is treated as a mixture of P models and $\boldsymbol{\delta}$ represents the mixture probabilities. In this case, the posterior mean of T_{eff} is $\bar{\mathrm{T}}_{\text{eff}} = \sum_{p=1}^{P} \hat{\delta}_p \mathrm{T}_{eff_p}$. Note that, compared to the hierarchical model in Chapter 13, for this model we do not need to select the best model among the P candidate models, but we make inference sbased on the posterior distribution of $\boldsymbol{\delta}$.

Linear Mixed Smoothers

In Chapter 12 we smoothed the observational error using linear mixed models and we have shown that, for a given value of the smoothing parameter, the BLUP is a cubic smoothing spline. Following Speed (1991), we have shown that for the design matrix given in (12.39) the solution for the penalized least squares in (12.14) cab be expressed as linear mixed model. The question that arose is what happens if we replace the design matrix for the random effects, defined in (12.39), by another matrix. Obviously, the BLUP will not be a cubic smoothing spline anymore but the model $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}$ can still be used to smooth the scatterplot. For example, Figure 14.2 shows the example function discussed in Chapter 12 (dotted line) and two smoothers. the first (LMM(1), represented by dotted-dashed line) is a cubic smoothing splines which fitted as LMM using the design matrix for the random effects as defined in (12.39). The solid line show a second linear mixed model (LMM(2)) for which the design matrix is a kernel matrix, shown in Figure 14.3. Note that the solid line is not a cubic smoothing spline but rather a linear smoother. In fact, it is a linear mixed smoother since the model includes random effects, smoothed with the kernel matrix. Figure 14.3 shows that the concept "smoothing a scatterplot with four parameters" holds not just for cubic smoothing spline defined as LMM but in fact it holds in general. Of course, this issue needs to be investigated further.

References

Ades, A.E. and Nokes, D.J. (1993), Modeling age- and time specific incidence from seroprevalence: toxoplasmosis, *American Journal of Epidemiology.* 137, 1022–1034.

Akaike, H. (1974), A new look at the statistical identification model, *IEEE transactions on automatic control*, **19**, 716–723.

Anderson, R.M. and May, R.M. (1985), Age-related changes in the rate of disease transmission: implication for the design of vaccination programmes Oxford University Press, *J. Hyg. Camb.* 94, 365-436.

Anderson, R.M. and May, R.M. (1991), *Infectious diseases of humans: dynamic and control*, Oxford university press, Oxford.

Anderson, R.M. (1982), *Population dynamics of infectious diseases, theory and applications*, Chapman and Hall.

Babu, G.J. and Feigelson, E.D. (1997), *Statistical challenges in modern astronomy II*, Springer.

Bailey, N.T.J. (1988) Simplified modeling of the population dynamics of HIV/AIDS *J. R. Statistic Soc.* A 151, 31-43.

Bailey, N.T.J. (1982), *The biomathematic of the malaria,* Charels Griffin.

Barlow, R.E., Bartholomew, D.J., Bremner, M.J. and Brunk, H.D. (1972), *Statistical inference under order restriction*, New York: Wiley.

Becker, N.G. (1989), *Analysis of infectious disease data.* London: Chapman and Hall.

Beutels, M., Van Damme, P., Aelvoet, W., Desmyter, J., Dondeyne, F., Goilav, C., Mak, R., Muylle, L., Pierard, D., Stroobant, A., Van Loock, F., Waumans, P. and Vranckx, R. (1997), Prevalence of Hepatitis A, B and C in the Flemish Population. *Eur. J. Epidem.* , **13**, 275–280 .

Beutels, M., Van Damme, P., Vranckx, R. and Meheus, A. (1998), The shift in prevalence of hepatitis A immunity in Flanders, Belgium. *Acta Gastro-enterologica Belgica*, **61**, 4–7.

Bowman A.W. and Azzalini A. (1997), *Applied smoothing techniques for data analysis*, Clarendon press, Oxford.

Carlin, B.P., Gelfand, A.E. and Smith, A.F.M. (1992), Hierarchical Bayesian analysis of changepoint problems, *Applied Statistics* **41**, 389-405.

Carlin, B.P. and Louis, T.A. (1996), *Bayes and Empirical bayes methods for data analysis.* Chapman and Hall/CRC, London.

Cleveland, W.S., (1979) Robust locally weighted regression and smoothing scatterplots, *JASA*, 74, 829-836.

Collins, G.W. (1989) *The fundamentals of stellar astrophysics.* Freeman and Co.

Cowles, M.K. and Carlin, B.P. (1996), Markov chain monte carlo convergence diagnostic: a comparative review, *Journal of the American Statistical Association* **91**, 883-904.

Craven, P. and Wahba, G. (1979), Smoothing noisy data with spline functions, *Numer. Math.*, **31**, 377-403.

Davison, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and Their Application.* Cambridge University Press.

De Boor, C. (1978), *Practical guide to splines*,Springer-Verlag.

Decin, L. (2000), *Synthetic spectra of cool stars observed with short wavelength spectrometer: improving the models and calibration of the instrument,* Phd thesis, Leuven.

Decin, L., Waelkens C., Eriksson K., Gustafssom B, Plez B. Sauval A.J., Van Assche W., and Vandenbussche B., (2000a), Formation, structure and evolution of stars: ISO-SWS calibration and the accurate modeling of cool-star atmospheres. I. Method. A&A **364**,1, 137-156.

Decin, L., Waelkens C., Eriksson K., Gustafssom B, Plez B. Sauval A.J., Van Assche W., and Vandenbussche B., (2000b), Formation, structure and evolution of stars: ISO-SWS calibration and the accurate modeling of cool-star atmospheres. II. General results. submitted to A&A.

Decin, L., Vandenbussche B., Waelkens C., Eriksson K., Gustafsson B., Plez B., Sauval A.J., (2001a), ISO-SWS calibration and the accurate modeling of cool-star atmospheres: III. A0-G2 stars submitted to A&A.

Decin, L., Vandenbussche B., Waelkens C., Decin G., Eriksson K., Gustafsson B., Plez B., Sauval A.J., (2001b), ISO-SWS calibration and the accurate modeling of cool-star atmospheres: IV. G9-M2 stars, submitted to Astronomy and Astrophysics.

Decin, L., Shkedy, Z., Molenberghs, G. and Aerts, C. (2003), Estimating stellar parameters I. Nonparametric estimator for the spectrum and lack-of-fit test. submitted to Astronomy and Astrophysics.

Diamond, I.D. (1991), Discussion on : Age-specific incidence and prevalence: a statistical perspective, by Keiding N. (1991), *J. R. Statist. Soc.* A, **154**, 396–398.

Diamond, I.D. and McDonald, J.M. (1992), Analysis of current-status data. In *Demographic Application of Event History Analysis* (eds. J. Trussel, R. Hankinson and J. Tiltan), Ch. 12. Oxford University Press.

Dietz, K. (1993), The estimation of the basic reproductive number of infectious diseases, *Statistical methods in medical research*, **2**, 23–41.

Diggle, P.J (1990), *Time series, A biostatistical introduction* Oxford University Press.

Disch, D. (1981), Bayesian nonparametric inference for effective doses in a quantal-response experiment, *Biometrics* **37**, 713–722.

Edmunds, W.J., Gay, N.J., Kretzschmar , M., Pebody, R.G and Wachmann, H. (2000), The pre vaccination epidemiology of measles, mumps and rubella in Europe: implications for modeling studies, *Epidemiol. infect.*, **125**, 635-650.

Edmunds, W.J., Medley, G.F., Nokes, D.J., Hall, A.J. and Whittle H.C. (1993), The influence of age on the development of the hepatitis B carrier state, *Proc. R. Soc Lond. B* **253**, 197-201.

Edmunds, W.F., Medley, G.F. and Nokes, D.J. (1996), The Transmission Dynamic and Control of Hepatitis B Virus in the Gambia, *Statistics in Medicine,* Vol. 15, 2215-2233.

Efromovich S., (1999), *Nonparametric curve estimation.* New-York: Springer.

Erkanli, A., Soyer, R. and Costello, E.J., (1999), Bayesian inference for prevalence in longitudinal two-phase studies, *Biometrics***55** 1145-1150.

Fan, J. and Gijbels, I. (1996), *Local polynomial modeling and its application.* London: Chapman and Hall.

Fan, J., Heckman, N.E. and Wand, M.P. (1995), Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Statist. Assoc.*, **90**, 141–150.

Farrington, C.P. (1990), Modeling Forces of infection for measles, mumps and rubella. *Statist. Med.*, **9**, 953–967.

Farrington, C.P., Kanaan, M.N., Gay, N.J. (2001), Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Appl. Statist.*, **50**, 251–292.

Ferguson, N.M., Anderson, R.M and Garnett, G.P. (1996), Mass vaccination to control chikenpox: the influence of zoster, *Proc. Natl. Acad. Sci. USA* **93**, 7231–7235.

Ferguson, T.S. (1973), A Bayesian analysis of some nonparametric problems, *The annals of statistics* **1**, 209–230.

Friedman, J. and Tibshirani, R. (1984), The monotone smoothing of scatterplots. *Technometrics*, **31**, 3–39.

Friedman, J.H. and Silverman B.W. (1989), Flexible parsimonious smoothing and additive modeling. *Technometrics*, **26**, 243–247.

Gelfand, A.E. and Ghosh, A.K. (1998), Model choice : A minimum posterior predictive loss approach, *Biometrika* **85**, 1-11.

Gelfand, A.E. and Kuo, L. (1991), Nonparametric Bayesian bioassay including ordered polytomous response, *Biometrika* **78**, 657-666.

Gelfand, A.E., Smith, A.F.M, and Lee, T.M. (1992), Bayesian analysis of constrained parameters and truncated data problems using Gibbs sampling *Journal of the american statistical association.* **87**, 523-532.

Gelfand, A.E., Ecker, M.D. Chrisiansen, C., Mclaughlin, T.J. and Soumerai, S.B., (2000), Conditional categorical response with application to treatment of acute myocardial infraction, *Applied statistics* **49**, 171-186.

Gelman, A , Carlin, J.B, Stern, H.S and Rubin, D.B (1995) *Bayesian data analysis.* Chapman and Hall, London.

Gelman, A. (1996), Inference and monitoring convergence in *Markov Chain Monte Carlo in Practice.* (eds Gliks W.R, Richardson S. and Spiegelhalter D.J.) Chapman and Hall, London, 1996.

Gelman, A. and Rubin, D.B. (1992), Inference from iterative simulation using multiple sequences, *Statistical Science* **7**, 457-511.

Gemerman, D. (1997), *Markov chain monte carlo stochastic simulation for Bayesian inference,* Chapman and Hall.

Geweke, J., (1992), Evaluating the accuracy of sampling based approaches in the calculation of posterior moments, *Bayesian statistics 4* (Bernardo *et al.* Eds.), Oxford university press, page 169–193.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996), *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Goldstein, H. (1995), *Multilevel statistical models.* Arnold, London.

Green, P.J. and Silverman, B.W. (1994), *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall, London.

Greenhalgh, D. and Dietz, K. (1994), Some bounds on estimation for reproductive ratios derived from the age-specific force of infection, *Mathematical biosciences*, **124**, 9–57.

Grenfell, B.T. and Anderson, R.M. (1985), The estimation of age-related rates of infection from case notification and serological data, *J. Hyg. Camb.* **95**, 419-436.

Griffiths, D. (1974), A catalytic model of infection for measles. *Appl. Statist.*, **23**, 330–339.

Grummer-Strawn, L.M. , (1993), Regression analysis of current status data: an application to breast feeding, *Biometrika* **72**, 527–537.

Hart, J.D., (1997), *Nonparametric smoothing and lack-of-fit tests.* New-York: Springer.

Hadler, S.C. (1991), Global impact of hepatitis A virus infection: changing patterns. In *Viral hepatitis and Liver Disease* (eds F.B. Hollinger, S.M. Lemon, H.S. Margolis), pp. 14–20. Baltimore: Williams & Wilkins.

Halloran, M.E. (1998), Concept of infectious disease epidemiology, in Rothman and Greenland (Eds.)*Modern epidemiology, second edition*, Lippincott-Raven, page 529-554.

Halloran, M.E., Watelet, L. and Struchiner, C.J. (1994), Epidemiologic effects of vaccinations with complex direct effects in age-structured population, *Mathematical Biosciences* 121, 193–225.

Halloran, M.E., Cochi, S.L., Lieu, T.C., Wharton, M. and Fehrs, L. (1994), Theoretical epidemiologic effects and mobility effects of routine varicella immunization of preschool children in the united states, *American journal of epidemiology,* 140, 81–104.

Hart, J.D. (1997), *Nonparametric smoothing and lack-of-fit tests.* Springer.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized additive models.* Chapman and Hall.

Healy, M.J.R (1998), Short-term extrapolation of the AIDS Epidemic, *J. R. Statistic Soc.* A 151, 50-61, 1988.

Hinkley, D.V. and Hinkley, E.A. (1970), Inference about the change-point model in a sequence of binomial variables, *Biometrika* **57**, 477-488.

Isham, V. (1998), Mathematical modeling for the transmission dynamic of HIV infection and AIDS: a review, *J. R. Statistic Soc.* A 151, 5-30.

Jewell, N.P., and Van Der Leen, M., (1995), Generalizations of current status data with applications, *Lifetime data analysis*, **1**, 101–109.

Keiding, N. (1991) Age-specific incidence and prevalence: a statistical perspective. *J. R. Statist. Soc.* A, **154**, 371–412.

Keiding, N., Begtrup, K., Scheike, T.H., and Hasibeder, G. (1996), Estimation from current status data in continuous time, *Lifetime data analysis*, **2**, 119–129.

Kuo, L. (1988), Linear Bayes estimation of the potency curve in bioassay, *Biometrika*, **75**, 91–96.

Laud, P.W. and Ibrahim, J.G. (1995), Predictive model selection, *Journal of the Royal Statistical Society - Series B* **57**, 247-262.

Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. (2001), A general framework for constrained smoothing. *Statistical Science* , **16**, 232-248.

McCullagh, P. and Nelder, J.A (1989), *Generalized Linear Models.* Chapman and Hall. New York.

McCulloch, C.E. and Searle, S.R. (2001), *Generalized, linear and mixed models.* Wiley, New-York.

Morgem-Capner, P., Wright, J., Miller, C.L. and Miller, E. (1988), Surveillance of antibody to measles, mumps and rubella by age, *British medical journal* , 770–772.

Muench, H. (1934), Derivation of rates from summation data by the catalytic curve, *Journal of the American statistical association* **00**, 000, 25-38.

Muench, H. (1959), *Catalytic models in epidemiology.* Boston: Harvard University Press.

Nagelkerke, N., Heisterkamp, S., Borgdorff, M., Broekmans, J. and Van Houwelingen, H. (1999), Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statist. Med.*, **18**, 307–320.

O'Neill, P.D. (2001) discussion in Farrington, C.P., Kanaan, M.N., Gay, N.J. (2001), Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data, *Appl. Statist.*, **50**, 251–292.

Pinheiro, J.C. and Bates, D.M. (2000), *Mixed effects models in S and Splus*. Springer.

Qian S.S., Lavive, M. and Stow, C.A. (2000), Univariate Bayesian nonparametric binary regression with application in environmental management, *Environmental and ecological statistics* **7**, 77-91.

Rahman, H.j., Wakfield, J.C., Stephens, D.A. and Falcoz, C. (1999), The Bayesian analysis of pivotal pharmacokinetic study, *Statistical methods in medical research* **8**, 195–216.

Ramsey, F.L. (1972), A Bayesian approach to bioassay, *Biometrics* **28**, 841–858.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988), *Order restricted statistical inference*, Wiley.

Robinson, G.K. (1991), That blup is a good thing: the estimation of random effects, *Statistical science* **6**, 15–51.

Rossini, A.J. and Tsiatis, A.A. (1996), A semiparametric proportional odds regression model for the analysis of current status data, *Journal of the American statistical association* **91**, 423, 713-721.

Royston, P. and Altman, D.G. (1994), Regression using fractional polynomials of continuous covariates : parsimonious parametric modeling. *Appl. Statist.*, **43**, 429–467.

Rutherfurd, E. (1998), *London* Arrow.

Searle, S.R., Casella, G. and McCulloch C.E., (1992), *Variance components*. Wiley.

Shiboski, S.C. (1998), Generalized additive models for current status data. *Lifetime Data Analysis*, **4**, 29–50.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, Ph. and Van Damme, P. (2002), Modeling Age Dependent Force of Infection From Prevalence Data Using Fractional Polynomials. Submitted to Statistics in Medicine.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, Ph. and Van Damme, P. (2002), Modeling Hepatitis A force of infection using monotone local polynomials. Submitted to Applied Statistics.

Shkedy, Z., Decin, L., Molenberghs, G. and Aerts, C. (2003), Estimating stellar parameters II. Hierarchical Bayesian approach. submitted to Astronomy and Astrophysics.

Silverman, B.W. (1985), Some aspects of the spline smoothing approach to non-parametric regression curve fitting, *J.R.Statist.Soc.*, **47**, 1-45.

Smith A.F.M (1975), A Bayesian approach to inference about a changepoint in a sequence of random variables, *Biometrika* **62**, 407-416.

Speed, T. (1991), coments in Robinson (1991) That blup is a good thing: the estimation of random effects, *Statistical science* **6**, 15–51.

Spiegelhalter D.J, Best N.G. and Carlin B.P. (1998), Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models, *Research Report 98-009, Division of Biostatistics, University of minisota*.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der linde, A. (2002), Bayesian measures of model complexity and fit, *J. R. Statistical Soc. B* **64**, 1-34.

Thiry, N., Beutels, Ph., Van Damme, P. and Vranckx, R (2002), The seroepidemiology of primary varicella-zoster virus (vzv) infection in Flanders (Belgium), submitted.

Trimble, V.(1997) Late nights thoughts of s classical astronomer, in Babu, G.J. and Feigelson, E.D. (Eds.) (1997), *Statistical challenges in modern astronomy II,* Springer.

Van Damme, P., Van der Wielen, M., Beutels, M., Van Herck, K., Vercauteren, A. and Meheus, A. Hepatitis B : A public health problem, *Arch Publ Health,* 56, 209–231.

Van Damme, P., Van der Wielen, M., Beutels, M., Van Herck, K., Vercauteren, A. and Meheus, A. (1997) Integration of hepatitis B vaccination into national immunisation programmes. *Brit Med J* 314, 1033–1037.

Verbeke, G. and Molenberghs, G. (1977), *Linear mixed models in practice: a SAS oriented approach.* Lecture notes in statistics 126. Springer.

Verbeke, G. and Molenberghs, G. (2000), *Linear mixed models for longitudinal data.* Springer.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999), The analysis of designed experiments and longitudinal dada using smoothing splines, *J. R. Statistical Soc. C* **48**, 269-311.

Wand, M.P. (2002), Smoothing and mixed models, *Thecnical report, school of public health, Harvard*

Wang, Y. (1998), Smoothing splines models with correlated random errors, *Journal of the American statistical association* **93**, 423, 341-348.

Wang, Y. (1998), Mixed effects smoothing spline analysis of variance, *J. R. Statistical Soc. B* **60**, 159-174.

Whaba, G. (1978), Improper priors, spline smoothing and the problem of guarding against model errors in regression, *Journal of the Royal Statistical Society - Series B* **40**, 364-372.

Whaba, G. (1983), Bayesian confidence intervals for the cross validate smoothing spline, *Journal of the Royal Statistical Society - Series B* **45**, 133-150.

Whaba, G. (1990) *Spline models for observational data.* Society for industrial and applied mathematics, Philadelphia.

Wilks, S. (1962), *Mathematical statistics*, Wiley.

Samenvatting

Gegevens over besmettelijke ziekten modelleren

Edward Rutherfurd (1998) wijdde, in zijn historisch werk *London*, een hoofdstuk aan de pestepidemie die uitbrak in de zomer van 1665. Rutherfurd (1998) beschreef de uitbraak en verspreiding van de ziekte vanuit het perspectief van een arts, Dr. Meredith. Hij verbleef in Londen gedurende die vreselijke zomer, in een poging zijn patiënten te genezen. Aan het begin van de zomer maakte hij zich nog geen zorgen. Rutherford (1998) schreef:

"...Doctor Meredith had not taken much notice of the trouble when a few cases appeared in May. Sporadic visits like this had been a feature of summer in London for centuries...No significant outbreak had occurred, he remained himself, for nearly twenty years and nothing really major since the regime of King James I..."
Enkele weken later, toen hij het overlijdensrapport las, besefte Dr. Meredith dat de ziekte uitgebroken was:

"...The Bill of Mortality was a document produced every week. In two long columns it noted the numbers who had died, of each of some fifty causes, in the city and surrounding parishes of London. Most of the numbers were small. Apoplexy: 1. Dropsy: 40. Infants: 21. But near the top of the second column, the clerk had pointed to one, terrifying number: 1843. And besides it a single, awful word: Plague..."

En, zoals we op pagina 896 kunnen lezen, verspreidde de ziekte zich vlug:

"..By mid-August the Mortality Bill was four thousand a week; by the end of August, six thousand..."

De werkelijke aantallen lagen nog hoger, zoals Dr. Meredith pas later ontdekte. Het overlijdensrapport vertoonde systematische onder-rapportering:

"...Pepys was an official at the Navy Board and, Meredith knew, had access to information of all kind. The real number of deaths is higher than the Mortality Bills shows, Pepys told him. The clerks are falsifying the accounts and some of the poor aren't being counted. The bills show seven and a half last week. And the real figures ? Neared ten, Pepys replied grimly..."

Wie het betreffende hoofdstuk in *London* leest kan zich een beeld vormen van de paniek waaraan de Londense bevolking moet ten prooi gevallen zijn: grote groepen stierven aan de pest, terwijl niemand een idee had hoe de ziekte moest gestopt worden, terwijl zelfs niemand verstond hoe ze zich verspreidde.

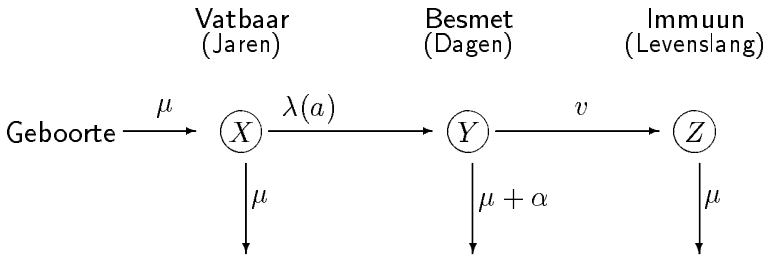


Figure 1: *Grafische voorstelling van het SIR model. Individuen komen het model binnen via de vatbare groep, gaan dan over naar de besmette categorie om dan door te gaan, na herstel, naar de immune klasse. De parameters μ, λ, α en v worden in Sectie 2.1 besproken.*

Het eerste deel van dit proefschrift is gewijd aan de statistische modellering van besmettelijke ziekten. Het accent ligt daarbij op éénmalige seroprevalentie gegevens. De term besmettelijke ziekte of infectieziekte verwijst naar een ziekte die besmettelijk is in de zin dat een geïnfecteerde gastheer een stadium doorloopt, de zogenaamde infectieuze periode, gedurende dewelke hij of zij de ziekte kan doorgeven aan andere, vatbare individuen. Figuur 1 toont de evolutie van een individu doorheen de verschillende stadia. Dit is een typisch patroon voor een zogenaamde kinderziekte (Anderson en May 1991). Een individu komt binnen via de vatbare groep (X), bij de geboorte, geraakt dan besmet (Y), om vervolgens te genezen en levenslange immuniteit te genieten (Z). Uiteraard bestaan er heel wat complexere modellen, waarbij bijvoorbeeld afweer meegegeven via de moeder en/of latentieperioden in rekening worden gebracht. Een zeer fundamentele parameter, nuttig om infectieziekten te beschrijven en te karakteriseren, is de infectiedruk (*force of infection*). De parameter wordt aangeduid met $\lambda(a)$ in Figuur 1. Men gaat ervan uit dat de infectiedruk varieert met leeftijd. Bijvoorbeeld, de infectiedruk voor rubella is veel groter bij kinderen onder de tien jaar dan bij tieners. Het eerste deel van dit proefschrift is gericht op het statistisch modelleren van leeftijdsafhankelijke infectiedruk. We vertrekken daarbij van drie fundamenteel belangrijke manuscripten (Muench 1934, Griffiths 1974, Grenfell en Anderson, 1985). Alle drie modelleren de infectiedruk, uitgaande van het onderliggende katalytische model. De modellering evolueert van het constante model van Muench (1994), over het lineaire model van Griffiths (1974) naar het polynomiale model van Grenfell en Anderson (1985).

In Hoofdstuk 2 wordt een overzicht gegeven van het wiskundig model, gebruikt om de transmissiedynamica van infectieziekten te beschrijven. Het tweede deel van het hoofdstuk legt een verband tussen het wiskundig model wat de mechanismen achter de gegevens beschrijft, aan de ene kant, en de statistische modellen die aangewend worden om de infectiedruk te beschrijven aan de hand van serologische gegevens, aan de andere kant. Het hoofdstuk eindigt met de beschrijving van vijf datasets die gebruikt worden om de analysemethoden te illustreren in latere hoofdstukken.

Hoofdstuk 3 voert fractionele polynomen in (Royston en Altman, 1994), in de context van binaire regressie, als een mogelijk parametrisch model voor infectiedruk.

In Hoofdstuk 4 wordt een niet-parametrische beschouwd, gebaseerd op locale polynomen (Fan en Gijbels, 1996), om de infectiedruk te schatten. Locale lineaire en kwadratische veeltermen leiden tot consistente schatters voor de kans om geïnfecteerd te worden vóór

leeftijd a. Dit geldt evenzeer voor de eerste afgeleide ervan, als functie van a. Daarom kan men via deze techniek een consistente schatter construeren voor de locale infectiedruk. Gebaseerd op de asymptotische verdeling van de schatter voor de infectiedruk kan men de optimale bandbreedte afleiden, waarbij de asymptotische MSE geminimiseerd wordt. In Hoofdstuk 5 wordt een simulatiestudie beschreven die de performantie van de locale polynomiale aanpak vergelijkt met het isotone regressiemodel.

Daar waar Hoofdstukken 3 en 4 de frequentistische filosofie aanhangen, wordt in Hoofdstuk 6 overgestapt naar het hiërarchisch Bayesiaans kader. Niet-lineaire en veralgemeend lineaire modellen worden ingeleid in het eerste deel van het hoofdstuk. Gezien verscheidene parametrische modellen beschouwd worden voor dezelfde gegevens, voeren we modelselectie uit aan de hand van het zogenaamde *deviance information criterion*. Het beta-binomiale model, wat opgevat kan worden als een Bayesiaanse versie van het isotone regressiemodel van Keiding (1991), wordt voorgesteld in het tweede deel van het hoofdstuk.

In het zevende en laatste hoofdstuk over infectiedruk, wenden we een Bayesiaans model aan met Dirichlet *a priori* procesverdeling voor de prevalentie. We tonen aan dat een dergelijke aanpak het gebruik van parametrische modellen (zoals niet-lineaire of veralgemeend lineair modellen) als *a priori* verdeling toelaat voor zowel de prevalentie als de infectiedruk.

In Hoofdstuk 8 tenslotte stellen we een meta-analyse voor die werd uitgevoerd om de kans te schatten om hepatitis B drager te worden. De onderzoeksvraag verschuift dus inderdaad van de infectiedruk naar de kans om drager te worden. Deze kans, in het geval van hepatitis B, wordt verondersteld van leeftijd af te hangen. Ze wordt gebruikt om de transitie te beschrijven waarmee individuen overgaan van de besmette groep naar de groep van dragers. De gegevens werden reeds door Edmunds *et al.* (1993) geanalyseerd. Deze auteurs veronderstelden dat de kans constant bleef gedurende de perinatale periode (vanaf de geboorte tot aan de leeftijd van 6 maanden) om daarna exponentieel te dalen. Wij maken gebruik van een hiërarchisch Bayesiaans breekpunt model (Carlin *et al.* 1992) om het breekpunt te schatten in plaats van vooraf vast te leggen. Dit betekent dus dat het einde van de perinatale periode niet vastligt maar een modelparameter is.

Sterrenkundige gegevens modelleren

In haar Hoofdstuk *"Late night thoughts of a classical astronomer"* (Babu and Feigelson 1997), vatte Virginia Trimble (1997) haar gedachtengang samen alvorens naar een gemeenschappelijk congres voor statistici en sterrenkundigen te gaan, met als thema *Statistical Challenges in Modern Astronomy* en gehouden in Penn State University, in juni 1996, als volgt samen:

"*I arrived at Penn State with a prediction at hand. Nobody is going to learn anything at this meeting*".

De laatste zin in haar hoofdstuk is:

"*How can we foster collaborations between statisticians and astronomers that will be attractive to both in the sense of advancing basic knowledge in both fields so that each collaborator has something significant to add to his CV at the end ? Participants nodded solemnly and promised to Think About it All*".

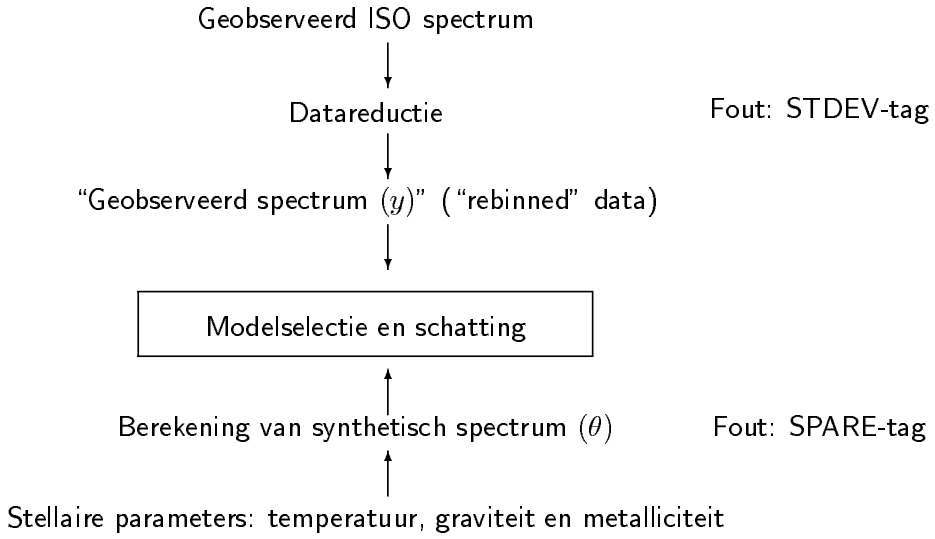


Figure 2: *Schatting van stellaire parameters. 'STDEV-tag' verwijst naar de meetfout t.g.v. rebinning. 'SPARE-tag' verwijst naar de onzekerheid afkomstig van de berekening van een synthetisch spectrum. Datareductie verwijst naar het proces waarbij de ruwe, niet-binned gegevens worden omgezet in binned, interpreteerbare gegevens.*

Het tweede deel van dit proefschrift is een verslag van een onderzoek wat deze samenwerkingsgedachte in de praktijk brengt. We schatten stellaire parameters van een koele ster. De schattingsmethode bestaat uit een vergelijking van theoretische en geobserveerde spectra van een ster. Het berekenen van het theoretisch spectrum is een tijdsintensieve bezigheid. Concreet wordt de modelmatige atmosfeer van de ster berekend en het theoretisch spectrum wordt berekend uit de relatieve transfert vergelijkingen. Een schematische weergave van het hele proces wordt gegeven in Figuur 2. Een theoretisch spectrum bestaat uit tientallen parameters die de structuur van de beschouwde ster beschrijven. Eens een synthetisch spectrum berekend, dient het vergeleken te worden met het geobserveerde spectrum. Het is de bedoeling het synthetisch spectrum te kiezen wat de beste modelaanpassing geeft en van daarna de parameters te schatten van de ster waarop dit model gebaseerd is.

Trimble (1997) formuleerde drie vragen die verband houden met modelaanpassing. We gebruiken deze vragen om de structuur van het tweede deel van het proefschrift aan te geven.

1. *Hoe vinden we onze weg, in een meerdimensionale parameterruimte, naar de beste aanpassing ?*

   In dit proefschrift volgen we de aanpak van Decin (2000), de zogenaamde *zoom in* method. De methode concentreert zich op de drie meest belangrijke parameters: de effectieve temperatuur, de graviteit, en de metalliciteit. De andere parameters worden vastgehouden. Op die manier ontstaat een reeks synthetische spectra

waarbij datgene gekozen wordt wat de beste modelaanpassing geeft. Daarna wordt een nieuwe reeks synthetische spectra berekend op basis van een gevoeliger rooster omheen de effectieve temperatuur, graviteit en metalliciteit. Het nieuwe rooster wordt vastgelegd op basis van de resultaten uit de vorige stap. Eens de nieuwe collectie van spectra berekend, dienen ze vergeleken te worden met het waargenomen spectrum om de beste aanpassing te vinden.

2. *Waarmee vervangen we de chi-kwadraat als toets voor modelaanpassing ?*

   Decin (2000) stelde het gebruik van de Kolmogorov-Smirnov statistiek voor om aan modelselectie te doen. Hoofdstuk 9 schetst de sterrenkundige en statistische problemen bij het schatten van stellaire parameters. In Hoofdstukken 10 en 11 ligt de klemtoon op sensitiviteitsanalyse. Men dient zich rekenschap te geven dat de hoge resolutie waargenomen spectra niet gebruikt worden in de analyse omdat ze niet dezelfde resolutie hebben als de synthetische spectra. Om een dergelijke vergelijking toch mogelijk te maken worden de gegevens eerst aan zogenaamde *rebinning* onderworpen. Een kubische smoothing spline methode om een dergelijke vergelijking mogelijk te maken wordt voorgesteld in Hoofdstuk 10. De nieuwe methode wordt vergeleken met de klassieke rebinning. We vergelijken ook de performantie van de Kolmogorov-Smirnov statistiek met deze van het kleinste kwadraten criterium. In Hoofdstuk 11 tonen we aan dat de Kolmogorov-Smirnov statistiek een locale versie heeft in de familie van toetsen voor aanpassing.

3. *Veronderstel dat het model wat men gebruikt om aanpassing te bereiken zelf onzerkheden vertoont, zoals in atomaire gegevens en spectrale lijnen, hoe kan dit dan ingebed worden in de foutschattingen bij zowel Bayesiaanse als frequentistische methoden ?*

   Hoofdstukken 12 en 13 zijn gewijd aan Bayesiaanse analyse. In Hoofdstuk 12 bespreken we het gebruik van lineaire gemengde modellen als scatterplot smoothers. De methode wordt toegepast op waarnemingsfouten. In Hoofdstuk 13 wordt een hiërarchisch Bayesiaanse methode voorgesteld voor het spectrum. Het model houdt rekening met twee bronnen van onzekerheid: (1) STDEV-tag, meetfout ten gevolge van het waarnemen en (2) SPARE-tag, onzekerheid op het berekende, synthetisch spectrum. STDEV-tag komt voor als variantie in de likelihood, terwijl SPARE-tag vorkomt als variantie in de *a priori* verdeling van het werkelijke spectrum. We tonen aan dat het *a posteriori* gemiddelde voor het spectrum kan uitgedrukt worden als een gewogen gemiddelde van het waargenomen en het synthetisch spectrum. We beschouwen 125 synthetische spectra en we hebben dus nood aan een selectieprocedure, gebaseerd op de predictieve verdeling van het spectrum, om het best aanpassende model te selecteren.